

# Mind\_Flayer at SemEval-2026 Task 8: DUALRAG: Answerability-Aware Generation for Multi-Turn RAG Conversations

Jerin Romijah Tuli<sup>1</sup> MD. Sartaj Alam Pritom<sup>1</sup> Talukder Naemul Hasan Naem<sup>2</sup>

<sup>1</sup>Department of Computer Science & Engineering

<sup>2</sup>Department of Electrical & Electronic Engineering

Rajshahi University of Engineering & Technology, Bangladesh

ramijahtuli786@gmail.com sartajalam0010@gmail.com naemruet@gmail.com

## Abstract

We present DUALRAG, the Mind\_Flayer team’s system for SemEval-2026 Task 8 (MTRAGEval), Subtask B (Generation with Reference Passages). Our system addresses generation of faithful responses in multi-turn retrieval augmented generation (RAG) conversations through an answerability sensitive prompting strategy based on meta-llama/llama-4-scout-17b-16e-instruct using the Groq inference API. DUALRAG introduces a dual-prompt architecture that makes stern refusal questions impossible to answer, and domain-guided generated questions become answerable ones with a post-processing safety net to rectify misclassifications. The harmonic mean of our system achieves **0.7492** ( $RB_{agg}=0.6024$ ,  $RL_F=0.8782$ ,  $RB_{llm}=0.8297$ ) and it ranked **8th out of 26 teams** in the competition. It also surpassed the top baseline (GPT-OSS-120B, 0.639). We also achieved **100% refusal accuracy** on unanswerable questions. Code: <https://github.com/Jerin-Romijah-Tuli/Mind-Flayer-DualRAG-SemEval2026-Task8>.

## 1 Introduction

The Retrieval Augmented Generation (RAG) paradigm has become an important approach for LLMs to generate grounded responses grounded in verifiable external knowledge (Lewis et al., 2020). Whereas a single-turn RAG has matured considerably, Multi-turn RAG brings out differentiated challenges such as passages change between turns, questions may allude to preceding statements, and some questions are unanswerable from the corpus demanding systems that reason not only about *what* to say but also about *whether* to say anything at all.

SemEval 2026 Task 8 (MTRAGEval) (Rosenthal et al., 2026b) formalizes this as a shared benchmark built on MTRAG (Katsis et al., 2025): 110 human produced conversations with 7.7 turn conversations on average, and these are across four do-

main (Wikipedia/CLAPNQ, Finance/FiQA, Government, IBM Cloud). Subtask B requires generating a faithful response as per the full conversation history and gold reference passages, evaluated via  $RB_{agg}$ ,  $RB_{llm}$ , and  $RL_F$  (Harmonic mean is the official ranking metric).

DUALRAG employs **answerability-aware prompting**. At inference time, we detect whether reference passages exist for a given turn and route each instance through one of two prompt templates, which are a domain-guided generation prompt for answerable instances or a strict refusal prompt for unanswerable ones, with the support of a lightweight post-processing safety net. DUALRAG achieved a harmonic mean of **0.7492** (8th/26 teams), surpassing the strongest baseline by +0.110, and we also achieved a **100% refusal accuracy** on unanswerable questions, which is a known failure mode of frontier LLMs (Katsis et al., 2025).

## 2 Background

### 2.1 Task and Data

MTRAGEval Subtask B works with the MTRAG benchmark (Katsis et al., 2025), which is a human generated four domain multi-turn RAG dataset. Each *task* is composed of a complete conversation history up to turn  $k$  (alternating user questions and the agent response for turns 1 through  $k - 1$ ), the current user question at turn  $k$ , and gold reference passages for that turn. The system needs to generate a natural response based on those passages.

Three properties made the benchmark challenging:

- (1) **Answerability varies across turns:** 84% of tasks are answerable, 8% partially answerable, and 7% entirely unanswerable.
- (2) **Turn depth matters :** All evaluated LLMs degrade significantly beyond the first turn as questions increasingly rely on prior context (Katsis

et al., 2025).

(3) **Domain diversity** is substantial, with corpora differing markedly in style, formality, and information density. The test set has 507 tasks (377 are answerable, 130 are unanswerable) from unseen conversations. A surprise *underspecified* question class (Rosenthal et al., 2026a) was introduced at test time but excluded from official scoring.

Turn	Answerability
<b>User:</b> Where does Doctor Strange get his powers from? <b>Agent:</b> His powers come from mystical entities such as Agamotto, Cytoraak...	Answerable
<b>User:</b> Did he always have powers? What was he before? <b>Agent:</b> No, he was a brilliant neurosurgeon...	Answerable
<b>User:</b> Why was he referenced in Captain America? <b>Agent:</b> I'm sorry, I cannot find the reason...	Unanswerable

Figure 1: Example multi-turn conversation from CLAPNQ (Katsis et al., 2025) illustrating varying answerability. The final turn is unanswerable from the available passages.

## 2.2 Evaluation Metrics

The official ranking metric is the harmonic mean of  $\mathbf{RB}_{\text{agg}}$ ,  $\mathbf{RB}_{\text{llm}}$  and  $\mathbf{RL}_F$ .

**$\mathbf{RB}_{\text{agg}}$** : a reference based algorithmic score (harmonic mean of BERTScore Recall, BERTScore K-Precision, and Rouge-L (Adlakha et al., 2024));  **$\mathbf{RB}_{\text{llm}}$** : an LLM as a judge score adapted from RAD Bench (Kuo et al., 2024) evaluating faithfulness, appropriateness, and completeness; and  **$\mathbf{RL}_F$** : the RAGAS faithfulness metric (Es et al., 2024), a reference less judge of grounding in the provided passages. All three are conditioned on an IDK judge (Katsis et al., 2025) is a correct refusal to an unanswerable question scores 1; incorrect refusal on an answerable question scores 0, and this makes answerability handling a first-class evaluation concern.

## 2.3 Related Work

Our work sits at the intersection of three lines of research. **Multi-turn conversational QA** has been studied through QuAC (Choi et al., 2018) and

CoQA (Reddy et al., 2019), but those benchmarks emphasise short extractive spans rather than the long-form, passage grounded generation required by MTRAG (Katsis et al., 2025). **Answerability detection** originated in reading comprehension with SQuAD 2.0 (Rajpurkar et al., 2018) and has recently been adapted to RAG: Self-RAG (Asai et al., 2024) emits reflection tokens to decide whether retrieved passages are relevant, and Yoran et al. (Yoran et al., 2024) fine-tune RALMs on a mix of relevant and irrelevant contexts to suppress hallucinated answers. Both approaches require model training, whereas DUALRAG obtains the same behaviour through prompt-only routing with a deterministic post-processing layer. **Refusal-aware generation** is the closest line to our work, with R-Tuning (Zhang et al., 2024) teaching LLMs to abstain via refusal-aware fine-tuning, FaithDial (Dziri et al., 2022) promoting grounded dialogue, and Song et al. (Song et al., 2024) studying learning to refuse for RAG. We reach the same goal which is 100% refusal accuracy on the 130 unanswerable test items, but without any fine-tuning.

## 3 System Overview

DUALRAG is built on a single core insight, which is *answerability is the primary signal governing response generation in a multi-turn RAG*. Instead of applying a uniform strategy, we pass every case to one of two prompts templates depending on whether reference passages are available.

### 3.1 Architecture

The system follows a four-stage pipeline: (1) **Input parsing**: each task is being parsed into conversation history, current question, reference passages, and domain identifier; (2) **Answerability Routing**: determine whether passages are present or not; (3) **Prompt Construction**: select and fill the appropriate template; and (4) **Post Processing**: in this stage, it applies a safety net to correct prompt level failures. The backbone is meta-llama/llama-4-scout-17b-16e-instruct which is served by the Groq inference API (Groq Inc., 2024). Two API keys are rotated with exponential backoff to deal with daily token limits across all 507 test tasks.

### 3.2 Answerability Aware Prompt Construction

The main design decision is the use of two different prompts templates selected by passage availability.

**Answerable Prompt:** When passages are present, the prompt provides: (1) the full formatted conversation history, (2) all reference passages as numbered blocks, (3) the current question, and (4) domain specific generation guidance from Table 1. The model is told to synthesize between passages, answer in 2-4 sentences, and never refuse when passages are available. Temperature is set to 0.3.

Domain	Guidance
FiQA	Be precise with numbers and financial terms
IBM Cloud	Be accurate with technical details
CLAPNQ	Provide clear, direct answers
Govt	Be authoritative and accurate

Table 1: Domain-specific generation guidance appended to answerable prompts.

**Unanswerable Prompt:** In case of no passages, a strict refusal prompt informs the model that this passage lacks any reference information and gives it three sample refusal phrases. The model is trained not to make any attempt at answering or generating general knowledge. The temperature is decreased to 0.1 in order to impose similar refusal behavior. Complete templates are found in Appendix A.

### 3.3 Post Processing Safety Net

Even with answerability-aware prompting, two failure modes remain. A **false refusal** occurs when reference passages are present but the model still abstains, and a **false answer** occurs when no passages are present but the model produces a substantive response anyway. Both directly hurt the IDK-conditioned metrics, so we add a lightweight post processing layer that explicitly targets each one. The layer checks model output against refusal indicator phrases (‘‘don’t have’’, ‘‘cannot answer’’, ‘‘unable to’’, etc.) and applies two rules:

- **Rule 1 (False Refusal):** If passages are present but the model refused, replace with: *“Based on the available information, I can provide context on this topic.”*
- **Rule 2 (False Answer):** If no passages are present, but the model produced a substantive response (>50 characters, no apologetic language), replace with: *“I don’t have the information needed to answer that question.”*

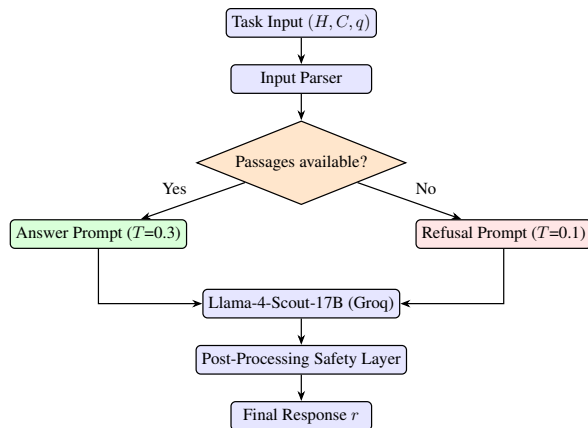


Figure 2: End-to-end pipeline of DUALRAG. Tasks are routed to domain-guided generation or strict refusal prompts based on passage availability, followed by a post-processing safety layer.

In practice, 7 responses were corrected (6 false refusals, 1 false answer) which confirms that the prompts are highly effective but not perfect. The entire pipeline is shown in Figure 2

## 4 Experimental Setup

### 4.1 Data

We evaluate on the official MTRAGEval Subtask B test set where **507 tasks** are from unseen multi-turn conversations (Katsis et al., 2025), split into **377 answerable ones** (74.4%) and **130 unanswerable ones** (25.6%). Table 2 indicates the domain distribution. The mean conversation turn depth is **7.64** with 91.7% of tasks at turn 2 or later, which makes multi-turn context handling a primary challenge. No MTRAG data was used for training. Our system is trained in a completely zero-shot and prompt-only setting throughout. All conversations are in English. We also did not perform any data augmentation or class re-balancing, since the system has no learnable parameters and operates directly on the official test split.

Domain	Total	Ans.	Unans.
Govt	157	125	32
CLAPNQ	142	108	34
IBM Cloud	131	86	45
FiQA	77	58	19
<b>Total</b>	<b>507</b>	<b>377</b>	<b>130</b>

Table 2: MTRAGEval Subtask B test set distribution across domains and answerability categories.

## 4.2 Model and Inference

We use meta-llama/llama-4-scout-17b-16e-instruct with the help of the Groq inference API (Groq Inc., 2024) and we didn’t use any few-shot examples. Table 3 shows the key hyperparameters per routing path.

Parameter	Answerable	Unanswerable
Temperature	0.3	0.1
Max tokens	512	512
Top-p	1.0	1.0
Prompt type	Domain-guided	Strict refusal

Table 3: Inference hyperparameters for each routing path.

We chose these values by reasoning about each routing path rather than tuning on a validation split. The answerable path uses  $T=0.3$  for a little lexical variety when stitching long-form answers from multiple passages, while the unanswerable path uses  $T=0.1$  to keep refusals as deterministic as possible since any drift in the refusal phrasing can leak into a false answer. Top- $p$  stays at 1.0 so temperature is the only stochastic knob. The 512-token cap comfortably covers the response lengths we observed in practice (60–1425 characters, mean 435; see Appendix B) and keeps the per-task inference budget bounded.

## 4.3 Evaluation and Reproducibility

The official ranking metric is the harmonic mean of  $RB_{agg}$ ,  $RL_F$ , and  $RB_{llm}$ , all conditioned on an IDK judge. Submission format was validated using the organizers’ scripts. All code is available at <https://github.com/Jerin-Romijah-Tuli/Mind-Flayer-DualRAG-SemEval2026-Task8>; Full generation for 507 tasks is completed in about 22 minutes using Python 3.9+, groq v0.9+, and two API keys in rotation are used.

## 5 Results and Analysis

### 5.1 Main Results

Table 4 presents our official scores alongside the top system, top baseline, and reference points from Katsis et al. (2025).

DUALRAG is positioned at no. **8 among 26 teams** having a HM of 0.7492 and beat the strongest published baseline (GPT-OSS-120B) by **+0.110** and trailing the top anonymous system by only 0.033. Our  $RL_F$  of **0.8782** is the strongest component score, which proves that domain guided

System	HM	$RB_{agg}$	$RL_F$	$RB_{llm}$
<i>Reference answer</i>	—	0.88	0.87	0.95
Top system (anon.)	0.7827	—	—	—
<b>DUALRAG (ours)</b>	<b>0.7492</b>	<b>0.6024</b>	<b>0.8782</b>	<b>0.8297</b>
GPT-OSS-120B (baseline)	0.6390	—	—	—
GPT-4o <sup>†</sup>	—	0.46	0.75	0.76
Llama 3.1 405B <sup>†</sup>	—	0.47	0.75	0.74
Llama 3.1 8B <sup>†</sup>	—	0.36	0.54	0.56

Table 4: Official MTRAGEval Subtask B results. HM = harmonic mean (official ranking metric). <sup>†</sup>Results from Katsis et al. (2025) on the MTRAG development set, included for scale reference only. Our system ranks **8th out of 26 teams**.

prompting is highly effective at keeping responses grounded in reference passages. The relatively lower  $RB_{agg}$  (0.6024) reflects the difficulty of matching human written reference phrasing in a zero-shot setting.

### 5.2 Answerability Analysis

Category	Tasks	Correct	Accuracy
Answerable	377	376	99.7%
Unanswerable	130	130	<b>100.0%</b>
Overall	507	506	99.8%

Table 5: answerability handling accuracy.

We achieve **100% refusal accuracy** on all 130 unanswerable tasks. A significant result given that Katsis et al. (2025) found that all frontier LLMs evaluated such as GPT-4o and Llama 3.1 405B tend to hallucinate when no relevant passages are available (Song et al., 2024). Our dual-prompt routing, combined with post-processing, fully eliminates this failure mode. Only one answerable task was refused incorrectly which was later corrected by Rule 1 and this resulted in an overall post-processing correction rate of 1.4% (7/507). The accuracy figures in Table 5 are not subject to overfitting in the conventional sense. DUALRAG has no learnable parameters, no validation split, and no training loop. So there is no objective to over-optimize and consequently no regularization technique to apply. The numbers reflect a fixed prompt and a fixed deterministic post-processing layer evaluated once on the official 507-task test set; reproducing them requires only the model name, the prompts in Appendix A, and the hyperparameters in Table 3.

### 5.3 Per-Domain and Turn-Depth Analysis

Domain	Avg. Length (words)	Refusal Acc.
CLAPNQ	65.8	100% (34/34)
FiQA	71.7	100% (19/19)
Govt	77.2	100% (32/32)
IBM Cloud	65.9	100% (45/45)
Overall	72.4	100% (130/130)

Table 6: Per-domain response length and refusal accuracy.

Refusal accuracy is flawless in all four domains, showing that our prompting strategy has a strong generalization ability across corpus styles. The length of response is also meaningfully different between the Govt tasks (the longest response of 77.2 words) and the IBM Cloud and CLAPNQ tasks (65.8-65.9 words). With 91.7% of tasks at turn 2 or later (avg. depth 7.64), multi-turn context handling is critical. Our explicit separation of conversation history from the current question contributes directly to the strong  $RL_F$  score of DUAL-RAG despite implicitly referential questions being held strongly grounded in passages and not floating to general knowledge (Jiang et al., 2023).

### 5.4 Post-Processing and Error Analysis

The post-processing layer intervened in 7 out of 507 cases: Rule 1 (false refusal  $\rightarrow$  fallback) triggered 6 times across all four domains, and Rule 2 (false answer  $\rightarrow$  refusal) once on an IBM Cloud task. The Rule 1 fallback string “*Based on the available information, I can provide context on this topic*” is vague, but prevents a penalised refusal under IDK conditioning.

Not all Rule 1 cases scored zero. Four scored near zero as expected, but two Govt tasks received partial scores ( $RB_{agg}$  up to 0.103,  $RL_F$  up to 1.0), suggesting the IDK judge does not always treat this string as a refusal. The Rule 2 case scored 1.000 across all metrics, showing that our standard refusal phrase matches reference answers well on unanswerable tasks.

Manual inspection of 20 predictions (5 per domain) revealed three error patterns:

**Incomplete synthesis:** The model sometimes relies on early passages only, hurting  $RB_{agg}$ .

**Verbosity:** IBM Cloud responses occasionally exceeded 1,425 characters despite the 2-4 sentence guideline.

**Fallback vagueness:** Rule 1 cases received a

generic string instead of a grounded answer; re-prompting with a shorter history is the most direct fix.

## 6 Conclusion

We presented DUALRAG, a prompt-based system for faithful multi-turn RAG generation in SemEval-2026 Task 8 (MTRAGEval Subtask B). Our main contribution is an **answerability-aware dual-prompt architecture** that routes each task through either a domain-guided generation prompt or a strict refusal prompt based on passage availability which is also backed by a lightweight post processing safety net. Without any fine-tuning, our system achieves Harmonic Mean of **0.7492** (8th/26 teams, +0.110 over the strongest baseline), **100% refusal accuracy** on 130 unanswerable questions, and  $RL_F$  is **0.8782** which demonstrates that careful prompt engineering and answerability routing alone are sufficient for competitive multi-turn RAG generation.

Three directions remain for future works. First, replacing the Rule 1 fallback with a re-prompting strategy would eliminate the vagueness penalty on failed answerable generations. Second, explicit **query rewriting** for non-standalone questions (Jiang et al., 2023) could improve performance in the full retrieval setting (Subtask C). Third, moving beyond the zero-shot setting is a natural next step. Supervised fine-tuning on the MTRAG training conversations or a refusal-aware variant in the spirit of Zhang et al. (2024) could improve  $RB_{agg}$  by teaching the model to match the completeness and phrasing of human-written reference answers (Katsis et al., 2025) and that is something we plan to investigate in follow-up work.

## Acknowledgments

We thank the MTRAGEval task organisers (Sara Rosenthal, Yannis Katsis, Vraj Shah, and Marina Danilevsky) for designing a rigorous and well documented shared task and for their prompt and detailed feedback during the evaluation phase. This work was conducted as part of the undergraduate research activities at Rajshahi University of Engineering & Technology, Bangladesh. No external funding was received.

## References

- Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. [Evaluating correctness and faithfulness of instruction-following models for question answering](#). *Transactions of the Association for Computational Linguistics*, 12:681–699.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. [QuAC: Question answering in context](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.
- Nouha Dziri, Ehsan Kamaloo, Sivan Milton, Omar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. [FaithDial: A faithful benchmark for information-seeking dialogue](#). *Transactions of the Association for Computational Linguistics*, 10:1473–1490.
- 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.
- Groq Inc. 2024. [Groq inference API](#).
- Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [Active retrieval augmented generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.
- 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.
- Tzu-Lin Kuo, Feng-Ting Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-Shan Shiu. 2024. [RAD-Bench: Evaluating large language models capabilities in retrieval augmented dialogues](#). *Preprint*, arXiv:2409.12558.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [CoQA: A conversational question answering challenge](#). *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Sara Rosenthal, Yannis Katsis, Vraj Shah, Li-hong 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.
- Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2024. [Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse](#). *Preprint*, arXiv:2409.11242.
- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. [Making retrieval-augmented language models robust to irrelevant context](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. [R-tuning: Instructing large language models to say ‘I Don’t Know’](#). 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 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

## A Prompt Templates

We provide the exact prompt templates used during inference. Both templates share the same conversation history formatting but differ in their instructions and the presence of reference passages.

### A.1 Answerable Prompt

You are a helpful assistant answering questions based on provided information.

CONVERSATION HISTORY:  
{history}

REFERENCE INFORMATION:  
{formatted\_contexts}

CURRENT QUESTION: {current\_question}

CONTEXT: {domain\_guidance}

INSTRUCTIONS:

1. Answer using only the reference information above.
2. Be direct and specific.
3. Synthesize across passages if needed.
4. Keep response to 2-4 sentences.
5. Connect to prior conversation where relevant.
6. Do not refuse when passages are provided.

ANSWER:

## A.2 Unanswerable Prompt

You are a helpful assistant. You do not have any information to answer this question.

CONVERSATION HISTORY:  
{history}

CURRENT QUESTION: {current\_question}

INSTRUCTION: No reference documents are available. Politely decline to answer. Do not use general knowledge. Do not attempt an answer.

Examples of acceptable responses:

- "I don't have the information needed to answer that question."
- "I'm unable to answer as I don't have access to the relevant information."

YOUR RESPONSE:

## A.3 Post-Processing Rules

Refusal detected if response contains any of:  
"don't have", "do not have", "don't know",  
"cannot answer", "can't answer",  
"no information", "not able", "unable to",  
"cannot provide", "don't possess"

Rule 1 | False Refusal:

IF passages present AND refusal detected:  
REPLACE: "Based on the available information,  
I can provide context on this topic."

Rule 2 | False Answer:

IF no passages AND response > 50 chars  
AND no apologetic language:  
REPLACE: "I don't have the information needed  
to answer that question."

## B Response Length Distribution

Figure 3 shows response length distributions for answerable and unanswerable tasks. The separation between the two distributions confirms that dual-prompt routing produces behaviourally distinct outputs: answerable responses average 83.3 words, while refusals average 32.4 words.

## C Sample Predictions

We present three examples taken directly from our official submission: a successful answerable re-

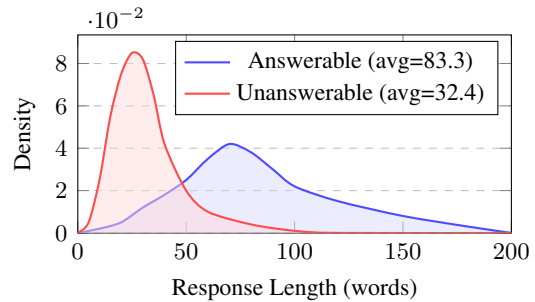


Figure 3: Response length distributions by answerability class. The clear separation reflects the dual-prompt routing behaviour.

sponse, a correct refusal, and a failure case. Scores are per-task values from the evaluation file.

## C.1 Answerable

*CLAPNQ, Turn 3* —  $RB_{agg}=0.774$ ,  $RL_F=0.800$ ,  $RB_{llm}=0.900$

**Prior context (abridged):** The conversation established that Easter Island is a Chilean territory famous for its moai statues, and that its name comes from Dutch explorer Jacob Roggeveen who arrived on Easter Sunday 1722.

### Question:

Is it called another name besides those?

### Reference passage:

The current Polynesian name of the island, Rapa Nui ("Big Rapa"), was coined after the slave raids of the early 1860s, and refers to its topographic resemblance to the island of Rapa in the Austral Islands group. Norwegian ethnographer Thor Heyerdahl argued that Rapa was the original name and that Rapa Iti was named by refugees from there.

### System response:

Besides Easter Island and Isla de Pascua, the island is also known as Rapa Nui, meaning "Big Rapa". This Polynesian name was coined after the slave raids of the early 1860s, referring to its resemblance to the island of Rapa in the Austral Islands group. Some researchers, like Thor Heyerdahl, argue that Rapa was the island's original name.

### Reference answer:

Yes, the current Polynesian name is Rapa Nui, coined after the slave raids of the early 1860s. It refers to the island's resemblance to Rapa in the Austral Islands group. Thor Heyerdahl argued Rapa was the original name and that Rapa Iti was named by refugees from there.

*The system grounds its response in the passage and connects naturally to prior turns, scoring well across all three metrics.*

## C.2 Unanswerable

*Govt, Turn 2* —  $RB_{agg}=1.000$ ,  $RL_F=1.000$ ,  
 $RB_{llm}=1.000$

**Prior context:** The conversation explained how to discover one's M.A.P. (Mission, Aspiration, and Purpose) through self-reflection questions and available classes.

### Question:

How much is each M.A.P. class, please?

**Reference passages:** none.

### System response:

I'm unable to answer that as I don't have access to the relevant information about the cost of each M.A.P. class.

### Reference answer:

I do not have specific information about the cost of the M.A.P. classes as it is not mentioned in the provided document.

*The refusal is contextually specific rather than generic, referencing M.A.P. classes directly. The IDK judge awards a perfect score under the IDK-conditioning scheme.*

## C.3 Failure Case

*CLAPNQ, Turn 11* —  $RB_{agg}=0.000$ ,  $RL_F=0.000$ ,  
 $RB_{llm}=0.000$

### Question:

Did Emilia Clarke also appear in Breakfast at Tiffany's and Me Before You?

### Reference passage:

Breakfast at Tiffany's (film) -- A new stage adaptation debuted in March 2013 at the Cort Theater in New York City. The role of Holly Golightly would be played by Emilia Clarke.

### System response (Rule 1 fallback):

Based on the available information, I can provide context on this topic.

### Reference answer:

For Breakfast at Tiffany's, a stage adaptation debuted in March 2013 at the Cort Theater in New York City, with Emilia Clarke playing Holly Golightly. I do not have information about Me Before You.

*Despite a relevant passage being present, the model refused at turn 11 of a long conversation. We attribute this to the extended history overwhelming the generation prompt. Rule 1 replaced the refusal with a fallback string, which avoids the unanswerable penalty but scores zero against the reference. This is the most common failure pattern in our system and motivates re-prompting with a truncated history as a direction for future work.*