Well Begun Is Half Solved: Future-Aware Route Optimization for Long Narrative Question Answering

Junhyung Kim[*], **Chanhee Lee**[*], **Sungjun Ha, Yoonsoo Kim, Hayoung Oh**[†]

Sungkyunkwan University

[†]Corresponding author

{leechanhye, at100, hhssjj0521, arthur0303, hyoh79}@skku.edu

Abstract

Long narrative question answering requires identifying sparse and delayed evidence across extended story contexts. Existing chunk- or agent-based methods often explore multiple paths but resolve uncertainty only after each path has generated a complete answer. We argue that this late answer aggregation is ill-suited for narrative QA, where the key decision is which evidence route should be trusted before answer generation. We propose **FARO**, **F**uture-**A**ware **R**oute **O**ptimization, an inference-time framework that constructs complementary evidence routes, probes their future answerability, and commits to a single route before final reading. FARO then generates the answer only from the selected route, without post-hoc consensus, majority voting, or route-answer pooling. On DetectiveQA with Llama3.1-8B-Instruct, FARO achieves 56.49% accuracy on 154 instances, outperforming the strongest reported baseline while reducing the none-rate. Ablations show that the recency anchor and conservative commitment rule are essential for stable performance.

Code — https://github.com/Bellissimo-AI/FARO

1 Introduction

Long narrative question answering requires a model to answer questions over extended story contexts, where relevant clues are often sparse, indirect, and separated by long narrative distances. In this setting, the main challenge is not simply that the input is long. The model must decide which evidence path should be followed before a reliable answer can be produced. This challenge is amplified in long narrative contexts because the evidential value of a passage often depends on future context. Locally salient information may later turn out to be misleading, while seemingly minor details may become decisive only when connected to later events.

Recent long context reasoning methods often address long inputs by decomposing a document into smaller chunks. This decomposition is necessary because many long documents exceed the finite context window of language models, and even when the input can fit into the context window, processing it as a single flat sequence can make evidence access unreliable. Chunk-based methods therefore process multiple local paths, agents, or retrieved contexts to improve evidence coverage.

However, this strategy introduces another problem. These approaches often resolve uncertainty only after each path has already generated a complete answer. This creates a structural limitation for long narrative QA. A route based on incomplete or misleading evidence can still produce a plausible answer, and that answer may survive into the final aggregation stage. Majority voting, answer pooling, or post hoc consensus can compare completed answers, but they do not directly solve the earlier problem of choosing a reliable evidence route before answer generation.

We argue that the central decision in long narrative QA is not which completed answer should win after aggregation, but which evidence route should be trusted before answer generation. In other words, route uncertainty should be resolved before final reading, rather than after multiple routes have already produced answers. This shifts the focus from late answer aggregation to early evidence-route commitment.

To this end, we propose **FARO**, **F**uture-**A**ware **R**oute **O**ptimization, an inference-time framework for long narrative question answering. FARO treats candidate routes not as answer candidates, but as possible evidence futures. It first constructs a small set of complementary evidence routes, probes each route using only early evidence, forecasts whether the route is likely to become answerable after deeper reading, and commits to a single route before final answer generation. The final reader then receives only the selected route, while all unselected routes and their tentative probe outputs are discarded.

FARO instantiates this idea with three evidence routes: (1) an answer-neighborhood route that anchors the current narrative state, (2) a question-option salience route that retrieves locally relevant evidence, and (3) an option-bridge route that preserves potential links between candidate answers and distant narrative events. A conservative online commitment rule switches away from the anchor only when another route is both sufficiently strong and sufficiently better than the anchor. This allows FARO to perform lightweight lookahead over evidence trajectories without expanding complete reasoning trees or generating competing final answers from each route.

[*]These authors contributed equally.

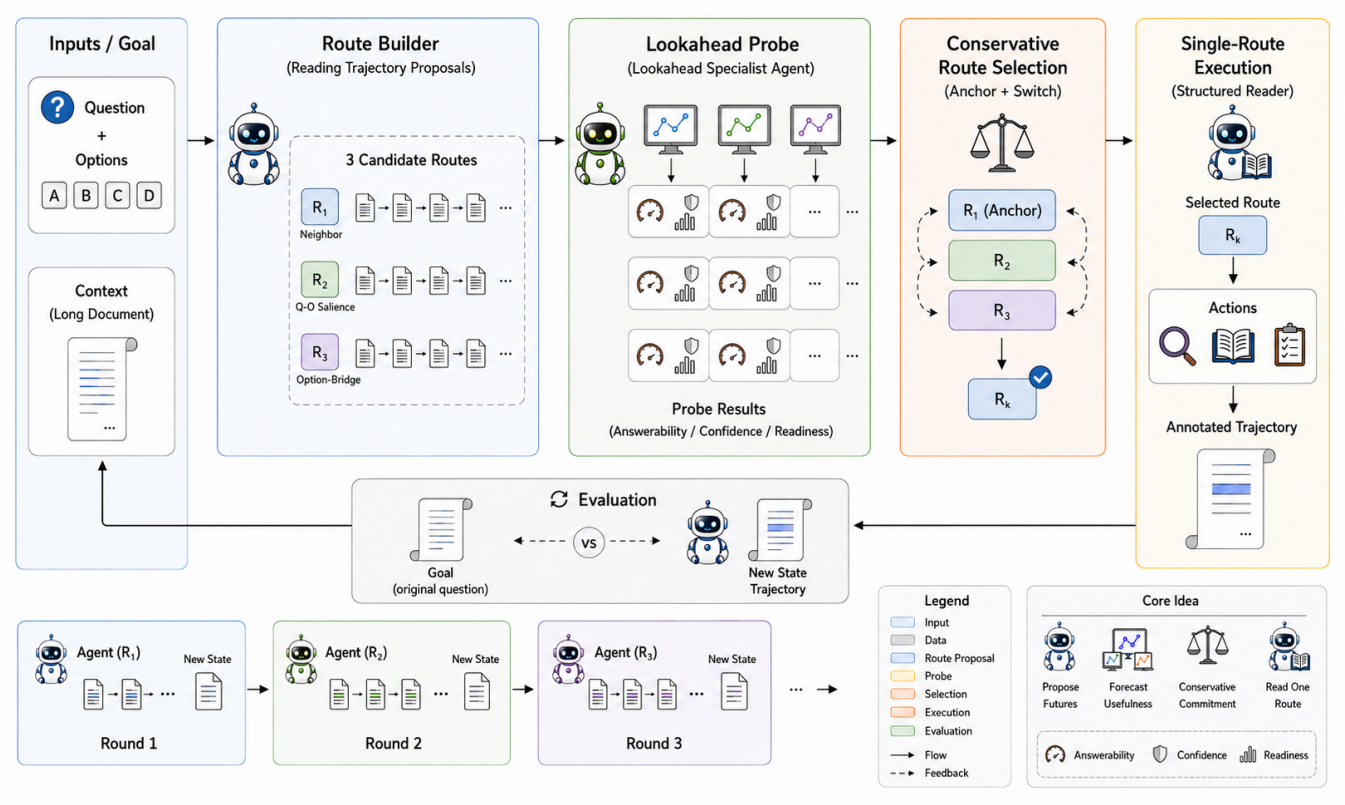

Figure 1: Overall architecture of FARO. FARO constructs multiple candidate evidence routes, probes their future answerability, commits to one route before final answer generation, and performs structured reading only on the selected route.

We evaluate FARO on DetectiveQA using Llama3.1-8B-Instruct as the backbone model. FARO achieves 56.49% accuracy on the evaluation set, outperforming the strongest reported baseline (e.g., ToA (Yu et al. 2025) while reducing the none-rate. Ablation results show that the answer-neighborhood route provides an important recency anchor and that conservative online commitment is essential for stable performance. These results support our central claim that long narrative QA benefits from resolving evidence-path uncertainty before answer generation.

Our contributions are summarized as follows:

- We propose **FARO**, an inference-time framework that forecasts route answerability from early evidence and performs final reasoning only on the selected route.
- We introduce a conservative anchor-based commitment rule that enables lightweight lookahead over possible evidence futures without expanding complete reasoning trees or aggregating route-level answers.
- We show through experiments and ablations on DetectiveQA that FARO improves long narrative QA by committing to a useful evidence trajectory before final answer generation.

Related Work

Long Context Reasoning and Chunk Based Processing

Long context reasoning requires models to identify and integrate evidence distributed across distant parts of an input document. However, standard LLMs have a predefined context window, which limits the amount of text that can be processed in a single forward pass. For example, LLaMA3.1-8B-Instruct (Grattafiori et al. 2024) supports a context length of 128K tokens, which is long but still finite (Meta AI 2024; Meta Llama 2024). Moreover, even when a long input fits within the context window, prior works have shown that LLMs may fail to use information uniformly across the input (Yu et al. 2025), especially when relevant evidence appears in the middle of the context (Liu et al. 2024). These limitations suggest that long context reasoning is not only a matter of increasing context length, but also a matter of deciding how to access and organize distributed evidence.

Chunk-based processing addresses this problem by decomposing a long document into smaller evidence units and performing local reasoning before global answer synthesis. Recent agent frameworks extend this idea by assigning different document segments to separate agents or processing units. Chain of Agents (Wei et al. 2022) uses worker agents for local context processing and a manager agent for final synthesis (Zhang et al. 2024), while Tree of Agents (Yu et al.

2025) explores multiple reasoning paths over chunk permutations with caching and pruning. XpandA (Xiao et al. 2025) further improves long context processing through dynamic partitioning, question driven collaboration, and selective replay. Although these methods improve evidence coverage, they still leave an important planning problem unresolved: the system must decide which evidence units should be considered, how they should be organized, and when enough information has been gathered for answer generation.

Evidence Representation for Route Planning

Effective long context reasoning requires more than splitting a document into smaller segments. Each segment must be represented in a form that helps the model decide whether it can contribute to the final answer. Existing chunk-based agent methods use local summaries, extracted evidence, or inter-agent messages to transfer information across chunks (Zhang et al. 2024; Yu et al. 2025; Xiao et al. 2025). However, these representations are mainly used for communication or final answer synthesis, rather than for predicting which evidence path should be followed before answer generation.

This issue is related to document and prompt compression, where long inputs are compressed into task useful representations before downstream inference (Xu, Shi, and Choi 2023; Li et al. 2023; Jiang et al. 2023; Pan et al. 2024). However, our goal is different from simply reducing input length. We use compact evidence representations as planning states for route selection. In our framework, paragraphs are treated as basic evidence units, and candidate routes are constructed to capture different types of answer bearing evidence, such as late local evidence, question option salience, and option bridge evidence. This allows the model to evaluate not only whether a paragraph is relevant, but also whether an early part of a route is likely to lead to an answerable evidence path.

Efficient Path Selection for Long Context Reasoning

Search based reasoning methods show that exploring multiple candidate paths can improve robustness, but they often require repeated expansion, scoring, rollout, or backtracking (Yao et al. 2023; Hao et al. 2023; Zhou et al. 2024). Adaptive branching and value guided methods reduce unnecessary exploration by estimating whether a branch is worth expanding (Li 2025; Zhang 2025; Lee et al. 2024). Nevertheless, these methods still generally rely on maintaining and comparing multiple candidate paths, which can be expensive in long context reasoning because each path may require additional evidence reading and answer refinement.

Our work addresses this limitation through future aware route selection. Instead of expanding many chunk trajectories or aggregating answers after multiple routes have already been read, we construct a small set of candidate evidence routes over paragraph level units and estimate their future answerability from early evidence. The model then commits to a single route before final answer generation and closes the remaining routes. This design preserves the benefit of chunk based evidence access while avoiding exhaustive route expansion. Compared with prior chunk based agent frameworks, our architecture shifts the key decision from late answer aggregation to early evidence route commitment.

Method

Figure 1 illustrates the overall architecture of FARO. Given a question, four answer options, and a long narrative context, FARO first constructs multiple candidate evidence routes that represent different hypotheses about where useful clues may appear. Each route is then evaluated with a lightweight early probe that estimates whether the route is likely to become answerable after deeper reading. Based on this route-level forecast, FARO commits to a single route before final answer generation. The selected route is passed to a structured reader, while all other routes are discarded. Thus, FARO performs future-aware route selection without aggregating multiple completed answers.

Problem Setup

We consider long-form multiple-choice question answering over a long narrative context. Given a context C, we represent it as an ordered sequence of paragraphs, $C = \{p_1, p_2, \ldots, p_n\}$. Each instance consists of the context C, a question q, and four answer options $O = \{A, B, C, D\}$. The goal is to predict the correct answer $\hat{y} \in O$.

We follow the inference setting of DetectiveQA. For each question, the context contains only the story text available before the answer point, so the model must reason from the information that would be accessible at that point in the narrative. During inference, FARO does not use gold reasoning annotations, clue positions, clue contexts, distractor annotations, or answer labels. Thus, FARO is an inference-only procedure built on top of an instruction-tuned language model. In our experiments, we instantiate the backbone with Llama-3.1-8B-Instruct (Grattafiori et al. 2024).

Under this setting, FARO formulates long-context reasoning as an online evidence-routing problem. Instead of reading the entire narrative as a flat input or allowing multiple routes to independently generate final answers, FARO estimates which evidence route is likely to become useful and commits to that route before final answer generation. This design is motivated by the observation that long-form detective questions often depend on sparse but decisive evidence. A useful inference procedure should therefore decide not only which evidence is locally relevant, but also which evidence path is likely to become answerable after further reading.

Future Evidence Route Construction

FARO first constructs a small set of candidate evidence routes:

$$R = \{r_1, r_2, r_3\}. \tag{1}$$

Each route corresponds to a different hypothesis about where answer-supporting evidence may appear. In this work, r_1 denotes an answer-neighborhood route that anchors the current narrative state, r_2 denotes a question-relevant route

that retrieves evidence directly related to the query, and r_3 denotes a bridge route that preserves potentially useful intermediate evidence connecting earlier context to the current situation. These routes are not answer candidates. Rather, they are candidate reading trajectories that preserve different possible evidence futures before the model commits to one path.

The first route is the *answer-neighborhood route*. It contains the final k_{recent} paragraphs of the provided context:

$$r_1 = \text{AN}(C). \tag{2}$$

This route is used as a recency anchor because the last part of the available context defines the immediate narrative state at the question point. Rather than assuming that the answer is located near the end, this route provides a stable reference for the current characters, events, and unresolved situation before exploring alternative evidence paths.

The second route is the *question-option salience route*:

$$r_2 = \text{QS}(q, O, C). \tag{3}$$

This route selects paragraphs that overlap with the question and answer-option terms. It is intended to capture explicit evidence when important entities, events, or phrases are directly mentioned in the narrative.

The third route is the *option-bridge route*:

$$r_3 = \text{OB}(O, C). \tag{4}$$

This route selects option-balanced paragraphs that may connect candidate answers to supporting or refuting narrative events. It is useful when the decisive evidence is not contained in a single local span, but emerges through a bridge between an earlier option-related clue and a later narrative event.

Here, AN, QS, and OB denote the answer-neighborhood, question-option salience, and option-bridge route constructors, respectively. The route set is intentionally small so that FARO can preserve complementary evidence hypotheses without performing expensive tree expansion.

Early Route Probing

After constructing the route set, FARO probes each route using only a short prefix of its evidence. For a route r_i, let E_i^{probe} denote the partial evidence used for probing. The backbone model produces a structured route forecast:

$$z_i = M_\theta^{\text{probe}}(q, O, E_i^{\text{probe}}), \tag{5}$$

where M_θ is the backbone language model.

The forecast predicts whether the partial evidence in route r_i is a promising starting point for obtaining sufficient evidence through further reading. From the forecast z_i, FARO extracts two route-level quantities:

$$a_i = \text{Ans}(z_i), \qquad c_i = \text{Conf}(z_i). \tag{6}$$

Here, a_i is the answerability score and c_i is the confidence score. The answerability score measures how likely the route is to become answerable after additional evidence is read, while the confidence score measures the reliability of this estimate.

Although the probe may produce a tentative answer for traceability, FARO does not use that answer for final prediction. The probe is used to decide which route to read, not to vote over answer options. This distinction is central to FARO: early route probing forecasts future evidence utility rather than generating competing final answers.

Online Route Commitment

FARO next commits to exactly one route before final answer generation. For each route r_i, FARO computes a route-readiness score $s_i = a_i + \lambda c_i$, where λ controls the contribution of confidence. We set $\lambda = 0.5$.

The answer-neighborhood route r_1 is used as the anchor route, denoted by $r_{\text{anchor}} = r_1$. Among the remaining routes, FARO identifies the strongest challenger by selecting the route index with the highest readiness score:

$$j = \underset{i:\, r_i \in R \setminus \{r_{\text{anchor}}\}}{\arg\max} s_i. \tag{7}$$

The strongest challenger is then given by r_j.

The model switches away from the anchor only when the challenger is both sufficiently strong and sufficiently better than the anchor. The selected route is defined as:

$$r_{\text{sel}} = \begin{cases} r_j, & \Delta_j \geq \delta \text{ and } s_j \geq \gamma, \\ r_{\text{anchor}}, & \text{otherwise}, \end{cases} \tag{8}$$

where $\Delta_j = s_j - s_{\text{anchor}}$. In our implementation, we set $\gamma = 1.15$ and $\delta = 0.65$.

This rule is deliberately conservative. Because the answer-neighborhood route provides a recency anchor for the current narrative state, FARO keeps the anchor unless another route shows clear evidence of higher future utility. At the same time, the commitment rule allows FARO to switch to the question-option salience route or the option-bridge route when either route appears substantially more promising. Thus, FARO performs lightweight lookahead over possible evidence trajectories without expanding complete reasoning trees or generating competing final answers from each route.

Structured Single-Route Reading

Once a route is selected, FARO closes all other routes. The final reader receives only the question, the answer options, and the full evidence from the committed route:

$$E_{\text{sel}} = \text{Full}(r_{\text{sel}}). \tag{9}$$

It does not receive tentative answers, rationales, confidence scores, or probe outputs from discarded routes. This prevents preliminary route-level predictions from being reused as answer votes.

The final reader is prompted to produce a structured response. It first extracts compact evidence facts from the selected route, then assesses each answer option against those facts, and finally returns one answer with confidence and a short explanation:

$$o_{\text{read}} = M_\theta^{\text{read}}(q, O, E_{\text{sel}}). \tag{10}$$

Algorithm 1: FARO Inference

Require: Question q, options O, context C, backbone model M_θ
Ensure: Final answer $\hat{y} \in O$

$r_1 \leftarrow \mathrm{AN}(C)$
$r_2 \leftarrow \mathrm{QS}(q, O, C)$
$r_3 \leftarrow \mathrm{OB}(O, C)$
$R \leftarrow \{r_1, r_2, r_3\}$
for $r_i \in R$ **do**
  $E_i^{\mathrm{probe}} \leftarrow \mathrm{Prefix}(r_i)$
  $z_i \leftarrow M_\theta^{\mathrm{probe}}(q, O, E_i^{\mathrm{probe}})$
  $a_i \leftarrow \mathrm{Ans}(z_i)$
  $c_i \leftarrow \mathrm{Conf}(z_i)$
  $s_i \leftarrow a_i + 0.5c_i$
end for
$r_{\mathrm{anchor}} \leftarrow r_1$
$s_{\mathrm{anchor}} \leftarrow s_1$
$j \leftarrow \arg\max_{i \in \{2,3\}} s_i$
if $s_j \geq 1.15$ and $s_j - s_{\mathrm{anchor}} \geq 0.65$ **then**
  $r_{\mathrm{sel}} \leftarrow r_j$
else
  $r_{\mathrm{sel}} \leftarrow r_{\mathrm{anchor}}$
end if
$E_{\mathrm{sel}} \leftarrow \mathrm{Full}(r_{\mathrm{sel}})$
$o_{\mathrm{read}} \leftarrow M_\theta^{\mathrm{read}}(q, O, E_{\mathrm{sel}})$
$\hat{y} \leftarrow \mathrm{Answer}(o_{\mathrm{read}})$
return $\hat{y}$

The final prediction is obtained directly from this reader output:

$$\hat{y} = \mathrm{Answer}(o_{\mathrm{read}}). \tag{11}$$

No post-hoc consensus, majority voting, route-answer pooling, self-consistency over independently generated answers, or deterministic lexical thresholding is applied. Once a route is selected, FARO performs final reasoning only over the committed route, making it a strictly single-route reasoning method after route commitment.

Implementation Details

In our implementation, the answer-neighborhood route uses the final 16 paragraphs of the provided context. The question-option salience route selects up to 18 paragraphs, and the option-bridge route selects up to 24 paragraphs. FARO probes each route using 5 paragraphs and limits the final selected route to 4200 tokens. We set the confidence weight to $\lambda = 0.5$, the challenger readiness threshold to $\gamma = 1.15$, and the margin threshold to $\delta = 0.65$, based on empirical tuning on development examples. The final system uses three routes, anchor-score commitment, generation-based structured reading, and no reader critic.

Experiments

In this section, we describe the evaluation dataset, metrics, and experimental results used to evaluate FARO. In this version, we focus our empirical evaluation on DetectiveQA, a long-form narrative question answering benchmark that requires understanding and reasoning over long story contexts. We use DetectiveQA as the primary benchmark to assess FARO under a realistic long-context narrative QA setting.

Evaluation Datasets

Experiments are designed around two long-context reasoning datasets and one long-context retrieval benchmark: DetectiveQA.

DetectiveQA (Xu et al. 2024) is a long-document multiple-choice question answering dataset based on detective novels. Since detective narratives contain scattered clues, misleading events, and delayed revelations, the dataset is well suited for evaluating whether a model can identify useful evidence paths before producing a final answer. Following the DetectiveQA inference setting, each context contains only the story text available before the answer point, and FARO does not use gold reasoning annotations, clue positions, clue contexts, distractor annotations, or answer labels during inference.

In this version, we focus our evaluation on DetectiveQA and leave broader cross-dataset validation for future work. Although NovelQA (Wang et al. 2025) and Needle-in-a-Haystack (Nelson et al. 2024) are relevant long-context benchmarks, they pose different practical and methodological constraints. NovelQA requires large scale inference over full length novels and official leaderboard based evaluation, while Needle-in-a-Haystack mainly tests controlled retrieval rather than narrative route commitment. Since our primary goal is to validate whether early evidence route commitment improves long narrative QA, DetectiveQA provides the most directly aligned evaluation setting. Nevertheless, evaluating FARO on broader narrative and retrieval benchmarks remains an important direction for future work.

Evaluation Metrics

For DetectiveQA, we evaluate performance using two metrics: **Accuracy** and **None-rate**. Accuracy measures the percentage of questions for which the model selects the correct answer among the four answer options. None-rate measures the percentage of instances where the model explicitly indicates that it cannot retrieve sufficient evidence to answer the question. A lower none-rate indicates that the model more consistently produces answerable predictions rather than abstaining due to missing evidence.

Main Results

Table 1 reports the main results on DetectiveQA using Llama3.1-8B. FARO achieves an accuracy of 0.565, outperforming the strongest reported baseline, TOA, which achieves 0.543. FARO also obtains the lowest none-rate, 0.008, compared with 0.017 for TOA. These results indicate that FARO improves answer accuracy while reducing the frequency of noncommittal outputs.

The improvement is notable because FARO does not aggregate independently generated route answers. Unlike vote-style methods or self-consistency-based approaches, FARO commits to a single evidence route before final answer generation and performs structured reading only on that route. This suggests that forecasting route utility before answer generation can be more effective than generating multiple answers and resolving them afterward. In other words,

Table 1: Main results on DetectiveQA using Llama3.1-8B. Accuracy and none-rate are reported following the four-option QA evaluation protocol. Higher accuracy is better, and lower none-rate is better.

Method	Acc. ↑	None ↓
COA	0.253 ± 0.012	0.370
LONGAGENT	0.487 ± 0.031	0.157
LongLLMLingua	0.307 ± 0.005	0.260
LongRAG	0.370 ± 0.000	0.217
TOA	$\underline{0.543 \pm 0.009}$	$\underline{0.017}$
Sequential	0.400 ± 0.028	0.143
Vote	0.330 ± 0.022	0.023
FARO (ours)	$\mathbf{0.565 \pm 0.000}$	**0.008**

Table 2: Ablation results on DetectiveQA. Accuracy measures the proportion of correct four-option predictions, and None denotes the none-rate. The best result is shown in bold, and the second-best result is underlined.

Ablation	Acc. ↑	None ↓
FARO full	$\underline{0.565}$	**0.008**
w/o Answer-Neighborhood Route	0.325	0.013
w/o Question-Option Salience Route	**0.571**	**0.008**
w/o Option-Bridge Route	$\underline{0.565}$	$\underline{0.009}$
w/o Conservative Gate	0.351	0.019
LLM-based Route Commitment	0.422	0.016

FARO improves long-context QA not by increasing answer-level aggregation, but by selecting a more useful evidence trajectory before final reasoning.

Ablation Study

We conduct an ablation study to examine how each component of FARO contributes to the final performance. Following the method design, we evaluate two groups of ablations: route ablations and commitment ablations. The route ablations remove one of the three evidence routes used in FARO, while the commitment ablations replace the conservative online route commitment rule with less constrained alternatives. Table 2 summarizes the results on the DetectiveQA evaluation set.

The route ablations show that the answer-neighborhood route is the most important evidence route in FARO. Removing this route reduces accuracy from 0.565 to 0.325, producing the largest drop among all route ablations. This result supports the role of the answer-neighborhood route as a recency anchor for the current narrative state. The route does not assume that the decisive evidence is located near the end of the context; rather, it provides a stable reference for the current characters, events, and unresolved situation at the question point. Without this anchor, FARO must rely only on salience-based or option-balanced evidence, which is insufficient for many questions that require interpreting earlier clues in relation to the current narrative state.

Removing the question-option salience route slightly improves accuracy from 0.565 to 0.571, while keeping the none-rate unchanged. This suggests that lexical relevance is not consistently beneficial under the current routing policy. Although the salience route can retrieve paragraphs that overlap with the question and answer options, such overlap does not always correspond to decisive evidence. In detective narratives, paragraphs that mention important entities or option terms can still function as distractors. Therefore, the question-option salience route may occasionally introduce noisy evidence, and its benefit appears to depend on more precise salience estimation.

Removing the option-bridge route has almost no effect on accuracy, which remains 0.565. The none-rate increases only slightly from 0.008 to 0.009. This indicates that the option-bridge route contributes limited additional information under the current setting. Although the route is designed to preserve option-level coverage by connecting candidate answers to supporting or refuting events, its selected paragraphs often do not add useful evidence beyond the answer-neighborhood route. This suggests that bridge construction may require stronger evidence matching or more refined cross-span linking to become consistently useful.

The commitment ablations demonstrate the importance of conservative online route commitment. When the conservative gate is removed and FARO simply selects the route with the highest probe-readiness score, accuracy drops sharply to 0.351, and the none-rate increases to 0.019. This shows that raw probe scores are unstable when used as a direct route-selection criterion. Because early route probing observes only partial evidence, a high probe score does not necessarily imply that the full route will support the correct answer. The conservative gate therefore plays a stabilizing role by preserving the recency anchor unless a challenger route is sufficiently strong and sufficiently better than the anchor.

Replacing the explicit commitment rule with LLM-based route commitment also degrades performance, achieving 0.422 accuracy and a 0.016 none-rate. This result suggests that asking the LLM to freely choose a route is less reliable than using an explicit commitment rule. Unconstrained route selection may overreact to locally salient but incomplete evidence, whereas the conservative gate restricts switching to cases where the challenger route shows clear future utility.

Overall, the ablation results support the main design of FARO. The framework benefits most from the combination of an answer-neighborhood recency anchor and conservative future-aware commitment. The auxiliary salience and bridge routes provide alternative evidence hypotheses, but they must be controlled by a stable route-selection mechanism. Naive probe-score maximization or unconstrained LLM route selection substantially weakens performance, confirming that FARO's effectiveness comes not only from constructing multiple routes, but also from committing to them conservatively before final single-route reading.

Limitation

FARO has several limitations. First, its route design is currently developed for long narrative multiple-choice question answering. In particular, the answer-neighborhood route relies on the final part of the provided context as a recency an-

chor for the current narrative state. This anchor may be less effective when the decisive evidence is located much earlier, when the narrative state cannot be summarized by recent context, or when the answer requires evidence distributed across the entire document. Second, FARO uses three hand-designed route constructors and empirically tuned commitment thresholds. Although this design keeps the inference procedure simple and lightweight, the same route definitions and thresholds may not transfer directly to other datasets, domains, or question types. Third, FARO commits to a single route before final answer generation. This prevents post-hoc answer aggregation, but it can fail when the correct answer requires complementary evidence that is separated across multiple routes. Future work can address these limitations by learning adaptive route constructors, calibrating route-commitment thresholds across datasets, and allowing controlled multi-route reading when evidence is genuinely distributed.

Conclusion

We proposed FARO, an inference-time framework for long narrative question answering. FARO reframes long-context reasoning from late answer aggregation to early evidence-route commitment. Instead of generating multiple route answers and resolving them through voting or consensus, FARO forecasts the future utility of candidate evidence routes, commits to a single route, and performs structured reading only on the selected evidence. Experiments on DetectiveQA show that FARO improves accuracy while reducing none-rate compared with existing baselines. Ablation results further show that the recency anchor and conservative commitment rule are important for stable performance. These findings suggest that resolving evidence-path uncertainty before answer generation is a promising direction for efficient long-context narrative reasoning.

References

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hao, S.; Gu, Y.; Ma, H.; Hong, J. J.; Wang, Z.; Wang, D. Z.; and Hu, Z. 2023. Reasoning with Language Model is Planning with World Model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Jiang, H.; Wu, Q.; Luo, X.; Li, D.; Lin, C.-Y.; Yang, Y.; and Qiu, L. 2023. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. *arXiv preprint arXiv:2310.06839*.

Lee, J. H.; Yang, J. Y.; Heo, B.; Han, D.; and Yoo, K. M. 2024. Token-Supervised Value Models for Enhancing Mathematical Reasoning Capabilities of Large Language Models. *arXiv preprint arXiv:2407.12863*.

Li, X. 2025. Chain-in-Tree: Back to Sequential Reasoning in LLM Tree Search. *arXiv preprint arXiv:2509.25835*.

Li, Y.; Dong, B.; Lin, C.; and Guerin, F. 2023. Compressing Context to Enhance Inference Efficiency of Large Language Models. *arXiv preprint arXiv:2310.06201*.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Meta AI. 2024. Introducing Llama 3.1: Our Most Capable Models to Date. https://ai.meta.com/blog/meta-llama-3-1/. Accessed: 2026-06-10.

Meta Llama. 2024. meta-llama/Llama-3.1-8B-Instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Hugging Face model card. Accessed: 2026-06-10.

Nelson, E.; Kollias, G.; Das, P.; Chaudhury, S.; and Dan, S. 2024. Needle in the haystack for memory based large language models. *arXiv preprint arXiv:2407.01437*.

Pan, Z.; Wu, Q.; Jiang, H.; Xia, M.; Luo, X.; Zhang, J.; Lin, Q.; Rühle, V.; Yang, Y.; Lin, C.-Y.; Zhao, H. V.; Qiu, L.; and Zhang, D. 2024. LLMLingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression. *arXiv preprint arXiv:2403.12968*.

Wang, C.; Ning, R.; Pan, B.; Wu, T.; Guo, Q.; Deng, C.; Bao, G.; Hu, X.; Zhang, Z.; Wang, Q.; et al. 2025. Novelqa: Benchmarking question answering on documents exceeding 200k tokens. In *International Conference on Learning Representations*, volume 2025, 23444–23466.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Xiao, S.; Lin, Z.; Gao, W.; Chen, H.; and Zhang, Y. 2025. Long Context Scaling: Divide and Conquer via Multi-Agent Question-driven Collaboration. *arXiv preprint arXiv:2505.20625*.

Xu, F.; Shi, W.; and Choi, E. 2023. RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation. *arXiv preprint arXiv:2310.04408*.

Xu, Z.; Ye, J.; Liu, X.; Liu, X.; Sun, T.; Liu, Z.; Guo, Q.; Li, L.; Liu, Q.; Huang, X.; et al. 2024. Detectiveqa: Evaluating long-context reasoning on detective novels. *arXiv preprint arXiv:2409.02465*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T. L.; Cao, Y.; and Narasimhan, K. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*.

Yu, S.; Xu, X.; Deng, K.; Li, L.; and Tian, L. 2025. Tree of Agents: Improving Long-Context Capabilities of Large Language Models through Multi-Perspective Reasoning. *arXiv preprint arXiv:2509.06436*.

Zhang, J. 2025. Entropy-based Exploration Conduction for Multi-step Reasoning. *arXiv preprint arXiv:2503.15848*.

Zhang, Y.; Sun, R.; Chen, Y.; Pfister, T.; Zhang, R.; and Arik, S. O. 2024. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks. In *Advances in Neural Information Processing Systems*.

Zhou, A.; Yan, K.; Shlapentokh-Rothman, M.; Wang, H.; and Wang, Y.-X. 2024. Language Agent Tree Search Unifies Reasoning, Acting, and Planning in Language Models. In *Proceedings of the 41st International Conference on Machine Learning*.