Statistical Early Stopping for Reasoning Models

Yangxinyu Xie[1], **Tao Wang**[1], **Soham Mallick**[1], **Yan Sun**[2], **Georgy Noarov**[1], **Mengxin Yu**[3]
Tanwi Mallick[4], **Weijie J. Su**[1], **Edgar Dobriban**[1]
[1]University of Pennsylvania
[2] New Jersey Institute of Technology
[3] Washington University in St. Louis
[4]Argonne National Laboratory

Abstract

While LLMs have seen substantial improvement in reasoning capabilities, they also sometimes overthink, generating unnecessary reasoning steps, particularly under uncertainty given ill-posed or ambiguous queries. We introduce statistically principled early stopping methods that monitor uncertainty signals during generation to mitigate this issue. Our first approach is nonparametric and provides finite-sample guarantees on the probability of halting too early on well-posed queries. Our second approach is parametric: it models inter-arrival times of uncertainty keywords as a renewal process and applies sequential testing for stopping. We conduct empirical evaluations on reasoning tasks across several domains and models. Our results indicate that uncertainty-aware early stopping can improve both efficiency and reliability in LLM reasoning. The performance varies across domains, and we observe especially significant gains for math reasoning.

1 Introduction

Large language models (LLMs) have made remarkable progress in multi-step reasoning, yet they sometimes still struggle when faced with ill-posed or ambiguous queries, see e.g., [12, 15], etc. Instead of abstaining or clarifying, models often attempt to provide definitive answers [12, 15]. This tendency can undermine reliability and waste computation on answers that should not have been generated. A related failure mode is *overthinking*: producing unnecessarily long reasoning traces that do not improve accuracy. Though reasoning can improve performance, empirical evidence suggests that verbose reasoning sometimes correlates with incorrect or uncertain predictions [19, 5], and that reasoning models may verbalize uncertainty without abstaining [16, 5, 15]. Together, these findings point to a central challenge: reasoning models lack principled mechanisms to regulate reasoning dynamically in response to uncertainty.

To address this, we propose *statistically principled early stopping methods* that monitor uncertainty signals during token generation. Our first approach is nonparametric: a conformal prediction–based method with finite-sample guarantees, ensuring that the probability of halting too early on well-posed queries is controlled at a pre-selected false positive rate. Next, we introduce a parametric alternative grounded in renewal process theory. It models the inter-arrival times of uncertainty keywords and leverages asymptotic properties of renewal processes to construct sequential tests for stopping.

Through systematic experiments across math, science, and medical reasoning tasks, we show that our approaches improve efficiency (cutting unnecessary tokens on ill-posed queries), while avoiding premature halts on well-posed queries and maintaining accuracy. Performance varies across tasks, and is especially promising for math. Our findings suggest that uncertainty-aware early stopping is a promising mechanism for reducing computational cost in reasoning models.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

Background Our work adapts conformal anomaly detection e.g., [20, 13] etc, to address overthinking when queries are ambiguous or ill-posed. Previous work suggests that reasoning models generate longer responses for ill-posed queries [5, 15, 19]. As a baseline, we therefore limit the trace length uniformly across all examples. Given calibration traces, a length threshold τ is computed as the $(1-\alpha)(1+1/n)$-quantile of trace lengths $|T_i|$, $i = 1, \ldots, n$ in our calibration dataset. For new queries, generation continues until the length exceeds τ, at which point token generation stops. This method effectively controls the false positive rate (FPR)—defined here as the probability of halting too early on a well-posed query [20, 13].[1]

2 Methods

Qualitative analyses from [12, 5, 15] suggest that models often verbalize uncertainty in their reasoning chains, yet still provide definitive final answers. Inspired by these findings, we propose algorithms that leverage such uncertainty signals for early stopping.

Uncertainty Keyword Set Construction We construct a domain-specific lexicon of uncertainty keywords using a semi-supervised approach (details in Appendix A). Specifically, we compare reasoning traces generated by three reasoning models[2] for a subset of math problems from AbstentionBench [12]. A random forest classifier is trained on k-gram features to distinguish traces generated for original and ill-posed questions. The most informative k-grams are extracted as seed keywords, then manually curated and expanded into a categorized lexicon summarized in Table 1, with full details in Appendix A. There are four main categories: Doubt/Speculation, Questioning the Premise, Missing Information, and Contrasting Possibilities. The keywords within these categories serve as interpretable signals of uncertainty in reasoning traces. Importantly, keyword matching is far more efficient than using a secondary LLM to judge uncertainty, making it practical for large-scale or real-time applications.

Category	Examples
Doubt / Speculation	"maybe," "not entirely sure"
Questioning the Premise	"maybe the question," "the problem could be"
Missing Information	"not provided," "without details"
Contrasting Possibilities	"but maybe," "alternatively it could be"

Table 1: Overview of uncertainty keywords. See Appendix A for the full taxonomy.

Uncertainty Score. For a tokenized trace prefix $T[1:\ell]$ of length $\ell > 0$, we define a (fast-to-compute and scalable) uncertainty score $u(T;\ell)$, the frequency of uncertainty expressions:

$$u(T;\ell) = \#\{\text{uncertainty phrases in } T[1:\ell]\}/\ell. \tag{1}$$

Maxwise Conformal Stopping Our first approach computes the maximum of the uncertainty score over the traces. We partition traces into bins of tokens of size[3] B, with boundaries $L_j = j \cdot B$, $j = 1, \ldots$. We collect calibration traces T_i, $i = 1, \ldots, n$, on well-posed problems, representing reasoning paths of the model we are interested in. For each calibration trace T_i and boundary $L_j \leq |T_i|$, we compute the score $u(T_i; L_j)$ on the prefix $T_i[1:L_j]$. We then define a global threshold of "maximal uncertainty" as follows. For each calibration trace T_i, compute $M_i = \max_{j:L_j \leq |T_i|} u(T_i; L_j)$. Then, to quantify how large these scores typically get during normal reasoning, we calculate their $(1-\alpha)(1+1/n)$ quantile, for some user-specified $\alpha \in (0,1)$, i.e., $\tau^\star = \text{Quantile}_{(1-\alpha)(1+1/n)}(\{M_i\}_{i=1}^n)$. For a new query T at inference/test-time, we monitor $u(T; L_j)$, $j = 1, \ldots$ and stop as soon as $u(T, L_j) > \tau^\star$. This controls the probability of early stopping on a well-posed query at level α (see Appendix B for a formal statement).

Renewal Process Stopping Our second approach is a parametric rule based on renewal process theory [6]. Here, each occurrence of an uncertainty phrase is treated as an "arrival" in a renewal process. This rule stops decoding when the observed arrival rate of uncertainty phrases is significantly higher than expected. From calibration traces, we extract inter-arrival times A_j, $j = 1, \ldots$. We estimate the

[1] For completeness, a proof is provided in Appendix B.1.

[2] `Qwen/QwQ-32B`, `microsoft/Phi-4-reasoning`, and `deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`

[3] Here, B is treated as a hyperparameter and we choose $B = 100$; see Section 4 about its choice.

mean and variance of their distribution as[4] $\hat{\mu} = \sum_{j=1}^{M} A_j/M$ and $\hat{\sigma}^2 = \sum_{j=1}^{M}(A_j - \hat{\mu})^2/(M-1)$. For a test trace, let N_t be the number of arrivals up to position t. Renewal theory [see 6, Sec. 10.2, p. 417] implies that the normalized statistic $Z_t = (N_t - t/\hat{\mu}) \Big/ \sqrt{t\hat{\sigma}^2/\hat{\mu}^3}$ converges in distribution to a standard normal[5] as $t \to \infty$. If Z_t exceeds $z_{1-\alpha'}$, the $(1-\alpha')$ standard normal quantile, we halt generation at step t, for $t = k \cdot B$ for some $k = 1, \ldots$ and B is the same as before. To account for repeated testing, we adjust the significance level using the Šidák correction [18]: $\alpha' = 1-(1-\alpha)^{1/T}$, where $T = \left\lceil \frac{L_{\max}}{B} \right\rceil$ is the maximum number of tests performed for this trace.

3 Experiments

Prompting Baselines We distill two prompting baselines from prior work: 1. Confidence Dampening Prompt: Following [9], append: *"Answer only if you are confident. Otherwise, say 'I am not sure.' "* to encourage abstention. 2. Critical Reasoning Prompt: Following [4], append: *"Please solve these problems with criticism. If the problem is reasonable, think step by step and put your final answer in a box. If the problem is unreasonable, highlight the issues clearly and provide a succinct explanation."* to encourage critique of unreasonable inputs while maintaining structured reasoning.

Experimental Setup We set `max_new_tokens` = 32,768 with default temperature and decoding method defined by model providers. For simplicity, we use identical zero-shot prompts and do not apply any model-specific tuning. We evaluate 6 RL-tuned and 3 distilled reasoning models, spanning families such as DeepSeek [7], Qwen [21], Phi [2]; the full list is in Appendix C.1.

Datasets & Benchmarks We build on **AbstentionBench** [12], which constructs benchmark subsets from existing datasets by pairing original well-posed problems with ill-posed or unanswerable variants. We follow their subsets and also cap the evaluation size at 200 samples per benchmark for comparability. The resulting benchmarks span mathematical, scientific, and medical reasoning. Each dataset is split 50/50 into calibration and test sets. The datasets are summarized in Appendix C.2.

Evaluation Metrics We evaluate each method for *false positive control* and *power/efficiency*. All metrics are averaged across models and datasets within each domain.

False Positive Rate (FPR) Control: Defined as the probability of stopping too early on an well-posed query (operationalized as the *early stopping rate on original benchmarks*). Lower values indicate stronger FPR control. For the prompting baselines, we use the drop in accuracy as a proxy.

Power / Efficiency: The ability to stop early in ill-posed cases while not truncating useful reasoning otherwise. We measure this using: (a) *Early stopping rate on ill-posed benchmarks* (Power; higher values are better): the frequency of early stopping on ill-posed queries. For the prompting baselines, we use the abstention rate as a proxy. (b) *Token savings on ill-posed benchmarks* (Efficiency): the percentage of tokens saved relative to the full trace length.

Generalization to MIP. To test out-of-distribution robustness, we additionally evaluate on the *MiP benchmark (Missing Premise)* [5], which introduces unanswerable variants of math problems, constructed differently from AbstentionBench. Models are calibrated on AbstentionBench GSM8K and evaluated on MiP. The dataset sizes for all MiP variants are reported in Appendix C.3.

3.1 Results

Effective FPR Control. Across math, science, and medical domains, uncertainty-based rules achieve the target FPR control. This indicates that our proposed early stopping methods effectively control the false positive rate.

Power in Math (In-Domain). On math tasks, uncertainty-based rules exhibit high power. The Maxwise Rule achieves the highest early stopping rate (90.99%), closely followed by the Renewal Process Rule (89.45%). Both far outperform the prompting and length baselines. This suggests

[4]For ease of exposition, we continue to use the term "calibration phase" to describe parameter estimation.

[5]While the theoretical result applies when we have a renewal process, e.g., when the sequence of tokens is a Markov chain, in experiments, we have found that it provides a reasonable approximation even when the sequence of tokens is generated from a language model.

that some component of math reasoning traces—possibly their logical structure—make uncertainty signals powerful.

Domain Generalization. When moving to science and medical reasoning, power drops. For example, on the Science dataset, the Renewal Process Rule achieves only 45.71% early stopping. On the Medical dataset, the Maxwise Rule reaches 19.67%. Thus, while uncertainty-based rules generalize well in terms of FPR control, their power is domain-sensitive.[6] Despite the decrease in drop, our proposed stopping rules outperform all baselines.

Stopping Rule	Math		Science		Medical	
	FPR	Power	FPR	Power	FPR	Power
No Intervention	0.00%	9.42%	0.00%	0.56%	0.00%	0.00%
Confidence	1.05%	17.15%	1.67%	12.22%	0.70%	0.56%
Criticism	1.94%	17.81%	1.11%	16.11%	1.11%	1.00%
Length	3.91%	20.85%	1.11%	1.11%	4.20%	6.60%
Maxwise	4.08%	90.99%	6.11%	37.78%	3.59%	19.67%
Renewal	3.52%	89.45%	3.89%	41.67%	4.42%	15.73%

Table 2: Average early stopping rates across domains. "Original (FPR)" = stopping too early on well-posed queries. "AbstentionBench (Power)" = stopping on unanswerable/ambiguous queries.

Efficiency (Tokens Saved). Efficiency patterns mirror early stopping rates. The Maxwise Rule saves the most tokens in Math (85.25%), but efficiency declines in Science (35.63%) and Medical (12.82%). The Renewal Process Rule follows a similar trend. Length truncation saves negligible tokens across all domains.

Generalization to MiP. When calibrated on GSM8K and tested on MiP, only uncertainty-driven rules preserves FPR control. For MiP queries, the Maxwise Rule (81.48%) and Renewal Process Rule (81.39%) rules dominate again, outperforming the Length rule (45.97%).

Stopping Rule	Math	Science	Medical
Maxwise	85.25%	35.63%	12.82%
Renewal	77.88%	32.06%	8.01%
Length	5.98%	0.00%	1.17%

(a) Efficiency (Tokens Saved).

Stopping Rule	Original (FPR)	MiP (Power)
Maxwise	3.87%	81.48%
Renewal	1.96%	81.39%
Length	12.91%	45.97%

(b) Generalization to MiP.

Table 3: Left: Average percentage of tokens saved on ill-posed queries. Right: Generalization to MiP. "Original (FPR)" = early stopping on well-posed GSM8K calibration queries. "MiP (Power)" = early stopping on MiP unanswerable variants.

4 Discussion and Conclusion

We introduce two efficient mechanisms for truncating reasoning traces based on detecting uncertainty cues, with guaranteed false-positive control and good performance across diverse datasets spanning math, science and medical domains. In contrast, the baselines fail to produce meaningful improvements. Our results are thus complementary to previous findings [12, 5] that unreasonable problems inherently trigger longer reasoning. Instead, our data show that some models generate shorter but equally flawed traces when premises are missing, making length an unreliable stopping signal and requiring more nuanced uncertainty-based strategies.

These are preliminary results, and several directions require further investigation. First, our method shows reduced power on medical tasks, highlighting the need for domain-specific adaptations. Second, some hyperparameters, such as the bin size B in our maxwise conformal stopping rule, needs to be carefully decided. Third, broader benchmarking—beyond simple prompting methods—will be

[6]There are a number of possible reasons, including the uncertainty keywords we constructed were extracted from mathematical reasoning traces, making them especially tailored to mathematical reasoning. While a detailed investigation of the reasons for this is beyond the scope of our current paper, they are an important topic for future research.

important for further evaluation. We leave these directions, along with more comprehensive ablation studies to strengthen generalization, to future work.

5 Acknowledgment

This work was supported in part by the US NSF, DOE, ARO, AFOSR, ONR, the Simons Foundation and the Sloan Foundation. The opinions expressed in this document are solely those of the authors and do not represent the views of the above institutions.

References

[1] Github copilot. `https://en.wikipedia.org/wiki/GitHub_Copilot`. Accessed on September 8, 2025.

[2] Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, et al. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*, 2025.

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[4] Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

[5] Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *arXiv preprint arXiv:2504.06514*, 2025.

[6] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford University Press, Oxford; New York, 2001.

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

[8] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

[9] Yin Huang, Yifan Ethan Xu, Kai Sun, Vera Yan, Alicia Sun, Haidar Khan, Jimmy Nguyen, Mohammad Kachuee, Zhaojiang Lin, Yue Liu, et al. Confqa: Answer only if you are confident. *arXiv preprint arXiv:2506.07309*, 2025.

[10] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[11] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Craft-md: A conversational evaluation framework for comprehensive assessment of clinical llms. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

[12] Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

[13] Rikard Laxhammar and Göran Falkman. Sequential conformal anomaly detection in trajectories based on hausdorff distance. In *14th international conference on information fusion*, pages 1–8. IEEE, 2011.

[14] Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.

[15] Jingyuan Ma, Damai Dai, Zihang Yuan, Weilin Luo, Bin Wang, Qun Liu, Lei Sha, Zhifang Sui, et al. Large language models struggle with unreasonability in math problems. *arXiv preprint arXiv:2403.19346*, 2024.

[16] Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know? *arXiv preprint arXiv:2506.18183*, 2025.

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

[18] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American statistical association*, 62(318):626–633, 1967.

[19] Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. *arXiv preprint arXiv:2505.00127*, 2025.

[20] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

[21] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A Construction of Uncertainty Keywords

A.1 Semi-Supervised Keyword Identification

To construct a domain-relevant and interpretable set of uncertainty keywords, we employ a semi-supervised procedure that combines model outputs, feature extraction, and manual curation. The process begins by comparing reasoning traces on benchmark datasets (`gsm8k`, `mmlu`) across several reasoning-oriented LLMs, including `Qwen/QwQ-32B`, `microsoft/Phi-4-reasoning`, and `deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`.

We create a small training set by pairing original reasoning traces with corresponding unanswerable ones from AbstentionBench [12]. Using only the training split to avoid contamination, we train a Random Forest classifier with k-gram features ($k = 2, 3, 4$) to discriminate between the two types of traces. Then, we extract the top-20 most informative k-gram features from each model-dataset combination. Running the procedure across both datasets and all three models produces multiple candidate feature lists, which are merged into a single sorted list of unique candidates.

A.2 Augmenting the Keyword Set

The merged candidate features serve as seeds for constructing a more comprehensive uncertainty keyword set. Following the methodology of AbstentionBench, we manually categorize and expand these seed phrases, ensuring semantic coverage across multiple ways models express doubt or hesitation. To enrich the set further, we leverage automated code completion tools to suggest semantically similar variants, which are curated for relevance and precision [1].

A.3 Categorization of Uncertainty Expressions

To support interpretability and modular use, we organize the keyword set into semantic categories:

- **Doubt/Speculation:** explicit signals of uncertainty or hedging, e.g., "maybe," "not entirely sure."
- **Questioning the Premise:** expressions that highlight potential flaws in the query itself, e.g., "maybe the question," "the problem could be."

- **Missing Information:** indicators of insufficient or absent data, e.g., "without specific details," "don't have any information."
- **Contrasting Possibilities:** hedged alternatives framed with doubt, e.g., "but maybe," "alternatively it could be."

This structured taxonomy makes the keyword set directly usable as a feature for uncertainty scoring functions. For example, the density of matched keywords in a reasoning prefix can serve as a nonparametric measure of uncertainty, which is then integrated into our conformal and renewal-based early stopping frameworks.

B Conformal Early Stopping Methods

B.1 Length-Based Early Stopping

Algorithm 1 Conformal Stopping Threshold for Reasoning Traces (Length-based)

Require: Calibration set $\{(X_i, T_i)\}_{i=1}^n$, confidence level $\alpha \in (0, 1)$
Ensure: Estimated stopping threshold $\hat{\tau}$
Initialize list of stopping steps $S \leftarrow [\,]$
for each (X_i, T_i) in calibration set **do**
    $\ell_i \leftarrow |T_i|$ ▷ Compute the length of each reasoning trace
end for
Sort the lengths in ascending order: $\ell_{(1)} \leq \ell_{(2)} \leq \cdots \leq \ell_{(n)}$
Let $k = \lceil (n+1)(1-\alpha) \rceil$
Set the threshold $\hat{\tau} = \ell_{(k)}$
Return $\hat{\tau}$

Prediction Phase:
For a new input X, generate reasoning trace $T = (t_1, t_2, \ldots)$ token by token.
Once the length of the generated trace exceeds $\hat{\tau}$, stop. Otherwise, continue generating.

Proposition B.1. *Assume $(\ell_1, \ldots, \ell_n, \ell_{n+1})$ are exchangeable, where ℓ_i is the reasoning trace length for query X_i. Let $\hat{\tau} = \ell_{(k)}, k = \lceil (n+1)(1-\alpha) \rceil$. Then $\mathbb{P}(\ell_{n+1} > \hat{\tau}) \leq \alpha$. Equivalently, the false positive rate of stopping too early on a well-posed query is controlled at level α.*

Proof. Let R be the rank of ℓ_{n+1} among $\{\ell_1, \ldots, \ell_n, \ell_{n+1}\}$ when sorted in nondecreasing order (ties broken deterministically or randomly). By exchangeability, R is marginally uniform on $\{1, \ldots, n+1\}$. By construction, the event $\{\ell_{n+1} \leq \ell_{(k)}\}$ is equivalent to $\{R \leq k\}$. Therefore, $\mathbb{P}(\ell_{n+1} \leq \hat{\tau}) = \mathbb{P}(R \leq k) \geq \frac{k}{n+1} \geq 1 - \alpha$. Equivalently, $\mathbb{P}(\ell_{n+1} > \hat{\tau}) \leq \alpha$. Since our procedure stops reasoning whenever $\ell_{n+1} > \hat{\tau}$, the probability of prematurely stopping on a well-posed query—i.e., the false positive rate—is at most α. This completes the proof. □

B.2 Maxwise Conformal Stopping

Proposition B.2. *Let $M_i = \max_{j: L_j \leq |T_i|} u_i(L_j)$ for calibration traces, and M_{n+1} for the test trace. Define $\tau^\star$ as in Algorithm 2. Then*

$$\mathbb{P}(\exists j : u_{n+1}(L_j) > \tau^\star) = \mathbb{P}(M_{n+1} > \tau^\star) \leq \alpha.$$

Proof. The values $(M_1, \ldots, M_n, M_{n+1})$ are exchangeable since each M_i is a deterministic functional of the corresponding trace. By conformal validity, the test statistic M_{n+1} falls below the $(1-\alpha)(1 + 1/n)$ quantile $\tau^\star$ with probability at least $1 - \alpha$. Hence $\mathbb{P}(M_{n+1} > \tau^\star) \leq \alpha$. Equivalently, the chance that the test trace ever crosses the global threshold across all bins is controlled at level α. □

Algorithm 2 Maxwise Conformal Stopping (Uncertainty-based)

Require: Calibration set $\{(X_i, T_i)\}_{i=1}^n$, bin size B, confidence level $\alpha \in (0, 1)$
Ensure: Global stopping threshold $\tau^\star$
1: Define bin boundaries $L_j = j \cdot B$ up to max trace length.
2: **for** each calibration trace T_i **do**
3: Compute $M_i = \max_{j:L_j \leq |T_i|} u_i(L_j)$.
4: **end for**
5: Set global threshold $\tau^\star = \text{Quantile}_{(1-\alpha)(1+1/n)}(\{M_i\}_{i=1}^n)$.
6: **Return** $\tau^\star$.

Prediction Phase:
7: For a new trace T, after the first bin boundary L_1, check uncertainty at bin boundaries.
8: Stop if $u(L_j) > \tau^\star$ for some j.

C More on Experiments

C.1 Models Evaluated

The list of models evaluated are included in Table C.1. Inference is performed using `vLLM` on four NVIDIA A100-SXM4 GPUs (40GB VRAM each).

Category	Model	Size
RL-Tuned	QwQ-32B	32B
	Qwen3-32B	32B
	Phi-4-reasoning-plus	14B
	AceReason-Nemotron-14B	14B
	MiMo-7B-RL-0530	7B
	Skywork-OR1-32B	32B
Distilled	DeepSeek-R1-Distill-Qwen-32B	32B
	Qwen3-8B	8B
	Phi-4-reasoning	14B

Table 4: Models evaluated in our experiments, covering both RL-tuned and distilled reasoning families.

C.2 Datasets Evaluated

Domain	Dataset	Size	Description
Math	GSM8K	200	Grade-school math word problems [3].
	MMLU	133	Math subsets (college math, abstract algebra, high school math) [8].
Science	GPQA	40	Graduate-level, Google-proof QA. Diamond subset [17].
Medical	MEDQA	200	Patient records followed by a multiple-choice question [14, 10].
	CRAFT-MD	137	Dermatology patient records in the MEDQA format [14, 11].

Table 5: AbstentionBench-derived benchmark subsets used in our experiments.

C.3 MIP Dataset Sizes

Table 6 reports the sizes of the different subsets of the MIP benchmark used in our experiments. For the MiP-GSM8K split, we remove all instances identical to the original GSM8K problems to avoid data leakage.

MIP Subset	Size
MIP-Formula	50
MIP-Math500	52
MIP-GSM8K	492
MIP-SVAMP	300

Table 6: Dataset sizes for the MIP benchmark subsets.