Confidence Leaps in LLMs: Enabling Early Stopping and Cross-Model Reasoning Transfer

Anonymous ACL submission

Abstract

We challenge the assumption that large language models (LLMs) build confidence gradually while reasoning. Instead, we find that confidence often appears in a discrete "moment of insight" — a sharp increase in the probability of the correct answer, which we call a confidence leap. Leveraging this finding, we introduce a training-free, model-agnostic early-stopping heuristic. When a confidence leap is present, we can halt the generation, reducing reasoning tokens by up to 31% without accuracy loss. We further show that the reasoning preceding a leap is semantically transferable: providing this partial reasoning to a different model family substantially improves its accuracy. These results suggest that confidence leaps mark shared, interpretable reasoning milestones rather than model-specific artifacts.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in multi-step reasoning, largely driven by techniques like Chain-of-Thought (CoT) prompting (Wei et al., 2022). By generating intermediate reasoning steps, LLMs can decompose complex problems and arrive at more accurate solutions. However, the internal dynamics of how confidence evolves during this reasoning process remain poorly understood.

A common assumption is that model confidence builds gradually as the reasoning chain progresses. In this work, we challenge this view and argue that conviction is often not reached through gradual accumulation, but through a discrete "confidence leap": a sudden and sharp increase in the probability of an answer. This "moment of insight" signals a critical juncture where the model has effectively processed the core of the problem. Understanding this phenomenon has direct implications for optimizing inference costs and interpreting reasoning mechanisms.

Based on this discovery, we propose a simple yet powerful early-stopping heuristic. Recent efforts in this direction have explored training auxiliary classifiers on internal model activations (Fu et al., 2023; Snell et al., 2023; Chen et al., 2024; Zhang et al., 2025; Afzal et al., 2025), requiring architectural access and model-specific training, as well as answer consistency heuristics that monitor output stability across consecutive steps (Liu and Wang, 2025; Mao et al., 2025). Our training-free, model-agnostic method monitors the output probability distribution and terminates generation upon detecting a confidence leap, significantly reducing generation length while maintaining or even improving accuracy.

To further probe the nature of these leaps, we connect our work to the interpretability and transfer of reasoning traces (Lanham et al., 2023; Lampinen et al., 2024) and investigate whether they represent genuine semantic insights or are merely model-specific artifacts. We demonstrate that the reasoning text leading up to a confidence leap is transferable: feeding this partial reasoning to a model from a different family dramatically increases its likelihood of arriving at the correct answer. This suggests that the "moment of insight" is not a statistical quirk but a portable, interpretable reasoning milestone.

Our contributions are three-fold:

- We identify and characterize "confidence leaps", sudden leaps in answer probability that mark key moments in CoT reasoning.
- We propose an effective early-stopping heuristic based on this phenomenon that reduces computation and can improve accuracy.
- We provide evidence that the reasoning chunks causing these leaps contain transferable semantic insights, not just model-specific artifacts.

2 Data and Models

We use *GPQA* (Rein et al., 2024), a graduate-level multiple-choice benchmark, and *LogiQA* (Liu et al., 2020), a logical reasoning dataset. We collect reasoning traces from Qwen3-32B (Yang et al., 2025), QwQ-32B (QwenTeam, 2025) and DeepSeek-R1 Distill Llama-70B (Guo et al., 2025).

We segment reasoning traces into chunks by splitting on trigger words (e.g., *"Wait"*, *"However"*, *"Hmm"*) that signal reasoning shifts. A new chunk begins when a paragraph starts with a trigger word, yielding a sequence $C_1, C_2, \ldots, C_N$ for each trace.

For each chunk C_i, we truncate the reasoning chain, append `</think>` followed by `\boxed{`, and extract the predicted answer and probability distribution over choices (A, B, C, D). This yields the sequence $P_1, P_2, \ldots, P_N$.

3 Analysis of Reasoning Dynamics

We analyze how often models revise their predictions. A "change" is a modification of the predicted answer from one chunk to the next. Figure 2 shows accuracy as a function of length, broken down by number of changes.

Notably, across both datasets, a substantial fraction of examples involve at least one answer change. The most frequent type of revision is a single change.

3.1 The Critical Leap in Confidence

Given the prevalence of single-change traces, we now ask: *what happens within the reasoning process at the moment of this change*? To investigate this, we focus on the subset of traces with exactly one change. We define the "critical chunk" (k) as the point of change and analyze the model's

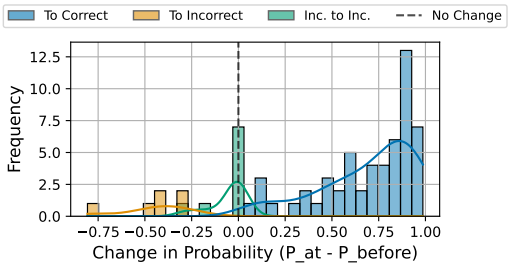

Figure 1: Distribution of the probability change for the correct answer at the critical chunk for Qwen3-32B on GPQA. A large positive change is observed for "Switch to Correct" events.

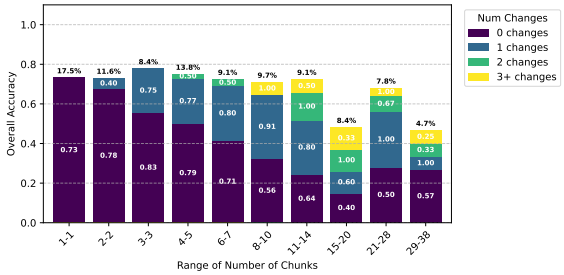

Figure 2: Accuracy vs. reasoning chain length for Qwen3-32B on GPQA, stratified by number of answer changes. Each stacked segment represents a different change category, with segment height reflecting its contribution to overall accuracy. The percentage on top of each bar shows the proportion of examples in that bin. A substantial portion of reasoning chains involve one or more changes. Similar patterns are observed for other models and datasets (see Appendix D).

probability distribution over the correct answer immediately before (at chunk $k-1$) and at the critical chunk (k).

Figure 1 shows the distribution of this probability change ($P_k(\text{correct}) - P_{k-1}(\text{correct})$) for Qwen3-32B on GPQA. The distribution is clearly bimodal. When the model switches *to* the correct answer, the probability change is sharply positive, with a mean of +0.69 for Qwen3-32B. Conversely, when the model switches *from* the correct answer, the change is negative. This sharp increase, which we term a "critical leap", demonstrates that the model's confidence in the correct answer rises dramatically and decisively upon processing the critical chunk. Similar patterns are observed for other models and on LogiQA (see Appendix A).

3.2 Location of the Critical Leap

Having established the existence of a critical leap, a natural next question is: *where in the reasoning chain does this leap tend to occur?* We analyzed the relative position of the critical chunk (k/N) across all single-change instances. Figure 3 shows the distribution of this relative position for Qwen3-32B on GPQA. We observe that the critical chunk does not appear at a fixed location; instead, it is distributed broadly across the entire length of the reasoning chain. This pattern holds across both datasets and all models (Appendix B). It implies that any potential optimization, such as early stopping, cannot rely on a fixed-length prefix of the reasoning. The crucial insight can emerge at any point.

Metric	Qwen3-32B		DeepSeek-R1		QwQ-32B	
	GPQA	LogiQA	GPQA	LogiQA	GPQA	LogiQA
Optimal Threshold (τ)	0.81	0.56	0.21	0.26	0.56	0.46
Baseline Accuracy (%)	69.7	78.4	68.9	67.3	74.8	79.0
Heuristic Accuracy (%)	69.7	79.0	77.6	68.3	74.8	79.5
Accuracy Change (%)	**0.0**	**+0.6**	**+8.7**	**+1.0**	**0.0**	**+0.5**
Coverage (%)	10.6	3.5	54.0	12.9	27.6	4.7
Token Savings (%)	31.5	19.2	24.2	26.4	32.7	28.5

Table 1: Performance of the leap heuristic at the optimal threshold across two datasets. The heuristic achieves substantial token savings (19-33%) while maintaining or even improving accuracy.

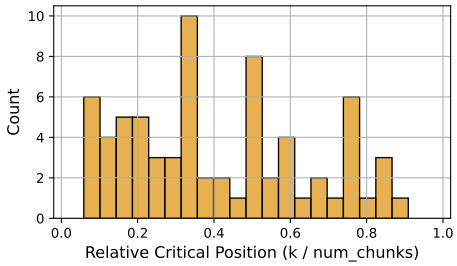

Figure 3: Distribution of the relative position of the critical chunk for Qwen3-32B on GPQA. The leap in confidence can happen at nearly any point in the reasoning process.

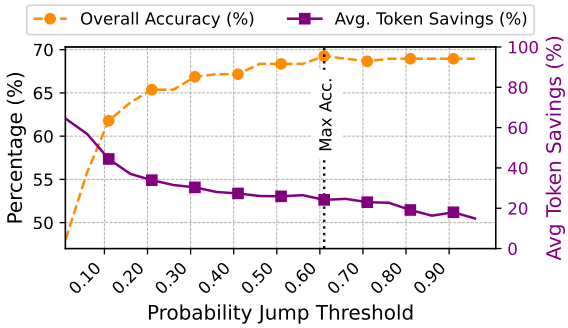

Figure 4: Performance of the leap heuristic for DeepSeek-R1 on GPQA. The plot shows overall accuracy and token savings percentage as a function of the probability leap threshold. The vertical line indicates the threshold that maximizes accuracy.

4 A Heuristic for Early Stopping

4.1 The Probability Leap Heuristic

The existence of a decisive probability leap motivates a simple, training-free heuristic for early stopping: terminate the reasoning process as soon as the probability for any answer choice increases by more than a predefined threshold τ from one chunk to the next. The predicted answer is the one that experienced the leap. If no such leap occurs, the model's final answer after the full chain is used.

4.2 Performance

We evaluated this heuristic across a range of thresholds. For each model, we selected the threshold that maximized overall accuracy. The results are summarized in Table 1.

The heuristic demonstrates substantial token savings across all models and datasets. For Qwen3-32B on GPQA, it achieves 31.5% reduction in reasoning tokens on a subset of samples where confidence leap is present, with no loss in accuracy. For DeepSeek-R1 on GPQA, the heuristic saves 24.2% of tokens while simultaneously *improving* accuracy by a substantial 8.7 percentage points. This suggests that for some models, later reasoning steps can be detrimental, and our heuristic effectively identifies the point at which the core insight has been reached. QwQ-32B demonstrates consistent savings of around 30% across both datasets. Similar patterns are observed on LogiQA, with token savings ranging from 19% to 28% alongside modest accuracy improvements. A visualization of this trade-off for DeepSeek-R1 on GPQA is shown in Figure 4. Additional visualizations for all models and datasets are provided in Appendix C.

5 Transferability of Reasoning

Having identified critical chunks that precipitate answer changes, we now ask: *do these chunks contain model-specific computational artifacts, or do they encode semantic insights that can transfer across model families?* To answer this question, we conduct a cross-model transfer experiment, providing reasoning chains from a source model (Qwen3-32B) to a target model from a different architecture (gpt-oss-20b).

5.1 Experimental Setup

We select the subset of traces with exactly one answer change from Qwen3-32B. For each trace,

Table 2: Accuracy and agreement of gpt-oss-20b when provided with reasoning chains from Qwen3-32B.

	Direct Transfer		With Continuation	
Scenario	**Accuracy (%)**	**Agreement (%)**	**Accuracy (%)**	**Agreement (%)**
At pre-pre-critical chunk ($k-2$)	33.33	35.71	42.86	40.48
At pre-critical chunk ($k-1$)	41.79	44.78	56.72	58.21
At critical chunk (k)	**70.59**	**88.24**	72.46	84.06

we extract the reasoning chunks $C_1, \ldots, C_N$ and identify the critical chunk index k. We then construct prompts for the target model by including the original question followed by varying amounts of the source model's reasoning. The target model is then forced to produce an answer by appending a special stopping sequence that terminates the reasoning process.

We evaluate six scenarios that test both direct information transfer and the target model's ability to build upon partial reasoning:

- **Direct transfer scenarios**: Provide reasoning up to chunk $k-2$, $k-1$, or k, then immediately force an answer.
- **Continuation scenarios**: Provide reasoning up to chunk $k-2$, $k-1$, or k, let the target model generate 128 tokens of its own reasoning, then force an answer.

5.2 Progressive Information Transfer

Table 2 shows the target model's accuracy across all scenarios. We first examine the direct transfer cases. When provided with reasoning only up to chunk $k-2$ (two steps before the critical change), the target model achieves 33.33% accuracy. Advancing to chunk $k-1$ improves this to 41.79%. However, including the critical chunk k yields a dramatic increase to 70.59% accuracy.

This progressive improvement demonstrates that the reasoning chain is not simply a binary "insightful or not" signal. Rather, earlier chunks lay groundwork that becomes fully actionable only when the critical chunk is present with a significant leap at chunk k.

5.3 The Role of Continuation

The continuation scenarios reveal two critical insights. First, when the target model is given partial reasoning and allowed to continue with its own generation, accuracy consistently improves: from $k-2$, accuracy increases by +9.53 (33.33% to 42.86%); from $k-1$, by +14.93 (41.79% to 56.72%); and from k, by +1.87 (70.59% to 72.46%). This consistent improvement demonstrates that the target model is *capable of continuing the reasoning trajectory initiated by a different model family*, providing strong evidence that the transferred reasoning encodes genuine semantic content rather than model-specific artifacts.

Second, and most revealing, is the asymmetry in these improvements: continuation from $k-1$ yields a substantial +14.93 gain, while continuation from k adds only +1.87. This dramatic reduction in the benefit of continuation suggests that chunk k is genuinely *critical*—it is self-contained and model-agnostic, already encoding the essential insight needed to reach the correct answer. Once this chunk is provided, there is little left for the target model to "fill in", confirming that the confidence leap marks a portable reasoning milestone rather than an incomplete intermediate state.

5.4 Cross-Model Agreement as Evidence of Semantic Content

Beyond absolute accuracy, we measure the target model's agreement with the source model's final answer. At chunk k, the target model agrees with Qwen3-32B in 88.24% of cases, compared to only 44.78% at chunk $k-1$. The target model become more accurate and converges on the *same answer* as the source model. This convergence provides strong evidence that the critical chunk encodes semantic reasoning rather than model-specific computational patterns.

6 Conclusion

We identified "critical leaps" in confidence during LLM reasoning and leveraged them for an effective early-stopping heuristic that reduces computational cost while maintaining or improving accuracy. We showed that these critical reasoning steps transfer across model families, highlighting their semantic value and opening avenues for more efficient inference.

Limitations

Our method relies on extracting probability distributions over discrete answer choices (A, B, C, D) at each reasoning step. Consequently, it is currently applicable only to multiple-choice question answering tasks where the answer space is finite and enumerable.

Additionally, our chunking methodology depends on linguistic trigger words that signal reasoning transitions. The effectiveness of this approach may vary across languages and domains where such markers are less prevalent or have different distributions.

In future work, we plan to extend our approach to open-ended generation tasks by developing alternative confidence metrics that do not rely on discrete answer distributions.

Ethical considerations

This study examines the working mechanisms of large language models and, therefore, does not introduce risks beyond those typically associated with natural language processing or computational linguistics research.

We utilize the GPQA dataset (Rein et al., 2024), which is licensed under the CC BY-NC 4.0 license and LogiQA (Liu et al., 2020), which is licensed under CC-BY-SA-4.0, both license suitable for research purposes.

Use of AI Assistants We utilize Grammarly to enhance and proofread the text of this paper, correcting grammatical, spelling, and stylistic errors, as well as rephrasing sentences. Consequently, certain sections of our publication may be identified as AI-generated, AI-edited, or a combination of human and AI contributions.

References

Anum Afzal, Florian Matthes, Gal Chechik, and Yftah Ziser. 2025. Knowing before saying: Llm representations encode information about chain-of-thought success before completion. *arXiv preprint arXiv:2505.24362*.

Yifan Chen, Zhaojiang Liu, Shuohang Li, and 1 others. 2024. Rethinking the cost of reasoning: Adaptive depth in chain-of-thought generation. *arXiv preprint arXiv:2401.02412*.

Yao Fu, Xiaochuang Tang, Hao Peng, and 1 others. 2023. Chain-of-thought early exit: When to stop thinking in large language models. *arXiv preprint arXiv:2310.04369*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Andrew K. Lampinen, Ishita Dasgupta, James L. McClelland, and Noah D. Goodman. 2024. Can language models learn from explanations in context? *arXiv preprint arXiv:2401.10968*.

Tom Lanham, Yuhuai Wu, and Noah D. Goodman. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Xin Liu and Lu Wang. 2025. Answer convergence as a signal for early stopping in reasoning. *arXiv preprint arXiv:2506.02536*.

Minjia Mao, Bowen Yin, Yu Zhu, and Xiao Fang. 2025. Early stopping chain-of-thoughts in large language models. *arXiv preprint arXiv:2509.14004*.

QwenTeam. 2025. Qwq-32b: Embracing the power of reinforcement learning. https://qwen.ai/blog?id=6aed6aa257238a0b6c77a6753f180350c2fecc4a&from=research.research-list. Accessed: 2025-10-04.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Charlie Snell, James Thorne, and Roger Grosse. 2023. Reducing the cost of reasoning via confidence-based early stopping. *arXiv preprint arXiv:2310.08838*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*.

A Probability Change at the Critical Chunk

This section contains figures analyzing the change in probability for the correct answer at the identified critical chunk.

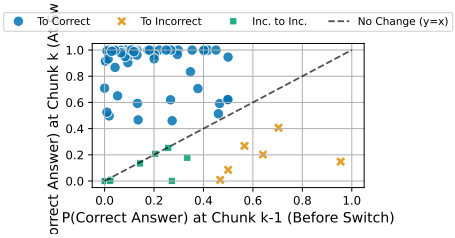

Figure 5: Scatter plot of probability change at the critical chunk for Qwen3-32B on GPQA. Points above the diagonal (y=x) indicate probability increases.

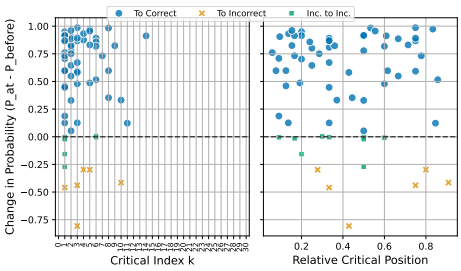

Figure 6: Probability change as a function of critical chunk position (absolute and relative) for Qwen3-32B on GPQA.

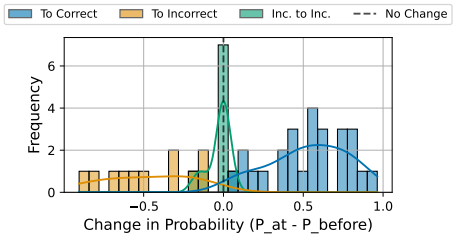

Figure 7: Distribution of probability change at the critical chunk for Qwen3-32B on LogiQA.

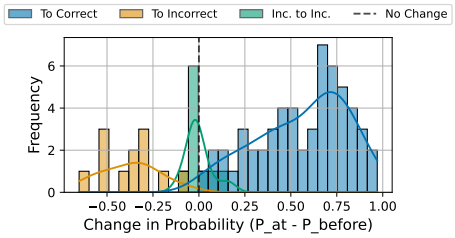

Figure 8: Distribution of probability change at the critical chunk for DeepSeek-R1 on GPQA.

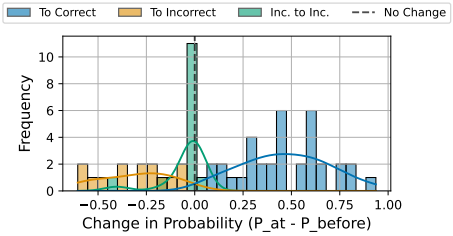

Figure 9: Distribution of probability change at the critical chunk for DeepSeek-R1 on LogiQA.

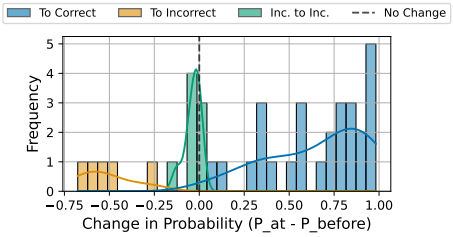

Figure 10: Distribution of probability change at the critical chunk for QwQ-32B on GPQA.

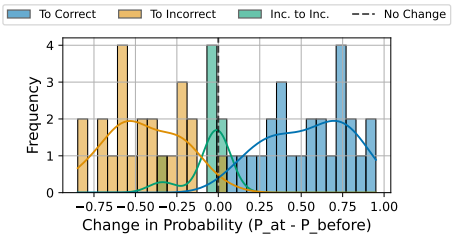

Figure 11: Distribution of probability change at the critical chunk for QwQ-32B on LogiQA.

B Position of the Critical Chunk

This section shows the distribution of the position of the critical chunk within the reasoning trace.

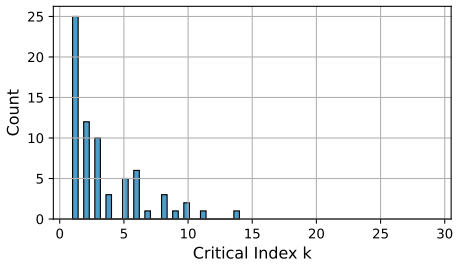

Figure 12: Distribution of the absolute position of the critical chunk for Qwen3-32B on GPQA.

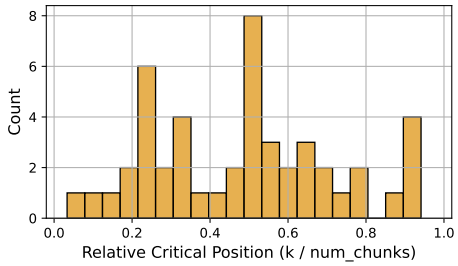

Figure 13: Distribution of the relative position of the critical chunk for Qwen3-32B on LogiQA.

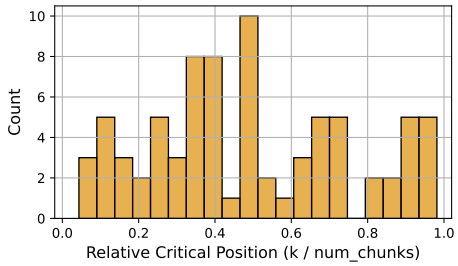

Figure 14: Distribution of the relative position of the critical chunk for DeepSeek-R1 on GPQA.

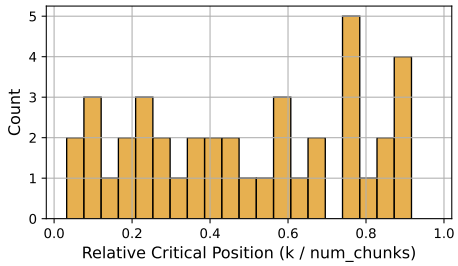

Figure 15: Distribution of the relative position of the critical chunk for QwQ-32B on GPQA.

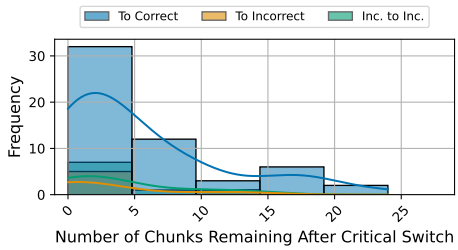

Figure 16: Distribution of chunks remaining after the critical switch for Qwen3-32B on GPQA.

C Leap Heuristic Performance

This section evaluates the performance of the leap heuristic across all models and datasets.

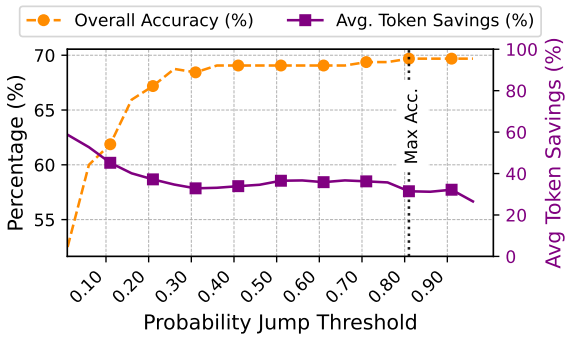

Figure 17: Performance of the leap heuristic for Qwen3-32B on GPQA.

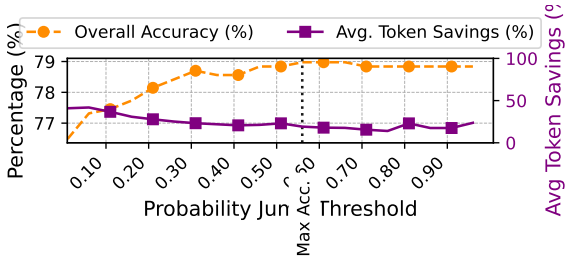

Figure 18: Performance of the leap heuristic for Qwen3-32B on LogiQA.

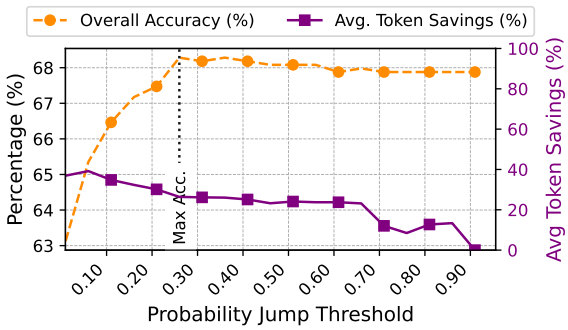

Figure 19: Performance of the leap heuristic for DeepSeek-R1 on LogiQA.

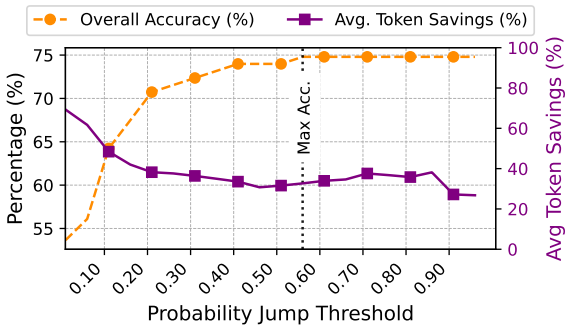

Figure 20: Performance of the leap heuristic for QwQ-32B on GPQA.

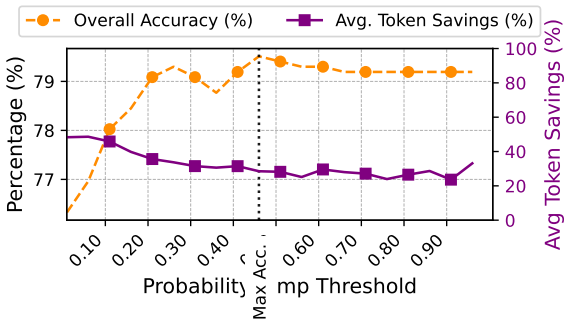

Figure 21: Performance of the leap heuristic for QwQ-32B on LogiQA.

D Accuracy vs. Reasoning Length

This section shows the relationship between accuracy and reasoning chain length across all models and datasets, stratified by the number of answer changes.

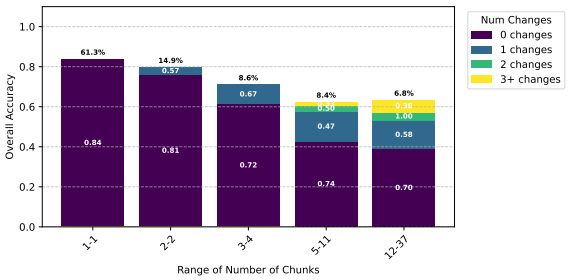

Figure 22: Accuracy vs. reasoning chain length for Qwen3-32B on LogiQA.

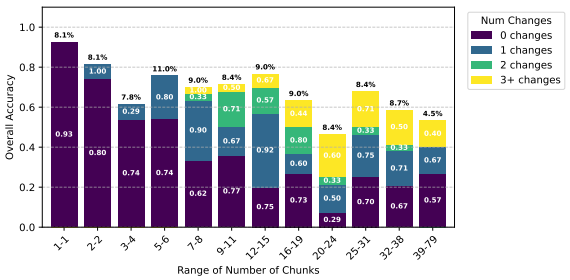

Figure 23: Accuracy vs. reasoning chain length for DeepSeek-R1 on GPQA.

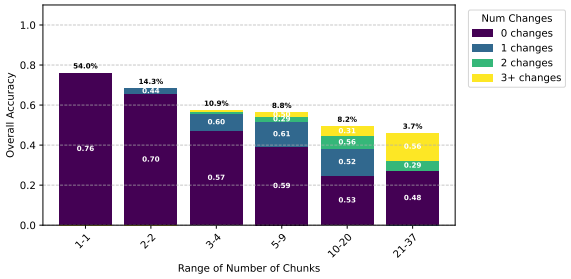

Figure 24: Accuracy vs. reasoning chain length for DeepSeek-R1 on LogiQA.

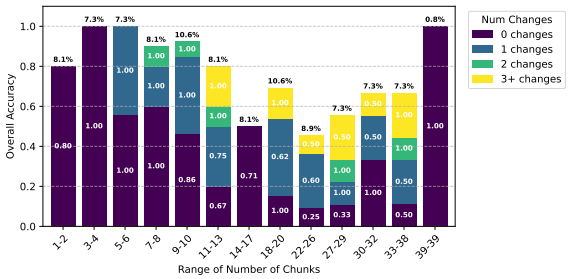

Figure 25: Accuracy vs. reasoning chain length for QwQ-32B on GPQA.

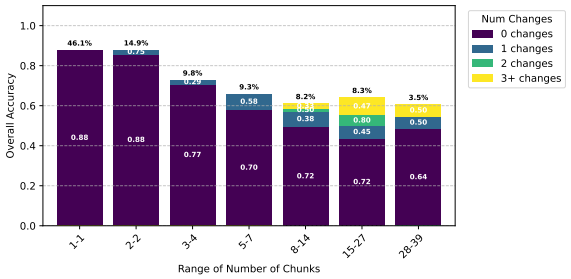

Figure 26: Accuracy vs. reasoning chain length for QwQ-32B on LogiQA.

E Chunking Statistics

This section provides basic statistics on the chunking process, such as the number and length of chunks.

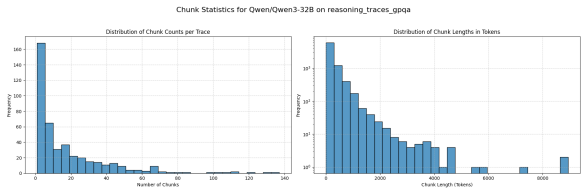

Figure 27: Distribution of chunk counts and lengths for Qwen3-32B on GPQA.

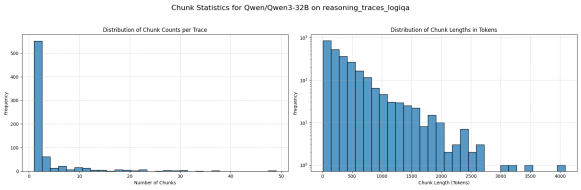

Figure 28: Distribution of chunk counts and lengths for Qwen3-32B on LogiQA.