PFW Task 8 at SemEval-2026 Task 8: Lightweight Tri-Fusion Retrieval with Prompt-Engineered Faithful Generation for Multi-Turn RAG

Taleef Tamsal* **and Jonathan Rusert**
Purdue University Fort Wayne
{tamst01, jrusert}@pfw.edu

Abstract

We describe PFW Task 8's system for SemEval-2026 Task 8 (MTRAGEval), a benchmark for multi-turn retrieval-augmented generation across four English-language corpora. Our submission combines BM25, SPLADE-v3, and Jina Embeddings v4 with weighted reciprocal rank fusion for retrieval, plus zero-shot GPT-4o/GPT-4o-mini prompting for generation. Officially, our system ranks 6th of 26 on Task B ($H = 0.756$), 14th of 29 on Task C ($H = 0.533$), and 20th of 38 on Task A ($\mathrm{nDCG@5} = 0.433$). For the camera-ready analysis, we rerun retrieval at the official $\mathrm{nDCG@5}$ cutoff, strengthen the prompt ablation with per-domain statistics and exact tests, and analyze official outputs by answerability and domain. On a balanced 100-example development sample, explicit citation-format instructions—not chain-of-thought alone—raise citation use from 4% to 93%, and a fixed-context Task C control improves from $H = 0.463$ with GPT-4o-mini to $H = 0.523$ with GPT-4o. Official analytics also show near-perfect UNANSWERABLE handling ($H = 0.990$) but weak behavior on UNDERSPECIFIED turns, where the system answers or refuses instead of clarifying. Our code is publicly available.[1]

1 Introduction

Retrieval-augmented generation (RAG) becomes substantially harder in multi-turn settings, where user turns are often elliptical, coreferential, or underspecified relative to the previous conversation (Katsis et al., 2025). This makes retrieval and generation tightly coupled: weak retrieval misses the latent referent, while overly eager generation can still produce fluent but poorly grounded answers.

MTRAGEval (Rosenthal et al., 2026b) provides a shared benchmark for this setting across three subtasks: retrieval only (Task A), generation from gold passages (Task B), and end-to-end RAG (Task C). The benchmark is especially useful because it includes diverse domains and a surprise UNDERSPECIFIED answerability class in the hidden test set, forcing systems to confront ambiguity rather than only answerability (Rosenthal et al., 2026a).

Our system was designed around a pragmatic constraint: we wanted a submission that remained lightweight, easy to reproduce, and feasible without task-specific fine-tuning. For retrieval, we therefore combined one lexical, one learned sparse, and one dense retriever with weighted reciprocal rank fusion. For generation, we used API-based zero-shot prompting with explicit citation constraints instead of supervised adaptation. This design worked well for Task B, where we ranked 6/26 and exceeded the strongest baseline by 18.3%, but it underperformed on Task A and left Task C near baseline.

For the camera-ready revision, we use the available development artifacts and official per-instance analytics to tighten the paper's main claims without revising the official leaderboard results. Our contributions are:

1. We correct the retrieval analysis at the official $\mathrm{nDCG@5}$ cutoff, add a clean query-rewrite comparison, and clarify that the SPLADE-only result is a post-hoc development finding rather than the basis of the submitted run or a revised leaderboard claim.

2. We strengthen the generation ablation by reporting sampling details, per-domain citation rates, Wilson confidence intervals, and exact McNemar tests, showing that citation-format rules—not chain-of-thought alone—drive faithful citation behavior.

3. We expand the official analytics with per-domain and answerability breakdowns, showing near-perfect UNANSWERABLE handling but systematic failure on UNDERSPECIFIED

[1] https://github.com/Taleef7/semeval-2026-task8

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1526–1532

turns, where the model rarely asks clarifying questions.

The remainder of the paper describes our system (§4), experimental setup (§5), and analysis (§6).

2 Task Setting

MTRAGEval evaluates multi-turn RAG on four English-language corpora: ClapNQ, IBM Cloud, FIQA, and Govt. The development set contains 842 tasks and the hidden test set contains 507 tasks. Task A is ranked by nDCG@5. Tasks B and C are ranked by the harmonic mean of three official metrics: **RL_F** (RAGAS faithfulness to the provided passages) (Es et al., 2024), **RB_llm** (LLM-judge answer quality relative to the reference), and **RB_agg** (an algorithmic aggregate of overlap and extractiveness). All three are conditioned by an IDK judge for cases where refusal is the correct behavior (Katsis et al., 2025; Rosenthal et al., 2026b). The hidden test set also introduces UNDERSPECIFIED turns, which are excluded from official ranking but are central to our error analysis (Rosenthal et al., 2026a).

3 Related Work

Recent benchmark work has shown that multi-turn RAG remains difficult even for strong retrievers and LLM generators. MTRAG (Katsis et al., 2025) established the core evaluation setting, while MTRAG-UN broadened the focus to answerability, underspecification, and other open failure modes that standard single-turn benchmarks miss (Rosenthal et al., 2026a). Our paper builds on this evaluation perspective by analyzing where a lightweight system fails rather than only reporting aggregate leaderboard numbers.

The most relevant retrieval literature for our system is conversational query reformulation. Methods such as ConvGQR (Mo et al., 2023), LLM-aided informative rewriting (Ye et al., 2023), and IterCQR (Jang et al., 2024) explicitly rewrite context-dependent turns into stand-alone queries, often improving off-the-shelf retrievers. Our submitted system intentionally omitted this family of methods in order to isolate a simple last-turn retriever, and the camera-ready analysis shows that this simplification likely explains a meaningful part of the Task A gap.

On the retrieval side, our system combines classical lexical retrieval, learned sparse retrieval, and dense retrieval. SPLADE (Formal et al., 2021) and reciprocal rank fusion (Cormack et al., 2009) are natural components for a no-fine-tuning system, while zero-shot dense retrieval approaches such as HyDE (Gao et al., 2023) highlight alternative ways to reduce label dependence. On the generation side, we focus on prompt-level controls rather than new model training. Chain-of-thought prompting (Wei et al., 2022) and structured response constraints are widely used, but our ablation asks a narrower practical question: which prompt elements actually force citation-grounded answers in this benchmark?

4 System Overview

4.1 Retrieval Pipeline (Tasks A & C)

We retrieve from the benchmark-provided passage corpora without re-chunking or document rewriting. In the submitted system, the retrieval query is only the **last user utterance** from the conversation. We chose this simplification deliberately to isolate retriever behavior from conversational reformulation. It should not be read as a claim that last-turn retrieval is best practice for multi-turn search. As the camera-ready development analysis shows, organizer-provided rewrite queries improve both SPLADE-only and tri-fusion retrieval, which is consistent with the stronger query-rewriting baselines reported by the organizers.

We combine three retrieval paradigms via weighted Reciprocal Rank Fusion (RRF):

BM25 ($k_1 = 0.9$, $b = 0.6$): lexical matching with default tokenization. We selected these parameters by grid search over $k_1 \in \{0.9, 1.2, 1.5, 1.8, 2.0\}$ and $b \in \{0.50, 0.60, 0.75, 0.85\}$ on the development set.

SPLADE-v3 (Formal et al., 2021): Learned sparse model (`naver/splade-v3`), max query length 512 tokens. Documents encoded offline; queries at inference.

Jina Embeddings v4: Dense bi-encoder (`jinaai/jina-embeddings-v4`), 1024-dim embeddings. FAISS flat inner-product index (Johnson et al., 2021).

Each retriever returns top-100 candidates. Rankings are fused:

$$\text{score}(d) = \sum_{i=1}^{3} \frac{w_i}{k + r_i(d)} \quad (1)$$

where w_i are the fusion weights, $k = 60$, and $r_i(d)$ is document d's rank in retriever i. We searched 14 valid weight combinations from the

medium grid used in our development scripts, with BM25 weights in $\{0.2, 0.3, 0.4, 0.5, 0.6\}$, SPLADE weights in $\{0.2, 0.3, 0.4, 0.5\}$, and dense weights in $\{0.1, 0.2, 0.3, 0.4\}$, subject to summing to 1.0. Under that development procedure, the best configuration was $(0.20, 0.50, 0.30)$ for BM25, SPLADE, and dense retrieval respectively. We return top 10 documents for Task A and top 5 for Task C. As shown in §6.2, the camera-ready rerun at the official cutoff suggests that SPLADE-v3 is the strongest single retriever and SPLADE+dense is the best last-turn variant. Both are post-hoc development findings and were not part of the original submission decision path.

4.2 Faithful Generation (Task B)

We use GPT-4o (`gpt-4o`, OpenAI API) with a structured three-step prompt (full text in Appendix A). The structure arose from an empirical observation during development: early responses often contained relevant content but omitted citations or cited inconsistently. We therefore separated the prompt into analysis, fact extraction, and synthesis so that citation attachment becomes an explicit intermediate step rather than an implicit style preference.

1. **Analyze** each passage for relevant information.
2. **Extract facts with citations**, noting source passages.
3. **Synthesize** a response with `[Passage X]` on every claim.

Refusal policy: the refusal threshold is implemented *only* as a prompt instruction, not as a separate classifier or heuristic. The model is instructed to refuse (*"insufficient information"*) only when the passages are completely unrelated. If the passages contain partially relevant information, the prompt tells the model to answer with that evidence and cite it. For zero-context tasks, we hard-code refusal. We do not allow parametric knowledge. The final settings are temperature 0.1, max_tokens 500, and top_p 0.95.

4.3 End-to-End RAG (Task C)

Task C combines the retrieval pipeline from §4.1 with **GPT-4o-mini**, temperature 0.3, and max_tokens 300. We also cap the retrieved evidence at five passages to reduce noisy context and prioritize faithfulness. We switched from GPT-4o in Task B to GPT-4o-mini in Task C for cost and throughput reasons during end-to-end generation. To isolate this design choice in the camera-ready analysis, we additionally ran a fixed-context development experiment on a balanced 100-example sample from the organizer-provided `RAG.jsonl` file, keeping the prompt and top-5 evidence constant while varying only the generator. Under identical evidence, GPT-4o improves overall Task C harmonic mean from 0.463 to 0.523 ($\Delta = +0.0588$, 95% CI [0.0186, 0.0981], $p = 0.0052$), with gains in RB_agg (0.349→0.390), RL_F (0.712→0.810), and RB_llm (0.454→0.516). The largest per-domain gains appear on ClapNQ (0.490→0.580), FiQA (0.406→0.523), and Govt (0.483→0.523), while Cloud is slightly lower (0.471→0.461). The model switch therefore matters, but it does not change the broader conclusion that retrieval quality is the larger end-to-end bottleneck.

5 Experimental Setup

Hardware and cost: Retrieval indexing and ablations were run on Purdue Gilbreth GPUs (A100-80GB for indexing; A30 for later analysis). Generation used the OpenAI API at approximately $5 for the full official test run.

Data usage: We used the 842-task development set for retrieval tuning, prompt design, and post-hoc analysis. No supervised training or fine-tuning was performed.

Prompt ablation protocol: For the Task B ablation, we use a fixed balanced 100-example development sample with seed 42, taking 25 examples from each domain. The resulting sample contains 84 ANSWERABLE, 10 PARTIAL, 5 UNANSWERABLE, and 1 CONVERSATIONAL item. For each prompt variant we report citation rate, multi-citation rate, IDK rate, and average response length, then add Wilson 95% confidence intervals and paired exact McNemar tests for citation presence.

6 Results and Analysis

6.1 Official Results

Table 1 presents our official results, provided by the task organizers.

Our strongest result is Task B, where zero-shot GPT-4o outperforms the 120B baseline by 18.3%. RL_F (passage faithfulness) and RB_llm (response quality) are strong; the algorithmic ag-

Task	Score	Rank	Top	BL
A (nDCG@5)	0.433	20/38	0.578	0.480
B (H-mean)	**0.756**	**6/26**	0.783	0.639
C (H-mean)	0.533	14/29	0.586	0.537

	RB_agg	RL_F	RB_llm	H-mean
Task B	0.608	0.892	0.833	**0.756**
Task C	0.420	0.666	0.574	0.533

Table 1: Top: official scores/ranks. BL = top baseline (Task A: ELSER w/ GPT-OSS-20b query rewrite; Task B: gpt-oss-120b; Task C: qwen-30b-a3b-thinking). Bottom: component score breakdown. Task A falls below the baseline; see §6.2.

System	CQ	CL	FI	GV	All
BM25	.183	.202	.073	.228	.174
Dense (Jina-v4)	.358	.242	.230	.262	.276
SPLADE-v3	.442	.356	.305	.379	.373
SPLADE+Dense	**.444**	.356	**.314**	**.379**	**.376**
Tri-opt†	.415	**.361**	.287	.376	.362

Table 2: Retrieval ablation on the development set at the official Task A cutoff (nDCG@5). CQ/CL/FI/GV = ClapNQ/Cloud/FiQA/Govt. †Our official submission. SPLADE+dense is the strongest last-turn variant; SPLADE-v3 remains the strongest single retriever.

gregate RB_agg is the bottleneck. Task A underperformance is analyzed in §6.2.

6.2 Retrieval Ablation

To correct the mismatch in the original submission, we re-ran the retrieval ablation at the official nDCG@5 cutoff on the full development set. Table 2 shows that the earlier *"SPLADE beats fusion"* wording was too strong at this cutoff. The best last-turn configuration is SPLADE+dense (0.376), with SPLADE-v3 alone nearly tied at 0.373. Both outperform the submitted tri-fusion (0.362). The SPLADE+dense gain over tri-fusion is significant under paired bootstrap resampling ($\Delta = +0.0135$, 95% CI [0.0013, 0.0257], $p = 0.0298$). By contrast, the gap between SPLADE+dense and SPLADE-v3 alone is not significant ($\Delta = +0.0028$, 95% CI [-0.0127, 0.0181], $p = 0.694$). The main retrieval lesson is therefore narrower: learned sparse retrieval is the key signal in our setup, while modest dense complementarity can help at the official cutoff.

The submitted tri-fusion setting remained weaker than these simpler variants because BM25 contributes little positive signal on the hardest domains, especially FiQA. Although our camera-ready nDCG@5 rerun clarifies the ranking more cleanly, an earlier development ablation at nDCG@10 had already suggested that SPLADE-only could outperform the submitted tri-fusion setting. We did not treat that diagnostic as the final selection criterion during submission. In retrospect, that was a workflow mistake, and we now make that limitation explicit. The stronger nDCG@5 ablation reported here was computed post hoc for the camera-ready analysis and did not drive the official submission.

Config	Cite%	Multi%	IDK%	Len
Basic	4.0	0.0	1.0	696
+ CoT	4.0	2.0	1.0	976
+ Cite rules	**93.0**	**75.0**	11.0	887
Full v2	**93.0**	**75.0**	11.0	**607**

Table 3: Prompt ablation on a balanced 100-example development sample (25 per domain, GPT-4o). Cite%: responses with ≥1 `[Passage X]` citation.

We also evaluated organizer-provided query rewrites with the same qrels. Rewriting substantially improves both SPLADE-only (0.373→0.419) and tri-fusion (0.362→0.410), with paired-bootstrap 95% confidence intervals entirely above zero for both systems. This development-set evidence supports the reviewers' concern that last-turn retrieval was a major architectural weakness in our Task A system. It is also consistent with prior MTRAG evidence that conversational rewriting helps retrieval on context-dependent turns. We did not run ELSER ourselves. Prior benchmark work on MTRAG reports ELSER as the strongest among the compared lexical, dense, and sparse retrievers, achieving nDCG@5 = 0.45 with last-turn queries and 0.48 with query rewriting (Katsis et al., 2025). That external benchmark evidence is consistent with the official leaderboard gap, but it is not part of our own ablation table.

6.3 Prompt Engineering Ablation

Table 3 reproduces the same qualitative result as in the submission, but now on a stronger balanced development sample. CoT alone leaves the citation rate unchanged at 4% (Wilson 95% CI [1.6, 9.8] for both settings), and the paired McNemar test finds no effect ($p = 1.0$). By contrast, adding explicit citation-format rules raises citation use from 4% to 93% ($p = 3.23 \times 10^{-27}$). The gain appears in every domain: the citation-aware prompt reaches 96% on ClapNQ, 84% on Cloud, 96% on FiQA,

Class	N	RB_a	RL_F	RB_l	H
Answer.	285	.523	.903	.820	.708
Unans.	97	**.990**	**.990**	**.990**	**.990**
Partial	47	.334	.624	.586	.476
Undersp.	78	.039	.039	.039	.039

Table 4: Task B by answerability. H = harmonic mean. RB_a = RB_agg, RB_l = RB_llm. UNDERSPECIFIED is not in official rankings.

Domain	Task B H	Task C H
ClapNQ	.564	.412
Cloud	**.753**	**.556**
FiQA	.638	.408
Govt	.633	.424

Table 5: Official per-domain harmonic means from the organizer analytics. Cloud is the strongest domain in both generation settings; Task C drops relative to Task B in every domain.

and 96% on Govt. The final brevity-constrained `full_v2` prompt preserves the same citation behavior ($p = 1.0$ versus `cot_smartidk`) while substantially shortening responses (887→607 characters on average). This development-set evidence indicates that citation formatting—not CoT alone—is the main driver of grounded generation behavior, while the last prompt revision mainly improves concision.

6.4 Answerability Analysis

Table 4 confirms one of the clearest strengths of our system: near-perfect UNANSWERABLE handling. The same prompt logic is much weaker on UNDERSPECIFIED turns, where the correct action is neither a direct answer nor a flat refusal but a clarification request. On these 78 hidden examples, the official harmonic mean drops to 0.039.

Table 5 shows that the system is strongest on Cloud and weakest on ClapNQ/FiQA, which is consistent with the retrieval difficulty of open-domain and conversational follow-up questions. The end-to-end Task C setting reduces performance in every domain, suggesting that retrieval errors compound even when the Task B prompt remains reasonably faithful once good passages are given.

Underspecified failures. A coarse pass over the 78 UNDERSPECIFIED Task B cases found two dominant behaviors: 43 direct answers and 35 refusals, with no reliable clarification-seeking pattern. Two examples illustrate the failure mode. In an ambiguous-anaphora case, the user asks *"Do you believe this ocean currents play a crucial role in global climate regulation?"* and the model begins *"Yes, ocean currents play a crucial role . . . the North Atlantic Subpolar Gyre . . . "* instead of asking which ocean is meant. In an ambiguous-entity case, the user asks *"Can you tell me what is the secret behind the widespread acclamation of this burger in the fast food industry?"* and the model answers about the *Big Boy hamburger* rather than clarifying which burger is under discussion. In both cases, the model collapses ambiguity to the most salient entity in the retrieved passages. This reinforces that our prompt implements only a binary answer/refuse policy; it has no mechanism for asking clarifying follow-up questions.

7 Conclusion

We presented a lightweight multi-turn RAG system that performed strongly on Task B while remaining below baseline on Task A and near baseline on Task C. The camera-ready revision adds post-hoc development analyses and official-output diagnostics, rather than revised leaderboard claims, to clarify three points that were underdeveloped in the original submission. First, the retrieval story is not that fusion is universally helpful. At the official cutoff, learned sparse retrieval remains the key signal in our setup, and query rewriting matters substantially. Second, our most practically useful generation finding is not about CoT in general, but about explicit citation-format constraints that reliably force grounded answers on the development sample. Third, the official analytics show that our system is competent at refusing truly unanswerable questions but structurally unable to handle underspecified ones, because it does not know how to ask for clarification. A controlled Task C development experiment further shows that GPT-4o is materially stronger than GPT-4o-mini under fixed evidence, but the gain is only partial. Future work should therefore prioritize conversational query reformulation and clarification-aware answerability modeling alongside generator choice.

Limitations

Our post-hoc retrieval analyses, prompt ablation, and Task C generator control are development-set studies rather than hidden-test reruns. They should be interpreted as explanatory camera-ready analyses, not revised leaderboard claims. The fixed-context Task C control isolates generator differ-

ences, but it does not replace a full end-to-end comparison with rewritten retrieval and stronger open models. Finally, generation depends on a proprietary API, and our system still lacks an explicit clarification policy for UNDERSPECIFIED turns.

Acknowledgments

We thank the MTRAGEval organizers for the benchmark and detailed evaluation reports, and Purdue University Research Computing for HPC resources.

References

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 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.

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.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

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, Toronto, Canada. Association for Computational Linguistics.

Yunah Jang, Kang-il Lee, Hyunkyung Bae, Hwanhee Lee, and Kyomin Jung. 2024. IterCQR: Iterative conversational query reformulation with retrieval guidance. 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 8121–8138, Mexico City, Mexico. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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.

Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative query reformulation for conversational search. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4998–5012, Toronto, Canada. Association for Computational Linguistics.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5985–6006, Singapore. Association for Computational Linguistics.

A Prompt Templates

Task B prompt (GPT-4o, system: "You are a helpful assistant that answers questions faithfully based on provided passages. Always cite your sources."):

```
You are a precise, helpful assistant
that answers questions using ONLY
the provided reference passages.

## INSTRUCTIONS
### STEP 1: Analyze the passages
For each passage, identify information
relevant to the question.
### STEP 2: Extract facts with citations
List key facts from the passages,
noting which passage each comes from.
### STEP 3: Synthesize your answer
Combine the extracted facts into a
coherent response. EVERY factual claim
MUST include a citation in
[Passage X] format.

### CRITICAL RULES
1. Use ONLY information from the
```

```
   passages - no external knowledge
2. ALWAYS cite sources: "According to
   [Passage 1], ..." or "... [Passage 2]"
3. If passages don't directly answer,
   STILL extract and cite any RELEVANT
   information
4. Only say "insufficient information"
   if passages are completely unrelated
5. Aim for 2-4 sentences with citations
6. Be helpful - provide what you CAN
   answer based on the passages

## REFERENCE PASSAGES
{passages}
## CONVERSATION CONTEXT
{history}
## CURRENT QUESTION
{question}
## YOUR RESPONSE (with citations):
```

Task C uses the same prompt with GPT-4o-mini (temp=0.3, max_tokens=300).