G Supplementary Material

G.1 Temporality analysis for previous horizon from $h_v = 1$ to $h_v = 8$

In Table 8, we can see the absolute percentages of networks which develop temporal references.

The full plots for all metrics are shown for previous horizon from $h_v = 1$ to $h_v = 8$. The $M_{\ominus^h}$ metric values for the Base, Temporal, and TemporalR agents are available in Figures 10 to 12. The correctness of the messages used as the $\ominus^h$ operator for each agent type is available in Figures 13 and 14. No Base networks feature in the plots, as they do not develop temporal references, and so there are no messages to measure the correctness of.

Table 8. Emergence of temporal references for a given horizon.

Network Type	$h_v = 1$	$h_v = 2$	$h_v = 3$	$h_v = 4$	$h_v = 5$	$h_v = 6$	$h_v = 7$	$h_v = 8$
Base	0%	0%	0%	0%	0%	0%	0%	0%
Base+L	0%	0%	0%	0%	0%	0%	0%	0%
Temporal	100%	100%	100%	100%	100%	100%	100%	100%
Temporal+L	100%	100%	100%	100%	100%	100%	100%	100%
TemporalR	98.89%	100%	100%	99.44%	97.22%	100%	99.44%	97.78%
TemporalR+L	99.44%	100%	99.44%	98.89%	98.89%	99.44%	99.44%	98.89%

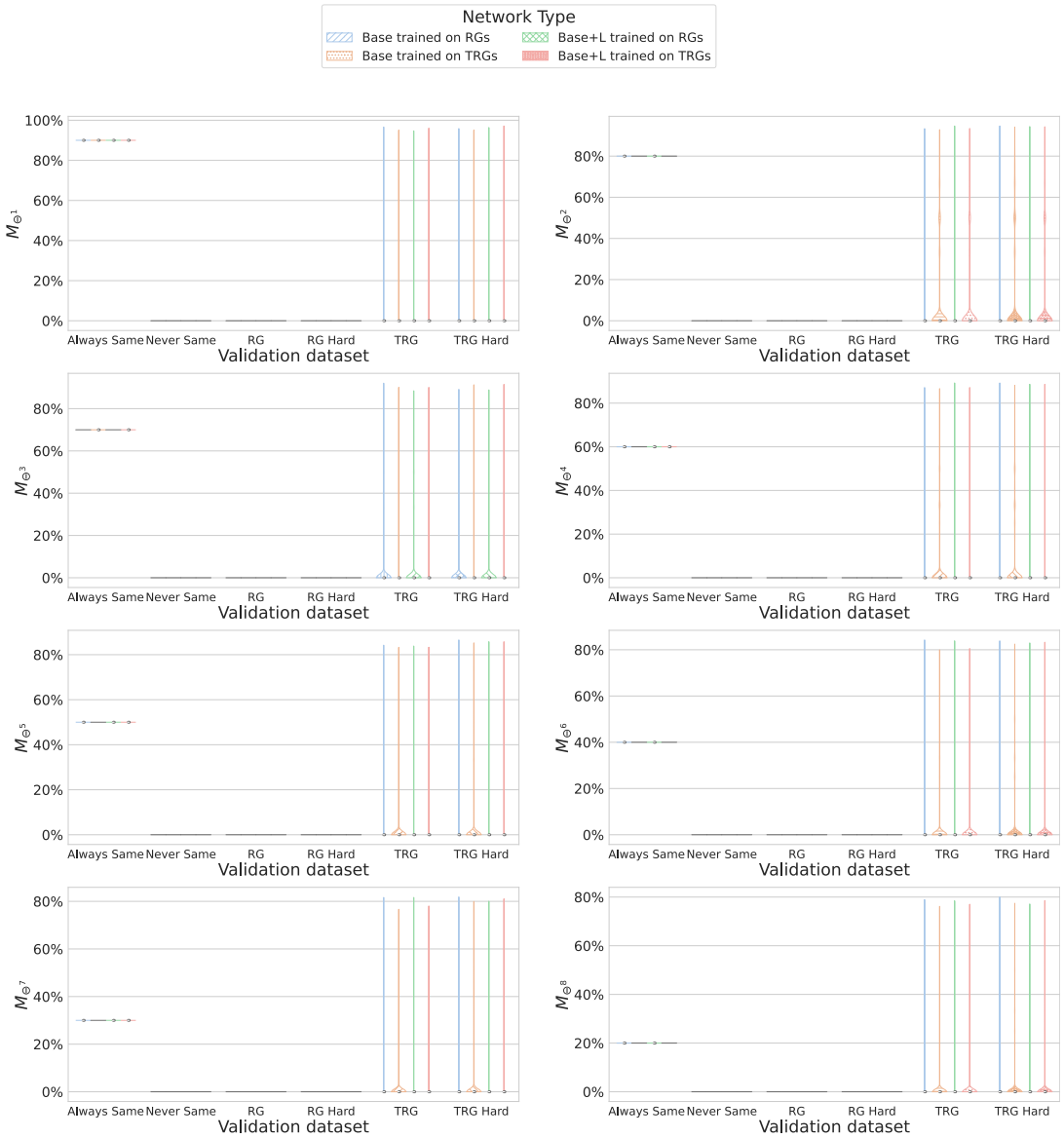

Fig. 10. The $M_{\ominus^h}$ metric values per message for the *Base* agents, for all environments.

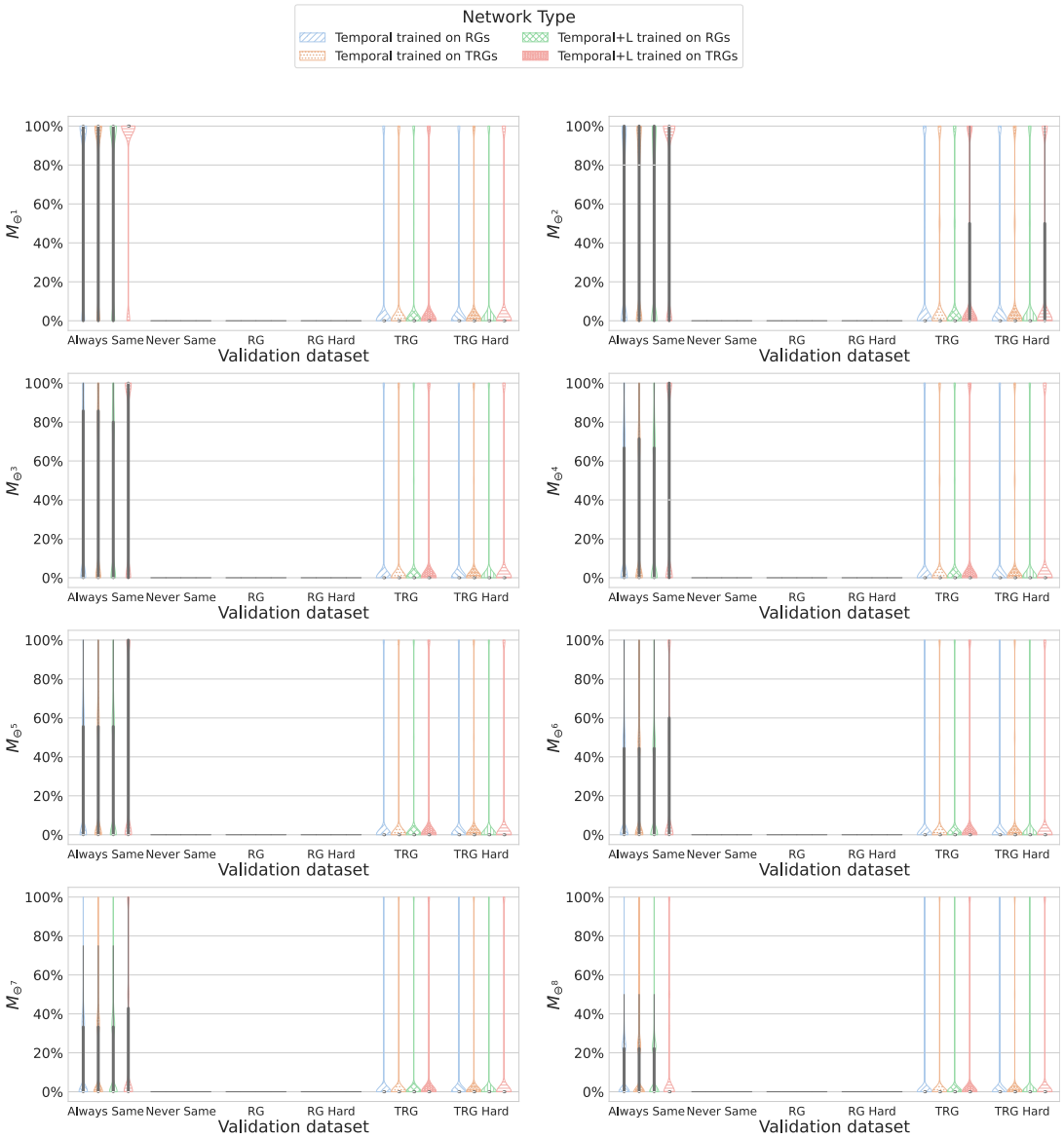

Fig. 11. The $M_{\ominus h}$ metric values per message for the *Temporal* agents, for all environments.

Fig. 12. The $M_{\ominus^h}$ metric values per message for the *TemporalR* agents, for all environments.

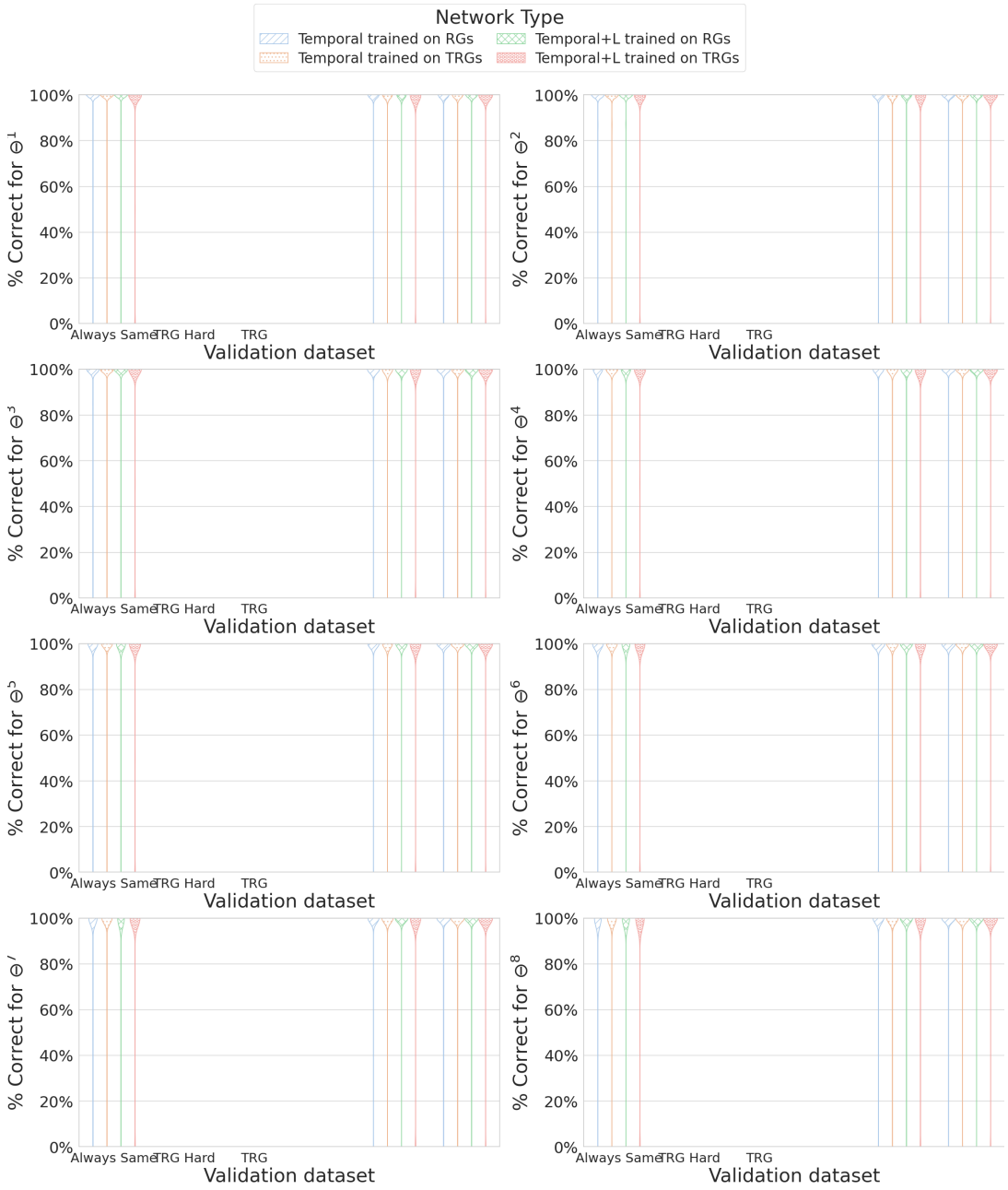

Fig. 13. Correctness of messages used as the $\ominus^h$ operator for the *Temporal* agents.

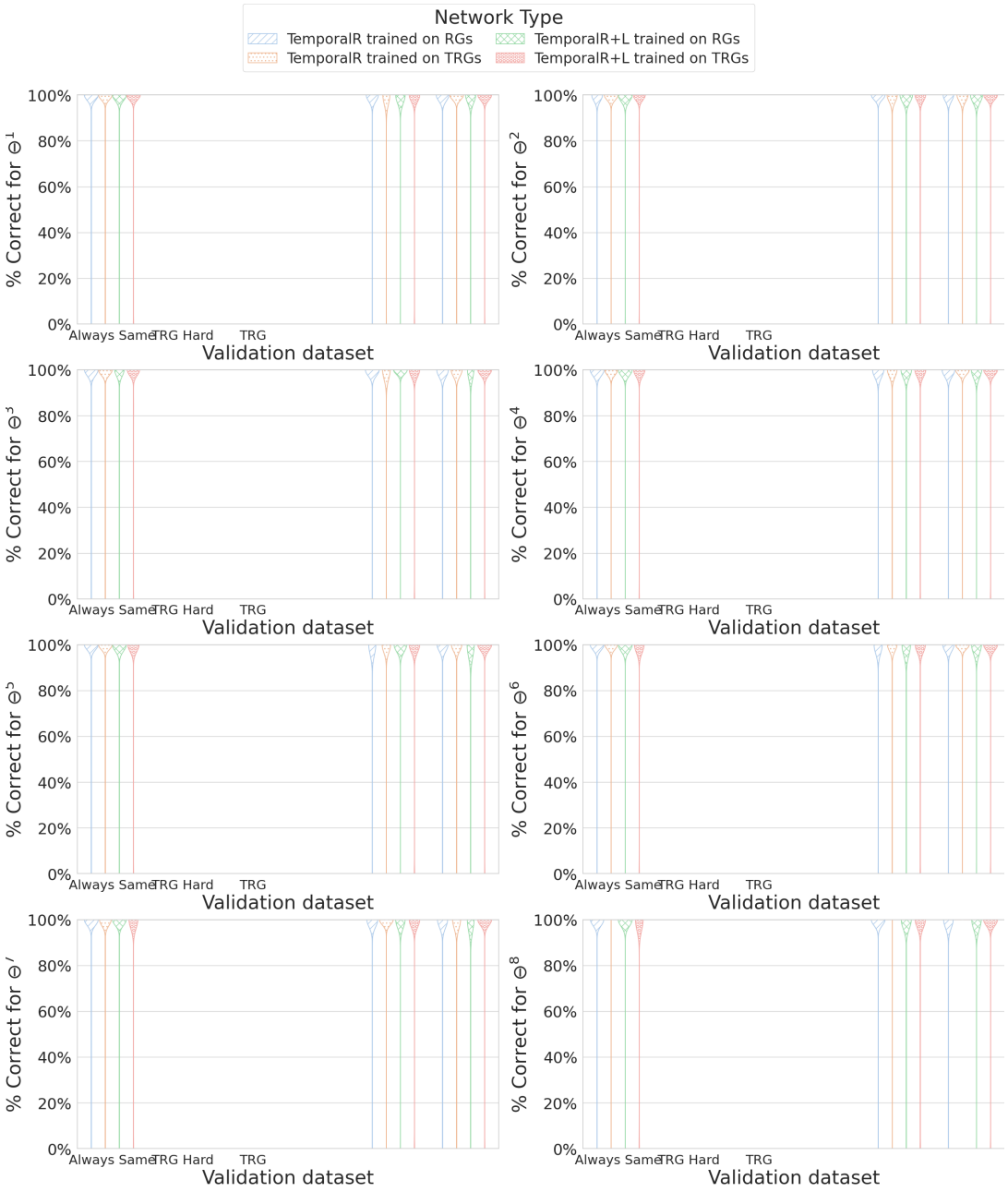

Fig. 14. Correctness of messages used as the $\ominus^h$ operator for the *TemporalR* agents.

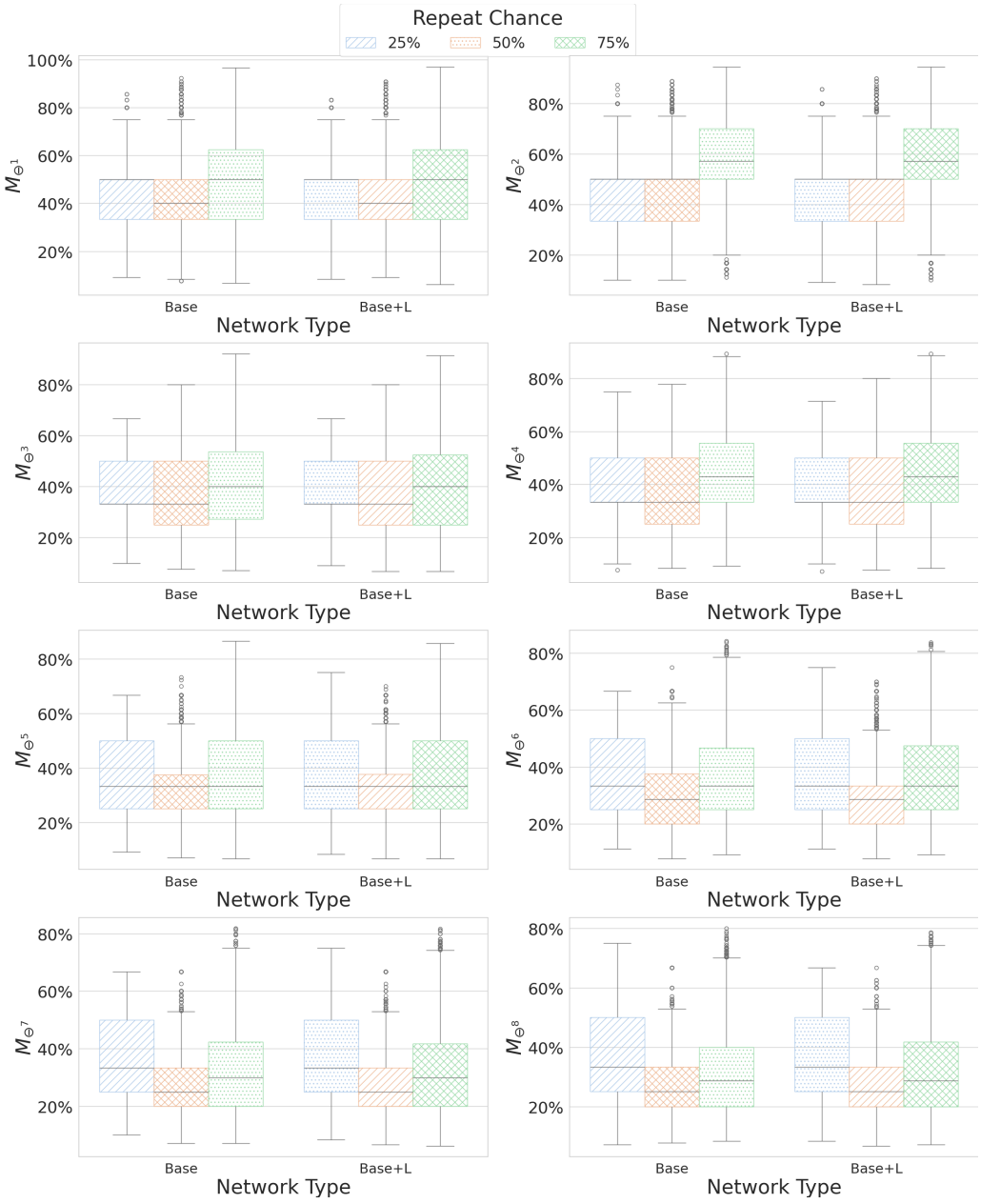

Fig. 15. The $M_{\ominus^h}$ value when varying the loss and the chance of repetition for the *Base* agents.

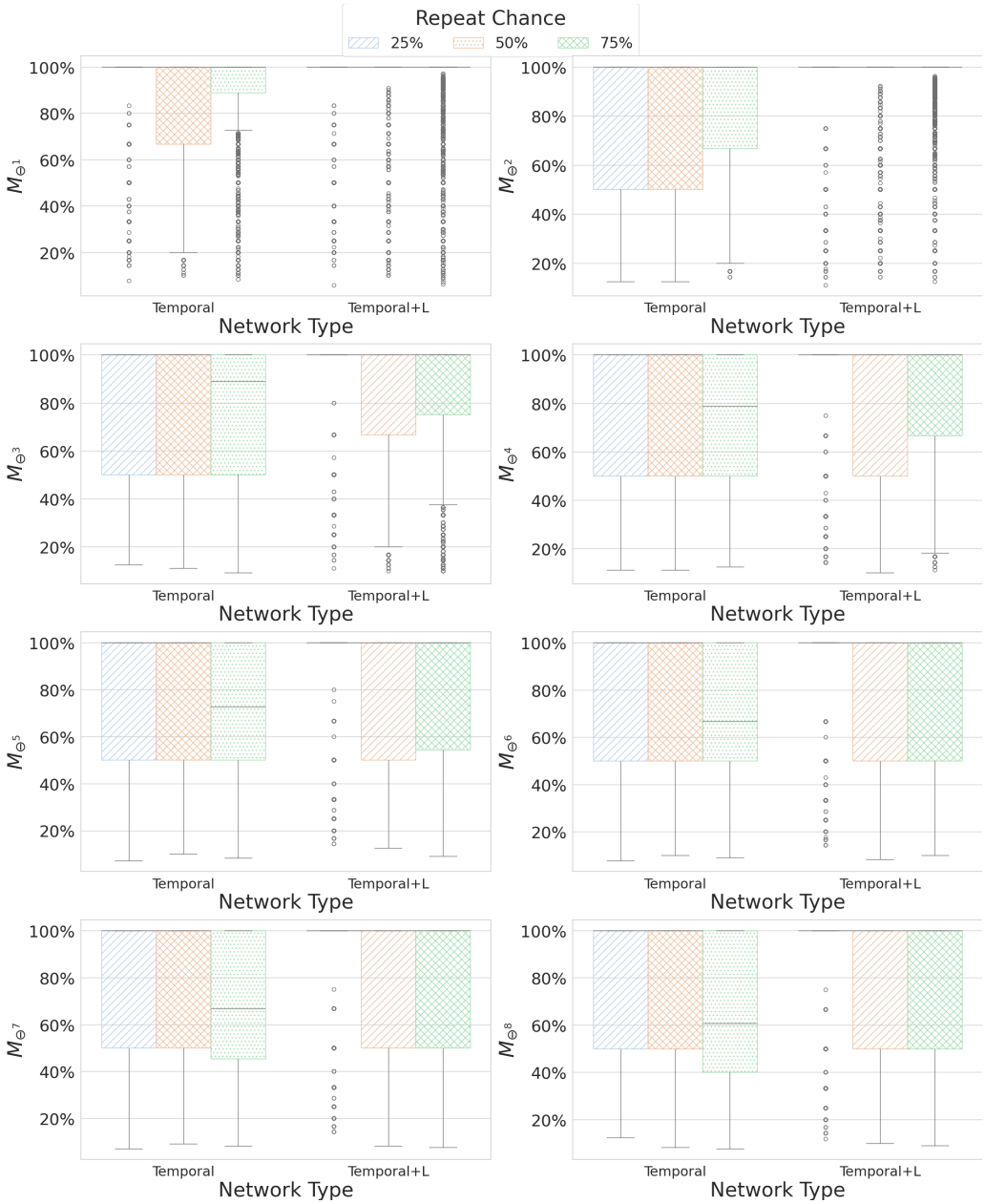

Fig. 16. The $M_{\ominus^h}$ value when varying the loss and the chance of repetition for the *Temporal* agents.

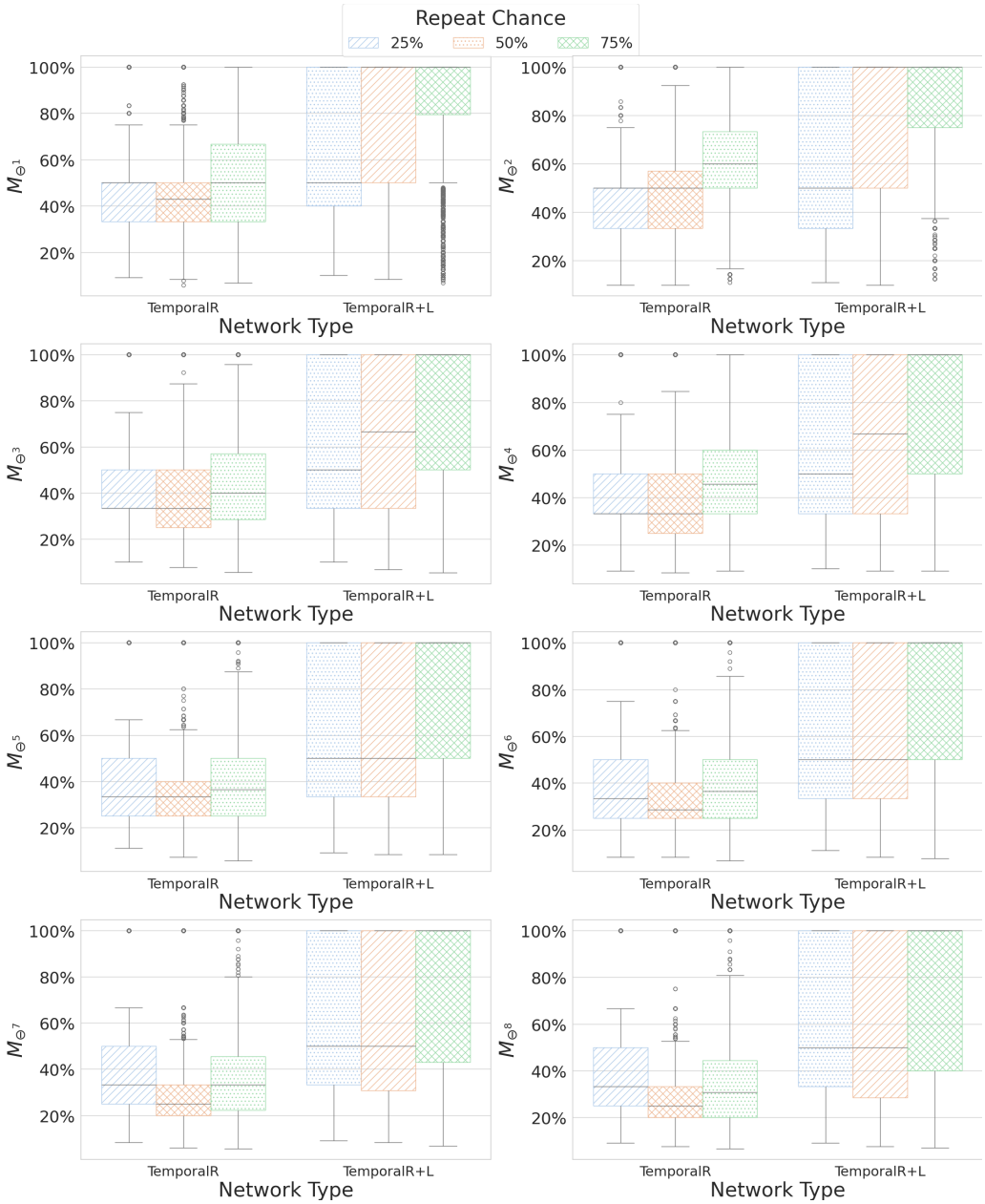

Fig. 17. The $M_{\ominus^h}$ value when varying the loss and the chance of repetition for the *TemporalR* agents.

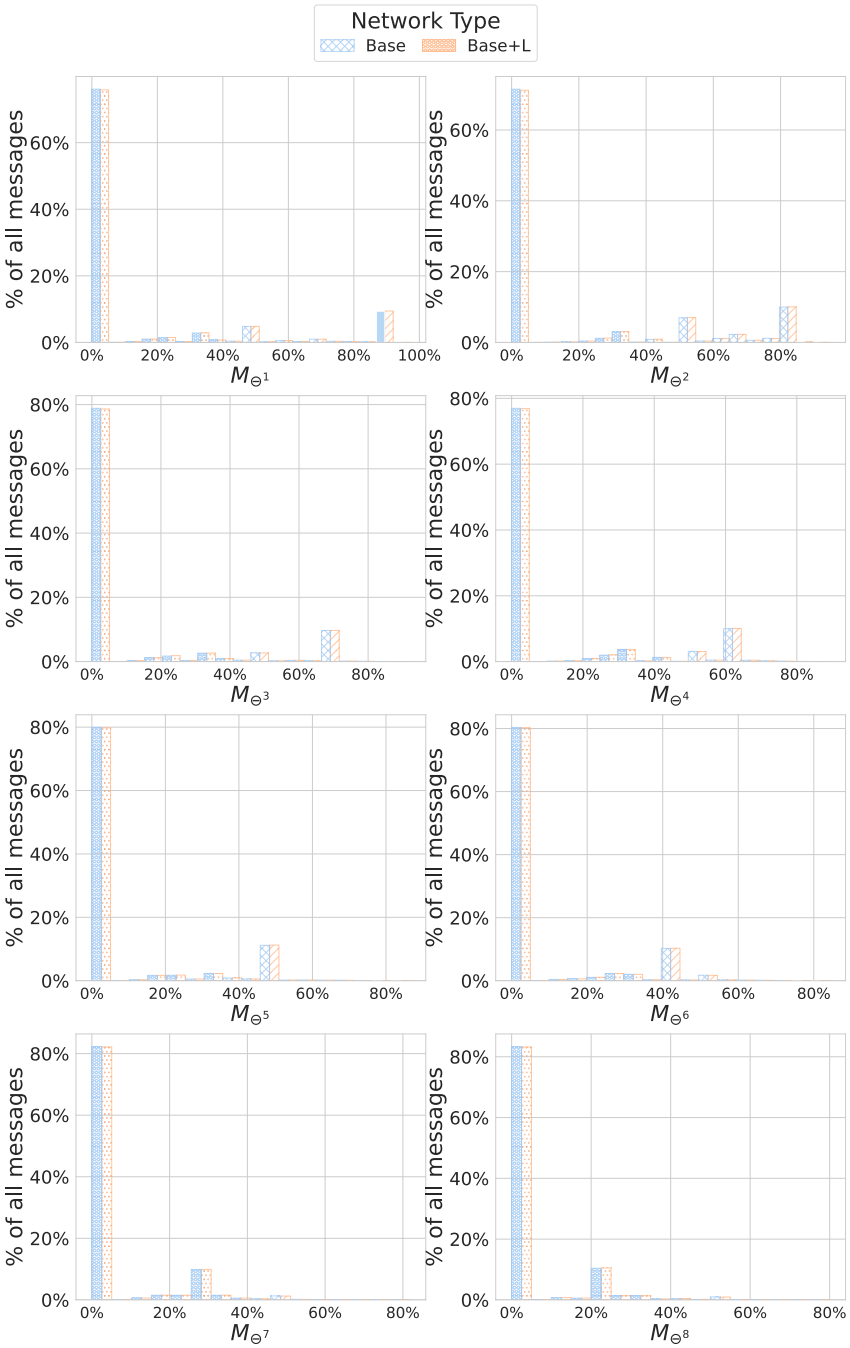

Fig. 18. Usage of messages compared to their $M_{\ominus^h}$ value for the *Base* agents.

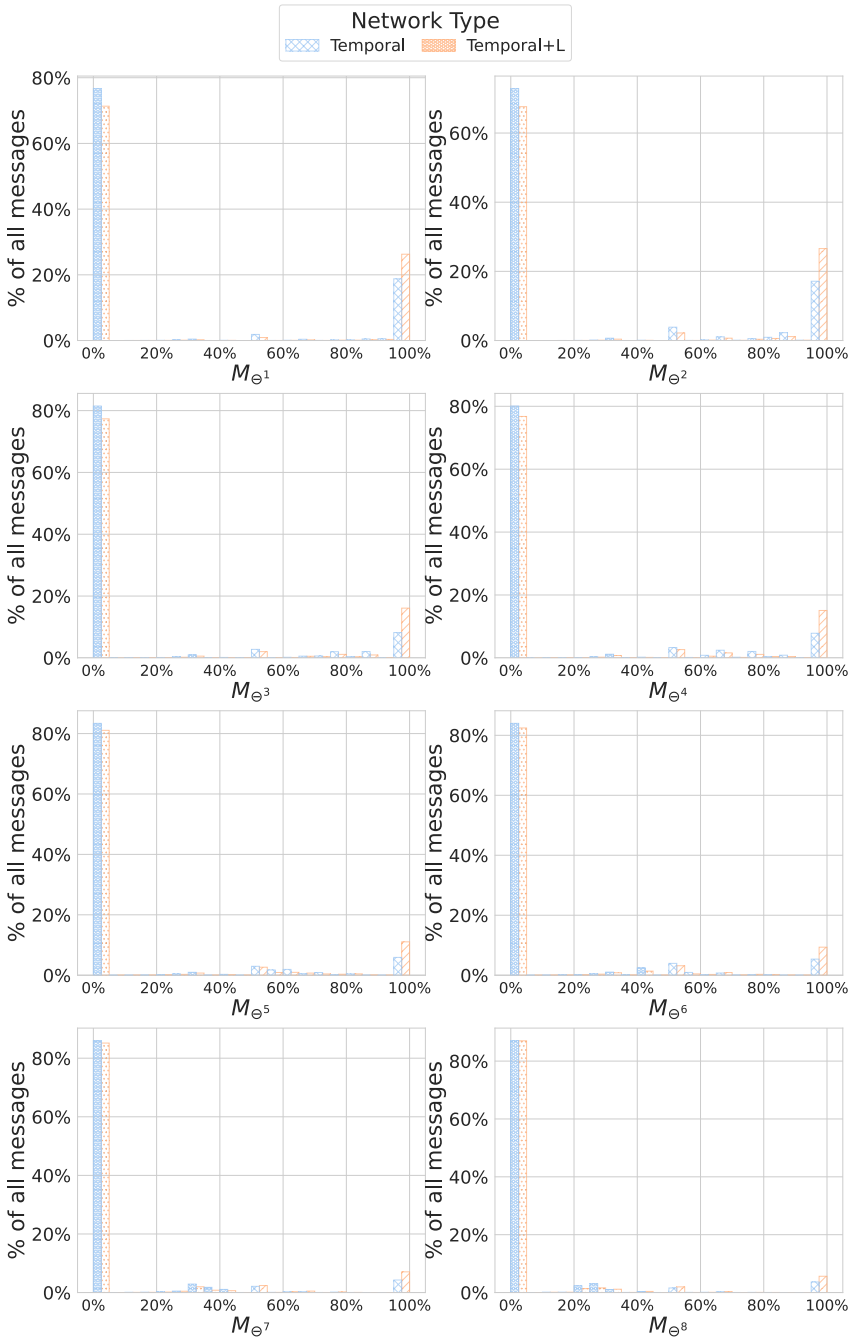

Fig. 19. Usage of messages compared to their $M_{\ominus^h}$ value for the *Temporal* agents.

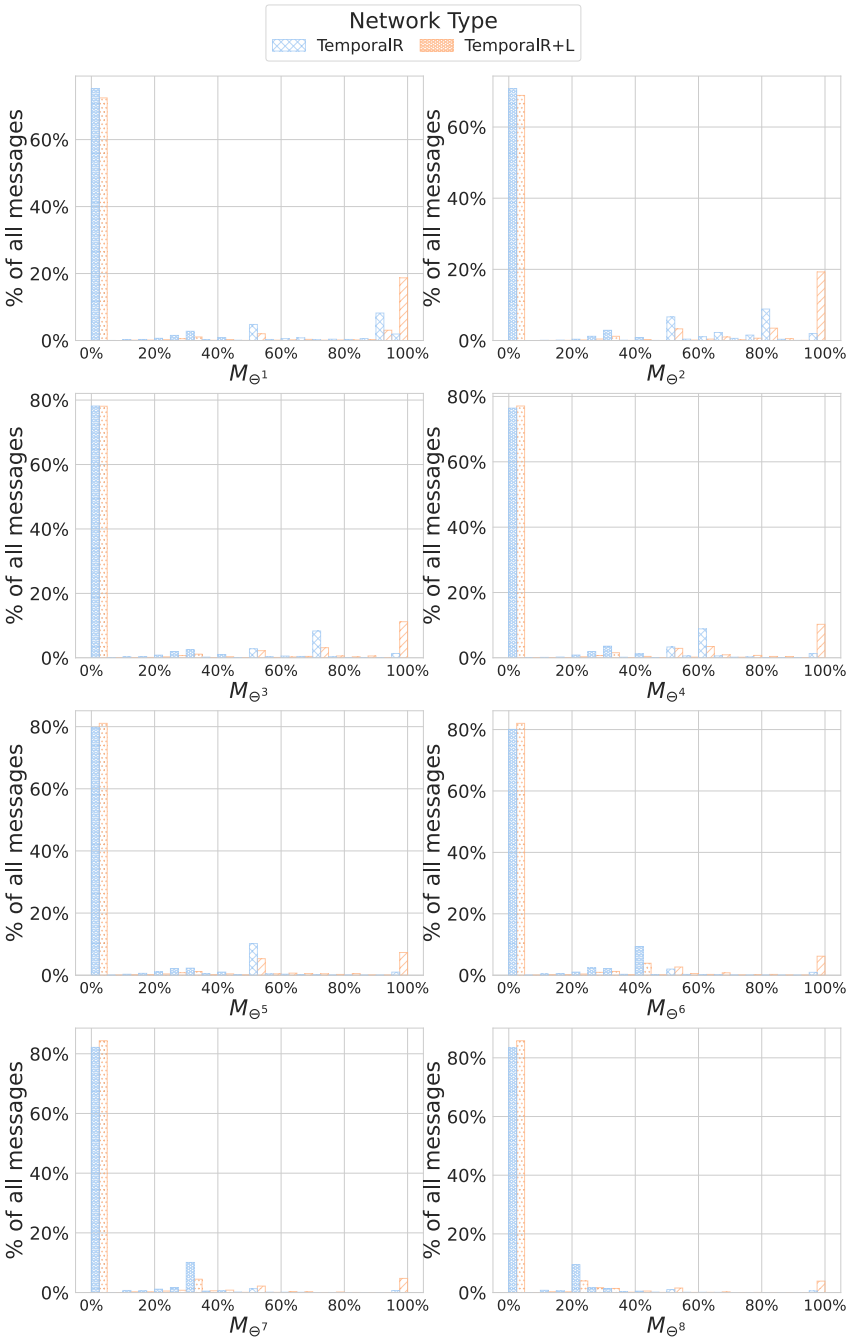

Fig. 20. Usage of messages compared to their $M_{\ominus^h}$ value for the *TemporalR* agents.

G.2 Analysis within Network Types

In this section, the detailed values of all statistical significance tests are provided.

Network Type	Training Env	Always Same	Never Same	RG	RG Hard	TRG	TRG Hard
Base	RG	0.01	0.01	0.01	0.02	0.01	0.02
	TRG	0.39	0.04	0.02	0.01	0.01	0.01
Base+L	RG	0.01	0.01	0.01	0.73	0.08	0.01
	TRG	0.18	0.25	0.98	0.42	0.24	0.01
Temporal	RG	0.11	0.65	0.22	0.86	0.69	0.01
	TRG	0.01	0.01	0.01	0.02	0.01	0.01
Temporal+L	RG	0.02	0.07	0.08	0.46	0.48	0.01
	TRG	0.03	0.78	0.82	0.80	0.60	0.01
TemporalR	RG	0.09	0.68	0.27	0.04	0.16	0.01
	TRG	0.02	0.01	0.02	0.05	0.25	0.01
TemporalR+L	RG	0.03	0.01	0.37	0.28	0.63	0.01
	TRG	0.06	0.10	0.21	0.30	0.34	0.01

Table 9. p values for the normality test on the accuracy scores.

Network Type	Training Env	Always Same	Never Same	RG	RG Hard	TRG	TRG Hard
Base	RG	NaN	N/A	NaN	NaN	0	0
	TRG	NaN	N/A	NaN	NaN	0	0
Base+L	RG	NaN	N/A	NaN	NaN	0	0
	TRG	NaN	N/A	NaN	NaN	0	0
Temporal	RG	0	N/A	NaN	NaN	0	0
	TRG	0	N/A	NaN	NaN	0	0
Temporal+L	RG	0	N/A	NaN	NaN	0	0
	TRG	0	N/A	NaN	NaN	0	0
TemporalR	RG	0	N/A	NaN	NaN	0	0
	TRG	0	N/A	NaN	NaN	0	0
TemporalR+L	RG	0	N/A	NaN	NaN	0	0
	TRG	0	N/A	NaN	NaN	0	0

Table 10. p values for the normality test on the $M_{\ominus^n}$ scores.

G.2.1 Normality Tests.

Network Type	Training Env	TopSim	Posdis	Bosdis
Base	RG	0.01	0.01	0.75
	TRG	0.44	0.01	0.90
Base+L	RG	0.31	0.02	0.05
	TRG	0.91	0.01	0.01
Temporal	RG	0.11	0.01	0.02
	TRG	0.38	0.02	0.01
Temporal+L	RG	0.05	0.01	0.14
	TRG	0.38	0.01	0.80
TemporalR	RG	0.97	0.10	0.32
	TRG	0.34	0.08	0.52
TemporalR+L	RG	0.02	0.01	0.83
	TRG	0.01	0.02	0.01

Table 11. p values for the normality test on the compositionality scores.

	Always Same	Never Same	RG	RG Hard	TRG	TRG Hard
Base	0	0	0	0	0	0.060617
Temporal	0	0	0	0	0	0.000197
TemporalR	0	0	0	0	0	0.004879

Table 12. p values of the Kruskal-Wallis H-test for the accuracy scores within a given network type.

G.2.2 Accuracy Kruskal-Wallis H-test P-Values within Network Types.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	0.433259	0	0.498274
Base RG	0.433259	1	0	0.171310
Base+L TRG	0	0	1	0
Base RG	0.498274	0.171310	0	1

Table 13. p values of the post-hoc Conover test for the accuracy scores on the Always Same environment.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	0.433899	0	0.582182
Base RG	0.433899	1	0	0.226648
Base+L TRG	0	0	1	0
Base RG	0.582182	0.226648	0	1

Table 14. p values of the post-hoc Conover test for the accuracy scores on the Never Same environment.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	0.149439	0	0.668939
Base RG	0.149439	1	0	0.083620
Base+L TRG	0	0	1	0
Base RG	0.668939	0.083620	0	1

Table 15. p values of the post-hoc Conover test for the accuracy scores on the RG environment.

G.2.3 Accuracy Post-Hoc Conover Analysis within Network Types.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	0.346337	0	0.346337
Base RG	0.346337	1	0	0.040696
Base+L TRG	0	0	1	0
Base RG	0.346337	0.040696	0	1

Table 16. p values of the post-hoc Conover test for the accuracy scores on the RG Hard environment.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	0.185363	0	0.185363
Base RG	0.185363	1	0	0.005564
Base+L TRG	0	0	1	0
Base RG	0.185363	0.005564	0	1

Table 17. p values of the post-hoc Conover test for the accuracy scores on the TRG environment.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	1	0.175554	1
Base RG	1	1	0.087409	1
Base+L TRG	0.175554	0.087409	1	0.175554
Base RG	1	1	0.175554	1

Table 18. p values of the post-hoc Conover test for the accuracy scores on the TRG Hard environment.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	1	0	1
Temporal RG	1	1	0	1
Temporal+L TRG	0	0	1	0
Temporal RG	1	1	0	1

Table 19. p values of the post-hoc Conover test for the accuracy scores on the Always Same environment.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	0.876314	0	0.114027
Temporal RG	0.876314	1	0	0.114027
Temporal+L TRG	0	0	1	0
Temporal RG	0.114027	0.114027	0	1

Table 20. p values of the post-hoc Conover test for the accuracy scores on the Never Same environment.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	0.384963	0	0.635388
Temporal RG	0.384963	1	0	0.229984
Temporal+L TRG	0	0	1	0
Temporal RG	0.635388	0.229984	0	1

Table 21. p values of the post-hoc Conover test for the accuracy scores on the RG environment.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	0.056940	0	0.000033
Temporal RG	0.056940	1	0	0.017182
Temporal+L TRG	0	0	1	0
Temporal RG	0.000033	0.017182	0	1

Table 22. p values of the post-hoc Conover test for the accuracy scores on the RG Hard environment.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	1	0	1
Temporal RG	1	1	0	1
Temporal+L TRG	0	0	1	0
Temporal RG	1	1	0	1

Table 23. p values of the post-hoc Conover test for the accuracy scores on the TRG environment.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	0.993318	0.000135	0.691546
Temporal RG	0.993318	1	0.001456	0.993318
Temporal+L TRG	0.000135	0.001456	1	0.006804
Temporal RG	0.691546	0.993318	0.006804	1

Table 24. p values of the post-hoc Conover test for the accuracy scores on the TRG Hard environment.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	0.442920	0	0.442920
TemporalR RG	0.442920	1	0	0.071027
TemporalR+L TRG	0	0	1	0
TemporalR RG	0.442920	0.071027	0	1

Table 25. p values of the post-hoc Conover test for the accuracy scores on the Always Same environment.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	0.224358	0	0.084598
TemporalR RG	0.224358	1	0	0.004183
TemporalR+L TRG	0	0	1	0
TemporalR RG	0.084598	0.004183	0	1

Table 26. p values of the post-hoc Conover test for the accuracy scores on the Never Same environment.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	0.009688	0	0.512991
TemporalR RG	0.009688	1	0	0.040779
TemporalR+L TRG	0	0	1	0
TemporalR RG	0.512991	0.040779	0	1

Table 27. p values of the post-hoc Conover test for the accuracy scores on the RG environment.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	0.589040	0	0.871791
TemporalR RG	0.589040	1	0	0.589040
TemporalR+L TRG	0	0	1	0
TemporalR RG	0.871791	0.589040	0	1

Table 28. p values of the post-hoc Conover test for the accuracy scores on the RG Hard environment.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	0.015392	0	0.481321
TemporalR RG	0.015392	1	0	0.002605
TemporalR+L TRG	0	0	1	0
TemporalR RG	0.481321	0.002605	0	1

Table 29. p values of the post-hoc Conover test for the accuracy scores on the TRG environment.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	1	0.011330	1
TemporalR RG	1	1	0.012271	1
TemporalR+L TRG	0.011330	0.012271	1	0.018447
TemporalR RG	1	1	0.018447	1

Table 30. p values of the post-hoc Conover test for the accuracy scores on the TRG Hard environment.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	1	0	1
Base RG	1	1	0	1
Base+L TRG	0	0	1	0
Base RG	1	1	0	1

Table 31. p values of the post-hoc Conover test for the topographic similarity scores.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	0.271644	0	0.075936
Temporal RG	0.271644	1	0	0.454652
Temporal+L TRG	0	0	1	0
Temporal RG	0.075936	0.454652	0	1

Table 32. p values of the post-hoc Conover test for the topographic similarity scores.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	0.000081	0	0
TemporalR RG	0.000081	1	0	0.131500
TemporalR+L TRG	0	0	1	0
TemporalR RG	0	0.131500	0	1

Table 33. p values of the post-hoc Conover test for the topographic similarity scores.

G.2.4 Topographic Similarity Post-Hoc Conover Analysis within Network Types.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	1	0	0.967127
Base RG	1	1	0	1
Base+L TRG	0	0	1	0
Base RG	0.967127	1	0	1

Table 34. p values of the post-hoc Conover test for the posdis scores.

G.2.5 Positional Disentanglement Post-Hoc Conover Analysis within Network Types.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	1	1	0.082919
Temporal RG	1	1	1	0.015836
Temporal+L TRG	1	1	1	0.045470
Temporal RG	0.082919	0.015836	0.045470	1

Table 35. p values of the post-hoc Conover test for the posdis scores.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	1	0.149973	1
TemporalR RG	1	1	0.042950	1
TemporalR+L TRG	0.149973	0.042950	1	0.248096
TemporalR RG	1	1	0.248096	1

Table 36. p values of the post-hoc Conover test for the posdis scores.

	Base+L RG	Base RG	Base+L TRG	Base RG
Base+L RG	1	1	0.001331	0.843365
Base RG	1	1	0.008259	1
Base+L TRG	0.001331	0.008259	1	0.034571
Base RG	0.843365	1	0.034571	1

Table 37. p values of the post-hoc Conover test for the bosdis scores.

	Temporal+L RG	Temporal RG	Temporal+L TRG	Temporal RG
Temporal+L RG	1	0.023856	0	0.000197
Temporal RG	0.023856	1	0	0.135339
Temporal+L TRG	0	0	1	0
Temporal RG	0.000197	0.135339	0	1

Table 38. p values of the post-hoc Conover test for the bosdis scores.

	TemporalR+L RG	TemporalR RG	TemporalR+L TRG	TemporalR RG
TemporalR+L RG	1	0.043151	0	0.878130
TemporalR RG	0.043151	1	0	0.043455
TemporalR+L TRG	0	0	1	0
TemporalR RG	0.878130	0.043455	0	1

Table 39. p values of the post-hoc Conover test for the bosdis scores.

G.2.6 Bag-of-Words Disentanglement Post-Hoc Conover Analysis within Network Types.

G.3 Analysis Between Network Types

G.3.1 Accuracy Post-Hoc Conover Analysis across Network Types. We do not list all the Kruskal-Wallis H-test p values individually, as they are all below the threshold value of 0.001.

	Base	Temporal	TemporalR
Base	1	0.000317	0.330549
Temporal	0.000317	1	0.006967
TemporalR	0.330549	0.006967	1

Table 40. p values of the post-hoc Conover test for the accuracy scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0.000005	0.235654
Temporal	0.000005	1	0.000513
TemporalR	0.235654	0.000513	1

Table 41. p values of the post-hoc Conover test for the accuracy scores on the Never Same environment.

	Base	Temporal	TemporalR
Base	1	0.000001	0.213130
Temporal	0.000001	1	0.000186
TemporalR	0.213130	0.000186	1

Table 42. p values of the post-hoc Conover test for the accuracy scores on the RG environment.

	Base	Temporal	TemporalR
Base	1	0.000184	0.909507
Temporal	0.000184	1	0.000195
TemporalR	0.909507	0.000195	1

Table 43. p values of the post-hoc Conover test for the accuracy scores on the RG Hard environment.

	Base	Temporal	TemporalR
Base	1	0.000011	0.337959
Temporal	0.000011	1	0.000422
TemporalR	0.337959	0.000422	1

Table 44. p values of the post-hoc Conover test for the accuracy scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0.676719	0.763523
Temporal	0.676719	1	0.724056
TemporalR	0.763523	0.724056	1

Table 45. p values of the post-hoc Conover test for the accuracy scores on the TRG Hard environment.

G.3.2 Compositionality Post-Hoc Conover Analysis across Network Types. We provide a full breakdown of the post-hoc Conover analysis of the compositionality scores. However, the p value of the Kruskal-Wallis H-test for the posdis metric was above the 0.001 threshold ($p = 0.10$), and so this analysis may not be statistically significant.

	Base	Temporal	TemporalR
Base	1	0	0.000020
Temporal	0	1	0.001046
TemporalR	0.000020	0.001046	1

Table 46. p values of the post-hoc Conover test for the topographic similarity scores.

	Base	Temporal	TemporalR
Base	1	0.274577	1
Temporal	0.274577	1	0.294687
TemporalR	1	0.294687	1

Table 47. p values of the post-hoc Conover test for the posdis scores.

	Base	Temporal	TemporalR
Base	1	0.003442	0.813492
Temporal	0.003442	1	0.032353
TemporalR	0.813492	0.032353	1

Table 48. p values of the post-hoc Conover test for the bosdis scores.

G.3.3 Temporality Post-Hoc Conover Analysis across Network Types. We do not list all the Kruskal-Wallis H-test p values individually, as they are all below the threshold value of 0.001.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 49. p values of the post-hoc Conover test for the $M_{\ominus^1}$ scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0.049232
TemporalR	0	0.049232	1

Table 50. p values of the post-hoc Conover test for the $M_{\ominus^2}$ scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0	0.000001
Temporal	0	1	0
TemporalR	0.000001	0	1

Table 51. p values of the post-hoc Conover test for the $M_{\ominus^3}$ scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 52. p values of the post-hoc Conover test for the $M_{\ominus^4}$ scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 53. p values of the post-hoc Conover test for the $M_{\ominus^5}$ scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 54. p values of the post-hoc Conover test for the $M_{\ominus^6}$ scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 55. p values of the post-hoc Conover test for the $M_{\ominus^7}$ scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 56. p values of the post-hoc Conover test for the $M_{\ominus^8}$ scores on the Always Same environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0.031283
TemporalR	0	0.031283	1

Table 57. p values of the post-hoc Conover test for the $M_{\ominus^1}$ scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 58. p values of the post-hoc Conover test for the $M_{\ominus^2}$ scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 59. p values of the post-hoc Conover test for the $M_{\ominus^3}$ scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 60. p values of the post-hoc Conover test for the $M_{\ominus^4}$ scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 61. p values of the post-hoc Conover test for the $M_{\ominus^5}$ scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 62. p values of the post-hoc Conover test for the $M_{\ominus^6}$ scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0.000006
TemporalR	0	0.000006	1

Table 63. p values of the post-hoc Conover test for the $M_{\ominus^7}$ scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0.000001
TemporalR	0	0.000001	1

Table 64. p values of the post-hoc Conover test for the $M_{\ominus^8}$ scores on the TRG environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0.000077
TemporalR	0	0.000077	1

Table 65. p values of the post-hoc Conover test for the $M_{\ominus^1}$ scores on the TRG Hard environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 66. p values of the post-hoc Conover test for the $M_{\ominus^2}$ scores on the TRG Hard environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 67. p values of the post-hoc Conover test for the $M_{\ominus^3}$ scores on the TRG Hard environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 68. p values of the post-hoc Conover test for the $M_{\ominus^4}$ scores on the TRG Hard environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 69. p values of the post-hoc Conover test for the $M_{\ominus^5}$ scores on the TRG Hard environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0
TemporalR	0	0	1

Table 70. p values of the post-hoc Conover test for the $M_{\ominus^6}$ scores on the TRG Hard environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0.000041
TemporalR	0	0.000041	1

Table 71. p values of the post-hoc Conover test for the $M_{\ominus^7}$ scores on the TRG Hard environment.

	Base	Temporal	TemporalR
Base	1	0	0
Temporal	0	1	0.000119
TemporalR	0	0.000119	1

Table 72. p values of the post-hoc Conover test for the $M_{\ominus^8}$ scores on the TRG Hard environment.