Computational Approaches to Automatic Poetry Generation and Evaluation: A Survey

ILYA KOZIEV*, SberAI, Russia
LEONID SINEV, SberAI, Russia

This survey provides a comprehensive synthesis of research on automatic poetry generation and evaluation from 2017 to 2025. We examine computational approaches that leverage pre-trained LLMs, multimodal architectures, and specialized algorithms for handling poetic constraints such as meter, rhyme, and stanza structure. In addition to surveying generative methods, we analyze practices in data engineering, including corpus construction, annotation, and preprocessing tools tailored to poetry. Evaluation receives particular attention: we review automatic metrics, LLM-as-a-judge methods, and human-centered protocols, discussing their strengths and limitations. Compared with prior surveys, our work emphasizes (1) the dominant role of LLMs in both generation and evaluation, (2) a taxonomy of poetry generation tasks categorized by interaction modality, (3) systematic coverage of dataset engineering challenges, and (4) a comprehensive analysis of automatic and human evaluation approaches, highlighting their drawbacks. By consolidating advances across diverse research lines, we show how poetry serves as a challenging benchmark for controllable text generation, multimodal grounding, and human-aligned evaluation. Building on this perspective, the survey summarizes current methods and open challenges in the generation, control, and evaluation of poetic and lyrical text.

JAIR Track: Surveys

JAIR Associate Editor: Valerio Basile

JAIR Reference Format:
Ilya Koziev and Leonid Sinev. 2026. Computational Approaches to Automatic Poetry Generation and Evaluation: A Survey. *Journal of Artificial Intelligence Research* 86, Article 4 (May 2026), 76 pages. DOI: 10.1613/jair.1.20584

1 Introduction

Text generation has become one of the most dynamic areas of research in natural language processing (NLP) and artificial intelligence (AI). While recent years have seen the rapid progress of large language models (LLMs) in general-purpose tasks, poetry generation represents a particularly demanding domain. Unlike ordinary text generation, poetry requires not only semantic coherence but also sensitivity to form, rhythm, rhyme, imagery, and cultural nuance. As such, it serves as a challenging benchmark for controllability, multimodal grounding, and evaluation — issues of broad relevance to the NLP and generative AI communities.

Research on automatic poetry generation has expanded considerably since 2017, encompassing diverse approaches such as rhyme- and meter-controlled generation, cross-lingual poetic translation, multimodal conditioning (e.g., image-to-poem), and interactive human-in-the-loop systems. These efforts provide insights into fundamental challenges in generative modeling, including dataset scarcity, alignment with human preferences, reproducibility of evaluation, and the integration of symbolic constraints with neural architectures.

*Corresponding Author.

Authors' Contact Information: Ilya Koziev, ORCID: 0009-0004-4447-132X, inkoziev@gmail.com, SberAI, Moscow, Russia; Leonid Sinev, ORCID: 0000-0003-2097-9036, leonid.sinev@yandex.ru, SberAI, Moscow, Russia.

DOI: 10.1613/jair.1.20584

For the purposes of this survey, we adopt an operational definition of poetry grounded in literary theory and computational practice. Following Attridge (1995), poetry can be treated as text in which formal properties such as rhythm, rhyme, meter, line structure, or patterned repetition play a central role in meaning construction, which distinguishes it from ordinary prose. This view is consistent with formal accounts of poetry that emphasize constraint and structure, particularly in fixed-form verse (Fussell 1979, page 127).

Accordingly, this survey focuses on systems that explicitly model or evaluate at least one poetic constraint, such as meter, rhyme, stanza structure, or line-level organization. We include both traditional poetic forms (e.g., sonnets, haiku, quatrains, limericks) and closely related forms such as song lyrics, including rap, where similar formal constraints apply. Texts that lack explicit poetic structure or are treated purely as unrestricted prose are outside the scope of this review.

Poetry is not only of literary interest but also a useful testbed for challenges in generative AI. It combines structural constraints such as rhyme and meter with requirements for meaning and style, making it a compact example of controllable text generation. The lack of large, consistent poetry datasets mirrors data limitations in many other domains. Evaluating poems is also difficult, raising broader questions about human-aligned evaluation and the use of LLMs as judges. For these reasons, the survey is intended to support both researchers in generative poetry and the wider NLP and AI communities interested in control, alignment, and evaluation.

Our contributions are as follows:

- We analyze and systematize research on generative poetry published between 2017 and 2025, extending and updating earlier surveys such as Gonçalo Oliveira (2017).
- We introduce a taxonomy of poetry generation tasks, organized by user control modality (structural, semantic, stylistic, multimodal), highlighting how different interaction formats shape the generation problem.
- We provide systematic coverage of data engineering practices, including open datasets, preprocessing pipelines, and tools for rhyme, stress, and scansion — resources that remain underexplored in prior surveys.
- We analyze model architectures and decoding strategies, with special attention to the role of LLMs in shaping the state of the art.
- We conduct a comprehensive review of evaluation practices, comparing automatic metrics, LLM-based evaluation methods, and human protocols, and we highlight their limitations.
- Finally, we identify open challenges and future research opportunities that connect poetry-specific problems to broader issues in controllability, alignment, and evaluation in generative AI.

By consolidating and synthesizing recent advances, this survey is intended not only as a technical reference for computational creativity but also as a resource for the wider NLP and AI communities seeking to address fundamental challenges in creative and controllable text generation.

Given the substantial engineering innovations in recent years, this survey aims to provide an updated synthesis of methods, models, and results that have not been adequately covered in previous surveys. We begin by briefly summarizing these earlier works to contextualize our contribution.

Compared to previous surveys, our work makes several distinct contributions. Gonçalo Oliveira (2017) provided valuable coverage of rule-based and pre-neural approaches but offered limited synthesis on datasets, evaluation practices, and task taxonomies. Our survey extends this line of work by presenting a high-level analysis of datasets (Section 3.1), task formulations (Section 2), and assessment methods (Section 5), with particular emphasis on how LLMs now shape both poem generation and evaluation. More recently, Franceschelli and Musolesi (2024b) provided an in-depth discussion of creativity and figurative language, but expressed skepticism about the role of transformer-based architectures (Vaswani et al. 2017) in creative systems. In contrast, we highlight the dominant role of transformers in state-of-the-art generative poetry systems (Section 4.2.3). Furthermore, while their review takes a primarily theoretical perspective on creativity and evaluation, our survey approaches evaluation from an

engineering standpoint, analyzing concrete metrics and methods used in published systems. These perspectives are complementary: together they provide both conceptual and practical insights, whereas our contribution is to consolidate advances in methods, data, and evaluation with a focus on how LLMs reshape the field.

Our analysis focuses primarily on research articles published between 2017 and 2025, drawing in particular on works indexed in the ACL Anthology[1]. Supporting statistics are reported in Appendix A. In a small number of cases, we also include papers published before 2017 when this helps to contextualize trends in areas with limited prior work.

This survey is organized into the following sections. Section 2 introduces a taxonomy for classifying poetry generation systems. Of particular interest in this part of the survey are the new methods of controlling text generation that have emerged in recent years with the rise of multimodal language models (LMs), such as generating song lyrics from audio or generating poems based on a given image. Each subsection focuses on a specific group of tasks, categorized by the tools they provide for managing generation or the input data used to form the final text. We discuss tasks such as keyword-conditioned poem generation (Section 2.2), instructive prompting (Section 2.3), poetic translation (Section 2.4), poetry style imitation and parody generation (Section 2.5), prose-to-poem conversion (Section 2.6), acrostic generation (Section 2.7), image-conditioned poem generation (Section 2.8), melody-to-lyric generation (Section 2.9), controlled generation approaches with other modalities (Section 2.10), and poem writing assistance (Section 2.11). Section 3 examines approaches to preparing training data, addressing challenges such as scraped data quality control and tasks specific to poem text preparation, including syllabification (Section 3.2), stress assignment (Section 3.3), rhyme detection (Section 3.4), and part-of-speech tagging (Section 3.5). In Section 4 of this survey, we aim to highlight both mainstream architectures used for poetry generation and contemporary alternatives that, while underrepresented, hold potential interest for researchers. This section explores the design of poetry generation systems through three orthogonal factors: (1) text representation, for example, character-level (Section 4.1.1), syllable-level (Section 4.1.2), word-level (Section 4.1.3), or phonetic (Section 4.1.4) representation; (2) the architecture of the LM (Section 4.2); and (3) the LM decoding algorithm (Section 4.3). Each factor requires a choice among alternatives. Selecting the right combination often depends on the researcher's intuition and domain expertise. Section 5 focuses on the assessment of generated poems, emphasizing the importance of research in this area. We provide a high-level analysis of popular approaches to automatic evaluation, then examine individual metrics and their calculation methods, and offer recommendations for their use. After that, we review specific metrics, including meter, rhyme, and form (Section 5.1.2); grammaticality and meaningfulness (Section 5.1.3); diversity (Section 5.1.4); novelty and creativity (Section 5.1.5); naturalness (Section 5.1.6); and others. In a separate section (Section 5.2), we systematize approaches to human assessment of generative poetry. Finally, in Section 6, we outline open challenges in the field of generative poetry and propose a research agenda for the coming years.

2 Generative Poetry Tasks

In this survey, we categorize generative poetry tasks according to the type of user control exerted during the generation process. Control can take structural forms, such as specifying rhyme schemes or metrical patterns; semantic forms, such as guiding themes, keywords, or emotional tones; stylistic forms, such as imitating the voice of a specific poet or performing cross-lingual poetic translation; or multimodal forms, where inputs like images or melodies influence poetic output. This taxonomy emphasizes how different interaction modalities shape the generation problem, rather than grouping systems by underlying architectures. By organizing tasks around user control, we highlight the ways in which poets, researchers, and end-users can steer creative output, providing a coherent framework that unifies diverse approaches under a common perspective. Table 1 presents the result of using this taxonomy construction principle based on works published between 2017 and 2025.

[1] https://aclanthology.org/

Table 1. Generative poetry tasks in publications (2017–2025). Implementation complexity reflects the difficulty of building systems given available datasets. Target languages indicate natural languages with implementations.

Task	Implementation complexity	Target languages	Papers
First line continuation (Section 2.1)	Lowest (none in case of decoder LM)	Chinese, English	Aguiar and Liao (2019), Badura et al. (2022), Boggia et al. (2022), D'Souza and Mimno (2023), Köbis and Mossink (2021), D. Liu et al. (2018), Lo et al. (2022), Ram et al. (2021), Y. Song (2022a), Uthus et al. (2022), and K. Yang and Klein (2021)
Keywords-conditioned poem generation (Section 2.2)	Low (requires keyword extraction pipeline)	Chinese, English, French, Portuguese, Spanish	Boggia et al. (2022), Chang et al. (2023), Gonçalo Oliveira, Mendes, et al. (2017), Z. Guo, X. Yi, et al. (2019), Hämäläinen, Alnajjar, and Poibeau (2022), Hegade et al. (2021), and Z. Liu et al. (2019)
Instruction-tuned prompting (Section 2.3)	Medium to high (requires human involvement or LLM-powered synthetic prompt generation pipeline)	Chinese, English, Russian, Vietnamese	Agirrezabal and Oliveira (2024), Chakrabarty, Padmakumar, et al. (2022), Y. Chen, Gröner, et al. (2024), D'Souza and Mimno (2023), Davis (2024), Horishny (2022), Z. Hu et al. (2024), Huynh and Bao (2024), Koziev and Fenogenova (2025), Porter and Machery (2024), Shalevska (2024), Tian and Peng (2022), Walsh, Preus, et al. (2024), C. Yu et al. (2024), and R. Zhang and Eger (2024)
Poetic translation (Section 2.4)	High (there is little quality data for training, the assessment approach must take into account many criteria simultaneously)	Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish	Cespedosa Vázquez and Mitkov (2023), Chakrabarty, Saakyan, et al. (2021), A. Chen et al. (2024), Cho et al. (2025), Ghazvininejad, Y. Choi, et al. (2018), Moreno (2021), Mukherjee et al. (2024), Resende and Hadley (2024), W. L. Song et al. (2023), and S. Wang, Wong, et al. (2024)
Poetry style imitation and parody (Section 2.5)	Medium to high (there are no datasets with parodies or imitation of style)	Chinese, English, Russian	Chang et al. (2023), Nalci et al. (2025), Sawicki, Grzes, Goes, Brown, Peeperkorn, and Khatun (2023), Sawicki, Grzes, Goes, Brown, Peeperkorn, Khatun, and Paraskevopoulou (2023), Shihadeh and Ackerman (2020), and Tikhonov and Yamshchikov (2018a)
Prose-to-poem (Section 2.6)	Medium (requires LLM-powered summarization pipeline)	English, French, Persian	Khanmohammadi et al. (2023) and Van de Cruys (2020)
Acrostics (Section 2.7)	Medium (no ready-to-use datasets available)	English	Agarwal and Kann (2020), Fan et al. (2019), and Fedchin et al. (2025)
Image-conditioned poem generation (Section 2.8)	Medium	Chinese, English	Jiang et al. (2025), B. Liu et al. (2018), L. Liu et al. (2018), Loller-Andersen and Gambäck (2018), Nalci et al. (2025), Xu et al. (2018), and D. Zhu et al. (2024)
Melody-to-lyric (Section 2.9)	Medium (few datasets available, web scraping and data preprocessing required)	Chinese, English	Y. Chen and Lerch (2020), Elzohbi and R. Zhao (2024), Lu, J. Wang, et al. (2019), Tian, Narayan-Chen, et al. (2023), Vechtomova et al. (2021), and Y. Yu et al. (2021)
Other controlled generation tasks (Section 2.10)	The complexity varies greatly for different subtasks	Chinese, English, Spanish	Cerdas (2025), H. Chen et al. (2019), McCormack et al. (2024), Røstvold and Gambäck (2020), X. Song et al. (2025), Thölke et al. (2024), Tonra et al. (2019), Wöckener et al. (2021), X. Yang et al. (2018), and X. Yi, R. Li, C. Yang, et al. (2020)

We begin by detailing common prompting strategies before discussing their application to specific generative tasks.

The prevalence of instruction-tuned transformer models has led most modern systems to adopt a single-turn, prompt-to-poem paradigm. This approach leverages widely available LLMs and inference tools. It also requires minimal user training, as the chat-based interface has become a de facto standard for generative AI. Within this paradigm, we can distinguish several core prompting strategies, which often correspond to distinct historical tasks:

- Generating from a keyword or set of keywords.
- Continuing from a given first line provided by the user or generated by another LM.
- Following a complex instruction (which may include keywords, first lines, or other constraints).

Historically, systems built for keyword-based generation or line continuation predate modern instructional LLMs. They often relied on more straightforward and easier-to-implement controls within recurrent neural architectures. The technical implementation for each strategy also dictates different data engineering requirements. For instance, fine-tuning a decoder-only transformer on a poetry corpus is a natural fit for a line-continuation task, as it requires no special instruction dataset. In contrast, building a keyword-based system necessitates either a ready-made dataset with poems marked by genre, era, style,[2] or building a pipeline to extract keywords from text using standard NLP methods, though this offers greater flexibility by allowing control over attributes like genre, emotion, or authorial style.

Due to these fundamental differences in their historical development and technical underpinnings, we describe keyword-based and line-continuation approaches separately from modern instructional prompting. The advent of powerful, publicly available LLMs like GPT-2 (Radford et al. 2019) and T5 (Raffel et al. 2020) has enabled all these strategies to be implemented under a unified prompting paradigm.

In the following subsections, we will discuss the approaches mentioned and also discuss other options, highlighting their challenges for researchers and practitioners. In Section 2.10 we highlight some notable examples of approaches to certain control modalities that do not neatly fit into other categories, as they often encompass unique or hybrid mechanisms that defy straightforward classification. Finally, in Section 2.11 we discuss interactive systems that can implement different approaches to generation control, providing a user interface to access the relevant tools.

We'll start with the easiest approach to implement — poem completion based on the first line given.

2.1 First Line Continuation

The task of continuing a poem from a given first line simplifies system implementation, particularly when using decoder transformer models. These models can operate directly in text completion mode after being trained on poetic texts, with the first line provided either by the user or selected from existing human-written poems. The initial line implicitly provides structural constraints — such as meter or syllable patterns — that guide the LM in generating subsequent lines that adhere to poetic form requirements.

While this approach offers implementation advantages, it limits user control over other aspects of generation compared to instructive prompting methods described in Section 2.3. Users cannot explicitly specify additional constraints or stylistic preferences beyond what is embedded in the initial line.

Using human-written first lines: The initial line can be sourced from a corpus of human-written poems or provided directly by the user. While simple to implement, this approach carries risks of line-level plagiarism and, on a broader scale, potential reproduction of memorized content if the generative model lacks sufficient originality. Another potential issue may be the inability of the generative model to complete the poem in the intended style, especially in cases where the model's training data did not contain enough poems in the required style. Examples

[2]See examples of such datasets in Section 3.

of this method include: interactive poetry generation in Uthus et al. (2022); Chinese poem generation in D. Liu et al. (2018); GPT-2 fine-tuning for English poetry in Köbis and Mossink (2021); controlled generation testing for couplet completion in K. Yang and Klein (2021); haiku generation in Aguiar and Liao (2019); and generation of Shakespeare-styled poems with the first line provided by the user in Badura et al. (2022). D'Souza and Mimno (2023) used a first-line completion approach to test the ability of LLM Pythia (Biderman et al. 2023) and GPT-2 to memorize and reproduce poem texts.

Employing a dedicated first-line generator: A separate LM, potentially with a different architecture or reduced capacity, can be trained specifically to generate the first line. This increases the number of models involved in the generation process, raising resource requirements for inference. For instance, Boggia et al. (2022) describe such an approach. Additionally, Lo et al. (2022) propose a two-stage generation technique for limericks. In the first stage, a fine-tuned GPT-2 model generates the first line. The second stage uses a separate model trained on a corpus of limericks with reverse-ordered (right-to-left) tokens within each line, preserving the overall line order. This reverse tokenization facilitates rhyme selection, ensuring adherence to the AABBA[3] rhyme scheme.

Collaborative line-level generation: The system generates multiple alternative variants of the next line, from which the user selects one. This process iterates, with the system generating subsequent lines based on the extended text. An example of this approach is described in Ram et al. (2021), where the T5 transformer model generates the next line based on user-provided initial lines. Uthus et al. (2022) implemented an interactive system in which the user can enter a string and receive continuation options with specified properties (number of syllables, rhyme), from which to choose the best one.

Generating poems from a given first line offers limited control over content, particularly for longer compositions. A common alternative is an approach where one or more keywords serve as input to guide the generative model. The following subsection details this method.

2.2 Keywords-Conditioned Poem Generation

Generating poems from a user-specified list of keywords is a popular approach for defining generative poetry tasks, particularly before the widespread adoption of LLMs. This method places the responsibility on the user to select keywords that best capture the desired theme of the poem. However, it offers significant flexibility to the generation algorithm in terms of lexical choices and rhetorical devices, as the set of keywords does not impose the level of detail achievable with modern instructional LMs. Additionally, this approach simplifies the creation of training datasets, as researchers can leverage existing NLP tools for topic modeling and keyword extraction from documents.

A notable example of this approach is the system for automatic French poetry generation described by Hämäläinen, Alnajjar, and Poibeau (2022). Their system employs an encoder-decoder architecture, where the encoder is a RoBERTa (Y. Liu, Ott, et al. 2019) model and the decoder is a GPT-2 model. The system uses Spacy[4] to extract up to four keywords from each poem in a corpus of 8,500 poems. The first line of the poem is generated based on one to four keywords, and subsequent lines are generated sequentially, conditioned on the previously generated lines.

Po-Miner by Hegade et al. (2021) is another example of the poem generation system accepting a single keyword as user input. Its architecture diverges from mainstream approaches by eschewing generative LMs in favor of a rule-based system that incorporates syllabification and rhyme selection. A unique aspect of this work is its proposed application: generating hack-resistant one-time passwords.

[3]The notation AABBA indicates a five-line stanza where the first, second, and fifth lines rhyme with each other, while the third and fourth lines share a separate rhyme.
[4]https://spacy.io

Some interactive systems[5] provided the ability to enter a set of keywords as a way to control the generation of the next line or an entire poem. The Jiuge system (Z. Guo, X. Yi, et al. 2019), designed as a poetry creation assistant with interactive editing capabilities, accepts both keywords and images as input modalities for poem generation. Another interactive poetry generation assistant that uses seed words for content control is CoPoeTryMe (Gonçalo Oliveira, Mendes, et al. 2017). In this system, seed words do not necessarily appear verbatim in the generated poem. Instead, they define a semantic field of related words from which the system selects appropriate terms to satisfy poetic constraints.

A keyword-based generation scenario can serve as an internal system component rather than a user-facing feature. For example, the Sudowoodo system for Chinese lyric generation (Chang et al. 2023) employs an auxiliary model that generates lyrics from keywords. This model creates synthetic training data by first extracting keywords from target songs, then generating corresponding lyrics. These synthetic pairs are subsequently used to build the final training dataset.

In the system proposed by Boggia et al. (2022), one of the two models generates a poem's first line from user-provided keywords. To train this model, the authors extract nouns, adjectives, and verbs from the first lines of poems in their corpus. These extracted terms serve as keywords, forming `(keyword, first-line)` pairs for the training dataset.

The limited expressiveness of keyword-based input makes this approach less suitable for contemporary LLM-powered systems, which typically rely on the more sophisticated control offered by instructional prompts. In the next subsection, we will consider examples of the implementation of poetry generation systems with this approach to control.

2.3 Instruction-tuned Prompting

Instruction tuning (supervised fine-tuning on instruction, output pairs) enables LLMs to accept direct natural-language constraints (e.g., "Write a 8-line poem about spring sunsets, with ABAB rhyme for each quatrain") and produce usable poetic outputs. This approach offers maximum flexibility to users, enabling control over various properties of the generated poem, such as genre, topic, keywords, emotional tonality, and more (Chakrabarty, Padmakumar, et al. 2022). Off-the-shelf instruction-tuned chat models (ChatGPT (OpenAI 2022), DeepSeek (DeepSeek-AI 2024), and other virtual assistants) respond well to zero- and few-shot instructive prompts and are widely used in empirical studies; researchers used these models with instructive prompts to generate poems for human rating. For instance, Shalevska (2024) conducted a comprehensive study on the creativity and rhetorical devices employed in poems generated by ChatGPT and found that this system can mirror human poetic expression, producing complex poems with numerous literary devices, but its creativity is constrained by training data and mastery of prompt engineering. Furthermore, Walsh, Preus, et al. (2024) provided a comparative analysis of poems generated by ChatGPT and those authored by human poets from the Poetry Foundation and the Academy of American Poets. They found that the ChatGPT models can successfully produce poems in a range of both common and uncommon English-language forms, but are much more constrained and uniform than human poetry.

Another example of using ChatGPT to generate poems using few-shot prompts is the paper by Tian and Peng (2022): in each prompt, they specified two sonnet examples and additional information about the topic of the required generation, using this approach as a baseline for evaluating the proposed approach.

One of the key advantages of instructive prompting is its ability to control poetic parameters like meter and rhyme directly through prompt instructions. In contrast, other systems often rely on separate components to manage these aspects during generation. For instance, Z. Hu et al. (2024) employ a specialized module to control diffusion processes, whereas instructive prompting integrates such control directly into the prompt.

[5] See more on interactive systems in Section 2.11.

However, a significant challenge with instructive prompting is the requirement for a training dataset that includes prompts for every sample. While C. Yu et al. (2024) used human-written prompts, this approach becomes impractical for datasets with tens of thousands of samples due to its labor-intensive nature. To address this, Huynh and Bao (2024) propose generating prompts automatically using LLMs.

Several studies have systematically evaluated instruction-tuned models. Porter and Machery (2024) conducted a large human study using ChatGPT to generate poems in the style of various poets. Walsh, Preus, et al. (2024) evaluated GPT-3.5 and GPT-4 (OpenAI 2023) across 24 poetic forms using instructional prompts. Y. Chen, Gröner, et al. (2024) quantitatively compared the diversity of poetic outputs from general-purpose instruction-tuned LLMs and poetry-specialized models. R. Zhang and Eger (2024) explored multi-agent poetry generation where agents follow instructional prompts to produce diverse outputs. Agirrezabal and Oliveira (2024) tested whether off-the-shelf LLama 2 (Touvron, Martin, et al. 2023) can generate metrically controlled verse via zero-shot instructional prompts. D'Souza and Mimno (2023) used custom-designed prompts to test the ability of the ChatGPT and PaLM (Chowdhery et al. 2023) LLMs to accurately reproduce poems by famous authors.

Instruction-tuned prompting has become the dominant method for controlling poetry generation, largely supplanting earlier techniques. This approach enables flexible control over both form and content in poetry writing assistance systems (Section 2.11), as demonstrated by CoPoet (Chakrabarty, Padmakumar, et al. 2022) — a T5-based model fine-tuned on instruction-output pairs for poetry generation. User studies confirm that instruction tuning enhances collaborative poem writing.

Many generation paradigms — including first line continuation (Section 2.1), keyword-conditioned (Section 2.2), and image-conditioned (Section 2.8) approaches — can be implemented as specialized cases of instruction-tuned prompting without sacrificing versatility. Additionally, a number of specialized tasks, such as poetic translation, poetry style imitation, and converting prose into poetry, can also be implemented using this approach. Due to the specific nature of these tasks, we decided to consider them separately in the following subsections.

2.4 Poetic Translation

Poetic translation requires reconciling semantic fidelity with formal and aesthetic constraints such as rhyme, rhythm, and imagery, making it one of the most technically demanding tasks in generative poetry. Due to its complexity, this task is well-suited for benchmarking multilingual LLMs (A. Chen et al. 2024) and all kinds of neural MT systems (Cespedosa Vázquez and Mitkov 2023; Mukherjee et al. 2024).

An analysis of the poetry translation studies cited in Table 1 reveal several consistent trends in computational approaches to poetry translation.

Training one's own LM from scratch or retraining open LLMs (Chakrabarty, Saakyan, et al. 2021; Ghazvininejad, Y. Choi, et al. 2018; Moreno 2021; W. L. Song et al. 2023) is less popular than zero-shot experiments with LLMs, including proprietary systems and open LLMs (Cespedosa Vázquez and Mitkov 2023; A. Chen et al. 2024; Cho et al. 2025; Mukherjee et al. 2024; Resende and Hadley 2024; S. Wang, Wong, et al. 2024), especially in the period after 2023. Two studies examine the effect of adding information about the poem being translated to the prompt (A. Chen et al. 2024; S. Wang, Wong, et al. 2024).

For many language pairs, the task of poetry translation is low-resource, forcing researchers to use different approaches with training data augmentation by adding prose translations or translations for an auxiliary language pair (Ghazvininejad, Y. Choi, et al. 2018; Moreno 2021).

Evaluation remains dominated by n-gram similarity metrics, though LLM-as-a-judge methods are beginning to emerge as effective alternatives to human evaluation (A. Chen et al. 2024; S. Wang, Wong, et al. 2024). Novel metrics such as Levenshtein distance (Levenshtein 1966) have also been explored (Cho et al. 2025), but no consensus has yet been reached on how to best capture poetic fidelity.

The development of new models, datasets, and quality assessment algorithms for poetry translation represents a promising research direction. Even when constrained to proprietary LLMs like ChatGPT, researchers can explore innovative approaches, such as prompt engineering, to enhance poetry translation, as demonstrated by S. Wang, Wong, et al. (2024). However, this area remains underexplored within the broader field of MT, as current models and evaluation protocols often fail to address the unique features of poetic texts. Insights from translation studies also offer valuable perspectives for computational approaches. Robinson (2010) highlights the philosophical and creative dilemmas inherent to poetry translation. Y. Ma and B. Wang (2020) proposed a systemic functional linguistics (SFL) framework for analyzing translations across multiple levels, from phonology to context. Herbert et al. (2024) introduced the collaborative "poettrio" method, where a source-language poet, a target-language poet, and a language advisor jointly refine translations — a process reminiscent of multi-agent approaches in LLM research.

Overall, research on poetic translation remains fragmented, with promising but underexplored directions in both model design and evaluation. Stronger integration with insights from human translation studies may provide a path forward.

2.5 Poetry Style Imitation and Parody

Like poetic translation, style imitation and parody require preserving stylistic fidelity, but within the same language rather than across languages. These tasks demand that models generate fluent verse while capturing the characteristic voice, rhythm, and creative intent of a source author. We identify two main approaches.

Author-stylized generation. This approach targets the style of a specific poet or a closed set of poets. Notable examples include EMILY (Shihadeh and Ackerman 2020), which imitates Emily Dickinson using a Markov model trained on 444 poems; Sounds Wilde (Tikhonov and Yamshchikov 2018a), which mimics several English and Russian poets; and William ShakesBlake 2.0 (Nalci et al. 2025), a GPT-3.5 model fine-tuned on 400 sonnets by Shakespeare and Blake that can generate hybrid, image-prompted verse.

A. Dai (2021) studies the imitation of Emily Dickinson's poetic style using GPT-2 further trained on a dataset of 586 stanzas from her poems, totaling fewer than 7,000 lines. Their evaluation is limited to qualitative analysis, without quantitative metrics, which makes comparison with other studies difficult.

Sawicki, Grzes, A. Jordanous, et al. (2022) study authorial style reproduction in poetry by fine-tuning GPT-2 on poems by Byron and Shelley collected from Project Gutenberg. Their evaluation combines BLEU-based similarity to reference poems, a BERT-based style classifier, and a qualitative assessment of grammaticality and fluency.

Open style transfer. In this setting, models imitate the style of an arbitrary prototype text rather than a fixed author. For example, Sudowoodo (Chang et al. 2023) adapts to the style of Chinese song lyrics provided at inference time.

Parody generation can be addressed by applying controlled transformations to an existing text, such as substituting words with synonyms or semantically related alternatives (Bay et al. 2017). This approach avoids the need to train LMs on limited domain-specific data.

Gonçalo Oliveira (2020) adopts a transformation-based method, mainly driven by analogy, to generate new song lyrics from an original song and a target theme word. This system computes analogies in a distributional semantic space based on GloVe word vectors (Pennington et al. 2014) and shifts the theme by replacing selected words in the original text. To improve output quality, the substitution process is constrained using part-of-speech information and phonetic data from the CMU Pronouncing Dictionary.[6]

An alternative strategy for selecting replacement words uses masked LMs. Gonçalo Oliveira (2021) applies this approach to Portuguese song lyric generation using BERT (Devlin et al. 2019), with additional constraints to support grammatical correctness, meter, and singability.

[6] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

When the parody task is narrowly defined in terms of the target effect and input data, it may be possible to avoid the use of LMs. This is shown by Gatti et al. (2017), who present a system for generating parodies of popular English-language songs. Their approach starts from a corpus of 100 well-known songs and a user-provided text that defines the parody context. The system replaces words using lexical resources such as WordNet (Fellbaum 2010), the Oxford Thesaurus (Urdang 1993), and WikiData (Vrandečić and Krötzsch 2014). Only song choruses are modified, and word selection is guided by heuristic rules together with part-of-speech and phonetic constraints, including rhyme and syllable count.

Several strategies address the data scarcity that characterizes the parody task. One approach employs models that can learn from small corpora, though limited training and evaluation samples often undermine reliability. A more common direction adapts LLMs through supervised fine-tuning or few-shot prompting. Sawicki, Grzes, Goes, Brown, Peeperkorn, and Khatun (2023) report weak results with GPT-3.5 and GPT-4 under few-shot prompting, while other work (Sawicki, Grzes, Goes, Brown, Peeperkorn, Khatun, and Paraskevopoulou 2023) shows that fine-tuning GPT-3 on approximately 300 poems per author yields more convincing stylistic fidelity.

The small number of studies prevents reliable trend analysis in evaluation. Most rely on human judgments (Chang et al. 2023; Gatti et al. 2017; Nalci et al. 2025; Shihadeh and Ackerman 2020), sometimes combined with automatic metrics (Tikhonov and Yamshchikov 2018a). Automatic evaluation remains rare, with only Sawicki, Grzes, Goes, Brown, Peeperkorn, and Khatun (2023) experimenting with a GPT-3 classifier.

The scarcity of work on parody generation underscores the difficulty of collecting suitable data, evaluating results reproducibly, and modeling humor. These challenges define important directions for future research. Moving beyond style imitation, the next task category considers how prose can be transformed into poetry, compressing narrative content into verse.

2.6 Prose-To-Poem: From Narrative to Verse

Transforming prose into poetry challenges systems to compress narrative content into verse, introducing rhythm, imagery, and figurative language without losing essential meaning. There are several examples of solutions to this problem. Khanmohammadi et al. (2023) apply the NMT approach for the generation of ancient Persian poetry given the prose source. The peculiarity of the experiment was the limited data available. As a parallel corpus of translations, they had 5,191 samples, of which 1,720 were set aside for testing, and the remaining samples were augmented to 28,820 pairs. On this data, an encoder-decoder model was trained with the Gated Recurrent Unit (GRU) in the encoder and decoder with an attention mechanism. A larger dataset (289,422 couplets in the training part) with poems in Old Persian was used to train the masked LM (MLM), which generates poetry by substituting mask tokens into templates prepared using a special algorithm collected from the translation results. The authors experimented with BERT, RoBERTa, ALBERT (Z. Lan et al. 2020), and DistilBERT (Sanh et al. 2020) as MLMs for this stage. The evaluation protocol included automated metrics and human evaluation. The automated metrics are BLEU and ROUGE. The human evaluation stage involved two groups with different levels of expertise who rated the final texts on a scale of 1–5 based on the following criteria: fluency, coherence, meaningfulness, poeticness, and translation quality.

Van de Cruys (2020) presents a system for generating poetry in English and French. The authors use an encoder-decoder model with GRU and attention. The text is presented as words (a dictionary of 17 thousand words), and the words in a line go from right to left to improve rhyme generation. Poem generation is implemented by introducing additional restrictions on the distribution of words in the decoder. The authors did not use automatic metrics to evaluate the results. Instead, a group of 22 people evaluated a small number of texts on a 5-point scale according to a set of criteria: fluency, coherence, meaningfulness, and poeticity. In addition, people tried to determine whether the text of the poem was written by a person or generated by a model.

From the researcher's perspective, translating prose into a semantically equivalent poetic form can be regarded as a form of MT within a single language. This process comes with additional strict requirements, including the necessity to incorporate rhyme, maintain a poetic meter, and meet softer criteria such as embedding deeper metaphors, bold epithets, and allusions that are characteristic of poetry.

2.7 Acrostic Poetry Generation

Acrostic generation is often treated as a separate task in the generative poetry literature due to the explicit structural constraints it imposes (see more on acrostic at R. Greene et al. 2012, p. 4; and Dunphy 2010, p. 8) and the way it is addressed in prior work. Several studies (Agarwal and Kann 2020; L.-H. Shen et al. 2019) focus on acrostic generation as their primary research problem, proposing models designed to enforce strict positional constraints on line-initial letters. In multi-genre poetry generation systems, acrostic-related constraints are sometimes implemented as a dedicated module that complements other generation mechanisms, as in Jiuge (Z. Guo, X. Yi, et al. 2019). Although acrostic generation can also be framed within instruction-based prompting approaches (Section 2.3), its recurring treatment as either a standalone focus or a specialized system component motivates its separate discussion in this survey.

One of the challenges in creating acrostic generation systems, especially those based on transformer LLMs, is the limited number of available datasets. To address this, Fedchin et al. (2025) developed AcrosticSleuth, a tool for detecting acrostics in corpora of French, English, and Russian poems, enabling the creation of training datasets. The authors also released an example dataset, called the Acrostic Identification Dataset.

Agarwal and Kann (2020) propose a method for generating English acrostics up to 8 lines long. Their approach modifies the token probability distribution for the first words of each line, guiding the decoding algorithm to select tokens with the required initial letter. This eliminates the need for a specialized acrostic dataset. Additionally, the authors used an auxiliary model to generate line-ending words, reducing reliance on the main model's ability to produce rhymes.

L.-H. Shen et al. (2019) describe an acrostic generation system for both English and Chinese based on a transformer model. The training data were constructed from web-scraped texts, comprising approximately 651,000 Chinese samples and 1,000,000 English samples. A key feature of their approach is the use of user-defined control codes that specify properties such as text length and rhyme scheme, allowing explicit control over the generation process.

The acrostic constraint, which specifies the first letters of poem lines, provides a valuable test of an LM's character-level control. This is particularly relevant for models using Byte Pair Encoding (BPE) tokenization (Sennrich et al. 2016), where character-level manipulation can be challenging.[7] Since verifying correct acrostic generation can be implemented as a formal analysis of line-initial letters, this task could be effectively incorporated into LLM benchmarks, as suggested for sonnets by Walsh, Antoniak, et al. (2024).

2.8 Image-Conditioned Poem Generation

Beyond text-based control methods, multimodal inputs such as images have become increasingly important. Image-conditioned poetry generation explores multimodality, requiring models to interpret visual inputs and translate them into poetic language that conveys both description and emotion.

Traditionally, image-conditioned poem generation relied on either image captioning models to produce short descriptions of a given image (Nalci et al. 2025) or image classifiers to detect the depicted object class. These outputs were then used to select keywords, effectively reducing the task to a keyword-to-poem generation approach (Section 2.2). An example of this method is described by Loller-Andersen and Gambäck (2018).

[7] See discussion in Section 4.1.

Several additional papers on image-to-poem generation were published in 2018, prior to the availability of open multimodal LLMs:

(1) B. Liu et al. (2018) use an GRU-based generator trained in an adversarial setting. The key idea of this work is a common space of embeddings for images and poem texts, formed at the training stage and then used to condition the generation of a poem by a recurrent LM. As part of this work, the authors published a dataset with image-poem pairs (more than 8,200), available in a repository on GitHub. Unfortunately, all texts in this dataset are converted to lowercase and are more like free verse than metrical poems.
(2) Xu et al. (2018) describe a system for generating Chinese poems from images. The architecture employs a multi-stage pipeline: for an input image, visual features are extracted using a VGG-19 (Simonyan and Zisserman 2015) model, and keywords are subsequently identified. These features and keywords are then used to control the generation process in a GRU-based decoder.
(3) L. Liu et al. (2018) present another image-to-poem generation system built upon Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997)-based LM. A distinctive feature of this work is its handling of image sequences, where the model automatically selects the most suitable image for feature extraction.
(4) The system proposed by Loller-Andersen and Gambäck (2018) processes input images through the following pipeline: an Inception network classifies the image, the top-5 predicted classes are selected, relevant keywords are identified for these classes, rhyme candidates are generated, and finally, the collected words guide poem generation in an LSTM-based LM.

The systems described above share common architectural features, including the use of recurrent LMs and earlier-generation encoders like VGG for visual processing. Since 2019, the availability of pre-trained generative LMs has enabled more advanced approaches. For instance, MiniGPT-4 (D. Zhu et al. 2024) combines the Q-Former visual encoder from BLIP-2 (J. Li et al. 2023) with the Vicuna LM (Zheng et al. 2023) (a LLaMa (Touvron, Lavril, et al. 2023) variant), using a two-stage fine-tuning process on a custom dataset. Although the paper does not detail the evaluation protocol, it suggests that the developers used a binary scale to assess whether generated texts matched the prompt and image content.

The release of open-source multimodal LLMs in 2023 — such as LLaVA (H. Liu et al. 2023) and Qwen-VL[8] — simplified the development of image-to-poem systems by providing high-quality text generation capabilities. An example is VisuCraft by Jiang et al. (2025), which integrates LLaVA and InstructBLIP (W. Dai et al. 2023). Its architecture comprises: (1) a multimodal structured information extractor that generates detailed image descriptions, and (2) a dynamic prompt generation module that constructs prompts from these descriptions and user instructions. Human evaluation results reported in the paper indicate that enriching user queries with extracted image information improves performance across several metrics (Perceived Visual Relevance, Human Creativity Score, Human Instruction Adherence, and Overall Quality), all rated on a 5-point Likert scale.

Given the rapid progress of open multimodal LMs, we can expect to see more research in generative poetry with more complex user interaction scenarios inherent in virtual assistants described in Section 2.11.

2.9 Melody-Conditioned Lyric Generation

Beyond image-to-poem generation, another multimodal task involves generating song lyrics for a given musical accompaniment. A critical requirement for this task is the alignment of musical rhythm with the prosodic properties of the lyrics, ensuring compatibility in terms of rhythm, syllable count, and overall musicality. This section provides a brief analysis of relevant research in this area.

Y. Chen and Lerch (2020) propose an end-to-end system based on the SeqGAN (L. Yu et al. 2017) architecture, capable of generating a line of song lyrics given a melody and a topic as input. The model was trained on a dataset[9]

[8] https://qwenlm.github.io/blog/qwen-vl
[9] https://github.com/yy1lab/Lyrics-Conditioned-Neural-Melody-Generation

containing 12,197 MIDI songs, each with paired lyrics and melody alignment. This dataset was originally created by Y. Yu et al. (2021) for a lyrics-to-melody system using a combination of LSTM and generative adversarial network (GAN) architectures. Unfortunately, at the time of writing this survey, the dataset in the repository appears to be incomplete and lacks MIDI samples.

Vechtomova et al. (2021) developed a real-time system that generates lyric lines congruent with live music during a jam session. Their approach uses a variational autoencoder (VAE) (Kingma and Welling 2014) to learn representations of mel-spectrograms from audio clips and a conditional VAE to learn representations of lyric lines. These representations are aligned so that the text representation can be inferred from the melody and then decoded into a line of lyrics using a text VAE decoder.

Tian, Narayan-Chen, et al. (2023) designed a hierarchical lyric generation framework that first generated a song outline and then the complete lyrics. To overcome the limitations of copyrighted data, the authors trained their model solely on song lyrics. The alignment of lyrics to the melody was achieved during decoding by imposing constraints on the generation algorithm. The framework operated in two steps: (1) a fine-tuned BART-large (M. Lewis et al. 2020) model generated a song outline based on the song title, genre, and salient words, and (2) a fine-tuned GPT-2 large model generated the full song text from the outline. The system was evaluated using automatic metrics (topic relevance, diversity, and fluency) and human evaluation. Human annotators assessed the generated songs on a 1–5 Likert scale for singability, intelligibility, coherence, creativity, and rhyme match.

Elzohbi and R. Zhao (2024) presented a system for generating English song lyrics with stress patterns adapted to musical rhythms. Their approach uses a ByT5 (Xue, Barua, et al. 2022) transformer encoder-decoder with byte-level tokenization. The model, fine-tuned on a dataset of over 1 million samples, takes as input a sequence of 0s and 1s representing musical beats and generates corresponding lyrics. For training data preparation, lyrics were converted to beat patterns through phonemicization using DeepPhonemizer,[10] a text-to-phoneme conversion library. For evaluation, following common practice in generative poetry research, the authors developed a custom protocol measuring both the perplexity of generated texts against reference samples and the accuracy of stress pattern alignment with given beats. The full system source code is available in their repository.[11]

Y. Chen and Teufel (2024) addressed the lack of large parallel melody-lyrics datasets for Mandarin song generation by introducing an intermediate text representation that captures key phonetic and rhythmic properties of a song. An mBART (Y. Liu, J. Gu, et al. 2020)-based generative model is trained to produce Mandarin song lyrics conditioned on this representation. The training data are derived from a corpus of 36,891 Mandarin song lyrics available in a public GitHub repository.[12] At inference time, an input melody provided in MIDI format is first converted into the intermediate representation, which is then used as input to the mBART model.

Qian et al. (2023) address Chinese melody-conditioned lyric generation by decomposing the task into two stages using an intermediate representation. This representation consists of compound templates that combine acoustic parameters of the melody with textual attributes. The system is based on the mT5 model (Xue, Constant, et al. 2021), which is further fine-tuned on a synthetic Lyric-Template dataset containing 249,007 samples.

Given the growing popularity of song generation models like Suno,[13] this research direction holds significant potential. Future work could explore combining melody-to-lyric generation with other control methods, such as incorporating author style or emotional tone, as well as integrating cross-task ideas, such as simulating a song based on a given sample of an author's performance.

[10] https://github.com/spring-media/DeepPhonemizer
[11] https://github.com/melzohbi/poem-rhythm
[12] https://github.com/gaussic/Chinese-Lyric-Corpus
[13] https://suno.com

2.10 Other Controlled Generation Tasks

In the field of automatic poetry generation, certain control modalities do not neatly fit into the previously discussed categories, as they often encompass unique or hybrid mechanisms that defy straightforward classification. These modalities are frequently represented by individual examples that embody a specific control method, making it challenging to group them into broader, separate sections. Consequently, these approaches are best described within this section to preserve their distinctiveness and to avoid fragmenting their representation across multiple categories. Below, we highlight some notable examples of such approaches, which demonstrate the diversity and creativity in controlling poetic output.

Elzohbi and R. Zhao (2025a) explored the generation of English song lyrics by use of ByT5 byte-level transformer model (Xue, Barua, et al. 2022) finetuned on LLM-generated poetry line paraphrase dataset, to rephrase poetry lines according to a given beat template.

Røstvold and Gambäck (2020) explore the generation of English poetry with a specified sentiment. Their system employs an LSTM as the generative model, augmented with auxiliary components for rhyme selection and evaluation of the generated text. A unique feature of this system is its representation of text: words are processed from right to left within each line. Before generating a line, the system selects a pair of rhyming words with the desired sentiment, and the LM then generates the line from the rhyme backward to the beginning.

Another approach to controlling sentiment and style in Chinese poetry generation is presented by Shao et al. (2021). They propose a GPT-2-based system for Chinese classical poetry in which control is achieved through input tags. Evaluation focuses on the model's ability to follow the specified attributes using automatic metrics.

Modern transformer LMs can generate coherent text through prompt-based instruction following. However, prompts alone may not reliably enforce explicit structural or lexical constraints. To address this limitation, Roush et al. (2022) proposed a plug-and-play approach implemented in the Constrained Text Generation Studio. Their framework applies token-level filters during decoding to enforce constraints such as banning specific letters, requiring a fixed number of syllables, or restricting words to partial anagrams of a given term. To evaluate this approach, the authors fine-tuned GPT-2 on the *Lipogram-e* dataset,[14] which consists of texts that avoid the letter "e." They compared generations from the fine-tuned model with outputs from the base GPT-2 model combined with their constraint filtering method. The results show that the plug-and-play constraints effectively enforce the desired properties without requiring task-specific retraining.

Tonra et al. (2019) propose a system that uses biometric data from wearable fitness devices to generate poetry. The approach correlates the wearer's physiological states (e.g., heart rate and activity level) with poetic content, automatically creating and publishing poems that reflect real-time physiological changes. A related approach using brain activity data is implemented in Bio-Mechanical Poet (Thölke et al. 2024), described by its authors as an adaptive brain-computer interface that integrates real-time electroencephalography (EEG) with generative AI to create immersive audiovisual poetic experiences. The system calculates a spectrum from recorded electrical activity, extracts dominant frequencies, and computes powers for six canonical frequency bands along with additional neuroscience metrics. These signals are transformed into prompts every 20–30 seconds for the OpenAI GPT model, which generates corresponding poems. This system enables users to explore their internal cognitive states through real-time poetic generation.

Cerdas (2025) presents a method for generating poems using X-ray intensity maps as an input. While this technique employs a simple Markov model as the generator and does not provide significant NLP or generative modeling results, we include this work in our survey due to its original interdisciplinary approach.

McCormack et al. (2024) introduce "The Mimetic Poet Machine" — a physical device designed for interactive poetry creation. Users interact with ChatGPT through a set of built-in prompt templates by selecting word cards from a collection of approximately 200 words and special control markers. The authors argue that this

[14] https://huggingface.co/datasets/Hellisotherpeople/Lipogram-e

device fosters creative ideation, inspiration, and reflective thought. The Mimetic Poet was evaluated through a two-week installation in the researchers' laboratory, where lab members were invited to use it freely. Data on user interactions was collected and supplemented with a focus group discussion. Key findings included: (1) a positive reception of the physical interface, with participants describing the tangible, magnetic poetry approach as "playful" and "intuitive;" (2) an appreciation for poetic responses, as users valued the AI's ambiguous, poetic outputs over purely factual ones, finding them more engaging and thought-provoking; (3) creative constraints, where some participants found the limited vocabulary stimulating, while others felt constrained; (4) time for reflection, as the slower pace of interaction was generally appreciated for encouraging thoughtful engagement. The paper also presents a methodology for analyzing human-AI interaction sessions, along with the authors' findings and conclusions.

These examples illustrate the diversity of control modalities in poetry generation. While the methods discussed here are not exhaustive, they highlight the potential for exploring new modalities. Future research could focus on identifying novel control mechanisms, necessitating the full research cycle — from data collection and model architecture design to the development of evaluation protocols.

2.11 Interactive System and Poem Writing Assistance

The problems and approaches discussed in previous subsections primarily support offline operation, including batch generation for statistical analysis. In contrast, a distinct class of generative poetry systems prioritizes interactivity and user control over individual steps in the generation process. These are often categorized in the literature as writing assistance tools, artistic support tools, or creativity support tools. Their objective is to assist the poet at various stages — such as drafting, rhyme selection, or text revision — enabling the user to remain an active participant in the creative process rather than a passive consumer of generated text. In this section, we examine several examples of such systems.

Poem Machine is an interactive online tool for co-authoring Finnish poetry (Hämäläinen 2018b) primarily targeted as an educational tool for children. The system does not use generative LMs, relying instead on dictionaries, tools for working with the morphology and syntax of the Finnish language, and a set of rules for generating lines of poetry.

Jiuge is a human-machine collaborative system for generating Chinese classical poetry (Z. Guo, X. Yi, et al. 2019). The features of this system are its multimodality — both text prompts (keywords, texts) and images can be used as input data. The described pipeline of the system includes an encoder-decoder LM for line-by-line generation of poems and a ranker for selecting the best option.

Boggia et al. (2022) describe a human-computer co-creative poetry writing system that combines several approaches: the first model generates an initial line based on keywords, while the second model proposes continuations based on the existing text. The user selects a continuation, and the process repeats iteratively.

Y. Sun et al. (2023) propose a song rewriting system based on an encoder-decoder transformer architecture for generating Chinese lyrics from existing text prompts. The system allows users to control text generation by specifying keywords for masked fragments, which can range from individual tokens to entire lines. This flexibility enables both partial editing and full lyric generation from scratch. The authors implement a sophisticated rhyme generation scheme:

- Lines are generated right-to-left to prioritize rhyme selection at the start of each line.
- Special positional embeddings indicate token positions relative to line beginnings.
- A custom *restricted vowel loss* component during training reinforces internal rhyme patterns.

For evaluation, the authors developed a specialized protocol assessing rhyme quality and diversity. Their results show superior rhyme performance compared to GPT-2, though the custom evaluation metrics limit direct comparison with other approaches.

Phraselette is an artistic support tool designed to assist poets in word and phrase selection (Calderwood, Chung, et al. 2025). The authors highlight the system's ability to support various low-level search scenarios while allowing flexible control over search constraints. Most of the paper focuses on detailed descriptions of these usage scenarios with practical examples. While the authors repeatedly claim their approach is more user-friendly than ChatGPT, they do not specify which LMs power the system.

Roush et al. (2022) introduce "Constrained Text Generation Studio," an interactive AI tool for composing poetry and other texts in English. The paper's main contribution is a method for vocabulary control in LMs to enable constrained text generation. A notable omission is the lack of any systematic evaluation of the poetry generation capability.

J. Ma et al. (2023) present an interactive system for Chinese song lyric generation. Their approach allows users to begin with an initial draft and make guided improvements to text fragments. Technically, the system employs GPT-2-based models for initial draft generation and a masked token regeneration model for word replacement. While the paper includes evaluation results, the use of a custom evaluation protocol limits comparability with other approaches.

Gonçalo Oliveira, Mendes, et al. (2017) present a multilingual assistant (English, Spanish, Portuguese) for poetry generation that supports semi-automatic text refinement. The system allows users to control textual aspects through a graphical interface, including syllable count per line, rhyme scheme, and sentiment. Additional functionality includes edit history tracking, social media sharing capabilities, and saving/loading poems at any stage of composition.

Research on writing assistance systems should extend beyond technical aspects to examine their impact on professional writers' productivity and creative authorship. This perspective is explored by Reza et al. (2025), who conducted (1) a systematic review of 100+ works on human-AI collaborative writing and (2) interviews with 15 system users across multiple domains (including poetry). Similar investigations of writing tool applications appear in works by Calderwood, Qiu, et al. (2020) and A. Guo et al. (2025).

The rapid development of chatbots in recent years has led to many interactive scenarios for working with text being implemented in "general-purpose" virtual assistants. On the one hand, this reduces the motivation to create writing assistance systems specialized in creating poetry. On the other hand, we believe that this opens up the possibility of studying user behavior, as was done, for example, by Booten and Gero (2021).

Similar to earlier tools ranging from typewriters to grammar checkers, modern writing assistance systems should minimize disruptions to the creative process. However, the significant influence of current LLM-powered systems on all forms of linguistic expression necessitates careful examination of their effects. Studies by Padmakumar and He (2024) and Doshi and Hauser (2024) exemplify this research direction through their investigations of impacts on linguistic diversity. Comparable studies focusing on domains like song lyric composition will prove equally valuable for advancing the field of AI-assisted creative writing.

We also suppose that the development of interactive poetry creation systems can have a socially useful effect, for example, to support children with autism spectrum disorders (Shabani Minaabad 2020) or to develop creativity for children (Coles 2017).

Taken together, all task formulations discussed in the subsections above illustrate the diverse ways in which user control — through structure, semantics, style, or modality — shapes the design and evaluation of generative poetry systems. Realizing these task formulations in practice depends heavily on the availability and quality of datasets, which provide training material for generative models and benchmarks for their evaluation, as well as on supporting data-engineering tools such as syllabification and rhyme detection.

3 Data Engineering

Generative poetry tasks depend on both carefully curated datasets and specialized preprocessing. This section surveys the data-engineering foundations of the field, focusing on available corpora as well as tools for their annotation and analysis.

Modern LMs are highly data-dependent. While some studies (Hämäläinen 2018a; Rashel and R. Manurung 2014; Tian and Peng 2022) rely on prose corpora, which are more readily available, the mainstream approach uses poetry-specific datasets in the target language and genre. Consequently, one of the researcher's first tasks is to locate suitable training data, preprocess it, and, where necessary, augment it to enable control mechanisms.[15] Key criteria for selection include poem length, genre or form, and language. To simplify experimentation, many projects constrain poem length, often reducing even lengthier genres (e.g., sonnets) to a single quatrain.

Among the key factors influencing the choice of training data, the issue of copyright and data licensing occupies a special place. Modern LLMs require large volumes of poetic text for training, much of which is protected by copyright, particularly in the case of contemporary works by professional poets, songwriters, and other authors. The use of text collected without explicit permission, and sometimes in violation of usage restrictions, has raised regulatory and legal concerns (Buick 2024; Lucchi 2023). Consequently, access to modern poetry for academic model training is often legally constrained or requires costly licensing arrangements (K. Li et al. 2024). As a result, many studies rely on public-domain or classical poetry — such as Shakespearean verse or pre-modern Chinese poetic forms — as a legally safer alternative. While this approach facilitates reproducibility and legal compliance, it also introduces stylistic and temporal biases that limit the diversity of forms, themes, and language patterns captured by current generative poetry systems.

Among poetic genres, sonnets are the most frequently studied (Agnew et al. 2023; Ghazvininejad, X. Shi, et al. 2016; Hopkins and Kiela 2017; Walsh, Antoniak, et al. 2024), likely due to the availability of open datasets and the absence of copyright restrictions. Other explored genres include limericks (J. Wang et al. 2021), acrostic poetry (Agarwal and Kann 2020), and blackout poetry (Baral et al. 2021). Multi-genre systems exist, but remain less common (J. Hu and M. Sun 2020; Koziev and Fenogenova 2025).

In lyric generation, researchers have explored rap (Xue, K. Song, et al. 2021) and pop lyrics (Ram et al. 2021). Multi-genre systems are facilitated by large-scale lyric datasets, such as the "Genius Song Lyrics" corpus,[16] which contains over 5 million songs across 100 languages and is scraped from platforms like Genius.[17]

Language choice is often constrained by corpus availability. Multilingual resources such as PoeTree (Plecháč et al. 2024) provide deduplicated poems in nine languages, enriched with Universal Dependencies annotations (Marneffe et al. 2021) in machine-readable form. Most open datasets, however, are monolingual and vary in metadata quality, an issue that must be carefully addressed during data engineering.

The availability of suitable training data remains a major limitation in generative poetry research, given the language and genre restrictions of most corpora. To provide researchers with practical guidance, we next describe open datasets in detail (Section 3.1). Where additional processing is required — such as stress placement, syllabification, or rhyme detection — we discuss the corresponding tools in Sections 3.2–3.5.

3.1 Datasets

The use of existing datasets can significantly reduce research costs. Well-constructed datasets serve two primary purposes in generative poetry. Small, high-quality datasets with carefully defined evaluation and human labeling protocols enable precise assessment of generated poems. Conversely, large, diverse corpora of high-quality texts facilitate the rapid prototyping of experimental machine learning pipelines. This section reviews both types

[15] See more on control mechanisms at Section 2.
[16] https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information
[17] https://genius.com

Table 2. Open datasets for solving generative poetry problems.

Language	Poetry genres	Datasets
English	multi-genre	Gutenberg Poetry Corpus (github.com/aparrish/gutenberg-poetry-corpus) Poems dataset (www.kaggle.com/datasets/michaelarman/poemsdataset) Poetry Foundation Poems (www.kaggle.com/datasets/tgdivy/poetry-foundation-poems) Public Domain Poetry (huggingface.co/datasets/DanFosing/public-domain-poetry) Aoyama et al. (2023), Haider (2021), Haider, Eger, et al. (2020), Mahbub et al. (2023), and Sreeja and Mahalakshmi (2019)
English	limericks	Abdibayev, Tikhonov, et al. (2021)
Chinese	multi-genre	THUNLP-AIPoet Datasets (github.com/THUNLP-AIPoet/Datasets) Chinese Classical Poetry Matching Dataset (CCPM) (github.com/THUNLP-AIPoet/CCPM) Comprehensive Database of Chinese Poetry (github.com/chinese-poetry/chinese-poetry) C.-L. Liu et al. (2022)
Chinese	song lyrics	Crothers et al. (2023)
Spanish	multi-genre	Garzón and Pérez (2020)
Spanish	sonnets	Barbado et al. (2022)
Persian	multi-genre	Shereno: A dataset of Persian modernist poetry (www.kaggle.com/datasets/elhamaghakhani/persian-poems) Khanmohammadi et al. (2023)
Arabic	multi-genre	Abboushi and Azzeh (2023), Alyafeai et al. (2023), and Shahriar et al. (2023)
German	multi-genre	Haider (2024)
French	multi-genre	french_poetry (huggingface.co/datasets/manu/french_poetry)
Turkish	multi-genre	turkish-poems (huggingface.co/datasets/okg/turkish-poems)
Kurdish	lyrics and songs	Ahmadi et al. (2020)
Hindi	multi-genre	Hindi_Poems (huggingface.co/datasets/Sourabh2/Hindi_Poems
Bulgarian	multi-genre	bulgarian_poems (huggingface.co/datasets/Dilyana56/bulgarian_poems)
Czech	multi-genre	R. Rosa et al. (2025)
Italian	multi-genre	biblioteca_italiana (github.com/linhd-postdata/biblioteca_italiana)
Multilingual	song lyrics	Bertin-Mahieux et al. (2011) and Meseguer-Brocal et al. (2018)

of datasets available for researchers. A complete catalog of poetry datasets with detailed statistics and links is provided in Appendix B and summary information by language and genre is available in Table 2.

The quality of the training datasets significantly impacts the performance of the resulting systems. Therefore, careful consideration must be given to several key points: (1) defining what constitutes defects in poetic texts collected from the internet, (2) cleaning the collected poems to remove noise, and (3) annotating texts for training various model architectures. Addressing these issues ensures the creation of high-quality datasets, which are essential for developing robust and effective generative poetry systems.

An analysis of the data presented in Table 2 shows that 1) in absolute terms, two languages — English and Chinese — dominate, possessing datasets suitable for both LLM training and evaluation (Aoyama et al. 2023;

Haider, Eger, et al. 2020; C.-L. Liu et al. 2022; Sreeja and Mahalakshmi 2019; X. Yi, M. Sun, et al. 2018). Human-labeled data can also be found for other languages (Barbado et al. 2022; Garzón and Pérez 2020; Shahriar et al. 2023), but in most cases, the choice is significantly narrower.

Public datasets from other NLP domains can provide valuable data for generative poetry research. Conversational datasets like WildChat (Y. Deng et al. 2024) and the OpenAssistant Conversations Dataset[18] contain user interactions with virtual assistants that can be filtered for poetry-related queries. This approach could yield data for both instructive prompting (Section 2.3) and interactive poetry systems (Section 2.11). Initial filtering experiments using regular expressions suggest approximately 1474 English prompts and 626 Russian prompts from WildChat are poetry-related, potentially providing several hundreds examples for further analysis.

Several tasks that have gained attention in the broader NLP community remain under-explored in generative poetry, representing potential research gaps. These include applying Chain-of-Thought (Wei et al. 2022) (CoT) and reasoning approaches to poetry generation, where models would explicitly plan content while satisfying poetic constraints. While datasets like OpenMathReasoning (Moshkov et al. 2025) exist for reasoning tasks, and preliminary work like Tian and Peng (2022) demonstrates constraint-aware generation, comprehensive studies integrating cognitive processes of poetry composition (Hanauer 2010; Peskin and Ellenbogen 2019) with modern reasoning techniques are lacking.

Data scarcity also limits research on user preferences for generated poetry. While a small Chinese dataset exists (X. Yi, M. Sun, et al. 2018) with 173 samples, much larger preference datasets are available for general text generation, such as HelpSteer3-Preference (Z. Wang, Zeng, et al. 2025) with 40,000 human-annotated samples. These could serve as prototypes for developing poetry-specific preference datasets.

Poetry datasets are few in number and often inconsistent in quality. This makes it difficult to study tasks that require strict formal control or rely on human judgments of quality. Similar challenges are likely to appear in other creative areas of NLP, such as storytelling, style transfer, or humor, where good benchmarks are also scarce and expensive to build. In short, poetry datasets illustrate how data scarcity and quality issues slow progress on creative NLP tasks, a challenge shared with areas like storytelling, style transfer, and humor generation.

Datasets containing synthetic (machine-generated) poetry, primarily developed for training and evaluating detectors of LLM-generated text as well as for solving highly specialized poetry-related problems, warrant separate discussion.

Hayawi et al. (2024) released a small dataset of LLM-generated texts that includes 502 poems produced by GPT-3.5 and BARD.[19] The dataset is publicly available via a GitHub repository.[20]

S. Wang, J. Wu, et al. (2025) study the detection of synthetic poetry in Chinese and describe a dataset consisting of 800 poems written by professional poets and 41,600 poems generated by multiple virtual assistants. At the time of writing, the dataset referenced in the paper was not publicly accessible.

A smaller publicly available dataset hosted on Kaggle[21] contains 776 English-language poems generated by LLMs, including GPT, Phi (Microsoft 2024), LLaMA, and Gemini, alongside 375 poems authored by humans. While limited in scale, this resource enables comparative analysis of human- and machine-generated poetry.

Elzohbi and R. Zhao (2025a) released the ParaPoetry dataset,[22] which contains approximately 4.7 million English poetry lines along with GPT-3.5-generated paraphrases. The dataset was originally designed to address the task of paraphrasing song lyrics to fit a specified musical beat.

A more detailed discussion of algorithms and models for detecting LLM-generated poetry is provided in Section 5.1.6.

[18] https://huggingface.co/datasets/OpenAssistant/oasst1
[19] BARD was a conversational AI service developed by Google DeepMind.
[20] https://github.com/sakibsh/LLM
[21] https://www.kaggle.com/datasets/armandszokoly/human-and-language-model-generated-poems
[22] https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/WMU1BC

In the following subsections, we address the challenges associated with preparing data for training generative poetry systems, including segmenting text into syllables (Section 3.2), determining poetic meter and placing stresses (Section 3.3), detecting rhymes (Section 3.4), and part-of-speech tagging (Section 3.5). Certain data engineering issues remain underexplored in the context of generative poetry due to the lack of relevant research. Examples include extracting poems from web pages, correcting spelling errors, and addressing punctuation and grammar issues. These tasks, while potentially valuable, are not discussed in this survey.

3.2 Syllabification

Some generative models operate on a syllabic representation of text, where tokens correspond to syllables.[23] This approach necessitates the use of a syllabification tool, typically tailored to a specific language. For instance, Agirrezabal, Alegria, Arrieta, et al. (2012) describe a syllabifier for Basque, while Chudoba and R. Rosa (2024) discuss experiments with syllabified Czech poetry using the tool Sekáček.[24] Similarly, Zugarini, Melacci, et al. (2019) employ a syllabifier based on Italian hyphenation rules, and Haider (2021) uses hyphenation information for syllabifying English and German poetry. For Russian poetry, tools such as rusyllab[25] and the syllabification module in rupo[26] are available. Additionally, De Sisto et al. (2024) mention syllabifiers for several Indo-European languages.

However, syllabification tools are not readily available for many languages. When selecting an architecture that relies on syllabic text representation, researchers must either allocate resources to develop a custom syllabifier or consider alternative architectures that do not require syllabic segmentation. This decision is critical, as the availability of language-specific tools can significantly influence the feasibility and effectiveness of the chosen approach.

3.3 Scansion and Stress Assignment

Scansion and stress assignment are important components in poetry generation systems, as they ensure the generated text adheres to the metrical and rhythmic patterns of the target language. The complexity of stress placement depends largely on the target language, which can be categorized into two types: (1) languages with fixed stress and (2) languages with variable stress. In fixed-stress languages, such as Czech, Finnish, and Hungarian, stress placement follows predictable rules. For instance, in Czech, stress always falls on the first syllable. In contrast, variable-stress languages, such as English and Russian, lack straightforward rules for predicting stress position. In English, stress shifts can alter a word's morphological category (e.g., noun vs. verb). Similarly, in Russian, stress placement distinguishes grammatical forms or parts of speech in homographs. For variable-stress languages, accurate stress prediction typically requires a stress dictionary and part-of-speech analysis, either explicit or implicit. The inherent complexity of this task has led to the development of benchmarks to evaluate the performance of LLMs in related tasks (Suvarna et al. 2024).

A common approach to solving this problem is to combine a dictionary, if memorization of non-rule-predicted cases is needed, and some set of rules to resolve ambiguous cases. For instance, Ram et al. (2021) utilize phonetic transcription libraries such as eSpeak[27] and Festival[28] to generate pop lyrics. Many English-language systems rely on the Carnegie Mellon University (CMU) Pronouncing Dictionary[29] for stress placement, as seen in Agirrezabal, Gonçalo Oliveira, et al. (2023). However, the effectiveness of this tool is questionable due to its limited coverage

[23] See Section 4.1 on text representation approaches.
[24] https://github.com/Gldkslfmsd/sekacek
[25] https://github.com/Koziev/rusyllab
[26] https://github.com/IlyaGusev/rupo
[27] https://espeak.sourceforge.net
[28] https://www.cstr.ed.ac.uk/projects/festival
[29] https://www.speech.cs.cmu.edu/cgi-bin/cmudict

and inability to account for the prosodic variability of English. E. Greene et al. (2010) address this limitation by proposing an unsupervised approach to stress placement based on finite-state automata (FSA).

Ghazvininejad, X. Shi, et al. (2016) provide a detailed discussion of stress placement, highlighting the importance of secondary stress (included in the CMU Pronouncing Dictionary) and the challenges of adjusting stress patterns in poetic texts. For Spanish-language poetry, Marco et al. (2021) present algorithmic approaches to stress assignment. Agirrezabal, Alegria, and Hulden (2017) describe an approach to stress assignment in lines of English and Spanish poetry using a combination of LSTM, conditional random fields (Lafferty et al. 2001), and a text representation of word embeddings enriched with character-wise embeddings of those words. A comprehensive survey of tools for annotating poetic texts in Indo-European languages, as well as Finnish and Basque, is provided by De Sisto et al. (2024), including links to relevant code repositories.

Navarro et al. (2016) present an automatic scansion system for Spanish poetry and introduce a large, metrically annotated corpus of Spanish sonnets. Their work also evaluates the quality of the markup through two key analyses: (1) a comparison between automatic and manual markup, and (2) an assessment of inter-annotator agreement for the manual markup. Both the annotated corpus and the results of the annotation evaluation are particularly valuable to generative poetry researchers, as they provide essential resources for developing and validating systems that require precise metrical understanding and high-quality poetic data. Rantanplan[30] is another poetry scansion tool for Spanish language, accompanied by evaluation scripts.

Researchers aiming to incorporate stress information into their systems should first evaluate existing libraries for the target language, testing their effectiveness on poetic texts. Developing a custom tool may require significant effort, as it involves both algorithmic design and the preparation of evaluation datasets. This step is crucial for ensuring the accuracy and reliability of stress assignment in poetry generation systems.

3.4 Rhyme Detection

Rhyme is one of the most distinctive features of poetic texts and song lyrics, though it is worth noting that certain poetic genres explicitly avoid rhymes. The requirement for rhyme imposes significant constraints on vocabulary selection, making it a challenging aspect of poetic composition. Inexperienced poets often resort to homogeneous rhymes — such as indefinite or conjugated verb forms in Russian, where suffixes are tightly bound to verb morphology, simplifying the selection of rhyming pairs. When training data for generative poetry systems is sourced from open internet collections (for example, amateur poetry), the prevalence of such low-quality rhymes makes it essential to evaluate the rhyme quality in poems and songs. This necessitates a tool capable of verifying whether two words rhyme, particularly in languages with shifting stress patterns, while accounting for stress placement.

Such tools are highly specialized, and their availability is limited. The following is an analysis of rhyme detection approaches used in research papers on generative poetry, including relevant pre-2017 papers not covered in previous surveys.

- Hirjee and Brown (2009) proposed an algorithmic approach to detecting partial and internal rhymes in rap lyrics.
- Reddy and Knight (2011) discussed the compilation of a rhyme dictionary using n-gram statistics.
- Haider and Kuhn (2018) presented a neural network-based method for rhyme detection in English, German, and French poetry.
- Haider (2021) explored corpus-driven neural models that evaluate prosodic features at both syllable and line levels, leveraging rhythmically diverse datasets.
- Abdibayev, Igarashi, et al. (2021) investigated methods for detecting limericks by analyzing key features such as syllable count, poetic meter, and rhyme scheme.

[30] https://github.com/linhd-postdata/rantanplan

For evaluating rhyme and poetry detection tools, the BPoMP dataset (Abdibayev, Riddell, et al. 2021) provides a notable resource. Constructed using the BLiMP methodology (Warstadt et al. 2020), it contains paired limericks where one text represents a standard poem in the genre while its counterpart presents a deliberately distorted version.

Rhyme detection relies heavily on phonetic information, making phonemecization models and libraries for the target language valuable resources for this task. However, due to the scarcity of ready-made tools, researchers aiming to filter datasets based on rhyme quality must either develop custom solutions or adopt simplified approaches, such as using rhyme dictionaries or stress pattern matching, depending on the phonetic structure of the target language.

3.5 Part-of-Speech Tagging and Syntax Analysis of Poetry

While NLP research has increasingly focused on the architecture and training of LLMs, part-of-speech (POS) tagging and syntactic analysis remain important tools for data engineering and generation analysis. For instance, in languages with frequent homography, a combination of POS tagging and a stress dictionary is a practical approach for resolving homographs when assigning stress marks to poems (see Section 3.3).

Poetic texts often exhibit significant differences in vocabulary and syntax compared to prose, leading to a strong domain shift that can degrade the performance of conventional POS tagging tools. One example of a poetic-specific phenomenon is enjambment (R. Greene et al. 2012, p. 241), which is further explored by Ruiz Fabo et al. (2017). While POS tagging is comprehensively reviewed by Chiche and Yitagesu (2022), we focus here on recent studies that address the unique challenges of the poetic domain.

Kanerva and Ginter (2022) analyze the performance of a parser for Finnish texts across multiple domains, including poetry. Their work introduces a new treebank for Finnish in the Universal Dependencies format and provides a comparative evaluation of UDPipe across several treebanks.

Several studies focus on building treebanks for specific languages, expanding resources for POS tagging and syntactic analysis. Aoyama et al. (2023) present GENTLE, a new mixed-genre English corpus totaling 17K tokens and consisting of 8 unusual text types for out-of-domain evaluation, including poetry. This dataset was manually annotated for a variety of popular NLP tasks, including syntactic dependency parsing. J. Lee and Kong (2012) detail the creation of a dependency treebank for Chinese poetry, comprising 521 poems by ancient Chinese authors.

Colhon et al. (2017) describe the development of a treebank for Romanian, covering multiple domains, including poetry. For Arabic, Habash et al. (2022) introduce CAMELTB, a treebank containing 188K words, with a focus on pre-Islamic poetry. The authors also compare lexicon and syntax across different text domains within the corpus. Additionally, Al-Ghamdi et al. (2021) present a dependency treebank specifically for Arabic poetry.

Finally, we highlight several studies that explore syntactic analysis of poetry in various languages: Latin (Behr 2024; Calvi et al. 2024), Italian (Corbetta et al. 2024), and Chinese (P. Wang, S. Zhang, et al. 2023). The methodologies and insights from these works can inform the development of data preparation pipelines for generative poetry systems.

While data quality and volume are critical determinants of performance, architectural factors — such as tokenization strategies, model design, and decoding algorithms — play an equally central role. We next turn to these implementation aspects in Section 4.

4 Implementation Aspects

This section discusses the options available to researchers for building a LM-based poetry generation system, covering alternative tokenization approaches (Section 4.1), LM architectures (Section 4.2), and LM decoding algorithms (Section 4.3). From an engineering perspective, two main approaches exist: (1) building a generative

system from scratch, beginning with LM pre-training, (2) utilizing a pre-trained LM and fine-tuning it on a custom poetry dataset.

The first approach provides complete control over tokenization, architecture, and training objectives. Such an approach allows developers to carefully balance creativity, linguistic accuracy, and adherence to poetic constraints. However, this method requires substantial computational resources for effective pre-training.

The second, more common approach, inherently constrains these choices. The pre-trained LM typically determines the tokenization algorithm (BPE, for example), though some researchers have explored post-hoc modifications to character (C. Yu et al. 2024) or syllable-level (Chudoba and R. Rosa 2024) tokenization through additional training. Consequently, this approach offers two primary degrees of freedom for the researcher: (1) the data engineering process for constructing the training corpus, and (2) the selection of decoding algorithms during generation. The main disadvantages of using pre-trained LLMs include potential negative effects of subword tokenization on metrical poetry generation and the often unclear influence of pre-training corpus composition on literary quality. Since LLM developers typically prioritize metrics from domains like mathematics or reasoning, their models may be suboptimal for poetic generation tasks.

When selecting a pre-trained LM from available alternatives, researchers should consider several factors:

- Target language(s) and expected generation quality for this language. Some languages may be underrepresented in the LM pretraining corpus, resulting in generation quality degradation for those languages.
- Available computational budget, which constrains model capacity.
- Maximum context length requirements for longer poetic forms or tasks like prose-to-poem.[31]
- Potential benefits of multimodality for certain applications.
- Architectural variations (e.g., mixture-of-experts vs. dense models, layer count, vocabulary size) that may affect poetry generation metrics.

Several model families (e.g., Qwen[32]) offer multiple variants along these parameters, enabling researchers to begin with smaller prototypes and scale their solutions as needed.

4.1 Tokenization

Most work on generative poetry, particularly within the period covered by this survey, relies on LMs. These models present two inherent challenges for poetry generation: (1) they typically operate on subword tokenizations, which can misalign with phonological units, and (2) pre-trained LMs are generally trained on written text, which lacks explicit phonetic or prosodic information. Consequently, using standard LMs in poetry generation pipelines creates clear difficulties for maintaining metrical structure and modeling rhyme. In this section, we discuss text tokenization approaches designed to mitigate the aforementioned specific issues.

An alternative approach involves moving beyond discrete representations by adopting continuous space methodologies, as seen in chain-of-thought research (S. Hao et al. 2025). However, no current work on generative poetry has adopted this methodology. We also exclude discussions of traditional vector representations used in generative poetry systems for tasks such as measuring text similarity, finding parodic samples, or controlling generation flow. These standard representations typically do not encode poetic attributes and thus fall outside the scope of this survey.

For transformer architectures, which are currently dominant in generative poetry, subword tokenization methods such as BPE and SentencePiece (Kudo and Richardson 2018) are the most widely adopted. Many generative poetry systems fine-tune popular LMs on poetry or song lyrics datasets while retaining the base model's tokenization (Hämäläinen, Alnajjar, and Poibeau 2022; Panahandeh et al. 2023; Yee-king et al. 2023).

[31] See more on prose-to-poem task at Section 2.6.
[32] https://huggingface.co/Qwen/collections

However, some systems adapt the tokenization process to better align with poetry constraints. For example, Lo et al. (2022) reverse the order of tokens within lines to improve rhyme generation. This approach of reversing the order of tokens in each line of the poem has been repeatedly utilized in other works (Ou et al. 2023; Van de Cruys 2020; Xue, K. Song, et al. 2021). While this technique facilitates rhyme generation, it inevitably restricts the pretrained model's ability to leverage transfer learning effectively. This limitation arises because the model was exclusively exposed to left-to-right word order during pretraining.

An alternative approach, proposed by Pasini et al. (2024), addresses this issue by introducing a specialized method for rhyme specification. In this approach, the rhyme is explicitly indicated at the beginning of each line. The model first selects a rhyme during generation and then proceeds to generate the entire line, ensuring it concludes with the preselected rhyme.

Recent work on Claude 3.5 Haiku (Transformer Circuits Thread 2025) demonstrates that LLMs can employ a planning mechanism to generate rhyming poetry, rather than relying solely on improvisation during autoregressive decoding. While LMs typically predict text sequentially, the model exhibits evidence of pre-planning line-final rhyme words before composing the full line. Specifically, feature activation patterns suggest that the model generates candidate rhyming words early in the decoding process, integrates these candidates into its latent state, and uses this information to guide the generation of semantically coherent lines that satisfy the rhyme scheme.

A notable challenge arises when combining subword unit tokenization with Chinese word segmentation, as highlighted by X. Zhang et al. (2023). Haslett (2025) examines how suboptimal segmentation of Chinese text affects semantic representation in models such as GPT-4, GPT-4o (OpenAI 2024), and Llama 3 (Llama Team 2024). To address the mismatch between tokenization and poetry text segmentation, C. Yu et al. (2024) propose transitioning the Qwen model to a token-free tokenization approach.

In the context of poetry, syllable-level tokenization offers a promising alternative to subword tokenization methods. We discuss its advantages and limitations in Section 4.1.2.

The following subsections detail alternative tokenization methods to BPE for poetic texts.

4.1.1 Character-level Text Representation. Tokenization methods relying solely on subword frequency information often yield inconsistent performance for multilingual texts or rare words. Character- or byte-level tokenization can address these limitations, particularly benefiting generative tasks requiring fine-grained character-level processing. Poetry generation exemplifies such tasks, as models must accurately capture language phonetics. In many languages, pronunciation correlates significantly with spelling, making character-level representations especially valuable for poetry systems.

Several studies in generative poetry support this approach. C. Yu et al. (2024) demonstrate that character- and byte-level tokenization outperforms subword unit tokenization for Chinese poetry due to the language's unique writing characteristics. Similarly, Y. Chen, Gröner, et al. (2024) find character-level models superior for generating diverse poetry, while Belouadi and Eger (2023) show that such models outperform transformer architectures based on subword unit tokenization (GPT-2 and T5) in English and German quatrain generation.

A specific application appears in work with Russian accentual-syllabic poetry by Koziev and Fenogenova (2025). Their approach involved: (1) training character-level LMs of varying capacities on mixed poetry/prose corpora, then (2) fine-tuning them on instructional datasets containing poetry samples.

Beyond phonetics, linguistic creativity — such as the ability to coin new words from morphemes (Kitzlerová 2022) — is crucial for generative poetry. For morphologically rich languages, subword unit tokenization often misaligns with morpheme boundaries (Ismayilzada, Circi, et al. 2025), making character-level tokenization a preferable alternative.

Before the widespread adoption of transformer architectures, character-level representations were commonly used with recurrent LMs (Alyafeai et al. 2023; Hopkins and Kiela 2017; Lau et al. 2018; Popescu-Belis et al. 2022;

Tikhonov and Yamshchikov 2018b; S. Xie et al. 2017). With the rise of transformer architecture, researchers have explored character- and byte-level tokenization in this framework. Pre-trained models like CANINE (J. H. Clark et al. 2022) and ByT5 are now available for research in generative poetry. CANINE, an encoder-decoder Transformer, uses Unicode codepoints as tokens, while ByT5 employs byte-level tokenization, making it well-suited for generative tasks, including poetry. For example, Z. Zhang et al. (2024) fine-tuned CANINE to weight syllable tokens for melody-to-lyrics generation. Additionally, Belouadi and Eger (2023) trained character-level decoder transformers from scratch, comparing them to ByT5 and GPT-2.

However, byte-level models face challenges in poetry generation for languages with non-Latin alphabets. For instance, Cyrillic letters are encoded as two bytes in UTF-8 encoding, effectively doubling the token count and reducing the effective context length during both training and inference. Similarly, character-level approaches suffer from increased fertility,[33] which results in significantly longer token sequences. Longer sequences, in turn, negatively impact the computational budget required for pre-training (due to the quadratic complexity of self-attention mechanisms in many LLMs), as well as generation time and memory consumption.

To address the limitations of character-level representations, researchers have proposed innovative solutions. Pagnoni et al. (2025) combine the computational efficiency of BPE with character-level granularity using dynamic byte chain packing via an auxiliary transformer. Deiseroth et al. (2024) introduced T-FREE, which represents words as combinations of character trigrams, enabling morpheme-level processing and improved performance on low-resource languages. L. Sun, Luisier, et al. (2023) proposed a two-stage pipeline with a shallow transformer to learn word representations from characters, and Tay et al. (2022) developed CharFormer, which uses a gradient-based subword tokenization module to learn latent subword representations from byte sequences.

Despite the dominance of BPE and unigram tokenization in mainstream LMs, character- and byte-level approaches continue to evolve. Given their strong performance in poetry generation, experimenting with these newer character-level tokenization methods, even in smaller LMs, holds promise for researchers in generative poetry.

4.1.2 Syllable-level Tokenization. Syllable-level tokenization arises naturally when it comes to syllabic or accentual-syllabic versification, and also in some cases when working with song lyrics (D. Zhang et al. 2022). This type of tokenization can be considered a specialized form of subword tokenization, where each token corresponds to a syllable. For example, Zugarini, Melacci, et al. (2019) successfully employed syllable-level tokenization for generating Italian poetry. Koziev and Fenogenova (2025) describe experiments with generating Russian accentual-syllabic poetry using small LMs with syllable-level tokenization, trained from scratch. The results show that a small LM with syllable-level tokenization can compete in the task of generating poetry with large models with subword unit tokenization.

Unfortunately, despite all the obvious advantages in terms of compatibility with syllabic and accentual-syllabic poetry, syllabic tokenization is not without its drawbacks, which make it difficult to use with LMs and hinder its widespread adoption. First of all, syllable-level tokenization is not well-suited for general-purpose language modeling tasks, and as a result, there are no open pre-trained LMs available. Researchers must train such models from scratch.

Chudoba and R. Rosa (2024) propose a workaround by combining GPT-2 tokenization with syllable-level text splitting. Their evaluation of the Czech poetry generation shows promising results. However, it would be beneficial to include metrics that can assess the impact of the modified tokenization on the grammaticality of the generated text. A similar approach, using standard GPT-2 tokenization with syllable-level text splitting, is mentioned in a study on Vietnamese poetry generation by Nguyen et al. (2021).

[33]The term "tokenization fertility" was coined in https://juditacs.github.io/2019/02/19/bert-tokenization-stats.html.

Second, syllable-level tokenization is computationally more expensive than other tokenization methods, especially for languages where syllable boundaries are not easily inferred from surface characters.[34]

To summarize the above, it can be said that syllable-level tokenization, on the one hand, can be a very effective representation of text for the tasks of generative poetry, but its inherent technical shortcomings should be carefully taken into account when choosing a tokenization method for a poetic LM.

4.1.3 Word-level Tokenization. Using word-level tokenization in generative models with a large or unbounded dictionary is challenging, if not impractical, due to the open-ended nature of natural language. Consequently, word-level text representation is rarely used in poetry generation systems in papers published after 2017. One notable exception is S. Xie et al. (2017), who explored a hybrid approach combining word-level and character-level representations in recurrent and convolutional models.

Another example is Tikhonov and Yamshchikov (2018b), which employs word-level representation using vectors of connected embeddings. These embeddings encode semantic, phonetic, and character-level information, along with additional features to capture the style of specific poets.

EMILY (Shihadeh and Ackerman 2020) is a pipeline for generating poems in the style of Emily Dickinson. The system uses a generative Markov chain model, operating at the word level. This design represents a practical compromise, chosen due to the extremely limited training data (10,178 lines from 444 poems) and the goal of replicating the author's style without employing an LLM.

Given the inherent limitations of word-level tokenization, such as difficulties handling neologisms, occasionalisms, and the challenges posed by agglutinative languages like Turkish and Finnish — where word derivation is highly frequent due to their morphological structure — this approach is generally not recommended for new studies. If considered, the implications of this choice must be carefully evaluated to ensure the model's ability to handle creative and unconventional language use.

4.1.4 Phonetic Representation of Text. To complete our discussion of text representation options for poetic LMs, we briefly address *phonemization*, a rarely used but intriguing approach. Phonemization represents text as a sequence of phonemes using a specialized phonetic alphabet. Libraries like Phonemizer (Bernard and Titeux 2021) and Transphone (X. Li et al. 2022) facilitate this process, enabling the grapheme-to-phoneme (G2P) conversion.

There are a few works that study LMs with phonemic representation of text, mainly for acoustic speech recognition (ASR) and text-to-speech (TTS) tasks. This list includes several encoder models: Phoneme-BERT (Sundararaman et al. 2021), Mixed-Phoneme BERT (G. Zhang et al. 2022), XPhoneBERT (The Nguyen et al. 2023), BORT (Gale et al. 2023). An example of the use of phonemic representation of text for the tasks of generative poetry is provided by Hopkins and Kiela (2017), who trained two stacked LSTMs on English poetry texts encoded phonetically using the CMU dictionary[35] and rules for out-of-vocabulary words. The model was trained on sequences of 150 phonemes corresponding to four lines of a sonnet quatrain.

Unfortunately, phonemization has notable drawbacks for generative poetry tasks. First, the process is cumbersome, requiring third-party libraries and additional computational resources that may be unavailable for the target language. Second, converting phonetic representations back into standard text is challenging, particularly for languages like English. This difficulty arises in part due to the abundance of homophones in English — words that are pronounced the same but spelled differently (for example, "to," "too," and "two"). Such linguistic features complicate the disambiguation process during conversion, as phonetic representations alone often lack sufficient context to determine the correct orthographic form.

[34] The complexity of syllabification for such languages is discussed in detail in Section 3.2.
[35] https://www.speech.cs.cmu.edu/cgi-bin/cmudict

Given the limited research employing phonetic representation, its inclusion in new architectures should be approached with caution. However, it is important to recognize that certain tasks critically depend on phonetic information, making studies based on this approach potentially valuable. Such research can yield unexpected and insightful results, not only in the context of poetry generation but also as interdisciplinary scientific contributions that explore connections to cognitive processes, such as brain function. For instance, the generation of tongue twisters (Loakman et al. 2024) exemplifies a task where phonetic considerations are essential, much like in generative poetry, while also requiring a balance with creative generation.

4.2 Models and Algorithms

Research on generative poetry has explored a wide range of language modeling paradigms. Early systems relied on explicit structural representations, such as templates and handcrafted rules. Later work introduced evolutionary algorithms, reinforcement learning (RL), and recurrent neural networks (RNN). More recently, transformer-based LMs and diffusion models, often also using the transform model for the task of text denoising, have become the dominant approaches.

Figure 1 shows the distribution of the algorithmic approaches discussed in this survey over time. Template-based methods are mainly found in earlier work and show little growth, likely due to their dependence on manually defined structures and resources. RNN-based approaches, especially those using LSTM and GRU architectures, are most common between 2017 and 2020, in line with their broader use in neural language generation during that period. From 2021 onward, most new studies adopt transformer-based models, reflecting general trends in natural language generation. The figure is intended as a descriptive summary of the surveyed literature rather than a measure of research impact or quality. Details on the construction of the visualization are provided in Appendix C.

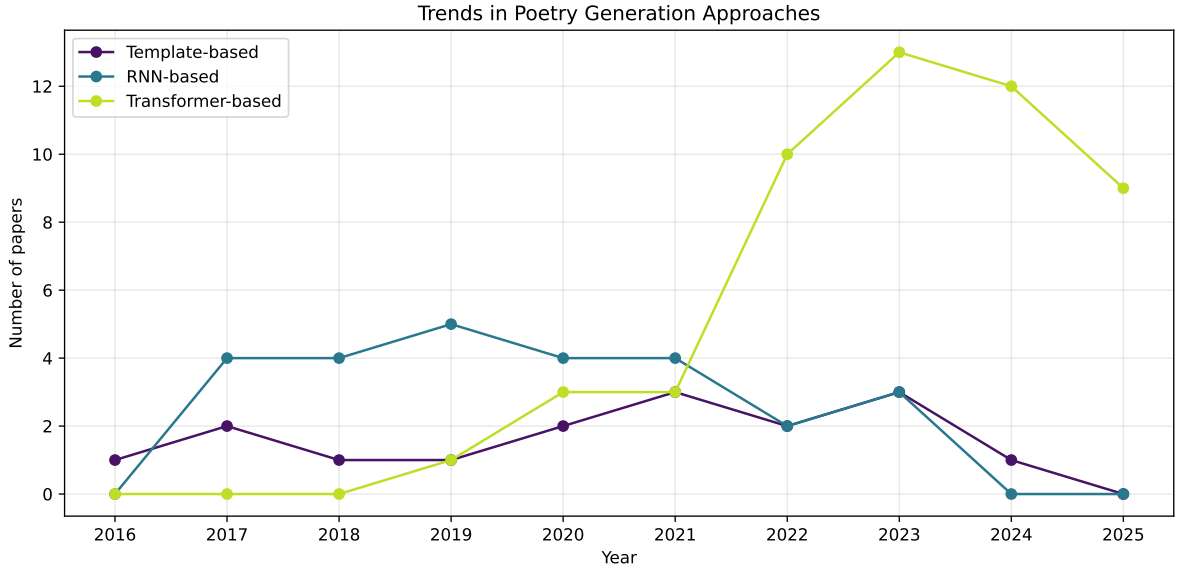

Fig. 1. Number of publications per year employing template-based, RNN-based, and transformer-based approaches for poetry and song lyric generation between 2016 and 2025. Counts are based on the papers included in this survey and are intended to illustrate broad methodological trends rather than provide an exhaustive bibliometric analysis.

This section reviews mentioned methods in a structured manner, beginning with template-based approaches (Section 4.2.1), then covering RNN (Section 4.2.2), transformer (Section 4.2.3) and text diffusion methods (Section 4.2.4), and finally mentioning other approaches like RL, adversarial networks and evolutionary algorithms (Section 4.2.5).

4.2.1 Template-based Approaches. Before the introduction of transformer-based LMs, poetry generation often relied on manually designed templates or those automatically derived from reference poems or song texts. Representative systems from this period include:

- **Full FACE** (Colton, Goodwin, et al. 2012), which generates English-language poems using templates constrained by features such as rhyme, meter, and word choice.
- **PoeTryMe** (Gonçalo Oliveira 2012) and its adaptation to Spanish (Gonçalo Oliveira, Hervás, et al. 2014), which use user-defined template structures for lines and stanzas to generate Portuguese and Spanish poetry.
- **Tra-la-Lyrics** (Gonçalo Oliveira 2015), which extends the PoeTryMe framework by integrating a rhythm module with semantic templates for song lyric generation.
- **InkWell** (Gabriel 2016), which focuses on generating English-language haiku.

For a more detailed historical overview of template- and rule-based poetry generation systems, including those relying on human-written text corpora, we refer readers to Lamb, Brown, and C. L. A. Clarke (2017).

In addition to templates, these systems rely on linguistic resources such as simile collections, frequency dictionaries, thesauri, and knowledge bases. In recent work, template-based approaches have been largely replaced by systems based on LLMs.

Although template-based methods are no longer the dominant approach in poetry and song lyric generation research, templates are still used as a core or auxiliary mechanism for controlling generation in some recent systems. For example, Bay et al. (2017) describe the Lyrist system for generating song lyrics and poems, which performs word replacement in human-written texts based on vector word similarity. In this system, templates are constructed dynamically and filled during generation.

Manjavacas et al. (2019) propose a hip-hop lyrics generation system that extracts conditional templates from text snippets to guide LM decoding. These templates are designed to enforce rhythmic and rhyming constraints and are used together with LSTM-based character-level and word-level LMs. The system relies on syllabified text and phonological representations from the CMU Pronouncing Dictionary.

Hybrid approaches that combine templates with neural LMs have also been explored. J. Wang et al. (2021) present a system that derives templates from a small set of examples and uses GPT-2 to fill the resulting slots. A related approach is described by P. Li et al. (2020), who generate poems in hard forms (Fussell 1979, page 127), such as English sonnets (Lotman 2013) and Chinese SongCi poetry (Lang and Liangzhi 2024), using templates to constrain generation. N. Liu et al. (2022) also use a combination of user-provided templates describing the structure of song lyrics and a transformer model that generates song lyrics based on a given template. Similarly, Murakami and Terai (2023) combine automatically extracted templates with BERT in a Japanese song lyric generation pipeline.

PoeLM (Ormazabal et al. 2022) is another example of a hybrid approach that combines templates with a GPT-2-based transformer LM. The system controls generation at the line level using special code prefixes, whose sequence defines the high-level structure of the poem, including the number of syllables and the rhyme scheme.

Similarly, Agnew et al. (2023) propose a sonnet generation system that integrates approximately 120 handcrafted templates with part-of-speech slots, while relying on GPT-2 for slot filling. Finally, T. Zhang, M. Lee, et al. (2023) introduce a framework for generating text from automatically constructed templates. Although not directly focused on poetry, this approach may be relevant in settings where predictable and controllable poetic output is required.

Qian et al. (2023) use automatically constructed compound templates as an intermediate representation for Chinese song lyric generation. These templates encode high-level information about both the melody and the lyrics, and the generative model produces the final lyrics by conditioning on this representation.

Template-based methods remain a common foundation for systems that generate jokes and short humorous texts (Amin and Burghardt 2020; Goel et al. 2024; Winters et al. 2018), riddles (Gonçalo Oliveira and Rodrigues 2018; Terai et al. 2020), slogans (Alnajjar and Toivonen 2021; Repar et al. 2018), and puns. Several of these genres partially overlap with poetic forms, particularly in the case of riddles and humorous verse, and often rely on structured mechanisms for constructing metaphors, similes, and associative mappings. As a result, techniques developed for template-based humor generation may be relevant to generative poetry research. This connection is further strengthened by the use of modern LLMs as flexible template instantiation mechanisms capable of producing linguistically varied realizations from fixed structural patterns (T. Zhang, M. Lee, et al. 2023).

Template-based approaches alone were limited in the diversity and flexibility they could provide in generated poetry. RNN-based neural networks were the first widely used models to address this limitation, with LSTMs dominating early NLG research (Gatt and Krahmer 2018). Although transformers are now standard, RNNs remain relevant and are briefly reviewed in the next subsection.

4.2.2 RNN-based Approaches. Transformer-based LMs have largely replaced RNN-based solutions due to their superior scalability and pretraining performance. However, some post-2017 poetry generation systems still utilize LSTM or GRU architectures. Taking this into account, as well as the unclear prospects of this architecture, we decided to mention these works in this survey without going into details of the implementation and results:

- Lau et al. (2018) describe the Deep-speare system, which jointly models language, meter, and rhyme using word- and character-level LSTMs.
- X. Yi, R. Li, and M. Sun (2018) proposed a Chinese poetry generation system based on LSTM with attention.
- B. Liu et al. (2018) used a GRU-based model to generate poem texts conditioned by an input image.
- Van de Cruys (2019) described a GRU-based system for generating French poetry.
- Aguiar and Liao (2019) implemented a haiku generation system that uses beam search and an LSTM model.
- Manjavacas et al. (2019) uses character- and word(syllable)-level LSTMs along with control templates to generate hip-hop lyrics.
- Agarwal and Kann (2020) used an LSTM network to generate English acrostic poems.
- Røstvold and Gambäck (2020) explored a poetry generator combining a bidirectional LSTM with rhyme pair generation, rule-based word prediction, and tree search to expand generation possibilities.
- Van de Cruys (2020) presented an encoder-decoder RNN architecture for generating English and French poetry. Their system incorporates unique features, such as training on non-poetic texts, reversing token order within lines to improve rhyme selection, and controlling token probabilities to enforce poetic constraints.
- Wöckener et al. (2021) conducted experiments with RNNs to generate English poetry using a small training corpus.
- Heerden and Bas (2021) used a two-layer LSTM architecture with 50 units per layer, trained on a 208,616-word contemporary novel, to generate Afrikaans poetry in a machine-in-the-loop setting.

4.2.3 Transformer-based Approaches. Currently, transformer models for poetry generation are typically used as pre-trained architectures fine-tuned on poetry datasets of various formats. These models, pre-trained on hundreds of billions of tokens, excel at generating highly grammatical texts. However, shortly after the introduction of the transformer architecture, researchers also explored custom models incorporating transformers. For example, Takeishi et al. (2022) developed a Japanese poetry generation system combining a transformer with a variational autoencoder. This system accepts user-specified keywords and generates tanka (Ishikawa 2016) — a type of

fixed-form poetry. Another example is Y. Sun et al. (2023), who proposed a song rewriting system based on a transformer encoder-decoder model. Their model, pre-trained on Baidu Encyclopedia[36] texts and fine-tuned on scraped song lyrics, assists users in creating song lyrics from existing text.

Many poetry generation tasks[37] are naturally framed as sequence-to-sequence problems, making encoder-decoder architectures a convenient choice. Typically, researchers opt for pre-trained T5 models (Chakrabarty, Padmakumar, et al. 2022). However, the simplicity and efficiency of decoder-only architectures, along with the availability of open pre-trained models, have also made them popular for such tasks despite research papers (Qorib et al. 2024) showing the advantages of encoder-decoder architectures on many similar problems.

Upon its public release in 2019, the GPT-2 model offered state-of-the-art English text generation capabilities alongside the facility for further training. These features led to its widespread adoption in generative poetry research. As a result, research papers employing GPT-2 continued to appear through 2023.

For instance, Pardinas et al. (2023) fine-tuned the 140M version of GPT-2 to generate haiku (Harr 1975). They implemented a RLHF scheme: readers rated the generated haiku, and a reward model was trained based on this feedback to improve the generative model. Sawicki, Grzes, A. Jordanous, et al. (2022) fine-tuned GPT-2 to replicate the poetic style of two selected English-language authors.

For non-English poetry generation, variants of GPT-2 trained on additional language-specific data are often employed. For example, Hämäläinen, Alnajjar, and Poibeau (2022) developed a French poetry generation system based on RoBERTa and BelGPT-2 (Louis 2020). Abboushi and Azzeh (2023) fine-tuned the AraGPT2 (Antoun et al. 2021) model on an Arabic poetry corpus and evaluated the results using both automatic metrics and human experts. Chudoba and R. Rosa (2024) experimented with a small GPT-2 model trained on Czech texts for Czech poetry generation. Additionally, Tian and Peng (2022) proposed a method for fine-tuning LMs to generate Spanish syllabic poetry without requiring poetry-specific training data. Their approach involves training the model to generate lines with a specified number of syllables and clauses, combined with ranking the generated results.

Following the release of more capable, advanced, and multilingual publicly available LLMs — such as LLaMa and Qwen — researchers began transitioning to these more powerful alternatives. For example, C. Yu et al. (2024) fine-tuned a token-free model based on Qwen-7B-Chat (Bai et al. 2023) to generate Chinese classical poetry following complex instructions. Huynh and Bao (2024) explored the applicability of the BLOOM (BigScience Workshop 2023) model for Vietnamese poetry generation. An evaluation of the quality of poetry, namely diversity for quatrains generated using LlaMa3-8B, can be found in the paper by Y. Chen, Gröner, et al. (2024). The SongComposer system for music and song lyrics generation, described by Ding et al. (2025), uses the InternLM2-7B (Z. Cai et al. 2024) as base LLM.

Other transformer-based models have also been applied to poetry generation tasks. For example, Boggia et al. (2022) used the mBART model fine-tuned on the Gutenberg Poetry Corpus (Parrish 2016) to generate poetry line by line. Their system employs two models: one generates the starting line based on keywords, and the other generates subsequent lines based on previous ones. Similarly, Z. Wang, Guan, et al. (2023) used BART to generate Chinese classical poetry, with inputs including the first line of the poem, theme words, and a list of keywords to be embedded in the output. Riedl (2020) utilized XLNet (Z. Yang, Z. Dai, et al. 2019) for slot-filling to generate parody lyrics. Qian et al. (2023) use the mT5 model for Chinese song lyric generation.

Classical encoder architectures, particularly BERT, are also widely used for tasks such as classifying themes in Chinese poetry (Hou and S. Zhang 2024), embedding poetic texts (Z. Guo, J. Hu, et al. 2020), metaphor retrieval (M. Choi et al. 2021), and multilingual poetry analysis, as demonstrated by the ALBERTI model (J. d. l. Rosa et al. 2023), trained on the PULPO corpus containing poems in 12 languages. Baral et al. (2021) applied RoBERTa to generate "blackout poetry" by leveraging its masked language modeling capabilities.

[36] https://baike.baidu.com

[37] See Section 2 on poetry generation tasks.

4.2.4 Diffusion Language Models. While transformer-based architectures dominate the field of NLP, alternative approaches such as diffusion LMs (Z. Hu et al. 2024), state-space models such as Mamba (A. Gu and Dao 2024), and novel LSTM variants (Beck et al. 2024) continue to show promise, particularly in the context of poetry generation systems. Among these, diffusion models have garnered significant attention due to their remarkable success in image-related tasks, inspiring their adaptation to language modeling (Q. Yi et al. 2024).

Recent work by Z. Hu et al. (2024) introduced PoetryDiffusion, a novel framework that integrates a diffusion model for semantic generation with a metrical controller to enforce structural constraints such as rhythm and format. This approach has demonstrated impressive results in generating sonnets and Chinese SongCi poetry, highlighting the potential of diffusion models for creative text generation.

Further advancing this line of research, Nie et al. (2025) explored the capabilities of generative diffusion models in the context of poetry composition. Their work leverages the LLaDA (Nie et al. 2025) model to tackle the challenging task of generating both subsequent and preceding lines of a poem, showcasing the versatility and robustness of diffusion-based architectures in handling sequential and context-sensitive text generation tasks.

Beyond template-based, neural, and diffusion models, a few alternative methods have been explored in the generative poetry domain. We discuss them below.

4.2.5 Other Approaches. In addition to template-based, recurrent, transformer-based, and diffusion models, a number of alternative approaches to poetry and song lyric generation have been explored. These methods are used relatively rarely and have not developed into dominant paradigms, but they address specific aspects of creative text generation and therefore deserve mention in this survey. In particular, evolutionary algorithms, RL, and adversarial training frameworks have been applied to poetry generation, often with the goal of incorporating explicit quality objectives, user feedback, or external evaluation signals into the model training process.

Research employing evolutionary approaches for generative poetry has remained relatively limited. Notable examples include the Finnish poetry generation system proposed by Hämäläinen and Alnajjar (2019a) and an evolutionary approach to Tang poetry generation described by Mu et al. (2020). Even prior to the time span covered by this survey, work based on genetic algorithms was confined to a small number of studies (Levy 2001; H. M. Manurung 2003; R. Manurung et al. 2012; Soo et al. 2015).

Another approach to poetry generation involves RL, though its applications remain limited. RL is primarily used to enhance generative LMs through user feedback, as demonstrated in haiku generation (Pardinas et al. 2023) and Chinese poetry generation (Y. Song 2022b). In this scheme, reinforcement learning from human feedback (RLHF) acts as an alternative way of retraining an LM whose architecture is not limited to transformers.

An alternative RL approach employs automatic evaluation metrics as reward signals. For instance, X. Yi, M. Sun, et al. (2018) trained two GRU-based generators using rewards based on poeticity, grammaticality, coherence, and meaningfulness. Zugarini, Pasqualini, et al. (2021) proposed a different method where poems are iteratively refined through word replacement. Their architecture uses separate models: a "detector" to identify problematic words and a "promoter" to suggest replacements. Only the detector participates directly in RL, learning to select words that require modification. This iterative refinement process mirrors human poetic composition (Oliver 1994, p. 109; Addonizio and Laux 1997, p. 186).

Another way to use external evaluation of the generation quality to improve the efficiency of the LM is generative adversarial networks (GANs) (Goodfellow et al. 2014). This approach assumes that there are reference texts that have a certain quality, for example, creativity. The discriminator learns to distinguish between the LM generations and reference texts, and the generative part learns to "fool" the discriminator. Due to the complexity of applying GANs to the text domain, research in the field of generative poetry based on this method is limited to a few works (Y. Chen and Lerch 2020; Saeed et al. 2019; L. Yu et al. 2017; Y. Yu et al. 2021). Based on these works, it is difficult to draw conclusions about the potential effectiveness of GANs, but it can be stated that offline RL

methods and their implementation in ready-to-use libraries (Werra et al. 2020) allow solving at least the same problems.

4.3 Text Generation

The task of generating text for poems and song lyrics, as a subset of NLG, involves several unique challenges. A key challenge is the presence of strict constraints on the text, such as meter, rhyme, and structure (P. Li et al. 2020). Generative transformer models, which have become the standard tool for poetry generation, use autoregressive text generation. In this approach, the text of a poem is produced token by token, from the beginning to the end. At each step, the next token is chosen based on a probability distribution over the entire vocabulary. Below, we discuss various methods for selecting tokens.

Common LM decoding methods include nucleus sampling (top-p) and top-k sampling, along with their variants such as typical-p (Meister et al. 2023). While numerous new decoding algorithms and their modifications exist in the literature (Das et al. 2025; Garces Arias et al. 2024; Gareev et al. 2024; J. Guo et al. 2025; Luo et al. 2025; Minh et al. 2025; Pynadath and R. Zhang 2025; Tu et al. 2024; H. Yang et al. 2024; W. Zhu, H. Hao, et al. 2024), their effectiveness for generative poetry tasks remains largely unverified. The following discussion is therefore limited to alternative decoding methods for which specific evaluations on poetry generation are available.

X. Zhang et al. (2023) propose a new decoding method called "nucleus sampling with flattened head (NS-FH)" for modern Chinese poetry generation. This approach modifies the predicted word distribution by randomizing high-frequency words while preserving low-frequency selections, improving the novelty of generated poems compared to standard sampling techniques. The authors developed NS-FH to address limitations in conventional top-p and top-k sampling when used with their transformer-based LM. These standard methods often produced repetitive and predictable outputs, negatively impacting poetic quality. Through frequency analysis, they demonstrate how different NS-FH configurations influence the characteristics of the generated poetry.

FUDGE (K. Yang and Klein 2021) is a controlled text generation method that uses a plug-and-play approach. It modifies the token probability distribution from a base LM using additional discriminators, steering the autoregressive generation toward texts with desired properties. The method works by adjusting the LM's next-token probabilities to favor tokens that are more likely to produce the target text property in subsequent generation steps. This adjustment is calculated using a pre-trained binary classifier that evaluates token significance. The authors demonstrate FUDGE's effectiveness on several tasks, including couplet completion.

Panahandeh et al. (2023) propose dynamic temperature control during decoding: the temperature is set higher at the beginning of each line to encourage exploration of the search space, then gradually decreases as the line progresses to improve adherence to metrical and rhyme constraints. Dhuliawala et al. (2024) extend this idea by developing a framework for automatically controlling the decoding temperature at the token or text level, demonstrating its effectiveness in story generation and mathematical problem-solving tasks.

Chudoba and R. Rosa (2024) describe a line-by-line generation approach using "forced generation," where a service prompt (for example, rhyme information) is provided before each line to enforce a specific rhyme scheme. Similarly, Ormazabal et al. (2022) propose a method for generating Spanish and Basque poetry without fine-tuning GPT-2 on poetry data. The model is trained to generate text from a given list of string descriptors that specify the required number of syllables and the string representation of the last syllable in the string. This proves sufficient to generate poems in the specified languages. Amina et al. (2025) extended this approach to generate poetry in Bangla.

Autoregressive generation can also be modified to incorporate additional constraints beyond the model's token probability distribution. For instance, Geng et al. (2023) propose grammar-constrained decoding, which uses formal grammar to guide the generation of structured text. Agnew et al. (2023) apply a similar approach to poetry generation, using beam search constrained by handcrafted templates that specify fixed text fragments and slots

to be filled. Riedl (2020) uses XLNet to fill template slots (for example, mask tokens) after selecting rhyme words. Zou (2025) proposes block inverse prompting, a constrained generation framework designed to emulate the human text-writing process. This framework leverages block generative models to enhance zero-shot generation quality, particularly for challenging tasks such as open-domain traditional-form Chinese poem generation.

Roush et al. (2022) propose a plug-and-play approach for generating English poetry. Their method does not require modifications to the underlying LM, which they highlight as its key advantage. The authors classify poetic text constraints into two categories:

- **Soft constraints**: Style-related requirements (e.g., generating text in a specific poetic form).
- **Hard constraints**: Token-level requirements, particularly phonetic ones.

At each generation step, tokens violating hard constraints are suppressed in the probability distribution, while the remaining tokens undergo standard sampling. Phonetic constraints are enforced using CMUdict. The authors identify one limitation: the method performs poorly with LMs that use small vocabularies, as low-frequency words are often split into subword tokens, disrupting the constraint system. For evaluation, they tested GPT-2-medium on a synthetic task (generating text without the letter "e"). However, the study does not assess the quality of the generated poems, making it difficult to compare with other poetry generation approaches.

Beam search is a widely used text generation method; however, it is less popular in automatic poetry generation systems due to its tendency to produce repetitive or locally optimal outputs. This limitation significantly reduces the diversity of generated texts, resulting in less effective use of the inference budget for creative tasks. To address this issue, several variants of beam search have been proposed, such as diverse beam search (Vijayakumar et al. 2018), grid beam search (Hokamp and Q. Liu 2017), self-evaluation guided beam search (Y. Xie et al. 2023), and creative beam search (Franceschelli and Musolesi 2024a). These methods aim to enhance the diversity and creativity of generated outputs while mitigating the limitations of standard beam search.

Improving text generation can be achieved not only by modifying low-level token selection algorithms based on the probability distributions produced by the LM. Another approach involves multi-step, plan-and-solve schemes (L. Wang et al. 2023), where the LM iteratively refines the conditions specified in the user prompt, ultimately generating a poem with the desired properties. For example, Young et al. (2024) demonstrate how this approach enhances diversity metrics in both poetry and code generation tasks.

The choice of decoding algorithm and its parameters significantly impacts the quality of generated poems. For example, dynamically adjusting decoding parameters based on the poem's structure could improve overall quality, as poems often exhibit a clear two-dimensional structure distinct from prose. Additionally, exploring speculative decoding or similar techniques to optimize inference speed presents an interesting avenue for future research.

5 Evaluation of Generative Poetry

Evaluating how effectively a system composes poems or lyrics is a critical step in generative poetry research. This section provides an overview of evaluation metrics, their calculation methods, and the applicability of current practices in the field. One consequence of the proliferation of open-source LMs capable of text generation is the need to develop automatic methods and protocols for objectively evaluating their outputs. For generative models that solve mathematical problems, answer exam questions, or write code, there has been active development of diverse datasets, benchmarks, leaderboards, and even the use of LMs themselves for evaluation and comparison. However, in the case of automatic poetry generation, such tools for rapid comparison and metric measurement of new models are virtually nonexistent. Even side-by-side human evaluations across different studies are often incomparable due to significant differences in protocol details.

Poetry evaluation approaches can be categorized as follows:

Offline vs. online evaluation: Offline evaluation assesses pre-collected poems, either to filter low-quality samples[38] or to compare outputs from different models, architectures, and training configurations. Online evaluation (generation ranking) selects the best outputs in real time by comparing multiple generations for a given prompt.

Automatic metrics vs. human evaluation: Automatic metrics are widely used for evaluating generative poetry (Section 5.1), yet human assessment (Section 5.2) remains essential for judging poetic quality. However, human evaluation faces challenges of high cost, limited reproducibility, and poor scalability. These limitations have motivated the development of automated alternatives.

A promising compromise is the LLM-as-a-judge approach, whose popularity grows alongside improvements in both open-source instruction-tuned models and proprietary API-accessible LLMs. Section 5.1.1 examines applications of this method for evaluating generative poetry.

The need for automated, reproducible evaluation of LMs on open-ended tasks has driven the development of LLM-based benchmarks. However, current benchmarks remain limited for generative poetry evaluation. For example, the BiGGen Bench framework (Kim, Suk, et al. 2025) — which employs LMs to conduct fine-grained, instance-specific evaluations instead of coarse metrics like "helpfulness" — contains fewer than ten poetry-specific prompts, only one of which targets Urdu poetry generation. Similarly, while the FLASK benchmark (Ye et al. 2024) includes poetry in its "Language" section (approximately 10% of tasks), a search for "poem" returns fewer than ten prompts. These limitations suggest significant opportunities for developing more comprehensive benchmarks for multilingual generative poetry evaluation.

Human evaluation methodologies vary across studies. Some employ side-by-side comparisons, where evaluators choose the better poem from a pair (similar to A/B testing, as described by Ghazvininejad, X. Shi, et al. 2016). Others use scoring systems, where evaluators rate each poem on a scale (for example, 3, 5, 7, or 10 points) based on predefined criteria. Examples of human evaluation studies that used discrete scales for individual poem generations include Nguyen et al. (2021) (overall quality scored from 0 to 5), Van de Cruys (2020) (five-point scales for fluency, coherence, meaningfulness, and poeticness), C.-C. Wu et al. (2019) (five-point assessment of image-to-poem generations), and B. Liu et al. (2018) (annotators give a 0–10 scale score to a poem given an image, considering their relevance).

A specialized task in human evaluation is determining the authorship of poems. Here, evaluators assess whether a poem was written by a human or a generative system. This task assumes a positive correlation between the "naturalness" of a text and its poetic quality, implying that human-written poems are generally superior. Z. Deng et al. (2024) used this approach to evaluate whether LLMs can generate classical Chinese poems indistinguishable from human-written ones. Similarly, J. Wang et al. (2021) assessed the quality of generated limericks through authorship evaluation, and Zugarini, Melacci, et al. (2019) used it to evaluate a system imitating the style of Dante's *Divine Comedy*.

Several important poetic properties were excluded from our summary table due to limited research coverage. Metaphor detection and other figures of speech (R. Greene et al. 2012, page 273) represent valuable evaluation dimensions for poetic language, yet we found only one relevant study in generative poetry (Chakrabarty, X. Zhang, et al. 2021). For comprehensive coverage of figurative language processing, we refer readers to Lai and Nissim (2024).

Similarly, humor detection was omitted despite its relevance to certain poetic genres. The only work where the human evaluation protocol mentions binary classification based on the criterion of humor presence is Gatti et al. (2017). Research on humor detection outside poetry is surveyed by Ren et al. (2024).

[38] See Section 3 on data engineering.

Poetry translation[39] evaluation presents particular challenges when considering literary translation standards (Matusov 2019). We identified only one work (Chakrabarty, Saakyan, et al. 2021) that employs both automatic metrics (BLEU (Papineni et al. 2002), BERTScore (T. Zhang, Kishore, et al. 2020), COMET (Rei et al. 2020)) and human evaluation of meaning preservation and poetic style.

In Section 5.1, we present a holistic analysis of current approaches to automated assessment of generative poetry, drawing on papers published between 2017 and 2025.

5.1 Automatic Evaluation

An analysis of generative poetry papers from 2017 to 2025 shows that researchers use about a dozen metrics to evaluate their systems and approaches, less than half of which are the most frequently used. Table 3 summarizes the relevant statistics for groups of metrics that are close or overlapping in terms of the measured value.

In the next subsection, we will analyze the algorithmic approaches to calculating the most popular automatic metrics and provide our conclusions on their application.

5.1.1 Automatic Metric Computation Methods. One of the most popular methods for various evaluations of generative poetry is the use of n-gram frequency characteristics obtained for generated texts. Notable implementations for poetry evaluation include style imitation assessment (Sawicki, Grzes, A. Jordanous, et al. 2022; Tikhonov and Yamshchikov 2018b), assessing the similarity of generated text to a ground-truth poem in the image-to-poem system (B. Liu et al. 2018), measuring similarity of generated poems to the training data (Ram et al. 2021), novelty computation (Agirrezabal, Gonçalo Oliveira, et al. 2023), measuring the average fluency of generated poems using the 5-gram character based LM trained on poetry corpus (L. Shen et al. 2020), topic relevance (Tian, Narayan-Chen, et al. 2023), measuring diversity by count distinct unigrams and bigrams (Tian, Narayan-Chen, et al. 2023).

A large number of papers directly use n-gram metrics such as BLEU (Papineni et al. 2002) and ROUGE (C.-Y. Lin 2004), originally developed for machine translation. Calculating these metrics requires a reference text, typically human-written poems. This method is used by many researchers to assess how well a generative system addresses the requirements specified in the input prompt or reproduces the style and vocabulary of the target genre.

For open-ended tasks like poetry generation, where no single reference text exists, the methodological validity of these metrics is often questionable. Several authors acknowledge their limited usefulness for evaluating generative poetry (Z. Zhang et al. 2024). S. Wang, Wong, et al. (2024) demonstrate this limitation by comparing automatic metrics (SacreBLEU, BERTScore, COMET) with human evaluations for poetry translation, revealing poor correlation between the two assessment methods.

An analysis of the studies reveals that the comparison of n-gram frequency distributions of generated texts and the target poem corpus, depending on the interpretation, is presented by the authors as one of the following metrics: (1) novelty (the model's ability to generate word combinations not present in the training data), (2) plagiarism (the model's tendency to literally reproduce the training data), or (3) reproduction of the style of the target genre or author. It can be concluded that authors, when applying such assessments, do not always realize that n-gram frequency estimates can be indicators of all of the above characteristics simultaneously.

Another limitation of the n-gram approach is that it is based on surface text properties and can only assess the superficial similarity of two texts. For texts in Chinese and, to some extent, English, inflection has virtually no effect on the results, ignoring compound words. However, for languages with complex morphology, such as Russian, Finnish, and Turkish, inflection must be taken into account. For agglutinative languages (Turkish), a stemmer is sufficient, but for fusional languages (Russian), a stemmer can be unreliable, so a lemmatizer appears to be a more reliable way to eliminate the influence of inflection. However, we have not encountered any analysis

[39] See Section 2.4 on poetic translation task.

Table 3. Automatic evaluation metrics for generated poetry (2017–2025). The "Evaluated Property" column indicates the poetic quality being measured.

Metrics group	Evaluated Property	Papers
Meter, Rhyme, Form (Section 5.1.2)	Are there any rhymes? Is the meter correct? Are the line and syllable counts correct?	Agirrezabal, Gonçalo Oliveira, et al. (2023), Alyafeai et al. (2023), Amina et al. (2025), Belouadi and Eger (2023), Chakrabarty, Padmakumar, et al. (2022), Chudoba and R. Rosa (2024), De Araujo Possi et al. (2023), Ferraz de Arruda et al. (2022), Gokirmak (2021), Hämäläinen and Alnajjar (2019b), Horishny (2022), J. Hu and M. Sun (2020), Z. Hu et al. (2024), Huynh and Bao (2024), P. Li et al. (2020), Nguyen et al. (2021), Nikolov et al. (2020), Ormazabal et al. (2022), Ou et al. (2023), Ram et al. (2021), Tian and Peng (2022), Z. Xie et al. (2019), Xue, K. Song, et al. (2021), C. Yu et al. (2024), and J. Zhao and H. J. Lee (2022)
Grammaticality, Meaningfulness, Fluency (Section 5.1.3)	Does the text comply with grammar rules? Is the text meaningful?	Amina et al. (2025), Baral et al. (2021), Che et al. (2017), A. Chen et al. (2024), Z. Guo, X. Yi, et al. (2019), Horishny (2022), Z. Hu et al. (2024), P. Li et al. (2020), Z. Liu et al. (2019), Lo et al. (2022), Saeed et al. (2019), L. Shen et al. (2020), Takeishi et al. (2022), Tian, Narayan-Chen, et al. (2023), Xue, K. Song, et al. (2021), Yan (2016), K. Yang and Klein (2021), X. Yi, R. Li, C. Yang, et al. (2020), C. Yu et al. (2024), X. Zhang et al. (2023), J. Zhao and H. J. Lee (2022), W. Zhu and Bhat (2020), and Zugarini, Melacci, et al. (2019)
Novelty, Plagiarism, Originality	Difference from training set; Training data reproducibility	Agirrezabal, Gonçalo Oliveira, et al. (2023), Belouadi and Eger (2023), D'Souza and Mimno (2023), Lu, Sclar, et al. (2025), Nguyen et al. (2021), Ormazabal et al. (2022), Ram et al. (2021), Sawicki, Grzes, Goes, Brown, Peeperkorn, and Khatun (2023), L. Shen et al. (2020), Takeishi et al. (2022), Tian and Peng (2022), R. Zhang and Eger (2024), and Z. Zhang et al. (2024)
Similarity to ground-truth texts	How close are the vocabulary and style of the generations to the ground-truth poems?	Abboushi and Azzeh (2023), Cespedosa Vázquez and Mitkov (2023), Chakrabarty, Saakyan, et al. (2021), Chiang et al. (2021), Fan et al. (2019), Gokirmak (2021), Z. Hu et al. (2024), Khanmohammadi et al. (2023), B. Liu et al. (2018), Z. Liu et al. (2019), Lu, J. Wang, et al. (2019), Moreno (2021), Nikolov et al. (2020), Panahandeh et al. (2023), Ram et al. (2021), Resende and Hadley (2024), Sawicki, Grzes, Goes, Brown, Peeperkorn, Khatun, and Paraskevopoulou (2023), L. Shen et al. (2020), W. L. Song et al. (2023), Tikhonov and Yamshchikov (2018a), Tikhonov and Yamshchikov (2018b), Z. Xie et al. (2019), and Z. Yang, P. Cai, et al. (2019)
Diversity (Section 5.1.4)	How big is the active vocabulary?	Boggia et al. (2022), Y. Chen, Gröner, et al. (2024), Hamat (2024), Z. Liu et al. (2019), Lo et al. (2022), Ram et al. (2021), L. Shen et al. (2020), Takeishi et al. (2022), Tian, Narayan-Chen, et al. (2023), K. Yang and Klein (2021), X. Yi, R. Li, C. Yang, et al. (2020), R. Zhang and Eger (2024), and X. Zhang et al. (2023)
Naturalness (Section 5.1.6)	Signs that text was generated	Z. Zhang et al. (2024)
Relevance	Input conditions matching	Agirrezabal, Gonçalo Oliveira, et al. (2023), Boggia et al. (2022), Tian, Narayan-Chen, et al. (2023), and C. Yu et al. (2024)
Coherency	Is the poem text coherent?	Belouadi and Eger (2023), Boggia et al. (2022), and Z. Zhang et al. (2024)
Aesthetics, Imagery, Elegance, Poetic value	Evokes emotion; Aesthetically pleasing; Has poetic value	A. Chen et al. (2024), Hämäläinen and Alnajjar (2019b), C. Yu et al. (2024), and Z. Zhang et al. (2024)
Conformity to the author's style	How well does the style of a given author appear?	Sawicki, Grzes, Goes, Brown, Peeperkorn, and Khatun (2023)

of these issues in studies of generative poetry in languages other than the aforementioned English and Chinese, which also raises several questions about the methodological purity of the demonstrated results.

The advantage of n-gram metrics, in addition to their ease of implementation, is the good interpretability of their results. Other evaluation approaches, such as LLM perplexity or cosine distance for embeddings, typically lack a simple way to understand why a metric deviates from expectations in a given case.

In general, the use of n-gram metrics in poetry evaluation should be carefully justified and, if possible, verified against expert annotation (see more on BLEU verification on a variety of tasks at Reiter 2018). Tools such as BERT for text similarity estimation, LLM-based perplexity calculations, and increasingly popular LLM-as-a-judge approaches, which have become standard in many NLP fields, can effectively replace n-gram metrics.

Another popular tool for evaluating generated poems is the LLM perplexity calculation on these texts. The availability of open-source, high-quality multilingual models and supporting libraries allows for calculations with minimal researcher effort. The resulting perplexity is typically considered an indicator of grammaticality or fluency. Examples of such applications can be found in many works (Che et al. 2017; Z. Hu et al. 2024; Yan 2016; X. Zhang et al. 2023; Zugarini, Melacci, et al. 2019). It should be noted that, as a universal method for text evaluation, LLM model perplexity, like n-gram metrics mentioned above, requires methodological care when interpreting the results. We will discuss all of this in detail in Section 5.1.3.

A complementary evaluation approach employs embedder LMs such as BERT and RoBERTa and measures cosine similarity between text embeddings. This approach can be used to assess the relevance of the generated text to the input prompt or a given topic (Agirrezabal, Gonçalo Oliveira, et al. 2023), the coherence of text fragments (W. Zhu and Bhat 2020), and grammaticality (K. Yang and Klein 2021). Furthermore, the ability of BERT-like models to produce context-sensitive word embeddings can be used to detect metaphors (Y. Li et al. 2023).

The constantly improving ability of advanced LMs to analyze literary works (Z. Yang, Z. Liu, et al. 2024), generated poetry (A. Chen et al. 2024; Sawicki, Grzes, Brown, et al. 2025; C. Yu et al. 2024; Z. Zhang et al. 2024; C. Zhao et al. 2024), poetic translations (S. Wang, Wong, et al. 2024) suggests promising progress in automating the evaluation of generative poetry by use of LLM-as-a-Judge and similar approaches. For example, Z. Zhang et al. (2024) explored ChatGPT-based evaluation for generated lyrics, while Z. Li et al. (2024) provided a systematic overview of using LLMs for evaluating generative models. Researchers planning to use **LLMs** to replace human evaluation protocols for generative poetry should be aware of the tool's inherent limitations. For instance, in side-by-side evaluation scenarios, the order in which generated texts are presented within a prompt can systematically bias the model's selection, as demonstrated by L. Shi et al. (2025).

However, many properties of poetic texts cannot be assessed using universal tools such as perplexity calculations, as these properties require a specialized approach. An example is the assessment of poetic attributes such as rhyme and meter. In Section 5.1.2, we will discuss approaches for assessing them.

5.1.2 Meter, Rhyme, Form. Evaluating the quality of generated poetry is a challenging task. Many researchers emphasize that only humans can comprehensively assess poetic quality, making automation difficult, time-consuming, and costly.

Typically, a poetic text must adhere to specific prosodia-related requirements, such as rules for alternating stressed and unstressed syllables (in accentual-syllabic versification) or tone patterns (in Chinese poetry). Additionally, the text often needs to conform to structural constraints, such as a fixed number of lines and syllables per line. A well-known example is the haiku, a three-line poem following a strict 5-7-5 syllable structure, traditionally incorporating a seasonal reference (Higginson and Harter 1985, p. 104).

Some genres and forms also require specific rhyme schemes between lines. A well-known example is the varieties of sonnet (Fuller 2017, p. 2).[40]

[40]These properties of poetic texts can be automatically assessed using tools described in Section 3.

Automatic evaluation of rhyme- and rhythm-related properties of poetry is employed in various projects. For example, De Araujo Possi et al. (2023) propose metrics for assessing rhyme and meter automatically. J. Zhao and H. J. Lee (2022) introduce a "tone-checker" module with rules tailored to classical Chinese poetry. Chudoba and R. Rosa (2024) use a set of automatic metrics to evaluate the form of generated poems, including compliance with meter, rhyme, and syllable count. Similarly, Z. Hu et al. (2024) incorporate a metrical controller to constrain their diffusion-based poetry model. Nguyen et al. (2021) apply automatic quantitative evaluation to assess rhyme and tone rule conformance in Vietnamese poetry generation. Agirrezabal, Gonçalo Oliveira, et al. (2023) describe a suite of automated tests that evaluate poeticity (for example, rhyme richness), novelty (using ROUGE as a similarity metric), and topicality (using BERT-based embeddings). Ferraz de Arruda et al. (2022) propose an approach that uses binary classification models — such as random forest, k-nearest neighbors, support vector machines, and multi-layer perceptron — along with phonetic features to distinguish between English-language prose and poetry. The texts were sourced from Project Gutenberg. Their results show that the classifiers achieved an accuracy of at least 75% in identifying the type of text.

While these metrics offer algorithmic and interpretable benefits, their use depends on the availability of tools capable of analyzing poetic meter and rhyme and detecting defects. Such tools are not universally available for all languages and may require significant development effort. An alternative approach, albeit potentially less precise, could involve leveraging audio modality models capable of detecting rhythm and consonance independently of language. Such an approach could provide a more generalizable solution for assessing poetic form.

Rhyme and meter are important properties of a poetic text. On the other hand, a poem remains a text that must be written correctly, coherently, and meaningfully. The assessment of these properties is described in Section 5.1.3.

5.1.3 Grammaticality and Meaningfulness. H. M. Manurung (2003) identifies three key properties of well-formed poetry: poeticness, grammaticality, and meaningfulness. These criteria have been directly assessed by human judges in several studies (Agarwal and Kann 2020; Røstvold and Gambäck 2020). Similarly, Ram et al. (2021) employed human evaluators to assess the grammaticality and other properties of generated song lyrics.

The grammaticality and meaningfulness of generated texts can also be approximated computationally. One common approach is to calculate the probability of the text, as demonstrated by X. Yi, M. Sun, et al. (2018). This is often operationalized using the perplexity of LMs (Che et al. 2017; Z. Hu et al. 2024; Yan 2016; X. Zhang et al. 2023; Zugarini, Melacci, et al. 2019). However, perplexity can be an unreliable metric, as noted by Kuribayashi et al. (2021) and Y. Wang et al. (2023). Additionally, low perplexity may indicate overly simplistic or repetitive texts, which can detract from the perceived quality of poetry.

J. Zhao and H. J. Lee (2022) propose an alternative approach by training a separate model, termed a "text-fluency-checker." They use a synthetic dataset for training, where positive samples consist of human-written poems, while negative samples are generated by applying distortion rules to these texts.

W. Zhu and Bhat (2020) introduce a BERT-based model to evaluate the linguistic quality of generated poems across multiple criteria: (1) grammaticality, (2) non-redundancy, (3) focus (semantic relatedness between adjacent sentences), and (4) structure and coherence. Similarly, Boggia et al. (2022) measure line coherence in generated and human-written texts using cosine similarity between Word2Vec vectors.

The assessment of grammaticality and linguistic acceptability in poetry presents an interesting research direction, particularly in exploring how algorithms, models, and datasets developed for prose perform in the poetry domain.

A generative poetry system that produces literate, coherent, rhyming, and metered texts must possess one more property to be engaging to the user: the ability to create lexically diverse texts. We will examine the metrics used to evaluate this property in Section 5.1.4.

5.1.4 Diversity Assessment. Linguistic diversity in poetry refers to the range and variability of linguistic features present in a text. It may involve lexical, syntactic, stylistic, and semantic variation, and is often interpreted as an

indicator of expressive richness or non-redundancy. In the context of generative poetry, diversity is typically assessed through lexicographic analysis at the word or *n*-gram level, which serves as a practical proxy for estimating the size and variability of a model's active vocabulary.

A variety of metrics have been proposed in computational linguistics to quantify lexical diversity, including:

- MTLD (Measure of Textual Lexical Diversity) (McCarthy 2005), which estimates lexical diversity based on the stability of type–token ratios across a text;
- TTR (Type–Token Ratio) (Johnson 1944), a simple measure that is strongly influenced by text length;
- MATTR (Moving Average TTR) (Covington and McFall 2010), which mitigates length effects by averaging TTR over sliding windows;
- HD-D (Hypergeometric Distribution D) (McCarthy and Jarvis 2010, p. 383), a probabilistic measure of lexical diversity;
- the hapax legomenon ratio (Baayen and Lieber 1991), which quantifies the proportion of words that occur only once in a text or corpus.

The application of these metrics to both human-written and generated poetry is discussed in, for example, Hamat (2024) and Y. Chen, Gröner, et al. (2024).

In addition to these general measures, diversity estimation methods used in generative poetry papers published between 2017 and 2025 can be grouped into two broad categories.

(1) **Lexical diversity ratios.** Diversity is computed as the ratio of unique *n*-grams (typically unigrams or bigrams) to the total number of *n*-grams. For unigrams, this metric is commonly referred to as Diversity-1, while for bigrams it is known as Diversity-2. These measures are widely used in poetry generation studies, including (Z. Liu et al. 2019; Lo et al. 2022; Ram et al. 2021; L. Shen et al. 2020; K. Yang and Klein 2021).
(2) **Inter-text similarity measures.** Diversity is assessed by computing the average or maximum pairwise similarity among generated texts. Similarity is estimated using measures such as Jaccard similarity over unigrams or bigrams (Boggia et al. 2022; X. Yi, R. Li, C. Yang, et al. 2020), the Dice–Sørensen coefficient (Takeishi et al. 2022), BLEU (X. Zhang et al. 2023), or embedding-based measures such as BERTScore (Y. Chen, Gröner, et al. 2024).

As suggested by the definition at the beginning of this section, linguistic diversity extends beyond lexical variation. Developing evaluation methods that account for additional dimensions, such as syntactic or stylistic diversity, remains an open and promising direction for future research in generative poetry.

5.1.5 Creativity Assessment. There is a view, supported by experimental evidence, that creativity is very closely related to human intelligence (Frith et al. 2021). This connection makes computational creativity interesting from the point of view of building systems with artificial general intelligence (Colton and Wiggins 2012).

Automatic poetry generation is an example of an NLG task where we expect the system being created to not only follow the formal rules and constraints inherent to the poetry genre but also to exhibit a certain creativity, even if the system developers do not explicitly set such a task, limiting themselves to assessments of novelty or the absence of plagiarism in generations. We will leave aside the controversial question of whether a generative LM can be creative without being capable of human perception and reflection, although we consider it necessary in this regard to mention the opinion of Kreminski (2024) that "machines are often usefully creative because they fail to see things completely as humans do: their oversights and inabilities lead them to mix human-like with non-human-like creative decisions in unanticipated ways, and thereby to supply human creators with ideas that they otherwise never would have considered." In this section, we discuss what creativity is and what practical methods and approaches for assessing the level of creativity exist in the domain of generative poetry and in NLG in general.

The main difficulty in assessing the creativity of poetry generation models is the lack of a single, generally accepted definition of creativity, on the basis of which one could select automatic metrics or develop protocols for objective human assessment. Different authors decompose this concept differently and offer different approaches to assessing the identified key elements, which makes it difficult to compare the results. For example, Nguyen et al. (2021) define creativity as a value complementary to the level of plagiarism, that is, the share of texts that the model reproduced from the training data. A similar approach to automatic assessment of creativity is being developed by Lu, Sclar, et al. (2025): they introduce a "creativity index," that is the share of the words that are covered by verbatim and near-verbatim matches of n-grams from a given poem against the web. According to Elzohbi and R. Zhao (2023), the main components of creativity are (1) originality, (2) unpredictability, and (3) sociability. The first two components, if we follow their author's definition, are directly related to the novelty of the work. The last component, called sociability, should reflect the interest that the given work will arouse in readers and critics and will encourage them to evaluate and analyze the text. In our view, this third component is directly related to the value mentioned in the following paper. A good discussion of alternative definitions of creativity and its components is given in Ismayilzada, Paul, et al. (2024).

When developing methods to assess the creativity of generative poetry systems, it is useful to draw on established creativity assessment frameworks. Below, we briefly summarize several influential approaches that are commonly referenced in the computational creativity literature.

Rhodes (1961) conceptualizes creativity through four complementary aspects, known as the Four Ps: *Person*, which concerns individual traits and dispositions; *Process*, which covers cognitive and motivational mechanisms involved in creation; *Press*, which refers to environmental and contextual factors; and *Product*, which denotes the creative artifact itself. This framework has been widely adopted as a theoretical foundation in computational creativity research. Relevant examples include the following works:

(1) Kantosalo (2019), which applies creativity theory to the development and evaluation of co-creative poetry systems;
(2) A. Jordanous (2016), which analyzes the applicability of the Four Ps model to computational creativity assessment, with emphasis on novelty and value;
(3) Carnovalini and Rodà (2020), which discusses creativity evaluation in the context of generative music systems;
(4) Lamb, Brown, and C. L. A. Clarke (2018), which provides a tutorial-style overview of methods for evaluating creative computational systems.

Beyond the Four Ps framework, several other approaches to creativity assessment are relevant to generative poetry research. One example is SPECS, a structured methodology for evaluating computational creativity proposed by A. K. Jordanous (2012). Another is the Creative Tripod framework, consisting of *Skill*, *Appreciation*, and *Imagination*, introduced by Colton (2008) to assess systems that generate creative artifacts such as poems, paintings, melodies, and mathematical results. Karimi et al. (2018) present a framework for evaluating computational creativity based on four questions: (1) who evaluates creativity, (2) what is being evaluated, (3) when the evaluation takes place, and (4) how the evaluation is performed. The framework is illustrated with several practical examples.

The following examples illustrate how creativity is assessed in generative poetry systems in practice, where automatic metrics often rely on simplified definitions of creativity compared to the frameworks discussed above.

Franceschelli and Musolesi (2022) describe an approach to assessing two components of creativity — novelty and value, using neural network methods. Value is assessed using a discriminator from a GAN, which learns to distinguish real (and valuable) poems from generated ones. Novelty is proposed to be assessed using a generative adversarial network. Bena and Kalita (2019) investigate the ability of a retrained GPT-2 LM to generate creative English-language poems, defining creativity as a combination of (1) the ability to evoke certain emotions and

feelings in the reader, (2) the ability to imitate dream language in poetry (see more on dream poetry in Spearing 1976).

The novelty of a literary text includes the absence of plagiarism, that is, the inclusion of fragments of other people's texts without indicating the source. This must be taken into account in the case of generative poetry systems that use LLMs, since transformer-based generative models are known for their ability to memorize and then reproduce verbatim poem texts from the training set, as shown by D'Souza and Mimno (2023). Reproducing fragments of human-written verse during generation is generally undesirable, except in parody, where limited quotation can be used to support recognition of the source text. To reduce the effect of memorization at the training stage, various approaches have been proposed. For example, Hans et al. (2024) suggest masking losses for some of the tokens in the batch. To assess the model's tendency to reproduce texts from the training set, some works on poetry generation utilize plagiarism rate evaluation. Nguyen et al. (2021) count how often lines from the training dataset occur in generated poems.

While the above-mentioned works use n-gram approaches to calculate novelty or plagiarism level to assess LLM creativity, Olson et al. (2024) proposes a fundamentally different way to simultaneously calculate creativity and improve LLM creativity. The key idea is to calculate the "direction of creativity" as an average vector obtained through the difference in activations of the selected internal LLM layer for pairs of texts that differ only in the level of creativity specified in the prompts. Then, the resulting creativity vector can be used to (1) evaluate the creativity of new generations by calculating the cosine proximity of activations on the selected layer and the "direction of creativity," (2) improve the creativity of generations by mixing the creativity vector with the corresponding activations.

There are also a number of works devoted to the analysis of poetry written by people from the point of view of various linguistic and statistical properties. Although these works do not mention generative poetry as an object of study or comparison, we admit that the ideas from these works can be useful for poetry generation tasks. For example, Kao and Jurafsky (2012) examine the statistical and linguistic correlates of good poetry.

Publicly available LLM-based chatbots have enabled comparative studies of model creativity across a range of generation tasks (L. Sun, Yuan, et al. 2025). A broader overview of LLM creativity is provided by Franceschelli and Musolesi (2024c). This survey focuses instead on practical approaches to assessing creativity in generative poetry.

The assessment of the creativity of poems is one of the most poorly formalized and controversial procedures, starting from the lack of a generally accepted definition of creativity as such and ending with the lack of generally accepted standard protocols for its assessment. Probably, an important and interesting direction of research may be the automation of creativity assessment, including the assessment and improvement of the alignment of automatic assessment with human assessments.

5.1.6 Naturalness Assessment. A key objective for automatic poetry generation systems is to produce text that is indistinguishable from human-written poetry, focusing on the naturalness or human-like quality of the output. Detecting AI-generated poetry can be achieved through human evaluation or algorithmic approaches, including artificial text detection techniques. However, assessing naturalness presents several challenges:

- **Plagiarism and memorization**: Models may reproduce training texts, leading to outputs that are effectively human-written, which can bias results in favor of the model. Thus, it is essential to evaluate the level of plagiarism or memorization.
- **Evaluator bias**: Evaluators' familiarity with human-written works may lead them to recognize and favor human-authored texts, skewing results in favor of humans.

S. Ma and Q. Wang (2024) propose **token cohesiveness** as a criterion for distinguishing LLM-authored texts from human-written ones. Token cohesiveness is measured by randomly deleting tokens and assessing the semantic difference. The authors demonstrate that LLM-generated texts exhibit higher token cohesiveness than human-written texts.

Hayawi et al. (2024) introduced a dataset comprising human-authored and LLM-generated texts across various genres, such as essays, stories, and poetry. They utilized several machine learning models to classify the texts within the dataset. The study found that while binary classification tasks (distinguishing human-authored from LLM-generated texts) achieved strong performance, more complex multiclass tasks — such as differentiating between human-generated texts and outputs from multiple LLMs — yielded significantly poorer results.

Kushnareva et al. (2021) propose a method based on topological data analysis (TDA) for analyzing attention maps in a BERT model. Differences in the topological properties of attention maps between human and machine texts enable authorship identification with reasonable accuracy. P. Wang, L. Li, et al. (2023) address the artificial text detection (ATD) task at the sentence level, focusing on documents that mix human- and LLM-authored sentences.

E. Clark et al. (2021) analyze the challenges faced by untrained evaluators in determining the authorship of texts across domains such as news, recipes, and stories. They note that evaluators often rely on text fluency, which modern LMs can replicate effectively, making detection difficult.

The MAUVE metric, introduced by Pillutla et al. (2021), quantifies the divergence between the distributions of human-written and machine-generated texts. Unlike ATD, MAUVE requires a large number of text samples (approximately 5,000) to achieve statistically reliable results. The code for MAUVE calculations is available in the repository.[41] When using MAUVE, it should be noted that some studies Garces Arias et al. (2024) show poor correlation of this metric with human assessments.

Future research in this area should explore how methods for detecting machine-generated content perform under domain shift conditions in poetry, considering the unique features of this domain.

5.2 Human Evaluation

The automated evaluation metrics discussed previously offer advantages in speed, scalability, and reproducibility. However, they are often insufficient for a comprehensive assessment. Certain aspects of poetry perception are inherently subjective and difficult to formalize. Additionally, the correlation between automated metrics and human judgment is often poorly established. For these reasons, human evaluation remains a critical component in most generative poetry research, frequently used alongside automated metrics within the same study.

This section examines popular human evaluation approaches, protocol design considerations, and common challenges. Our analysis covers papers published between 2017 and 2025. Table 4 provides relevant statistics, with metrics grouped by similarity or overlapping properties. We have tried to form groups of metrics that are as consistent with those presented in Table 3 as possible, given the wide variety of approaches.

Human evaluation differs from automated metrics in two key aspects. First, assessment instructions are critical for human judges. These instructions must be detailed, cover edge cases thoroughly, and ideally include checks for assessor understanding. This requirement parallels the careful prompt engineering needed for LLM-as-a-judge approaches. Second, human evaluation outcomes should include not only average scores but also an analysis of inter-annotator agreement. Variability between assessors can significantly affect the reliability of the results. Unfortunately, many studies omit this information, complicating cross-study comparisons.

In addition to the differences mentioned above, human metrics often produce values not on a continuous interval, but on an ordinal-scale metric, as in the case of a Likert scale. Table 5 provides summary statistics for different scale types in human assessments. We have allocated evaluation through side-by-side comparisons and choosing the best of two or more options to a separate group.

One challenge in human evaluation of generative poetry is the potential for systematic differences across assessor populations. In particular, crowd-sourced evaluations have been shown to produce more favorable judgments of generated poetry than evaluations conducted by domain experts, as reported by Lau et al. (2018).

[41] https://github.com/krishnap25/mauve

Table 4. Human evaluation metrics for generated poetry (2017–2025). The "Evaluated Property" column specifies the poetic quality assessed by human judges.

Metrics group	Evaluated Property	Papers
Rhyme, Meter, Singability	The text is checked for rhymes and meter.	Belouadi and Eger (2023), H. Chen et al. (2019), Gonçalo Oliveira (2020), Lau et al. (2018), Y. Liu, D. Liu, et al. (2020), Nikolov et al. (2020), Ram et al. (2021), Xue, K. Song, et al. (2021), and Z. Zhang et al. (2024)
Grammaticality, Meaningfulness, Fluency	Compliance with grammar rules is assessed. The text is checked to ensure it is meaningful and fluent.	Abboushi and Azzeh (2023), Agarwal and Kann (2020), Agnew et al. (2023), Amina et al. (2025), Chakrabarty, Padmakumar, et al. (2022), Chang et al. (2023), H. Chen et al. (2019), Elzohbi and R. Zhao (2025a), Gonçalo Oliveira (2020), Hämäläinen, Alnajjar, and Poibeau (2022), Z. Hu et al. (2024), Khanmohammadi et al. (2023), Lau et al. (2018), Y. Liu, D. Liu, et al. (2020), Lu, J. Wang, et al. (2019), Nguyen et al. (2021), Panahandeh et al. (2023), Ram et al. (2021), Røstvold and Gambäck (2020), Sawicki, Grzes, A. Jordanous, et al. (2022), Shao et al. (2021), L. Shen et al. (2020), Z. Sheng et al. (2021), W. L. Song et al. (2023), Takeishi et al. (2022), Tian and Peng (2022), Van de Cruys (2020), Xue, K. Song, et al. (2021), L. Yi (2023), X. Zhang et al. (2023), Z. Zhang et al. (2024), and J. Zhao and H. J. Lee (2022)
Originality, Creativity, Novelty	How original or banal is the text?	Elzohbi and R. Zhao (2025a), Gonçalo Oliveira (2020), Nguyen et al. (2021), Tian and Peng (2022), L. Yi (2023), X. Zhang et al. (2023), Z. Zhang et al. (2024), and J. Zhao and H. J. Lee (2022)
Imagery, Poetic value, Emotions	Does the poem evoke emotion? Does it have poetic value?	Abboushi and Azzeh (2023), Agarwal and Kann (2020), Agnew et al. (2023), Elzohbi and R. Zhao (2025a), Hämäläinen, Alnajjar, and Poibeau (2022), Z. Hu et al. (2024), Khanmohammadi et al. (2023), Lau et al. (2018), Nguyen et al. (2021), Panahandeh et al. (2023), Røstvold and Gambäck (2020), Shao et al. (2021), L. Shen et al. (2020), W. L. Song et al. (2023), Tian and Peng (2022), and Van de Cruys (2020)
Diversity	Do generated sentences often show similar phrases, words, or rhymes?	Lu, J. Wang, et al. (2019), Shao et al. (2021), and Xue, K. Song, et al. (2021)
Naturalness	Turing test. Does the text feel like it was written by a human?	Agnew et al. (2023), Baral et al. (2021), Belouadi and Eger (2023), Lau et al. (2018), D. Lewis et al. (2021), Pascual (2021), Van de Cruys (2020), J. Wang et al. (2021), Z. Wang, Guan, et al. (2023), J. Zhao and H. J. Lee (2022), and Zugarini, Melacci, et al. (2019)
Relevance	Are all conditions specified in the input request taken into account?	Agnew et al. (2023), Chakrabarty, Padmakumar, et al. (2022), Chang et al. (2023), and Z. Zhang et al. (2024)
Translation quality	How fully does the translation reflect the meaning, form, and poetic attributes of the original?	A. Chen et al. (2024) and Khanmohammadi et al. (2023)
Style	How well does the style of a given author appear?	Tikhonov and Yamshchikov (2018a) and J. Zhao and H. J. Lee (2022)
Overall quality	General assessment without details.	Benhardt et al. (2019), Chakrabarty, Padmakumar, et al. (2022), Chang et al. (2023), H. Chen et al. (2019), Gatti et al. (2017), Gonçalo Oliveira (2020), Hämäläinen, Alnajjar, and Poibeau (2022), Köbis and Mossink (2021), Lu, J. Wang, et al. (2019), Ormazabal et al. (2022), Popescu-Belis et al. (2022), Z. Sheng et al. (2021), and W. L. Song et al. (2023)

This issue is examined more systematically by Lamb, Brown, and C. Clarke (2015), who analyze how assessor expertise influences judgments of creative text.

When using human authorship assessment, in which annotators judge whether a text was written by a human or generated by a system, several factors should be considered. Even relatively small LMs such as GPT-2 can produce texts that are difficult to distinguish from human writing, and annotators often overestimate their confidence in these judgments, as shown by Gunser et al. (2022).

Rahmeh (2023) presents a comparative human evaluation of LLM-generated and human-authored poetry and explicitly report demographic characteristics of the assessor group, including gender and age, as potential sources of bias. Their study is based on a survey of 80 graduate students in English language and literature from a Lebanese university and compares Shakespeare's Sonnet 18 with a thematically similar sonnet generated by ChatGPT. Participants rated enjoyment, emotional engagement, and perceived linguistic complexity on a ten-point scale. The results show that the human-authored poem is preferred overall, while a subset of participants reports positive reception of the generated sonnet. A limitation of the study is that the evaluation relies on only two poems whose authorship was known to the assessors.

The absence of prior information about text authorship is an important methodological condition in studies where assessors are asked to judge whether a text was written by a human or generated by a system, a setup often compared to a Turing-style test. Examples of relevant work are provided below.

Porter and Machery (2024) explore people's inability to distinguish LLM-generated poetry from human-written poetry. Their study reveals an intriguing finding: "*the simplicity of AI-generated poems may be easier for non-experts to understand, leading them to prefer AI-generated poetry and misinterpret the complexity of human poems as incoherence generated by AI.*" Similarly, Köbis and Mossink (2021) evaluate people's ability to distinguish GPT-generated poems. They generate continuations of human-written poems using GPT-2, based on the original poem's headline. Two evaluation settings are used:

- **Human-in-the-loop**: The best GPT-generated continuation is selected and compared to the human-written poem. In this setting, evaluators struggle to reliably distinguish between AI-generated and human-written poems.
- **Human-out-of-the-loop**: A random GPT-generated continuation is compared to the human-written poem. Here, evaluators confidently identify the AI-generated text.

Another example of a Turing-style evaluation is presented by B. Liu et al. (2018), who develop a GRU-based system for generating English-language poems conditioned on images. Their human evaluation involves 500 assessors, including 30 poetry experts, and uses 1,500 images from the dataset. In addition to relevance and quality ratings, assessors are asked to judge whether each poem was written by a human or generated by a system. The results show that both non-expert and expert assessors are sometimes unable to reliably distinguish generated poems from human-authored ones, although experts perform better on this task.

For a detailed overview of the relevant features of the human assessment protocols, we recommend the reader to refer to the survey by Hämäläinen and Alnajjar (2021).

While human evaluation remains the gold standard, it is often costly, time-consuming, and inconsistent across studies. This has motivated interest in understanding how well automatic metrics, and especially LLM-based judgment, align with human assessments. We therefore next review what is known about the correlation between automatic metrics and human evaluation in generative poetry, as well as insights from adjacent domains.

5.3 Automatic Metrics vs. Human Judgments

The concurrent use of automatic metrics and human evaluation in generative poetry research raises important questions about their correlation. This is particularly relevant for LLM-as-a-judge approaches, which aim to replicate human assessment.

Table 5. Statistics on scale variants used in human assessment protocols for generative poetry in papers from 2017 to 2025.

Scale	Papers
Likert	Abboushi and Azzeh (2023), Agarwal and Kann (2020), Amina et al. (2025), Chakrabarty, Padmakumar, et al. (2022), Chang et al. (2023), H. Chen et al. (2019), Hämäläinen, Alnajjar, and Poibeau (2022), Z. Hu et al. (2024), Khanmohammadi et al. (2023), Lau et al. (2018), Nikolov et al. (2020), Panahandeh et al. (2023), L. Shen et al. (2020), Z. Sheng et al. (2021), Shihadeh and Ackerman (2020), Takeishi et al. (2022), Tian and Peng (2022), Van de Cruys (2020), Xue, K. Song, et al. (2021), L. Yi (2023), X. Zhang et al. (2023), and J. Zhao and H. J. Lee (2022)
Binary	Baral et al. (2021), Benhardt et al. (2019), Chakrabarty, Padmakumar, et al. (2022), Gatti et al. (2017), D. Lewis et al. (2021), Nikolov et al. (2020), Pascual (2021), J. Wang et al. (2021), and Zugarini, Melacci, et al. (2019)
Side-by-side	Agnew et al. (2023), Belouadi and Eger (2023), Köbis and Mossink (2021), and Ormazabal et al. (2022)
0 to 5	Lu, J. Wang, et al. (2019), Nguyen et al. (2021), Q. Wang et al. (2016), and Z. Zhang et al. (2024)
0 to 1	A. Chen et al. (2024)
1 to 3	Nalci et al. (2025) and Røstvold and Gambäck (2020)
1 to 10	Nalci et al. (2025)
0 to 100	Elzohbi and R. Zhao (2025a)

At the time of writing, only Sawicki, Grzes, Brown, et al. (2025) provide a systematic comparison of LLM-based judges for poetry evaluation. Their study uses a corpus of 90 poems with gold standard annotations and follows principles inspired by the Consensual Assessment Technique (Amabile 1983), which is the generally accepted approach for assessments in psychology. Claude-3-Opus and GPT-4o are used as surrogate judges, with rater diversity simulated through temperature variation. The results show strong correlations with the gold standard and higher accuracy than nonexpert human judges, while maintaining high agreement across repeated evaluations. These findings suggest that LLM-based evaluation of poetry is increasingly feasible.

W. Wang et al. (2025) study the relationship between automated and human evaluation for literary fiction and report good agreement between human ratings and their LLM-based framework. However, their focus on long-form narratives limits direct applicability to poetry.

Similarly, Kim, Shin, et al. (2024) demonstrate that an LLM-based judge performs well across a range of evaluation tasks and correlates with human judgments. Their evaluation does not include poetry generation, leaving its relevance to this domain open.

Complementary evidence comes from Walsh, Antoniak, et al. (2024), who evaluate how well different LLMs can detect specific poetic properties, including form, rhyme, meter, line repetitions, thematic adherence, and length constraints. Because their dataset was annotated by human experts, the study effectively measures how well LLM-as-judge can replicate human labeling across fine-grained poetic features, though not holistic judgments of quality.

Our analysis of the reviewed literature reveals several noteworthy patterns, albeit less systematic than the fiction study.

Importance of prompt engineering: Multiple studies emphasize that careful prompt construction is crucial for reliable LLM assessment. C. Yu et al. (2024) and Z. Zhang et al. (2024) specifically highlight the need to specify all relevant evaluation details in the prompt.

Varying correlation levels: The agreement between automated and human evaluation appears inconsistent across different poetic qualities. Belouadi and Eger (2023) found that automatic rhyme evaluation significantly diverged from human judgment, altering model rankings. By contrast, naturalness assessment showed better alignment. Similarly, Z. Zhang et al. (2024) report good correlation between human evaluations and LLM-as-a-judge scores, suggesting this approach could reduce reliance on costly human assessment. A. Chen et al. (2024) also demonstrate strong agreement between GPT-4 judgments and human ratings for translation adequacy in poetry, though without providing explicit correlation coefficients.

Taken together, these findings suggest that while LLM-based evaluation holds promise as a scalable proxy for human judgment, its reliability remains uneven across poetic qualities and task types. More systematic studies are needed to establish when such metrics can replace or complement human evaluation. This gap is not only critical for generative poetry but also reflects a broader challenge in generative AI: ensuring that automatic evaluation faithfully captures human-centered criteria such as creativity, coherence, and poetic value.

6 Conclusion and Future Insights

This survey reviewed recent research on generative poetry published between 2017 and 2025, with particular attention to model architectures, task formulations, data resources, and evaluation practices. Several consistent trends emerge from the surveyed literature.

Observed trends. First, there is a clear and sustained shift toward transformer-based LMs as the dominant architecture for poetry generation (see Figure 1). Most recent systems rely on subword tokenization schemes such as BPE, with only limited exploration of syllable-, character-, or byte-level representations. Second, research activity around alternative paradigms — including RL, evolutionary methods, purely template-based systems, and non-transformer architectures — has declined considerably. Third, LLMs are increasingly used as evaluators, either to complement or partially replace traditional automatic metrics and human assessment. Finally, several studies report that, under specific conditions, LLM-generated poems can approach human-written poetry in perceived quality, although other works report mixed or contrasting results. Taken together, these findings suggest substantial improvements in fluency and stylistic control, while also highlighting the variability of evaluation outcomes.

Current limitations. Despite this progress, the field faces several methodological and practical limitations. A major challenge is the absence of standardized evaluation protocols, shared benchmarks, and public leaderboards for poetry generation. This stands in contrast to other areas of NLP, where large-scale benchmarks and evaluation toolkits are widely available (Gao et al. 2025; Hendrycks et al. 2021; Y.-T. Lin and Y.-N. Chen 2023). Although LLM-as-a-judge approaches show promise, robust evaluation methods that account for genre, style, language, and versification systems remain largely undeveloped.

A second limitation is the lack of systematic comparative studies of decoding strategies[42] for poetry generation. While recent work has proposed a wide range of decoding algorithms aimed at improving diversity, constraint satisfaction, and fluency (Das et al. 2025; Garces Arias et al. 2024; Gareev et al. 2024; J. Guo et al. 2025; Luo et al. 2025; Pynadath and R. Zhang 2025; Tu et al. 2024; W. Zhu, H. Hao, et al. 2024), their impact on poetic quality has not yet been systematically evaluated.

A third gap concerns linguistic creativity. Poetic phenomena such as nonce word formation, compounding, and unconventional morphology remain underexplored, particularly in languages with productive word-formation systems (Benigni and Masini 2009; Crystal 2008). Although related work exists (Coil and Shwartz 2023; Ismayilzada, Circi, et al. 2025; Lencione et al. 2022; Lynott and Keane 2005), generative poetry offers a distinct setting in which to study these forms of creativity.

[42] Decoding methods are discussed in Section 4.3.

Finally, the limited availability of high-quality, publicly accessible poetry and song lyric datasets — especially for non-English and Chinese languages and less common poetic forms — continues to constrain progress. This limitation is particularly relevant for transformer-based models, whose performance depends heavily on data scale and quality.

Future research directions. Looking ahead, several research directions appear especially promising. Chain-of-thought methods, which have proven effective in structured reasoning tasks (Xia et al. 2025), may offer a way to model planning, drafting, and revision processes observed in human poetic composition (Addonizio and Laux 1997; Flower and Hayes 1981; Oliver 1994). While early attempts to apply CoT to poetry evaluation have yielded mixed results (Tomizawa et al. 2025), further exploration remains warranted.

Retrieval-augmented generation (RAG) also presents clear opportunities for poetry generation. Potential applications include integrating factual knowledge, stylistic exemplars, metaphor databases, rhyme dictionaries, and genre-specific constraints (A. Chen et al. 2024; Y. Liu, L. Lan, et al. 2025). Recent systems demonstrate that incorporating external retrieval can substantially improve generation quality in poetry-specific settings (Chatzikyriakidis and Natsina 2025; Y. Liu, L. Lan, et al. 2025).

Finally, agent-based approaches remain largely unexplored in this domain. Preliminary work suggests that multi-agent systems may improve diversity and novelty, though trade-offs with coherence and lexical control remain (R. Zhang and Eger 2024). Combining agent-based methods with CoT and RAG frameworks may offer new ways to balance structure, creativity, and evaluation in future poetry generation systems.

Broader perspective. Although this survey has focused primarily on technical methods and empirical findings, poetry generation is also situated within the broader field of computational creativity. Conceptual frameworks such as Boden's taxonomy of creativity (Boden 2004) provide useful perspectives for interpreting recent advances. Connecting empirical progress in generative poetry with such theoretical models may help clarify both the capabilities and the limitations of current systems, and guide future work at the intersection of NLP and creativity research.

7 Limitations

The primary limitation of this survey stems from the wide variety of approaches used for similar tasks, often with little methodological overlap. This diversity makes it difficult to compare the effectiveness of different systems. For instance, nearly every study employs a unique evaluation protocol, making direct comparison of results challenging even when similar metrics are reported for similar poetry genre and target language. These same factors limited our analysis of tokenization methods, as differences in model architecture, target languages, and evaluation design prevent meaningful generalization of results — for example, findings on syllable-level tokenization in one context may not transfer to others.

Another issue in the generative poetry domain is the strong focus on English and Chinese as the main target languages (Table 7). These languages share several similarities: (1) they use word order to show grammatical relationships, that is, analytic syntax; (2) they have limited inflectional morphology; and (3) they rely heavily on function words. As a result, languages with different grammatical structures, such as Turkish or Russian, are rarely studied in the context of poetry generation. For example, our analysis reveals only 2 papers related to the Russian poetry generation in the ACL Anthology (see ACL Anthology analysis procedure and results in Appendix A), compared to over 40 for both English and Chinese. This lack of representation makes it harder to assess how well the discussed methods work for such languages.

8 Ethical Considerations

This section discusses ethical aspects of generative poetry systems and the creation of this survey.

8.1 Ethics in Generative Poetry Systems

Ethical challenges emerge throughout the development pipeline of generative poetry systems, including data collection, preprocessing, model training, inference and evaluation. We highlight key concerns below, referring readers to Ungless et al. (2025) for detailed discussion.

Data collection raises ethical questions because poetic texts are typically not created for machine learning purposes. Authors may not anticipate their work being used to train models, particularly regarding potential reproduction of text fragments or style imitation by LMs.

Data representation presents another concern, as training datasets may unevenly represent different social groups, languages, or age demographics. Such imbalances can propagate through model outputs.

Data preprocessing introduces ethical considerations through:

- Filtering methods (n-grams, keyword lists, content detectors).
- Potential biases in hate speech detection models.
- Unintended side effects of text cleaning.

Model training involves ethical challenges in tokenization (see more on tokenization in Section 4.1). Tokenizer choices affect text representation across languages and registers, potentially reinforcing linguistic inequalities.

Model inference, particularly in interactive generative systems, may exhibit various forms of unintended bias depending on the content of user prompts. For example, generated outputs can reflect social biases related to gender, ethnicity, or other personal attributes, which are often difficult to control at inference time. An analysis of the sources of such biases, along with bias detection, evaluation and mitigation methods, is provided by E. Sheng and Uthus (2020).

Related concerns about the effects of implicit biases in LLMs on user creativity are discussed by Olatunji (2023). They note that sociocultural, demographic, and age-related biases present in training data can influence model outputs and may limit users' creative expression. As a possible mitigation strategy in poetry co-creation systems, the authors suggest broadening training data to include underrepresented genres, such as rap.

Evaluation presents distinct ethical concerns for automated versus human assessment:

- Automated metrics may favor texts with simpler vocabulary or grammar, and LLM-based evaluators can be implicitly biased toward outputs that resemble the distributions or stylistic preferences of the evaluating model itself, leading to inflated or misleading scores.
- Human evaluation introduces subjectivity through individual preferences and cultural background.
- Assessment protocols may inadvertently introduce systematic biases.

8.2 Ethics of This Survey

We used DeepSeek Assistant[43] for proofreading and improving readability. While some text may be flagged as AI-processed by detection tools, all intellectual content, research, and ideas remain our own. We assume full responsibility for this work.

References

O. Abboushi and M. Azzeh. 2023. "Toward fluent Arabic poem generation based on fine-tuning AraGPT2 transformer." *Arabian Journal for Science and Engineering*, 48, 8, 10537–10549. doi:10.1007/s13369-023-07692-1.

A. Abdibayev, Y. Igarashi, A. Riddell, and D. Rockmore. Nov. 2021. "Automating the Detection of Poetic Features: The Limerick as Model Organism." In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, and S. Szpakowicz. Association for Computational Linguistics, Punta Cana, Dominican Republic (online), (Nov. 2021), 80–90. doi:10.18653/v1/2021.latechclfl-1.9.

[43] https://chat.deepseek.com

A. Abdibayev, A. Riddell, and D. Rockmore. Sept. 2021. "BPoMP: The Benchmark of Poetic Minimal Pairs – Limericks, Rhyme, and Narrative Coherence." In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Ed. by R. Mitkov and G. Angelova. INCOMA Ltd., Online, (Sept. 2021), 1–9. doi:10.26615/978-954-452-072-4_001.

A. Abdibayev, A. Tikhonov, and I. P. Yamshchikov. Sept. 2021. *Dataset of Limericks for Computational Poetics*. Version 3. Zenodo, (Sept. 2021). doi:10.5281/zenodo.5722527.

K. Addonizio and D. Laux. 1997. *The Poet's Companion: A Guide to the Pleasures of Writing Poetry*. W. W. Norton & Company. ISBN: 978-0-393-31654-4. https://aldenhebronenglish.weebly.com/uploads/3/7/7/1/37712587/the_poets_companion_1.pdf.

R. Agarwal and K. Kann. Nov. 2020. "Acrostic Poem Generation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Association for Computational Linguistics, Online, (Nov. 2020), 1230–1240. doi:10.18653/v1/2020.emnlp-main.94.

M. Agirrezabal, I. Alegria, B. Arrieta, and M. Hulden. July 2012. "Finite-State Technology in a Verse-Making Tool." In: *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*. Ed. by I. Alegria and M. Hulden. Association for Computational Linguistics, Donostia–San Sebastián, (July 2012), 35–39. https://aclanthology.org/W12-6206.

M. Agirrezabal, I. Alegria, and M. Hulden. Sept. 2017. "A Comparison of Feature-Based and Neural Scansion of Poetry." In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Ed. by R. Mitkov and G. Angelova. INCOMA Ltd., Varna, Bulgaria, (Sept. 2017), 18–23. doi:10.26615/978-954-452-049-6_003.

M. Agirrezabal, H. Gonçalo Oliveira, and A. Ormazabal. 2023. "Erato: Automatizing Poetry Evaluation." In: *Lecture Notes in Computer Science*. Vol. 14116: *Progress in Artificial Intelligence* (EPIA 2023). Ed. by N. Moniz, Z. Vale, J. Cascalho, C. Silva, and R. Sebastião. Springer Nature Switzerland, Cham, 3–14. ISBN: 978-3-031-49011-8. doi:10.1007/978-3-031-49011-8_1.

M. Agirrezabal and H. G. Oliveira. 2024. "Zero-Shot Metrical Poetry Generation with Open Language Models: a Quantitative Analysis." In: *Proceedings of the 15th International Conference on Computational Creativity* (ICCC 2024) (Jönköping, Sweden, June 17–21, 2024). Ed. by K. Grace, M. T. Llano, P. Martins, and M. M. Hedblom. Association for Computational Creativity (ACC), 272–277. ISBN: 978-989-54160-6-6. https://computationalcreativity.net/iccc24/papers/ICCC24_paper_164.pdf.

E. Agnew, M. Qiu, L. Zhu, S. Wiseman, and C. Rudin. July 2023. "The Mechanical Bard: An Interpretable Machine Learning Approach to Shakespearean Sonnet Generation." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 1627–1638. doi:10.18653/v1/2023.acl-short.140.

R. Aguiar and K. Liao. 2019. *Autonomous Haiku Generation*. (2019). arXiv: 1906.08733 [cs.CL].

S. Ahmadi, H. Hassani, and K. Abedi. May 2020. "A Corpus of the Sorani Kurdish Folkloric Lyrics." eng. In: *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*. Ed. by D. Beermann, L. Besacier, S. Sakti, and C. Soria. European Language Resources association, Marseille, France, (May 2020), 330–335. ISBN: 979-10-95546-35-1. https://aclanthology.org/2020.sltu-1.46.

K. Alnajjar and H. Toivonen. 2021. "Computational generation of slogans." *Natural Language Engineering*, 27, 5, 575–607. doi:10.1017/S1351324920000236.

Z. Alyafeai, M. S. Al-Shaibani, and M. Ahmed. 2023. *Ashaar: Automatic Analysis and Generation of Arabic Poetry Using Deep Learning Approaches*. arXiv: 2307.06218 [cs.CL].

T. M. Amabile. 1983. "A Consensual Technique for Creativity Assessment." In: *The Social Psychology of Creativity*. Springer New York, New York, NY, 37–63. ISBN: 978-1-4612-5533-8. doi:10.1007/978-1-4612-5533-8_3.

M. Amin and M. Burghardt. Dec. 2020. "A Survey on Approaches to Computational Humor Generation." In: *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by S. DeGaetano, A. Kazantseva, N. Reiter, and S. Szpakowicz. International Committee on Computational Linguistics, Online, (Dec. 2020), 29–41. https://aclanthology.org/2020.latechclfl-1.4.

Amina, Abdullah, M. Al Mushabbir, and S. Ahmed. Dec. 2025. "Form-aware Poetic Generation for Bangla." In: *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)*. Ed. by F. Alam, S. Kar, S. A. Chowdhury, N. Hassan, E. H. Prince, M. Tasnim, M. R. A. H. Rony, and M. T. R. Rahman. Association for Computational Linguistics, Mumbai, India, (Dec. 2025), 366–372. ISBN: 979-8-89176-314-2. https://aclanthology.org/2025.banglalp-1.30.

W. Antoun, F. Baly, and H. Hajj. Apr. 2021. "AraGPT2: Pre-Trained Transformer for Arabic Language Generation." In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Ed. by N. Habash, H. Bouamor, H. Hajj, W. Magdy, W. Zaghouani, F. Bougares, N. Tomeh, I. Abu Farha, and S. Touileb. Association for Computational Linguistics, Kyiv, Ukraine (Virtual), (Apr. 2021), 196–207. https://aclanthology.org/2021.wanlp-1.21.

T. Aoyama, S. Behzad, L. Gessler, L. Levine, J. Lin, Y. J. Liu, S. Peng, Y. Zhu, and A. Zeldes. July 2023. "GENTLE: A Genre-Diverse Multilayer Challenge Set for English NLP and Linguistic Evaluation." In: *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*. Ed. by J. Prange and A. Friedrich. Association for Computational Linguistics, Toronto, Canada, (July 2023), 166–178. doi:10.18653/v1/2023.law-1.17.

D. Attridge. 1995. *Poetic rhythm: an introduction*. Cambridge University Press. ISBN: 978-0-521-42369-4.

H. Baayen and R. Lieber. 1991. "Productivity and English derivation: a corpus-based study." *Linguistics*, 29, 5, 801–844. doi:10.1515/ling.1991.29.5.801.

M. Badura, M. Lampert, and R. Dreżewski. 2022. "System Supporting Poetry Generation Using Text Generation and Style Transfer Methods." *Procedia Computer Science*, 207, 3310–3319. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022. doi:10.1016/j.procs.2022.09.389.

J. Bai et al.. 2023. *Qwen Technical Report*. arXiv: 2309.16609 (cs.CL).

A. Baral, H. Jain, D. D, and D. M. H. R. Nov. 2021. "MAPLE – MAsking words to generate blackout Poetry using sequence-to-sequence LEarning." In: *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*. Ed. by M. Abbas and A. A. Freihat. Association for Computational Linguistics, Trento, Italy, (Nov. 2021), 47–54. https://aclanthology.org/2021.icnlsp-1.6.

A. Barbado, V. Fresno, Á. M. Riesco, and S. Ros. June 2022. "DISCO PAL: Diachronic Spanish sonnet corpus with psychological and affective labels." *Language Resources and Evaluation*, 56, 2, (June 2022), 501–542. doi:10.1007/s10579-021-09557-1.

B. Bay, P. Bodily, and D. Ventura. 2017. "Text Transformation Via Constraints and Word Embedding." In: *Proceedings of the Eighth International Conference on Computational Creativity* (ICCC 2017). Ed. by A. Goel, A. Jordanous, and A. Pease. ACC, Atlanta, GA, 49–56. http://computationalcreativity.net/iccc2017/ICCC_17_accepted_submissions/ICCC-17_paper_59.pdf.

M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter. 2024. "xLSTM: Extended Long Short-Term Memory." In: *Advances in Neural Information Processing Systems (NeurIPS 2024)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. Curran Associates, Inc., 107547–107603. https://proceedings.neurips.cc/paper_files/paper/2024/hash/c2ce2f2701c10a2b2f2ea0bfa43cfaa3-Abstract-Conference.html.

R. Behr. May 2024. "Behr at EvaLatin 2024: Latin Dependency Parsing Using Historical Sentence Embeddings." In: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*. Ed. by R. Sprugnoli and M. Passarotti. ELRA and ICCL, Torino, Italia, (May 2024), 198–202. https://aclanthology.org/2024.lt4hala-1.22.

J. Belouadi and S. Eger. July 2023. "ByGPT5: End-to-End Style-conditioned Poetry Generation with Token-free Language Models." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 7364–7381. doi:10.18653/v1/2023.acl-long.406.

B. Bena and J. Kalita. Dec. 2019. "Introducing Aspects of Creativity in Automatic Poetry Generation." In: *Proceedings of the 16th International Conference on Natural Language Processing*. Ed. by D. M. Sharma and P. Bhattacharya. NLP Association of India, International Institute of Information Technology, Hyderabad, India, (Dec. 2019), 26–35. https://aclanthology.org/2019.icon-1.4.

J. Benhardt, P. Hase, L. Zhu, and C. Rudin. 2019. *Shall I Compare Thee to a Machine-Written Sonnet? An Approach to Algorithmic Sonnet Generation*. (2019). arXiv: 1811.05067 (cs.AI).

V. Benigni and F. Masini. 2009. "Compounds in Russian." *Lingue e linguaggio*, 8, 2, 171–194. doi:10.1418/30926.

M. Bernard and H. Titeux. 2021. "Phonemizer: Text to Phones Transcription for Multiple Languages in Python." *Journal of Open Source Software*, 6, 68, 3958. doi:10.21105/joss.03958.

T. Bertin-Mahieux, D. P. W. Ellis, B. Whitman, and P. Lamere. 2011. "The Million Song Dataset." In: *Proceedings of the 12th International Conference on Music Information Retrieval* (ISMIR 2011) (Miami, Florida, USA, Oct. 24–28, 2011), 591–596. https://ismir2011.ismir.net/papers/OS6-1.pdf.

S. Biderman et al.. 2023. "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling." In: *Proceedings of the 40th International Conference on Machine Learning* (Proceedings of Machine Learning Research) (July 23–29, 2023). Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. PMLR, 2397–2430. https://proceedings.mlr.press/v202/biderman23a.html.

BigScience Workshop. 2023. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. arXiv: 2211.05100 (cs.CL).

M. A. Boden. 2004. "In a Nutshell." In: *The Creative Mind: Myths and Mechanisms*. (2nd ed.). Routledge, London, 1–10. ISBN: 978-1-134-37958-3. doi:10.4324/9780203508527.

M. Boggia, S. Ivanova, S. Linkola, A. Kantosalo, and H. Toivonen. 2022. "One Line at a Time — Generation and Internal Evaluation of Interactive Poetry." In: *Proceedings of the 13th International Conference on Computational Creativity* (ICCC 2022) (Bozen-Bolzano, Italy, June 27–July 1, 2022). Ed. by M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz, and T. Veale. Association for Computational Creativity (ACC), 7–11. ISBN: 978-989-54160-4-2. https://computationalcreativity.net/iccc22/papers/ICCC-2022_paper_129.pdf.

K. Booten and K. I. Gero. 2021. "Poetry Machines: Eliciting Designs for Interactive Writing Tools from Poets." In: *Proceedings of the 13th Conference on Creativity and Cognition* (C&C '21) Article 51 (Virtual Event, Italy). Association for Computing Machinery, New York, NY, USA, 5 pages. ISBN: 978-1-4503-8376-9. doi:10.1145/3450741.3466813.

A. Buick. 2024. "Copyright and AI training data—transparency to the rescue?" *Journal of Intellectual Property Law & Practice*, 20, 3, 182–192. doi:10.1093/jiplp/jpae102.

Z. Cai et al.. 2024. *InternLM2 Technical Report*. arXiv: 2403.17297 (cs.CL).

A. Calderwood, J. J. Y. Chung, Y. Sun, M. Roemmele, and M. Kreminski. 2025. "Phraselette: A Poet's Procedural Palette." In: *Proceedings of the 2025 ACM Designing Interactive Systems Conference* (DIS '25). Association for Computing Machinery, New York, NY, USA, 2701–2717. ISBN: 979-8-4007-1485-6. doi:10.1145/3715336.3735832.

A. Calderwood, V. Qiu, K. I. Gero, and L. B. Chilton. 2020. "How novelists use generative language models: An exploratory user study." In: *Joint Proceedings of the Workshops on Human-AI Co-Creation with Generative Models and User-Aware Conversational Agents co-located with 25th International Conference on Intelligent User Interfaces* (IUI 2020) (Cagliari, Italy, Mar. 17, 2020). Ed. by W. Geyer, Y. Khazaeni, and M. Shmueli-Scheuer. https://ceur-ws.org/Vol-2848/HAI-GEN-Paper-3.pdf.

G. Calvi, R. Ginevra, and F. Iurescia. Dec. 2024. "Combining Universal Dependencies and FrameNet to Identify Constructions in a Poetic Corpus: Syntax and Semantics of Latin Felix and Infelix in Virgilian Poetics." In: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*. Ed. by F. Dell'Orletta, A. Lenci, S. Montemagni, and R. Sprugnoli. CEUR Workshop Proceedings, Pisa, Italy, (Dec. 2024), 141–147. ISBN: 979-12-210-7060-6. https://aclanthology.org/2024.clicit-1.18.

F. Carnovalini and A. Rodà. 2020. "Computational creativity and music generation systems: An introduction to the state of the art." *Frontiers in Artificial Intelligence*, 3, Article 14, 20 pages. doi:10.3389/frai.2020.00014.

C. D. B. Cerdas. 2025. *Astrophysical Narratives: Poetic Representations of Gamma-Ray Emission from FermiLAT via Markov Chains*. arXiv: 2501.05692 (astro-ph.HE).

A. I. Cespedosa Vázquez and R. Mitkov. Sept. 2023. "Machine Translation of literary texts: genres, times and systems." In: *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*. Ed. by R. L. Gutiérrez, A. Pareja, and R. Mitkov. INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, (Sept. 2023), 48–53. https://aclanthology.org/2023.nlp4tia-1.7.

T. Chakrabarty, V. Padmakumar, and H. He. Dec. 2022. "Help me write a Poem: Instruction Tuning as a Vehicle for Collaborative Poetry Writing." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, and Y. Zhang. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, (Dec. 2022), 6848–6863. doi:10.18653/v1/2022.emnlp-main.460.

T. Chakrabarty, A. Saakyan, and S. Muresan. Nov. 2021. "Don't Go Far Off: An Empirical Study on Neural Poetry Translation." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, (Nov. 2021), 7253–7265. doi:10.18653/v1/2021.emnlp-main.577.

T. Chakrabarty, X. Zhang, S. Muresan, and N. Peng. June 2021. "MERMAID: Metaphor Generation with Symbolism and Discriminative Decoding." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Association for Computational Linguistics, Online, (June 2021), 4250–4261. doi:10.18653/v1/2021.naacl-main.336.

Y. Chang, R. Zhang, L. Jiang, Q. Chen, L. Zhang, and J. Pu. Dec. 2023. "Sudowoodo: A Chinese Lyric Imitation System with Source Lyrics." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Y. Feng and E. Lefever. Association for Computational Linguistics, Singapore, (Dec. 2023), 99–105. doi:10.18653/v1/2023.emnlp-demo.8.

S. Chatzikyriakidis and A. Natsina. May 2025. "Poetry in RAGs: Modern Greek interwar poetry generation using RAG and contrastive training." In: *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. Ed. by M. Hämäläinen, E. Öhman, Y. Bizzoni, S. Miyagawa, and K. Alnajjar. Association for Computational Linguistics, Albuquerque, USA, (May 2025), 257–264. ISBN: 979-8-89176-234-3. doi:10.18653/v1/2025.nlp4dh-1.22.

T. Che, Y. Li, R. Zhang, R. D. Hjelm, W. Li, Y. Song, and Y. Bengio. 2017. *Maximum-Likelihood Augmented Discrete Generative Adversarial Networks*. arXiv: 1702.07983 (cs.AI).

A. Chen, L. Lou, K. Chen, X. Bai, Y. Xiang, M. Yang, T. Zhao, and M. Zhang. 2024. *Large Language Models for Classical Chinese Poetry Translation: Benchmarking, Evaluating, and Improving*. arXiv: 2408.09945 (cs.CL).

H. Chen, X. Yi, M. Sun, W. Li, C. Yang, and Z. Guo. 2019. "Sentiment-Controllable Chinese Poetry Generation." In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (Macao, Aug. 10–16, 2019). Ed. by S. Kraus. International Joint Conferences on Artificial Intelligence, 4925–4931. ISBN: 978-0-9992411-4-1. doi:10.24963/ijcai.2019/684.

Y. Chen, H. Gröner, S. Zarrieß, and S. Eger. Nov. 2024. "Evaluating Diversity in Automatic Poetry Generation." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 19671–19692. doi:10.18653/v1/2024.emnlp-main.1097.

Y. Chen and A. Lerch. 2020. "Melody-Conditioned Lyrics Generation with SeqGANs." In: *2020 IEEE International Symposium on Multimedia (ISM)* (Naples, Italy (virtual), Dec. 2–4, 2020). IEEE, 189–196. ISBN: 978-1-7281-8697-9. doi:10.1109/ism.2020.00040.

Y. Chen and S. Teufel. May 2024. "Scansion-based Lyrics Generation." In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue. ELRA and ICCL, Torino, Italia, (May 2024), 14370–14381. https://aclanthology.org/2024.lrec-main.1252.

J. Cheng, C. Pan, and S. Li. Nov. 2025. "ToneCraft: Cantonese Lyrics Generation with Harmony of Tones and Pitches." In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng. Association for Computational Linguistics, Suzhou, China, (Nov. 2025), 335–353. ISBN: 979-8-89176-332-6. doi:10.18653/v1/2025.emnlp-main.18.

K.-Y. Chiang, S. Lin, J. Chen, Q. Yin, and Q. Jin. 2021. *TransCouplet: Transformer based Chinese Couplet Generation.* (2021). arXiv: 2112.01707 (cs.CL).

A. Chiche and B. Yitagesu. Jan. 2022. "Part of speech tagging: a systematic review of deep learning and machine learning approaches." *Journal of Big Data*, 9, 1, (Jan. 2022), 10. doi:10.1186/s40537-022-00561-y.

W. Cho, Y. Kim, S. Lee, and Y. Yu. Nov. 2025. "MAVL: A Multilingual Audio-Video Lyrics Dataset for Animated Song Translation." In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng. Association for Computational Linguistics, Suzhou, China, (Nov. 2025), 13640–13668. ISBN: 979-8-89176-332-6. doi:10.18653/v1/2025.emnlp-main.689.

M. Choi, S. Lee, E. Choi, H. Park, J. Lee, D. Lee, and J. Lee. June 2021. "MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Association for Computational Linguistics, Online, (June 2021), 1763–1773. doi:10.18653/v1/2021.naacl-main.141.

A. Chowdhery et al.. 2023. "PaLM: Scaling Language Modeling with Pathways." *Journal of Machine Learning Research*, 24, 240, 1–113. http://jmlr.org/papers/v24/22-1144.html.

M. Chudoba and R. Rosa. 2024. *GPT Czech Poet: Generation of Czech Poetic Strophes with Language Models.* arXiv: 2407.12790 (cs.CL).

E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith. Aug. 2021. "All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Association for Computational Linguistics, Online, (Aug. 2021), 7282–7296. doi:10.18653/v1/2021.acl-long.565.

J. H. Clark, D. Garrette, I. Turc, and J. Wieting. Jan. 2022. "CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation." *Transactions of the Association for Computational Linguistics*, 10, (Jan. 2022), 73–91. doi:10.1162/tacl_a_00448.

A. Coil and V. Shwartz. July 2023. "From chocolate bunny to chocolate crocodile: Do Language Models Understand Noun Compounds?" In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 2698–2710. doi:10.18653/v1/2023.findings-acl.169.

J. Coles. Oct. 2017. "Planting Poetry: Sowing Seeds of Creativity in a Year 5 Class." *Changing English*, 24, 4, (Oct. 2017), 386–398. doi:10.1080/1358684x.2017.1308806.

M. Colhon, C. Mărănduc, and C. Mititelu. Sept. 2017. "A Multiform Balanced Dependency Treebank for Romanian." In: *Proceedings of the Workshop Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP 2017*. Ed. by K. Zervanou, P. Osenova, E. Wandl-Vogt, and D. Cristea. INCOMA Inc., Varna, (Sept. 2017), 9–18. doi:10.26615/978-954-452-040-3_002.

S. Colton. 2008. "Creativity Versus the Perception of Creativity in Computational Systems." In: *Papers from the 2008 AAAI Spring Symposium. Creative intelligent systems* (Palo Alto, California, Mar. 26–28, 2008). Association for the Advancement of Artificial Intelligence. https://cdn.aaai.org/Symposia/Spring/2008/SS-08-03/SS08-03-003.pdf.

S. Colton, J. Goodwin, and T. Veale. May 2012. "Full-FACE Poetry Generation." In: *Proceedings of the Third International Conference on Computational Creativity* (ICCC 2012). Ed. by M. L. Maher, K. Hammond, A. Pease, R. Pérez, D. Ventura, and G. Wiggins. Dublin, Ireland, (May 2012), 95–102. http://computationalcreativity.net/iccc2012/wp-content/uploads/2012/05/095-Colton.pdf.

S. Colton and G. A. Wiggins. 2012. "Computational creativity: The final frontier?" In: *20th European Conference on Artificial Intelligence (ECAI'12)*. IOS Press, Montpellier, France, 21–26. doi:10.3233/978-1-61499-098-7-21.

C. Corbetta, M. Passarotti, and G. Moretti. May 2024. "The Rise and Fall of Dependency Parsing in Dante Alighieri's Divine Comedy." In: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*. Ed. by R. Sprugnoli and M. Passarotti. ELRA and ICCL, Torino, Italia, (May 2024), 50–56. https://aclanthology.org/2024.lt4hala-1.7.

M. A. Covington and J. D. McFall. 2010. "Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR)." *Journal of Quantitative Linguistics*, 17, 2, 94–100. doi:10.1080/09296171003643098.

E. Crothers, H. L. Viktor, and N. Japkowicz. 2023. "In BLOOM: Creativity and Affinity in Artificial Lyrics and Art." In: *The AAAI-23 Workshop on Creative AI Across Modalities* (Washington DC, USA). https://openreview.net/forum?id=3hk5PFxQSG.

D. Crystal. 2008. *A dictionary of linguistics and phonetics*. Blackwell Publishing. ISBN: 978-1-4443-0277-6. doi:10.1002/9781444302776.

L. D'Souza and D. Mimno. Dec. 2023. "The Chatbot and the Canon: Poetry Memorization in LLMs." In: *Proceedings of the Computational Humanities Research Conference 2023*. Ed. by A. Sela, F. Jannidis, and I. Romanowska. Vol. 3558. Paris, France, (Dec. 2023), 475–489. https://ceur-ws.org/Vol-3558/paper5712.pdf.

A. Dai. 2021. *GPT-2 for Emily Dickinson poetry generation.* Course Project Report for Fall 2021 CS230: Deep Learning. Department of Computer Science, Stanford University. https://cs230.stanford.edu/projects_fall_2021/reports/103051256.pdf.

W. Dai, J. Li, D. O. N. G. X. U. LI, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. 2023. "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning." In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 49250–49267. https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

S. Das, L. Jin, L. Song, H. Mi, B. Peng, and D. Yu. Jan. 2025. "Entropy Guided Extrapolative Decoding to Improve Factuality in Large Language Models." In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert. Association for Computational Linguistics, Abu Dhabi, UAE, (Jan. 2025), 6589–6600. https://aclanthology.org/2025.coling-main.439.

E. Davis. Nov. 2024. *ChatGPT's Poetry is Incompetent and Banal: A Discussion of (Porter and Machery, 2024)*. Unpublished. (Nov. 2024). https://cs.nyu.edu/~davise/papers/GPT-Poetry.pdf.

M. De Araujo Possi, A. De Paiva Oliveira, A. Moreira, and L. M. Costa. 2023. "CARMEN: A Method for Automatic Evaluation of Poems." In: *2023 5th International Conference on Natural Language Processing (ICNLP)*. IEEE, 244–247. doi:10.1109/icnlp58431.2023.00051.

M. De Sisto, L. Hernández-Lorenzo, J. De la Rosa, S. Ros, and E. González-Blanco. Feb. 2024. "Understanding poetry using natural language processing tools: a survey." *Digital Scholarship in the Humanities*, 39, 2, (Feb. 2024), 500–521. doi:10.1093/llc/fqae001.

DeepSeek-AI. 2024. *DeepSeek LLM: Scaling Open-Source Language Models with Longtermism*. (2024). arXiv: 2401.02954 (cs.CL).

B. Deiseroth, M. Brack, P. Schramowski, K. Kersting, and S. Weinbach. Nov. 2024. "T-FREE: Subword Tokenizer-Free Generative LLMs via Sparse Representations for Memory-Efficient Embeddings." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 21829–21851. doi:10.18653/v1/2024.emnlp-main.1217.

Y. Deng, W. Zhao, J. Hessel, X. Ren, C. Cardie, and Y. Choi. Nov. 2024. "WildVis: Open Source Visualizer for Million-Scale Chat Logs in the Wild." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by D. I. Hernandez Farias, T. Hope, and M. Li. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 497–506. doi:10.18653/v1/2024.emnlp-demo.50.

Z. Deng, H. Yang, and J. Wang. 2024. *Can AI Write Classical Chinese Poetry like Humans? An Empirical Study Inspired by Turing Test*. arXiv: 2401.04952 (cs.CL).

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. June 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 4171–4186. doi:10.18653/v1/N19-1423.

S. Dhuliawala, I. Kulikov, P. Yu, A. Celikyilmaz, J. Weston, S. Sukhbaatar, and J. Lanchantin. 2024. *Adaptive Decoding via Latent Preference Optimization*. arXiv: 2411.09661 (cs.CL).

S. Ding, Z. Liu, X. Dong, P. Zhang, R. Qian, J. Huang, C. He, D. Lin, and J. Wang. July 2025. "SongComposer: A Large Language Model for Lyric and Melody Generation in Song Composition." In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Association for Computational Linguistics, Vienna, Austria, (July 2025), 7108–7127. ISBN: 979-8-89176-251-0. doi:10.18653/v1/2025.acl-long.352.

A. R. Doshi and O. P. Hauser. July 2024. "Generative AI enhances individual creativity but reduces the collective diversity of novel content." *Science Advances*, 10, 28, (July 2024), eadn5290. doi:10.1126/sciadv.adn5290.

R. G. Dunphy, (Ed.) . 2010. *Encyclopedia of the Medieval Chronicle*. Vol. 1. Brill Leiden. ISBN: 978-90-04-18464-0. Retrieved July 8, 2024 from https://archive.org/details/encyclopedia-of-medieval-chronicle-v/Encyclopedia%20of%20Medieval%20Chronicle%20I/mode/1up.

M. Elzohbi and R. Zhao. June 2023. "Creative data generation: A review focusing on text and poetry." In: *Proceedings of the 14th International Conference on Computational Creativity* (ICCC 2023). Association for Computational Creativity (ACC), Ontario, Canada, (June 2023), 29–38. https://computationalcreativity.net/iccc23/papers/ICCC-2023_paper_10.pdf.

M. Elzohbi and R. Zhao. 2024. "Let the Poem Hit the Rhythm: Using a Byte-Based Transformer for Beat-Aligned Poetry Generation." In: *Proceedings of the 15th International Conference on Computational Creativity* (ICCC 2024) (Jönköping, Sweden, June 17–21, 2024). Ed. by K. Grace, M. T. Llano, P. Martins, and M. M. Hedblom. Association for Computational Creativity (ACC), 407–411. ISBN: 978-989-54160-6-6. https://computationalcreativity.net/iccc24/papers/ICCC24_paper_192.pdf.

M. Elzohbi and R. Zhao. 2025a. "Poems to Lyrics: Automated Rephrasing with Beat Alignment." In: *Proceedings of the 16th International Conference on Computational Creativity* (ICCC 2025) (Campinas, Brazil, June 23–27, 2025). https://computationalcreativity.net/iccc25/wp-content/uploads/papers/iccc25-elzohbi2025poems.pdf.

M. Elzohbi and R. Zhao. Nov. 2025b. "Tahḍīb: A Rhythm-Aware Phrase Insertion for Classical Arabic Poetry Composition." In: *Proceedings of The Third Arabic Natural Language Processing Conference*. Ed. by K. Darwish et al. Association for Computational Linguistics, Suzhou, China, (Nov. 2025), 194–202. ISBN: 979-8-89176-352-4. doi:10.18653/v1/2025.arabicnlp-main.15.

H. Fan, J. Wang, B. Zhuang, S. Wang, and J. Xiao. 2019. "Automatic Acrostic Couplet Generation with Three-Stage Neural Network Pipelines." In: *Lecture Notes in Computer Science*. Vol. 11670: *PRICAI 2019: Trends in Artificial Intelligence. Proceedings of 16th Pacific Rim International Conference on Artificial Intelligence* (Cuvu, Yanuca Island, Fiji, Aug. 26–30, 2019). Ed. by A. C. Nayak and A. Sharma. Springer International Publishing, Cham, 314–324. ISBN: 978-3-030-29908-8. doi:10.1007/978-3-030-29908-8_25.

A. Fedchin, I. Cooperman, P. Chaudhuri, and J. P. Dexter. Apr. 2025. "AcrosticSleuth: Probabilistic Identification and Ranking of Acrostics in Multilingual Corpora." In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Ed. by L. Chiruzzo, A. Ritter, and

L. Wang. Association for Computational Linguistics, Albuquerque, New Mexico, (Apr. 2025), 7430–7437. ISBN: 979-8-89176-195-7. doi:10.18653/v1/2025.findings-naacl.414.

C. Fellbaum. 2010. "WordNet." In: *Theory and Applications of Ontology: Computer Applications*. Ed. by R. Poli, M. Healy, and A. Kameas. Springer Netherlands, Dordrecht, 231–243. ISBN: 978-90-481-8847-5. doi:10.1007/978-90-481-8847-5_10.

H. Ferraz de Arruda, S. M. Reia, F. N. Silva, D. R. Amancio, and L. da Fontoura Costa. 2022. "Finding contrasting patterns in rhythmic properties between prose and poetry." *Physica A: Statistical Mechanics and its Applications*, 598, 127387. doi:10.1016/j.physa.2022.127387.

L. Flower and J. R. Hayes. Dec. 1981. "A Cognitive Process Theory of Writing." *College Composition & Communication*, 32, 4, (Dec. 1981), 365–387. doi:10.2307/356600.

G. Franceschelli and M. Musolesi. 2024a. "Creative Beam Search: LLM-as-a-Judge for Improving Response Generation." In: *Proceedings of the 15th International Conference on Computational Creativity* (ICCC 2024) (Jönköping, Sweden, June 17–21, 2024). Ed. by K. Grace, M. T. Llano, P. Martins, and M. M. Hedblom. Association for Computational Creativity (ACC), 364–368. ISBN: 978-989-54160-6-6. https://computationalcreativity.net/iccc24/papers/ICCC24_paper_161.pdf.

G. Franceschelli and M. Musolesi. June 2024b. "Creativity and Machine Learning: A Survey." *ACM Computing Surveys*, 56, 11, Article 283, (June 2024), 41 pages. doi:10.1145/3664595.

G. Franceschelli and M. Musolesi. Dec. 2022. "Deepcreativity: measuring creativity with deep learning techniques." *Intelligenza Artificiale*, 16, 2, (Dec. 2022), 151–163. doi:10.3233/ia-220136.

G. Franceschelli and M. Musolesi. Nov. 2024c. "On the creativity of large language models." *AI & SOCIETY*, (Nov. 2024). doi:10.1007/s00146-024-02127-3.

E. Frith, D. B. Elbich, A. P. Christensen, M. D. Rosenberg, Q. Chen, M. J. Kane, P. J. Silvia, P. Seli, and R. E. Beaty. Apr. 2021. "Intelligence and creativity share a common cognitive and neural basis." *Journal of Experimental Psychology: General*, 150, 4, (Apr. 2021), 609–632. doi:10.1037/xge0000958.

J. Fuller. July 2017. *The Sonnet*. The Critical Idiom Reissued. Taylor & Francis, (July 2017). ISBN: 978-1-315-11543-6. doi:10.4324/9781315115436.

P. Fussell. 1979. *Poetic Meter and Poetic Form*. McGraw Hill, New York. 188 pp. ISBN: 978-0-394-32120-2.

R. P. Gabriel. 2016. "In the control room of the banquet." In: *Proceedings of the 2016 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software* (Onward! 2016). Association for Computing Machinery, Amsterdam, Netherlands, 250–268. ISBN: 9781450340762. doi:10.1145/2986012.2986028.

R. C. Gale, A. C. Salem, G. Fergadiotis, and S. Bedrick. July 2023. "Mixed Orthographic/Phonemic Language Modeling: Beyond Orthographically Restricted Transformers (BORT)." In: *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*. Ed. by B. Can et al. Association for Computational Linguistics, Toronto, Canada, (July 2023), 212–225. doi:10.18653/v1/2023.repl4nlp-1.18.

[SW] L. Gao et al., *A framework for few-shot language model evaluation* 2025. doi:10.5281/zenodo.5371628, URL: https://github.com/EleutherAI/lm-evaluation-harness.

E. Garces Arias, J. Rodemann, M. Li, C. Heumann, and M. Aßenmacher. Nov. 2024. "Adaptive Contrastive Search: Uncertainty-Guided Decoding for Open-Ended Text Generation." In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 15060–15080. doi:10.18653/v1/2024.findings-emnlp.885.

D. Gareev, T. Hofmann, E. Krishnasamy, and T. Pimentel. Nov. 2024. "Local and Global Decoding in Text Generation." In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 14577–14597. doi:10.18653/v1/2024.findings-emnlp.854.

A. M. Garzón and M. L. Pérez. 2020. "EDA and NLP basics: Exploring the innards of the Spanish poetry." In: Python Conference Spain. PyConES (Oct. 2–3, 2020). Asociación Python España. https://2020.es.pycon.org/speaker_4.html.

A. Gatt and E. Krahmer. Jan. 2018. "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation." *Journal of Artificial Intelligence Research*, 61, (Jan. 2018), 65–170. doi:10.1613/jair.5477.

L. Gatti, G. Özbal, O. Stock, and C. Strapparava. Apr. 2017. "To Sing like a Mockingbird." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by M. Lapata, P. Blunsom, and A. Koller. Association for Computational Linguistics, Valencia, Spain, (Apr. 2017), 298–304. doi:10.18653/v1/e17-2048.

S. Geng, M. Josifoski, M. Peyrard, and R. West. Dec. 2023. "Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Association for Computational Linguistics, Singapore, (Dec. 2023), 10932–10952. doi:10.18653/v1/2023.emnlp-main.674.

S. Al-Ghamdi, H. Al-Khalifa, and A. Al-Salman. Dec. 2021. "A Dependency Treebank for Classical Arabic Poetry." In: *Proceedings of the Sixth International Conference on Dependency Linguistics (Depling, SyntaxFest 2021)*. Ed. by N. Mazziotta and S. Mille. Association for Computational Linguistics, Sofia, Bulgaria, (Dec. 2021), 1–9. https://aclanthology.org/2021.depling-1.1.

M. Ghazvininejad, Y. Choi, and K. Knight. June 2018. "Neural Poetry Translation." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by M. Walker, H. Ji, and A. Stent. Association for Computational Linguistics, New Orleans, Louisiana, (June 2018), 67–71. doi:10.18653/v1/N18-2011.

M. Ghazvininejad, X. Shi, Y. Choi, and K. Knight. Nov. 2016. "Generating Topical Poetry." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by J. Su, K. Duh, and X. Carreras. Association for Computational Linguistics, Austin, Texas, (Nov. 2016), 1183–1191. doi:10.18653/v1/d16-1126.

M. Goel, P. Krishnamurthy, and R. Mamidi. Dec. 2024. "Automating Humor: A Novel Approach to Joke Generation Using Template Extraction and Infilling." In: *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*. Ed. by S. Lalitha Devi and K. Arora. NLP Association of India (NLPAI), AU-KBC Research Centre, Chennai, India, (Dec. 2024), 442–448. https://aclanthology.org/2024.icon-1.51.

M. Gokirmak. 2021. "Converting prose into poetry using neural networks." Master's thesis. Univerzita Karlova, Matematicko-fyzikální fakulta, Prague. http://hdl.handle.net/20.500.11956/148157.

H. Gonçalo Oliveira. Sept. 2017. "A Survey on Intelligent Poetry Generation: Languages, Features, Techniques, Reutilisation and Evaluation." In: *Proceedings of the 10th International Conference on Natural Language Generation*. Ed. by J. M. Alonso, A. Bugarín, and E. Reiter. Association for Computational Linguistics, Santiago de Compostela, Spain, (Sept. 2017), 11–20. doi:10.18653/v1/W17-3502.

H. Gonçalo Oliveira. 2021. "Exploring a Masked Language Model for Creative Text Transformation." In: *Proceedings of the 12th International Conference on Computational Creativity* (ICCC 2021). Ed. by A. G. de Silva Garza, T. Veale, W. Aguilar, and R. P. y Pérez. Association for Computational Creativity, México City, México (Virtual), 62–71. ISBN: 978-989-54160-3-5. https://computationalcreativity.net/iccc21/wp-content/uploads/2021/09/ICCC_2021_paper_39.pdf.

H. Gonçalo Oliveira. 2012. "PoeTryMe: a versatile platform for poetry generation." In: *Proceedings of the ECAI 2012 Workshop on Computational Creativity, Concept Invention, and General Intelligence* (C3GI at ECAI 2012) (Montpellier, France, Aug. 27–28, 2012), 21–27. http://eden.dei.uc.pt/~hroliv/pubs/GoncaloOliveira2012_c3gi_CRC.pdf.

H. Gonçalo Oliveira. 2015. "Tra-la-Lyrics 2.0: Automatic Generation of Song Lyrics on a Semantic Domain." *Journal of Artificial General Intelligence*, 6, 1, 87–110. doi:10.1515/jagi-2015-0005.

H. Gonçalo Oliveira. 2020. "WeirdAnalogyMatic: Experimenting with Analogy for Lyrics Transformation." In: *Proceedings of the 11th International Conference on Computational Creativity* (ICCC 2020). Ed. by F. A. Cardoso, P. Machado, T. Veale, and J. M. Cunha. Association for Computational Creativity, Coimbra, Portugal, 228–235. ISBN: 978-989-54160-2-8. http://computationalcreativity.net/iccc20/papers/047-iccc20.pdf.

H. Gonçalo Oliveira, R. Hervás, A. D'ıaz, and P. Gervás. June 2014. "Adapting a Generic Platform for Poetry Generation to Produce Spanish Poem." In: *Proceedings of the Fifth International Conference on Computational Creativity* (ICCC 2014). Ed. by S. Colton, D. Ventura, N. Lavrač, and M. Cook. Ljubljana, Slovenia, (June 2014), 63–71. http://computationalcreativity.net/iccc2014/wp-content/uploads/2014/06//6.2_Oliveira.pdf.

H. Gonçalo Oliveira, T. Mendes, and A. Boavida. Sept. 2017. "Co-PoeTryMe: a Co-Creative Interface for the Composition of Poetry." In: *Proceedings of the 10th International Conference on Natural Language Generation*. Ed. by J. M. Alonso, A. Bugarín, and E. Reiter. Association for Computational Linguistics, Santiago de Compostela, Spain, (Sept. 2017), 70–71. doi:10.18653/v1/W17-3508.

H. Gonçalo Oliveira and R. Rodrigues. Nov. 2018. "Exploring Lexical-Semantic Knowledge in the Generation of Novel Riddles in Portuguese." In: *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018)*. Ed. by H. Gonçalo Oliveira, B. Burtenshaw, and R. Hervás. Association for Computational Linguistics, Tilburg, the Netherlands, (Nov. 2018), 17–25. doi:10.18653/v1/W18-6604.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf.

E. Greene, T. Bodrumlu, and K. Knight. Oct. 2010. "Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation." In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Li and L. Màrquez. Association for Computational Linguistics, Cambridge, MA, (Oct. 2010), 524–533. https://aclanthology.org/D10-1051.

R. Greene, S. Cushman, C. Cavanagh, J. Ramazani, and P. Rouzer. 2012. *The Princeton encyclopedia of poetry and poetics*. Princeton University Press.

A. Gu and T. Dao. 2024. "Mamba: Linear-Time Sequence Modeling with Selective State Spaces." In: *First Conference on Language Modeling* (COLM 2024) (Philadelphia, Pennsylvania, USA, Oct. 7–9, 2024). https://openreview.net/forum?id=tEYskw1VY2.

V. E. Gunser, S. Gottschling, B. Brucker, S. Richter, D. C. Çakir, and P. Gerjets. May 2022. "The Pure Poet: How Good is the Subjective Credibility and Stylistic Quality of Literary Short Texts Written with an Artificial Intelligence Tool as Compared to Texts Written by Human Authors?" In: *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Ed. by T.-H. '. Huang, V. Raheja, D. Kang, J. J. Y. Chung, D. Gissin, M. Lee, and K. I. Gero. Association for Computational Linguistics, Dublin, Ireland, (May 2022), 60–61. doi:10.18653/v1/2022.in2writing-1.8.

A. Guo, S. Sathyanarayanan, L. Wang, J. Heer, and A. X. Zhang. 2025. "From Pen to Prompt: How Creative Writers Integrate AI into their Writing Practice." In: *Proceedings of the 2025 Conference on Creativity and Cognition* (C&C '25). Association for Computing Machinery, New York, NY, USA, 527–545. ISBN: 979-8-4007-1289-0. doi:10.1145/3698061.3726910.

J. Guo et al.. Nov. 2025. "M-Ped: Multi-Prompt Ensemble Decoding for Large Language Models." In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng. Association for Computational Linguistics, Suzhou, China, (Nov. 2025), 16693–16711. ISBN: 979-8-89176-335-7. doi:10.18653/v1/2025.findings-emnlp.906.

Z. Guo, J. Hu, and M. Sun. 2020. *BERT-CCPoem, a pre-trained model for Chinese classical poetry*. Research Center for Natural Language Processing, Computational Humanities and Social Sciences, Tsinghua University. https://github.com/THUNLP-AIPoet/BERT-CCPoem.

Z. Guo, X. Yi, M. Sun, W. Li, C. Yang, J. Liang, H. Chen, Y. Zhang, and R. Li. July 2019. "Jiuge: A Human-Machine Collaborative Chinese Classical Poetry Generation System." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by M. R. Costa-jussà and E. Alfonseca. Association for Computational Linguistics, Florence, Italy, (July 2019), 25–30. doi:10.18653/v1/P19-3005.

N. Habash, M. AbuOdeh, D. Taji, R. Faraj, J. El Gizuli, and O. Kallas. June 2022. "Camel Treebank: An Open Multi-genre Arabic Dependency Treebank." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by N. Calzolari et al. European Language Resources Association, Marseille, France, (June 2022), 2672–2681. https://aclanthology.org/2022.lrec-1.286.

T. Haider. May 2024. "A Large Annotated Reference Corpus of New High German Poetry." In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue. ELRA and ICCL, Torino, Italia, (May 2024), 677–683. https://aclanthology.org/2024.lrec-main.59.

T. Haider. Apr. 2021. "Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by P. Merlo, J. Tiedemann, and R. Tsarfaty. Association for Computational Linguistics, Online, (Apr. 2021), 3715–3725. doi:10.18653/v1/2021.eacl-main.325.

T. Haider, S. Eger, E. Kim, R. Klinger, and W. Menninghaus. May 2020. "PO-EMO: Conceptualization, Annotation, and Modeling of Aesthetic Emotions in German and English Poetry." In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari et al. European Language Resources Association, Marseille, France, (May 2020), 1652–1663. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.205.

T. Haider and J. Kuhn. Aug. 2018. "Supervised Rhyme Detection with Siamese Recurrent Networks." In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by B. Alex, S. Degaetano-Ortlieb, A. Feldman, A. Kazantseva, N. Reiter, and S. Szpakowicz. Association for Computational Linguistics, Santa Fe, New Mexico, (Aug. 2018), 81–86. https://aclanthology.org/W18-4509.

M. Hämäläinen. 2018a. "Harnessing NLG to create Finnish poetry automatically." In: *Proceedings of the 9th International Conference on Computational Creativity* (ICCC 2018) (Salamanca, Spain, June 25–29, 2018). Ed. by F. Pachet, A. Jordanous, and C. León. Association for Computational Creativity (ACC), 9–16. ISBN: 978-989-54160-0-4. https://computationalcreativity.net/iccc2018/sites/default/files/papers/IICCC_2018_paper_6.pdf.

M. Hämäläinen. Nov. 2018b. "Poem Machine - a Co-creative NLG Web Application for Poem Writing." In: *Proceedings of the 11th International Conference on Natural Language Generation*. Ed. by E. Krahmer, A. Gatt, and M. Goudbeek. Association for Computational Linguistics, Tilburg University, The Netherlands, (Nov. 2018), 195–196. doi:10.18653/v1/W18-6525.

M. Hämäläinen and K. Alnajjar. Nov. 2019a. "Generating Modern Poetry Automatically in Finnish." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Association for Computational Linguistics, Hong Kong, China, (Nov. 2019), 5999–6004. doi:10.18653/v1/D19-1617.

M. Hämäläinen and K. Alnajjar. Aug. 2021. "Human Evaluation of Creative NLG Systems: An Interdisciplinary Survey on Recent Papers." In: *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Ed. by A. Bosselut, E. Durmus, V. P. Gangal, S. Gehrmann, Y. Jernite, L. Perez-Beltrachini, S. Shaikh, and W. Xu. Association for Computational Linguistics, Online, (Aug. 2021), 84–95. doi:10.18653/v1/2021.gem-1.9.

M. Hämäläinen and K. Alnajjar. Oct. 2019b. "Let's FACE it. Finnish Poetry Generation with Aesthetics and Framing." In: *Proceedings of the 12th International Conference on Natural Language Generation*. Ed. by K. van Deemter, C. Lin, and H. Takamura. Association for Computational Linguistics, Tokyo, Japan, (Oct. 2019), 290–300. doi:10.18653/v1/W19-8637.

M. Hämäläinen, K. Alnajjar, and T. Poibeau. 2022. "Modern French Poetry Generation with RoBERTa and GPT-2." In: *Proceedings of the 13th International Conference on Computational Creativity* (ICCC 2022) (Bozen-Bolzano, Italy, June 27–July 1, 2022). Ed. by M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz, and T. Veale. Association for Computational Creativity (ACC), 12–16. ISBN: 978-989-54160-4-2. https://computationalcreativity.net/iccc22/wp-content/uploads/2022/06/ICCC-2022_10S_H%C3%A4m%C3%A4l%C3%A4inen-et-al..pdf.

A. Hamat. June 2024. "The Language of AI and Human Poetry: A Comparative Lexicometric Study." *3L The Southeast Asian Journal of English Language Studies*, 30, 2, (June 2024), 1–20. doi:10.17576/3l-2024-3002-01.

D. I. Hanauer. 2010. *Poetry as Research: Exploring Second Language Poetry Writing*. John Benjamins. ISBN: 9789027288301. doi:10.1075/lal.9.

A. Hans et al.. 2024. "Be like a Goldfish, Don't Memorize! Mitigating Memorization in Generative LLMs." In: *Advances in Neural Information Processing Systems (NeurIPS 2024)*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. Curran Associates, Inc., 24022–24045. https://proceedings.neurips.cc/paper_files/paper/2024/hash/2ad2dffba5079687651226ac8752df97-Abstract-Conference.html.

S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. E. Weston, and Y. Tian. 2025. "Training Large Language Models to Reason in a Continuous Latent Space." In: *ICLR 2025 Workshop on Reasoning and Planning for Large Language Models* (Singapore, Apr. 28, 2025). https://openreview.net/forum?id=KrWSrrYGpT.

L. E. Harr. July 1975. "Haiku Poetry." *Journal of Aesthetic Education*, 9, 3, (July 1975), 112–119. doi:10.2307/3331909.

D. A. Haslett. Mar. 2025. "Tokenization Changes Meaning in Large Language Models: Evidence from Chinese." *Computational Linguistics*, (Mar. 2025), 1–30. doi:10.1162/coli_a_00557.

K. Hayawi, S. Shahriar, and S. S. Mathew. Feb. 2024. "The imitation game: Detecting human and AI-generated texts in the era of ChatGPT and BARD." *Journal of Information Science*, (Feb. 2024), 1–36. doi:10.1177/01655515241227531.

I. van Heerden and A. Bas. Apr. 2021. "AfriKI: Machine-in-the-Loop Afrikaans Poetry Generation." In: *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Ed. by S. L. Blodgett, M. Madaio, B. O'Connor, H. Wallach, and Q. Yang. Association for Computational Linguistics, Online, (Apr. 2021), 74–80. https://aclanthology.org/2021.hcinlp-1.12.

P. J. Hegade, V. G. K. Rajaram M.and Hegde, and S. Tejaswiniand Basavaraddi. Aug. 2021. "Po-Miner: A Web Mining Poem Generator and its Security Model." *SN Computer Science*, 2, 5, (Aug. 2021), 401. doi:10.1007/s42979-021-00802-6.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. 2021. "Measuring Massive Multitask Language Understanding." In: *9th International Conference on Learning Representations* (ICLR 2021) (Virtual Event, Austria, May 3–7, 2021). https://openreview.net/forum?id=d7KBjmI3GmQ.

W. N. Herbert, F. R. Jones, and F. Sampson. June 2024. *Collaborative Poetry Translation: Processes, Priorities, and Relationships in the Poettrio Method*. Routledge, New York, NY, (June 2024). 1227 pp. ISBN: 978-0-429-03029-1. doi:10.4324/9780429030291.

W. J. Higginson and P. Harter. 1985. *The Haiku Handbook: How to Write, Share, and Teach Haiku*. McGraw-Hill. 331 pp. ISBN: 978-0-07-028786-0.

H. Hirjee and D. G. Brown. 2009. "Automatic Detection of Internal and Imperfect Rhymes in Rap Lyrics." In: *Proceedings of the 10th International Society for Music Information Retrieval Conference* (ISMIR 2009) (Kobe, Japan, Oct. 26–30, 2009). International Society for Music Information Retrieval, 711–716. https://ismir2009.ismir.net/proceedings/OS8-1.pdf.

S. Hochreiter and J. Schmidhuber. 1997. "Long Short-Term Memory." *Neural computation*, 9, 8, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

C. Hokamp and Q. Liu. July 2017. "Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by R. Barzilay and M.-Y. Kan. Association for Computational Linguistics, Vancouver, Canada, (July 2017), 1535–1546. doi:10.18653/v1/P17-1141.

J. Hopkins and D. Kiela. July 2017. "Automatically Generating Rhythmic Verse with Neural Networks." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by R. Barzilay and M.-Y. Kan. Association for Computational Linguistics, Vancouver, Canada, (July 2017), 168–178. doi:10.18653/v1/P17-1016.

E. Horishny. 2022. *Romantic-Computing*. (2022). arXiv: 2206.11864 (cs.CL).

J. Hou and S. Zhang. Dec. 2024. "Exploring Thematic Diversity in Classical Chinese Poetry: A Novel Dataset and a BERT-enhanced Ensemble Learning Approach." *ACM Journal on Computing and Cultural Heritage*, 17, 4, Article 60, (Dec. 2024), 19 pages. doi:10.1145/3685679.

J. Hu and M. Sun. May 2020. "Generating Major Types of Chinese Classical Poetry in a Uniformed Framework." In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by N. Calzolari et al. European Language Resources Association, Marseille, France, (May 2020), 4658–4663. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.573.

Z. Hu, C. Liu, Y. Feng, A. T. Luu, and B. Hooi. 2024. "PoetryDiffusion: Towards Joint Semantic and Metrical Manipulation in Poetry Generation." In: *Proceedings of the 38th AAAI Conference on Artificial Intelligence* (AAAI-24) (Vancouver, Canada, Feb. 20–27, 2024). Ed. by M. Wooldridge, J. Dy, and S. Natarajan. Vol. 38. AAAI Press, Washington, DC, USA, 18279–18288. ISBN: 978-1-57735-887-9. doi:10.1609/aaai.v38i16.29787.

C. Huang and X. Shen. Jan. 2025. "PoemBERT: A Dynamic Masking Content and Ratio Based Semantic Language Model For Chinese Poem Generation." In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert. Association for Computational Linguistics, Abu Dhabi, UAE, (Jan. 2025), 50–60. https://aclanthology.org/2025.coling-main.5.

J. D. Hunter. 2007. "Matplotlib: A 2D graphics environment." *Computing in Science & Engineering*, 9, 3, 90–95. doi:10.1109/MCSE.2007.55.

T. M. Huynh and Q. L. Bao. 2024. *Vietnamese Poem Generation & The Prospect Of Cross-Language Poem-To-Poem Translation*. arXiv: 2401.01078 (cs.CL).

M. Ishikawa. 2016. "About Tanka." *Jung Journal*, 10, 1, 32–36. doi:10.1080/19342039.2016.1120610.

M. Ismayilzada, D. Circi, J. Sälevä, H. Sirin, A. Köksal, B. Dhingra, A. Bosselut, D. Ataman, and L. V. D. Plas. Apr. 2025. "Evaluating Morphological Compositional Generalization in Large Language Models." In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by L. Chiruzzo, A. Ritter, and L. Wang. Association for Computational Linguistics, Albuquerque, New Mexico, (Apr. 2025), 1270–1305. ISBN: 979-8-89176-189-6. doi:10.18653/v1/2025.naacl-long.59.

M. Ismayilzada, D. Paul, A. Bosselut, and L. van der Plas. 2024. *Creativity in AI: Progresses and Challenges*. arXiv: 2410.17218 (cs.AI).

R. Jiang, R. Long, C. Gu, and M. Yan. 2025. *VisuCraft: Enhancing Large Vision-Language Models for Complex Visual-Guided Creative Content Generation via Structured Information Extraction*. arXiv: 2508.02890 (cs.CV).

W. Johnson. 1944. “A program of research.” In: *Psychological Monographs*. Studies in language behavior. Vol. 56.2. Ed. by J. F. Dashiell. American Psychological Association (APA), 1–15. https://archive.org/details/sim_psychological-monographs_1944_56_2/page/n3/mode/2up.

A. Jordanous. 2016. “Four PPPPerspectives on computational creativity in theory and in practice.” *Connection Science*, 28, 2, 194–216. doi:10.1080/09540091.2016.1151860.

A. K. Jordanous. Dec. 2012. “Evaluating computational creativity: a standardised procedure for evaluating creative systems and its application.” Ph.D. Dissertation. Department of Informatics, University of Sussex, UK, (Dec. 2012). https://kar.kent.ac.uk/id/eprint/42388. KAR id:42388.

J. Kanerva and F. Ginter. June 2022. “Out-of-Domain Evaluation of Finnish Dependency Parsing.” In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by N. Calzolari et al. European Language Resources Association, Marseille, France, (June 2022), 1114–1124. https://aclanthology.org/2022.lrec-1.120.

A. Kantosalo. Aug. 2019. “Human-computer co-creativity — Designing, evaluating and modelling computational collaborators for poetry writing.” Ph.D. Dissertation. Department of Computer Science, University of Helsinki, Finland, (Aug. 2019). ISBN: 978-951-51-5337-1. https://urn.fi/URN:ISBN:978-951-51-5337-1. Report A-2019-3.

J. Kao and D. Jurafsky. June 2012. “A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry.” In: *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Ed. by D. Elson, A. Kazantseva, R. Mihalcea, and S. Szpakowicz. Association for Computational Linguistics, Montréal, Canada, (June 2012), 8–17. https://aclanthology.org/W12-2502.

P. Karimi, K. Grace, M. L. Maher, and N. Davis. 2018. “Evaluating Creativity in Computational Co-Creative Systems.” In: *Proceedings of the 9th International Conference on Computational Creativity* (ICCC 2018) (Salamanca, Spain, June 25–29, 2018). Ed. by F. Pachet, A. Jordanous, and C. León. Association for Computational Creativity (ACC), 104–111. ISBN: 978-989-54160-0-4. https://computationalcreativity.net/iccc2018/sites/default/files/papers/ICCC_2018_paper_26.pdf.

R. Khanmohammadi, M. S. Mirshafiee, Y. Rezaee Jouryabi, and S. A. Mirroshandel. June 2023. “Prose2Poem: The Blessing of Transformers in Translating Prose to Persian Poetry.” *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22, 6, Article 170, (June 2023), 18 pages. doi:10.1145/3592791.

S. Kim, J. Shin, et al.. 2024. “Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models.” In: *The Twelfth International Conference on Learning Representations* (ICLR 2024) (Vienna, Austria, May 7–11, 2024). https://openreview.net/forum?id=8euJaTveKw.

S. Kim, J. Suk, et al.. Apr. 2025. “The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models.” In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by L. Chiruzzo, A. Ritter, and L. Wang. Association for Computational Linguistics, Albuquerque, New Mexico, (Apr. 2025), 5877–5919. ISBN: 979-8-89176-189-6. doi:10.18653/v1/2025.naacl-long.303.

D. P. Kingma and M. Welling. 2014. “Auto-Encoding Variational Bayes.” In: *Conference Track Proceedings*. 2nd International Conference on Learning Representations. ICLR 2014 (Banff, Canada, Apr. 14–16, 2014). arXiv: 1312.6114 [stat.ML].

J. Kitzlerová. Mar. 2022. “Mayakovsky’s Neologisms: Word-Formation Models, Functions, Afterlife.” *Russian Literature*, 128, (Mar. 2022), 1–30. doi:10.1016/j.ruslit.2021.12.003.

N. Köbis and L. D. Mossink. 2021. “Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry.” *Computers in human behavior*, 114, 106553. doi:10.1016/j.chb.2020.106553.

I. Koziev and A. Fenogenova. May 2025. “Generation of Russian Poetry of Different Genres and Styles Using Neural Networks with Character-Level Tokenization.” In: *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*. Ed. by A. Kazantseva, S. Szpakowicz, S. Degaetano-Ortlieb, Y. Bizzoni, and J. Pagel. Association for Computational Linguistics, Albuquerque, New Mexico, (May 2025), 47–63. ISBN: 979-8-89176-241-1. doi:10.18653/v1/2025.latechclfl-1.6.

M. Kreminski. 2024. “Computational Poetry is Lost Poetry.” In: *Proceedings of the Halfway to the Future Symposium* (HttF ’24) Article 9 (Santa Cruz, CA, USA, Oct. 21–23, 2024). Association for Computing Machinery, New York, NY, USA, 4 pages. ISBN: 979-8-4007-1042-1. doi:10.1145/3686169.3686179.

T. Kudo and J. Richardson. Nov. 2018. “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.” In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by E. Blanco and W. Lu. Association for Computational Linguistics, Brussels, Belgium, (Nov. 2018), 66–71. doi:10.18653/v1/D18-2012.

T. Kuribayashi, Y. Oseki, T. Ito, R. Yoshida, M. Asahara, and K. Inui. Aug. 2021. “Lower Perplexity is Not Always Human-Like.” In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Association for Computational Linguistics, Online, (Aug. 2021), 5203–5217. doi:10.18653/v1/2021.acl-long.405.

L. Kushnareva, D. Cherniavskii, V. Mikhailov, E. Artemova, S. Barannikov, A. Bernstein, I. Piontkovskaya, D. Piontkovski, and E. Burnaev. Nov. 2021. “Artificial Text Detection via Examining the Topology of Attention Maps.” In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, (Nov. 2021), 635–649. doi:10.18653/v1/2021.emnlp-main.50.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. International Conference on Machine Learning.” In: *Proceedings of the Eighteenth International Conference on Machine*

Learning (ICML '01). Ed. by C. E. Brodley. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 282–289. ISBN: 978-1-55860-778-1. https://dl.acm.org/doi/10.5555/645530.655813.

H. Lai and M. Nissim. May 2024. "A Survey on Automatic Generation of Figurative Language: From Rule-based Systems to Large Language Models." *ACM Computing Surveys*, 56, 10, Article 244, (May 2024), 34 pages. doi:10.1145/3654795.

C. Lamb, D. G. Brown, and C. Clarke. 2015. "Human Competence in Creativity Evaluation." In: *Proceedings of the Sixth International Conference on Computational Creativity* (ICCC 2015) (Park City, Utah, USA, June 29–July 2, 2015). Ed. by H. Toivonen, S. Colton, M. Cook, and D. Ventura. Brigham Young University, 102–109. http://computationalcreativity.net/iccc2015/proceedings/5_2Lamb.pdf.

C. Lamb, D. G. Brown, and C. L. A. Clarke. 2017. "A taxonomy of generative poetry techniques." *Journal of Mathematics and the Arts*, 11, 3, 159–179. doi:10.1080/17513472.2017.1373561.

C. Lamb, D. G. Brown, and C. L. A. Clarke. Feb. 2018. "Evaluating Computational Creativity: An Interdisciplinary Tutorial." *ACM Computing Surveys (CSUR)*, 51, 2, Article 28, (Feb. 2018), 34 pages. doi:10.1145/3167476.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." In: *International Conference on Learning Representations*. https://openreview.net/forum?id=H1eA7AEtvS.

Y. Lang and Z. Liangzhi. 2024. "Ci Poetry of the Song Dynasty: Chanting of the Soul." In: *Insights into Chinese Culture*. Springer Nature Singapore, Singapore, 265–285. ISBN: 978-981-97-4511-1. doi:10.1007/978-981-97-4511-1_23.

J. H. Lau, T. Cohn, T. Baldwin, J. Brooke, and A. Hammond. July 2018. "Deep-speare: A joint neural model of poetic language, meter and rhyme." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych and Y. Miyao. Association for Computational Linguistics, Melbourne, Australia, (July 2018), 1948–1958. doi:10.18653/v1/P18-1181.

J. Lee and Y. H. Kong. June 2012. "A Dependency Treebank of Classical Chinese Poems." In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by E. Fosler-Lussier, E. Riloff, and S. Bangalore. Association for Computational Linguistics, Montréal, Canada, (June 2012), 191–199. https://aclanthology.org/N12-1020.

G. R. Lencione, R. F. Nogueira, and P. Y. Pasqualini. 2022. *Nameling: Creative Neologism Generation with Transfer Learning*. Research summary at Doctoral Consortium Program of 13th International Conference on Computational Creativity. https://computationalcreativity.net/iccc22/wp-content/uploads/2022/06/GABRIEL_LENCIONE_DC_research_summary.pdf.

V. I. Levenshtein. 1966. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." *Soviet Physics–Doklady*, 10, 8, 707–710. https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf. Trans. of "Binary codes capable of correcting deletions, insertions, and reversals." *Doklady Akademii Nauk SSSR*, 163, 4, 845–848. https://www.mathnet.ru/eng/dan31411.

R. P. Levy. 2001. "A computational model of poetic creativity with neural network as measure of adaptive fitness." In: *Proceedings of the ICCBR-01 Workshop on Creative Systems* (Simon Fraser University at Harbour Centre, Vancouver, British Columbia, Canada, July 31, 2001). https://web.archive.org/web/20040815051716/http://www.aic.nrl.navy.mil/papers/2001/AIC-01-003/ws4/ws4toc9.zip.

D. Lewis, A. Zugarini, and E. Alonso. 2021. "Syllable Neural Language Models for English Poem Generation." In: *Proceedings of the 12th International Conference on Computational Creativity* (ICCC 2021) (México City, México (Virtual), Sept. 14–18, 2021). Ed. by A. G. de Silva Garza, T. Veale, W. Aguilar, and R. P. y Pérez. Association for Computational Creativity (ACC), 350–356. ISBN: 978-989-54160-3-5. https://computationalcreativity.net/iccc21/wp-content/uploads/2021/09/ICCC_2021_paper_31.pdf.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. July 2020. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Association for Computational Linguistics, Online, (July 2020), 7871–7880. doi:10.18653/v1/2020.acl-main.703.

J. Li, D. Li, S. Savarese, and S. Hoi. July 2023. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." In: *Proceedings of the 40th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Ed. by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett. Vol. 202. PMLR, (July 2023), 19730–19742. https://proceedings.mlr.press/v202/li23q.html.

K. Li, H. Wu, and Y. Dong. 2024. "Copyright protection during the training stage of generative AI: Industry-oriented U.S. law, rights-oriented EU law, and fair remuneration rights for generative AI training under the UN's international governance regime for AI." *Computer Law & Security Review*, 55, 106056. doi:https://doi.org/10.1016/j.clsr.2024.106056.

P. Li, H. Zhang, X. Liu, and S. Shi. July 2020. "Rigid Formats Controlled Text Generation." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Association for Computational Linguistics, Online, (July 2020), 742–751. doi:10.18653/v1/2020.acl-main.68.

W. Li, F. Qi, M. Sun, X. Yi, and J. Zhang. 2021. *CCPM: A Chinese Classical Poetry Matching Dataset*. arXiv: 2106.01979 [cs.CL].

X. Li, F. Metze, D. Mortensen, S. Watanabe, and A. Black. May 2022. "Zero-shot Learning for Grapheme to Phoneme Conversion with Language Ensemble." In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Association for Computational Linguistics, Dublin, Ireland, (May 2022), 2106–2115. doi:10.18653/v1/2022.findings-acl.166.

Y. Li, S. Wang, C. Lin, and F. Guerin. July 2023. "Metaphor Detection via Explicit Basic Meanings Modelling." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 91–100. doi:10.18653/v1/2023.acl-short.9.

Z. Li, X. Xu, T. Shen, C. Xu, J.-C. Gu, Y. Lai, C. Tao, and S. Ma. Nov. 2024. "Leveraging Large Language Models for NLG Evaluation: Advances and Challenges." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 16028–16045. doi:10.18653/v1/2024.emnlp-main.896.

C.-Y. Lin. July 2004. "ROUGE: A Package for Automatic Evaluation of Summaries." In: *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, (July 2004), 74–81. https://aclanthology.org/W04-1013.

Y.-T. Lin and Y.-N. Chen. July 2023. "LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models." In: *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*. Ed. by Y.-N. Chen and A. Rastogi. Association for Computational Linguistics, Toronto, Canada, (July 2023), 47–58. doi:10.18653/v1/2023.nlp4convai-1.5.

B. Liu, J. Fu, M. P. Kato, M. Yoshikawa, B. Liu, J. Fu, M. P. Kato, and M. Yoshikawa. 2018. "Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training." In: *Proceedings of the 26th ACM International Conference on Multimedia* (MM '18) (Seoul, Republic of Korea). Association for Computing Machinery, New York, NY, USA, 783–791. ISBN: 978-1-4503-5665-7. doi:10.1145/3240508.3240587.

C.-L. Liu, T.-Y. Zheng, K.-C. Chen, and M.-H. Chung. Nov. 2022. "Introducing a Large Corpus of Tokenized Classical Chinese Poems of Tang and Song Dynasties." In: *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*. Ed. by M. Hämäläinen, K. Alnajjar, N. Partanen, and J. Rueter. Association for Computational Linguistics, Taipei, Taiwan, (Nov. 2022), 135–144. doi:10.18653/v1/2022.nlp4dh-1.17.

D. Liu, Q. Guo, W. Li, and J. Lv. 2018. "A Multi-Modal Chinese Poetry Generation Model." In: *2018 International Joint Conference on Neural Networks (IJCNN)* (Rio de Janeiro, Brazil, July 8–13, 2018). IEEE, 3438–3445. doi:10.1109/IJCNN.2018.8489579.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. 2023. "Visual Instruction Tuning." In: *Advances in Neural Information Processing Systems (NeurIPS 2023)*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 34892–34916. https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

L. Liu, X. Wan, and Z. Guo. 2018. "Images2Poem: Generating Chinese Poetry from Image Streams." In: *Proceedings of the 26th ACM international conference on Multimedia* (MM '18) (Seoul, Republic of Korea). Association for Computing Machinery, New York, NY, USA, 1967–1975. ISBN: 978-1-4503-5665-7. doi:10.1145/3240508.3241910.

N. Liu, W. Han, G. Liu, D. Peng, R. Zhang, X. Wang, and H. Ruan. May 2022. "ChipSong: A Controllable Lyric Generation System for Chinese Popular Song." In: *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Ed. by T.-H. '. Huang, V. Raheja, D. Kang, J. J. Y. Chung, D. Gissin, M. Lee, and K. I. Gero. Association for Computational Linguistics, Dublin, Ireland, (May 2022), 85–95. doi:10.18653/v1/2022.in2writing-1.13.

Y. Liu, L. Lan, J. Cao, H. Cheng, K. Ding, and L. Jin. Apr. 2025. "Large-Scale Corpus Construction and Retrieval-Augmented Generation for Ancient Chinese Poetry: New Method and Data Insights." In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Ed. by L. Chiruzzo, A. Ritter, and L. Wang. Association for Computational Linguistics, Albuquerque, New Mexico, (Apr. 2025), 779–817. ISBN: 979-8-89176-195-7. doi:10.18653/v1/2025.findings-naacl.46.

Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Nov. 2020. "Multilingual Denoising Pre-training for Neural Machine Translation." *Transactions of the Association for Computational Linguistics*, 8, (Nov. 2020), 726–742. doi:10.1162/tacl_a_00343.

Y. Liu, M. Ott, et al.. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. (2019). arXiv: 1907.11692 [cs.CL].

Y. Liu, D. Liu, J. Lv, and Y. Sang. 2020. "Generating Chinese Poetry from Images via Concrete and Abstract Information." In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8. doi:10.1109/IJCNN48605.2020.9206952.

Z. Liu, Z. Fu, J. Cao, G. de Melo, Y.-C. Tam, C. Niu, and J. Zhou. July 2019. "Rhetorically Controlled Encoder-Decoder for Modern Chinese Poetry Generation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by A. Korhonen, D. Traum, and L. Màrquez. Association for Computational Linguistics, Florence, Italy, (July 2019), 1992–2001. doi:10.18653/v1/P19-1192.

Llama Team. 2024. *The Llama 3 Herd of Models*. (2024). arXiv: 2407.21783 [cs.AI].

K.-L. Lo, R. Ariss, and P. Kurz. 2022. *GPoeT-2: A GPT-2 based poem generator*. arXiv: 2205.08847 [cs.CL].

T. Loakman, C. Tang, and C. Lin. Dec. 2024. "Train and Constrain: Phonologically Informed Tongue Twister Generation from Topics and Paraphrases." *Computational Linguistics*, (Dec. 2024), 1–52. doi:10.1162/coli_a_00544.

M. Loller-Andersen and B. Gambäck. 2018. "Deep Learning-based Poetry Generation Given Visual Input." In: *Proceedings of the 9th International Conference on Computational Creativity* (ICCC 2018) (Salamanca, Spain, June 25–29, 2018). Ed. by F. Pachet, A. Jordanous, and C. León. Association for Computational Creativity (ACC), 240–247. ISBN: 978-989-54160-0-4. https://computationalcreativity.net/iccc2018/sites/default/files/papers/ICCC_2018_paper_59.pdf.

R. Lotman. Dec. 2013. "Sonnet as Closed Form and Open Process." *Interlitteraria*, 18, 2, (Dec. 2013), 317–334. doi:10.12697/il.2013.18.2.03.

A. Louis. 2020. *BelGPT-2: A GPT-2 Model Pre-trained on French Corpora*. https://github.com/ant-louis/belgpt2.

X. Lu, M. Sclar, et al.. 2025. "AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text." In: *The Thirteenth International Conference on Learning Representations* (ICLR 2025) (Singapore, Apr. 24–28, 2025). https://openreview.net/forum?id=ilOEOIqolQ.

X. Lu, J. Wang, B. Zhuang, S. Wang, and J. Xiao. 2019. "A Syllable-Structured, Contextually-Based Conditionally Generation of Chinese Lyrics." In: *Lecture Notes in Computer Science*. Vol. 11672: *PRICAI 2019: Trends in Artificial Intelligence. Proceedings of 16th Pacific Rim*

International Conference on Artificial Intelligence (Cuvu, Yanuca Island, Fiji, Aug. 26–30, 2019). Ed. by A. C. Nayak and A. Sharma. Springer International Publishing, Cham, 257–265. ISBN: 978-3-030-29894-4. doi:10.1007/978-3-030-29894-4_20.

N. Lucchi. 2023. "ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems." *European Journal of Risk Regulation*, 15, 3, 602–624. doi:10.1017/err.2023.59.

W. Luo, F. Song, W. Li, G. Peng, S. Wei, and H. Wang. July 2025. "Odysseus Navigates the Sirens' Song: Dynamic Focus Decoding for Factual and Diverse Open-Ended Text Generation." In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Association for Computational Linguistics, Vienna, Austria, (July 2025), 27200–27218. ISBN: 979-8-89176-251-0. doi:10.18653/v1/2025.acl-long.1320.

D. Lynott and M. T. Keane. 2005. "Familiarity and creativity in novel compound production." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 27, 1367–1372. https://escholarship.org/uc/item/2sg4n336.

J. Ma, R. Zhan, and D. F. Wong. May 2023. "Yu Sheng: Human-in-Loop Classical Chinese Poetry Generation System." In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Ed. by D. Croce and L. Soldaini. Association for Computational Linguistics, Dubrovnik, Croatia, (May 2023), 57–66. doi:10.18653/v1/2023.eacl-demo.8.

S. Ma and Q. Wang. Nov. 2024. "Zero-Shot Detection of LLM-Generated Text using Token Cohesiveness." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 17538–17553. doi:10.18653/v1/2024.emnlp-main.971.

Y. Ma and B. Wang. Nov. 2020. "Description and Quality Assessment of Poetry Translation: Application of a Linguistic Model." *Contrastive Pragmatics*, 3, 1, (Nov. 2020), 89–111. doi:10.1163/26660393-bja10015.

R. Mahbub, I. Khan, S. Anuva, M. S. Shahriar, M. T. R. Laskar, and S. Ahmed. Dec. 2023. "Unveiling the Essence of Poetry: Introducing a Comprehensive Dataset and Benchmark for Poem Summarization." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Association for Computational Linguistics, Singapore, (Dec. 2023), 14878–14886. doi:10.18653/v1/2023.emnlp-main.920.

E. Manjavacas, M. Kestemont, and F. Karsdorp. Oct. 2019. "Generation of Hip-Hop Lyrics with Hierarchical Modeling and Conditional Templates." In: *Proceedings of the 12th International Conference on Natural Language Generation*. Ed. by K. van Deemter, C. Lin, and H. Takamura. Association for Computational Linguistics, Tokyo, Japan, (Oct. 2019), 301–310. doi:10.18653/v1/W19-8638.

H. M. Manurung. 2003. "An evolutionary algorithm approach to poetry generation." Ph.D. Dissertation. University of Edinburgh. College of Science and Engineering. School of Informatics. http://hdl.handle.net/1842/314.

R. Manurung, G. Ritchie, and H. Thompson. 2012. "Using genetic algorithms to create meaningful poetic text." *Journal of Experimental & Theoretical Artificial Intelligence*, 24, 1, 43–64. doi:10.1080/0952813X.2010.539029.

G. Marco, J. de la Rosa, J. Gonzalo, S. Ros, and E. González-Blanco. 2021. "Automated Metric Analysis of Spanish Poetry: Two Complementary Approaches." *IEEE Access*, 9, 51734–51746. doi:10.1109/access.2021.3069635.

M.-C. de Marneffe, C. D. Manning, J. Nivre, and D. Zeman. July 2021. "Universal Dependencies." *Computational Linguistics*, 47, 2, (July 2021), 255–308. doi:10.1162/coli_a_00402.

E. Matusov. Aug. 2019. "The Challenges of Using Neural Machine Translation for Literature." In: *Proceedings of the Qualities of Literary Machine Translation*. Ed. by J. Hadley, M. Popović, H. Afli, and A. Way. European Association for Machine Translation, Dublin, Ireland, (Aug. 2019), 10–19. https://aclanthology.org/W19-7302.

P. M. McCarthy. Aug. 2005. "An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)." Ph.D. Dissertation. University of Memphis, Memphis, TN, USA, (Aug. 2005).

P. M. McCarthy and S. Jarvis. May 2010. "MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment." *Behavior Research Methods*, 42, 2, (May 2010), 381–392. doi:10.3758/BRM.42.2.381.

J. McCormack, E. Wilson, N. Rajcic, and M. T. Llano. 2024. "Mimetic Poet." In: *Proceedings of the 15th International Conference on Computational Creativity* (ICCC 2024) (Jönköping, Sweden, June 17–21, 2024). Ed. by K. Grace, M. T. Llano, P. Martins, and M. M. Hedblom. Association for Computational Creativity (ACC), 279–288. ISBN: 978-989-54160-6-6. https://computationalcreativity.net/iccc24/papers/ICCC24_paper_133.pdf.

C. Meister, T. Pimentel, G. Wiher, and R. Cotterell. 2023. "Locally Typical Sampling." *Transactions of the Association for Computational Linguistics*, 11, 102–121. doi:10.1162/tacl_a_00536.

G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters. 2018. "DALI: A Large Dataset of Synchronized Audio, Lyrics and notes, Automatically Created using Teacher-student Machine Learning Paradigm." In: *Proceedings of the 19th International Society for Music Information Retrieval Conference* (ISMIR 2018) (Paris, France, Sept. 23–27, 2018). ISMIR, 431–437. doi:10.5281/zenodo.1492443.

Microsoft. 2024. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. Tech. rep. Microsoft. arXiv: 2404.14219 (cs.CL).

N. N. Minh, A. Baker, C. Neo, A. G. Roush, A. Kirsch, and R. Shwartz-Ziv. 2025. "Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs." In: *The Thirteenth International Conference on Learning Representations* (ICLR 2025) (Apr. 24–28, 2025). Singapore. https://openreview.net/forum?id=FBkpCyujtS.

O. Moreno. June 2021. “The REPU CS’ Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation.” In: *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Ed. by M. Mager, A. Oncevay, A. Rios, I. V. M. Ruiz, A. Palmer, G. Neubig, and K. Kann. Association for Computational Linguistics, Online, (June 2021), 241–247. doi:10.18653/v1/2021.americasnlp-1.27.

I. Moshkov, D. Hanley, I. Sorokin, S. Toshniwal, C. Henkel, B. Schifferer, W. Du, and I. Gitman. 2025. *AIMO-2 Winning Solution: Building State-of-the-Art Mathematical Reasoning Models with OpenMathReasoning dataset*. (2025). arXiv: 2504.16891 (cs.AI).

Z. Mu, M. Liu, J. Sun, and C. Wang. 2020. “Research on a Tang Poetry automatic generation system based on an evolutionary algorithm.” *Journal of East China Normal University(Natural Science)*, 2020, 6, 129–139. doi:10.3969/j.issn.1000-5641.201921017.

A. Mukherjee, S. Yadav, and M. Shrivastava. Nov. 2024. “CoST of breaking the LLMs.” In: *Proceedings of the Ninth Conference on Machine Translation*. Ed. by B. Haddow, T. Kocmi, P. Koehn, and C. Monz. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 299–306. doi:10.18653/v1/2024.wmt-1.24.

K. Murakami and A. Terai. Dec. 2023. “Lyrics Generation Applying Metaphor Generation.” In: *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*. Ed. by C.-R. Huang et al. Association for Computational Linguistics, Hong Kong, China, (Dec. 2023), 543–551. https://aclanthology.org/2023.paclic-1.54.

D. Nalci, Z. Bilici, and S. Yeşilyurt. 2025. “William ShakesBlake 2.0: Fine-Tuned Language Models for Hybrid Poetic Style Generation.” In: *Proceedings of 5th International Artificial Intelligence and Data Science Congress* (ICADA 2025) (Zonguldak, Turkey, Apr. 24–25, 2025). Ed. by A. Alaybeyoğlu and S. Çakir. Izmir Katip Celebi University Press, 757–768. ISBN: 978-625-95496-4-4. https://drive.google.com/file/d/1Ic69np3acdOSk1Mdh2c1dX_mKXFEgaBv/view?usp=sharing.

B. Navarro, M. Ribes Lafoz, and N. Sánchez. May 2016. “Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation.” In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC‘16)*. Ed. by N. Calzolari et al. European Language Resources Association (ELRA), Portorož, Slovenia, (May 2016), 4360–4364. https://aclanthology.org/L16-1691.

T. Nguyen, P. Nguyen, H. Pham, T. Bui, T. Nguyen, and D. Luong. 2021. “SP-GPT2: semantics improvement in Vietnamese poetry generation.” In: *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1576–1581. doi:10.1109/icmla52953.2021.00252.

S. Nie et al.. 2025. “Large Language Diffusion Models.” In: *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy* (Singapore, Apr. 28, 2025). https://openreview.net/forum?id=wzl61tIUj6.

N. I. Nikolov, E. Malmi, C. Northcutt, and L. Parisi. Dec. 2020. “Rapformer: Conditional Rap Lyrics Generation with Denoising Autoencoders.” In: *Proceedings of the 13th International Conference on Natural Language Generation*. Ed. by B. Davis, Y. Graham, J. Kelleher, and Y. Sripada. Association for Computational Linguistics, Dublin, Ireland, (Dec. 2020), 360–373. doi:10.18653/v1/2020.inlg-1.42.

I. Olatunji. 2023. “Why try to build try to build a co-creative poetry system that makes people feel that they have “creative superpowers”?” In: *Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE co-located with the ACM International Conference on Intelligent User Interfaces* (IUI 2023) (Sydney, Australia, Mar. 27–31, 2023). Ed. by A. Smith-Renner and P. Taele. https://ceur-ws.org/Vol-3359/paper8.pdf.

M. Oliver. 1994. *A Poetry Handbook*. A Harvest Original Harcourt. 130 pp. ISBN: 978-0-15-672400-5.

M. L. Olson, N. Ratzlaff, M. Hinck, S.-y. Tseng, and V. Lal. Dec. 2024. “Steering Large Language Models to Evaluate and Amplify Creativity.” In: *NeurIPS 2024 Workshop on Creativity & Generative AI (Spotlight)*. (Dec. 2024). https://creativity-ai.github.io/assets/papers/67.pdf.

OpenAI. 2022. *ChatGPT: Optimizing Language Models for Dialogue*. (2022). Retrieved Sept. 5, 2025 from https://openai.com/blog/chatgpt/.

OpenAI. 2023. *GPT-4 Technical Report*. (2023). arXiv: 2303.08774 (cs.CL).

OpenAI. 2024. *GPT-4o System Card*. (2024). arXiv: 2410.21276 (cs.CL).

A. Ormazabal, M. Artetxe, M. Agirrezabal, A. Soroa, and E. Agirre. Dec. 2022. “PoeLM: A Meter- and Rhyme-Controllable Language Model for Unsupervised Poetry Generation.” In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Y. Goldberg, Z. Kozareva, and Y. Zhang. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, (Dec. 2022), 3655–3670. doi:10.18653/v1/2022.findings-emnlp.268.

L. Ou, X. Ma, M.-Y. Kan, and Y. Wang. July 2023. “Songs Across Borders: Singable and Controllable Neural Lyric Translation.” In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 447–467. doi:10.18653/v1/2023.acl-long.27.

V. Padmakumar and H. He. 2024. “Does Writing with Language Models Reduce Content Diversity?” In: *The Twelfth International Conference on Learning Representations* (ICLR 2024) (Vienna, Austria, May 7–11, 2024). https://openreview.net/forum?id=Feiz5HtCD0.

A. Pagnoni et al.. July 2025. “Byte Latent Transformer: Patches Scale Better Than Tokens.” In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Association for Computational Linguistics, Vienna, Austria, (July 2025), 9238–9258. ISBN: 979-8-89176-251-0. doi:10.18653/v1/2025.acl-long.453.

A. Panahandeh, H. Asemi, and E. Nourani. 2023. *TPPoet: Transformer-Based Persian Poem Generation using Minimal Data and Advanced Decoding Techniques*. arXiv: 2312.02125 (cs.CL).

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. July 2002. “Bleu: a Method for Automatic Evaluation of Machine Translation.” In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by P. Isabelle, E. Charniak, and D. Lin. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, (July 2002), 311–318. doi:10.3115/1073083.1073135.

R. Pardinas, G. Huang, D. Vazquez, and A. Piché. 2023. "Leveraging Human Preferences to Master Poetry." In: *The AAAI-23 Workshop on Creative AI Across Modalities* (Washington DC, USA). https://openreview.net/forum?id=9lmAR2NjTt.

A. Parrish. 2016. *Project Gutenberg Poetry Corpus*. Retrieved Mar. 4, 2025 from https://github.com/aparrish/gutenberg-poetry-corpus.

A. R. Pascual. 2021. *BACON: Deep-Learning Powered AI for Poetry Generation with Author Linguistic Style Transfer*. (2021). arXiv: 2112.11483 (cs.CL).

T. Pasini, A. López-Ávila, H. Quteineh, G. Lampouras, J. Du, Y. Wang, Z. Li, and Y. Sun. 2024. *Encoder-Decoder Framework for Interactive Free Verses with Generation with Controllable High-Quality Rhyming*. arXiv: 2405.05176 (cs.CL).

J. Pennington, R. Socher, and C. Manning. Oct. 2014. "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by A. Moschitti, B. Pang, and W. Daelemans. Association for Computational Linguistics, Doha, Qatar, (Oct. 2014), 1532–1543. doi:10.3115/v1/D14-1162.

J. Peskin and B. Ellenbogen. 2019. "Cognitive Processes While Writing Poetry: An Expert-Novice Study." *Cognition and Instruction*, 37, 2, 232–251. doi:10.1080/07370008.2019.1570931.

K. Pillutla, S. Swayamdipta, R. Zellers, J. Thickstun, S. Welleck, Y. Choi, and Z. Harchaoui. 2021. "MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers." In: *Advances in Neural Information Processing Systems (NeurIPS 2021)*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 4816–4828. https://proceedings.neurips.cc/paper_files/paper/2021/file/260c2432a0eecc28ce03c10dadc078a4-Paper.pdf.

P. Plecháč et al.. Apr. 2024. *PoeTree. Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian and Spanish*. Zenodo, (Apr. 2024). https://doi.org/10.5281/zenodo.10008458.

A. Popescu-Belis, À. Atrio, V. Minder, A. Xanthos, G. Luthier, S. Mattei, and A. Rodriguez. June 2022. "Constrained Language Models for Interactive Poem Generation." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by N. Calzolari et al. European Language Resources Association, Marseille, France, (June 2022), 3519–3529. https://aclanthology.org/2022.lrec-1.377.

B. Porter and E. Machery. Nov. 2024. "AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably." *Scientific Reports*, 14, 1, (Nov. 2024), 26133. doi:10.1038/s41598-024-76900-1.

P. Pynadath and R. Zhang. 2025. "Controlled LLM Decoding via Discrete Auto-regressive Biasing." In: *International Conference on Representation Learning*. Ed. by Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu. Vol. 2025, 75911–75932. https://proceedings.iclr.cc/paper_files/paper/2025/hash/bce52456a36be2be1abd95427139de37-Abstract-Conference.html.

T. Qian, F. Lou, J. Shi, Y. Wu, S. Guo, X. Yin, and Q. Jin. July 2023. "UniLG: A Unified Structure-aware Framework for Lyrics Generation." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 983–1001. doi:10.18653/v1/2023.acl-long.56.

M. Qorib, G. Moon, and H. T. Ng. Aug. 2024. "Are Decoder-Only Language Models Better than Encoder-Only Language Models in Understanding Word Meaning?" In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Association for Computational Linguistics, Bangkok, Thailand, (Aug. 2024), 16339–16347. doi:10.18653/v1/2024.findings-acl.967.

Qwen Team. 2025. *Qwen3 Technical Report*. Tech. rep. arXiv: 2505.09388 (cs.CL).

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." *OpenAI blog*. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research*, 21, 140, 1–67. https://www.jmlr.org/papers/v21/20-074.html.

H. Rahmeh. Sept. 2023. "Digital Verses Versus Inked Poetry: Exploring Readers' Response to AI-Generated and Human-Authored Sonnets." *Scholars International Journal of Linguistics and Literature*, 6, 09, (Sept. 2023), 372–382. doi:10.36348/sijll.2023.v06i09.002.

N. Ram, T. Gummadi, R. Bhethanabotla, R. J. Savery, and G. Weinberg. 2021. "Say what? Collaborative pop lyric generation using multitask transfer learning." In: *Proceedings of the 9th International Conference on Human-Agent Interaction*, 165–173. doi:10.1145/3472307.3484175.

F. Rashel and R. Manurung. June 2014. "Pemuisi: A constraint satisfaction-based generator of topical Indonesian poetry." In: *Proceedings of the 5th International Conference on Computational Creativity* (ICCC 2014). Ed. by S. Colton, D. Ventura, N. Lavrac, and M. Cook. Jozef Stefan Institute, Ljubljana, Slovenia, (June 2014), 82–90. https://computationalcreativity.net/iccc2014/wp-content/uploads/2014/06/6.4_Rashel.pdf.

S. Reddy and K. Knight. June 2011. "Unsupervised Discovery of Rhyme Schemes." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Ed. by D. Lin, Y. Matsumoto, and R. Mihalcea. Association for Computational Linguistics, Portland, Oregon, USA, (June 2011), 77–82. https://aclanthology.org/P11-2014.

R. Rei, C. Stewart, A. C. Farinha, and A. Lavie. Nov. 2020. "COMET: A Neural Framework for MT Evaluation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Association for Computational Linguistics, Online, (Nov. 2020), 2685–2702. doi:10.18653/v1/2020.emnlp-main.213.

E. Reiter. Sept. 2018. "A Structured Review of the Validity of BLEU." *Computational Linguistics*, 44, 3, (Sept. 2018), 393–401. doi:10.1162/coli_a_00322.

C. Ren, Z. Guo, P. Zhang, and Y. Gao. 2024. "Humor detection using deep learning in 10 years: A survey." *Métodos numéricos para cálculo y diseño en ingeniería: Revista internacional*, 40, 1, 1–13. doi:10.23967/j.rimni.2024.01.006.

A. Repar, M. Martinc, M. Znidarsic, and S. Pollak. 2018. "BISLON: BISociative SLOgaN generation based on stylistic literary devices." In: *Proceedings of the 9th International Conference on Computational Creativity* (ICCC 2018) (Salamanca, Spain, June 25–29, 2018). Ed. by F. Pachet, A. Jordanous, and C. León. Association for Computational Creativity (ACC), 248–255. ISBN: 978-989-54160-0-4. https://computationalcreativity.net/iccc2018/sites/default/files/papers/ICCC_2018_paper_60.pdf.

N. Resende and J. Hadley. Sept. 2024. "The Translator's Canvas: Using LLMs to Enhance Poetry Translation." In: *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*. Ed. by R. Knowles, A. Eriguchi, and S. Goel. Association for Machine Translation in the Americas, Chicago, USA, (Sept. 2024), 178–189. https://aclanthology.org/2024.amta-research.16.

M. Reza, J. Thomas-Mitchell, P. Dushniku, N. Laundry, J. J. Williams, and A. Kuzminykh. Oct. 2025. "Co-Writing with AI, on Human Terms: Aligning Research with User Demands Across the Writing Process." *Proceedings of the ACM on Human-Computer Interaction*, 9, 7, Article CSCW385, (Oct. 2025), 37 pages. doi:10.1145/3757566.

M. Rhodes. 1961. "An Analysis of Creativity." *The Phi Delta Kappan*, 42, 7, 305–310. http://www.jstor.org/stable/20342603.

M. Riedl. 2020. *Weird AI Yankovic: generating parody lyrics*. arXiv: 2009.12240 [cs.CL].

P. Robinson. 2010. *Poetry and Translation: The Art of the Impossible. The Art of the Impossible*. Liverpool University Press. ISBN: 978-1-80085-970-8.

J. de la Rosa, Á. Pérez Pozo, S. Ros, and E. González-Blanco García. Mar. 2023. "ALBERTI, a Multilingual Domain Specific Language Model for Poetry Analysis." *Procesamiento del Lenguaje Natural*, 71, (Mar. 2023), 215–225. doi:10.26342/2023-71-17.

R. Rosa, D. Mareček, T. Musil, M. Chudoba, and J. Landsperský. May 2025. "EduPo: Progress and Challenges of Automated Analysis and Generation of Czech Poetry." In: *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. Ed. by M. Hämäläinen, E. Öhman, Y. Bizzoni, S. Miyagawa, and K. Alnajjar. Association for Computational Linguistics, Albuquerque, USA, (May 2025), 524–542. ISBN: 979-8-89176-234-3. doi:10.18653/v1/2025.nlp4dh-1.45.

K. A. Røstvold and B. Gambäck. Dec. 2020. "Sentimental Poetry Generation." In: *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*. Ed. by P. Bhattacharyya, D. M. Sharma, and R. Sangal. NLP Association of India (NLPAI), Indian Institute of Technology Patna, Patna, India, (Dec. 2020), 246–256. https://aclanthology.org/2020.icon-main.33.

A. Roush, S. Basu, A. Moorthy, and D. Dubovoy. Oct. 2022. "Most Language Models can be Poets too: An AI Writing Assistant and Constrained Text Generation Studio." In: *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*. Ed. by X. Wu, P. Ruan, S. Li, and Y. Dong. Association for Computational Linguistics, Gyeongju, Republic of Korea, (Oct. 2022), 9–15. https://aclanthology.org/2022.cai-1.2.

P. Ruiz Fabo, C. Martínez Cantón, T. Poibeau, and E. González-Blanco. Aug. 2017. "Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets." In: *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by B. Alex, S. Degaetano-Ortlieb, A. Feldman, A. Kazantseva, N. Reiter, and S. Szpakowicz. Association for Computational Linguistics, Vancouver, Canada, (Aug. 2017), 27–32. doi:10.18653/v1/W17-2204.

A. Saeed, S. Ilić, and E. Zangerle. Dec. 2019. "Creative GANs for generating poems, lyrics, and metaphors." In: *Machine Learning for Creativity and Design NeurIPS 2019 Workshop*. Vancouver, Canada, (Dec. 2019). https://neurips2019creativity.github.io/doc/creative_gans.pdf.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2020. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *The 5th EMC2 - Energy Efficient Machine Learning and Cognitive Computing Workshop Co-located with the 33rd Conference on Neural Information Processing Systems NeurIPS 2019*. https://www.emc2-ai.org/assets/docs/neurips-19/emc2-neurips19-paper-33.pdf.

P. Sawicki, M. Grzes, D. G. Brown, and F. Goes. Nov. 2025. "Can Large Language Models Outperform Non-Experts in Poetry Evaluation? A Comparative Study Using the Consensual Assessment Technique." In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng. Association for Computational Linguistics, Suzhou, China, (Nov. 2025), 31901–31918. ISBN: 979-8-89176-332-6. doi:10.18653/v1/2025.emnlp-main.1625.

P. Sawicki, M. Grzes, F. Goes, D. G. Brown, M. Peeperkorn, and A. Khatun. June 2023. "Bits of Grass: Does GPT already know how to write like Whitman?" In: *Proceedings of the 14th International Conference on Computational Creativity* (ICCC 2023). Association for Computational Creativity (ACC), Ontario, Canada, (June 2023). https://computationalcreativity.net/iccc23/papers/ICCC-2023_paper_95.pdf.

P. Sawicki, M. Grzes, F. Goes, D. G. Brown, M. Peeperkorn, A. Khatun, and S. Paraskevopoulou. June 2023. "On the power of special-purpose GPT models to create and evaluate new poetry in old styles." In: *Proceedings of the 14th International Conference on Computational Creativity* (ICCC 2023). Association for Computational Creativity (ACC), Ontario, Canada, (June 2023), 10–19. https://computationalcreativity.net/iccc23/papers/ICCC-2023_paper_18.pdf.

P. Sawicki, M. Grzes, A. Jordanous, D. Brown, and M. Peeperkorn. 2022. "Training GPT-2 to represent two Romantic-era authors: challenges, evaluations and pitfalls." In: *Proceedings of the 13th International Conference on Computational Creativity* (ICCC 2022) (Bozen-Bolzano, Italy, June 27–July 1, 2022). Ed. by M. M. Hedblom, A. A. Kantosalo, R. Confalonieri, O. Kutz, and T. Veale. Association for Computational Creativity (ACC), 34–43. ISBN: 978-989-54160-4-2. https://computationalcreativity.net/iccc22/wp-content/uploads/2022/06/ICCC-2022_18L_Sawicki-et-al..pdf.

J. Secha, J. Cizhen, J. Cairang, and C. Huaguo. Oct. 2022. "Automatic Generation of Tibetan Poems based on Pre-training and Control Code Method." In: *Proceedings of the 21st Chinese National Conference on Computational Linguistics*. Ed. by M. Sun, Y. Liu, W. Che, Y. Feng, X. Qiu, G. Rao, and Y. Chen. in Chinese. Chinese Information Processing Society of China, Nanchang, China, (Oct. 2022), 366–373. https://aclanthology.org/2022.ccl-1.33.

R. Sennrich, B. Haddow, and A. Birch. Aug. 2016. "Neural Machine Translation of Rare Words with Subword Units." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by K. Erk and N. A. Smith. Association for Computational Linguistics, Berlin, Germany, (Aug. 2016), 1715–1725. doi:10.18653/v1/P16-1162.

M. Shabani Minaabad. 2020. "The Effect of Poetry Therapy on the Development of Language and Social Skills in Children with ASD." *Health Education and Health Promotion*, 8, 2, 79–86. http://hehp.modares.ac.ir/article-5-42048-en.html.

S. Shahriar, N. Al Roken, and I. Zualkernan. Apr. 2023. "Classification of Arabic Poetry Emotions Using Deep Learning." *Computers*, 12, 5, (Apr. 2023), 89. doi:10.3390/computers12050089.

E. Shalevska. Sept. 2024. "The digital laureate: Examining AI-generated poetry." *RATE Issues*, 31, 1, (Sept. 2024). doi:10.69475/RATEI.2024.1.3.

Y. Shao, T. Shao, M. Wang, P. Wang, and J. Gao. 2021. "A Sentiment and Style Controllable Approach for Chinese Poetry Generation." In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (CIKM '21). Association for Computing Machinery, Virtual Event, Queensland, Australia, 4784–4788. ISBN: 9781450384469. doi:10.1145/3459637.3481964.

L. Shen, X. Guo, and M. Chen. 2020. "Compose Like Humans: Jointly Improving the Coherence and Novelty for Modern Chinese Poetry Generation." In: *2020 international joint conference on neural networks (IJCNN)*. IEEE, 1–8. doi:10.1109/IJCNN48605.2020.9206888.

L.-H. Shen, P.-L. Tai, C.-C. Wu, and S.-D. Lin. Nov. 2019. "Controlling Sequence-to-Sequence Models - A Demonstration on Neural-based Acrostic Generator." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Ed. by S. Padó and R. Huang. Association for Computational Linguistics, Hong Kong, China, (Nov. 2019), 43–48. doi:10.18653/v1/D19-3008.

E. Sheng and D. Uthus. Dec. 2020. "Investigating Societal Biases in a Poetry Composition System." In: *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*. Ed. by M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster. Association for Computational Linguistics, Barcelona, Spain (Online), (Dec. 2020), 93–106. https://aclanthology.org/2020.gebnlp-1.9.

Z. Sheng, K. Song, X. Tan, Y. Ren, W. Ye, S. Zhang, and T. Qin. May 2021. "SongMASS: Automatic Song Writing with Pre-training and Alignment Constraint." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 15. Vol. 35. (May 2021), 13798–13805. doi:10.1609/aaai.v35i15.17626.

L. Shi, C. Ma, W. Liang, X. Diao, W. Ma, and S. Vosoughi. Dec. 2025. "Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge." In: *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Ed. by K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, and D. P. Singh. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Mumbai, India, (Dec. 2025), 292–314. ISBN: 979-8-89176-298-5. https://aclanthology.org/2025.ijcnlp-long.18.

J. Shihadeh and M. Ackerman. 2020. "EMILY: An Emily Dickinson Machine." In: *Proceedings of the 11th International Conference on Computational Creativity* (ICCC 2020) (Coimbra, Portugal, Sept. 7–11, 2020). Ed. by F. A. Cardoso, P. Machado, T. Veale, and J. M. Cunha. Association for Computational Creativity (ACC), 243–246. https://computationalcreativity.net/iccc20/papers/134-iccc20.pdf.

K. Simonyan and A. Zisserman. 2015. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *Conference Track Proceedings*. 3rd International Conference on Learning Representations. ICLR 2015 (San Diego, CA, USA, May 7–9, 2015). arXiv: 1409.1556 (cs.CV).

W. L. Song, H. Xu, D. F. Wong, R. Zhan, L. S. Chao, and S. Wang. Sept. 2023. "Towards Zero-Shot Multilingual Poetry Translation." In: *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*. Ed. by M. Utiyama and R. Wang. Asia-Pacific Association for Machine Translation, Macau SAR, China, (Sept. 2023), 324–335. https://aclanthology.org/2023.mtsummit-research.27.

X. Song, C. Song, H. Yu, Y. Zhu, and H. Yao. Mar. 2025. "MixSong: Diverse and Strictly Formatted Chinese Poetry Generation." *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24, 4, Article 37, (Mar. 2025), 16 pages. doi:10.1145/3718331.

Y. Song. Oct. 2022a. "Chinese Couplet Generation with Syntactic Information." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by N. Calzolari et al. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, (Oct. 2022), 6436–6446. https://aclanthology.org/2022.coling-1.560.

Y. Song. Dec. 2022b. "Composing Ci with Reinforced Non-autoregressive Text Generation." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Goldberg, Z. Kozareva, and Y. Zhang. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, (Dec. 2022), 7219–7229. doi:10.18653/v1/2022.emnlp-main.486.

V.-W. Soo, T.-Y. Lai, K.-J. Wu, and Y.-P. Hsu. 2015. "Generate modern style Chinese poems based on common sense and evolutionary computation." In: *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE. IEEE, 315–322. doi:10.1109/taai.2015.7407055.

A. C. Spearing. 1976. *Medieval dream-poetry*. Cambridge University Press. ISBN: 978-0-521-29069-2. Retrieved July 8, 2025 from https://archive.org/details/medievaldreampoe0000spea.

P. S. Sreeja and G. S. Mahalakshmi. 2019. "PERC-An Emotion Recognition Corpus for Cognitive Poems." In: *2019 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 0200–0207. doi:10.1109/ICCSP.2019.8698020.

L. Sun, F. Luisier, K. Batmanghelich, D. Florencio, and C. Zhang. July 2023. "From Characters to Words: Hierarchical Pre-trained Language Model for Open-vocabulary Language Understanding." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 3605–3620. doi:10.18653/v1/2023.acl-long.200.

L. Sun, Y. Yuan, Y. Yao, Y. Li, H. Zhang, X. Xie, X. Wang, F. Luo, and D. Stillwell. 2025. "Large Language Models Show Both Individual and Collective Creativity Comparable to Humans." *Thinking Skills and Creativity*, 101870. doi:10.1016/j.tsc.2025.101870.

Y. Sun, L. Li, Q. Liu, and D.-Y. Yeung. July 2023. "SongRewriter: A Chinese Song Rewriting System with Controllable Content and Rhyme Scheme." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 12863–12880. doi:10.18653/v1/2023.findings-acl.814.

M. N. Sundararaman, A. Kumar, and J. Vepa. Aug. 2021. "PhonemeBERT: Joint Language Modelling of Phoneme Sequence and ASR Transcript." In: *Proceedings of 22nd Annual Conference of the International Speech Communication Association* (Interspeech 2021) (Brno, Czechia, Aug. 30–Sept. 3, 2021). ISCA, (Aug. 2021), 3236–3240. doi:10.21437/interspeech.2021-1582.

A. Suvarna, H. Khandelwal, and N. Peng. Aug. 2024. "PhonologyBench: Evaluating Phonological Skills of Large Language Models." In: *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*. Ed. by S. Li, M. Li, M. J. Q. Zhang, E. Choi, M. Geva, P. Hase, and H. Ji. Association for Computational Linguistics, Bangkok, Thailand, (Aug. 2024), 1–14. doi:10.18653/v1/2024.knowllm-1.1.

Y. Takeishi, M. Niu, J. Luo, Z. Jin, and X. Yang. Apr. 2022. "WakaVT: A Sequential Variational Transformer for Waka Generation." *Neural Processing Letters*, 54, 2, (Apr. 2022), 731–750. doi:10.1007/s11063-021-10654-z.

Y. Tay et al.. 2022. "Charformer: Fast Character Transformers via Gradient-based Subword Tokenization." In: *International Conference on Learning Representations* (ICLR 2022) (Apr. 25–29, 2022). https://openreview.net/forum?id=JtBRnrlOEFN.

A. Terai, K. Yamashita, and S. Komagamine. 2020. "Computer Humor and Human Humor: Construction of Japanese "Nazokake" Riddle Generation Systems." *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 24, 2, 199–205. doi:10.20965/jaciii.2020.p0199.

L. The Nguyen, T. Pham, and D. Q. Nguyen. Aug. 2023. "XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech." In: *Interspeech 2023*. ISCA, (Aug. 2023), 5506–5510. doi:10.21437/interspeech.2023-444.

P. Thölke, A. Bellemare-Pépin, Y. Harel, F. Lespinasse, and K. Jerbi. 2024. "Bio-Mechanical Poet: An Immersive Audiovisual Playground for Brain Signals and Generative AI." In: *Proceedings of the 15th International Conference on Computational Creativity* (ICCC 2024) (Jönköping, Sweden, June 17–21, 2024). Ed. by K. Grace, M. T. Llano, P. Martins, and M. M. Hedblom. Association for Computational Creativity (ACC), 65–74. ISBN: 978-989-54160-6-6. https://computationalcreativity.net/iccc24/papers/ICCC24_paper_130.pdf.

Y. Tian, A. Narayan-Chen, et al.. July 2023. "Unsupervised Melody-to-Lyrics Generation." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 9235–9254. doi:10.18653/v1/2023.acl-long.513.

Y. Tian and N. Peng. July 2022. "Zero-shot Sonnet Generation with Discourse-level Planning and Aesthetics Features." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz. Association for Computational Linguistics, Seattle, United States, (July 2022), 3587–3597. doi:10.18653/v1/2022.naacl-main.262.

A. Tikhonov and I. P. Yamshchikov. Oct. 2018a. "Sounds Wilde. Phonetically Extended Embeddings for Author-Stylized Poetry Generation." In: *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Ed. by S. Kuebler and G. Nicolai. Association for Computational Linguistics, Brussels, Belgium, (Oct. 2018), 117–124. doi:10.18653/v1/w18-5813.

A. Tikhonov and I. P. Yamshchikov. 2018b. "Guess who? Multilingual approach for the automated generation of author-stylized poetry." In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 787–794. doi:10.1109/slt.2018.8639573.

S. Tomizawa, S. Yokoyama, T. Yamashita, and H. Kawamura. 2025. "Proposal of a Haiku Evaluation Method Using Large Language Model and Prompt Engineering." *IIAI Letters on Business and Decision Science*, 5, 1–15. doi:10.52731/lbds.v005.346.

J. Tonra, B. Davis, D. Kelly, and W. Khawaja. 2019. "Poetry In Motion: Quantified Self Data And Automated Poetry Generation." In: Digital Humanities Conference 2019. DH2019 (Utrecht, Netherlands, July 9–12, 2019). DataverseNL. doi:10.34894/Z41ZJA.

H. Touvron, T. Lavril, et al.. 2023. *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 (cs.CL).

H. Touvron, L. Martin, et al.. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. (2023). arXiv: 2307.09288 (cs.CL).

Transformer Circuits Thread. 2025. *Attribution Graphs in Biological Systems*. Retrieved Mar. 31, 2025 from https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-poems.

L. Tu, S. Yavuz, J. Qu, J. Xu, R. Meng, C. Xiong, and Y. Zhou. Nov. 2024. "Unlocking Anticipatory Text Generation: A Constrained Approach for Large Language Models Decoding." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 15532–15548. doi:10.18653/v1/2024.emnlp-main.870.

E. L. Ungless, N. Vitsakis, Z. Talat, J. Garforth, B. Ross, A. Onken, A. Kasirzadeh, and A. Birch. Jan. 2025. "The Only Way is Ethics: A Guide to Ethical Research with Large Language Models." In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert. Association for Computational Linguistics, Abu Dhabi, UAE, (Jan. 2025), 8992–9005. https://aclanthology.org/2025.coling-main.603.

L. Urdang. 1993. *The Oxford thesaurus: an A-Z dictionary of synonyms*. Clarendon Press, Oxford, UK. ISBN: 978-0-19-195801-4.

D. Uthus, M. Voitovich, and R. J. Mical. July 2022. "Augmenting Poetry Composition with Verse by Verse." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*. Ed. by

A. Loukina, R. Gangadharaiah, and B. Min. Association for Computational Linguistics, Hybrid: Seattle, Washington + Online, (July 2022), 18–26. doi:10.18653/v1/2022.naacl-industry.3.

T. Van de Cruys. July 2020. "Automatic Poetry Generation from Prosaic Text." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault. Association for Computational Linguistics, Online, (July 2020), 2471–2480. doi:10.18653/v1/2020.acl-main.223.

T. Van de Cruys. July 2019. "La génération automatique de poésie en français (Automatic Poetry Generation in French)." In: *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume I : Articles longs*. Ed. by E. Morin, S. Rosset, and P. Zweigenbaum. ATALA, Toulouse, France, (July 2019), 113–126. https://aclanthology.org/2019.jeptalnrecital-long.8.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. 31st Annual Conference on Neural Information Processing Systems. NIPS 2017 (Long Beach, CA, USA, Dec. 4–9, 2017). Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett. Vol. 30, 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

O. Vechtomova, G. Sahu, and D. Kumar. 2021. "LyricJam: A system for generating lyrics for live instrumental music." In: *Proceedings of the 12th International Conference on Computational Creativity* (ICCC 2021) (México City, México (Virtual), Sept. 14–18, 2021). Ed. by A. G. de Silva Garza, T. Veale, W. Aguilar, and R. P. y Pérez. Association for Computational Creativity (ACC), 122–130. ISBN: 978-989-54160-3-5. https://computationalcreativity.net/iccc21/wp-content/uploads/2021/09/ICCC_2021_paper_59.pdf.

A. Vijayakumar, M. Cogswell, R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. 2018. "Diverse Beam Search for Improved Description of Complex Scenes." In: *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI-18). Vol. 32. Association for the Advancement of Artificial Intelligence (AAAI), 7371–7379. doi:10.1609/aaai.v32i1.12340.

D. Vrandečić and M. Krötzsch. Sept. 2014. "Wikidata: a free collaborative knowledgebase." *Communications of the ACM*, 57, 10, (Sept. 2014), 78–85. doi:10.1145/2629489.

M. Walsh, M. Antoniak, and A. Preus. Nov. 2024. "Sonnet or Not, Bot? Poetry Evaluation for Large Models and Datasets." In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Association for Computational Linguistics, Miami, Florida, USA, (Nov. 2024), 15568–15603. doi:10.18653/v1/2024.findings-emnlp.914.

M. Walsh, A. Preus, and E. Gronski. Dec. 2024. "Does ChatGPT Have a Poetic Style?" In: *Proceedings of the Computational Humanities Research Conference 2024 (CHR 2024)*. Ed. by W. Haverals, M. Koolen, and L. Thompson. Vol. 3834. Aarhus, Denmark, (Dec. 2024), 1201–1219. https://2024.computational-humanities-research.org/papers/paper122.

J. Wang, X. Zhang, Y. Zhou, C. Suh, and C. Rudin. 2021. "There once was a really bad poet, it was automated but you didn't know it." *Transactions of the Association for Computational Linguistics*, 9, 605–620. Ed. by B. Roark and A. Nenkova. doi:10.1162/tacl_a_00387.

L. Wang, W. Xu, Y. Lan, Z. Hu, Y. Lan, R. K.-W. Lee, and E.-P. Lim. July 2023. "Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 2609–2634. doi:10.18653/v1/2023.acl-long.147.

P. Wang, L. Li, K. Ren, B. Jiang, D. Zhang, and X. Qiu. Dec. 2023. "SeqXGPT: Sentence-Level AI-Generated Text Detection." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Association for Computational Linguistics, Singapore, (Dec. 2023), 1144–1156. doi:10.18653/v1/2023.emnlp-main.73.

P. Wang, S. Zhang, Z. Li, and J. Hou. July 2023. "Enhancing Ancient Chinese Understanding with Derived Noisy Syntax Trees." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. Ed. by V. Padmakumar, G. Vallejo, and Y. Fu. Association for Computational Linguistics, Toronto, Canada, (July 2023), 83–92. doi:10.18653/v1/2023.acl-srw.15.

Q. Wang, T. Luo, and D. Wang. 2016. "Can Machine Generate Traditional Chinese Poetry? A Feigenbaum Test." In: *Lecture Notes in Computer Science*. Vol. 10023: *Advances in Brain Inspired Cognitive Systems* (BICS 2016) (Beijing, China, Nov. 28–30, 2016). Ed. by C.-L. Liu, A. Hussain, B. Luo, K. C. Tan, Y. Zeng, and Z. Zhang. Springer International Publishing, Cham, 34–46. ISBN: 978-3-319-49685-6. doi:10.1007/978-3-319-49685-6_4.

S. Wang, D. Wong, J. Yao, and L. Chao. Aug. 2024. "What is the Best Way for ChatGPT to Translate Poetry?" In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Association for Computational Linguistics, Bangkok, Thailand, (Aug. 2024), 14025–14043. doi:10.18653/v1/2024.acl-long.756.

S. Wang, J. Wu, F. Ye, D. F. Wong, J. Yao, and L. S. Chao. Nov. 2025. "Benchmarking the Detection of LLMs-Generated Modern Chinese Poetry." In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng. Association for Computational Linguistics, Suzhou, China, (Nov. 2025), 9533–9552. ISBN: 979-8-89176-335-7. doi:10.18653/v1/2025.findings-emnlp.507.

W. Wang, M. Gao, X. Hu, and X. Wan. July 2025. "Towards A "Novel" Benchmark: Evaluating Literary Fiction with Large Language Models." In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Association for Computational Linguistics, Vienna, Austria, (July 2025), 21648–21673. ISBN: 979-8-89176-256-5. doi:10.18653/v1/2025.findings-acl.1114.

Y. Wang, J. Deng, A. Sun, and X. Meng. 2023. *Perplexity from PLM Is Unreliable for Evaluating Text Quality*. arXiv: 2210.05892 (cs.CL).

Z. Wang, J. Zeng, O. Delalleau, H.-C. Shin, F. Soares, A. Bukharin, E. Evans, Y. Dong, and O. Kuchaiev. 2025. "HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages." In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=lovsIkZLnI.

Z. Wang, L. Guan, G. Liu, and J. Ma. 2023. "Generation of Chinese classical poetry based on pretrained model." In: *2023 2nd International Conference on Big Data, Information and Computer Network (BDICN)* (Xishuangbanna, China, Jan. 6–8, 2023). IEEE, 182–185. doi:10.1109/bdicn58493.2023.00045.

A. Warstadt, A. Parrish, H. Liu, A. Mohananey, W. Peng, S.-F. Wang, and S. R. Bowman. 2020. "BLiMP: The Benchmark of Linguistic Minimal Pairs for English." *Transactions of the Association for Computational Linguistics*, 8, 377–392. Ed. by M. Johnson, B. Roark, and A. Nenkova. doi:10.1162/tacl_a_00321.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter brian, F. Xia, E. Chi, Q. V. Le, and D. Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec. 2020. *TRL: Transformer Reinforcement Learning*. GitHub repository. https://github.com/huggingface/trl.

T. Winters, V. Nys, and D. De Schreye. 2018. "Automatic Joke Generation: Learning Humor from Examples." In: *Lecture Notes in Computer Science*. Vol. 10922: *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts. Proceedings of 6th International Conference* (DAPI 2018) (Las Vegas, NV, USA, July 15–20, 2018). Ed. by N. Streitz and S. Konomi. Springer, Cham, 360–377. ISBN: 978-3-319-91131-1. doi:10.1007/978-3-319-91131-1_28.

J. Wöckener, T. Haider, T. Miller, T.-K. Nguyen, T. T. L. Nguyen, M. V. Pham, J. Belouadi, and S. Eger. Nov. 2021. "End-to-end style-conditioned poetry generation: What does it take to learn from examples alone?" In: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, and S. Szpakowicz. Association for Computational Linguistics, Punta Cana, Dominican Republic (online), (Nov. 2021), 57–66. doi:10.18653/v1/2021.latechclfl-1.7.

C.-C. Wu, R. Song, T. Sakai, W.-F. Cheng, X. Xie, and S.-D. Lin. 2019. "Evaluating Image-Inspired Poetry Generation." In: *Lecture Notes in Computer Science*. Vol. 11838: *Proceedings of Natural Language Processing and Chinese Computing Conference* (NLPCC 2019) (Dunhuang, China, Oct. 9–14, 2019). Ed. by J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan. Springer International Publishing, 539–551. ISBN: 978-3-030-32233-5. doi:10.1007/978-3-030-32233-5_42.

Y. Xia, R. Wang, X. Liu, M. Li, T. Yu, X. Chen, J. McAuley, and S. Li. Jan. 2025. "Beyond Chain-of-Thought: A Survey of Chain-of-X Paradigms for LLMs." In: *Proceedings of the 31st International Conference on Computational Linguistics*. Ed. by O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert. Association for Computational Linguistics, Abu Dhabi, UAE, (Jan. 2025), 10795–10809. https://aclanthology.org/2025.coling-main.719.

S. Xie, R. Rastogi, and M. Chang. 2017. *Deep poetry: Word-level and character-level language models for Shakespearean sonnet generation*. Course Project Report for 2017 CS224n: Natural Language Processing with Deep Learning. Department of Computer Science, Stanford University. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2762063.pdf.

Y. Xie, K. Kawaguchi, Y. Zhao, J. X. Zhao, M.-Y. Kan, J. He, and M. Xie. 2023. "Self-Evaluation Guided Beam Search for Reasoning." In: *Advances in Neural Information Processing Systems (NeurIPS 2023)*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 41618–41650. https://proceedings.neurips.cc/paper_files/paper/2023/hash/81fde95c4dc79188a69ce5b24d63010b-Abstract-Conference.html.

Z. Xie, J. H. Lau, and T. Cohn. 2019. "From Shakespeare to Li-Bai: Adapting a Sonnet Model to Chinese Poetry." In: *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*. Ed. by M. Mistica, M. Piccardi, and A. MacKinlay. Australasian Language Technology Association, Sydney, Australia, 10–18. https://aclanthology.org/U19-1002.

L. Xu, L. Jiang, C. Qin, Z. Wang, and D. Du. 2018. "How Images Inspire Poems: Generating Classical Chinese Poetry from Images with Memory Networks." In: *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI-18). Vol. 32. Association for the Advancement of Artificial Intelligence (AAAI), 5618–5625. doi:10.1609/aaai.v32i1.12001.

L. Xue, K. Song, D. Wu, X. Tan, N. L. Zhang, T. Qin, W.-Q. Zhang, and T.-Y. Liu. Aug. 2021. "DeepRapper: Neural Rap Generation with Rhyme and Rhythm Modeling." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Association for Computational Linguistics, Online, (Aug. 2021), 69–81. doi:10.18653/v1/2021.acl-long.6.

L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. Mar. 2022. "ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models." *Transactions of the Association for Computational Linguistics*, 10, (Mar. 2022), 291–306. doi:10.1162/tacl_a_00461.

L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. June 2021. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard,

R. Cotterell, T. Chakraborty, and Y. Zhou. Association for Computational Linguistics, Online, (June 2021), 483–498. doi:10.18653/v1/2021.naacl-main.41.

R. Yan. 2016. "i, Poet: Automatic Poetry Composition through Recurrent Neural Networks with Iterative Polishing Schema." In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (IJCAI'16). Ed. by S. Kambhampati. AAAI Press, New York, New York, USA, 2238–2244. https://www.ijcai.org/Abstract/16/319.

H. Yang, D. Cai, H. Li, W. Bi, W. Lam, and S. Shi. May 2024. "A Frustratingly Simple Decoding Method for Neural Text Generation." In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue. ELRA and ICCL, Torino, Italia, (May 2024), 536–557. https://aclanthology.org/2024.lrec-main.47.

K. Yang and D. Klein. June 2021. "FUDGE: Controlled Text Generation With Future Discriminators." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou. Association for Computational Linguistics, Online, (June 2021), 3511–3535. doi:10.18653/v1/2021.naacl-main.276.

X. Yang, X. Lin, S. Suo, and M. Li. 2018. "Generating Thematic Chinese Poetry using Conditional Variational Autoencoders with Hybrid Decoders." In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (IJCAI-18) (Stockholm, July 13–19, 2018). Ed. by J. Lang. International Joint Conferences on Artificial Intelligence, 4539–4545. ISBN: 978-0-9992411-2-7. doi:10.24963/ijcai.2018/631.

Z. Yang, Z. Liu, et al.. 2024. *Analyzing Nobel Prize Literature with Large Language Models*. arXiv: 2410.18142 (cs.CL).

Z. Yang, P. Cai, Y. Feng, F. Li, W. Feng, E. S.-Y. Chiu, and H. Yu. Nov. 2019. "Generating Classical Chinese Poems from Vernacular Chinese." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Association for Computational Linguistics, Hong Kong, China, (Nov. 2019), 6155–6164. doi:10.18653/v1/D19-1637.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. 2019. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." In: *Advances in Neural Information Processing Systems* (NeurIPS 2019) (Vancouver, BC, Canada, Dec. 8–14, 2019). Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 5753–5763. https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

S. Ye, D. Kim, S. Kim, H. Hwang, S. Kim, Y. Jo, J. Thorne, J. Kim, and M. Seo. 2024. "FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets." In: *The Twelfth International Conference on Learning Representations* (ICLR 2024) (Vienna, Austria, May 7–11, 2024). https://openreview.net/forum?id=CYmF38ysDa.

M. Yee-king, A. Fiorucci, and M. d'Inverno. 2023. *The pop song generator: designing an online course to teach collaborative, creative AI*. arXiv: 2306.10069 (cs.CY).

L. Yi. 2023. *Controllable Ancient Chinese Lyrics Generation Based on Phrase Prototype Retrieving*. (2023). arXiv: 2303.11005 (cs.CL).

Q. Yi, X. Chen, C. Zhang, Z. Zhou, L. Zhu, and X. Kong. 2024. "Diffusion models in text generation: a survey." *PeerJ Computer Science*, 10, e1905. doi:10.7717/peerj-cs.1905.

X. Yi, R. Li, and M. Sun. Oct. 2018. "Chinese Poetry Generation with a Salient-Clue Mechanism." In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Ed. by A. Korhonen and I. Titov. Association for Computational Linguistics, Brussels, Belgium, (Oct. 2018), 241–250. doi:10.18653/v1/K18-1024.

X. Yi, R. Li, C. Yang, W. Li, and M. Sun. 2020. "MixPoet: Diverse Poetry Generation via Learning Controllable Mixed Latent Space." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 05 (New York, USA, Feb. 7–12, 2020). Vol. 34. AAAI Press, Palo Alto, California USA, 9450–9457. doi:10.1609/aaai.v34i05.6488.

X. Yi, M. Sun, R. Li, and W. Li. 2018. "Automatic Poetry Generation with Mutual Reinforcement Learning." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii. Association for Computational Linguistics, Brussels, Belgium, 3143–3153. doi:10.18653/v1/d18-1353.

H. Young, Y. Zeng, J. Gardner, and O. Bastani. 2024. *Improving Structural Diversity of Blackbox LLMs via Chain-of-Specification Prompting*. arXiv: 2408.06186 (cs.CL).

C. Yu, L. Zang, J. Wang, C. Zhuang, and J. Gu. Aug. 2024. "CharPoet: A Chinese Classical Poetry Generation System Based on Token-free LLM." In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Ed. by Y. Cao, Y. Feng, and D. Xiong. Association for Computational Linguistics, Bangkok, Thailand, (Aug. 2024), 315–325. doi:10.18653/v1/2024.acl-demos.30.

L. Yu, W. Zhang, J. Wang, and Y. Yu. 2017. "SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient." In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* 1 (San Francisco, California, USA, Feb. 4–9, 2017). Vol. 31. AAAI Press, Palo Alto, California USA, 2852–2858. doi:10.1609/aaai.v31i1.10804.

Y. Yu, A. Srivastava, and S. Canales. Feb. 2021. "Conditional LSTM-GAN for melody generation from lyrics." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17, 1, Article 35, (Feb. 2021), 20 pages. doi:10.1145/3424116.

D. Zhang, J.-C. Wang, K. Kosta, J. B. L. Smith, and S. Zhou. Dec. 2022. "Modeling the rhythm from lyrics for melody generation of pop songs." In: *Proceedings of the 23rd International Society for Music Information Retrieval Conference*. ISMIR, Bengaluru, India, (Dec. 2022), 141–148. doi:10.5281/zenodo.7316616.

G. Zhang et al.. Sept. 2022. "Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech." In: *Interspeech 2022* (Incheon, Korea). ISCA, (Sept. 2022), 456–460. doi:10.21437/interspeech.2022-621.

L. Zhang, R. Zhang, X. Mao, and Y. Chang. May 2022. "QiuNiu: A Chinese Lyrics Generation System with Passage-Level Input." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Ed. by V. Basile, Z. Kozareva, and S. Stajner. Association for Computational Linguistics, Dublin, Ireland, (May 2022), 76–82. doi:10.18653/v1/2022.acl-demo.7.

R. Zhang and S. Eger. 2024. *LLM-based multi-agent poetry generation in non-cooperative environments*. arXiv: 2409.03659 [cs.CL].

R. Zhang, X. Mao, L. Li, L. Jiang, L. Chen, Z. Hu, Y. Xi, C. Fan, and M. Huang. Oct. 2020. "Youling: an AI-assisted Lyrics Creation System." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Q. Liu and D. Schlangen. Association for Computational Linguistics, Online, (Oct. 2020), 85–91. doi:10.18653/v1/2020.emnlp-demos.12.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2020. "BERTScore: Evaluating Text Generation with BERT." In: *8th International Conference on Learning Representations* (ICLR 2020) (Virtual, Apr. 26–May 1, 2020). https://openreview.net/forum?id=SkeHuCVFDr.

T. Zhang, M. Lee, X. L. Li, E. Shen, and T. Hashimoto. July 2023. "TempLM: Distilling Language Models into Template-Based Generators." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, Toronto, Canada, (July 2023), 1970–1994. doi:10.18653/v1/2023.findings-acl.124.

X. Zhang, M. Sun, J. Liu, and X. Li. July 2023. "Lingxi: A Diversity-aware Chinese Modern Poetry Generation System." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Ed. by D. Bollegala, R. Huang, and A. Ritter. Association for Computational Linguistics, Toronto, Canada, (July 2023), 63–75. doi:10.18653/v1/2023.acl-demo.6.

Z. Zhang, K. Lasocki, Y. Yu, and A. Takasu. Mar. 2024. "Syllable-level lyrics generation from melody exploiting character-level language model." In: *Findings of the Association for Computational Linguistics: EACL 2024*. Ed. by Y. Graham and M. Purver. Association for Computational Linguistics, St. Julian's, Malta, (Mar. 2024), 1336–1346. https://aclanthology.org/2024.findings-eacl.89.

C. Zhao, B. Wang, and Z. Wang. 2024. *Understanding Literary Texts by LLMs: A Case Study of Ancient Chinese Poetry*. arXiv: 2409.00060 [cs.CL].

J. Zhao and H. J. Lee. 2022. "Automatic Generation and Evaluation of Chinese Classical Poetry with Attention-Based Deep Neural Network." *Applied Sciences*, 12, 13, 6497. doi:10.3390/app12136497.

L. Zheng et al.. 2023. "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena." In: *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 46595–46623. https://papers.nips.cc/paper%5C_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets%5C_and%5C_Benchmarks.html.

D. Zhu, j. chen jun, X. Shen, X. Li, and M. Elhoseiny. 2024. "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models." In: *International Conference on Representation Learning*. Ed. by B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun. Vol. 2024, 18378–18394. https://proceedings.iclr.cc/paper_files/paper/2024/hash/50623630a2372839c078474efa6c0cb8-Abstract-Conference.html.

W. Zhu and S. Bhat. Nov. 2020. "GRUEN for Evaluating Linguistic Quality of Generated Text." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by T. Cohn, Y. He, and Y. Liu. Association for Computational Linguistics, Online, (Nov. 2020), 94–108. doi:10.18653/v1/2020.findings-emnlp.9.

W. Zhu, H. Hao, Z. He, Y. Ai, and R. Wang. July 2024. "Improving Open-Ended Text Generation via Adaptive Decoding." In: *Proceedings of the 41st International Conference on Machine Learning* (Proceedings of Machine Learning Research). Ed. by R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp. Vol. 235. PMLR, (July 2024), 62386–62404. https://proceedings.mlr.press/v235/zhu24d.html.

X. Zou. July 2025. "BIPro: Zero-shot Chinese Poem Generation via Block Inverse Prompting Constrained Generation Framework." In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Association for Computational Linguistics, Vienna, Austria, (July 2025), 1116–1134. ISBN: 979-8-89176-251-0. doi:10.18653/v1/2025.acl-long.56.

A. Zugarini, S. Melacci, and M. Maggini. 2019. "Neural Poetry: Learning to Generate Poems Using Syllables." In: *Lecture Notes in Computer Science*. Vol. 11730: *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series. Proceedings of 28th International Conference on Artificial Neural Networks* (Munich, Germany, Sept. 17–19, 2019). Ed. by I. V. Tetko, V. Kůrková, P. Karpov, and F. Theis. Springer International Publishing, Cham, 313–325. ISBN: 978-3-030-30490-4. doi:10.1007/978-3-030-30490-4_26.

A. Zugarini, L. Pasqualini, S. Melacci, and M. Maggini. 2021. "Generate and Revise: Reinforcement Learning in Neural Poetry." In: *2021 International Joint Conference on Neural Networks (IJCNN)* (Shenzhen, China, July 18–22, 2021). IEEE, 1–8. doi:10.1109/ijcnn52387.2021.9533573.

Table 6. Statistics of works in the ACL Anthology from 2006 to 2025. Qwen3-8B model (Qwen Team 2025) with zero-shot prompting is used to determine whether a paper with title and abstract data belongs to the field of poetry generation, analysis, or evaluation.

Year	Total entries	Num. of poetry-related	Share of poetry-related, %
2006	2,174	1	0.046
2007	1,601	2	0.125
2008	2,232	3	0.134
2009	2,233	1	0.045
2010	3,069	4	0.13
2011	2,290	2	0.087
2012	3,391	4	0.118
2013	2,824	4	0.142
2014	3,639	4	0.11
2015	2,958	2	0.068
2016	4,220	7	0.166
2017	3,487	8	0.229
2018	4,822	8	0.166
2019	5,074	10	0.197
2020	7,245	9	0.124
2021	7,148	8	0.112
2022	8,649	12	0.139
2023	9,032	13	0.144
2024	12,098	5	0.041
2025	14,547	12	0.082

A ACL Anthology Analysis

An analysis of ACL Anthology database reveals that research interest in poetry-related topics, including the analysis of human-authored poems, has remained relatively stable over time compared to the overall volume of recorded works (see Table 6). However, it is important to note that a significant portion of these papers focuses on the analysis of human-written poetry rather than generative poetry. Consequently, the number of works directly relevant to the focus of this survey — generative poetry — is considerably smaller.

The data presented in Table 6 was collected by processing the ACL Anthology using the following automated procedure.

(1) The complete anthology dataset was downloaded from full ACL Anthology as BibTeX with abstracts archive[44] on January 21, 2026. This dataset contained 119,961 entries.
(2) The BibTeX data was parsed using the `pybtex` library.[45]
(3) For each entry, the title and abstract text were extracted. Curly braces, `{` and `}`, were removed from title fields to prevent reducing LLM performance due to LaTeX formatting.
(4) Each publication was classified using a zero-shot prompt with the following template, where the `title_text` and `abstract_text` placeholders were replaced with the extracted data:

```
Analyze the following title and abstract of an article from the ACL Anthology.
Determine whether this article is about generative poetry, poetry text analysis,
```

[44] https://aclanthology.org/anthology+abstracts.bib.gz
[45] https://pybtex.org

```
meter and rhyme definition, poetry evaluation metrics, or other topics related
to LLM-generated poetry.

If it is, output <result>True</result>
If it is not, output <result>False</result>

Title:
{title_text}

Abstract:
{abstract_text}
```

(5) Each assembled prompt was processed by the Qwen3-8B model[46] using model's chat template. Greedy decoding was used to generate the response.
(6) If the model's response contained the exact string `<result>True</result>`, a publication was categorized as poetry-related.

For entries classified as poetry by the above procedure, another analysis was performed using the same LLM Qwen to identify the target languages of poetry mentioned in the paper's title and abstract (if available). This analysis yields a rough distribution of the papers by language, summarized in a Table 7.

Table 7. Distribution of languages in the ACL Anthology poetry-related papers from 2006 to 2025

Language	Total entries
Chinese	27
English	18
French	6
Classical Chinese	3
Basque	3
Finnish	3
Czech	2
Russian	2
Persian	2
Arabic	2
Mandarin	2
Greek	1
Cantonese	1
Latin	1
Bangla	1
German	1
Spanish	1
Tibetan	1
Afrikaans	1
Portuguese	1
Romanian	1
Tamil	1
Japanese	1

[46] https://huggingface.co/Qwen/Qwen3-8B

The complete processing pipeline required approximately 8 hours to execute on an A100 GPU.

It is important to note that all classifications and language identifications reported in this appendix are produced fully automatically by the LLM-based procedure described above. We did not perform manual validation or post-processing of the model outputs. As a result, the reported statistics may contain a non-negligible rate of misclassification. This includes, for example, papers incorrectly labeled as poetry-related, papers on generative poetry that were not detected as such, or cases where the target language of the poetic text was misidentified or not mentioned explicitly in the abstract. The presented results should therefore be interpreted as approximate indicators of broader trends rather than as exact counts.

B Datasets

This section provides detailed parameters for the datasets mentioned in Table 2. These datasets can be used for generative poetry projects, eliminating the need for researchers to spend time collecting training or test data from online resources like Poetry Foundation,[47] Project Gutenberg,[48] Wikisource,[49] PoetryDB.[50]

For English-language poetry, a common solution is to use Project Gutenberg, such as the Gutenberg Poetry Corpus,[51] which contains about 3 million rows. This corpus does not store poems, but rather lines from them, without information about their order, so this data is only suitable for special generation tasks. In addition to this source, there are English-language datasets collected from various platforms, such as the "Poems dataset,"[52] "Poetry Foundation Poems"[53] (contains collections of poems in specific genres or forms, such as acrostics), "Public Domain Poetry."[54] Haider (2021) presented the dataset[55] of annotated verse for a varied sample of around 7,000 lines for German and English languages. Poems are given line-by-line meter information and syllable stress placement, which can be useful when developing poetry scansion tools or testing automated generative poetry assessments. In addition, we can mention PO-EMO[56] — a dataset with emotional tagging for English-language poems described by Haider, Eger, et al. (2020). Poem Emotion Recognition Corpus (PERC)[57] is a small dataset of English-language poems by Indian authors with emotion annotation described by (Sreeja and Mahalakshmi 2019). Emotional labeling can be used to create instructional datasets for training generative models, as well as for calibrating classifiers used to evaluate poem generations.

Abdibayev, Tikhonov, et al. (2021) released a large dataset of English-language limericks (R. Greene et al. 2012, page 449) containing 98,000 samples collected from The Omnificent English Dictionary In Limerick Form.[58]

Mahbub et al. (2023) introduced the "PoemSum" dataset[59] consisting of 3,011 samples of poetry and its corresponding summarized interpretation in the English language.

Aoyama et al. (2023) presented a manually annotated dataset of English texts covering several NLP tasks, such as syntactic dependency parsing, named entity recognition, coreference resolution, and discourse parsing. Because poetry is among the covered genres, the dataset can support the evaluation of data engineering tools (Section 3).

[47] https://www.poetryfoundation.org/
[48] https://gutenberg.org/
[49] https://wikisource.org/
[50] http://poetrydb.org/index.html
[51] https://github.com/aparrish/gutenberg-poetry-corpus
[52] https://www.kaggle.com/datasets/michaelarman/poemsdataset
[53] https://www.kaggle.com/datasets/tgdivy/poetry-foundation-poems
[54] https://huggingface.co/datasets/DanFosing/public-domain-poetry
[55] https://github.com/tnhaider/metrical-tagging-in-the-wild
[56] https://github.com/tnhaider/poetry-emotion
[57] https://huggingface.co/datasets/COMP0087-GROUP8-22-23/PERC
[58] https://www.oedilf.com/
[59] https://github.com/Ridwan230/PoemSum

For Spanish, several poetry datasets are available:

- A Spanish poetry dataset[60] containing 5,133 poems with metadata, collected by Garzón and Pérez (2020). This dataset, containing authorship information, can be helpful as a source for systems generating poems in a given style.
- The Diachronic Spanish Sonnet Corpus with Psychological and Affective Labels (DISCO PAL),[61] consisting of 4,530 sonnets and described by Barbado et al. (2022).

Available datasets for Chinese include:

- THUNLP-AIPoet Datasets[62] — released by THUAIPoet (Jiuge) group at Tsinghua University's Natural Language Processing Lab (THUNLP) for automatic poetry generation. The following datasets are included: a poetry quality evaluation dataset containing 173 poems labeled by fluency, coherence, and meaningfulness described in X. Yi, M. Sun, et al. (2018); a fine-grained sentimental poetry corpus containing 5,000 poems; a Chinese classical poetry corpus containing more than 127k samples divided into training, validation, and evaluation splits.
- Chinese Classical Poetry Matching Dataset (CCPM)[63] — a parallel corpus of poetry in modern and archaic Chinese (W. Li et al. 2021), containing over 27,000 poems.
- Comprehensive Database of Chinese Poetry,[64] containing over 330,000 poems from the Tang and Song dynasties.
- Mojim Popular Lyrics[65] — lyrics dataset (Crothers et al. 2023) from a lyrics sharing website, containing over 39,000 popular Chinese songs.
- Corpus of tokenized classical Chinese poems (C.-L. Liu et al. 2022) containing 32,399 poems labeled by human annotators — focuses on Tang and Song dynasty poetry.

For the Persian language, available datasets include:

- Shereno: A dataset of Persian modernist poetry[66] containing over 4,000 poems.
- Prose2Poem:[67] A parallel corpus of prose and ancient Persian poetry containing 1,319 poems, released by Khanmohammadi et al. (2023).

Several datasets are available for Arabic poetry research:

- Poem Comprehensive Dataset (PCD)[68] — used by Abboushi and Azzeh (2023) to train an AraGPT2-based Arabic poetry generation system. This dataset contains 1,831,771 poems.
- Ashaar dataset[69] — employed by Alyafeai et al. (2023) for Arabic poetry analysis and generation. The structure of this dataset is not described clearly enough, so we cannot provide information on the number of poems presented.
- Arabic Poetry Emotion dataset[70] — open-source dataset for emotion classification developed by Shahriar et al. (2023). This dataset includes 9,452 Arabic poems, labeled with emotions.

[60] https://huggingface.co/datasets/andreamorgar/spanish_poetry
[61] https://github.com/pruizf/disco
[62] https://github.com/THUNLP-AIPoet/Datasets
[63] https://github.com/THUNLP-AIPoet/CCPM
[64] https://github.com/chinese-poetry/chinese-poetry
[65] https://github.com/ecrows/mojim-lyrics
[66] https://www.kaggle.com/datasets/elhamaghakhani/persian-poems
[67] https://github.com/mitramir55/Prose2Poem
[68] https://hci-lab.github.io/ArabicPoetry-1-Private
[69] https://github.com/ARBML/Ashaar
[70] https://github.com/sakibsh/Arabic-Poem-Emotion

Biblioteca_italiana[71] is a subset (more than 18000 works from over 159 authors) of the available poems in the poetry corpus from the Biblioteca Italiana.

Poetry datasets are also available for other languages:

- German (Haider 2024): over 65,000 poems from the New High German period.
- French:[72] 1,821 poems with metadata.
- Turkish:[73] 4,691 poems with metadata.
- Kurdish (Ahmadi et al. 2020): folkloric lyrics and songs in Sorani Kurdish.
- Hindi:[74] 1,151 poems with metadata.
- Bulgarian:[75] 1,346 poems.

A detailed description of available datasets and tools for analyzing Czech poetic texts is provided by R. Rosa et al. (2025). In their paper, the authors summarize their work on analyzing and generating Czech poetry using the Corpus of Czech Verse, which contains 80,229 poems. The first part of the paper introduces a framework for the automatic analysis of poetry. This framework extracts various properties from a given poem, including phonetic transcriptions, syllabic features, morphological and syntactic annotations, versological annotations (reduplicants, rhymes, stresses, and metres), thematic motives, and stylometric analysis results. The second part presents a method for generating Czech poetry using finetuned Czech GPT-2 and Llama-3.1 LMs. The quality of the generated poetry was evaluated using the aforementioned analytical framework, with a focus on metrics such as rhyming consistency, metre consistency, and style accuracy.

For certain tasks, suitable datasets are either unavailable, incomplete, or require substantial preprocessing. This is particularly evident in poetry translation (Section 2.4), where obtaining high-quality professional translations involves navigating legal constraints, while scraping amateur translations introduces both copyright concerns and significant data cleaning overhead.

In melody-to-lyrics generation (Section 2.9), several small datasets exist for evaluation via side-by-side comparison. For instance, DALI (Meseguer-Brocal et al. 2018) provides approximately 5,000 songs with aligned audio and lyrics. For larger-scale training, researchers must assemble data from multiple sources. One approach involves combining the Million Song Dataset (Bertin-Mahieux et al. 2011) — which contains metadata and audio links — with lyric repositories such as the 5 Million Song Lyrics Dataset,[76] followed by downloading and processing hundreds of gigabytes of audio data.

C Visualization of Mainstream Approaches Over Time

This appendix describes the procedure used to produce Figure 1, which summarizes the number of papers on generative poetry that employ template-based (Section 4.2.1), RNN-based (Section 4.2.2), or transformer-based (Section 4.2.3) approaches.

First, we collected all relevant papers cited in the survey and grouped them into three lists according to their primary modeling approach. Several papers were included in more than one list because they describe hybrid approaches that combine template-based methods with either RNN-based (Manjavacas et al. 2019) or transformer-based (Qian et al. 2023) models. Diffusion-based models (Section 4.2.4) were grouped with transformer-based approaches, as they rely on similar large-scale transformer language modeling infrastructures. In addition, several papers (Cheng et al. 2025; Elzohbi and R. Zhao 2025b; Huang and X. Shen 2025; Murakami and Terai 2023; Secha

[71] https://github.com/linhd-postdata/biblioteca_italiana
[72] https://huggingface.co/datasets/manu/french_poetry
[73] https://huggingface.co/datasets/okg/turkish-poems
[74] https://huggingface.co/datasets/Sourabh2/Hindi_Poems
[75] https://huggingface.co/datasets/Dilyana56/bulgarian_poems
[76] https://www.kaggle.com/datasets/nikhilnayak123/5-million-song-lyrics-dataset

et al. 2022; L. Zhang et al. 2022; R. Zhang, Mao, et al. 2020) that were not discussed in detail in the main text but satisfied the same inclusion criteria were added to the lists.

The resulting lists were then used as input for Python-based processing and visualization. Publication years were obtained from a bibliography file parsed using the `bibtexparser` library.[77]

Papers published before 2016 were excluded, as earlier work is not systematically covered in this survey.

Finally, the processed publication-year lists were visualized using the `matplotlib` library (Hunter 2007).

Received 29 September 2025; accepted 01 April 2026

[77] https://github.com/sciunto-org/python-bibtexparser