NCL at SemEval-2026 Task 8: Deterministic Small-LLM RAG with Relation Classification

Zehao Liu, and Huizhi Liang
Newcastle University, Newcastle upon Tyne, UK
marshallcnliu@gmail.com, huizhi.liang@newcastle.ac.uk,

Abstract

We present NCL's system for SemEval-2026 Task 8B, the generation track for multi-turn retrieval-augmented dialogues. Our submission follows a compact and reproducible RAG pipeline: (1) global and local question rewriting with LLM-based multi-turn relation control, (2) passage reranking with BGE-M3, (3) context-level answerability filtering with strict binary LLM judgments ("yes"/"no"), and (4) deterministic inference with a small-llm (Qwen2.5-1.5B-Instruct) plus post-generation quality fallback (cleaning, bad-answer gate, one stricter retry, then an IDK fallback). On the official test set, our system achieved a harmonic mean score of 0.5973 (RB_{agg} 0.4993, RL_F 0.7235, RB_{llm} 0.6105), ranking 19th out of 26 teams on the leaderboard.

1 Introduction

Multi-turn retrieval-augmented generation (RAG) requires balancing three often competing goals: (i) faithfulness to retrieved evidence, (ii) completeness with respect to the user question and dialogue state, and (iii) fluency and helpfulness. The original MTRAG benchmark frames multi-turn RAG as an end-to-end setting over human-generated conversations with active retrieval, long-form answers, non-standalone follow-up questions, and varying answerability (Katsis et al., 2025). MTRAG-UN extends this setting with challenging tasks that emphasize unanswerable, underspecified, and non-standalone questions, as well as unclear responses across six domains (Rosenthal et al., 2026a). SemEval-2026 Task 8 operationalizes this benchmark family as MTRAGEval, covering retrieval, generation with reference passages, and full RAG, with both reference-based and reference-less metrics (Rosenthal et al., 2026b).

Our design goal is a *small-model, deterministic, zero-shot, and conservative* system that can be run locally while minimizing hallucinations. Instead of maximizing stylistic richness, we focus on evidence-grounded answers and a robust "I don't know" (IDK) behavior when the provided passages do not contain enough information.

From a broader perspective, multi-turn RAG is harder than single-turn RAG because retrieval quality depends on dialogue interpretation. In many turns, the latest user utterance is incomplete on its own and relies on references to entities, events, or constraints introduced earlier in the conversation. If a system retrieves only with the last surface-form query, it often misses key evidence and amplifies downstream generation errors. At the same time, naively appending all history can introduce topic drift, noisy lexical signals, and context-window pressure, especially for compact generators. This motivates explicit mechanisms for *history-aware retrieval* and *answerability control*.

Current multi-turn RAG methods usually combine three ideas. First, they reformulate the current question into a standalone query by resolving references from previous turns. Second, they select or summarize only relevant dialogue history rather than passing the full transcript. Third, they apply grounding instructions. These instructions separate dialogue context from retrieved passages. The former is used for interpretation, and the latter is used for factual support. Many systems also use an abstention rule when evidence is missing. The MTRAG benchmark highlights this trade-off. Systems must stay faithful to retrieved passages. At the same time, they must produce answers that are complete and contextually appropriate (Rosenthal et al., 2026a).

Related work in this space has emphasized both retrieval robustness and evaluation robustness. On the modeling side, dense multilingual embeddings and rerankers such as BGE-M3 improve retrieval and matching for diverse inputs (Chen et al., 2024; of Artificial Intelligence , BAAI). On the evaluation side, RAGAS-style faithfulness judges provide complementary reference-less signals when

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1477–1481

reference answers are incomplete or stylistically different (Es et al., 2024). The MTRAG evaluation protocol combines these views through RB_{agg}, RB_{llm}, and RL_F, encouraging systems that are not only fluent but also verifiably grounded (Rosenthal et al., 2026a,b).

Following this line, our submission intentionally prioritizes reliability over aggressiveness. We use deterministic inference with a small instruction model, explicit relevance filtering for both passages and dialogue turns, and a strict answerability policy. This design does not aim for maximal lexical overlap with references. Instead, it aims for stable behavior under noisy retrieval and limited compute, which is a realistic deployment scenario for many local or resource-constrained settings.

2 Task and Evaluation

Task 8B (generation track) provides a user–assistant dialogue history and a set of retrieved passages for each turn. Systems must generate a final answer for the last user query. The official score is the harmonic mean of three metrics (Rosenthal et al., 2026b):

- **RB_{agg}**: a reference-based aggregate metric combining algorithmic measures (including ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019)) as defined by the task.
- **RB_{llm}**: an LLM-as-a-judge reference-based score that evaluates faithfulness, appropriateness, and completeness (Rosenthal et al., 2026b).
- **RL_F**: a reference-less faithfulness judge adapted from RAGAS (Es et al., 2024).

3 System Overview

Figure 1 outlines our pipeline. In this version, the implementation follows a deterministic evidence-grounded generation workflow with an explicit LLM-based multi-turn controller. The main steps are: (1) evidence pre-check, (2) global question rewrite, (3) four-class relation classification against recent turns, (4) class-conditioned history expansion and local rewrite, (5) passage reranking and constrained prompting, and (6) deterministic inference with post-generation quality fallback.

3.1 Step 1: Evidence pre-check

For each sample, we first read the last user query from the dialogue. If the provided `contexts` field is empty, we directly output the fixed abstention string:

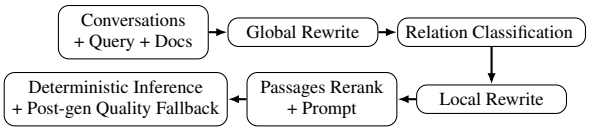

Figure 1: System pipeline for Task B generation.

I cannot find the answer in the given documents.

This enforces strict grounding in the reference setting, where an evidence-backed answer is only attempted when passages are provided.

3.2 Step 2: Global rewrite of the last question (pronoun-aware policy)

Before history classification, we rewrite the last question Q_t into an evidence-faithful standalone query, focusing on semantic completion and coreference resolution, without introducing any new facts. We first apply a regex-based pronoun/coreference detector on Q_t (e.g., *it, this, that, these, those, they, them, he, she, him, her, its, their, there, such, former, latter*). If a signal is detected, we rewrite using only the immediately previous turn; otherwise, we rewrite using the full dialogue history. The two cases use separate fixed prompt templates but share the same constraints, producing the normalized global query Q_t^g (or decomposed queries) for downstream decisions. We use the immediately previous turn in the pronoun-triggered case as a conservative choice to reduce topic drift and avoid injecting unrelated entities from older history. This can miss references whose antecedents appear two or more turns earlier, but the later relation classification and class-conditioned history expansion partially compensate for such cases.

3.3 Step 3: Four-class relation classification

The core controller compares the current question with the previous user question (Q_{t-1}) and previous assistant answer (A_{t-1}) using deterministic LLM judgment. We use four categories exactly as in the code:

- **Category 1**: related to both Q_{t-1} and A_{t-1} (continuation follow-up).
- **Category 2**: related to Q_{t-1} but not to A_{t-1} (new subtopic under same question thread).

- **Category 3**: not related to Q_{t-1} but related to A_{t-1} (answer-triggered follow-up).
- **Category 4**: related to neither (new topic).

This relation-first design explicitly models how multi-turn dependence enters retrieval and generation: dependence on prior user intent, prior assistant content, both, or neither.

3.4 Step 4: Class-conditioned history expansion and local rewrite

After classification, we apply a class-specific history policy:

- **Category 1**: include (Q_{t-1}, A_{t-1}) and recursively check older user turns $(Q_{t-2}, Q_{t-3}, \ldots)$, up to `HIST_USER_TURNS=3`; the recursion continues only if each newly checked turn is still classified as Category 1, and it stops immediately once a turn is classified as Category 2/3/4.
- **Category 2/3**: include only the most recent pair (Q_{t-1}, A_{t-1}) and stop.
- **Category 4**: include no history and stop.

If category is 1/2/3, we perform a second, local rewrite conditioned on the selected history and relation metadata. The local rewrite produces three outputs: (i) a brief summary of the selected history (a few short sentences describing what the user and the assistant discussed in the selected turns), (ii) an explicit intent analysis for the current query Q_t^g (what information and constraints the user is asking for), and (iii) the locally rewritten standalone question Q_t^l, which injects only the necessary constraints from the selected history to remove ambiguity. This Q_t^l is used for passage reranking and final answering. In all cases, history is marked as context-only (for reference resolution), not as factual evidence.

3.5 Step 5: Passage reranking and answerability-gated prompt construction

We rerank the provided contexts using BGE-M3 embeddings (of Artificial Intelligence , BAAI; Chen et al., 2024) with cosine similarity to the locally rewritten query Q_t^l, and keep the top-k passages ($k = 5$). We then concatenate the entire reranked top-k set into a single evaluation prompt and apply an LLM-based Answerability Gate conditioned on (i) the rewritten question Q_t^l and (ii) the selected dialogue history from Step 4. The gate judges whether the *top-k passages as a whole* contain sufficient information to satisfy the user intent, and is constrained to output a strict binary decision ("yes" or "no"). If the decision is "no", the system directly returns the fixed IDK response without entering final answer generation.

3.6 Step 6: Deterministic inference and post-generation quality fallback

We use Qwen2.5-1.5B-Instruct (Team, 2025) as the generator via `transformers` (Wolf et al., 2020). This compact instruction-tuned model balances generation quality with deployment efficiency, making it practical for zero-shot prompting in resource-constrained settings while maintaining strong instruction-following capabilities. To improve reproducibility and reduce variance, we use deterministic inference (`do_sample=false`, `temperature=0`, `top_p=1`) and limit output length (`max_new_tokens=96`).

The model is loaded with 4-bit NF4 quantization via bitsandbytes-style configuration, following the quantization strategy popularized by QLoRA (Dettmers et al., 2023). After generation, we run regex-based cleaning (remove chat markers/prompt echo/role prefixes) and a bad-answer gate (e.g., question restatement, too short, malformed ending, vague numeric response for count questions). If flagged, we execute one stricter retry; if still invalid, we output IDK. Combined with the pre-generation context answerability filter, this yields a two-stage abstention policy: evidence-level abstention before generation and quality-level abstention after generation.

4 Experimental Setup

Table 1 summarizes key hyperparameters (taken from our released implementation).

We run in deterministic mode end-to-end, including relation judgment and rewriting calls, to reduce run-to-run variance and simplify ablation analysis. Each sample therefore follows a fixed sequence: global rewrite $\rightarrow$ category classification $\rightarrow$ local rewrite $\rightarrow$ rerank $\rightarrow$ top-k context answerability filtering $\rightarrow$ answer generation $\rightarrow$ post-generation quality fallback. This makes error attribution easier because the stochastic component is minimized.

In this setup, added robustness mainly comes

Component	Setting
Generator	Qwen2.5-1.5B-Instruct
Quantization	4-bit NF4 (compute bf16)
Max tokens	96
Passage reranker	BGE-M3 embedding cosine
Top passages k	5
History depth cap	3 previous user turns
IDK trigger	Empty contexts, passages unanswerable, or failed retry

Table 1: Core configuration of the submitted system.

Method / Dataset	RB_{agg}	RL_F	RB_{llm}	HM
Proposed method / Competition	0.4993	0.7235	0.6105	0.5973
Proposed method / Reference	0.3221	0.7011	0.6764	0.4992
No Classification / Reference	0.3051	0.6833	0.6973	0.4858

Table 2: Official competition-set result and reference-set evaluation with an ablation that removes history classification/selection (*No Classification*).

from control logic rather than larger model size: two-stage rewriting improves standalone query quality, category-driven history control limits irrelevant context injection, and document-level answerability filtering blocks low-value evidence before generation. The post-generation quality fallback further prevents unstable surface-form failures from propagating to final outputs. The trade-off is higher inference latency due to extra LLM calls, especially for category-1 samples that trigger recursive backward checks and for samples with larger top-k filtering cost.

5 Results and Analysis

On the official competition set (507 tasks), our system obtained an overall score of 0.5973 (RB_{agg} 0.4993, RL_F 0.7235, RB_{llm} 0.6105), ranking 19 out of 26 submissions. We additionally evaluate on a separate reference dataset (842 tasks) to assess generalization, and we include an ablation that removes the history classification/selection step by directly feeding the full dialogue history to the local rewrite stage (denoted as *No Classification* in Table 2). For the proposed method, the reference-set scores are RB_{agg} 0.3221, RL_F 0.7011, RB_{llm} 0.6764, and Harmonic Mean (HM) 0.4992. The relatively strong RL_F reflects stable grounding behavior under the reference-less faithfulness judge, consistent with our strict evidence constraints and answerability gating. In contrast, RB_{agg} is lower because reference-based overlap metrics are sensitive to wording and to missing optional details, which can be amplified by our conservative abstention and concise answering strategy. On the reference dataset, the proposed method outperforms *No Classification* on RB_{agg} (0.3221 vs. 0.3051), RL_F (0.7011 vs. 0.6833), and HM (0.4992 vs. 0.4858), suggesting that class-conditioned history selection reduces prompt noise and helps retrieval and generation stay aligned with evidence, improving both overlap-oriented scoring and faithfulness. Meanwhile, *No Classification* achieves a slightly higher RB_{llm} (0.6973 vs. 0.6764), which we attribute to the judge rewarding answers that appear more complete or conversationally appropriate when the full dialogue context is provided, even though the extra context can also distract the model and slightly degrade evidence-faithful behavior (as reflected by RL_F).

In our manual inspection, the main error modes are:

- **Over-abstention**: the system outputs IDK when the passages contain partial but sufficient evidence, lowering RB_{agg} despite preserving grounding.
- **Underspecified answers**: short answers are often faithful but omit details expected by the reference answer, especially under the 96-token cap.
- **False answerability**: the gate or generator sometimes accepts evidence that is only topically related. For example, for "Which is more important?" the output defined NAV and market capitalization without answering the comparison; and for a phone-interaction security question, the answer drifted to a generic WebSocket description.

Overall, these results indicate a precision–recall trade-off: our proposed pipeline prioritizes faithfulness and controlled generation, which is useful for reducing history noise and improving intent alignment. However, when a task requires richer information, the model may produce under-detailed answers, abstain too early, or misjudge document answerability. The same control logic also increases latency relative to a direct generation baseline, because each non-empty sample can require multiple sequential small-LLM calls before the final answer is produced.

6 Conclusion

This paper presents a compact, deterministic, and reproducible multi-turn RAG system for SemEval-

2026 Task 8B. Our pipeline combines pronoun-aware global rewriting, four-class relation classification with class-conditioned history selection, BGE-M3 top-k reranking, binary answerability gating, and post-generation quality fallback with IDK control. On the official test set, the system reaches a harmonic mean of 0.5973 with strong reference-less faithfulness (RL_F = 0.7235), and on the reference-set ablation it improves HM over the no-classification variant (0.4992 vs. 0.4858), showing the benefit of structured history control. The main limitations are over-abstention, under-detailed answers, and occasional answerability misjudgment. Future work will focus on calibrating abstention thresholds, strengthening reranking/answerability modules, and adding lightweight evidence planning to improve completeness while preserving faithfulness.

Acknowledgements

We thank the MTRAG-UN organizers for building the benchmark and evaluation framework.

References

Jianlv Chen, Shitao Xiao, Peitian Wu, Dongze Lian, Wen Luo, Zhongxiao Wang, and Zheng Xing. 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 7552–7560, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Beijing Academy of Artificial Intelligence (BAAI). 2024. Baai/bge-m3. `https://huggingface.co/BAAI/bge-m3`. Hugging Face model card, accessed 2026-02-13.

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.

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. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.