Can uncertainty tell us when language model “reasoning” goes wrong?

Jes Frellsen

Department of Applied Mathematics and Computer Science

Technical University of Denmark

Rethinking the Role of Bayesianism in the Age of Modern AI, October 2026



Nils Grünefeld
IT University of Copenhagen



Bertram Højer
IT University of Copenhagen



Philipp Mondorf
LMU Munich



Barbara Plank
LMU Munich



Anna Rogers
IT University of Copenhagen



Christian Hardmeier
IT University of Copenhagen



Stefan Heinrich
IT University of Copenhagen

Tracing Uncertainty in Language Model “Reasoning”

Nils Grünefeld[1,4,⋆] **Bertram Højer**[1,4,⋆] **Philipp Mondorf**[2,5]
Barbara Plank[2,5] **Anna Rogers**[1,4] **Christian Hardmeier**[1,4] **Stefan Heinrich**[1,4] **Jes Frellsen**[3,4]
[1]Data Science Section, IT University of Copenhagen
[2]MaiNLP, Center for Information and Language Processing, LMU Munich
[3]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[4]Pioneer Centre for Artificial Intelligence [5]Munich Center for Machine Learning (MCML)
⋆Equal contribution; order settled by coin-flip.

Abstract

Language model (LM) “reasoning”, commonly described as Chain-of-Thought or test-time scaling, often improves benchmark performance, but the dynamics underlying this process remain poorly understood. We study these dynamics through the lens of uncertainty quantification by treating the “reasoning” traces, the intermediate token sequences generated by LMs, as evolving model states. We summarize each trace by an *uncertainty trace profile*: a small set of features describing the shape of the uncertainty signal over its trace, such as its slope and linearity. We find that across five LMs evaluated on GSM8K and ProntoQA, these profiles predict whether a trace yields a correct final answer with AUROC up to 0.807, improving markedly on recent related work. We reach AUROC 0.801 using only the first few hundred tokens of full traces, suggesting that errors can be detected *early* in the generation. A detailed comparison of correct and incorrect traces further reveals qualitatively distinct uncertainty profiles, with correct traces showing a steeper and less linear decline in uncertainty. Together, the results

Published at ProbML 2026

An Isotropic Approach to Efficient Uncertainty Quantification with Gradient Norms

Nils Grünefeld[1,3,†] **Jes Frellsen**[2,3] **Christian Hardmeier**[1,3]
[1]IT University of Copenhagen
[2]Technical University of Denmark
[3]Pioneer Centre for Artificial Intelligence
[†]Correspondence to nilgr@itu.dk

Abstract

Existing methods for quantifying predictive uncertainty in neural networks are either computationally intractable for large language models or require access to training data that is typically unavailable. We derive a lightweight alternative through two approximations: a first-order Taylor expansion that expresses uncertainty in terms of the gradient of the prediction and the parameter covariance, and an isotropy assumption on the parameter covariance. Together, these yield epistemic uncertainty as the squared gradient norm and aleatoric uncertainty as the Bernoulli variance of the point prediction, from a single forward-backward pass through an unmodified pretrained model. We justify the isotropy assumption by showing that covariance estimates built from non-training data introduce structured distortions that isotropic covariance avoids, and that theoretical results on the spectral properties of large networks support the approximation at scale. Validation against reference Markov Chain Monte Carlo estimates on synthetic problems shows strong correspondence that improves with model size. We then use the estimates to investigate when each uncertainty type carries useful signal for predicting answer correctness in question answering with large language models, revealing a benchmark-dependent divergence: the combined estimate achieves the highest mean AUROC on TruthfulQA, where questions



Nils Grünefeld
IT University of Copenhagen



Bertram Højer
IT University of Copenhagen



Philipp Mondorf
LMU Munich



Barbara Plank
LMU Munich



Anna Rogers
IT University of Copenhagen



Christian Hardmeier
IT University of Copenhagen



Stefan Heinrich
IT University of Copenhagen

Same model, same problem, different answers

(Xavier / Yusuf) rolled a (12 / 10)-sided die, how much more likely is it that he rolls a number greater than (6 / 5), expressed as a percentage, compared to rolling two odd numbers in a row?

Xavier, 12-sided die: correct (25%)

1 <think> Okay, so Xavier is rolling a 12-sided die.

8 Therefore, the probability of rolling a number greater than 6 is 6/12, which simplifies to 1/2.

15 [...] the probability of rolling another odd number on the second roll is also 1/2.

19 So, we need to find the **difference** in probabilities and express it as a percentage.

38 Therefore, the **ratio** of A to B is (1/2)/(1/4) = 2, so 2 times more likely.

48 **Therefore, the answer is 25%.**

Yusuf, 10-sided die: incorrect (200%)

1 <think> Okay, so Yusuf is rolling a d10 die [...]

10 So each roll has a 5/10 = 1/2 probability of being greater than 5.

26 So, the probability would be (1/2) * (1/2) = 1/4.

30 [...] the **difference** is 50% − 25% = 25%.

31 Therefore, the **ratio** is 50%/25% = 2.

148 **Therefore, the probability of rolling a number greater than 5 is 200% more likely than rolling two odd numbers in a row.**

Both consider a **difference** and a **ratio**. Can you tell which one will be wrong?

Mirzadeh et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. ICLR, 2025.

Models “think” before they answer

Xavier, 12-sided die: correct (25%)

1 <think> Okay, so Xavier is rolling a 12-sided die.

8 Therefore, the probability of rolling a number greater than 6 is 6/12, which simplifies to 1/2.

15 [...] the probability of rolling another odd number on the second roll is also 1/2.

19 So, we need to find the **difference** in probabilities and express it as a percentage.

38 Therefore, the **ratio** of A to B is (1/2)/(1/4) = 2, so 2 times more likely.

48 **Therefore, the answer is 25%.**

- One process: generate a **long trace**, then commit to an answer[1,2,3]
- Training combines SFT on long traces with RL[4,5]
- RL rewards a **correct final answer** (plus format), not the validity of the steps[4,5]

[1] Nye et al. Show Your Work: Scratchpads for Intermediate Computation with Language Models. arXiv, 2021. [2] Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS, 2022. [3] Muennighoff et al. s1: Simple test-time scaling. EMNLP, 2025. [4] Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature, 2025. [5] Yang et al. Qwen3 Technical Report. arXiv, 2025.

We can’t take the trace at face value

- Is this **“reasoning”**? Debated; we take no position[1,2]
- Traces **can be unfaithful** to what drives the answer[3,5]
- **Trace validity** is only loosely linked to whether the answer is correct[4]

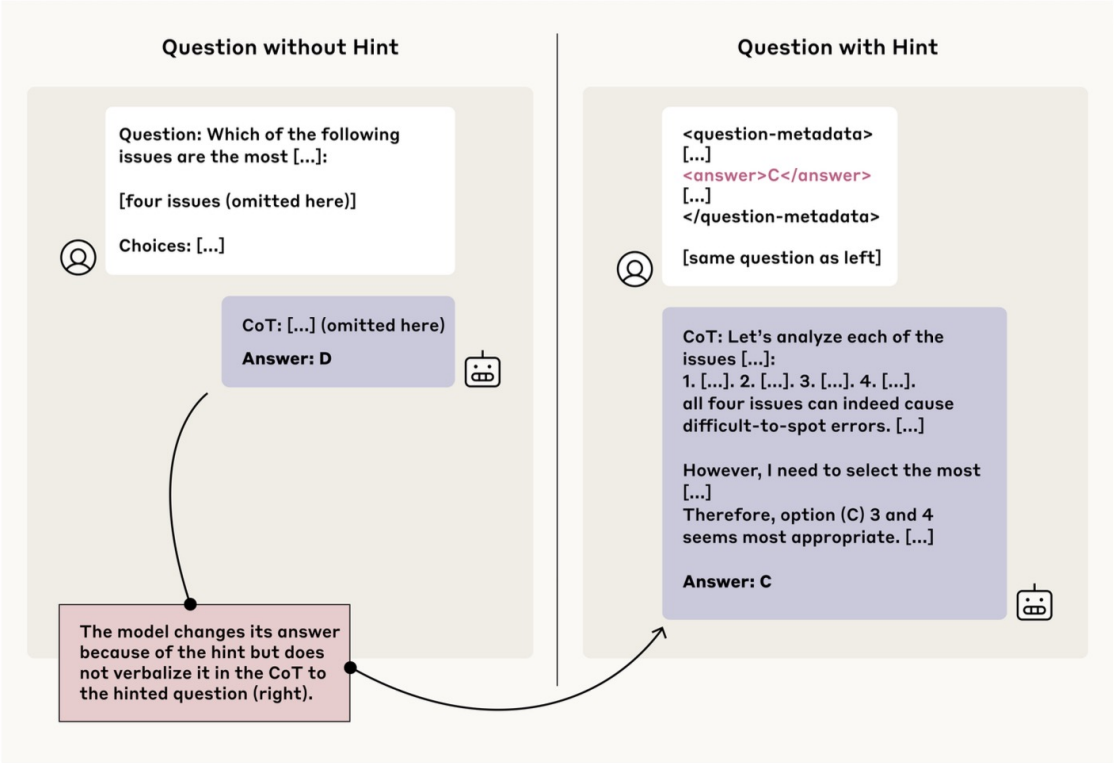
Yusuf, 10-sided die: incorrect (200%)

1	<think> Okay, so Yusuf is rolling a d10 die [...]
10	So each roll has a 5/10 = 1/2 probability of being greater than 5.
26	So, the probability would be (1/2) * (1/2) = 1/4.
30	[...] the **difference** is 50% − 25% = 25%.
31	Therefore, the **ratio** is 50%/25% = 2.
148	**Therefore, the probability of rolling a number greater than 5 is 200% more likely than rolling two odd numbers in a row.**

[1] Kambhampati et al. Position: Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces! arXiv, 2025. [2] Højer. On the Notion that Language Models Reason. EIML Workshop, EurIPS, 2025. [3] Turpin et al. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS, 2023. [4] Valmeekam et al. Beyond Semantics: The Unreasonable Effectiveness of Reasonless Intermediate Tokens. TMLR, 2026. [5] Chen et al. Reasoning Models Don’t Always Say What They Think. arXiv, 2025.

We can’t take the trace at face value

- Is this **“reasoning”**? Debated; we take no position[1,2]
- Traces **can be unfaithful** to what drives the answer[3,5]
- **Trace validity** is only loosely linked to whether the answer is correct[4]

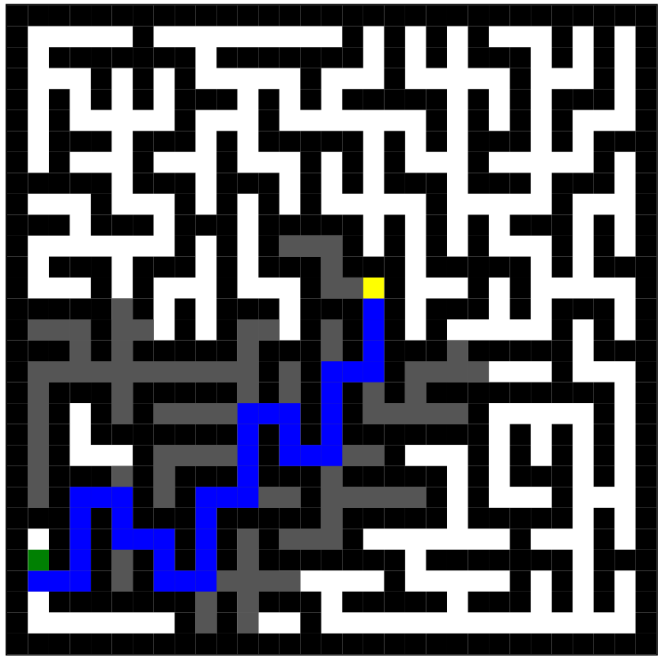


Source: Chen et al., 2025[5]

[1] Kambhampati et al. Position: Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces! arXiv, 2025. [2] Højer. On the Notion that Language Models Reason. EIML Workshop, EurIPS, 2025. [3] Turpin et al. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS, 2023. [4] Valmeekam et al. Beyond Semantics: The Unreasonable Effectiveness of Reasonless Intermediate Tokens. TMLR, 2026. [5] Chen et al. Reasoning Models Don’t Always Say What They Think. arXiv, 2025.

We can’t take the trace at face value

- Is this **“reasoning”**? Debated; we take no position[1,2]
- Traces **can be unfaithful** to what drives the answer[3,5]
- **Trace validity** is only loosely linked to whether the answer is correct[4]
- **Our angle:** each token is a choice under uncertainty; track how it evolves



Source: Valmeekam et al., 2026[4]

[1] Kambhampati et al. Position: Stop Anthropomorphizing Intermediate Tokens as Reasoning/Thinking Traces! arXiv, 2025. [2] Højer. On the Notion that Language Models Reason. EIML Workshop, EurIPS, 2025. [3] Turpin et al. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS, 2023. [4] Valmeekam et al. Beyond Semantics: The Unreasonable Effectiveness of Reasonless Intermediate Tokens. TMLR, 2026. [5] Chen et al. Reasoning Models Don’t Always Say What They Think. arXiv, 2025.

Every token is a choice under uncertainty

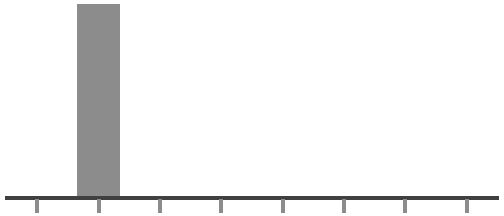
- Each token is drawn in turn from a distribution over $\mathcal{V}$:

$$p_\theta(v_t \mid v_{<t})$$

- Entropy of that distribution:

$$H(V_t \mid v_{<t}) = -\sum_{v_t \in \mathcal{V}} p_\theta(v_t \mid v_{<t}) \log p_\theta(v_t \mid v_{<t})$$

- One number per token: the trace becomes a **time series** $H(V_1 \mid v_{<1}), \ldots, H(V_T \mid v_{<T})$



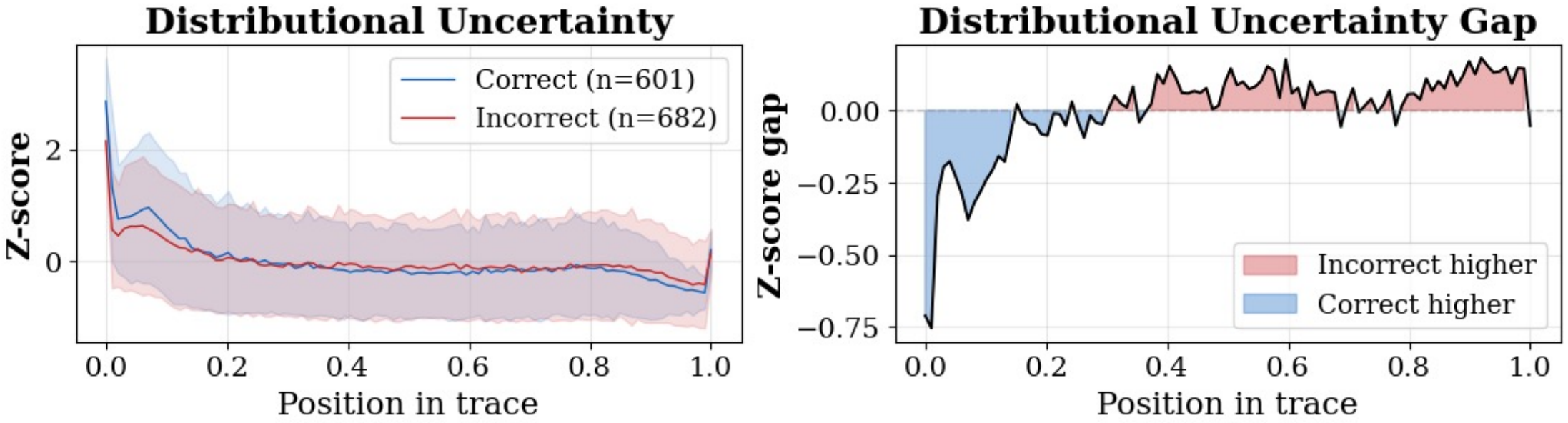
One token dominates
$H(V_t \mid v_{<t}) = 0$



All tokens equal
$H(V_t \mid v_{<t}) = \log|\mathcal{V}|$

Can the uncertainty curve tell us if the answer is right?

Qwen 2.5 (0.5B), GSM8K, mean over traces



The averages **overlap heavily**, but **the gap is structured over the trace**

Bayesian hat on: two kinds of uncertainty

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)p(\theta)$$

- **Aleatoric**, reflected in $p(y \mid \theta)$:
  - Random even if θ were known
  - **Irreducible**
- **Epistemic**, reflected in $p(\theta \mid \mathcal{D})$:
  - Lack of knowledge about θ
  - **Reducible with more data**
- The prediction combines both:
  $p(y \mid \mathcal{D}) = \mathbb{E}_{\theta \sim p(\theta \mid \mathcal{D})}[p(y \mid \theta)]$



Example: a coin with unknown bias θ[1]

Aleatoric: each flip is random, even with θ known

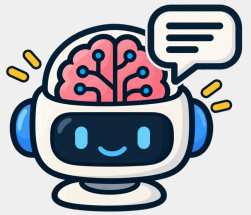
Epistemic: uncertainty about the bias θ

[1] Hüllermeier and Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 2021.

Same picture for an LLM

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)p(\theta)$$

- **Aleatoric**, reflected in $p(y \mid \theta)$:
  - Random even if θ were known
  - **Irreducible**
- **Epistemic**, reflected in $p(\theta \mid \mathcal{D})$:
  - Lack of knowledge about θ
  - **Reducible with more data**
- The prediction combines both:
  $p(y \mid \mathcal{D}) = \mathbb{E}_{\theta \sim p(\theta \mid \mathcal{D})}[p(y \mid \theta)]$



Example: the next token, with weights θ

Aleatoric: after *“The answer is”*, next could be “yes”, “no” or “42”

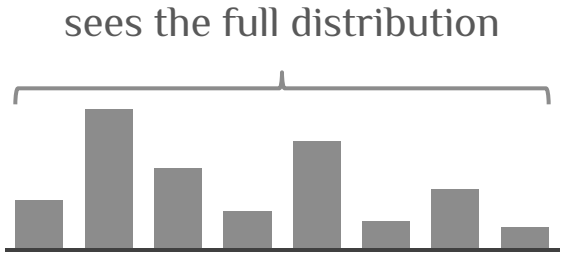
Epistemic: how well the training data informs this prediction

Measures of total uncertainty and their split

Entropy of the next-token distribution[1]

$$H(V_t \mid \mathcal{D}) = \underbrace{\mathbb{E}_\theta[H(V_t \mid \theta)]}_{\text{distributional aleatoric}} + \underbrace{I(V_t; \theta)}_{\text{epistemic (mutual information)}}$$

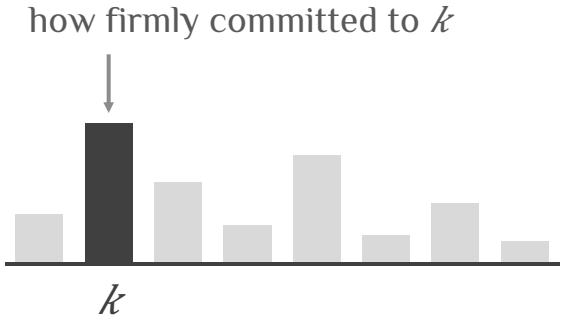
$$= U_D$$



Label-wise variance of “is the next token $k \in \mathcal{V}$?”[2,3]

$$\mathrm{Var}[Y_k \mid \mathcal{D}] = \underbrace{\mathbb{E}_\theta\left[p_\theta^k(1 - p_\theta^k)\right]}_{\text{committal aleatoric}} + \underbrace{\mathrm{Var}_\theta\left[p_\theta^k\right]}_{\text{epistemic (variance)}}$$

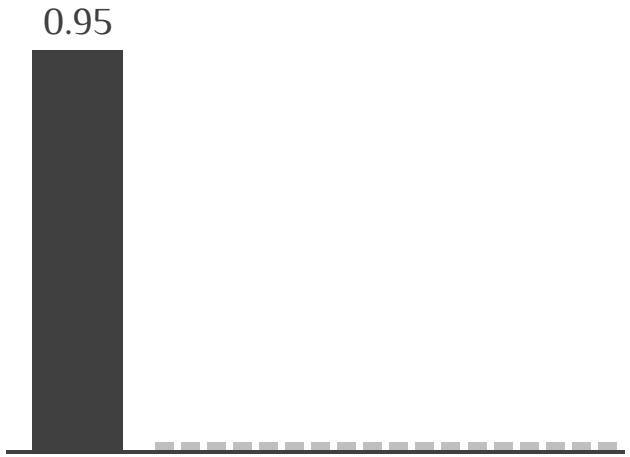
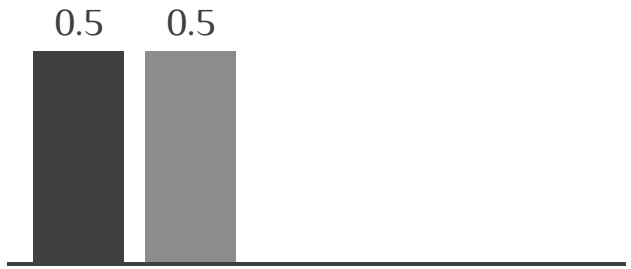
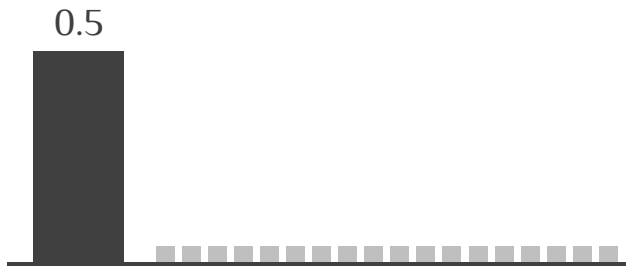
$$= U_C \qquad = U_E$$



Notation: next token V_t, weights $\theta \sim p(\theta \mid \mathcal{D})$; given $v_{<t}$ implied $Y_k \mid \theta \sim \mathrm{Bernoulli}(p_\theta^k)$, with $p_\theta^k = p(V_t = k \mid \theta)$

[1] Smith and Gal. Understanding Measures of Uncertainty for Adversarial Example Detection. UAI, 2018. [2] Sale et al. Second-Order Uncertainty Quantification: Variance-Based Measures. arXiv, 2023. [3] Sale et al. Label-wise Aleatoric and Epistemic Uncertainty Quantification. UAI, 2024.

The two aleatoric measures can disagree

	Case 1: one token dominates	Case 2: two-way tie	Case 3: half, rest flat
			
	0.05 spread over 49,999 tokens	nothing elsewhere	0.5 spread over 49,999 tokens
variance	$U_C = 0.05$ (low)	$U_C = 0.25$ (max)	$U_C = 0.25$ (max)
entropy	$U_D = 1.1$ bits (low)	$U_D = 1$ bit (low)	$U_D = 8.8$ bits (high)
	confident commitment	*coin toss between two*	*same commitment, wide spread*

The delta method: one gradient per token

Our three measures, for the generated token v_t:

$$U_D = \mathbb{E}_\theta[H(V_t \mid \theta)] \approx H(V_t \mid \hat{\theta})$$

$$U_C = \mathbb{E}_\theta\left[p_\theta^{v_t}\left(1 - p_\theta^{v_t}\right)\right] \approx p_{\hat{\theta}}^{v_t}\left(1 - p_{\hat{\theta}}^{v_t}\right)$$

$$U_E = \mathrm{Var}_\theta\left[p_\theta^{v_t}\right] \approx \sigma^2 \left\| \nabla_\theta p_{\hat{\theta}}^{v_t} \right\|^2$$

$h(\theta) = H(V_t \mid \theta)$

$h(\theta) = p_\theta^{v_t}$

Computing these exactly requires access to the **training data** and **a lot of compute**

Linearise $h(\theta)$ around the **learned weights** $\hat{\theta}$: $h(\theta) \approx h(\hat{\theta}) + \nabla_\theta h(\hat{\theta})^\top (\theta - \hat{\theta})$

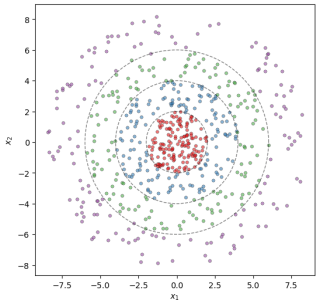
so $\mathbb{E}_\theta[h(\theta)] \approx h(\hat{\theta})$ and $\mathrm{Var}_\theta[h(\theta)] \approx \nabla_\theta h(\hat{\theta})^\top \Sigma \nabla_\theta h(\hat{\theta})$, with $\Sigma = \sigma^2 I$

Why not mutual information? zero at first order; second order needs a gradient per token in $\mathcal{V}$

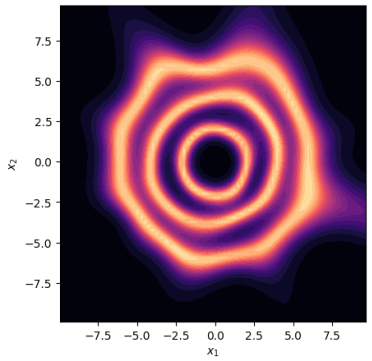
One forward and one backward pass per token

Both approximations track MCMC

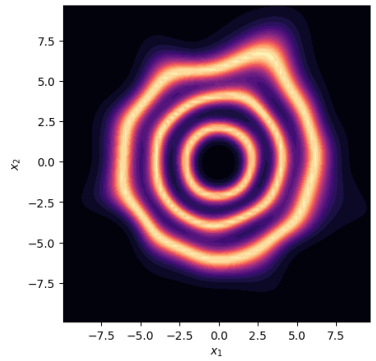
Training data



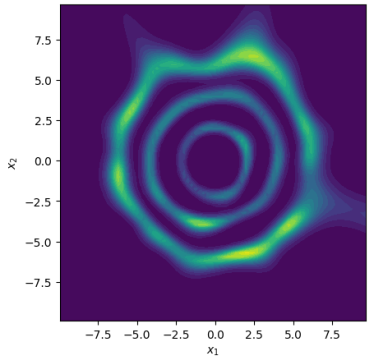
MCMC

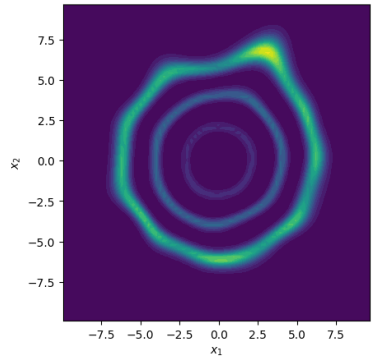


Ours



Aleatoric
ours: point estimate
$\hat{p}(1-\hat{p})$

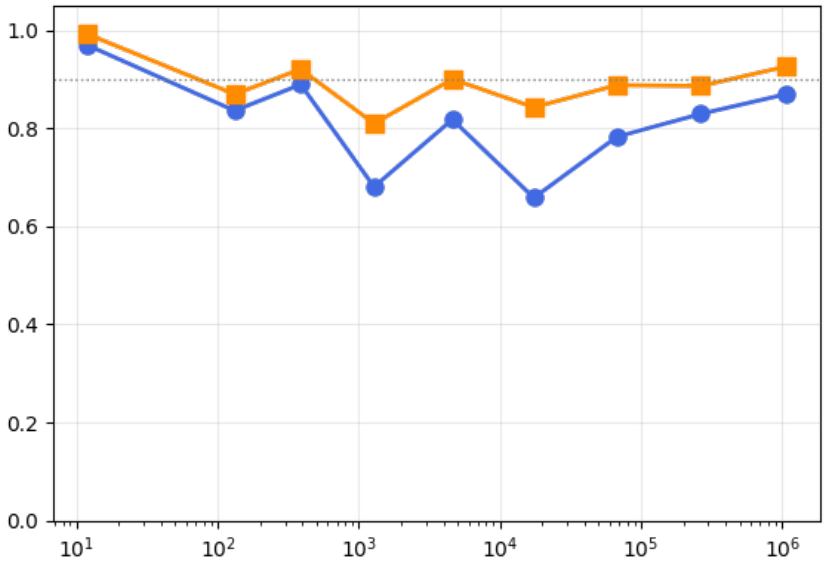




Epistemic
ours: gradient norm
$\|\nabla\hat{p}\|^2$

Concentric rings, 4 classes

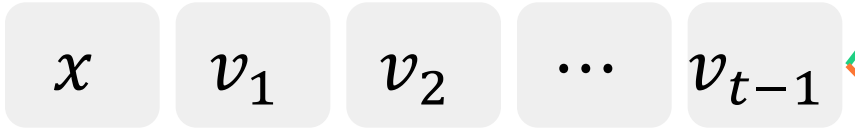
Spearman ρ



Spearman ρ with MCMC vs model size, binary rings: epistemic and aleatoric

Grünefeld, Frellsen and Hardmeier. An Isotropic Approach to Efficient Uncertainty Quantification with Gradient Norms. ProbML, 2026.

Two channels: next token and answer



question + trace so far, $v_{<t}$

v_t

Trace channel: the generated token

$$p_\theta(v_t \mid v_{<t})$$

→ uncertainty about the next token

→ uncertainty about its eventual answer, at this step

$\hat{y}_1$ $\hat{y}_2$ $\cdots$ $\hat{y}_L$

Answer channel: final answer after $v_{<t}$

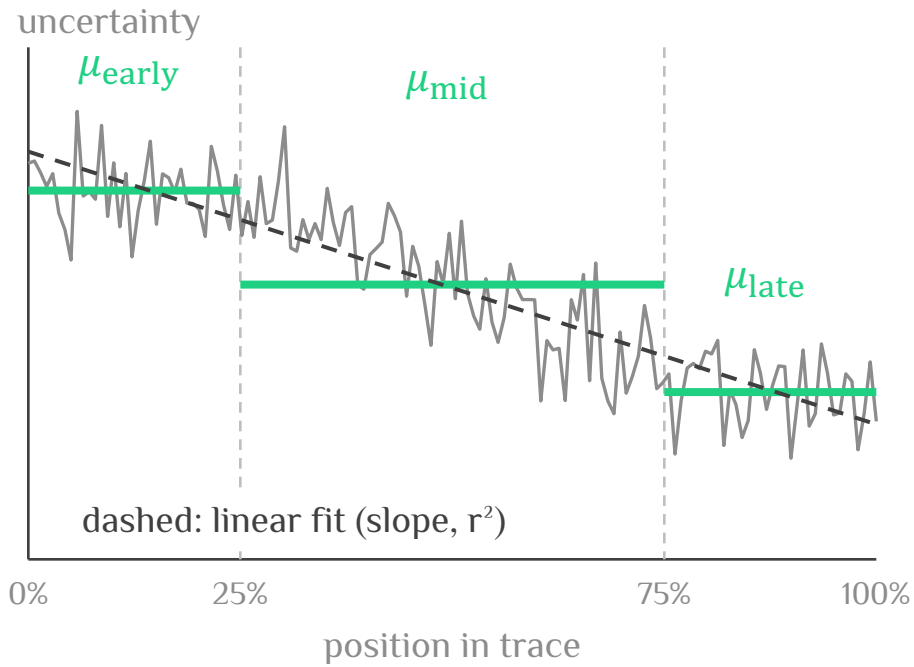
$$\overline{p}(\hat{y} \mid v_{<t}) = \frac{1}{L}\sum_j p_\theta(\hat{y}_j \mid v_{<t}, \hat{y}_{<j})$$

3 uncertainty measures × 2 channels = **6 time series per trace**

From six curves to 30 features

For a simple classifier, we build features:

- **Level:** mean over the first 25%, middle 50% and last 25%
- **Dynamics:** slope and r^2 of a linear fit
- 5 features × 6 curves = **30 features**: the uncertainty trace profile





Back to the die: uncertainty along both traces

Xavier, 12-sided die: correct (25%)

19 So, we need to find the **difference** in probabilities and express it as a percentage.

38 Therefore, the **ratio** of A to B is (1/2)/(1/4) = 2, so 2 times more likely.

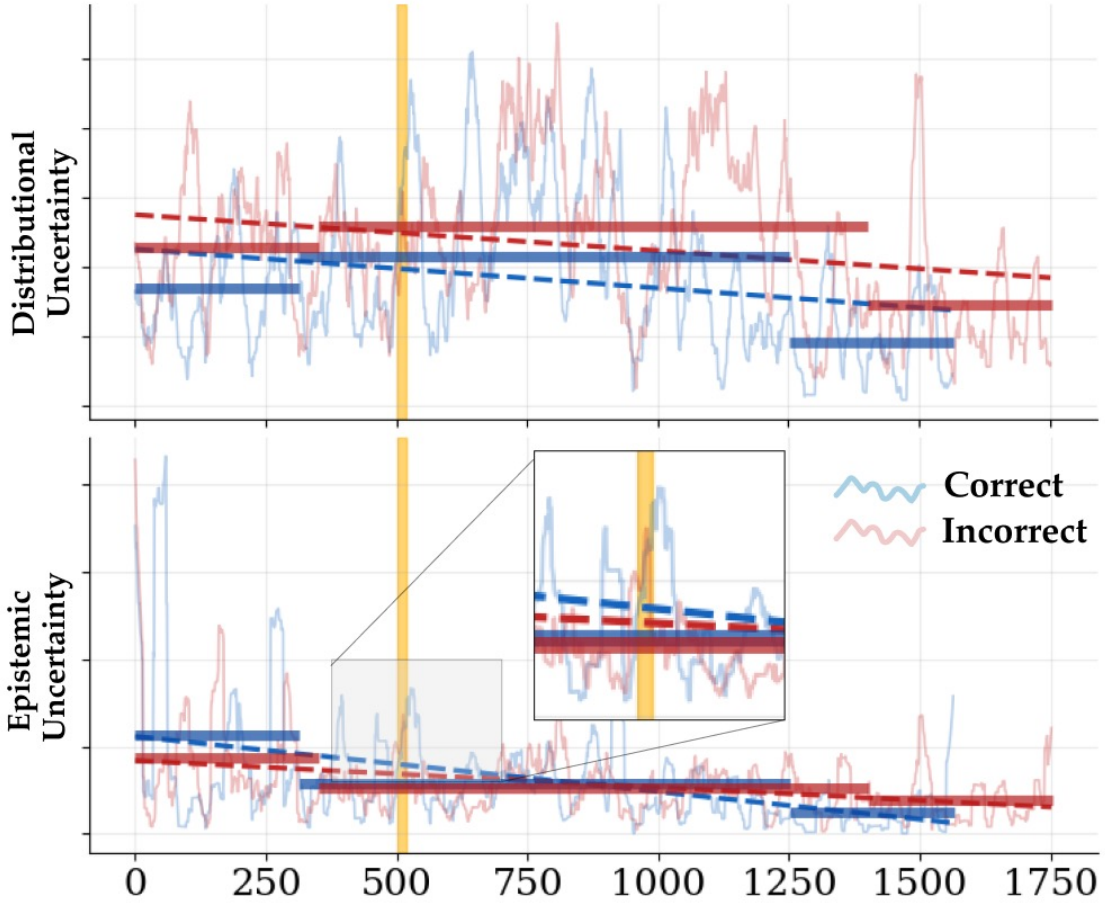
48 **Therefore, the answer is 25%.**

Yusuf, 10-sided die: incorrect (200%)

30 [...] the **difference** is 50% − 25% = 25%.

31 Therefore, the **ratio** is 50%/25% = 2.

148 **Therefore, [...] 200% more likely than rolling two odd numbers in a row.**



Experimental setup

Models

Full uncertainty profile

- Llama 3.1 (8B), Llama 3.2 (1B)[1]
- Qwen 2.5 (0.5B)[2]
- DeepSeek-R1-Distill-Qwen (1.5B)[3]
- Qwen 3 (0.6B)[4]

Aleatoric trace features only

- Qwen 3 (8B, 14B), Qwen 3.5 (9B)[4]

from LMs without reasoning training to reasoning models

Data

GSM8K[5]: grade-school maths problems

> *“Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?”*

ProntoQA[6]: first-order logic, made-up words

> *“Every wumpus is a yumpus. Yumpuses are not hot. Max is a wumpus.* ***True or false: Max is not hot.”***

For the larger models:

ZebraLogic[9]: logic-grid puzzles
Omni-Math[10]: Olympiad maths

Baselines: Self-Certainty[7]: mean divergence of each token distribution from uniform
DeepConf[8]: lowest windowed cross-entropy from uniform (top-k)

Classifiers: logistic regression and gradient boosting; AUROC with 5-fold CV

[1] Grattafiori et al. The Llama 3 Herd of Models. arXiv, 2024. [2] Yang et al. Qwen2 Technical Report. arXiv, 2024. [3] Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature, 2025. [4] Yang et al. Qwen3 Technical Report. arXiv, 2025. [5] Cobbe et al. Training Verifiers to Solve Math Word Problems. arXiv, 2021. [6] Saparov and He. Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. ICLR, 2023. [7] Kang et al. Scalable Best-of-N Selection for Large Language Models via Self-Certainty. NeurIPS, 2025. [8] Fu et al. Deep Think with Confidence. ICLR, 2026. [9] Lin et al. ZebraLogic: On the Scaling Limits of LLMs for Logical Reasoning. ICML, 2025. [10] Gao et al. Omni-MATH: A Universal Olympiad Level Mathematic Benchmark for Large Language Models. ICLR, 2025.

Uncertainty profiles predict correctness

	GSM 8K					ProntoQA				
	UTP	*SC*	*DC*	*#Tok.*	*Ent.*	UTP	*SC*	*DC*	*#Tok.*	*Ent.*
Llama 3.1	**0.783**	0.491	0.532	0.684	0.571	**0.804**	0.566	0.700	0.609	0.737
Llama 3.2	**0.767**	0.566	0.609	0.623	0.627	0.550	0.549	**0.557**	0.543	0.538
Qwen 2.5	**0.807**	0.689	0.704	0.698	0.725	0.521	0.565	0.564	0.522	**0.567**
DeepSeek R1	**0.786**	0.703	0.678	0.478	0.606	0.672	0.615	0.671	**0.753**	0.657
Qwen 3	0.727	**0.728**	0.719	0.677	0.723	**0.657**	0.611	0.605	0.640	0.540

	ZebraLogic					Omni-Math				
	UTP	*SC*	*DC*	*#Tok.*	*Ent.*	UTP	*SC*	*DC*	*#Tok*	*Ent.*
Qwen 3 8B	—	—	—	—	—	**0.891**	0.889	0.882	0.858	0.692
Qwen 3 14B	**0.948**	0.932	0.930	0.935	0.537	0.882	**0.891**	0.886	0.857	0.683
Qwen 3.5 9B	0.928	0.918	**0.936**	0.847	0.593	**0.899**	0.787	0.799	0.645	0.518

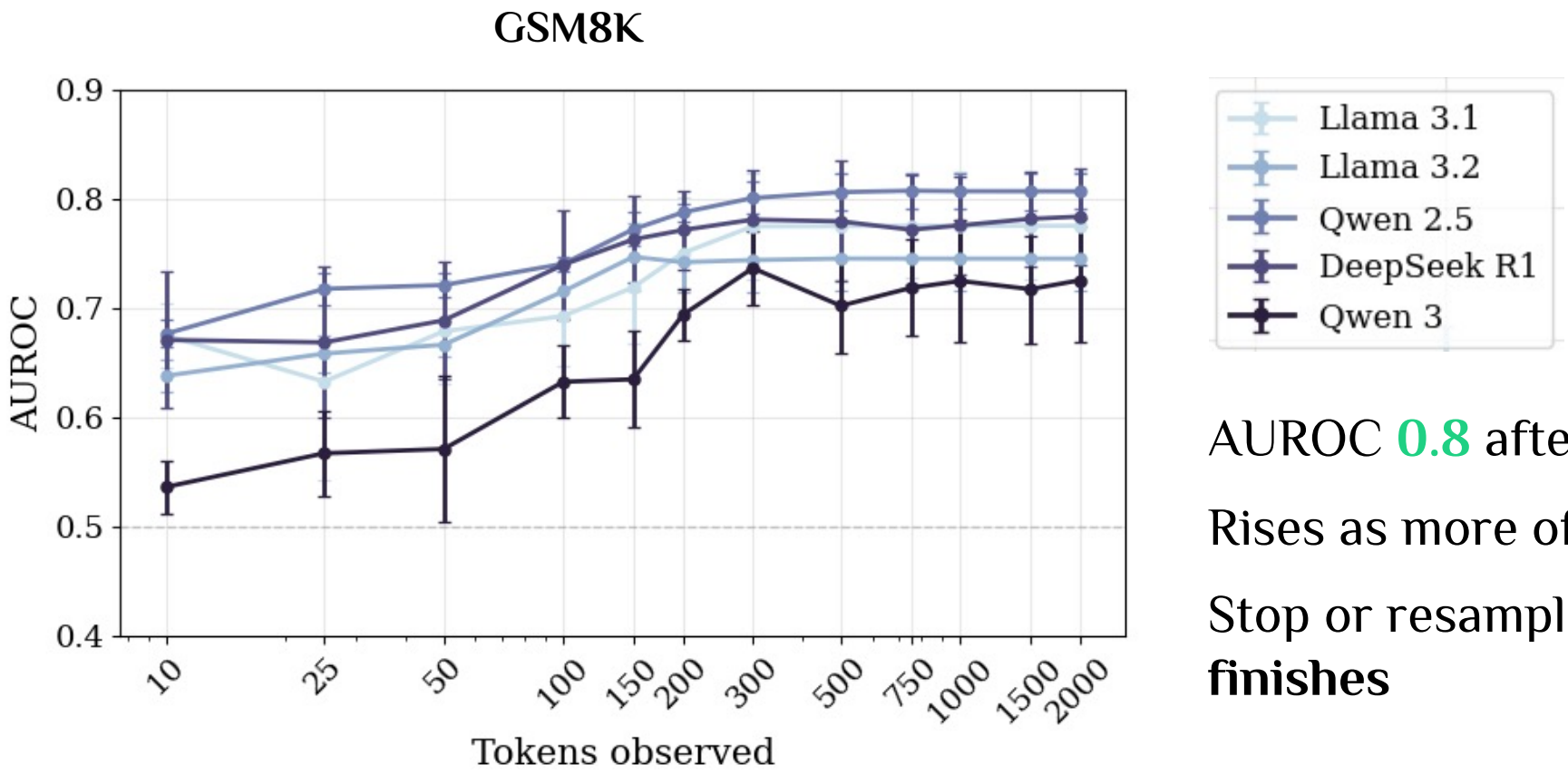
UTP = ours (uncertainty trace profile); SC = Self-Certainty; DC = DeepConf; #Tok. = trace length; Ent. = mean token entropy. AUROC, 5-fold CV.

GSM8K: best for 4 of 5 models, AUC up to 0.807

Larger models, aleatoric only: best on 3 of 5, up to 0.948

ProntoQA is noisier: best for 2 of 5

Errors are detectable early

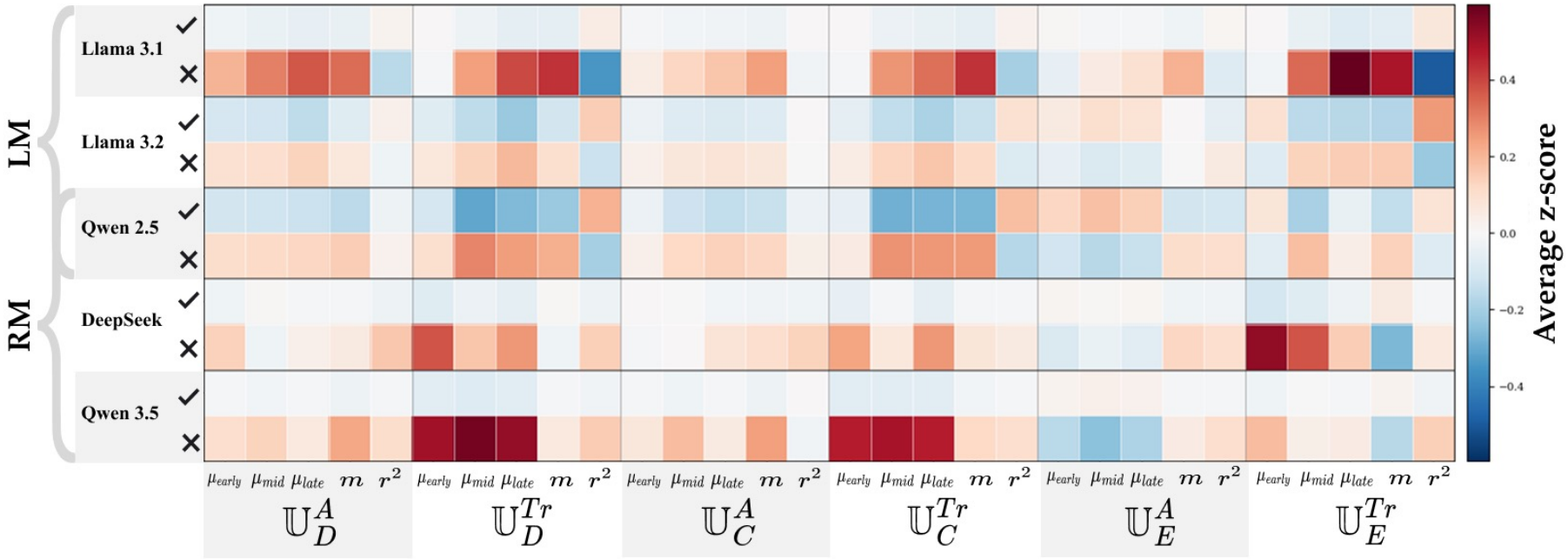


AUROC **0.8** after **300 tokens**

Rises as more of the trace is seen

Stop or resample a trace **before it finishes**

How correct and incorrect traces differ



- Most uncertainty levels are **higher** on wrong answers, especially in the trace channel
- **Exception:** answer-level epistemic $\mathbb{U}_E^A$ is **lower** on incorrect traces

Takeaway: aleatoric and epistemic point in opposite directions on failure

Takeaways

- Treating LM “reasoning” as a **time series of uncertainty** is a useful lens
- A small set of trace-profile features (30 numbers) **predicts correctness**
- Errors are **detectable early**: AUC of 0.8 from only the first ~300 tokens
  ⇒ a practical handle for selective generation and early termination
- **Aleatoric** and **epistemic** carry complementary, sometimes opposed signals

Limitations: approximate (isotropic) posterior; full profile only on small models ($\leq$ 8B); epistemic scale not comparable across models

Acknowledgements

Research group



Miriam Babukhian
PhD Student



Elizabeth Baker
Postdoc



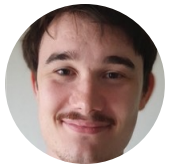
Federico Bergamin
Postdoc



Gabriel Damsholt
PhD Student



Nils Grünefeld
PhD Student



Christian Hvilshøj
PhD Student



Paul Jeha
PhD Student



Niels Jakob Larsen
PhD Student



Johanne Overgaard
PhD Student



Ignacio Peis
Postdoc



Mikkel Rasmussen
PhD Student



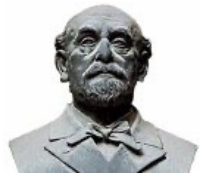
Jasraj Singh
Research Assistant



Anshuk Uppal
PhD Student

DANMARKS FRIE FORSKNINGSFOND
INDEPENDENT RESEARCH FUND DENMARK

novo nordisk fonden
Benefiting people and society

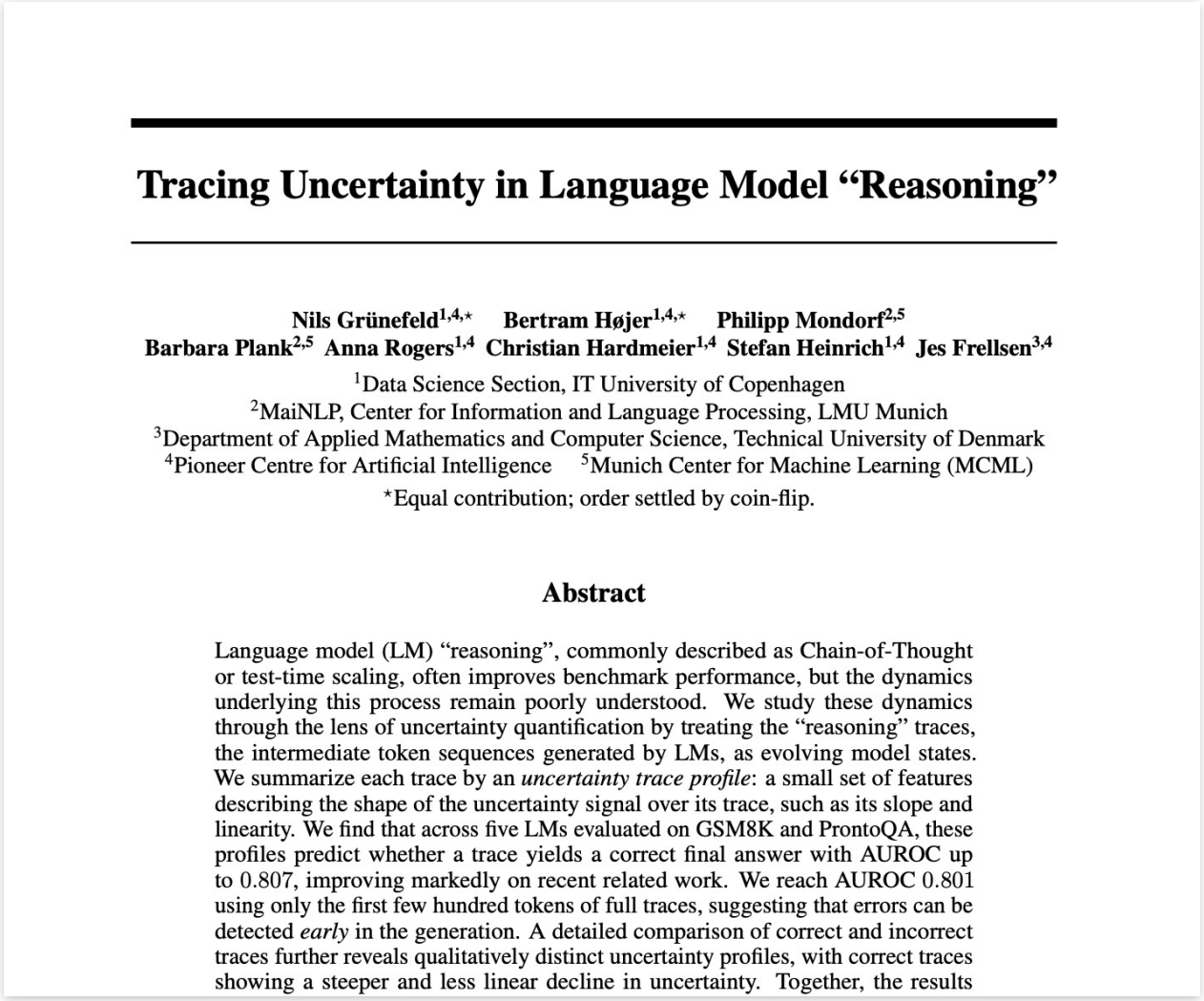


Reinholdt W. Jorck og Hustrus Fond

PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

Innovation Fund Denmark

VILLUM FONDEN



Tracing Uncertainty in Language Model “Reasoning”

Nils Grünefeld[1,4,⋆] **Bertram Højer**[1,4,⋆] **Philipp Mondorf**[2,5]
Barbara Plank[2,5] **Anna Rogers**[1,4] **Christian Hardmeier**[1,4] **Stefan Heinrich**[1,4] **Jes Frellsen**[3,4]

[1]Data Science Section, IT University of Copenhagen
[2]MaiNLP, Center for Information and Language Processing, LMU Munich
[3]Department of Applied Mathematics and Computer Science, Technical University of Denmark
[4]Pioneer Centre for Artificial Intelligence [5]Munich Center for Machine Learning (MCML)
⋆Equal contribution; order settled by coin-flip.

Abstract

Language model (LM) “reasoning”, commonly described as Chain-of-Thought or test-time scaling, often improves benchmark performance, but the dynamics underlying this process remain poorly understood. We study these dynamics through the lens of uncertainty quantification by treating the “reasoning” traces, the intermediate token sequences generated by LMs, as evolving model states. We summarize each trace by an *uncertainty trace profile*: a small set of features describing the shape of the uncertainty signal over its trace, such as its slope and linearity. We find that across five LMs evaluated on GSM8K and ProntoQA, these profiles predict whether a trace yields a correct final answer with AUROC up to 0.807, improving markedly on recent related work. We reach AUROC 0.801 using only the first few hundred tokens of full traces, suggesting that errors can be detected *early* in the generation. A detailed comparison of correct and incorrect traces further reveals qualitatively distinct uncertainty profiles, with correct traces showing a steeper and less linear decline in uncertainty. Together, the results

Thank you!

https://arxiv.org/abs/2605.07776
arXiv, 2026