Journal of Artificial Intelligence Research 80 (2024) 1271-1310 Submitted 09/2023; published 08/2024

Computational Argumentation-based Chatbots: a Survey

Federico Castagna FEDERICO.CASTAGNA@BRUNEL.AC.UK
Brunel University London, Kingston Lane, London, UB8 3PH, United Kingdom

Nadin Kökciyan NADIN.KOKCIYAN@ED.AC.UK
University of Edinburgh, Crichton St, Edinburgh EH8 9AB, United Kingdom

Isabel Sassoon ISABEL.SASSOON@BRUNEL.AC.UK
Brunel University London, Kingston Lane, London, UB8 3PH, United Kingdom

Simon Parsons SPARSONS@LINCOLN.AC.UK
Elizabeth Sklar ESKLAR@LINCOLN.AC.UK
University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, United Kingdom

Abstract

Chatbots are conversational software applications designed to interact dialectically with users for a plethora of different purposes. Surprisingly, these colloquial agents have only recently been coupled with computational models of arguments (i.e. computational argumentation), whose aim is to formalise, in a machine-readable format, the ordinary exchange of information that characterises human communications. Chatbots may employ argumentation with different degrees and in a variety of manners. The present survey sifts through the literature to review papers concerning this kind of argumentation-based bot, drawing conclusions about the benefits and drawbacks that this approach entails in comparison with standard chatbots, while also envisaging possible future development and integration with the Transformer-based architecture and state-of-the-art Large Language Models.

1. Introduction

Chatbots are conversational software applications designed to mimic human discourse mostly to enable automated online guidance and support (Caldarini et al., 2022). These computer programs generate responses based on given inputs, producing replies via text or speech format (Sojasingarayar, 2020; Bala et al., 2017). In addition, to be defined as such, chatbots must satisfy specific functions. As colloquial agents, they need to be able to understand the user (*comprehension*), have access to a knowledge base (*competence*) and provide an 'anthropomorphic effect' to increase the users' trust (*presence*) (Cahn, 2017; Sansonnet et al., 2006). Nowadays, these bots represent familiar tools that exist in our lives in the form of virtual agents. Their assistance ranges from answering inquiries to e-commerce, from information retrieval to educational tasks, and from developing new industrial solutions (Dale, 2016) to connecting smart objects (Kar & Haldar, 2016). The manifold investments of the past decade, the technological advancements (from both software and hardware viewpoints), and the development of more efficient Machine Learning (ML) models, including the latest Transformer-based architecture (Vaswani et al., 2017), have contributed to the

steady growth of the research field of chatbot design and implementation. Many steps forward have been taken since the release of *ELIZA* around sixty years ago, which is widely considered to be the first conversational agent (Weizenbaum, 1966).

The investigation of computational models of arguments in relation to chatbots has only recently received attention from researchers. Computational argumentation (Rahwan & Simari, 2009) has been applied in Artificial Intelligence (AI) as a mechanism for reasoning in which conclusions are drawn from evidence that supports the conclusions. Being an intuitive (i.e. closer to everyday human dialectical interplay), yet formal, approach for modelling conflicting information occurring during exchanges of arguments, computational argumentation should qualify as a highly appropriate methodology to enhance current bot behaviours. The benefits from such a combination include: more natural discourse, response coherence and strategical conveyance of information. Evaluating argumentation semantics would also provide the rationale for positing replies in a more transparent way than the black-box Large Language Models (LLMs) employed in today's state-of-the-art conversational agents. In recent years, cutting-edge technologies have produced implementations, such as the various versions of ChatGPT[1], which currently outperform argumentation-based conversational agents. Nonetheless, taking a closer look—as we do here—shows that there is plenty of room for improvement for these recent advanced models, and integration with the computational argumentation formalism may solve their present shortcomings (e.g. lack of explainability), thus potentially initiating a new generation of chatbots. To the best of our knowledge, this is the first survey that combines computational argumentation and chatbots[2]. Our main contribution involves an extensive examination of the relevant literature and the subsequent findings that can be drawn from such analysis.

The paper is structured as follows. We first start by introducing background information in Section 2 about the essential theoretical notions involved. In Section 3, we then discuss the methodology adopted for reviewing the relevant articles. A thorough classification and analysis of conversational agents leveraging computational argumentation is given in Section 4. Section 5 illustrates a comprehensive examination of our findings and potential future directions of the argumentation-based chatbot research field, and Section 6 concludes the survey with final remarks.

2. Background

The following background covers a concise summary of computational argumentation, along with a short overview of the history, classification and main features of chatbots. The information provided will prove useful for the analysis undertaken in the next sections, where each conversational agent will be classified according to the specific argumentation employment presented herein.

2.1 Computational Argumentation

The term 'computational model of arguments' encompasses a wide range of different approaches, each of which revolves around the notion of arguments and their employment. The

1. `https://chat.openai.com/`
2. Note that, for simplicity, we will often opt for the terminology 'argumentation-based chatbot' rather than 'computational argumentation-based chatbot', although the meaning will remain the same.

resulting research field, whose roots can be traced back to Pollock's and Dung's systematical account of arguments (Pollock, 1987; Dung, 1995), constitutes a rich interdisciplinary environment comprising subjects such as philosophy (Walton, 1990; Mercier & Sperber, 2011), jurisprudence (Bench-Capon et al., 2009), linguistics (Lawrence & Reed, 2020), formal logic (Lin & Shoham, 1989) and game theory (Rahwan & Larson, 2009). Within the scope of computational argumentation, it is possible to identify two main research goals: (a) understand argumentation as a cognitive phenomenon via computer program modelling; and (b) support the development of human-computer interaction by means of argumentation-related activities (Prakken et al., 2020; Dutilh Novaes, 2022). According to Dung's paradigm (Dung, 1995), arguments are considered suitable means to formalise non-monotonic reasoning, especially when showing how humans handle conflicting information in a dialectical way. The core notion of such an approach is underpinned by the definition of an argumentation framework, where arguments are intended as abstract entities:

Definition 1 (Abstract AFs (Dung, 1995)) *An argumentation framework* (AF) *is a pair:* $AF = \langle AR, \mathcal{C} \rangle$ *where* AR *is a set of arguments, and* $\mathcal{C}$ *is the 'attack' binary relation on* AR*, i.e.* $\mathcal{C} \subseteq AR \times AR$*.*

AFs can be rendered as graphs where each node is an argument, and every directed edge connects the conflicting arguments of the framework. The idea conveyed by this formalism is that correct reasoning is rendered via the acceptability of a statement: an argument is *justified* only if it is defended against any counterarguments.

Definition 2 (Semantics for Abstract AFs (Dung, 1995)) *Let* $AF = \langle AR, \mathcal{C} \rangle$*, and let* $\mathcal{S} \subseteq AR$ *be a set of arguments. Let also* $(X, Y) \in \mathcal{C}$ *denote the conflict existing between an argument* X *and its target* Y*:*

- $\mathcal{S}$ *is* conflict-free *iff* $\forall X, Y \in \mathcal{S}$*:* $(X, Y) \notin \mathcal{C}$*;*
- $X \in AR$ *is acceptable w.r.t.* $\mathcal{S}$ *iff* $\forall Y \in AR$ *such that* $(Y, X) \in \mathcal{C}$*:* $\exists Z \in \mathcal{S}$ *such that* $(Z, Y) \in \mathcal{C}$*;*
- *A conflict-free extension* $\mathcal{S}$ *is an* admissible *extension iff* $X \in \mathcal{S}$ *implies* X *is acceptable w.r.t.* $\mathcal{S}$*;*
- *An admissible extension* $\mathcal{S}$ *is a* complete *extension iff* $\forall X \in AR$*:* X *is acceptable w.r.t.* $\mathcal{S}$ *implies* $X \in \mathcal{S}$*. The minimal complete extension* (*with respect to set inclusion*) *is called the* grounded extension*, whereas a maximal complete extension* (*with respect to set inclusion*) *is called a* preferred extension*;*
- *A* stable extension $\mathcal{S}$ *is such that iff* $\forall Y \in AR$*, if* $Y \notin \mathcal{S}$*, then* $\exists X \in \mathcal{S}$ *such that* $(X, Y) \in \mathcal{C}$*.*

Furthermore, AFs can be instantiated by the formulae of some logical language. These instantiations paved the way for a plethora of different studies (e.g., Besnard and Hunter, 2008; Modgil and Prakken, 2013; Toni, 2014) concerning the so-called *structured argumentation*, as opposed to the previously introduced abstract approach. The internal structure of an argument is usually composed of (one or more) premises, a conclusion and a set

of inference rules (e.g. strict or defeasible) connecting premises to the conclusion. The same semantics described above can then be used to evaluate structured argumentation frameworks and compute justified arguments.

Example 1 *Let us consider the abstract AF depicted in Figure 1. Then, according to the semantics described in Definition 2, we can identify the following extensions:*

admissible $= \emptyset$, $\{a\}$, $\{b\}$, $\{e\}$, $\{a, e\}$, $\{b, e\}$;

complete $= \{e\}$, $\{a, e\}$, $\{b, e\}$;

grounded $= \{e\}$;

preferred $= \{a, e\}$, $\{b, e\}$;

stable $= \{a, e\}$.

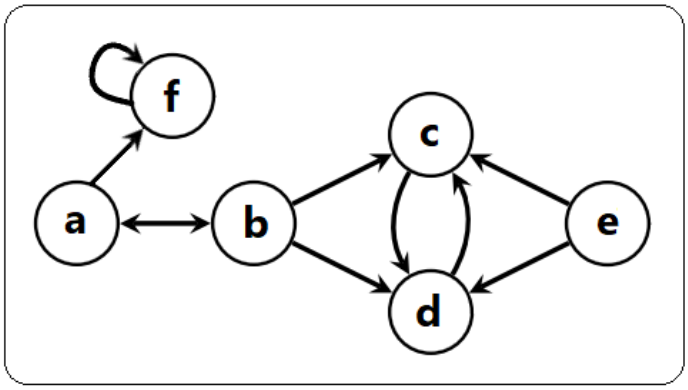

Figure 1: An abstract argumentation framework.

2.1.1 Argument Mining

Argument(ation) mining has been defined as *"the general task of analyzing discourse on the pragmatics level and applying a certain argumentation theory to model and automatically analyze the data at hand"* (Habernal & Gurevych, 2017). Argument mining (AM) can be considered the research area aimed at detecting natural language arguments and their relations in text, with the final goal of providing machine-processable structured data for computational models of argument (Cabrio & Villata, 2018). As depicted in Figure 5, an AM pipeline consists of two main stages: arguments' extraction and relations' prediction. We could delineate the AM framework by listing the tasks, in increasing order of complexity, that constitute such a framework. In short, moving from a preliminary textual segmentation and a classification of such elements as argumentative or not, it will then be possible to identify the single argument components (such as premises, claim, major claim, evidence, etc., as indicated in Mayer et al., 2020). The following steps envisage the recognition of clausal properties and relational properties with respect to the previously detected argument components (Lawrence & Reed, 2020). In particular, Saadat-Yazdi, Pan and Kökciyan

show how the use of external commonsense knowledge helps in identifying relations among arguments by uncovering implicit inferences (Saadat-Yazdi et al., 2023). Some of the models proposed in the literature include Long-Short Term Memory (LSTM) models (Cocarascu & Toni, 2017), pre-trained transformers (Ruiz-Dolz et al., 2021; Saadat-Yazdi et al., 2023) and logical rule-based systems (Jo et al., 2021). Overall, AM is useful in enabling the generation of an argumentation framework, or graph, from the mined corpus of texts. We now provide a more concrete analysis of the arguments' extraction stage within the AM pipeline.

Example 2 *Inspired by the political debate example illustrated in (Cabrio & Villata, 2018), we introduce an example to show how one can identify single arguments by following two distinct steps: (S1) the detection of argument components, such as premises and claims, and (S2) the recognition of their specific textual boundaries via the exclusion of any irrelevant words. In the following, we show how S1 and S2 could be applied to an example about the use of solar energy to extract an argument (Arg). Note that* ***(C)*** *and* ***(P)*** *distinguish the conclusion from the premises, whereas the bold and underlined fonts identify their respective boundaries.*

(S1) *"She talks about solar panels. We invested in a solar company, our country. That was a disaster* ***(C)****. They lost plenty of money on that one* ***(P)****. Now, look, I'm a great believer in all forms of energy* ***(P)****, but we're putting a lot of people out of work* ***(P)****."*

(S2) *"She talks about solar panels. We invested in a solar company, our country.* ***That was a disaster.*** *They lost plenty of money on that one. Now, look, I'm a great believer in all forms of energy, but we're putting a lot of people out of work."*

(Arg) *[Since] they lost plenty of money on that one, [even though] I'm a great believer in all forms of energy, we're [nonetheless] putting a lot of people out of work. [We can then conclude] that was a disaster.*

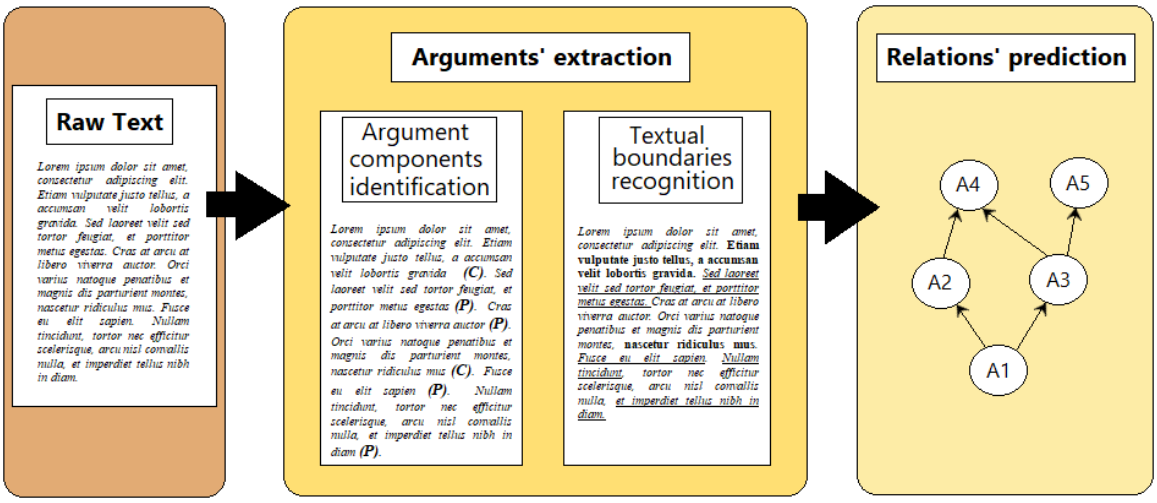

Figure 2: Example of an argumentation mining pipeline with fictitious text.

2.1.2 Argument Schemes

Argument schemes (AS) have been extensively investigated and employed in the AI literature as a way to directly convey presumptive reasoning in multi-agent interactions (e.g., Atkinson et al., 2006; Tolchinsky et al., 2012; Grando et al., 2013; Kökciyan et al., 2018; Kökciyan et al., 2021). Each AS is characterized by a unique set of critical questions (CQs), rendered as attacking arguments, whose purpose is to establish the validity of the scheme instantiations (which can then be evaluated by semantically computing their acceptability). Although the literature presents diverse classification systems for argument schemes (e.g., Walton et al., 2008; Walton and Macagno, 2015; Wagemans, 2016), they all share the idea that such schemes constitute reasoning patterns that may be harnessed to structure natural language text into rational and coherent arguments, thus generating systematic elements of dialogue.

Example 3 *As an example of AS in the healthcare domain, consider the argument scheme for proposed treatment (ASPT), as rendered in (Sassoon et al., 2021), and the respective critical questions: the validity of any potential ASPT instantiation depends upon the answers given to each critical question.*

ASPT
Premise : Given the patient's fact Ft *Premise* : In order to realise goal G *Premise* : Treatment T promotes goal G ——————————— *Conclusion* : Treatment T should be considered

CQ1: Has treatment T been unsuccessfully used on the patient in the past?

CQ2: Has treatment T caused side effects for the patient?

CQ3: Given the patient's fact Ft, are there counter-indications to treatment T?

CQ4: Are there alternative actions to achieve the same goal G?

The evaluation of AS via critical questions may take place in two different manners, according to the most accredited theories presented by Walton and Gordon (2011): (a) *initiative shifting* and (b) *backup evidence*.

(*a*) After having asked a critical question, the initiative immediately shifts to the proponent that has to provide an answer or else the argument is considered defeated. That is to say, asking the question is enough to temporarily defeat the argument. Nevertheless, the proponent has the capability of retaining the argument validity by providing an appropriate answer to the critical question.

(*b*) On the other hand, the second theory states that asking a critical question does not suffice to defeat the argument. The question, if challenged needs to be backed up with some evidence before it can shift any burden that would defeat the argument.

Finally, although the concept was developed for different purposes, the importance of argument schemes has found uptake within the computational argumentation community (Visser et al., 2018) also for textual mining tasks (Walton, 2012).

2.1.3 Argumentation Reasoning Engine

One of the main purposes of computational argumentation is to enable the resolution of conflicting knowledge, thus allowing for a selection of the most appropriate (i.e. justified) pieces of information. *"A decision is a choice between competing beliefs about the world or between alternative courses of action. [...] Inference processes generate arguments for and against each candidate [belief or action]. Decision making then ranks and evaluates candidates based on the underlying arguments and selects one candidate as the final decision. Finally, the decision commits to a new belief about a situation, or an intention to act in a particular way."* (Fox et al., 2007). Decision-making processes can be encoded as problems whose solutions are rendered by the computation and evaluation of AFs: an argumentation engine is essentially a reasoning tool driven by the same logic. The resulting acceptable entities provide a compelling rationale for and against a given choice, while also leaving space for further deliberations (Dix et al., 2009). Such an argumentative decision-making apparatus can be a useful addition to any real-world software application concerning defeasible reasoning, as advocated by the comprehensive study of Bryant and Krause (2008). We can distinguish two kinds of reasoning engines based on computational argumentation:

- 'Solvers', i.e. specialized pieces of software that encode and provide answers to distinct algorithmic problems. In particular, they address computational argumentation-related reasoning challenges according to a chosen semantics σ: for example, the enumeration of σ-extensions in the AF and the credulous and sceptical membership of a specific argument to at least one (credulous) or each (sceptical) σ-extensions. Examples of solvers are *AFGCN* (Malmqvist, 2021), *A-Folio DPDB* (Fichte et al., 2021), *ASPARTIX-V21* (Dvorák et al., 2021), *ConArg* (Bistarelli et al., 2021a), *FUDGE* (Thimm et al., 2021), *HARPER++* (Thimm, 2021), *MatrixX* (Heinrich, 2021), *μ-toksia* (Niskanen & Järvisalo, 2021), *PYGLAF* (Alviano, 2021).

- 'Panoptic Engines', i.e. solvers designed to implement additional functionalities and customisation tools, such as *ArguLab* (Podlaszewski et al., 2011), *ArgTrust* (Tang et al., 2012), *Argue tuProlog* (Bryant et al., 2006), *IACAS* (Vreeswijk, 1994), *CaSAPI* (Gartner & Toni, 2007), *Prengine* (Hung, 2017), *PyArg* (Borg et al., 2022), *NEXAS* (Dachselt et al., 2022).

Example 4 *The* ASP-Solver *ASPARTIX is an example of such an argumentation-driven reasoning engine. Starting from an AF as input, the Answer-Set-Programming solver will output the result of the specified reasoning task given a particular semantic (both encoded as ASP rules).*

It is worth mentioning that most of these engines also embed a planning component, which derives from their underlying employment of the AF formalism. Indeed, computing acceptable arguments enables 'argumentative paths' that lead to the achievement of the predetermined goal by deciding among possibly multiple options. Following edges that connect

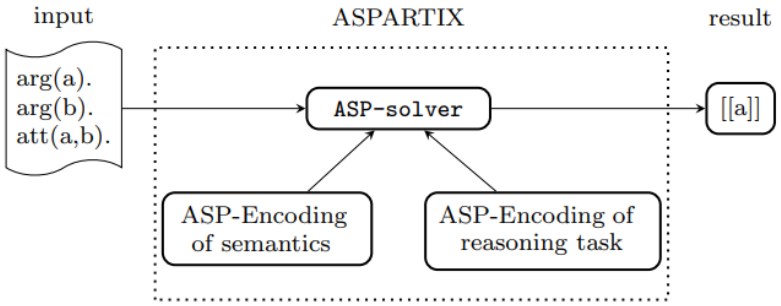

Figure 3: Example of an argumentation reasoning engine architecture (Dvořák et al., 2020).

justified nodes in an AF will exclude any potential rebuttals, thus ensuring a successful strategy. That is to say, each reasoning step, enclosed and rendered as an argument, is performed whilst having in mind the overall plan required for reaching a consistent decision.

2.1.4 Argumentation-based Dialogues

The view of computation as distributed cognition and interaction contributed to the rise of the multi-agent systems paradigm, where agents are intended as software entities capable of flexible autonomous action in dynamic and unpredictable domains (Luck et al., 2005). As a means of communication between such intelligent agents, formal dialogues were chosen due to their potential expressivity despite still being subject to specific restrictions (McBurney & Parsons, 2009). Argumentation-based dialogues are *rule-governed interactions* among participants (i.e. agents with their own beliefs, goals, desires and a limited amount of information regarding the other players) that take turns in making utterances. As shown in Table 1, these dialogues are usually categorized according to elements such as information possessed by the participants at the commencement of the interaction, their individual goals, and the knowledge and goals they share with other agents (Walton & Krabbe, 1995).

Dialogue type	**Description**	**Example**
Information-seeking	X seeks the answer to some question(s) from Y.	(Hulstijn, 2000)
Inquiry	X and Y collaborate to answer some question(s).	(Black & Hunter, 2007)
Persuasion	X seeks to persuade Y to accept a proposition.	(Prakken, 2006)
Negotiation	X and Y bargain over the division of some scarce resources.	(McBurney et al., 2003)
Deliberation	X and Y collaborate to decide what actions should be adopted.	(McBurney et al., 2007)
Eristic	X and Y quarrel verbally to vent perceived grievances.	(Blount, 2018)
Verification	X wants to verify the beliefs of Y.	(Cogan et al., 2005)
Query	X challenges Y since it is interested in Y's arguments.	(Cogan et al., 2005)
Command	X tells Y what to do.	(Girle, 1996)
Education	X wants to teach Y something.	(Sklar & Parsons, 2004)
Chance discovery	Ideas arise out of exchanges between X and Y.	(McBurney & Parsons, 2001)

Table 1: Description of existing dialogue types.

The selection and transitions between different dialogues can instead be rendered via a *Control Layer* (McBurney & Parsons, 2002; Sklar & Azhar, 2015), defined in terms of *atomic dialogue types* and *control dialogues*. The latter are meta-structures that have as their topics other dialogues and contribute to the management of the protocols' combinations and their transitions.

In general, the main components of argumentation-based dialogues can be identified as: (i) *syntax*, which handles the availability of and interaction between utterances; (ii) *semantics*, which differs according to the specific focus and final deployment of the dialogue; and (iii) *pragmatics*, which accounts for those aspects of the language that do not involve considerations about truth and falsity (e.g. the illocutionary force of the utterances) (McBurney & Parsons, 2013).

2.2 Chatbots

A chatbot must be able to parse the user input and interpret what it means before providing an appropriate response or output (and thus starting a 'chat'). The way in which the bot elaborates the replies to be delivered depends upon its *response architecture model*. Following the studies conducted in (Adamopoulou & Moussiades, 2020; Singh & Thakur, 2020; Klopfenstein et al., 2017; Codecademy, 2022), we can classify such models as:

- **Rule-based** chatbots employ the simplest response architecture model. The bots' replies are entirely predefined and returned to the user according to a series of rules. The internal structure of such rule-based software can be thought of as a decision tree that has a clear set of possible outputs defined for each step in the dialogue. Usually, this category of conversational agents handles those kinds of interactions where the user has a number of pre-compiled options to choose from. As an example of rule-based colloquial agents, we can consider ELIZA (Weizenbaum, 1966): deemed by scholars as the first implementation of a chatbot, it operates by harnessing linguistic rules in combination with recognized keywords from the users' inputs. Further development in the area resulted in PARRY (Colby et al., 1971), a chatbot that improved ELIZA via a conversational strategy embedded to simulate a person with paranoia. Jabberwacky (Carpenter, 1982) is also an instance of a rule-based bot that interacts through contextual pattern matching. It steadily expands its database by collecting tokens from previous conversations that occurred with different users.

- **Retrieval-based** chatbots represent a more complex response architecture model. The bots' replies are pulled from an existing corpus of stored sentences. Machine Learning and Natural Language Processing (NLP) models are used to interpret the user input (operation divided into *intent classification* and *entity recognition*) and determine the most fitting response to retrieve. As an example of retrieval-based colloquial agents, we can consider A.L.I.C.E. (Wallace, 2009) developed using the Artificial Intelligence Markup Language (AIML) (Wallace, 2003). Such a language comprises a class of data objects and partially describes the behaviour of computer programs that process them via stimulus-response templates. Furthermore, also IBM's Watson Assistant (IBM, 2006) and Microsoft's Cortana (Microsoft, 2014) represent other instances of the retrieval-based architecture. The first parses input to find statistically

relevant replies in its database by means of parallel algorithms. The second instead leverages the natural language processing capabilities of Tellme's Network (owned by Microsoft from 2007) and the Satori knowledge repository to provide responses (Marshall, 2014).

- **Generative** chatbots represent the most sophisticated response architecture model. These bots are capable of formulating their own original responses based on the user input rather than relying on existing text. The deployment of Deep Learning structures allows returning the appropriate response by calculating the likelihood of the next element(s) in a word sequence. However, training such models requires time, and it is not always clear what is used to produce replies, which may be repetitive or nonsensical. In addition, generative bots are not generally capable of accessing data other than what is embedded in their model parameters. One common approach to mitigate these problems is to combine both retrieval and generative operations in the chatbot (Roller et al., 2020). As an example of such a hybrid type of virtual assistant, we can consider Apple's Siri (Apple, 2011) and Amazon's Alexa (Amazon, 2014; Lopatovska et al., 2019). Both provide replies to users' questions (along with an additional wide array of possible functions) via Deep Learning procedures or delegating requests to a set of external providers, e.g., WolframAlpha (Heater, 2018).

  Generative-LLMs. Generative chatbots that hinge upon Large Language Models (LLMs) deserve special mention, given recent interest in such models. The design and deployment of the Transformer architecture (Vaswani et al., 2017) determined a paradigm shift towards 'pre-training' and 'fine-tuning' learnings (Zhao et al., 2023): scaling up pre-trained models led to the discovery of LLMs and their impressive capabilities (Brown et al., 2020; Touvron et al., 2023a, 2023b; Anil et al., 2023; Reid et al., 2024; Meta, 2024) . Leveraging these new technologies, conversational agents such as the famous ChatGPT[3] prove to outperform most of the previous benchmarks and predecessors in information extraction tasks (Li et al., 2023), natural language inference, question answering, dialogue tasks (Qin et al., 2023) and machine translation (Jiao et al., 2023). That being said, LLMs and the chatbots based on them also suffer from a number of downsides (Frieder et al., 2023; Bang et al., 2023; Wei et al., 2022a; Ji et al., 2023; Zhuo et al., 2023) including: faulty reasoning, inexplicable appearance of previously unknown abilities (a phenomenon denoted as *emergent abilities*[4]), nonsensical or unfaithful replies (i.e. *hallucination*), biased and toxic communications, expensive training costs and a high carbon footprint[5]. Finally, it has also been shown how underlying models such as GPT-3 (Brown et al., 2020) fall short of producing adequate and compelling arguments (Hinton & Wagemans, 2022). However the outputs of such models may prove particularly suited to support argument mining operations,

3. Other remarkable examples are DialoGPT (Zhang et al., 2019), BlenderBot 3x (Xu et al., 2023), Gemini 1.5 (`https://gemini.google.com/`), Claude 3.5 Sonnet (`https://claude.ai/`), Llama 3.1-Instruct (`https://www.meta.ai/`), Mistral-7b-Instruct (Jiang et al., 2023), and Zephyr-7b (Tunstall et al., 2023).
4. Emergent abilities constitute a controversial topic and some studies even argue against their existence (Schaeffer et al., 2023).
5. Although it has been argued that the adoption of best practices in model training should reduce carbon dioxide emissions by 2030 (Patterson et al., 2022).

given carefully conditioned (or an increased number of) inputs (de Wynter & Yuan, 2023; Chen et al., 2023).

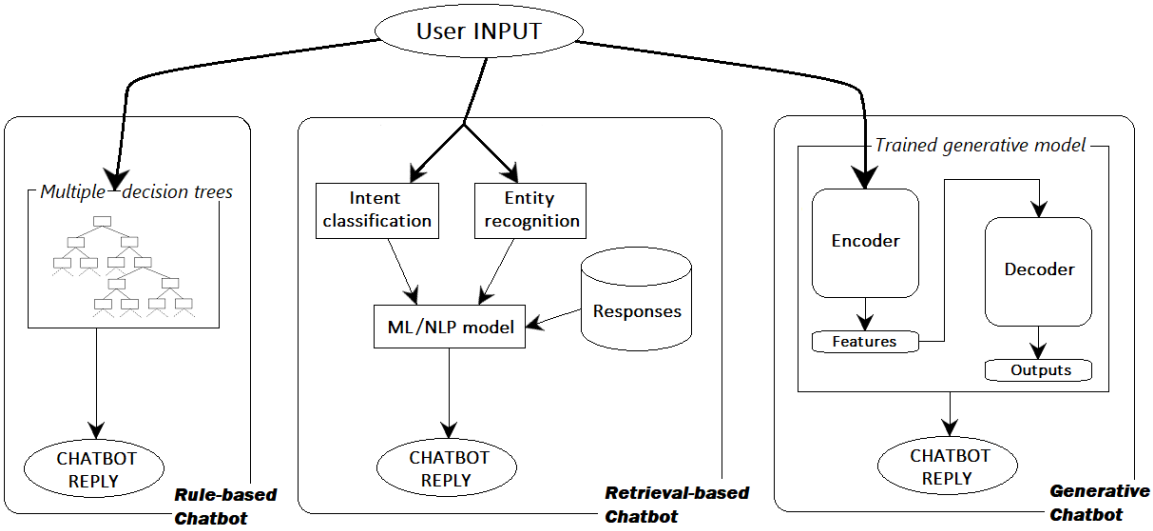

Figure 4: Comparisons of different response architecture models. Observe that generative chatbots are often presented in a decoder-only form and, more rarely, leverage an encoder-only form.

The different chatbot response architecture models and the corresponding high-level operations that characterise them are depicted in Figure 4. Notice, as previously anticipated, that is quite common for conversational agents to use a combination of different response models in order to produce optimal results. Furthermore, chatbots can be classified based on the conversation topics they are able to cover. *Closed domain* ones (e.g. bots focused on customer assistance or e-commerce) are restricted to providing responses within a particular matter. Due to their specific area of competence, usually, these agents are very efficient in delivering good-quality replies. On the other hand, *open domain* chatbots, e.g. the previously referenced Apple's Siri, Amazon's Alexa, Meta's Llama 3.1, Google's Gemini and OpenAI's ChatGPT, as well as Meena (Adiwardana et al., 2020), Mitsuku (Worswick, 2018) and Microsoft's XiaoIce (Zhou et al., 2020), should be able to explore any range of conversation topics, similar to how a real-world human-to-human interaction would be. However, it is not straightforward to implement such bots, and they prove to be more prone to errors, incoherent responses[6] or other issues similar to the aforementioned generative-LLMs.

2.2.1 The Knowledge Base Acquisition

Chatbots cannot automatically generate responses unless they are provided with a specific knowledge base from which those replies can be acquired. This limitation involves every type of conversational agent, not only the retrieval-based, as one may think. Indeed, rule-based architecture requires hard coding of data into the scripts of the chatbot, whereas generative

6. Notice these errors can have extreme and harmful consequences, such as a medical chatbot suggesting a patient kill themselves. (Daws, 2020)

models necessitate a corpus of information to be trained upon. However, extensive data collection is needed to obtain such knowledge bases, a procedure which is often costly and requires significant effort. In particular, anticipating a topic covered in the next sections, some argumentation-based chatbots are characterized by a knowledge base consisting of a set of arguments (alternatively, an argument graph) to collect which current approaches include argument mining from documents (e.g., Cocarascu et al., 2019; Trautmann et al., 2020) or hand coding of texts by researchers (e.g., Cerutti et al., 2016; Rosenfeld and Kraus, 2016). These operations can be complicated tasks to achieve, especially if we need to handle only real-world arguments rather than synthetic ones. That is to say, it may be difficult to retrieve high-quality arguments concerning a specific topic on the web, or it may be problematic to distinguish between the person (and, thus, account for her attributes) who posited a specific claim. Questionnaires or personal interviews may provide a solution, although such solutions are expensive and require a large amount of human effort. Interestingly, studies such as (Chalaguine & Hunter, 2018; Chalaguine et al., 2018) proposed an alternative method to face this potential issue. The results of their research show how a chatbot, with little to no domain expertise, may elicit arguments and counterarguments from different users, thus automating the process of argument acquisition. This procedure, called *argument harvesting* by the authors, allows for the generation of AFs that incorporate the knowledge base information over the required domain. Another alternative approach is provided by the work conducted in (Chalaguine & Hunter, 2019) that describes how to acquire a large number of (high-quality) arguments in a graph structure using crowd-sourcing.

3. Methodology

This survey hinges upon the collection and review of papers concerning argumentation-based chatbots. Before delving into the examination of our findings, it may be helpful to provide an uncontroversial definition of the subject of our investigation:

Definition 3 (Computational Argumentation-based Chatbots) *We consider* computational argumentation-based chatbots *those conversational agents that employ argumentative models to: (i)* extract *textual data via argument mining tools, (ii)* structure *information by means of argumentative templates, (iii)* reason *with argument semantics and/or (iv)* deliver *replies to users through argumentation-based dialogues.*

A schematic representation of the argumentation employment types within a conversational agent architecture is provided in Figure 5. Here it is specified the level at which each aspect operates in the overall chatbot design. Argument mining enables the construction of a database for model training or a knowledge base (KB) by *extracting* information from texts. KB data can be *structured* into argumentative patterns, which may then be *delivered* as argumentation-based dialogue replies to the interacting end-user after being selected through a *reasoning* step (that usually involves argument semantics computation). Notice that we strictly selected only papers involving such aspects, avoiding any other articles pertaining to chatbots or meanings of argumentation that differ from those introduced in Section 2. For example, we did not include the work of (Toniuc & Groza, 2017) among the surveyed papers since it does not account for computational argumentation as described herein (albeit the presented *textual entailments* relationship may be transformed into a

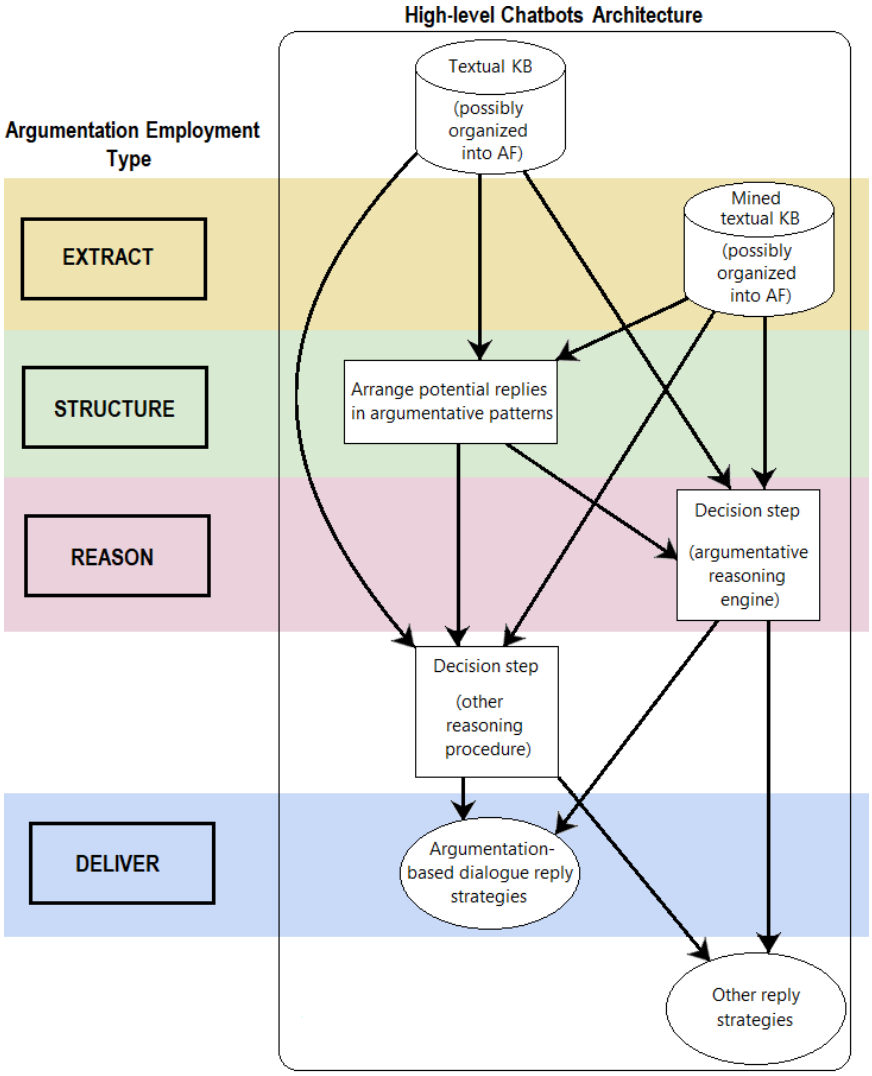

Figure 5: Schematic argumentation employment within the chatbot architecture.

form of premises-conclusion argument dependency). A similar issue can be observed in (Altay et al., 2022; Kulatska, 2019) where, although there is a reference to a general notion of arguments and counterarguments, it does not correspond to the one provided in Section 2. On the other hand, we also excluded research such as (Chalaguine et al., 2018; Chalaguine & Hunter, 2018) since their focus is more on the automated collection of a corpus of arguments and counterarguments rather than the implementation of a conversational agent that delivers argumentation-based dialogues. Furthermore, the pipeline delineated by (de Sousa et al., 2024), which outlines an approach that makes use of chatbot technologies to ameliorate the engagement of virtual systems with more sophisticated topics, does not qualify as an argumentation-based bot either. That is because the overall method is essentially equal to a classifier for a predetermined set of argument schemes rather than a fully interactive conversational agent.

To clarify, in our survey, we did not restrict the search according to particular chatbot types, or their final scope, nor did we distinguish between different bot denominations (e.g. 'argumentative dialogical agent', 'dialogue manager', 'automated persuasion system') or maturity of implementation, (e.g. fully-fledged or just sketched). Also, we did not account for specific time ranges and gathered articles independently of the year of publication. We have then analysed and organized the results in one concise comparative table (Table 2) that displays the classifications and main features of each conversational agent. In particular, we listed all the reviewed chatbots and distinguished between each bot's final purpose (e.g. persuade, explain, inform), response architecture model (the prevalent one is recorded in case of multiple models), and conversation domain (for which we mostly considered the topic specified in the corresponding paper examples). Additional data comprise also the way in which computational argumentation has been employed within the chatbot architecture (i.e. extraction, structure, reason, deliver). Finally, we inspected the arranged information and discussed our main findings.

4. Argumentation-based Chatbots

This section covers a concise description of all the reviewed chatbots according to their specific argumentation employment. We first outline each argumentation-based category before providing an account of the conversational agents pertinent to the class. Note that it may be the case for a bot to present components that fulfil specific tasks (e.g. extract, structure, reasoning, and deliver) without exploiting computational argumentation. The fact that we are not detailing such components does not undermine their presence or effectiveness but reflects the choice of strictly conferring an argumentative scope to the survey. Additionally, observe that when a chatbot falls within multiple argumentation employment categories, our description will focus on the one that primarily represents its unique features. We conclude by highlighting the evaluations of such chatbots (if any) as presented in their respective papers.

4.1 Argumentation-based Extraction

Starting from a corpus of natural language texts, argument mining procedures allow for the extraction of arguments, and the classification of their relations, within such documents. The mined data can then be further processed and organized in AFs[7], or simply be employed as replies according to the user's input. Unlike the latter, the former choice may lead to a reasoning operation upon the framework that will elicit specific output depending on the evaluation criteria of the captured semantics. For example, ADA, the argumentative dialogical agent introduced in (Cocarascu et al., 2019), extracts arguments from movie review snippets and mines the relations subsisting among them. The acquired data is then utilized to construct a Quantitative Bipolar argumentation framework (QBAF) experimentally eval-

7. We stipulate that constructing an AF from the utterances of an argumentation-based dialogue does not qualify as an 'argumentation-based extraction'. That is because the arguments and the attacks (respectively supports) are already given and do not require further parsing.

Paper(s)	Final Purpose	Response Architecture	Conversation Domain	Argumentation Employment
(Slonim et al., 2021)	Debate	Retrieval-based	Unspecified (semi-open domain)	Extract
(Galitsky, 2019) (Galitsky, 2018) (Galitsky, 2020)	Unspecified	Retrieval-based	Unspecified (closed domain)	Extract
(Wambsganss et al., 2021)	Explain	Unspecified	Unspecified (closed domain)	Extract
(Cocarascu et al., 2019)	Explain	Rule-based	Movie reviews	Extract, Reason Structure
(Bistarelli et al., 2021)	Converse	Retrieval-based	Unspecified (closed domain)	Reason
(Castagna et al., 2022, 2023)	Explain	Retrieval-based	Healthcare	Reason, Structure
(Rago et al., 2020)	Explain	Rule-based	Movie reviews	Reason, Structure
(Sassoon et al., 2019)	Explain	Retrieval-based	Healthcare	Reason, Structure, Deliver
(Dignum & Bex, 2017)	Converse	Retrieval-based	Healthcare	Reason, Deliver
(Fazzinga et al., 2021)	Inform	Retrieval-based	COVID-19 vaccine	Reason, Deliver
(Bex et al., 2016)	Inform	Retrieval-based	Fraud report	Reason, Deliver
(Rosenfeld & Kraus, 2016)	Persuade	Retrieval-based	Benefits of holding a Master's Degree	Reason, Deliver
(Sklar & Azhar, 2015) (Sklar & Azhar, 2018)	Explain	Retrieval-based	Treasure hunt game	Deliver
(Chalaguine et al., 2019)	Persuade	Retrieval-based	Meat consumption	Deliver
(Chalaguine & Hunter, 2020)	Persuade	Retrieval-based	UK university fees	Deliver
(Chalaguine & Hunter, 2021)	Persuade	Retrieval-based	COVID-19 vaccine	Deliver
(Hadoux & Hunter, 2019)	Persuade	Retrieval-based	Cycling in the city	Deliver
(Hadoux et al., 2021)	Persuade	Retrieval-based	UK university fees	Deliver
(Andrews et al., 2008)	Persuade	Retrieval-based	Desert survival	Deliver
(Guo et al., 2022)	Persuade	Retrieval-based	Nuclear energy	Deliver
(Hauptmann et al., 2024)	Converse	Retrieval-based	AI ethical challenges (closed domain)	Deliver

Table 2: Argumentation-based chatbot specifics.

uated against three gradual semantics: QuAD (Baroni et al., 2015), DF-QUAD (Rago et al., 2016) and the Restricted Euler-based semantics (Amgoud & Ben-Naim, 2018). Leveraging such a QBAF, the conversational agent will instantiate the reply templates stored within its system. Those replies will thus be delivered (no argumentation-based protocol is involved) to the interacting user when prompted for explanations about the selected movie recommendation.

Another example of an argumentation-based extraction chatbot is rendered by the conversational agent developed in (Slonim et al., 2021) whose purpose is to challenge humans with competitive debates. After having preprocessed a corpus of 400 million newspaper articles in order to create an index of meaningful concepts, the bot mines for arguments thus obtaining claims and evidence related to the selected dispute. In this process, the agent identifies the relations occurring between the mined arguments and takes advantage of these data to prepare counterarguments against different stances on the debate topic. The replies posited by the bot will then be retrieved among the mined arguments, or the

ones stored in a more general knowledge base, via a neural model. Notably, the interaction with the user occurs on a speech base and the speech-to-text conversion is performed by IBM's Watson[8].

A borderline case is constituted by the ArgueBot conversational agent (Wambsganss et al., 2021). Developed as a learning tool for providing adaptive feedback on students' logic argumentation, ArgueBot (a bot deployed within the Slack platform[9]) hinges on a BERT (Devlin et al., 2018) classifier to perform AM operations on the user's textual input before providing tailored comments on their argumentative writing. Although the chatbot may be equipped with specific reply templates, its exact response architecture is unclear and remains unspecified by the authors.

Finally, on a more abstract level, the research discussed in (Galitsky, 2020, 2019, 2018) describes the deployment of specific argument mining approaches to chatbots. Here, the conversational agent constructs a communicative discourse tree from a subset of text by matching each fragment of the subset that has a verb to a verb signature. The subsequent application of classification models allows the bot to detect arguments and their relations and then leverage that information to provide replies according to the user input. In a nutshell, by resorting to in-depth rhetorical analysis, the chatbot accounts for multiple features of the argument (e.g. embedded affective aspects, consistency with the domain clauses, etc.), which results in more precise user-bot reply matches.

4.2 Argumentation-based Reply Structures

Chatbot replies can be structured according to the traditional argumentative format: a claim derived from a set of premises by means of particular inference rules. This approach includes argument schemes and general frameworks for structured argumentation (e.g. ASPIC$^+$, ABA, etc.). In general, the organization of data within such an argumentative pattern occurs before the generation of an AF and the computation of its semantics. However, it may also be convenient to arrange the bot responses using specific templates, regardless of a further semantic evaluation. Indeed, providing replies with a precise structure serves to highlight the rationale underpinning the argument claim and enhance the overall clarity of the discourse.

As an example, we can consider the conversational agent presented in (Castagna et al., 2022, 2023), which may be seen as the final implementation of previous versions described in (Essers et al., 2018; Kökciyan et al., 2019; Balatsoukas et al., 2019; Chapman et al., 2019; Balatsoukas et al., 2020; Sassoon et al., 2020; Kökciyan et al., 2021; Drake et al., 2022). Harnessing the novel Explanation-Question-Response, or EQR, argument scheme (first envisaged as a dialogue protocol and fully-fledged by Castagna et al., 2024), this bot delivers tailored justified recommendations within the healthcare domain, helping users self-manage their conditions. These recommendations embed an additional layer of information: the rationale behind the instantiated scheme acceptability (i.e. its evaluation, automated via the ASPARTIX engine by Egly et al., 2008, according to the considered argumentation framework). Additional replies provided by the chatbot are then structured by harnessing

8. Once again, recall that we are emphasising the elements leveraging computational argumentation. Project Debater (Slonim et al., 2021) is a fully-fledged debating system, nonetheless, its employment of AM procedures is the only argumentation-related component, and this is why it is the one described.
9. `https://slack.com/`

the argument scheme (and respective CQs) templates instantiated by the bot's knowledge base. Another example is provided by the interactive recommender system delineated in (Rago et al., 2020), a partial extension of the work undertaken by (Rago et al., 2018), which clarifies its movie recommendations through argumentative explanations organised according to a series of predetermined premise-conclusion textual templates. The instantiations of such patterns ensue from the selection of explanations drawn from the Bipolar AF that embeds its underlying knowledge base. Notice, however, that the reasoning procedure underscoring the selection is only vaguely described in terms of argumentation semantics.

4.3 Argumentation-based Reasoning

As previously discussed, an argumentation engine can be employed as the underlying tool that drives a chatbot's *reasoning* operations. In such a circumstance, regardless of the chosen framework, e.g. Abstract AFs (Dung, 1995), Bipolar AFs (Cayrol & Lagasquie-Schiex, 2005), Weighted Bipolar AFs (Rosenfeld & Kraus, 2016), Quantitative Bipolar AFs (Cocarascu et al., 2019), Metalevel AFs (Kökciyan et al., 2021), etc., most of the decision-making processes involve the computation and semantic evaluation of the AF. Intuitively, starting from a knowledge base embedded in a set of arguments, the bot executes a reasoning procedure that usually results in a selection of acceptable arguments (which changes depending on the chosen semantics). When interacting with the user, the conversational agent will retrieve its replies, based upon the received input from its interlocutor, from the computed acceptable arguments. As such, we can generally assume that argumentation-based reasoning engines are intertwined with retrieval-based response architectures or hybrid models that include retrieval-based operations. For example, ArguBot (Bistarelli et al., 2021b), developed using Google DialogFlow[10], employs ASPARTIX (Egly et al., 2008) to compute arguments from an underlying Bipolar AF, to support (*pro-bot*) or challenge (*con-bot*) the user's opinion about the topic of dialogue.

The conversational agent presented in (Fazzinga et al., 2021)[11] retrieves its arguments from an underlying Bipolar AF as well, although it follows the semantics illustrated in (Fazzinga et al., 2018). The selected reply is, therefore, an argument acceptable with respect to an admissible extension computed over the overall framework, thus providing a strategy that also accounts for future developments of the chat. In addition, the bot is capable of formulating on-demand explanations about a particular reply, i.e. a sequence of natural language sentences that describes the facts supporting it, along with motivations against other possible conflicting arguments that the system discarded.

In contrast, the chatbot outlined in (Dignum & Bex, 2017) deploys computational argumentation as a means of evaluating completed phases of the ongoing dialogue, rather than starting with a previously generated AF. More precisely, an argument graph is constructed by incorporating the facts that emerge during the dialectical interaction with the user. Then, a formal assessment occurs by checking if those facts are members of acceptable extensions of the graph. Interestingly, this conversational agent harnesses social practices theory (Reckwitz, 2002; Shove et al., 2012) to contextualise the conversation and provide useful background information that facilitates the user's input interpretation. A similar

10. `https://cloud.google.com/dialogflow/docs/`
11. Subsequently embedded into a privacy-preserving dialogue system (Fazzinga et al., 2022).

deployment of computational argumentation is envisaged in (Bex et al., 2016), where an AI system that enhances the online report of trade frauds is outlined. A chatbot (the 'dialogue manager') exchanges arguments with the user parties (both fraud victims and police) eliciting, if needed, more information about the ongoing case whilst building a knowledge graph. The acquired data will then enable the matching of the graph with a typical criminal scenario known by the police. Subsequently, formal argumentation semantics will drive the reasoning with scenarios and pieces of evidence, i.e. the 'hybrid theory' (Bex et al., 2010; Bex, 2015).

Finally, the conversational agent (SPA) envisaged in (Rosenfeld & Kraus, 2016) also employs an argumentation-based reasoning engine. In particular, it embeds its knowledge base into a Weighted Bipolar AF (WBAF) and computes the argument that maximizes the framework evaluation function according to the user input. The score returned by the valuation function represents the reasoner's ability to support that argument and defend it against potential attacks. The dialectical interaction with the user follows a strategical persuasion dialogue protocol, optimized via Monte Carlo Planning (Silver & Veness, 2010), that might involve updating the argumentation frameworks of both the persuader and the persuadee.

4.4 Argumentation-based Reply Delivery

Chatbots may handle and deliver their responses to the user interacting with them by leveraging the protocols of argumentation-based dialogues. Harnessing the dialogue logic, the conversational agent can optimize its strategy and utter only the arguments that prove to be necessary for achieving its final goal. In a way, we could identify the delivery phase as a 'secondary reasoning step' where the bot chooses which arguments to move (strictly following the involved dialogue protocol instructions) among those available (possibly previously computed by the 'primary engine' described by the reasoning phase). Notice that the arguments licensed in a dialogue protocol follow a more flexible definition than the standard ones provided in the abstract or structure argumentation approach: *" [...] it is the idea of dialogue as an exchange between two or more individuals, an exchange which captures features of what would be informally called an "argument". That is, dialogue as the exchange of reasons [i.e. arguments] for or against some matter"*(Black et al., 2021).

As an example, we could examine the work introduced in (Hadoux et al., 2021), which expands upon (Hadoux & Hunter, 2019; Hunter, 2018; Hunter et al., 2019) and depicts an overall framework for modelling beliefs and concerns in a persuasion dialogue. An implementation of such a framework is then envisaged via an automated persuasion system (APS), a software application aiming at convincing the interacting agent to accept some arguments. Following the asymmetric persuasion dialogue protocol illustrated therein (i.e. unlike the system, the user is restricted in choosing replies among the provided options), the proposed chatbot proves to be capable of identifying, within its knowledge base embedded in an argument graph, the most appropriate argument to posit. Essentially, the APS performs a Monte Carlo Tree Search coupled with a reward function to maximize the addressing of concerns (paired with the arguments of the graph) and the user's beliefs.

Similarly, the bot presented in (Chalaguine & Hunter, 2020) aims at persuading the interlocutor via a free-text interaction where the user's inputs are matched (by vector ren-

dering and cosine similarity) with the (crowdsourced) arguments of the graph representing the knowledge base. The chatbot trains a classifier to detect the most common concerns of the persuadee and employs it to select counterarguments that will produce a result more compelling than a random choice. If no argument similarity is detected, the conversational agent will resort to a default reply based on the user's concerns. Furthermore, the same authors presented an analogous architecture for a persuasion bot in (Chalaguine & Hunter, 2021), with the addition of a particular concern-argument graph. By incorporating the knowledge base within such a small graph, it can be proved that no large amount of data is needed to generate effective persuasive dialogues. Interestingly, a preliminary analysis of the impact (appeal) of arguments addressing the users' concerns in a persuasion dialogue performed by a chatbot has also been conducted by the same authors in (Chalaguine et al., 2019). Another example of such a concern-based approach may be represented by Argumate, a chatbot designed to facilitate students' production of persuasive statements (Guo et al., 2022). To provide appropriate suggestions, the bot retrieves its replies from an underlying argument graph, whose edges denote attack and support relations, via a concern identification method. Notice that the interactions between Argumate and the users occur both by typing and selecting predefined options.

A common trait amongst all of the above argumentation-based conversational agents is that, although the corpus from which they extract their replies is organized as an argument graph, there is no interest in any particular acceptable semantics. That is to say, the knowledge base is organized and considered as a plain AF, where arguments and attacks are the only relevant features. In addition, most of these studies also account for a baseline chatbot which exploits a random strategy for selecting counterarguments from the available choices within the underlying knowledge base. The reason for this is to provide a means for comparing the developed bots which employ more fine-grained strategies for choosing their replies.

Finally, one last conversational agent that focuses on the delivery of persuasion dialogues is the chatbot designed in (Andrews et al., 2008). Implemented harnessing the AIML markup language (Wallace, 2003), the bot comprises a planning component that searches over an argumentation model for the optimal dialectical path to pursue in order to persuade the user. The agent records the user's beliefs and updates this information whenever its interlocutor agrees/disagrees during the interaction. Such belief revision plays an important role in the strategic view of the chatbot. Moving towards different topics, the conversational agent implemented in (Sassoon et al., 2019), within the context of explanation for wellness consultation, exploits multiple dialogue protocols (i.e. persuasion, deliberation and information seeking) whilst exchanging instantiations of acceptable argument schemes with its interlocutor. The adoption of diversified dialogue protocols (i.e. persuasion, inquiry and information seeking) characterises also the chatbot-equipped robot proposed in (Sklar & Azhar, 2015) and demonstrated in (Azhar & Sklar, 2017). Retrieving the most appropriate argument constructed from its beliefs, an operation facilitated by the restricted options available to the user, the robot communicates with its human interlocutor in order to strategize about a treasure-hunting game. We conclude the list with the bot introduced in (Hauptmann et al., 2024). This German-language conversational agent, following the formalisation of (Hadoux & Hunter, 2019), makes use of an argument graph to encode its knowledge base from which it retrieves main stances and counterarguments

to engage the users in discussions concerning the ethical challenges of AI implementations. The delivery strategy is somehow ambiguous but seems to balance a mixture of persuasion and information-seeking, according to the specific stage of the conversation.

4.5 Evaluation of the Chatbots

Thus far, we have described the reviewed argumentation-based chatbots, primarily focusing on their features in relation to argumentation employment. However, some of those conversational agents have also been evaluated via specifically designed user studies[12] whose results will be reported herein. For example, the virtual debater devised in (Slonim et al., 2021) exhibits a higher discussion quality than the compared artificial competitors, although it still fails to achieve a human-like level. Furthermore, (Balatsoukas et al., 2020) reported on the findings ensuing from the pilot study designed to assess a former version of the CONSULT system, which later informed the chatbot deployed in (Castagna et al., 2022, 2023). The outcome was a criticism concerning a lack of a more natural conversation flow when interacting with the bot. User studies have also been conducted to test the human-robot interaction presented within the ArgHRI system of (Sklar & Azhar, 2015; Azhar & Sklar, 2017). The results showed how argumentation-based dialogues contribute to enhancing trust towards the robots. Nonetheless, analysis of the dialogues themselves (Sklar & Azhar, 2018) highlighted how the possibility of interrogating the bot to obtain explanations did not lead to a significant increase in performance from the human-robot team, nor a boost in user satisfaction.

On the other hand, the SPA conversational agent introduced in (Rosenfeld & Kraus, 2016) outplayed the baseline chatbot (which harnessed a different, heuristic, strategy) when tested in its persuasion task, thus proving capable of delivering human-like level conversations. Similarly outperforming the baseline agent is the bot presented in (Chalaguine et al., 2019). Indeed, the paper includes an experiment that shows how such a chatbot, by positing arguments that address the users' concerns, is more likely to positively change the users' attitude in comparison with another agent that does not employ such a strategy. An analogous interest in users' concerns is encompassed in the study implemented in (Chalaguine & Hunter, 2020). The results, conjointly supported by the experiments in (Hadoux & Hunter, 2019) and confirmed by (Hadoux et al., 2021), conclude that a strategic chatbot accounting for concerns is more likely to provide relevant and cogent arguments. Moreover, it is also

12. A different (and outdated) way of evaluating the capability of a conversational agent would be through a discussion with a human end-user: the more natural and seamless the interaction, the more effective the chatbot. The Turing Test (or *Imitation Game*) is a proposal advanced by Alan Turing (Turing & Haugeland, 1950) whose idea was to present some sort of test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human. Hinging on the Imitation Game, the Loebner Prize is a contest started in 1980 to award computer programs that are the most human-like, i.e. that perform the best in the Turing Test. The winner of the contest is the one that tricks a judge the highest percentage of the time, and Mitsuku is the chatbot that won the largest number of such prizes (Worswick, 2013). The Loebner competition (considered defunct since 2020) has been subjected to a long list of criticisms. Among these was the alleged idea that entrants do not aim at understanding humans since deception and pretence are highly rewarded in this contest. Another criticism leveled against the Loebner Prize is that it confuses the Imitation Game with *proof of human-like intelligence*. However, machines cannot reason like humans, as claimed by Searle in 1980 with his famous 'Chinese Room experiment' (Searle, 1980; Cole, 2020).

worth mentioning the evaluation outcome of the other two persuasive agents presented in (Andrews et al., 2008; Chalaguine & Hunter, 2021). The former bot provides fluent conversations with its interlocutors performing generally better than a purely task-oriented system. The latter, instead, shows how an interactive chatbot yields more compelling information than a static webpage.

As another example, consider how the ArgueBot conversational agent underwent both quantitative and qualitative assessments (Wambsganss et al., 2021). The data collected from detailed feedback and Likert scale post-experiment forms yielded positive results. In particular, the participants perceived the chatbot as helpful, useful and easy to interact with. Resorting to pre and post-dialogues Likert-scale questionnaires is also the evaluation choice preferred in (Hauptmann et al., 2024). The results record successful shifts of the opinions of 40-50% of the participants after engaging with the chatbot. Overall, the users acknowledged the arguments' quality and the design of the conversational system. Lastly, we report also the experimental study outcome of the recommender system delineated in (Rago et al., 2020). The tentative conclusions seem to highlight how argumentative explanations improve trust and transparency towards the system, although preferences about their content and delivery may differ from one user to the other.

5. Discussion

Table 2 depicts an overview of our findings, with a quantitative summary of the sampled chatbots' features shown in Figure 6. As a first remark, it is surprising that argumentation-based chatbots are not well-represented in the literature. Indeed, the formal characterisation of real-world dialectical interactions provided by computational argumentation seems to be well-suited for agents whose role concerns conversing with users. This, however, may follow from the fact that the computational argumentation research field has experienced limited dissemination (especially outside of Europe) rather than deriving from the unsuitability of the argumentation formalism. Another possible explanation may be due to the fact that there has been an explosive interest in model-free methods in computer science in the last decades (Bringsjord & Govindarajulu, 2022), ignoring model-based methods (like computational argumentation), which are only now gaining favour again, for example, as a way of 'interpreting' the model-free output. Nevertheless, a number of considerations can be drawn from the outcome of our analysis. *Persuade* and *Explain* prove to be the most common goals of the examined chatbots. The latter stems from the recent interest in explainable AI and its link with computational models of arguments (Vassiliades et al., 2021; McBurney & Parsons, 2021; Čyras et al., 2021). Persuasion dialogues, instead, have been studied in papers such as (Hunter, 2015; Murphy et al., 2016), whose findings show how the use of argumentation-based formalisms may provide compelling strategies to induce belief change. One reason for such a number of persuasion-focused chatbots could indeed be related to the effectiveness of argumentation in delivering replies in such an area, as also advocated by the results of several user studies. To corroborate this, it can be noticed how persuasive conversational agents employ computational argumentation in such a way that falls under the (dialogical) *Deliver* category (which, as expected, turns out to be the most common class listed in Table 2). Observe also that the main features of such bots include the account of beliefs and concerns when positing cogent (argumentative) replies.

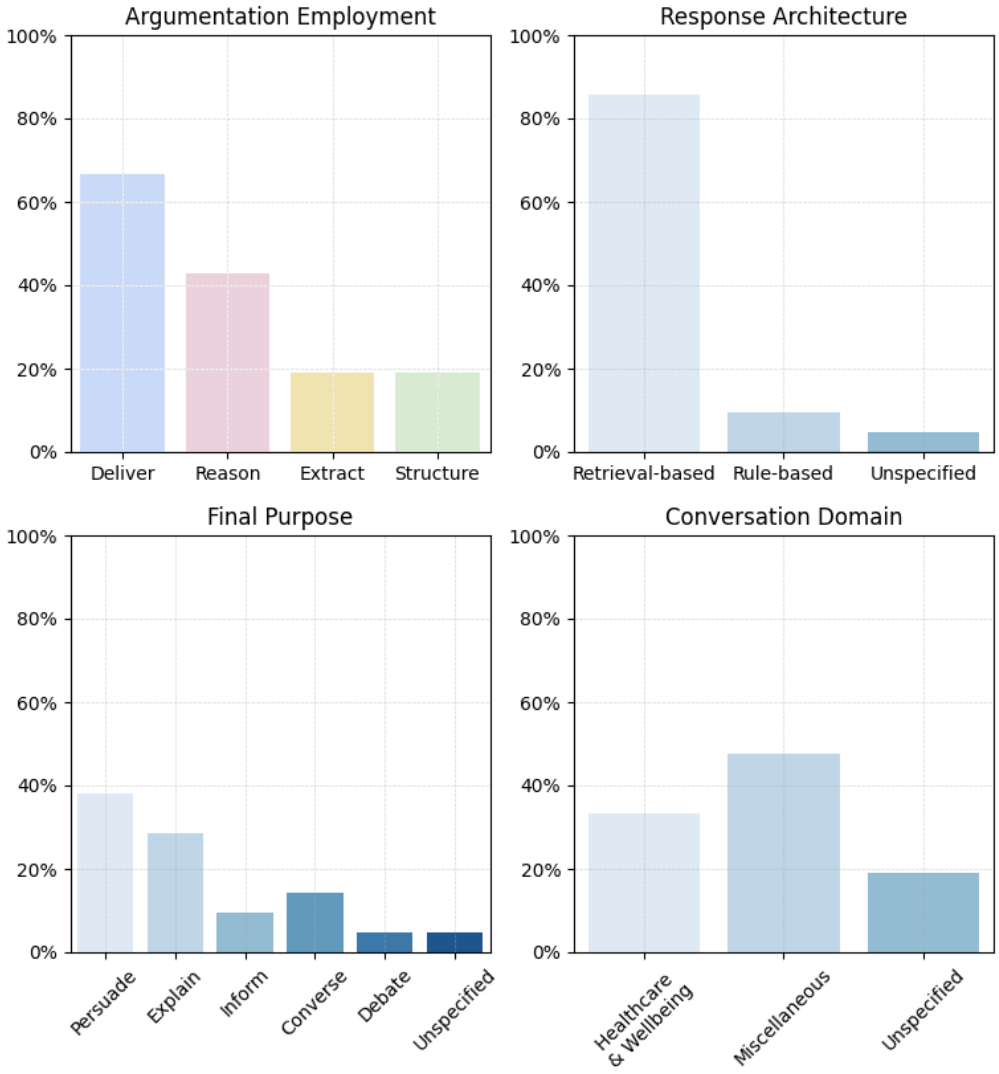

Figure 6: Percentage of sampled systems characterised by the argumentation employment type (top left), response architecture (top right), final purpose (bottom left) and conversation domain (bottom right) as described from the data of Table 2.

Continuing our analysis of different typologies of argumentation employment, it is worth emphasizing that *Structure* always appears together with *Reason* (though not vice versa), meaning that they are closely intertwined.

In general, it is less common that an argumentation-based chatbot employs argumentation solely for its reasoning engine. Indeed, after the semantics of the underlying AF have been computed, the bot usually leverages a dialogue protocol that handles the replies delivery. Speaking of the underlying argumentation framework, we realized that, when embedding a knowledge base into an AF, the Bipolar framework (and its variants QBAF and

WBAF) turns out to be the most popular option. This choice is related to the additional information provided by BAFs which encompass *support* relations rather than just *attacks*, allowing for an intuitive formalisation of both endorsements and conflicts between pieces of data.

Within our survey, we identified several conversation domains contemplated by the bots, ranging from *Healthcare* to *Nuclear energy*, with the former representing the prevailing domain (and also subsuming others). Notice that 'unspecified domain' could mean either that no conversational topic has been specified or that a sketched list of multiple topics has been presented. Interestingly, there is no argumentation-based chatbot eligible to be considered as open domain, although we might regard as 'semi-open domain' the agent discussed in (Slonim et al., 2021). Indeed, despite the absence of topic limitations in its debate delivery (due to a huge corpus upon which arguments are retrieved), the bot is not capable of handling small talk or other analogous trivial interactions. This also affects its discussions, each of which is modelled as a challenge towards opposite stances. Another peculiarity of the agent engineered by (Slonim et al., 2021) is that it allows for unconstrained speech in user input, whereas most chatbots only allow for free-text input, and the bots envisaged in (Bex et al., 2016) and (Guo et al., 2022) combines both free and limited textual prompts. Nonetheless, the proficiency in managing and processing unrestricted natural language sentences shows how argumentation-based chatbots can aptly mimic real-world-like discussions.

Finally, observe that almost every examined bot is equipped with a *retrieval-based* response model with the only exception envisaged in (Cocarascu et al., 2019; Rago et al., 2020). Indeed, the hybrid conversational agents proposed therein handle their dialogues mostly via a few tailored textual templates, hence harnessing their *rule-based* component. However, they may also resort to their retrieval-based model when in need of additional data (e.g., when the user questions the provided explanations). In general, it is also worth noticing that, unlike standard conversational agents, the surveyed literature revealed no *generative-type* argumentation-based chatbots[13]. Per se, this is not a major drawback, since generative response architecture may suffer from various issues such as lack of transparency about the origins of the produced replies, biased output, or creation of nonsensical responses. Nevertheless, this outlines a current limitation of argumentation-based bots, mostly due to an absence of studies on the matter. A possible solution to such a shortcoming may be provided, once again, by resorting to a hybrid approach that leverages state-of-the-art Transformer technologies. For example, embedding argumentation methodologies into current LLMs-based conversational agents would produce generative argumentation-based chatbots while also proving useful in mitigating those models' downsides.

5.1 Potential benefits of Leveraging Computational Argumentation Approaches in Generative-LLMs Chatbots Design

In the literature, the class of generative-LLMs chatbots (e.g. ChatGPT, Llama 3.1-Instruct, Gemini 1.5, Claude 3.5 Sonnet) is considered to be the present cutting-edge category of conversational agents. Having already listed the shortcomings that affect those models, we have

13. Recall, however, that we have no explicit information regarding the response architecture of ArgueBot (Wambsganss et al., 2021).

not yet discussed potential solutions on how to address such limitations. We argue that computational argumentation may prove to be an effective means capable of successfully handling and amending most of these weaknesses, especially (but not limited to) when they originate from the black-box nature of LLMs. Indeed, the thriving research field of *eXplainable AI (XAI)*, which studies ways to improve the interpretability of AI-driven systems, proposes also argumentative strategies as adequate forms of explanations to address the lack of models' transparency (Čyras et al., 2021; Vassiliades et al., 2021). These intuitions are backed by studies such as (McBurney & Parsons, 2021; Castagna, 2022; Castagna et al., 2024), where it is suggested that AI systems should adopt an argumentation-based approach to explanations consisting of dialogue protocols characterising the interactions between an explainer and an explainee. Embedded into LLMs, such a dialectical interplay would provide an informative post hoc method to deliver deliberated explanations to end-users while also ensuring detailed replies to follow-on queries.

On this matter, it is worth noticing that Microsoft conducted an analysis of the capability of GPT-4, one of the released GPT models (OpenAI, 2023), to provide clarifications regarding its output (Bubeck et al., 2023). Although it outperforms the ChatGPT version based on GPT-3.5, even GPT-4 has its drawbacks when dealing with the *process consistency* of its explanations: it provides a plausible account of the rationale behind the generation of its output, but it often fails in representing a more general justification able to predict the outcome of the model given similar inputs. An argumentative dialogue such as EQR (McBurney & Parsons, 2021; Castagna, 2022; Castagna et al., 2024), designed for explanation purposes, would solve the process-consistency issues by providing conversations where more information can be retrieved and thus eschewing the limited explanation length and language constraints deemed to be the leading causes of the problem (Bubeck et al., 2023).

Drawing from the usability of the aforementioned dialogue-based XAI, let us now delve into the possible ways in which computational argumentation may provide solutions (summarized in Table 3) to the current shortcomings of LLMs:

Emergent abilities. The puzzling appearance of such an unpredictable phenomenon consists of the sudden occurrence of specific competencies in large-scale models that do not manifest in smaller ones. Thus, it is not possible to anticipate the 'emergence' of these abilities (e.g. improved arithmetic, multi-task understanding, enhanced multi-lingual operations) by simply analysing smaller-scale models (Wei et al., 2022a). Among these capabilities, we can also identify Theory of Mind (ToM), i.e. the aptitude to impute mental state to others. Considered to be uniquely human, ToM may have spontaneously occurred in LLMs as a byproduct of their training (Kosinski, 2023). All of the aforementioned aspects contribute to the general mystery surrounding Transformer-based technology, which leads to mistrust among the general public, thus hampering LLMs' usability. Argumentative XAI could indirectly help as a post hoc solution: although it cannot identify the reasons why emergent abilities originate, it could nonetheless provide explanations that would clarify their functioning.

Hallucination. Defined as *'the generated content that is nonsensical or unfaithful to the provided source content'* (Ji et al., 2023) the phenomenon of hallucination in natural language generation can be divided into *intrinsic* and *extrinsic*. The former refers

to generated output that contradicts the source upon which the model was trained. The latter, instead, represents an output that cannot be verified. The employment of an argumentation reasoning engine can reduce the intrinsic hallucination kind by stipulating that only grounded arguments (hence, members of conflict-free sceptical extensions) will be output by the chatbot. On the other hand, extrinsic hallucinations can be probed by argumentative XAI methods, thus ensuring, in the worst-case scenario, the retrieval of additional information over the produced content.
Notice the differences: although inexplicable, emergent abilities usually characterise convenient competencies acquired by a model, whereas hallucinations only refer to contradictory or made-up content provided by the LLM as a reply to a user prompt.

Reasoning. Different scholars argue that, although LLMs provide a good representation of language generation, they lack reasoning skills and logical thinking (Mahowald et al., 2023; Bang et al., 2023; Frieder et al., 2023; Thorp, 2023). In an effort to offer effective solutions, various approaches have been developed, such as Chain, Tree and Graph of Thoughts (respectively, CoT, ToT and GoT). CoT consists of a prompting strategy that details a series of intermediate reasoning steps in order to achieve better performance in arithmetic, symbolic and commonsense inferences (Wei et al., 2022b). The limitations of this approach mostly concern the absence of a procedure to plan or analyse multiple reasoning paths before generating the output and this is exactly the enhancement yielded by ToT and GoT. Indeed, Tree of Thoughts frames each problem as a search over a tree, where each node is a partial solution (Yao et al., 2023). Graph of Thoughts, instead, envisages the information generated by an LLM as an arbitrary graph, distilling dependencies between such information units and enhancing reasoning by focusing on the core elements of the network (Besta et al., 2024). Against these three options, we argue that endowing generative-LLM-based chatbots with a reasoning engine driven by computational argumentation, similar to the work conducted in (Castagna et al., 2024), may provide a more intuitive, cheaper and comprehensive alternative (e.g. it does not require expensive resources to be implemented, and it is effective in a variety of topics, unlike ToT and GoT). Argumentative reasoning is particularly suited for models that parse, work and generate natural language. Recall that AFs are graphs whose edges represent paths determining the status of each node. Then, semantically computing an argumentation framework allows planning the most appropriate sequence of 'thoughts' (arguments) to achieve the desired result. Such sequences account for divergent information, thus also mimicking and (potentially) outperforming the CCoT (Contrastive Chain of Thought) prompting technique, which mostly handles only one contrastive sample at a time (Chia et al., 2023).

Biased and Toxic Output. Models have a tendency to reflect their training data, thus reproducing biased or toxic content that can harm the interacting user (Brown et al., 2020). This translates into the critical necessity of aligning LLMs towards human moral values, and even in this case, computational argumentation may prove useful to mitigate the problem. Indeed, a recent study investigates the use of computational argumentation as a tool for detecting unwanted bias in tabular data-driven binary classification decision-making systems (Waller, 2023). The proposed method

is model-agnostic and does not require access to labelled data or the specification of protected characteristics. Notice also that the steadfast progress in the field of argument mining could ensure the provision of algorithms capable of precisely detecting biased and toxic arguments in the underlying dataset and filtering them out. This would allow for the reduction of harmful data upon which generative models will be trained. Another potential solution envisages leveraging argument schemes and their taxonomies. Specifically, the instantiation of AS from AI systems enables a semantically richer approach capable of enhancing and leading LLMs-generated text into more realistic and ethically constructive debates (Bezou-Vrakatseli, 2023).

Generative LLMs Chatbot Shortcomings	**Potential Solutions**		
	Arg XAI	***Arg Engine***	***AM & AS***
Emergent Abilities	✓		
Hallucination	✓	✓	
Reasoning		✓	
Biased and Toxic Output			✓

Table 3: Computational argumentation means for addressing LLMs chatbots' downsides. Arg XAI (Argumentative XAI) refers to explanation procedures based on computational argumentation strategies and tools. Arg Engine (Argumentative Engine) concerns the reasoning capabilities of engines driven by computational argumentation (Section 2.1.3). Finally, AM indicates the Argument Mining operations of Section 2.1.1, whereas AS denotes the Argument Schemes structure of Section 2.1.2.

6. Conclusion

Conversational agents and computational argumentation are intrinsically connected by their shared focus on dialectical interactions. Combining both subjects, in this paper, we have sifted through the literature to review and analyse the existing argumentation-based chatbots. Around 70% of the bots we examined (recalling our constrained selection, as explained in Section 3) employ computational models of arguments as a way of delivering their replies to interacting users, following specific dialogue protocols. This implies that argumentative formalism proves to be particularly effective when handling exchanges of information in natural language, especially if a persuasion goal is involved. In addition, reasoning engines prove to be quite a common feature too. Harnessing argumentation extensions, those engines provide the rationale for selecting the most appropriate response to output, depending on the chosen semantics. Finally, unlike standard bots (i.e. non-argumentative ones), we discovered that there is no generative argumentation-based chatbot, nor an open-domain one, although there might be some ways of implementing such agents by embedding argumentation methodologies within LLM-driven conversational agents. Entangled with computational argumentation, chatbot design and their respective forthcoming progress, the research field of argumentation-based chatbots appears to have promising options to pursue in the coming years, including an interesting role to play in the recent Transformer-based turn of AI studies.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006.

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Altay, S., Schwartz, M., Hacquin, A.-S., Allard, A., Blancke, S., & Mercier, H. (2022). Scaling up interactive argumentation by providing counterarguments with a chatbot. *Nature Human Behaviour*, *6*(4), 579–592.

Alviano, M. (2021). The PYGLAF argumentation reasoner (ICCMA2021).. `http://argumentationcompetition.org/2021/downloads/pyglaf.pdf`, (last accessed 06/04/2024).

Amazon (2014). Alexa.. `https://developer.amazon.com/en-US/alexa`, (last accessed 06/04/2024).

Amgoud, L., & Ben-Naim, J. (2018). Evaluation of arguments in weighted bipolar graphs. *International Journal of Approximate Reasoning*, *99*, 39–55.

Andrews, P., Manandhar, S., & De Boni, M. (2008). Argumentative human computer dialogue for automated persuasion. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pp. 138–147.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Apple (2011). Siri.. `https://www.apple.com/siri/`, (last accessed 06/04/2024).

Atkinson, K., Bench-Capon, T., & Modgil, S. (2006). Argumentation for decision support. In *International Conference on Database and Expert Systems Applications*, pp. 822–831. Springer.

Azhar, M. Q., & Sklar, E. I. (2017). A study measuring the impact of shared decision making in a human-robot team. *International Journal of Robotics Research (IJRR)*, *36*, 461–482.

Bala, K., Kumar, M., Hulawale, S., & Pandita, S. (2017). Chat-bot for college management system using ai. *International Research Journal of Engineering and Technology*, *4*(11), 2030–2033.

Balatsoukas, P., Sassoon, I., Chapman, M., Kokciyan, N., Drake, A., Modgil, S., Ashworth, M., Curcin, V., Sklar, E., & Parsons, S. (2020). In the wild pilot usability assessment of a connected health system for stroke self management. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 1–3. IEEE.

Balatsoukas, P., Porat, T., Sassoon, I., Essers, K., Kökciyan, N., Chapman, M., Drake, A., Modgil, S., Ashworth, M., Sklar, E., et al. (2019). User involvement in the design of a data-driven self-management decision support tool for stroke survivors. In *IEEE EUROCON 2019-18th International Conference on Smart Technologies*, pp. 1–6. IEEE.

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al. (2023). A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Baroni, P., Romano, M., Toni, F., Aurisicchio, M., & Bertanza, G. (2015). Automatic evaluation of design alternatives with quantitative argumentation. *Argument & Computation*, *6*(1), 24–49.

Bench-Capon, T., Prakken, H., & Sartor, G. (2009). *Argumentation in legal reasoning*. Springer.

Besnard, P., & Hunter, A. (2008). *Elements of argumentation*, Vol. 47. MIT press Cambridge.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., et al. (2024). Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp. 17682–17690.

Bex, F. (2015). An integrated theory of causal stories and evidential arguments. In *Proceedings of the 15th international conference on artificial intelligence and law*, pp. 13–22.

Bex, F., Peters, J., & Testerink, B. (2016). A.I. for online criminal complaints: From natural dialogues to structured scenarios..

Bex, F. J., Van Koppen, P. J., Prakken, H., & Verheij, B. (2010). A hybrid formal theory of arguments, stories and criminal evidence. *Artificial Intelligence and Law*, *18*(2), 123–152.

Bezou-Vrakatseli, E. (2023). Evaluation of llm reasoning via argument schemes. In *Online Handbook of Argumentation for AI*, Vol. 4.

Bistarelli, S., Rossi, F., Santini, F., & Carlo, T. (2021a). CONARG: A constraint-programming solver for abstract argumentation problems.. `http://argumentationcompetition.org/2021/downloads/conarg.pdf`, (last accessed 06/04/2024).

Bistarelli, S., Taticchi, C., & Santini, F. (2021b). A chatbot extended with argumentation.. In *AI³@ AI* IA*.

Black, E., & Hunter, A. (2007). A generative inquiry dialogue system. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pp. 1–8. Association for Computing Machinery.

Black, E., Maudet, N., & Parsons, S. (2021). Argumentation-based dialogue. In *Handbook of Formal Argumentation, Volume 2*, p. 511. College Publications.

Blount, T. (2018). *Modelling eristic and rhetorical argumentation on the social web*. Ph.D. thesis, University of Southampton.

Borg, A., Odekerken, D., et al. (2022). PyArg for solving and explaining argumentation in python..

Bringsjord, S., & Govindarajulu, N. S. (2022). Artificial Intelligence. In Zalta, E. N., & Nodelman, U. (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 edition). Metaphysics Research Lab, Stanford University.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Bryant, D., & Krause, P. (2008). A review of current defeasible reasoning implementations. *The Knowledge Engineering Review*, *23*(3), 227–260.

Bryant, D., Krause, P. J., & Vreeswijk, G. (2006). Argue tuProlog: A lightweight argumentation engine for agent applications. *COMMA*, *144*, 27–32.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Cabrio, E., & Villata, S. (2018). Five years of argument mining: a data-driven analysis.. In *IJCAI*, Vol. 18, pp. 5427–5433.

Cahn, J. (2017). Chatbot: Architecture, design, & development. *University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science*.

Caldarini, G., Jaf, S., & McGarry, K. (2022). A literature survey of recent advances in chatbots. *Information*, *13*(1), 41.

Carpenter, R. (1982). Jabberwacky.. `https://web.archive.org/web/20050411013547/http://chat.jabberwacky.com/` (last accessed 06/04/2024).

Castagna, F. (2022). Towards a fully-fledged formal protocol for the Explanation-Question-Response dialogue. In *Online Handbook of Argumentation for AI*, pp. 17–21.

Castagna, F., Garton, A., McBurney, P., Parsons, S., Sassoon, I., & Sklar, E. I. (2023). EQRbot: A chatbot delivering EQR argument-based explanations. *Frontiers in Artificial Intelligence*, *6*.

Castagna, F., McBurney, P., & Parsons, S. (2024). Explanation–Question–Response dialogue: An argumentative tool for explainable AI. *Argument & Computation*, pp. 1–23.

Castagna, F., Parsons, S., Sassoon, I., & Sklar, E. I. (2022). Providing explanations via the EQR argument scheme. In *Computational Models of Argument: Proceedings of COMMA 2022*.

Castagna, F., Sassoon, I., & Parsons, S. (2024). Can formal argumentative reasoning enhance LLMs performances?..

Cayrol, C., & Lagasquie-Schiex, M.-C. (2005). On the acceptability of arguments in bipolar argumentation frameworks. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pp. 378–389. Springer.

Cerutti, F., Palmer, A., Rosenfeld, A., Šnajder, J., & Toni, F. (2016). A pilot study in using argumentation frameworks for online debates..

Chalaguine, L. A., Hamilton, F. L., Hunter, A., & Potts, H. W. W. (2018). Argument harvesting using chatbots. *Proceedings of COMMA*, 149.

Chalaguine, L. A., & Hunter, A. (2018). Chatbot design for argument harvesting. *Computational Models of Argument: Proceedings of COMMA 2018*, *305*, 457.

Chalaguine, L. A., & Hunter, A. (2019). Knowledge acquisition and corpus for argumentation-based chatbots. In *CEUR Workshop Proceedings*, Vol. 2528, pp. 1–14. CEUR Workshop Proceedings.

Chalaguine, L. A., & Hunter, A. (2020). A persuasive chatbot using a crowd-sourced argument graph and concerns. *Computational Models of Argument: Proceedings of COMMA 2020*, *326*, 9.

Chalaguine, L. A., & Hunter, A. (2021). Addressing popular concerns regarding COVID-19 vaccination with natural language argumentation dialogues. In *European Conference on Symbolic and Quantitative Approaches with Uncertainty*, pp. 59–73. Springer.

Chalaguine, L. A., Hunter, A., Potts, H., & Hamilton, F. (2019). Impact of argument type and concerns in argumentation with a chatbot. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1557–1562. IEEE.

Chapman, M., Balatsoukas, P., Kökciyan, N., Essers, K., Sassoon, I., Ashworth, M., Curcin, V., Modgil, S., Parsons, S., & Sklar, E. I. (2019). Computational argumentation-based clinical decision support. In *18th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019*, pp. 2345–2347. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

Chen, G., Cheng, L., Tuan, L. A., & Bing, L. (2023). Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Chia, Y. K., Chen, G., Tuan, L. A., Poria, S., & Bing, L. (2023). Contrastive chain-of-thought prompting. *arXiv preprint arXiv:2311.09277*.

Cocarascu, O., Rago, A., & Toni, F. (2019). Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 1261–1269. Association for Computing Machinery.

Cocarascu, O., & Toni, F. (2017). Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1374–1379, Copenhagen, Denmark. Association for Computational Linguistics.

Codecademy (2022). What are chatbots.. `https://www.codecademy.com/article/what-are-chatbots` (last accessed 06/04/2024).

Cogan, E., Parsons, S., & McBurney, P. (2005). New types of inter-agent dialogues. In *International Workshop on Argumentation in Multi-Agent Systems*, pp. 154–168. Springer.

Colby, K. M., Weber, S., & Hilf, F. D. (1971). Artificial paranoia. *Artificial Intelligence*, *2*(1), 1–25.

Cole, D. (2020). The Chinese Room Argument. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 edition). Metaphysics Research Lab, Stanford University.

Čyras, K., Rago, A., Albini, E., Baroni, P., & Toni, F. (2021). Argumentative xai: a survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) Survey Track*. International Joint Conferences on Artificial Intelligence.

Dachselt, R., Gaggl, S. A., Krötzsch, M., Mendez, J., Rusovac, D., & Yang, M. (2022). NEXAS: A visual tool for navigating and exploring argumentation solution spaces. *Computational Models of Argument: Proceedings of COMMA 2022*, *353*, 116.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811–817.

Daws, R. (2020). Medical chatbot using openai's gpt-3 told a fake patient to kill themselves.. `https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/` (last accessed 06/04/2024).

de Sousa, L. H. H., Trajano, G., Morales, A. S., Sarkadi, S., & Panisson, A. R. (2024). Using chatbot technologies to support argumentation. In *16th International Conference on Agents and Artificial Intelligence (ICAART 2024)*. SciTePress.

de Wynter, A., & Yuan, T. (2023). I wish to have an argument: Argumentative reasoning in large language models. *arXiv preprint arXiv:2309.16938*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dignum, F., & Bex, F. (2017). Creating dialogues using argumentation and social practices. In *International Conference on Internet Science*, pp. 223–235. Springer.

Dix, J., Parsons, S., Prakken, H., & Simari, G. R. (2009). Research challenges for argumentation.. *Comput. Sci. Res. Dev.*, *23*(1), 27–34.

Drake, A., Sassoon, I., Balatsoukas, P., Porat, T., Ashworth, M., Wright, E., Curcin, V., Chapman, M., Kokciyan, N., Sanjay, M., et al. (2022). The relationship of sociodemographic factors and patient attitudes to connected health technologies: a survey of stroke survivors.. *Health Informatics Journal*.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, *77*(2), 321–357.

Dutilh Novaes, C. (2022). Argument and Argumentation. In Zalta, E. N., & Nodelman, U. (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 edition). Metaphysics Research Lab, Stanford University.

Dvořák, W., Rapberger, A., Wallner, J. P., & Woltran, S. (2020). Aspartix-v19-an answer-set programming based system for abstract argumentation. In *International Symposium on Foundations of Information and Knowledge Systems*, pp. 79–89. Springer.

Dvorák, W., König, M., Wallner, J. P., & Woltran, S. (2021). ASPARTIX-V21.. http://argumentationcompetition.org/2021/downloads/aspartix-v21.pdf, (last accessed 06/04/2024).

Egly, U., Gaggl, S. A., & Woltran, S. (2008). Aspartix: Implementing argumentation frameworks using answer-set programming. In *International Conference on Logic Programming*, pp. 734–738. Springer.

Essers, K., Chapman, M., Kokciyan, N., Sassoon, I., Porat, T., Balatsoukas, P., Young, P., Ashworth, M., Curcin, V., Modgil, S., et al. (2018). The CONSULT system. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pp. 385–386.

Fazzinga, B., Flesca, S., & Furfaro, F. (2018). Probabilistic bipolar abstract argumentation frameworks: complexity results. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 1803–1809. International Joint Conferences on Artificial Intelligence Organization.

Fazzinga, B., Galassi, A., & Torroni, P. (2021). An argumentative dialogue system for covid-19 vaccine information. In *International Conference on Logic and Argumentation*, pp. 477–485. Springer.

Fazzinga, B., Galassi, A., & Torroni, P. (2022). A privacy-preserving dialogue system based on argumentation. *Intelligent Systems with Applications*, *16*, 200113.

Fichte, J. K., Hecher, M., Gorczyca, P., & Dewoprabowo, R. (2021). A-folio DPDB – system description for ICCMA 2021.. http://argumentationcompetition.org/2021/downloads/a-folio-dpdb.pdf, (last accessed 06/04/2024).

Fox, J., Glasspool, D., Grecu, D., Modgil, S., South, M., & Patkar, V. (2007). Argumentation-based inference and decision making–a medical perspective. *IEEE intelligent systems*, *22*(6), 34–41.

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*.

Galitsky, B. (2018). Enabling chatbots by detecting and supporting argumentation.. https://patents.google.com/patent/US10679011B2/en, (last accessed 06/04/2024).

Galitsky, B. (2019). Enabling chatbots by detecting and supporting affective argumentation.. https://patents.google.com/patent/US20190138595A1/en, (last accessed 06/04/2024).

Galitsky, B. (2020). Enabling chatbots by validating argumentation.. https://patents.google.com/patent/US10817670B2/en, (last accessed 06/04/2024).

Gartner, D., & Toni, F. (2007). CaSAPI: a system for credulous and sceptical argumentation. *Proc. of ArgNMR*, 80–95.

Girle, R. A. (1996). Commands in dialogue logic. In *International Conference on Formal and Applied Practical Reasoning*, pp. 246–260. Springer.

Grando, M. A., Moss, L., Sleeman, D., & Kinsella, J. (2013). Argumentation-logic for creating and explaining medical hypotheses. *Artificial intelligence in medicine*, *58*(1), 1–13.

Guo, K., Wang, J., & Chu, S. K. W. (2022). Using chatbots to scaffold efl students' argumentative writing. *Assessing Writing*, *54*, 100666.

Habernal, I., & Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, *43*(1), 125–179.

Hadoux, E., & Hunter, A. (2019). Comfort or safety? gathering and using the concerns of a participant for better persuasion. *Argument & Computation*, *10*(2), 113–147.

Hadoux, E., Hunter, A., & Polberg, S. (2021). Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. *arXiv preprint arXiv:2101.11870*.

Hauptmann, C., Krenzer, A., Völkel, J., & Puppe, F. (2024). Argumentation effect of a chatbot for ethical discussions about autonomous AI scenarios. *Knowledge and Information Systems*, 1–31.

Heater, B. (2018). Alexa gets access to Wolfram Alpha's knowledge engine.. `https://techcrunch.com/2018/12/20/alexa-gets-access-to-wolfram-alphas-knowledge-engine/`, (last accessed 06/04/2024).

Heinrich, M. (2021). The matrixx solver for argumentation frameworks.. `http://argumentationcompetition.org/2021/downloads/matrixx.pdf`, (last accessed 06/04/2024).

Hinton, M., & Wagemans, J. H. (2022). How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument & Computation*, pp. 1–16.

Hulstijn, J. (2000). Dialogue models for inquiry and transaction.. PhD thesis, Universiteit Twente, Enschede, The Netherlands.

Hung, N. D. (2017). Inference procedures and engine for probabilistic argumentation. *International Journal of Approximate Reasoning*, *90*, 163–191.

Hunter, A. (2015). Modelling the persuadee in asymmetric argumentation dialogues for persuasion. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Hunter, A. (2018). Towards a framework for computational persuasion with applications in behaviour change. *Argument & Computation*, *9*(1), 15–40.

Hunter, A., Chalaguine, L., Czernuszenko, T., Hadoux, E., & Polberg, S. (2019). Towards computational persuasion via natural language argumentation dialogues. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pp. 18–33. Springer.

IBM (2006). Watson.. `https://www.ibm.com/products/watson-assistant` (last accessed 06/04/2024).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, *55*(12).

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7b..

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Jo, Y., Bang, S., Reed, C., & Hovy, E. (2021). Classifying Argumentative Relations Using Logical Mechanisms and Argumentation Schemes. *Transactions of the Association for Computational Linguistics*, *9*, 721–739.

Kar, R., & Haldar, R. (2016). Applying chatbots to the internet of things: Opportunities and architectural elements. *International Journal of Advanced Computer Science and Applications*, *7*(11).

Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 conference on designing interactive systems*, pp. 555–565.

Kökciyan, N., Chapman, M., Balatsoukas, P., Sassoon, I., Essers, K., Ashworth, M., Curcin, V., Modgil, S., Parsons, S., & Sklar, E. I. (2019). A collaborative decision support tool for managing chronic conditions. In *The 17th World Congress of Medical and Health Informatics*.

Kökciyan, N., Sassoon, I., Sklar, E., Modgil, S., & Parsons, S. (2021). Applying metalevel argumentation frameworks to support medical decision making. *IEEE Intelligent Systems*, *36*(2), 64–71.

Kökciyan, N., Sassoon, I., Young, A., Chapman, M., Porat, T., Ashworth, M., Curcin, V., Modgil, S., Parsons, S., & Sklar, E. (2018). Towards an argumentation system for supporting patients in self-managing their chronic conditions. In *AAAI Joint Workshop on Health Intelligence (W3PHIAI)*.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Kulatska, I. (2019). Arguebot: Enabling debates through a hybrid retrieval-generation-based chatbot. Master's thesis, University of Twente.

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics*, *45*(4), 765–818.

Li, B., Fang, G., Yang, Y., Wang, Q., Ye, W., Zhao, W., & Zhang, S. (2023). Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Lin, F., & Shoham, Y. (1989). Argument systems: A uniform basis for nonmonotonic reasoning.. *KR*, *89*, 245–255.

Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., Sorsche, P., Hirsch, D., Li, Q., & Martinez, A. (2019). Talk to me: Exploring user interactions with the amazon alexa. *Journal of Librarianship and Information Science*, *51*(4), 984–997.

Luck, M., McBurney, P., Shehory, O., & Willmott, S. (2005). Agent technology: computing as interaction (a roadmap for agent based computing)..

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Malmqvist, L. (2021). AFGCN: An approximate abstract argumentation solver.. `http://argumentationcompetition.org/2021/downloads/afgcn.pdf`, (last accessed 06/04/2024).

Marshall, C. (2014). Cortana: everything you need to know about Microsoft's Siri rival.. `https://www.techradar.com/news/phone-and-communications/mobile-phones/cortana-everything-you-need-to-know-about-microsoft-s-siri-rival-1183607)` (last accessed 06/04/2024).

Mayer, T., Cabrio, E., & Villata, S. (2020). Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pp. 2108–2115. IOS Press.

McBurney, P., Hitchcock, D., & Parsons, S. (2007). The eightfold way of deliberation dialogue. In *International Journal of Intelligent Systems*, Vol. 22, pp. 95–132. Wiley Online Library.

McBurney, P., & Parsons, S. (2001). Chance discovery using dialectical argumentation. In *Annual Conference of the Japanese Society for Artificial Intelligence*, pp. 414–424. Springer.

McBurney, P., & Parsons, S. (2002). Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of logic, language and information*, *11*(3), 315–334.

McBurney, P., & Parsons, S. (2009). Dialogue games for agent argumentation. In *Argumentation in artificial intelligence*, pp. 261–280. Springer.

McBurney, P., & Parsons, S. (2013). Talking about doing. *From Knowledge Representation to Argumentation in AI, Law and Policy Making*, 151–166.

McBurney, P., & Parsons, S. (2021). Argument schemes and dialogue protocols: Doug walton's legacy in artificial intelligence. *Journal of Applied Logics*, *8*(1), 263–286.

McBurney, P., Van Eijk, R. M., Parsons, S., & Amgoud, L. (2003). A dialogue game protocol for agent purchase negotiations. In *Autonomous Agents and Multi-Agent Systems*, Vol. 7, pp. 235–273. Springer.

Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, *34*(2), 57–74.

Meta (2024). Introducing llama 3.1: Our most capable models to date. *Meta Blog*. `https://ai.meta.com/research/publications/the-llama-3-herd-of-models/` (last accessed 24/07/2024).

Microsoft (2014). Cortana.. https://www.microsoft.com/en-us/cortana (last accessed 06/04/2024).

Modgil, S., & Prakken, H. (2013). A general account of argumentation with preferences. *Artificial Intelligence*, *195*, 361–397.

Murphy, J., Black, E., & Luck, M. M. (2016). A heuristic strategy for persuasion dialogues. In *Computational Models of Argument: Proceedings of COMMA 2016*, pp. 411–418. IOS Press.

Niskanen, A., & Järvisalo, M. (2021). μ–toksia at ICCMA'21.. http://argumentationcompetition.org/2021/downloads/mu-toksia.pdf, (last accessed 06/04/2024).

OpenAI (2023). Gpt-4 technical report..

Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D. R., Texier, M., & Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. *Computer*, *55*(7), 18–28.

Podlaszewski, M., Caminada, M., & Pigozzi, G. (2011). An implementation of basic argumentation components. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pp. 1307–1308.

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive science*, *11*(4), 481–518.

Prakken, H. (2006). Formal systems for persuasion dialogue. In *The knowledge engineering review*, Vol. 21, pp. 163–188. Cambridge University Press.

Prakken, H., Bistarelli, S., & Santini, F. (2020). *Computational Models of Argument: Proceedings of COMMA 2020*, Vol. 326. IOS Press.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver?. *arXiv preprint arXiv:2302.06476*.

Rago, A., Cocarascu, O., Bechlivanidis, C., & Toni, F. (2020). Argumentation as a framework for interactive explanations for recommendations. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, Vol. 17, pp. 805–815.

Rago, A., Cocarascu, O., & Toni, F. (2018). Argumentation-based recommendations: Fantastic explanations and how to find them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 1949–1955.

Rago, A., Toni, F., Aurisicchio, M., & Baroni, P. (2016). Discontinuity-free decision support with quantitative argumentation debates. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Rahwan, I., & Larson, K. (2009). Argumentation and game theory. *Argumentation in artificial intelligence*, 321–339.

Rahwan, I., & Simari, G. R. (2009). *Argumentation in Artificial Intelligence*. Springer.

Reckwitz, A. (2002). Toward a theory of social practices: A development in culturalist theorizing. *European journal of social theory*, *5*(2), 243–263.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., et al. (2020). Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Rosenfeld, A., & Kraus, S. (2016). Strategical argumentative agent for human persuasion. In *ECAI 2016*, pp. 320–328. IOS Press.

Ruiz-Dolz, R., Alemany, J., Barbera, S., & Garcia-Fornes, A. (2021). Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, *36*(06), 62–70.

Saadat-Yazdi, A., Pan, J., & Kökciyan, N. (2023). Uncovering implicit inferences for improved relational argument mining. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2476–2487.

Sansonnet, J.-P., Leray, D., & Martin, J.-C. (2006). Architecture of a framework for generic assisting conversational agents. In *International Workshop on Intelligent Virtual Agents*, pp. 145–156. Springer.

Sassoon, I., Kökciyan, N., Chapman, M., Sklar, E., Curcin, V., Modgil, S., & Parsons, S. (2020). Implementing argument and explanation schemes in dialogue. *Computational Models of Argument: Proceedings of COMMA 2020*, *326*, 471.

Sassoon, I., Kökciyan, N., Modgil, S., & Parsons, S. (2021). Argumentation schemes for clinical decision support. *Argument & Computation*, pp. 1–27.

Sassoon, I., Kökciyan, N., Sklar, E., & Parsons, S. (2019). Explainable argumentation for wellness consultation. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pp. 186–202. Springer.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage?. *arXiv preprint arXiv:2304.15004*.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.

Shove, E., Pantzar, M., & Watson, M. (2012). *The Dynamics of Social Practice: Everyday Life and how it Changes*. SAGE.

Silver, D., & Veness, J. (2010). Monte-carlo planning in large pomdps. *Advances in neural information processing systems*, *23*.

Singh, S., & Thakur, H. K. (2020). Survey of various AI Chatbots based on technology used. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pp. 1074–1079. IEEE.

Sklar, E., & Parsons, S. (2004). Towards the application of argumentation-based dialogues for education. In *Autonomous Agents and Multiagent Systems, International Joint Conference on*, Vol. 4, pp. 1420–1421. IEEE Computer Society.

Sklar, E. I., & Azhar, M. Q. (2015). Argumentation-based dialogue games for shared control in human-robot systems. *Journal of Human-Robot Interaction*, *4*(3), 120–148.

Sklar, E. I., & Azhar, M. Q. (2018). Explanation through argumentation. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pp. 277–285.

Slonim, N., Bilu, Y., Alzate, C., Bar-Haim, R., Bogin, B., Bonin, F., Choshen, L., Cohen-Karlik, E., Dankin, L., Edelstein, L., et al. (2021). An autonomous debating system. *Nature*, *591*(7850), 379–384.

Sojasingarayar, A. (2020). Seq2seq ai chatbot with attention mechanism. *arXiv preprint arXiv:2006.02767*.

Tang, Y., Sklar, E., & Parsons, S. (2012). An argumentation engine: Argtrust. In *Ninth International Workshop on Argumentation in Multiagent Systems*.

Thimm, M. (2021). Harper+: Using grounded semantics for approximate reasoning in abstract argumentation.. `http://argumentationcompetition.org/2021/downloads/harper++.pdf`, (last accessed 06/04/2024).

Thimm, M., Cerutti, F., & Vallati, M. (2021). FUDGE: A light-weight solver for abstract argumentation based on sat reductions.. `http://argumentationcompetition.org/2021/downloads/fudge.pdf`, (last accessed 06/04/2024).

Thorp, H. H. (2023). Chatgpt is fun, but not an author. *Science*, *379*(6630), 313–313.

Tolchinsky, P., Modgil, S., Atkinson, K., McBurney, P., & Cortés, U. (2012). Deliberation dialogues for reasoning about safety critical actions. *Autonomous Agents and Multi-Agent Systems*, *25*(2), 209–259.

Toni, F. (2014). A tutorial on assumption-based argumentation. *Argument & Computation*, *5*(1), 89–117.

Toniuc, D., & Groza, A. (2017). Climebot: An argumentative agent for climate change. In *2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, pp. 63–70. IEEE.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trautmann, D., Daxenberger, J., Stab, C., Schütze, H., & Gurevych, I. (2020). Fine-grained argument unit recognition and classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 9048–9056.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., & Wolf, T. (2023). Zephyr: Direct distillation of lm alignment..

Turing, A. M., & Haugeland, J. (1950). Computing machinery and intelligence. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, 29–56.

Vassiliades, A., Bassiliades, N., & Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, *36*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Visser, J., Lawrence, J., Wagemans, J., & Reed, C. (2018). Revisiting computational models of argument schemes: Classification, annotation, comparison. In *7th International Conference on Computational Models of Argument, COMMA 2018*, pp. 313–324. ios Press.

Vreeswijk, G. (1994). IACAS: An interactive argumentation system. *Rapport technique CS*, *94*(03).

Wagemans, J. (2016). Constructing a periodic table of arguments. In *Argumentation, objectivity, and bias: Proceedings of the 11th international conference of the Ontario Society for the Study of Argumentation (OSSA), Windsor, ON: OSSA*, pp. 1–12.

Wallace, R. (2003). The elements of AIML style. *Alice AI Foundation*, *139*.

Wallace, R. S. (2009). The anatomy of A.L.I.C.E.. In *Parsing the turing test*, pp. 181–210. Springer.

Waller, M. (2023). An argumentation-based approach to bias detection in automated decision-making systems. In *Online Handbook of Argumentation for AI*, Vol. 4.

Walton, D. (2012). Argument mining by applying argumentation schemes. *Studies in Logic*, *4*(1), 2011.

Walton, D., & Krabbe, E. C. (1995). *Commitment in dialogue: Basic concepts of interpersonal reasoning*. SUNY press.

Walton, D., & Macagno, F. (2015). A classification system for argumentation schemes. *Argument & Computation*, *6*(3), 219–245.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.

Walton, D. N. (1990). What is reasoning? What is an argument?. *The journal of Philosophy*, *87*(8), 399–419.

Walton, D. N., & Gordon, T. F. (2011). Modeling Critical Questions as Additional Premises. In *Proceedings of the 8th International OSSA Conference*.

Wambsganss, T., Guggisberg, S., & Söllner, M. (2021). Arguebot: A conversational agent for adaptive argumentation feedback. In *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*, pp. 267–282. Springer.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022a). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Worswick, S. (2013). Interview - Loebner 2013 winner.. https://aidreams.co.uk/forum/index.php?page=Steve_Worswick_Interview_-_Loebner_2013_winner#.Y0IfoHZBxPY (last accessed 06/04/2024).

Worswick, S. (2018). Mitsuku wins Loebner Prize 2018!.. https://medium.com/pandorabots-blog/mitsuku-wins-loebner-prize-2018-3e8d98c5f2a7 (last accessed 06/04/2024).

Xu, J., Ju, D., Lane, J., Komeili, M., Smith, E. M., Ung, M., Behrooz, M., Ngan, W., Moritz, R., Sukhbaatar, S., Boureau, Y.-L., Weston, J., & Shuster, K. (2023). Improving open language models by learning from organic interactions..

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2019). Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics*, *46*(1), 53–93.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity..