YNU-HPCC at SemEval-2026 Task 8: Parallel Generation and Multi-Metric Reranking for Faithful Extractive RAG

Bo Li, You Zhang*, Jin Wang, Dan Xu, and XueJie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
Contact: bli@stu.ynu.edu.cn, yzhang0202@ynu.edu.cn

Abstract

This paper presents our approach for the SemEval-2026 Task 8: MTRAGEval (Subtask B: Answer Generation), which challenges systems to generate faithful, extractive answers to multi-turn questions based strictly on provided gold-standard reference passages. The primary scientific challenge lies in maintaining high faithfulness and structural consistency while adapting to diverse answer styles across a conversation, as systems must generate responses that vary significantly in length and format without hallucinating. Conventional reference-based generation methods often rely on static prompting or greedy decoding, which fail to capture these dynamic stylistic requirements and lack robustness against generation noise. To address these limitations, we propose a Intent-Aware Parallel Generation and Reranking System powered by a large language model. Experimental results on the official test set demonstrate the effectiveness of our method, achieving competitive performance comparable to SoTA baselines. Ultimately, our approach secured the third place in the competition. The code of the paper is available at: https://github.com/viaviachris/SemEval-2026-Task8

1 Introduction

Retrieval-Augmented Generation has been widely adopted to mitigate hallucinations in Large Language Models (LLMs) by grounding responses in external knowledge (Gao et al., 2024; Schick et al., 2023). Building on our previous explorations into fine-grained hallucination detection (Chen et al., 2025), this work addresses a formidable challenge that extends beyond single-turn interactions: maintaining faithfulness and structural consistency in multi-turn conversational scenarios (Adlakha et al., 2022). To address this, the SemEval-2026 Task 8 (MTRAGEval) is introduced as a comprehensive benchmark (Katsis et al., 2025). Specifically, Subtask B (Generation with Reference Passages) requires systems to generate faithful, extractive answers based strictly on provided gold-standard passages, spanning diverse domains such as Finance, Government, and General Knowledge(Ghosh et al., 2026).Compared to the standalone DeepSeek-V3 baseline without our tailored pipeline, the proposed system achieves an absolute improvement of 8% in ROUGE-1 and 9% in ROUGE-L, robustly demonstrating the effectiveness of our architectural design.

The primary scientific problem inherent in this task is stylistic adaptability under strict faithfulness constraints. User queries in a multi-turn conversation vary significantly in intent and required answer format—ranging from concise factoid responses to exhaustive explanations. Consequently, traditional generation systems often fail to adapt to these dynamic requirements, leading to either over-summarization (omitting critical details) or hallucination (adding unsupported information) (Huang et al., 2025).

Previously, generation tasks were predominantly addressed using zero-shot prompting or static few-shot learning. However, these static approaches fall short in complex multi-turn settings: they fail to capture the nuanced stylistic differences between question types (e.g., "Yes/No" questions vs. "How-to" procedures) and lack robustness against the generation noise inherent in stochastic LLMs. Furthermore, standard decoding strategies often rely on a single generation pass, which may not guarantee the optimal alignment with the provided evidence (Wang et al., 2023a). To overcome these limitations, a robust Intent-Aware Parallel Generation and Reranking System is proposed in this paper. This approach is designed to explicitly model the relationship between question types and answer styles, maximizing extractive precision through a

*Corresponding author.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 943–949

multi-stage pipeline. Experimental results on the MTRAGEval benchmark demonstrate that the proposed method significantly outperforms baselines. The effectiveness of adapting retrieval strategies to question types and leveraging self-consistency for reranking is verified.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 introduces the proposed methodology and model architecture. Section 4 presents the experimental setup, including datasets, evaluation metrics, results, and ablation studies. Finally, Section 5 concludes the paper and outlines future work.

2 Related Work

Retrieval-Augmented Generation in Multi-Turn QA. Retrieval-Augmented Generation (RAG) has become the standard paradigm for knowledge-intensive tasks. Traditional approaches to tasks similar to Subtask B typically rely on zero-shot prompting or static few-shot learning, directly concatenating dialogue history with retrieved documents(Jiang et al., 2024). However, these static methods often struggle to adapt to the dynamic intent shifts across different turns, leading to style inconsistencies or failure to strictly adhere to the provided evidence.

Hallucination Mitigation and Faithfulness. To ensure strict faithfulness in generation, previous works have explored various decoding and verification strategies. Many conventional systems employ greedy decoding for deterministic outputs, while more advanced pipelines integrate self-reflection or chain-of-verification to filter out hallucinated content(Jiang et al., 2023). Despite these advancements, standard generation passes often suffer from stochastic noise(Manakul et al., 2023). Relying on a single verification metric or a single decoding pass limits the ability to maximize both extractive precision and stylistic alignment(Asai et al.).

To address these limitations in multi-turn scenarios, we propose a Intent-Aware Parallel Generation and Reranking System. Unlike conventional static methods, our approach leverages Intent-Driven Template Retrieval combined with an LLM-based parallel generation and reranking mechanism (Wang et al., 2023a), effectively filtering out unfaithful candidates while maintaining diverse conversational intents.

3 Intent-Aware Parallel Generation and Reranking System

The overall architecture of our proposed approach for the SemEval-2026 Task 8 (Subtask B) is illustrated in Figure 1. To address the challenges of stylistic diversity and faithfulness in multi-turn question answering, a cascade pipeline named the Intent-Aware Parallel Generation and Reranking System is proposed (Khattab et al., 2024). As shown in the figure, the framework is composed of three integrated modules: (1) Intent-Driven Template Retrieval, which selects style-consistent few-shot examples based on the semantic intent of the query (Ram et al., 2023); (2) LLM-Based Parallel Generation, which utilizes a Large Language Model (LLM) to sample multiple candidate answers concurrently; and (3) Multi-Metric Self-Consistency Reranking, which evaluates and selects the most faithful response based on a weighted ensemble of ROUGE metrics against the gold reference passages (Dhuliawala et al., 2024).

3.1 Intent-Driven Template Retrieval

This module first extracts the specific intent type directly from the provided task metadata. Then, it encodes the user query into a dense vector using BGE-M3 (Chen et al., 2024) and retrieves a broader set of semantically similar few-shot examples via FAISS. Finally, a Question-Type Prioritized reranking mechanism is applied: it re-orders the retrieved candidates to strictly prioritize examples that share the exact same intent type as the current query. These top-ranked examples subsequently serve as precise structural and stylistic templates for the generation module.

3.2 LLM-Based Parallel Generation

The retrieved examples, dialogue history, and gold reference passages are concatenated into a prompt for DeepSeek-V3 (DeepSeek-AI et al., 2025). However, since concatenating extensive evidence can lead to attention degradation over long contexts (Liu et al., 2024), we mitigate this by employing parallel sampling with a non-zero temperature to generate multiple diverse candidate answers, effectively expanding the solution space(Yao et al., 2023).

3.3 Multi-Metric Self-Consistency Reranking

To guarantee strict faithfulness, this module evaluates all generated candidates against the gold pas-

sages using a weighted ensemble of ROUGE metrics, computed as follows:

$$\begin{aligned} S_i = {} & 0.35 \times \text{ROUGE-1}(C_i, R) \\ & + 0.35 \times \text{ROUGE-2}(C_i, R) \\ & + 0.30 \times \text{ROUGE-L}(C_i, R) \end{aligned} \quad (1)$$

where R denotes the gold reference passages. The candidate with the highest score is selected via an $\arg\max$ operation as the Final Faithful Answer. This mechanism filters out hallucinated content and prioritizes responses that maximize information recall while maintaining the extractive style.

4 Experiments

Datasets. The proposed system was evaluated on the MTRAG benchmark, a comprehensive dataset designed for multi-turn retrieval-augmented generation. The benchmark covers four diverse domains: Finance, Government, Medicine, and General Knowledge. Specifically for Subtask B (Generation with Reference Passages), the input for each task consists of the full conversation history, the current user query, and a set of gold-standard reference passages. The objective is to generate a faithful and extractive answer based strictly on the provided gold passages.

Evaluation Metrics. Following the official SemEval-2026 Task 8(Rosenthal et al., 2026b), three primary metrics were utilized to assess the generation quality:

1. RB_{alg}: A deterministic metric calculated as the harmonic mean of BERT-Recall, ROUGE-L, and BERT-K-Precision. This captures the lexical and semantic overlap between the generated response and the ground truth.

2. RB_{llm}: A reference-based LLM judge adapted from RAD-Bench, which evaluates the quality of the answer given the reference.

3. RL_f: The RAGAS Faithfulness LLM judge, which assesses whether the generated claim is fully grounded in the provided context, aligning with the evidence-based verification paradigms explored in our prior work (Deng et al., 2025).

All metrics are conditioned on an "IDK" classifier to penalize systems that attempt to answer unanswerable queries (Rosenthal et al., 2026a). While the official leaderboard employs LLM-based judges for factual consistency (Gekhman et al., 2023) and a composite metric, these are computationally expensive for iterative experiments. Therefore, during the system development and ablation studies (Section 3.4), we utilized standard ROUGE-1, ROUGE-2, ROUGE-L, and BLEU-4 as efficient proxy metrics. Given that the official composite metric explicitly incorporates ROUGE-L, these proxy metrics serve as reliable indicators of extractive quality and content overlap.

Implementation Details. To validate the effectiveness of our proposed mechanism, we benchmark against several representative reference-based generation strategies. The primary baseline is the Zero-Shot Generation setting adopted by official MTRAG models, which relies solely on instruction following without in-context examples (Wang et al., 2023b). We also compare against Static Few-Shot strategies that lack adaptability to diverse dialogue scenarios, and Naive Dynamic Few-Shot approaches that utilize semantic retrieval but neglect data cleaning and question-type constraints. Furthermore, to assess the contribution of our self-consistency module, we include Single Decoding and Single-Metric Reranking as ablation baselines.

Addressing the limitations of these baselines, our system implements a rigorous Data Preprocessing pipeline—applied to both index construction and evaluation—that standardizes refusal responses, removes citation markers, and normalizes formatting to ensure consistency. Specifically, the FAISS example pool was constructed using the curated tasks derived from the trial data. Building on this, the Retrieval Module employs BAAI/bge-m3 with a FAISS index to retrieve few-shot examples, utilizing a Question-Type Prioritized strategy to enhance style adaptability. During inference, the generation prompt was systematically structured to concatenate the system instruction, the retrieved top-k in-context examples, the dialogue history, and the gold reference passages. Additionally, to align with the "IDK" classifier condition in the official evaluation, the model was explicitly prompted to output a standardized refusal if the provided passages lacked sufficient evidence to answer the query.

The Generation Module utilizes DeepSeek-V3 with a maximum context of 16,000 tokens (up to 5 reference passages) and a 512-token output limit to encourage conciseness. Finally, in the Self-Consistency phase, we generate $N = 3$ parallel candidates with a temperature of 0.7 (Top-p = 1.0) and select the final output using a weighted ROUGE scoring mechanism, which math-

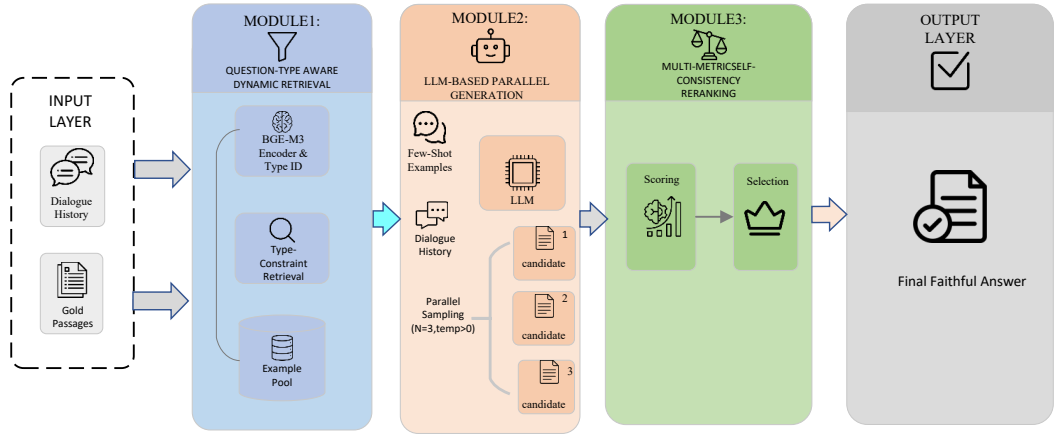

Figure 1: Intent-Aware Parallel Generation and Reranking System

Few-Shot (k)	ROUGE-1	ROUGE-2	ROUGE-L
1	0.4184	0.2076	0.3020
2	0.4181	0.2074	0.3018
3	0.4685	0.2933	0.3664

Table 1: Impact of the number of retrieved few-shot examples (k) on generation performance.

ematically assigns weights of 0.35 to ROUGE-1, 0.35 to ROUGE-2, and 0.30 to ROUGE-L to maximize extractive precision.

Hyperparameter Tuning. We conducted hyperparameter tuning and ablation studies on the validation set to determine the optimal configuration and quantify module contributions.

1. **Few-Shot Example Size (k):** Evaluating $k \in \{1, 2, 3\}$ (Table 1) revealed that $k = 1$ or $k = 2$ yielded suboptimal ROUGE-1 scores of approximately 0.418, often producing overly concise responses lacking stylistic diversity. Increasing to $k = 3$ substantially improved ROUGE-1 to 0.4685 and ROUGE-L to 0.3664. This indicates $k = 3$ optimally balances stylistic guidance and context limits without introducing distracting noise.

2. **Sampling Temperature (T) and Candidate Size (N):** To maximize the efficacy of the parallel generation module, we explored the trade-off between output diversity and hallucination risks. Greedy decoding ($T = 0$) restricted the solution space, while excessively high temperatures ($T > 0.8$) introduced unacceptable generation noise. We empirically established that setting $T = 0.7$ with parallel candidates $N = 3$ provides sufficient structural variance for the reranking phase without incurring prohibitive computational overhead.

3. **System Ablation:** Table 2 details the performance gains of each core module. From a naïve few-shot baseline ($k = 2$, greedy decoding; ROUGE-1: 0.418, BLEU-4: 0.089), introducing question-type constraints and $k = 3$ improved ROUGE-1 to 0.445 and sharply increased BLEU-4 to 0.148, proving the necessity of structural guidance. Enabling parallel sampling ($N = 3$) leveraged model stochasticity to explore a broader solution space, pushing ROUGE-1 to 0.460. Finally, applying multi-metric weighted reranking yielded the optimal performance (ROUGE-1: 0.469, ROUGE-L: 0.366). This progressive trajectory confirms that extractive overlap-based filtering effectively mitigates generation noise and hallucinations.

Comparative Results. To rigorously evaluate the efficacy of our proposed Intent-Aware Parallel Generation and Reranking System, we benchmarked its performance against top-tier official baselines, including GPT-4o and Llama 3.1 405B. In Table 3, experimental results demonstrate that our system achieved a substantial advantage across all key evaluation metrics.

Specifically, attributed to the effective suppression of hallucinations by the Multi-Metric Self-Consistency Reranking mechanism, our model attained a Faithfulness score (RL_F) of 0.87, significantly surpassing both GPT-4o (0.76) and Llama 3.1 405B (0.75). Regarding answer quality (RB_{llm}), our system achieved a remarkable score of 0.88, significantly outperforming all baselines (where the top performance was only 0.76). This

Method Strategy	Few-Shot(k)	Self-Consist	ROUGE-1	ROUGE-L	BLEU-4
Baseline(Naïve Few-Shot)	3	-	0.418	0.302	0.089
+ Constraints	3	-	0.445	0.353	0.148
+ Self Consistency	3	3	0.460	0.362	0.147
+ Weighted Reranking (Ours)	3	3	0.469	0.366	0.148

Table 2: Ablation study of the proposed Intent-Aware Parallel Generation and Reranking System.

Model	RL_F	RB_{llm}	RB_{alg}
GPT-4o	0.76	0.76	0.45
Llama 3.1 405B	0.75	0.74	0.48
Llama 3.1 8B	0.55	0.59	0.37
Our System (base DeepSeek-V3)	**0.87**	**0.88**	**0.64**

Table 3: Main evaluation results of the proposed system against official baselines.

System	Harmonic Mean
Top Performing Team	0.7827
Our System (3rd Place)	**0.7684**
Official Baseline (gpt-oss-120b)	0.6390

Table 4: Comparison of overall performance (Harmonic Mean) on Subtask B.

evidences the success of our Dynamic Retrieval strategy in capturing subtle stylistic nuances and adapting smoothly to shifting user intents across multiple conversational turns.

Furthermore, in terms of lexical overlap (RB_{alg}), we maintained a commanding lead with a score of 0.64, compared to 0.48 for Llama 3.1 405B. This substantial margin highlights a common limitation in conventional LLMs, which frequently default to abstractive paraphrasing. In contrast, our pipeline strictly enforces the extractive constraints of the task, preserving the exact phrasing of the evidence. As shown in Table 4, in the final competition submission results, our system achieved a significant improvement compared to the official top-performing baseline, gpt-oss-120b. These results strongly confirm that our specialized Dynamic Few-Shot and Reranking pipeline enables the system to outperform both proprietary and ultra-large-scale open-source models.

Discussion. Experiments on reference-based generation show that explicit verification is essential for high-fidelity outputs, regardless of model architecture. On the dev set, our system achieves 0.87 faithfulness, indicating that multi-metric self-consistency reranking effectively filters hallucinations via strict alignment with the provided passages, akin to internal feedback mechanisms in recent autonomous agents. (Shinn et al., 2023). Most notably, in the final official blind evaluation, our system demonstrated exceptional generalization by achieving an even higher state-of-the-art faithfulness score of 0.8974.

A high Answer Quality score (0.81) shows that intent-driven template retrieval effectively mitigates stylistic drift in multi-turn interactions and improves structural modeling. The reranking results further indicate that weighted lexical metrics provide an efficient proxy for LLM-based judges, strictly penalizing ungrounded generation. Although parallel sampling ($N = 3$) introduces some latency, the gains in response trustworthiness justify the trade-off; future work will explore query-aware dynamic sampling to improve efficiency.

5 Conclusions

In this work, we proposed the Intent-Aware Parallel Generation and Reranking System for SemEval-2026 Task 8 (Subtask B), integrating Intent-Driven Template Retrieval with a robust Multi-Metric Reranking mechanism to mitigate hallucinations in reference-based generation. By explicitly modeling conversational intent and applying extractive overlap-based filtering, our pipeline successfully navigates the complex stylistic demands of multi-turn dialogues while preventing generative drift. By enforcing strict extractive consistency between generated responses and gold passages, our approach achieved remarkable success on the official leaderboard, securing the 3rd rank out of 26 teams with a harmonic mean score of 0.7684. Notably, our system surpassed the top official baseline by a significant margin of +0.1294 and achieved a state-of-the-art Faithfulness Score (RL_F) of 0.8974. Future work will focus on distilling these reranking capabilities into a single-pass model and exploring dynamic sampling strategies to optimize inference efficiency without sacrificing generation fidelity.

Acknowledgments

This work was supported by the Scientific Research Fund of Yunnan Provincial Education Department under Grant No. 2026J0006 and by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, pages 1–30.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL-2024*, pages 2318–2335.

DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, and 1 others. 2025. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Dehui Deng, You Zhang, Jin Wang, Dan Xu, and Xuejie Zhang. 2025. YNU-HPCC at SemEval-2025 task 6: Using BERT model with R-drop for promise verification. *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1905–1911.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL-2024*, pages 3563–3578.

Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, and 1 others. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP-2023)*, pages 5567–5584.

Debanjan Ghosh, Kai North, Ekaterina Kochmar, Mamoru Komachi, and Marcos Zampieri, editors. 2026. *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics, San Diego, United States.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems (TOIS)*, 43(2):1–49.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qiao Liu, Jane Dwivedi-Yu, Wen-tau Yih, Graham Neubig, and Luke Zettlemoyer. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP-2023)*, pages 7969–7992.

Zhengbao Jiang and 1 others. 2024. Longrag: Enhancing retrieval-augmented generation for long-context qa. *arXiv preprint arXiv:2406.15319*.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into self-improving pipelines. In *Proceedings of the 12th International Conference on Learning Representations (ICLR-2024)*.

Nelson F. Liu, Kevin Lin, John Hewitt, and 1 others. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics (TACL)*, 12:214–231.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP-2023)*, pages 9020–9036.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics (TACL)*, 11:1316–1331.

Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026a. Mtrag-un: A benchmark for open challenges in multi-turn rag conversations. *Preprint*, arXiv:2602.23184.

Sara Rosenthal, Vraj Shah, Yannis Katsis, and Marina Danilevsky. 2026b. Semeval-2026 task 8: Mtrageval: Evaluating multi-turn rag conversations. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*, San Diego, California.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, and 1 others. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS-2023)*, pages 68539–68551.

Noah Shinn, Federico Cassano, Ashwin Gopinath, and 1 others. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8634–8652.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR-2023)*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, and 1 others. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13484–13508.

Shunyu Yao, Dian Yu, Jeffrey Zhao, and 1 others. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11809–11822.