The First Token Knows: Single-Decode Confidence for Hallucination Detection

Mina Gabriel[1,*], Pei Wang[1]

[1]Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA

Abstract

Sampling-based uncertainty methods detect hallucinations by generating multiple answers and measuring agreement, often with additional semantic clustering. We study whether uncertainty is already visible in the first content-bearing answer token of a single greedy decode. We define first-token confidence, ϕ_{first}, as one minus the normalized entropy of the top-K first-token distribution. Across PopQA and TriviaQA with three 7–8B instruction-tuned models, ϕ_{first} matches or modestly exceeds semantic self-consistency while requiring only one decode instead of one greedy decode plus ten sampled generations and NLI clustering. A subsumption analysis shows that ϕ_{first} is moderately to strongly correlated with semantic agreement, and combining both signals yields only small additional AUROC gains. We position ϕ_{first} as a cheap deployment-oriented baseline for factual QA uncertainty estimation, while noting its limits for high-confidence falsehoods and long-form generation.

Keywords

hallucination detection, uncertainty quantification, factual question answering, confidence estimation, trustworthy AI

1. Introduction

A common paradigm for uncertainty quantification in large language models is *self-consistency*: sample N responses for the same input and use disagreement among them as a proxy for uncertainty. Originally proposed as a decoding strategy for reasoning [1], the same sampling-based principle has become central to several hallucination-detection methods. Semantic uncertainty refines this idea by clustering generations into NLI-based equivalence classes and treating disagreement among clusters as evidence of model uncertainty [2, 3]. These methods provide strong baselines, but require multiple generations per question and a separate NLI-based clustering model.

We argue that, in closed-book short-answer factual QA, where the model answers from its parametric knowledge without retrieved documents, sampling-based methods act as expensive Monte Carlo probes of uncertainty that is already largely visible in the model's first-token logit distribution. For factual questions such as *"Who wrote Hamlet?"* or *"What is the capital of Australia?"*, the first generated answer token often marks the model's earliest commitment to an entity, name, or relation value. If most of the probability mass at this position is concentrated on one token, the model is making a confident early choice about how to begin the answer. If the probability mass is instead spread across several plausible first tokens, the model is unsure which answer to begin generating, even before the rest of the response has unfolded.

We define first-token confidence ϕ_{first} as the normalized entropy of the top-K logits at the first content-bearing answer token of a single greedy decode and compare it against semantic self-consistency, surface-form self-consistency, and verbalized confidence. We further test whether ϕ_{first} captures much of the same uncertainty information as semantic agreement, which requires multiple sampled generations.

Our contributions are: (i) we show that ϕ_{first} matches or modestly exceeds semantic agreement on PopQA and TriviaQA across three 7–8B models, at roughly $1/11$ of the generation cost, before accounting

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.

*Corresponding author.

mina.gabriel@temple.edu (M. Gabriel); pei.wang@temple.edu (P. Wang)

https://minagabriel.github.io/ (M. Gabriel); https://cis.temple.edu/~wangp/ (P. Wang)

0009-0003-2058-3485 (M. Gabriel); 0000-0002-1066-0454 (P. Wang)

for the additional NLI clustering required by semantic agreement; (ii) we provide a subsumption test showing that ϕ_{first} is moderately to strongly correlated with semantic agreement and that a logistic ensemble of the two adds only marginal AUROC over ϕ_{first} alone; and (iii) we show that the apparent relationship between ϕ_{first} and answer length is largely explained by correctness rather than answer length itself.[1]

Our use of the term hallucination detection is restricted to correctness prediction in closed-book factual QA, following common benchmark practice. We do not claim that first-token confidence detects all forms of hallucination. In particular, it is not expected to reliably identify high-confidence falsehoods, where the model assigns high probability to an incorrect answer. The proposed signal is therefore better viewed as a cheap uncertainty-based risk indicator: it can identify cases where the model's answer appears unstable or underdetermined, but it should be complemented by retrieval, external verification, or human review when confident errors are safety-critical.

2. Method

2.1. First-token confidence

Given a single greedy decode of a model's response, let $\ell_t \in \mathbb{R}^{|V|}$ denote the logits at decode step t and $p_{t,i}$ the corresponding softmax probabilities. Let $t^\star$ be the position of the first content-bearing answer token, identified by skipping whitespace, punctuation, and chat-template prefixes such as "Answer:". We take the top-K probabilities at position $t^\star$ (with $K = 100$), renormalize them to $\tilde{p}_{t^\star,1}, \ldots, \tilde{p}_{t^\star,K}$, and define

$$H_{t^\star} = -\sum_{i=1}^{K} \tilde{p}_{t^\star,i} \log \tilde{p}_{t^\star,i}, \qquad \phi_{\text{first}} = 1 - \frac{H_{t^\star}}{\log K}.$$

ϕ_{first} ranges from 0 (uniform top-K) to 1 (all mass on a single token). It is computed from a single greedy forward pass: no additional sampling, no external models.

2.2. Uncertainty baselines

We sample $N = 10$ completions per question using temperature 0.7 and top-$p = 0.95$. **AU-full** measures surface-form agreement by computing the fraction of sampled completions whose normalized full strings match the normalized greedy answer. **AU-3w** and **AU-1w** progressively relax this criterion to the first three words and the first word, providing increasingly strong surface-form baselines. **Semantic AU** performs meaning-level agreement by clustering the greedy answer and its N samples using bidirectional NLI entailment with DeBERTa-v3-large-mnli [4], following the procedure of [2], and reports the fraction of samples assigned to the greedy answer's cluster. **Verbalized confidence** prompts the model to output an integer from 0–100 reflecting its self-estimated correctness [5, 6]. We use the same sampling hyperparameters and scoring rules across all datasets and models. The resulting AUROC values should therefore be interpreted as untuned estimates rather than benchmark-specific optimized results.

2.3. Cost

ϕ_{first} requires one greedy forward pass per question. Semantic AU requires one greedy decode, $N = 10$ sampled generations, and representative-based bidirectional NLI clustering over the greedy and sampled answers. This requires $O(CN)$ NLI comparisons, where C is the number of discovered semantic clusters.

[1] Code, per-question run outputs, and analysis scripts: https://github.com/MinaGabriel/phi-signal.

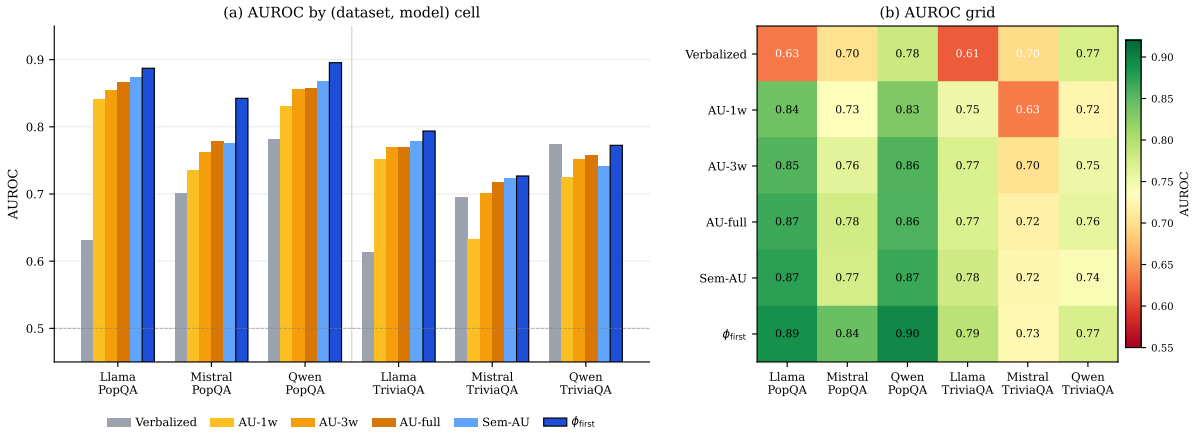

Figure 1: AUROC of correctness prediction across six dataset–model cells and six confidence signals (n=1000 per cell). **(a)** Grouped bars show AUROC per cell; the dashed line marks chance performance. **(b)** The same values are shown as a heatmap. ϕ_{first} achieves the highest AUROC in five of six cells and is within 0.002 AUROC of the strongest method in the remaining cell, while requiring only a single greedy decode.

3. Experiments

3.1. Setup

We evaluate on the test split of PopQA [7] and the validation split of TriviaQA [8], sampling $n = 1000$ examples per dataset with a fixed seed. The same 1000 examples are used across all three models so that all comparisons are paired at the example level. We choose $n = 1000$ as a compute–precision tradeoff. The standard error of an AUROC estimate decreases as $\Theta(1/\sqrt{n})$, so doubling to $n = 2000$ would only narrow each cell's bootstrap interval by about 0.007 AUROC points, while doubling all generation and NLI costs. We instead invest the saved compute in three models, two datasets, and the multi-method comparison, and report empirical 95% bootstrap confidence intervals and paired bootstrap tests for every cell.

We evaluate three instruction-tuned 7–8B models: Llama-3.1-8B-Instruct [9], Mistral-7B-Instruct-v0.3 [10], and Qwen2.5-7B-Instruct [11]. Correctness is determined by an automatic judge (Qwen2.5-14B-Instruct) given the question, the model's answer, and gold aliases.

3.2. Main results

In this subsection, we compare ϕ_{first} with verbalized confidence, surface-form self-consistency, and semantic self-consistency. The main question is whether a single-decode token-level confidence signal can match or exceed uncertainty signals that require multiple sampled generations.

Figure 1 summarizes our main result visually, and Table 1 reports the corresponding numbers. Panel (a) of Figure 1 shows AUROC as grouped bars per dataset–model cell, with ϕ_{first} highlighted; panel (b) presents the same values as a heatmap for at-a-glance comparison across methods. Both views show the same pattern: ϕ_{first} is the strongest method in five of six dataset–model cells and is within 0.002 AUROC of the strongest method in the remaining cell. The pattern is consistent across both datasets: ϕ_{first} improves the per-dataset mean by +0.036 AUROC on PopQA (0.875 vs. 0.839 for semantic AU) and by +0.016 on TriviaQA (0.764 vs. 0.748). The smaller TriviaQA gain suggests that longer and more lexically variable answers give sampling-based methods relatively more opportunity to recover useful agreement information; we return to this point in the limitations.

In the overall mean, ϕ_{first} reaches 0.820 AUROC, compared with 0.793 for semantic AU, 0.791 for AU-full, 0.782 for AU-3w, and 0.752 for AU-1w. Verbalized confidence is weaker, with a mean AUROC of 0.700, consistent with prior work showing that LLMs are often poorly calibrated when asked to

Table 1
AUROC for correctness prediction in closed-book factual QA across three 7–8B instruction-tuned models on PopQA and TriviaQA ($n = 1000$ each). Methods are grouped by inference cost. Best per row in **bold**; second-best underlined. Δ is the gap between ϕ_{first} and the strongest non-ϕ baseline.

		1 decode	$1{+}10$ **decodes (sampling)**				**1 decode**		
Dataset	**Model**	**Verb.**	**AU-1w**	**AU-3w**	**AU-full**	**Sem. AU**	$\boldsymbol{\phi_{\text{first}}}$	**Δ**	**Acc.**
PopQA	Llama-3.1-8B	0.632	0.840	0.854	0.866	0.874	**0.887**	+0.013	0.27
	Mistral-7B-v0.3	0.701	0.735	0.762	0.778	0.775	**0.842**	+0.064	0.25
	Qwen2.5-7B	0.782	0.831	0.856	0.857	0.867	**0.895**	+0.028	0.19
TriviaQA	Llama-3.1-8B	0.614	0.752	0.770	0.769	0.778	**0.794**	+0.016	0.64
	Mistral-7B-v0.3	0.696	0.632	0.701	0.718	0.724	**0.727**	+0.003	0.62
	Qwen2.5-7B	**0.774**	0.725	0.751	0.758	0.741	0.772	−0.002	0.52
PopQA mean		0.705	0.802	0.824	0.834	0.839	**0.875**	+0.036	0.24
TriviaQA mean		0.695	0.703	0.741	0.748	0.748	**0.764**	+0.016	0.59
Overall mean		0.700	0.752	0.782	0.791	0.793	**0.820**	+0.027	0.42

state their own confidence directly [5]. Thus, the advantage of ϕ_{first} over semantic AU is modest in absolute terms (+2.7 AUROC points on average), but it is obtained with a single greedy decode rather than multiple sampled generations and NLI-based semantic clustering.

3.3. Statistical reliability of the gains

The AUROC results show the size of the performance differences, but do not show by themselves whether those differences are stable between evaluation examples. We therefore use paired bootstrap resampling over questions to compare ϕ_{first} against the main baselines within each dataset–model cell. Because both methods are evaluated on the same questions, the test measures whether the observed AUROC gap is robust to resampling of the evaluation set.

Table 2
Paired bootstrap test of $\Delta\text{AUROC} > 0$ for ϕ_{first} vs. each baseline (B=1000 resamples; one-sided p). With $B = 1000$, the smallest resolvable p-value is ≈ 0.001, so cells reported as < 0.001 correspond to $0/1000$ resamples favoring the baseline. **Bold** indicates $p < 0.05$.

Dataset	**Model**	**vs. AU-full**	**vs. Sem. AU**	**vs. AU-1w**
PopQA	Llama	**0.018**	0.082	**<0.001**
	Mistral	**<0.001**	**<0.001**	**<0.001**
	Qwen	**0.001**	**0.008**	**<0.001**
TriviaQA	Llama	**0.025**	0.105	**<0.001**
	Mistral	0.301	0.416	**<0.001**
	Qwen	0.153	**0.014**	**<0.001**
Significant ($p < 0.05$)		4/6	3/6	6/6

Table 2 reports paired bootstrap tests over questions. These tests ask whether the AUROC advantage of ϕ_{first} over each baseline is stable under resampling of the same evaluation examples. The results show that ϕ_{first} significantly outperforms AU-full in four of six cells and semantic AU in three of six cells. The remaining semantic-AU differences are not statistically significant, so we frame ϕ_{first} as *matching* semantic self-consistency rather than uniformly outperforming it. Against AU-1w, the simplest surface-form baseline, the gain is significant in all six cells.

3.4. Subsumption analysis

We test whether ϕ_{first} already captures the information provided by semantic AU. For each cell we report two quantities: the Pearson correlation between ϕ_{first} and semantic AU, and the AUROC gain obtained by combining both signals in a standardized logistic regression over ϕ_{first} alone. A high correlation paired with a near-zero ensemble gain indicates that semantic AU adds little beyond ϕ_{first}.

Table 3
Subsumption analysis. **Pearson** r: correlation between ϕ_{first} and semantic AU. **Gain**: AUROC of the logistic ensemble of both signals minus AUROC of ϕ_{first} alone. Small gains indicate semantic AU adds little beyond ϕ_{first}.

Dataset	Model	Pearson r	Ensemble gain
PopQA	Llama	0.76	+0.017
	Mistral	0.59	+0.009
	Qwen	0.75	+0.012
TriviaQA	Llama	0.74	+0.019
	Mistral	0.54	+0.045
	Qwen	0.67	+0.024
Mean		**0.67**	**+0.021**

Three observations follow from Table 3. First, ϕ_{first} and semantic AU are moderately to strongly correlated, with Pearson r between 0.54 and 0.76 (mean 0.67). Second, combining the two signals improves AUROC by only $+0.021$ on average, and by less than $+0.025$ in five of six cells. Third, ϕ_{first} alone matches or exceeds semantic AU's standalone AUROC in every cell, so the residual ensemble gain reflects a small complementary contribution from semantic AU rather than a deficit in ϕ_{first}. Together, these results indicate that ϕ_{first} already captures most of the discriminative content that semantic agreement extracts at substantially higher inference cost.

3.5. Length confound

A natural concern is that ϕ_{first} may simply track the length of the generated answer. We test this in two stages. First, we compute the raw Pearson correlation r_{len} between ϕ_{first} and the number of generated answer tokens. Second, since wrong answers tend to be both longer and lower-confidence, we control for correctness by computing the partial Pearson correlation $r_{\text{len}}^{\text{partial}}$ between ϕ_{first} and answer length after removing the linear effect of the binary correctness label from both variables.

Table 4
Length confound. r_{len}: raw Pearson correlation between ϕ_{first} and answer length. $r_{\text{len}}^{\text{partial}}$: partial correlation controlling for correctness. Values close to zero after partialling indicate that length is not driving ϕ_{first}.

Dataset	Model	r_{len}	$r_{\text{len}}^{\text{partial}}$
PopQA	Llama	−0.16	−0.02
	Mistral	−0.13	−0.03
	Qwen	−0.14	−0.04
TriviaQA	Llama	−0.23	−0.18
	Mistral	−0.25	−0.17
	Qwen	−0.11	−0.05

Table 4 reports both quantities. The raw correlation ranges from -0.11 to -0.25 across cells, accounting for at most 6.5% of the variance in ϕ_{first}. On PopQA, the partial correlation shrinks substantially: from -0.16 to -0.02 for Llama and from -0.13 to -0.03 for Mistral. This suggests that the apparent length effect on PopQA is largely explained by correctness rather than answer length

itself. On TriviaQA, the partial correlation drops by less: a residual correlation of about -0.18 remains for Llama and Mistral. This indicates a small but non-trivial residual sensitivity to answer length on TriviaQA, which we list as a limitation.

4. Related work

Semantic self-consistency [2, 3] estimates uncertainty from disagreement among NLI-based equivalence classes of sampled generations. Surface-form variants instead measure agreement over normalized strings or answer prefixes. Other single-pass approaches include token probabilities, sequence-level likelihood [12], internal probes [13, 14], and verbalized confidence [5, 6]. Our work directly evaluates first-token entropy as a standalone signal against semantic self-consistency and tests how much semantic-agreement information it captures.

5. Implications for responsible LLM deployment

Because ϕ_{first} requires only one greedy decode, it can be logged or monitored in deployed factual QA systems at low cost. Low-confidence answers can trigger abstention, retrieval, verification, or human review before being shown to users.

The score should not be treated as a guarantee of correctness. Thresholds must be calibrated for the target domain, and systems should report error rates and coverage at different operating points. In high-stakes settings, ϕ_{first} is best used as an inexpensive first-stage signal within a broader assurance pipeline.

6. Discussion and conclusion

First-token confidence matches or modestly exceeds semantic self-consistency on closed-book factual QA across three 7–8B instruction-tuned models, while requiring roughly $1/11$ of the generation cost. The subsumption analysis shows that ϕ_{first} is moderately to strongly correlated with semantic agreement and recovers much of its discriminative content from a single greedy decode. We recommend reporting ϕ_{first} as a cheap baseline before claiming gains from sampling-based uncertainty methods.

Limitations. This study is limited to English closed-book short-answer factual QA with three open 7–8B models and two benchmarks. Results may not transfer to long-form generation, retrieval-augmented QA, multilingual settings, larger or proprietary models, or black-box APIs without token probabilities. We do not evaluate all single-pass baselines, such as sequence likelihood or P(True)-style prompting. We fix $K = 100$ rather than sweeping K, and use Qwen2.5-14B-Instruct as the judge, which may introduce model-family bias for the Qwen generator. Future work should test K sensitivity, additional baselines, non-Qwen judges, and human annotations.

7. Ethical Considerations

First-token confidence is not a complete hallucination detector. It is a low-cost uncertainty signal for closed-book factual QA and can help flag answers that need abstention, retrieval, verification, or human review. High confidence does not guarantee correctness, especially for high-confidence falsehoods.

For responsible deployment, ϕ_{first} should be used as one component of a broader reliability pipeline, not as a standalone safety mechanism. In high-stakes settings, it should be combined with external grounding, domain-specific evaluation, calibrated thresholds, and human oversight.

Declaration on Generative AI

During the preparation of this work, the author(s) used OpenAI ChatGPT for language editing, LaTeX formatting assistance, and revision planning. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: International Conference on Learning Representations (ICLR), 2023.

[2] L. Kuhn, Y. Gal, S. Farquhar, Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, in: International Conference on Learning Representations (ICLR), 2023.

[3] Z. Lin, S. Trivedi, J. Sun, Generating with confidence: Uncertainty quantification for black-box large language models, Transactions on Machine Learning Research (2024).

[4] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, in: International Conference on Learning Representations (ICLR), 2021.

[5] K. Tian, E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, C. D. Manning, Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023.

[6] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, B. Hooi, Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs, arXiv preprint arXiv:2306.13063 (2023).

[7] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, H. Hajishirzi, When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), 2023.

[8] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017.

[9] A. Dubey, et al., The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).

[10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7B, arXiv preprint arXiv:2310.06825 (2023).

[11] A. Yang, et al., Qwen2.5 technical report, arXiv preprint arXiv:2412.15115 (2024).

[12] A. Malinin, M. Gales, Uncertainty estimation in autoregressive structured prediction, in: International Conference on Learning Representations (ICLR), 2021.

[13] S. Kadavath, T. Conerly, A. Askell, et al., Language models (mostly) know what they know, arXiv preprint arXiv:2207.05221 (2022).

[14] A. Azaria, T. Mitchell, The internal state of an LLM knows when it's lying, in: Findings of the Association for Computational Linguistics: EMNLP, 2023.

A. Experimental Details

This appendix summarizes the key implementation details needed to reproduce the experiments. All code, prompts, and run scripts are available at https://github.com/MinaGabriel/phi-signal.

A.1. Hardware and software

All experiments were run on a single NVIDIA A100 40 GB GPU. Models were loaded in `bfloat16`. We used `transformers`, `accelerate`, and `datasets`; all runs used random seed 42.

A.2. Models and checkpoints

Table 5
Model checkpoints used. All models were loaded in `bfloat16`.

Role	HuggingFace path
Generator (Llama)	`meta-llama/Meta-Llama-3.1-8B-Instruct`
Generator (Mistral)	`mistralai/Mistral-7B-Instruct-v0.3`
Generator (Qwen)	`Qwen/Qwen2.5-7B-Instruct`
Judge	`Qwen/Qwen2.5-14B-Instruct`
NLI clustering	`MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli`

A.3. Datasets

We used `akariasai/PopQA` and `mandarjoshi/trivia_qa` with the `rc.nocontext` configuration. For each dataset, we used $n = 1000$ examples in a fixed deterministic order and reused the same indices across all generator models, making all comparisons paired at the example level.

A.4. Generation settings

Table 6
Generation settings used across datasets and generator models.

Parameter	Value
Greedy max new tokens	32
Sample max new tokens	32
Number of samples per question	$N = 10$
Sampling temperature	0.7
Sampling top-p	0.95
Top-K for ϕ_{first}	$K = 100$
Verbalized confidence max new tokens	6
Random seed	42

A.5. Prompts

For all generator models, we used each model's native chat template via `tokenizer.apply_chat_template` with `add_generation_prompt=True`.

Answering prompt.

> **System:** *You are a concise factual QA assistant. Answer in as few words as possible. No explanations.*
> **User:** *Question:* `{question}`
> *Answer:*

Verbalized confidence prompt.

> **System:** *You are calibrating the confidence of an answer. Reply with a single integer from 0 to 100 representing the probability that the answer is correct. No words.*
> **User:** *Question:* `{question}`
> *Answer:* `{model_answer}`
> *Confidence (0–100):*

The first integer in the model's reply, clipped to $[0, 100]$, is divided by 100 and used as the verbalized confidence score.

Judge prompt.

> **System:** *You are a strict factual grading assistant. Decide if the model's answer is factually correct given the question and gold answers. Respond with exactly one token: YES or NO.*
> **User:** *Question:* `{question}`
> *Model answer:* `{model_answer}`
> *Acceptable gold answers:* `{gold_list}`
> *Is the model answer correct? YES or NO.*

We greedy-decode up to 4 new tokens and set `is_correct` $= 1$ iff the decoded text begins with YES, case-insensitive.

A.6. First-token score

The first content-bearing answer token $t^\star$ is found by scanning the greedy decoded tokens and skipping leading punctuation, whitespace, and literal answer prefixes such as `Answer`. If no such token is found, we use the first decoded token. We then compute $\phi_{\text{first}} = 1 - H_{t^\star} / \log K$ using the renormalized top-K distribution at $t^\star$, with $K = 100$.

A.7. Agreement baselines

For surface-form agreement, sampled and greedy answers are normalized by lower-casing, removing articles, removing punctuation, and collapsing whitespace. **AU-full** compares the full normalized answer, **AU-3w** compares the first three words, and **AU-1w** compares the first word. Each score is the fraction of the $N = 10$ samples matching the greedy answer under that key.

For semantic agreement, we cluster the greedy answer and its samples using bidirectional NLI entailment following [2]. Semantic AU is the fraction of samples assigned to the greedy answer's semantic cluster.

A.8. Statistical procedures

For each model–dataset cell, we compute AUROC over the same $n = 1000$ questions. Confidence intervals and paired comparisons are estimated with $B = 1000$ bootstrap resamples over question indices.

For the subsumption experiment, we fit a standardized logistic regression on $(\phi_{\text{first}}, \text{Sem. AU})$ within each cell and compare its AUROC to ϕ_{first} alone. For the length analysis, we report the partial Pearson correlation between ϕ_{first} and answer length after residualizing both variables against correctness.