Tdnguyen at CQs-Gen 2025: Adapt Large Language Models with Multi-Step Reasoning for Critical Questions Generation

Tien-Dat Nguyen[1, 3], Duc-Vu Nguyen[2, 3]
[1]Faculty of Computer Science, University of Information Technology,
Ho Chi Minh City, Vietnam
[2]Laboratory of Multimedia Communications, University of Information Technology,
Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
`21520707@gm.uit.edu.vn` `vund@uit.edu.vn`

Abstract

This paper explores the generation of Critical Questions (CQs) from argumentative texts using multi-step reasoning techniques, specifically Chain-of-Thoughts (CoT) and Tree-of-Thoughts (ToT) prompting frameworks. CQs are essential for enhancing critical thinking and improving decision-making across various domains. Despite the promise of Large Language Models (LLMs) in this task, generating contextually relevant and logically sound questions remains a challenge. Our experiments show that CoT-based prompting strategies, including Zero-shot and One-shot methods, significantly outperform baseline models in generating high-quality CQs. While ToT prompting offers a more flexible reasoning structure, it was less effective than CoT in this task. We suggest exploring more advanced or computationally intense multi-step reasoning techniques, as well as alternative tree structures for the ToT framework, to further improve CQs-Gen systems.

1 Introduction

Generating Critical Questions (CQs) from argumentative texts plays a vital role in enhancing human understanding, promoting critical thinking, specifically in education (Santoso et al., 2018), and improving decision-making processes across domains such as education, law, and policy analysis. These questions aim to identify potential weaknesses, blind spots, or implicit assumptions within arguments, thereby encouraging deeper reflection and discourse. The task draws inspiration from argumentation theory, particularly Walton's argumentation schemes (Walton et al., 2008), where each type of argument is associated with a set of critical questions designed to test its validity (Calvo Figueras and Agerri, 2024).

Despite its importance, Critical Questions Generation (CQs-Gen) remains an underexplored and challenging task. It requires not only an understanding of the structure and content of arguments but also the ability to reason about unstated premises and hypothetical counterpoints. Previous approaches, including prompt-based generation using Large Language Models (LLMs), have shown promise but still struggle with generating questions that are both contextually relevant and logically sound (Calvo Figueras and Agerri, 2024).

In this paper, we explore the use of multi-step reasoning techniques, particularly Chain-of-Thoughts (CoT) and Tree-of-Thoughts (ToT) framework, to improve the performance of LLMs in CQs-Gen. These approaches aim to mimic the step-by-step reasoning process humans use when evaluating arguments, allowing the model to better capture underlying assumptions and generate more insightful critical questions. Through systematic experiments and evaluation, we demonstrate the effectiveness of multi-step prompting in generating high-quality, argument-sensitive critical questions.

2 Related Work

To contextualize our contribution, we examine three foundational pillars: the development of Large Language Models (LLMs), the task of CQs-Gen, and recent advancements in multi-step reasoning techniques.

2.1 Large Language Models

The advent of Large Language Models (LLMs) such as GPT-4o (OpenAI et al., 2024), Deepseek-R1 (DeepSeek-AI et al., 2025), and LLaMA 3 (Grattafiori et al., 2024) has brought significant progress in natural language processing. Trained on vast corpora of diverse text, these models exhibit strong zero-shot and few-shot performance across a wide array of tasks, including question answering, summarization, and reasoning. Notably, models like GPT-4o and Deepseek-R1 demonstrate implicit reasoning capabilities, often reasoning

Proceedings of the 12th Argument Mining Workshop, pages 265–280
July 31, 2025 ©2025 Association for Computational Linguistics

through complex prompts before producing outputs. These characteristics make them strong candidates for tasks that require nuanced comprehension and logical progression, such as multi-step inference and question generation.

2.2 Critical Questions Generation (CQs-Gen)

The CQs-Gen task, introduced by Calvo Figueras and Agerri (2024), involves generating questions that critically engage with the reasoning presented in a given text. CQs-Gen holds potential for applications in combating misinformation and promoting critical thinking, as evidenced by related studies on argument analysis and fallacy detection (Musi et al., 2023; Musi and Reed, 2022). However, the task remains underexplored. Existing LLMs often struggle with generating relevant and coherent critical questions. Common issues include producing questions that are (1) unrelated to the argument, (2) based on introduced but unsupported concepts, (3) overly generic, or (4) grounded in flawed reasoning. Furthermore, the absence of large-scale, high-quality datasets of reference critical questions poses an additional challenge, limiting the development and evaluation of robust CQs-Gen systems (Calvo Figueras and Agerri, 2024).

2.3 Multi-Step Reasoning Techniques

Recent advances in multi-step reasoning have introduced techniques such as CoT prompting (Wei et al., 2023), ToT (Yao et al., 2023; Long, 2023)[1], and Chain-of-Draft (Xu et al., 2025), all of which have demonstrated notable improvements in tasks like arithmetic reasoning, commonsense inference, and symbolic reasoning (e.g., the coin-flip task). More recent work has explored integrating these techniques with decision-making frameworks such as A* search or reinforcement learning to further enhance reasoning performance (Wang et al., 2024a,b).

In this study, we focus on two foundational methods – Chain-of-Thoughts and Tree-of-Thoughts – due to their simplicity, accessibility, and effectiveness, while leaving the exploration of more advanced or computationally intensive approaches to future work.

3 Methodology

In this section, we present several multi-step reasoning techniques for the CQs-Gen task. An overview of the different prompting strategies is illustrated in Figure 1. All prompts and sample responses can be found in appendix A.

3.1 Chain-of-Thoughts Prompting

To support the generation of critical questions that challenge arguments effectively, we adopt CoT prompting – a technique that decomposes reasoning into intermediate steps (Wei et al., 2023). This structure helps LLMs expose implicit assumptions and reasoning gaps. We evaluate several CoT prompting variants:

- **Zero-Shot CoT Prompting:** Following Kojima et al. (2023), this method appends a trigger phrase to the input to elicit structured reasoning without examples. We use the custom phrasing:

  > “Think about the argument step-by-step. Print your reasoning first in a structured step-by-step format.”

  to discourage models from skipping the reasoning phase.

- **One-Shot CoT Prompting:** As proposed in Wei et al. (2023), this variant includes an exemplar that illustrates structured reasoning and CQ derivation. For each intervention, we select a similar – but not identical – example from the validation set to promote unbiased performance (explained in Appendix A.3.2).

- **Least-to-Most CoT Prompting:** Introduced by Zhou et al. (2023), this approach breaks down the task into explicit subtasks, guiding the model step-by-step rather than relying on implicit decomposition.

3.2 Tree-of-Thoughts Framework

To complement linear CoT strategies, we explore Tree-of-Thoughts (ToT) prompting (Yao et al., 2023), which organizes reasoning as a search tree. Inspired by human problem-solving via branching deliberations (Newell, 1959), ToT enables multiple reasoning paths using heuristics such as BFS, DFS, or A*. Our implementation builds a shallow tree via BFS with three stages: (1) generate and rank initial plans; (2) expand top plans into sub-plans; (3) generate critical questions from sub-plans. At each

---

[1] For a prompt-based implementation of Tree-of-Thoughts, see `https://github.com/dave1010/tree-of-thought-prompting`

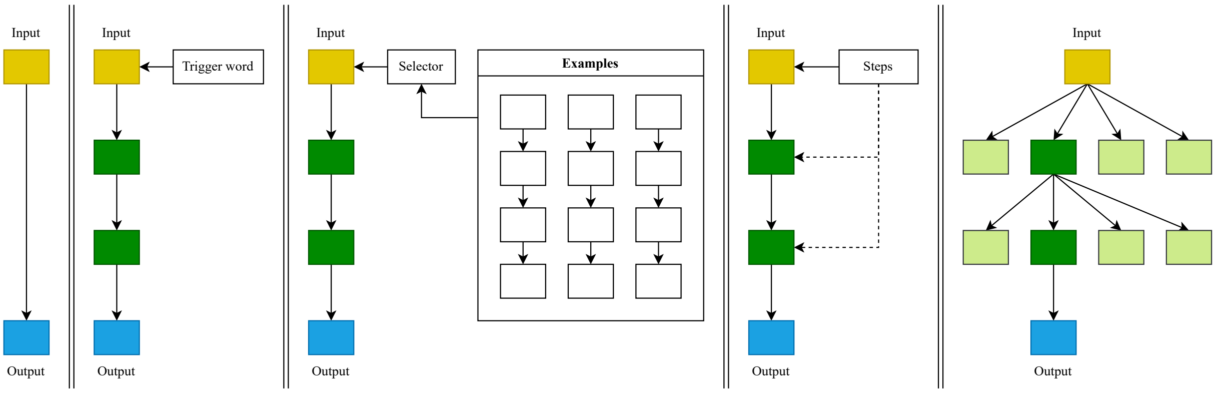

Figure 1: Overview of prompting strategies explored in this work. From left to right: (1) Standard input-output prompting (baseline), (2) Zero-Shot CoT prompting with a trigger phrase, (3) Few-Shot CoT prompting with selected exemplars, (4) Least-to-Most CoT prompting with explicit step guidance, and (5) Tree-of-Thoughts (ToT) prompting using breadth-first exploration of reasoning paths. **Yellow boxes** denote the input, **green boxes** represent intermediate reasoning steps or sub-tasks, **light green boxes** indicate reasoning paths that were pruned in the ToT framework, and **blue boxes** represent the final output (i.e., the generated critical question).

depth, a voting-based mechanism selects the most promising candidates. For simplicity and due to time constraints, we evaluate a single tree structure, leaving broader configurations for future work.

4 Experiments and Results

4.1 Experiments

In this section, we describe the experimental setup used to evaluate the multi-step reasoning strategies introduced in Section 3. We outline the dataset, model configurations, prompting procedures, evaluation protocol, and implementation details.

4.1.1 Dataset

We conduct all experiments using the validation set of the CQs-Gen Shared Task dataset (Figueras et al., 2025; Figueras and Agerri, 2025), which consists of argument-based scenarios paired with corresponding interventions. Each input scenario serves as the context for generating critical questions that challenge the underlying reasoning or assumptions of the argument. For One-shot CoT, for each intervention, we randomly selected three useful representative questions and then used DeepSeek-R1 to generate the reasoning steps that lead to those questions (prompt in Appendix A.3.1).

4.1.2 Model

We evaluate all prompting strategies using GPT-4o mini via the OpenAI API, keeping default parameters (temperature = 1.0) to encourage generation diversity. As baselines, we adopt LLaMA-3-8B and GPT-4o mini, both under standard prompting. While prior work used LLaMA-2-13B and Zephyr-13B (Calvo Figueras and Agerri, 2024), LLaMA-3-8B offers architectural improvements, including enhanced tokenization, instruction following, and reasoning capabilities (Grattafiori et al., 2024). Using GPT-4o mini additionally allows us to isolate the effect of prompting strategies from model capacity, and to assess whether structured reasoning benefits even strong proprietary models. In addition, we conducted tests using DeepSeek-R1, Mistral, and Qwen 3 (locally via Ollama) to explore generalization across different LLMs.

4.1.3 Tree-of-Thoughts Configuration

For the ToT strategy, we construct trees with a maximum depth of 4, where the root node (depth 1) represents the input scenario; we generate 5 candidate initial plans at depth 2; for each selected plan, we generate 3 sub-plans at depth 3, and each sub-plan is used to generate 3 candidate critical questions at the depth 4. At each depth level, we conduct 10 independent voting rounds using the model to evaluate and rank candidates. Based on these votes, we select the top 3 candidates to expand to the next level. The final output is selected from the pool of questions generated at depth 4.

4.1.4 Evaluation

In this study, we perform automatic evaluation due to time constraints. Following the CQs-Gen Shared Task setup (Calvo Figueras and Agerri, 2024), we use automatic evaluation based on similarity between generated and reference questions. For each generated question, we identify the reference question with the highest similarity. If this score is below 0.6, the output is marked as "not evaluable" and assigned a score of 0. Otherwise, we assign the reference label ("Useful," "Unhelpful," or "Invalid") to the generated question. Each intervention yields three questions. The intervention score is the sum of question scores, with "Useful" is 0.33, "Unhelpful" is 0.1, and others is 0. The system's overall score is the mean intervention score across the dataset. Finally, we submit the test set questions generated by the two most promising CoT prompting methods and the ToT framework for official evaluation by the CQs-Gen Shared Task organizers.

4.2 Results

4.2.1 Punctuation Evaluation Results

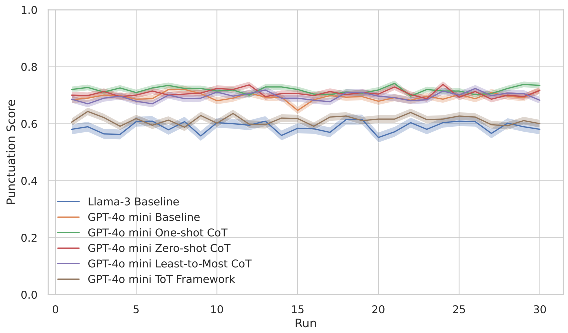

Figure 2: Punctuation score of all methods, including baseline.

Figure 2 and the table 1 presents the punctuation score across 30 runs on the validation set for all prompting strategies, including a baseline using GPT-4o mini. As expected, the baseline using LLaMA-3 8B lags significantly behind all other methods. Notably, the GPT-4o mini baseline demonstrates a strong improvement over the LLaMA-based baseline, confirming that model strength plays a substantial role in performance. However, prompting strategies such as One-shot CoT and Zero-shot CoT still outperform this strong baseline, indicating that reasoning scaffolds continue to provide benefits even when using advanced language models. Least-to-Most CoT performs comparably to GPT-4o mini baseline, while the ToT framework underperforms relative to other CoT-based methods in this setting. These results highlight that while newer models enable implicit reasoning, structured prompting continues to offer performance gains, especially in tasks requiring nuanced argumentative analysis. Results from additional model evaluations using DeepSeek-R1, Mistral and Qwen 3 are provided in Appendix B.

Interestingly, official evaluation on the hidden test set (also shown in table 1) reveals that Zero-shot CoT and the ToT Framework slightly outperform One-shot CoT, despite its superior performance on the validation set. This suggests that simpler or more diverse prompting strategies may generalize better in unseen scenarios.

Method	Validation	Test
Llama-3-8B Baseline	0.589 ± 0.019	
GPT-4o mini Baseline	0.694 ± 0.014	
GPT-4o mini One-shot CoT	0.719 ± 0.011	0.451
GPT-4o mini Zero-shot CoT	0.707 ± 0.013	0.490
GPT-4o mini Least-to-Most CoT	0.695 ± 0.014	
GPT-4o mini ToT Framework	0.613 ± 0.015	0.461

Table 1: Comparison of performance metrics of 30 runs across different prompting and reasoning strategies in validation set and test set.

4.2.2 Error Analysis

Our error analysis examined the top struggling interventions across various prompting strategies and model architectures. We identified several consistent patterns in failure cases: long and multi-topics interventions; emotionally charged or highly subjective content; too brief or ambiguous interventions; interventions involving sensitive themes such as race, health, or politics. Across methods, a common failure mode was the generation of descriptive rather than genuinely critical questions. Although the ToT approach demonstrated more structured reasoning, it continued to struggle with emotionally complex or contextually diffuse inputs. Representative examples of such interventions and model outputs are provided in Appendix C.

5 Conclusion

In this study, we investigated multi-step reasoning techniques, specifically CoT prompting and the ToT framework, for generating critical questions (CQs) from argumentative texts. Experiments show that CoT variants – especially Zero-shot CoT and One-shot CoT – outperform the LLaMA-3 baseline and slightly surpass GPT-4o mini, helping expose underlying assumptions. While the ToT framework supports structured exploration, it proved less effective in this context. Key challenges remain, such as handling emotionally charged or ambiguous content. Future work should explore more advanced multi-step strategies and alternative tree structures for ToT to enhance CQs-Gen performance.

Limitations

One limitation of our work lies in the inference cost associated with multi-step prompting using large language models (LLMs). Although GPT-4o-mini is relatively inexpensive, Table 2 shows that ToT incurred substantially higher cost across 30 runs ($21.10 total), compared to simpler strategies such as Zero-shot CoT ($0.98). While the per-sample cost remains low, the cumulative overhead may become prohibitive for larger-scale deployment. This motivates future exploration into cost-efficient prompting or model distillation techniques.

Method	Input	Output	Cost ($)
Zero-shot CoT	2.17M	1.10M	0.98
One-shot CoT	5.20M	0.90M	1.27
Least-to-Most CoT	2.39M	1.38M	1.19
ToT Framework	97.39M	10.94M	21.10

Table 2: Input/output token usage and total cost (in USD) for 30 runs of each multi-step reasoning technique using GPT-4o-mini on the validation set.

Acknowledgement

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund. We thank the anonymous reviewers for their time and helpful suggestions that improved the quality of the paper.

References

Blanca Calvo Figueras and Rodrigo Agerri. 2024. Critical questions generation: Motivation and challenges. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 105–116, Miami, FL, USA. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Banca Calvo Figueras and Rodrigo Agerri. 2025. Benchmarking critical questions generation: A challenging reasoning task for large language models. *Preprint*, arXiv:2505.11341.

Blanca Calvo Figueras, Jaione Bengoetxea, Maite Heredia, Ekaterina Sviridova, Serena Villata Elena Cabrio, and Rodrigo Agerri. 2025. Overview of the critical questions generation shared task 2025. In *Proceedings of the 12th Workshop on Argument Mining*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Jieyi Long. 2023. Large language model guided tree-of-thought. *Preprint*, arXiv:2305.08291.

Elena Musi, Elinor Carmi, Chris Reed, Simeon Yates, and Kay O'Halloran. 2023. Developing misinformation immunity: How to reason-check fallacious news in a human–computer interaction environment. *Social Media + Society*, 9(1):20563051221150407.

Elena Musi and Chris Reed. 2022. From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*, 33(3):349–370.

A. Newell. 1959. *Report on a General Problem-solving Program*. P (Rand Corporation). Rand Corporation.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

T Santoso, L Yuanita, and E Erman. 2018. The role of student's critical asking question in developing student's critical thinking skills. *Journal of Physics: Conference Series*, 953(1):012042.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Chaojie Wang, Yanchen Deng, Zhiyi Lyu, Liang Zeng, Jujie He, Shuicheng Yan, and Bo An. 2024a. Q*: Improving multi-step reasoning for llms with deliberative planning. *Preprint*, arXiv:2406.14283.

Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. 2024b. Offline reinforcement learning for llm multi-step reasoning. *Preprint*, arXiv:2412.16145.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Silei Xu, Wenhao Xie, Lingxiao Zhao, and Pengcheng He. 2025. Chain of draft: Thinking faster by writing less. *Preprint*, arXiv:2502.18600.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

A Prompts and Sample Responses

All sample responses presented in this section are generated using the following prompt with the specified intervention:

> **MT:**
> *"Claire's absolutely right about that. But then the problem is that that form of capitalism wasn't generating sufficient surpluses. And so therefore where did the money flow. It didn't flow into those industrial activities, because in the developed world that wasn't making enough money."*

A.1 Standard input-output prompting (baseline)

A.1.1 Prompt:

> Suggest exactly 3 critical questions that should be raised before accepting the arguments in this text:
> "{intervention}"
> Make the questions simple, and do not give any explanation regarding why the question is relevant.

A.1.2 Sample response:

1. What evidence supports the claim that that form of capitalism wasn't generating sufficient surpluses?
2. What specific industrial activities is the speaker referring to, and why were they not profitable?
3. Where did the money flow instead, and what were the consequences of that shift?

A.2 Zero-shot CoT

A.2.1 Prompt:

Think about the argument step-by-step. Print your reasoning first in a structured step-by-step format.

Then, from the reasoning, suggest exactly 3 critical questions that should be raised before accepting the arguments in this text:

"{intervention}"

Make the questions simple.

A.2.2 Sample response:

Step-by-Step Reasoning

1. **Identifying the Claim**: The speaker, MT, agrees with Claire regarding a point about capitalism. They assert that a particular form of capitalism ("that form of capitalism") is not generating sufficient surpluses.
2. **Analyzing the Problem**: MT highlights a problem: the insufficient generation of surpluses under the discussed form of capitalism. This implies a need to understand what "that form" refers to and why it fails in this aspect.
3. **Exploring the Consequence**: MT states that because of the lack of surpluses, there is a flow of money that does not enter industrial activities. There is an implication that perhaps money is flowing elsewhere but does not specify where exactly it is going.
4. **Contextualizing Development**: The mention of the "developed world" suggests that the argument is rooted in a specific economic context. It implies that in more economically advanced regions, industrial activities are not lucrative enough to attract investment.
5. **Inquiring About Implications**: MT's point suggests broader implications for capitalism and industrial economies. The argument may require further substantiation to understand the full picture of economic flow and investment in relation to industrial activities.

Critical Questions

1. What specific form of capitalism is being referenced, and why is it deemed ineffective in generating surpluses?
2. Where exactly is the money flowing if it is not going into industrial activities?
3. What evidence is there to support the claim that industrial activities in the developed world are not profitable enough to attract investment?

A.3 One-shot CoT

A.3.1 Prompt for generating contents of One-shot CoT:

Given the following intervention:
{intervention}
And the following reasoning guideline:
- Identify its core claim.
- Identify any assumptions it makes.
- Examine the logical connections between claims.
- Consider what evidence would be needed to support it.
- Identify any missing or unclear aspects.

Use this guideline to generate reasoning steps that lead to these questions:
{questions}

A.3.2 Prompt for generating questions:

* Task: Given an intervention, identify its core claim, assumptions, logical connections, required evidence, and missing aspects. Then, suggest 3 critical questions that should be raised before accepting the argument. Make the questions simple.

* Example:
- Intervention: "{example_intervention}"
- Core claim: {core_claim}
- Assumption: {assumptions}
- Logical connections: {logical_connections}
- Required evidence: {required_evidence}
- Missing or unclear aspects: {missing_or_unclear_aspects}
- From these reasoning, here are the questions: {questions}

* Intervention: "{intervention}"

In this prompt, we use a semantic similarity-based selection approach to retrieve an intervention that is similar – but not identical – to the input, as example intervention. For OpenAI GPT-4o mini, we compute cosine similarity directly between the input and all stored example embeddings. For locally run models (e.g., DeepSeek-R1, Mistral, Qwen3), due to time constraint, we use a FAISS-based vector store to perform similarity search. In both cases, we normalize and filter out exact matches and ensure the selected example is semantically relevant.

A.3.3 Sample response of generating questions:

Core Claim:
The current form of capitalism is not generating enough surplus to support industrial activities in the developed world, leading to a lack of investment in these sectors.

Assumption:
The speaker assumes that a correlation exists between surplus generation and investment in industrial activities, and that the current form of capitalism is inherently insufficient for fostering necessary economic growth within the developed world.

Logical Connections:
The speaker connects the inability to generate sufficient surpluses under the current form of capitalism directly to the lack of funding and investment in industrial sectors. They suggest a causal relationship where inadequate profitability leads to inadequate investment.

Required Evidence:
Evidence would include data comparing surplus generation under different forms of capitalism, examples of industrial investment trends in response to surplus levels, comparisons between developed and developing economies, and studies demonstrating the relationship between profit margins and investment in industrial activities.

Missing or Unclear Aspects:

The speaker does not clarify what specific factors contribute to the lack of surplus generation in current capitalist structures, how these factors impact different sectors, or what alternative economic models might better support surplus generation. The implications of this lack of surplus on long-term economic health are also not discussed.

From these reasoning, here are the questions:

1. What specific aspects of the current form of capitalism are preventing sufficient surplus generation?
2. How does the speaker define “sufficient surpluses,” and is there a consensus on what this should be?
3. Are there alternative economic models or strategies that could address this problem, and have they been considered?

A.4 Least-to-Most CoT

A.4.1 Prompt:

First, analyze the argument by breaking it down logically:
- Identify its core claim.
- Identify any assumptions it makes.
- Examine the logical connections between claims.
- Consider what evidence would be needed to support it.
- Identify any missing or unclear aspects.

Print your reasoning first in a structured step-by-step format. Then, suggest exactly 3 critical questions that should be raised before accepting the arguments in this text:

“{intervention}”

Make the questions simple.

A.4.2 Sample response:

Logical Breakdown of the Argument

1. **Core Claim:**
   - The speaker asserts that a specific form of capitalism is ineffective because it fails to generate sufficient surpluses, leading to a diversion of money away from industrial activities.
2. **Assumptions:**
   - There exists a specific form of capitalism being referenced.
   - The speaker believes that the failure to generate surpluses directly correlates with inadequate investment in industrial activities.
   - The claim assumes that sufficient surpluses are necessary for money to flow into industrial sectors.
3. **Logical Connections:**
   - The argument suggests a causal relationship between the insufficiency of surpluses and the lack of investment in industrial activities.
   - It implies that if capitalism were generating sufficient surpluses, then money would flow into the industrial sector.
4. **Evidence Needed:**
   - Data supporting the assertion that the current form of capitalism is inadequately generating surpluses.
   - Examples or statistics that demonstrate how lower surpluses correlate with reduced investment in industrial activities.
   - Analysis of the current financial flow patterns to industrial activities to substantiate claims of diversion.
5. **Missing or Unclear Aspects:**
   - The specific characteristics of the form of capitalism being discussed are not defined.
   - The speaker does not clarify what “sufficient surpluses” means or how it can be measured.
   - There’s no mention of potential alternative explanations for the lack of investment in industrial activities, which could dilute the strength of the claim.

Critical Questions to Raise

1. What specific form of capitalism is being referred to, and what differentiates it from other forms?
2. What criteria do you use to define "sufficient surpluses," and how do we measure them?
3. Are there other factors besides surplus generation that could explain why money is not flowing into industrial activities?

A.5 ToT Prompting

A.5.1 Generating plans

Here is the intervention:

"{intervention}"

Generate a one line plan on how you would suggest a critical question that should be raised before accepting the arguments in this text. Do not give the question yet.

A.5.2 Generating sub-plans

Here is the intervention:

"{intervention}"

To suggest a critical question, here is the plan:

{parent_plan}

Generate a one line sub-plan of this plan on how you would suggest a critical question that should be raised before accepting the arguments in this text. Do not give the question yet.

A.5.3 Generating questions

Here is the intervention:

"{intervention}"

To suggest a critical question, here is the plan:

{plan}

From the plan, suggest three critical questions that should be raised before accepting the arguments in that intervention. Make the questions simple, and do not give any explanation regarding why the question is relevant.

A.5.4 Voting for plans

Here is the intervention:

"{intervention}"

To suggest a critical question, here are the plans:

{plans}

Analyze these plans, then conclude which is the most promising for a critical question for the intervention.

A.5.5 Voting for questions

Here is the intervention:

"{intervention}"

Here are the critical questions:

{cqs}

Analyze these questions and choose the three most suitable critical questions that should be raised before accepting the arguments in the intervention.

B Additional Results with Deepseek-R1, Mistral-7B, and Qwen 3-7B

To assess the generalizability of our prompting strategies across different LLM families, we conducted supplementary experiments using DeepSeek-R1, Mistral-7B, and Qwen3-8B, all run locally via the Ollama framework. Due to extensive computational runtime and time constraints during local evaluation, we did not include the Tree-of-Thought (ToT) framework in these supplementary tests. The results are reported below.

Method	**DeepSeek-R1**	**Mistral-7B**	**Qwen3-8B**
Baseline	0.538 ± 0.022	0.631 ± 0.017	0.629 ± 0.018
One-shot CoT	0.472 ± 0.020	0.670 ± 0.018	0.650 ± 0.017
Zero-shot CoT	0.526 ± 0.020	0.677 ± 0.013	0.645 ± 0.021
Least-to-Most CoT	0.434 ± 0.021	0.532 ± 0.019	0.671 ± 0.017

Table 3: Validation results with DeepSeek-R1, Mistral-7B, and Qwen3-8B using local inference (with `enable_thinking=true` in `apply_chat_template` for Qwen3-8B only).

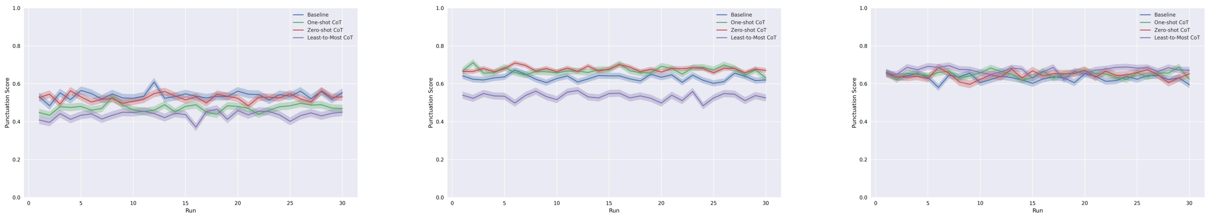

Figure 3: Performance comparison of prompting strategies. From left to right: DeepSeek-R1, Mistral-7B, and Qwen3-8B (with `enable_thinking=true` in `apply_chat_template` for Qwen3-8B only).

C Examples of Model Output on Challenging Interventions

C.1 Long, Multi-Topic Interventions

CLINTON
"The central question in this election is really what kind of country we want to be and what kind of future we'll build together. Today is my granddaughter's second birthday... We need new jobs, good jobs, with rising incomes... I also want to see more companies do profit-sharing."

- **GPT-4o mini Zero-shot CoT:** What specific policies or initiatives will be implemented to ensure new jobs in infrastructure and advanced manufacturing?
- **GPT-4o mini One-shot CoT:** What specific strategies will you implement to ensure new jobs created are of high quality and provide rising incomes for workers?
- **GPT-4o mini Least-to-Most CoT:** What evidence or data do you have to support the claim that increasing investments in certain sectors will lead to job growth and fairer economic conditions?
- **GPT-4o mini ToT Framework:** What impact will raising the national minimum wage have on small business profitability?
- **GPT-4o mini baseline:** What specific policies or measures will be implemented to build a fairer economy?

Despite their structured reasoning design, both CoT and ToT fail to produce useful critical questions in this example, as they focus on the most salient economic claims and neglect the discourse-level structure, resulting in unhelpful or invalid questions that do not fully challenge the argumentative richness of the intervention.

C.2 Emotionally Charged / Subjective Language

AllergyDad
"As a parent of a child with severe life-threatening allergies to peanuts and treenuts... a peanut can have the same result [as death]. I urge the DOT to issue the ban, and make traveling safe for everyone."

- **GPT-4o mini Zero-shot CoT:** What practical implications would banning peanuts and tree nuts on flights have on airlines and food service operations?
- **GPT-4o mini One-shot CoT:** How would the proposed ban on nut products be effectively enforced on flights?
- **GPT-4o mini ToT Framework:** How will a ban on peanut and tree nut products affect passengers with other dietary restrictions and allergies?
- **LLaMA-3 baseline:** What are the current airline policies regarding food allergies on flights?

CoT and ToT fail in this case because they generate operational or policy-related questions that accept the argument's premise, rather than critically examining whether an emotionally driven personal case justifies a general policy ban, rendering their outputs unhelpful or invalid.

C.3 Overly Short / Ambiguous Content

CLINTON
"It's a good one, though. It well describes the problem."

- **GPT-4o mini ToT Framework:** What specific problem is Clinton addressing in his statement?
- **GPT-4o mini baseline:** What specific problem is being referred to in the statement?

ToT fails in this case because the original statement lacks clear argumentative content, leading the model to produce a vague question that could be applied to any text – thus rendering it invalid.

C.4 Sensitive / Politically Charged Topics

CLINTON
"It's just a fact that if you're a young African-American man and you do the same thing as a young white man, you are more likely to be arrested... I want to see private prisons ended in the state system... we need commonsense gun safety measures."

- **LLaMA-3 baseline:** How would Clinton's plan to divert people from the criminal justice system work in practice, and what metrics would be used to measure its effectiveness?

C.5 Satirical / Figurative Language Misinterpreted Literally

Antanagoge

"Is it really possible that all blindly partisan advocates of the peanut... are overtaken by such a compulsion... These sound like symptoms of addiction... The only prudent course of action is to discontinue peanuts on airplanes."

howie

"Most food allergies are completely imagined... Peanuts are incredibly nutritious... Leave my peanuts alone!"

- **GPT-4o mini ToT Framework:** What are the documented cases of severe allergic reactions to peanuts occurring on airplanes?

ToT fails in this case because its multi-path reasoning converges prematurely on a literal interpretation of the text, overlooking its satirical tone.