GenAIus at SemEval-2026 Task 8: Beyond Retrieval with Relevance-Aware RAG for Faithful Multi-Turn Generation

Suveyda Yeniterzi
GenAIus Technologies
suveyda@genaius.tech

Reyyan Yeniterzi
GenAIus Technologies
reyyan@genaius.tech

Abstract

This paper describes our submission to SemEval-2026 Task 8 on multi-turn retrieval-augmented generation (RAG). We propose a hybrid multi-stage pipeline that combines high-recall lexical retrieval, dual-embedding dense re-ranking with reciprocal rank fusion, LLM-based relevance judging, and strictly constrained evidence-grounded generation. Our design emphasizes robustness and faithfulness across the full retrieval-to-generation pipeline. Our results suggest that relevance-aware filtering and constrained generation are important for improving faithfulness and overall RAG performance.

1 Introduction

Retrieval-Augmented Generation (RAG) systems have emerged as a widely adopted approach for enabling large language models to incorporate external knowledge sources during response generation. However, ensuring faithful, context-aware generation in multi-turn conversational settings remains a significant challenge. SemEval-2026 Task 8 addresses this problem by jointly evaluating retrieval quality and evidence-grounded generation in realistic dialogue scenarios (Rosenthal et al., 2026b). The task highlights a central issue in modern NLP systems: strong retrieval alone does not guarantee faithful downstream generation, and effective end-to-end performance requires careful integration of retrieval and reasoning components.

Our approach adopts a hybrid multi-stage architecture that explicitly separates high-recall retrieval from precision-oriented re-ranking and constrained generation. Rather than optimizing any single stage in isolation, our system is designed to improve robustness across the full retrieval-to-generation pipeline.

Our results demonstrate the effectiveness of this design. While achieving competitive performance in retrieval, our system ranked 4th in generation with reference passages and **1st place** in the full end-to-end RAG setting. This contrast suggests that relevance-aware filtering and constrained generation are important factors in downstream performance. In particular, we observe that balanced optimization across retrieval and generation yields stronger end-to-end robustness than maximizing retrieval metrics alone.

2 Background

SemEval-2026 Task 8 focuses on evaluating multi-turn RAG systems under realistic conversational settings (Rosenthal et al., 2026b). The task is built upon the MTRAG benchmark (Katsis et al., 2025; Rosenthal et al., 2026a), which consists of multi-turn user–assistant conversations paired with document corpora. The benchmark is designed to assess both retrieval quality and generation faithfulness in conversational question answering scenarios.

The task is divided into three sub-tasks. **Task A** evaluates retrieval quality by requiring systems to return a ranked list of relevant passages from the corpus. **Task B** evaluates generation quality using gold reference passages. **Task C** evaluates full RAG systems, where systems must first retrieve relevant passages and then generate a faithful answer grounded in those passages. We participated in all three sub-tasks.

3 System overview

We propose a multi-stage hybrid retrieval and generation pipeline combining lexical search, dense semantic re-ranking, reciprocal rank fusion (RRF), LLM-based relevance judgment, and constrained generation. The same core architecture is used across Tasks A, B, and C, with task-specific adaptations described in Section 4.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 2603–2610

3.1 Lexical Retrieval Stage

Our pipeline begins with a high-recall lexical retrieval stage designed to mitigate vocabulary mismatch and maximize candidate coverage.

3.1.1 Query Rewriting

We first prompt an LLM to generate 10 diverse keyword-based queries optimized for sparse retrieval methods such as BM25. The reformulation explicitly incorporates the full conversation history to ensure contextual completeness.

The objective is to maximize lexical recall by expanding the original query with semantically related terms, synonyms, abbreviations, aliases, spelling variations, and alternative phrasings that are likely to appear in relevant documents. By generating multiple diversified keyword queries, we increase term coverage and reduce lexical mismatch between user queries and the document corpus. The full prompt used for this step is provided in Figure 1.

3.1.2 HyDE

In addition to direct lexical rewriting, we apply Hypothetical Document Expansion (HyDE), which uses a generated hypothetical document or answer to improve retrieval for underspecified queries (Gao et al., 2023). We prompt the LLM to generate a synthetic answer passage to the user's question while incorporating the conversation history.

This generated passage functions as a pseudo-document that captures likely terminology, entities, and contextual expressions associated with relevant documents. The pseudo-document is then used as an alternative query for lexical retrieval. This approach enriches short or underspecified queries with contextualized content, further improving recall. The full prompt used for this step is provided in Figure 2.

3.1.3 BM25 and Candidate Pool Construction

We perform two independent BM25 retrieval runs: one using the set of lexical rewritten queries and one using the HyDE-generated pseudo-document. Each run retrieves the top 2000 documents from the corpus.

The resulting document sets are merged and deduplicated to form a high-recall candidate pool of up to 4000 unique documents. This two-stage lexical retrieval strategy balances query diversity (via multi-query rewriting) and contextual enrichment (via HyDE), ensuring broad coverage before semantic refinement.

3.2 Semantic Re-ranking Stage

After constructing a high-recall candidate pool via lexical search, we apply dense semantic re-ranking to improve precision (Karpukhin et al., 2020; Nogueira and Cho, 2019).

3.2.1 Query Rewriting

For dense re-ranking, we generate a self-contained semantic query tailored for embedding-based similarity search. Unlike lexical rewriting, this reformulation focuses on preserving the underlying intent and conceptual meaning of the user's request.

The semantic query integrates relevant conversational context and produces a natural-language representation optimized for embedding models. The reformulation explicitly avoids introducing new assumptions or facts, ensuring fidelity to the original user intent. The full prompt used for this step is provided in Figure 3.

3.2.2 Re-ranking

Re-ranking is conducted only over the lexical candidate pool rather than the full corpus. We perform two independent embedding-based re-ranking runs using: (1) *Qwen3-Embedding-0.6B*, an open-weight dense embedding model (Zhang et al., 2025) and (2) *text-embedding-3-small* (OpenAI, 2024).

Each embedding model produces a ranked list of documents within the lexical candidate set based on vector similarity. By employing two architecturally distinct embedding models, one open-weight model and one proprietary API-based model, we aim to capture complementary semantic representations. This dual-embedding strategy increases robustness against model-specific embedding biases and improves coverage of semantically relevant documents prior to rank fusion.

3.2.3 Reciprocal Rank Fusion (RRF)

To combine the rankings produced by the two embedding models, we apply Reciprocal Rank Fusion (RRF), a rank aggregation method designed to combine multiple retrieval runs into a unified ranking (Cormack et al., 2009). RRF aggregates the ranking signals from both dense re-ranking runs.

This ensemble strategy improves ranking robustness by leveraging complementary embedding spaces. After fusion, we select the top 20 passages for further refinement.

3.3 LLM-based Relevance Judging

To further refine retrieval precision, we employ an LLM-based relevance classifier that evaluates each retrieved passage independently (Faggioli et al., 2023; Zhuang et al., 2024; Saad-Falcon et al., 2024).

Each passage is assigned one of three relevance levels: highly relevant (2), partially relevant (1), or not relevant (0). The judgment considers the original user question, the full conversation history, and the passage content. Passages are assessed independently rather than comparatively.

The classifier outputs the relevance label, a numerical relevance score, a confidence score, and a brief justification. This step enables relevance-aware reranking and filtering before final submission or generation. The full prompt used for this step is provided in Figure 4.

3.4 Constrained Generation

For generation tasks, we adopt a strictly controlled RAG framework designed to maximize faithfulness and ensure alignment with the retrieved evidence.

The generation model receives the full conversation history, the latest user question, a semantically rewritten query, and a set of retrieved passages labeled with relevance levels.

We impose the following constraints:

- **Faithfulness**: The response must be entirely grounded in the provided passages.
- **No External Knowledge**: The model is explicitly prohibited from introducing external information, assumptions, or inferred facts.
- **Fallback Mechanism**: If the passages do not contain sufficient information to answer the question, the model must respond exactly with: "I do not have specific information."
- **Naturalness and Conciseness**: Responses must be fluent, focused, and free of redundancy.

The full prompt used for this step is provided in Figure 5.

4 Experimental Setup and Task Configuration

We outline the experimental setup for Tasks A, B, and C, covering data usage, model configuration, and evaluation methodology.

4.1 Data Usage and Splits

We used the official MTRAG benchmark training, validation, and trial data (Katsis et al., 2025; Rosenthal et al., 2026a) as released by the organizers. The benchmark includes metadata such as question type, answerability, and multi-turn classification; however, this metadata was not used in any of the tasks.

For Task A and Task C, the full document corpus was used as the retrieval space. For Task B, gold reference passages supplied by the organizers were directly used as input to the constrained generation module. No additional external datasets were used.

4.2 Task-Specific Configuration

Although the same core architecture (Section 3) was applied across all tasks, its usage differed per task:

Task A (Retrieval Only): We applied the full hybrid retrieval pipeline described in Section 3, including lexical query rewriting, HyDE-based expansion, dual BM25 retrieval (top-2000 per run), candidate merging and deduplication, constrained dense re-ranking using two embedding models, and RRF.

Following dense rank fusion, we applied the LLM-based relevance judging step to assess each candidate passage independently. Passages labeled as *highly relevant* and *partially relevant* were retained and subsequently reranked within their respective relevance groups to prioritize stronger evidence. The final top-10 passages after relevance-aware reranking were submitted.

Task B (Generation with Reference Passages): Retrieval was not performed. Instead, the constrained generation module (Section 3.4) was applied directly over the gold passages provided by the benchmark. The same faithfulness constraints and fallback mechanism were enforced.

Task C (Retrieval-Augmented Generation): For Task C, we applied the full end-to-end pipeline: hybrid retrieval (Sections 3.1–3.2), LLM-based relevance judging (Section 3.3), and constrained generation (Section 3.4). Although systems could use up to 10 passages, we passed only passages labeled as highly or partially relevant to the generator. After filtering, the generation module received an average of 8.04 passages per question, and 41.2% of examples used fewer than 10 passages. No passages were retained after filtering in 1.5% of exam-

ples; in these cases, the fallback mechanism was used. Overall, the fallback response was generated in 18.3% of cases.

4.3 Models and External Tools

Apart from the BM25 retrieval implementation and the embedding models described above (Qwen3-Embedding-0.6B and text-embedding-3-small), we employed GPT-4o (`2024-11-20`) for all LLM-based stages: lexical query rewriting, HyDE-based synthetic answer generation, semantic query rewriting, relevance judging, and constrained answer generation. All prompts used for these stages are provided in the Appendix.

No task-specific fine-tuning or parameter updates were conducted; all models were used in zero-shot prompting settings.

Our Task C pipeline uses LLM calls at five stages: lexical query rewriting, HyDE generation, semantic query rewriting, relevance judging, and final answer generation. The 10 lexical rewritten queries are generated in a single LLM call, and relevance judgments are batched over the retrieved passages rather than issued as one call per passage.

4.4 Evaluation Measures

The evaluation metrics were defined by the shared task and differ between the retrieval-only and generation settings (Rosenthal et al., 2026b).

Task A (Retrieval). Retrieval performance was evaluated using nDCG@5, with additional reporting at @1, @3, and @10.

Task B and Task C (Generation and RAG). For Tasks B and C, generation quality was evaluated using the harmonic mean of three complementary metrics (Katsis et al., 2025):

- **RB_{alg}**: Reference-based algorithmic metric computed as the harmonic mean of Bert-Recall, Bert-K-Precision, and Rouge-L.
- **RB_{llm}**: LLM-based reference comparison metric.
- **RL_F**: Reference-less faithfulness metric evaluating grounding in retrieved passages.

RB-based metrics compare the model response to the reference answer, while RL_F evaluates faithfulness with respect to the retrieved passages.

Metrics	Task B	Task C
RB_{alg}	0.6086	0.4343
RL_F	0.8896	0.7395
RB_{llm}	0.8603	0.6834
Harmonic Mean	**0.7634**	**0.5861**

Table 1: Official component scores and final harmonic mean for Tasks B and C.

The final system score for Tasks B and C is computed as the harmonic mean of these three components, encouraging balanced performance across completeness, appropriateness, and faithfulness.

5 Results

Official evaluation results for Tasks A, B, and C are reported below.

5.1 Task A: Retrieval Only

For Task A, systems were evaluated using nDCG@5. Our hybrid retrieval architecture achieved a score of **0.4957**, ranking **10th out of 38** participating teams. Our approach outperformed the strongest reported baseline, ELSER with GPT-OSS-20b query rewriting (0.4795).

Although this result indicates competitive retrieval performance, our system did not use ELSER in its retrieval pipeline, which provides useful context for interpreting its retrieval-only ranking. In their retrieval experiments, the official dataset paper reports that ELSER outperformed other retrieval methods, including lexical retrieval with BM25 and dense retrieval with BGE-base 1.5. The organizers therefore note a possible evaluation caveat: "Since we use Elser for retrieval during data creation, there may be some biases towards Elser" (Katsis et al., 2025). This suggests that ELSER alignment may be one possible factor affecting retrieval-only evaluation results.

More importantly, the contrast between our Task A ranking and our stronger Task C result suggests that retrieval-only metrics do not fully capture the effectiveness of an end-to-end RAG system.

5.2 Task B and Task C: Generation and RAG

Tasks B and C were evaluated using the harmonic mean of three metrics: RB_{alg}, RL_F, and RB_{llm}. Table 1 summarizes the component scores and final harmonic mean for both tasks.

Task B (Generation with Reference Passages). Our system achieved a harmonic mean score of

0.7634, ranking **4th out of 26** teams. This substantially outperformed the strongest baseline model, gpt-oss-120b (0.639).

The high RL_F (0.8896) and RB_{llm} (0.8603) scores indicate strong faithfulness and alignment with reference answers. These results suggest that our constrained generation framework effectively controls hallucination while maintaining completeness and appropriateness.

Task C (Retrieval-Augmented Generation). In the full RAG setting, our system achieved a harmonic mean score of **0.5861**, ranking **1st out of 29** teams. This score significantly outperforms the strongest baseline, qwen-30b-a3b-thinking (0.5366).

The contrast between our Task A ranking and Task C ranking suggests that end-to-end RAG performance depends not only on the initial retrieval ranking, but also on how retrieved evidence is filtered, selected, and used during generation. In our system, LLM-based relevance judging removes passages judged not relevant before generation, while the constrained generation prompt enforces passage-grounded answering and a fallback behavior when sufficient evidence is unavailable. These design choices appear to improve downstream faithfulness and answer appropriateness beyond what is captured by retrieval-only metrics.

5.3 Metric behavior across generation tasks

Across both Task B and Task C, our system obtains substantially higher RL_F and RB_{llm} scores than RB_{alg}. The official leaderboards show a similar tendency among top-ranked systems, suggesting that RB_{alg} is generally more conservative with respect to reference-answer coverage and lexical or semantic overlap (Rosenthal et al., 2026b). This pattern helps explain the RB_{alg} gap in our results: the 150-word constraint and strict grounding requirements favor concise, faithful answers, but may omit secondary details or wording present in longer or differently phrased reference answers.

5.4 Ablation Constraints and Future Analysis

The pipeline contains multiple interacting components, including lexical query rewriting, HyDE-based expansion, dual BM25 retrieval, dual-embedding re-ranking with RRF, LLM-based relevance judging, and constrained generation. Quantifying the marginal contribution of each stage requires controlled component-level ablations. In this shared-task setting, such ablations were constrained by the official evaluation protocol: Task B and Task C rely on LLM-based judging, and the organizers restricted teams to one evaluated run per task due to the resource-intensive nature of LLM-judge evaluation. Each ablated variant would therefore require generating additional system outputs and running repeated LLM-based evaluations at scale.

For this reason, we report the official scores for our submitted system and treat component-level ablation as future work. The most important future variants are: removing HyDE, passing the top RRF-ranked passages directly to generation without LLM-based relevance judging, comparing single-embedding and dual-embedding re-ranking, and varying the generation length limit to analyze the tradeoff between reference-answer coverage and faithfulness.

6 Conclusion

In this work, we presented a multi-stage hybrid retrieval and generation pipeline for SemEval-2026 Task 8. Our approach combines high-recall lexical retrieval through query rewriting and HyDE, dense semantic re-ranking with dual embedding models and reciprocal rank fusion, LLM-based relevance judging, and strictly controlled generation to enforce faithfulness.

Our system achieved strong results in generation with reference passages and obtained the best overall performance in the full retrieval-augmented generation setting. The contrast between our retrieval-only ranking and end-to-end RAG ranking highlights an important practical lesson: downstream RAG performance depends not only on retrieving relevant passages, but also on selecting reliable evidence and constraining generation to remain grounded in that evidence.

As discussed above, future work will focus on controlled ablations under reusable evaluation settings. Beyond these ablations, we plan to investigate adaptive context selection, more fine-grained relevance calibration, and tighter integration between retrieval signals and generation constraints to improve efficiency while maintaining faithfulness in multi-turn RAG systems.

References

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms

condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, Boston, MA, USA. ACM.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 39–50, Taipei, Taiwan. ACM.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. 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. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

OpenAI. 2024. text-embedding-3-small. https://developers.openai.com/api/docs/models/text-embedding-3-small. Accessed: 2026-03-01.

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. Association for Computational Linguistics.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2024. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 358–370, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix

```
You are an intelligent assistant tasked with
↪ rewriting the user's latest utterance into
↪ multiple high-quality lexical search queries
↪ optimized for keyword-based document
↪ retrieval.

You will be provided with:
- Conversation History
- Current User Utterance

Task:
Rewrite the current user utterance into 10
↪ related keyword-based queries:
- Maximize recall in BM25 / TF-IDF systems
- Incorporate relevant conversation context
- Diversify using synonyms, technical terms,
↪ acronyms, variations
```

Figure 1: Prompt used for rewriting queries for lexical search.

```
You are an intelligent assistant tasked with
↪ generating an answer of approximately 150
↪ words for the following query by also
↪ considering the previous conversation history
↪ to improve dense retrieval accuracy. The
↪ answer should be a natural passage containing
↪ information relevant to the query.

You will be provided with:
- Conversation History: all prior user and system
↪ messages.
- Current User Utterance: the user's most recent
↪ message.
```

Figure 2: Prompt used for HyDE to generate synthetic answers

```
You are an intelligent assistant tasked with
↪  rewriting the user's latest utterance into a
↪  high-quality search query considering the
↪  provided prior conversation history.
New query should be optimized for embedding-based
↪  semantic retrieval.

You will be provided with:
- Conversation History: all prior user and system
↪  messages.
- Current User Utterance: the user's most recent
↪  message.

Task:
- Generate a semantic query that captures the
↪  user's intent in clear, natural language and
↪  is suitable for semantic similarity search.
- Query must be fully self-contained and include
↪  any necessary context implied by the previous
↪  conversation.
- Do not introduce new concepts, assumptions, or
↪  facts.
- Allow light paraphrasing to better express
↪  intent, and focus on meaning and conceptual
↪  relationships.
```

Figure 3: Prompt used for rewriting queries for semantic search

```
You are an expert information-retrieval
↪  relevance judge.

Your task is to evaluate how relevant each
↪  passage independently is for answering the
↪  user's latest question, taking into account
↪  the full conversation context.
Although you are provided with a rewritten
↪  version of the user query and the full
↪  conversation history, you must base your
↪  judgment primarily on the original user
↪  intent and the most recent user query, not
↪  solely on the rewritten query.

Use the following relevance levels:

- NOT_RELEVANT, score = 0:
  The passage is off-topic, irrelevant, or cannot
  ↪  help answer the question.

- PARTIALLY_RELEVANT, score = 1:
  The passage is related to the topic but only
  ↪  indirectly useful, incomplete,
  or provides background without clearly
  ↪  supporting an answer.

- HIGHLY_RELEVANT, score = 2:
  The passage directly contains facts,
  ↪  explanations, or evidence that would be
  useful in a high-quality answer to the user's
  ↪  question.

Judge each passage independently.
Do not compare passages to each other.
Judge only based on the passage content and the
↪  user's question.

Output ONLY the following JSON object (no extra
↪  text), make sure output is a valid JSON:
{{
  "judgments": [
    {{
      "doc_id": "...",
      "relevance": "highly_relevant |
      ↪  partially_relevant | not_relevant",
      "relevance_score": 2 | 1 | 0,
      "confidence": 0.0-1.0,
      "reason": "short explanation"
    }}
  ]
}}
```

Figure 4: Prompt used for relevance judging of retrieved passages

```
You are an experienced retrieval-augmented
↪  generation (RAG) assistant.

Your task is to generate a response to the user's
↪  latest question using only the information
↪  contained in the provided passages.

Inputs you will receive:
- The full conversation history
- The user's most recent question
- A rewritten version of the user's question
↪  (based on conversation history)
- A list of retrieved passages, each labeled as
↪  HIGHLY_RELEVANT or PARTIALLY_RELEVANT

Constraints:
- The response must be no longer than 150 words
- Use only the provided passages and the
↪  conversation context
- Do NOT use any outside knowledge
- Do NOT add assumptions or inferred facts
- If the provided passages do not contain
↪  sufficient information to answer the
↪  question, respond exactly with:
 'I do not have specific information.'

Answer requirements:
- Faithfulness: Every statement must be directly
↪  supported by the passages
- Appropriateness: Address only the issues raised
↪  by the user's current question
- Completeness: Include all relevant information
↪  available in the passages
- Conciseness: Avoid redundancy or unnecessary
↪  details
- Relevance & Accuracy: Focus on the most
↪  relevant passages
- Naturalness: The answer should be fluent and
↪  readable

Additional rules:
- Do not mention passage labels or document
↪  metadata
- Do not explain your reasoning or cite sources
↪  explicitly
```

Figure 5: Prompt used for generating response