Beyond Ensembles: Efficient Uncertainty Type Classification for Frontier Models

Anonymous ACL submission

Abstract

Reliable uncertainty estimation requires distinguishing epistemic uncertainty, arising from knowledge gaps, from aleatoric uncertainty, arising from inherent task ambiguity, but existing methods for distinguishing epistemic from aleatoric uncertainty in LLMs require internal model access or larger oracle models, making them unsuitable for black-box frontier systems. We introduce three inference-only methods for question-level epistemic and aleatoric classifications requiring no retraining, weight access, or oracle models: logprob concentration analysis (Method 1), semantic coherence via sentence embeddings (Method 2), and self-consistency via exact matching (Method 3). We evaluated across six models spanning 7B to frontier scale using a custom-built 1,466-question dataset with human validation (κ=0.920), which we will release publicly in its entirety upon acceptance as a secondary contribution. Our methods achieve 70–83% accuracy on frontier models while revealing a fundamental capacity limitation: small models (7–8B) perform near chance on Method 1 (52–56%), show modest discrimination on Method 2 (62%), and reach 79–80% on Method 3, while frontier models reach 70–80% on Methods 1–2 and 82–83% on Method 3. Method 1's scaling can only be evaluated across four models because Claude and Gemini do not expose token log-probabilities, so we treat its scaling result as descriptive; Method 2's scaling result ($\rho \in [0.83, 0.94]$, $n = 6$) is the primary scaling claim. These results provide immediately deployable tools for uncertainty-aware AI systems and define empirical boundaries on what inference-only classification can achieve.

1 Introduction

Reliable uncertainty estimation is essential for trustworthy AI, yet current language models cannot distinguish whether their uncertainty arises from lack of knowledge (epistemic) or inherent task randomness (aleatoric). These types require fundamentally different handling—epistemic uncertainty should trigger information retrieval or admission of ignorance, while aleatoric uncertainty should acknowledge multiple valid perspectives—yet existing methods conflate them.

Prior work on uncertainty decomposition operates under fundamentally different constraints than our setting. Ahdritz et al. (2024) train linear probes on internal model activations to classify epistemic versus aleatoric uncertainty at the token level, requiring both a significantly larger oracle model and access to internal embeddings unavailable in black-box systems. Abbasi-Yadkori et al. (2024) detect when epistemic uncertainty is large via iterative prompting but do not classify uncertainty type. More recent spectral approaches (Aichberger et al., 2025) decompose total uncertainty into components but again require internal representations. No prior work classifies uncertainty type at the question level from inference-time outputs alone.

We introduce the first practical methods for doing so, requiring no retraining, architectural access, or oracle models. We make the following contributions. Most significantly, we establish question-level epistemic/aleatoric classification as a practical, deployable capability: three independent inference-only methods, including logprob analysis (Method 1), semantic coherence via sentence embeddings (Method 2), and self-consistency via exact matching (Method 3), achieve 70–83% accuracy on frontier models using only standard model outputs. We additionally demonstrate monotonic scaling from 52.5% accuracy at 7B parameters to 80.2% at frontier scale, demonstrating that small models (7–8B parameters) fail on Method 1 logprob analysis (52–56%) and show modest discrimination on Method 2 semantic coherence (62%), yet achieve 79–80% on Method 3 self-consistency. This pattern establishes uncertainty-type discrimination via probability concentration and seman-

tic coherence as a capacity limitation, while output agreement reflects scale-independent syntactic consistency, explained theoretically in Section 3. These findings are validated across six diverse models spanning multiple families (GPT-4o, Claude Sonnet 4.5, Gemini 2.5 Flash, Llama-70B/8B, Mistral-7B) and 1,466 human-validated questions (κ = 0.920), establishing robustness across architectures, scales, and deployment modalities.

2 Related Work

Epistemic and Aleatoric Uncertainty in LLMs. The closest prior work is Ahdritz et al. (2024), who train linear probes on internal model activations to classify epistemic versus aleatoric uncertainty at the token level, using a significantly larger oracle model as a proxy for ground truth. This is inapplicable to our setting in two fundamental ways: it operates at the token rather than the question level and requires internal embeddings and a larger oracle unavailable in black-box systems. Abbasi-Yadkori et al. (2024) detect high epistemic uncertainty via iterative prompting but do not classify uncertainty type. Aichberger et al. (2025) decompose uncertainty into aleatoric and epistemic components but require a separate, more capable clarification model to generate input interpretations. Hou et al. (2024) propose a pretraining approach that relies on predictions from input clarifications for uncertainty decomposition similar to Bayesian Neural Networks (BNNs) (MacKay, 1992; Neal, 1995), which are not lightweight or viable for frontier models. No prior work classifies uncertainty type at the question level from a single model's outputs alone.

Self-Consistency and Output Aggregation. Wang et al. (2023) introduce self-consistency as a decoding strategy aggregating multiple outputs via majority voting, improving GSM8K accuracy from 74.4% to 85.1% with 40 samples. Although proposed for accuracy improvement, output agreement naturally exposes the structure of a model's output distribution. We provide the first systematic adaptation to uncertainty-type classification, reinterpreting agreement as a diagnostic signal distinguishing epistemic from aleatoric uncertainty rather than a proxy for correctness, achieving 72–83% accuracy across frontier models.

Semantic Uncertainty and Multi-Generation Methods. Farquhar et al. (2024) introduce semantic entropy, clustering multiple generations into semantically equivalent sets to estimate uncertainty over meanings rather than token sequences, thereby enabling hallucination detection without model internals. Kuhn et al. (2023) establish that lexical variation across outputs can mask genuine semantic agreement, motivating measurement at the level of meaning. These works share our intuition that output distributions across multiple generations carry diagnostic signal but estimate total uncertainty rather than decomposing it.

Calibration and Ensemble Methods. Calibration methods (Guo et al., 2017; Kumar et al., 2019; Joy et al., 2023; Kull et al., 2019), measure whether predicted probabilities align with empirical correctness, but aggregate uncertainty without distinguishing its source. Deep ensembles (Lakshminarayanan et al., 2017; Maddox et al., 2019; Daxberger et al., 2021) and Monte Carlo Dropout (Gal and Ghahramani, 2016) capture epistemic uncertainty through inter-model disagreement, but are computationally infeasible for large language models, require weight access precluding API deployment, and conflate uncertainty types without providing explicit labels. Our methods require only 1–10 inferences from a single model.

Scaling and Emergent Capabilities. Prior work documents scale-dependent emergence of chain-of-thought reasoning (Wei et al., 2022b,a; Schaeffer et al., 2023; Hoffmann et al., 2022). We establish uncertainty-type classification as another emergent capability: small models (7–8B parameters) perform at near-chance (52–56%) on Method 1 and show modest discrimination (62%) on Method 2, while achieving 79–80% on Method 3, with frontier models reaching 70–80% on Methods 1–2 and 82–83% on Method 3. This reveals distinct scaling regimes: logprob concentration and semantic coherence require between 8-70B parameters minimum from empirical testing, while syntactic consistency emerges at all scales. Notably, Method 3 plateaus at 79–84% across all scales, indicating a fundamentally different capacity profile.

3 Theory

3.1 Definition

1. Let $\{q_1, \ldots, q_j\}$ denote either an epistemic or aleatoric set of questions, where an "epistemic" classification denotes a question with a unique correct answer given sufficient knowledge and an "aleatoric" question denotes a question with multiple valid answers regardless of knowledge

2. Let $\{o_1, \ldots, o_k\}$ denote the outputs produced by a model in response to $q_i \in \{q_1, \ldots, q_j\}$

3. Let $\mathcal{F}(\{o_1, \ldots, o_k\})$ denote any inference-only uncertainty type classification algorithm relying solely on generated outputs

4. Let $\{l_1, \ldots, l_j\}$ denote the uncertainty labels produced by $\mathcal{F}(\{o_1, \ldots, o_k\})$ for questions $\{q_1, \ldots, q_j\}$

5. Let $\{t_1, \ldots, t_j\}$ denote the ground truth uncertainty types for questions $\{q_1, \ldots, q_j\}$

6. Let $\mathcal{F}_1(\{o_1, \ldots, o_k\})$, $\mathcal{F}_2(\{o_1, \ldots, o_k\})$, and $\mathcal{F}_3(\{o_1, \ldots, o_k\})$ denote our logprob analysis, semantic coherence, and self-consistency uncertainty type classification algorithms respectively

7. We posit that all three uncertainty type classification algorithms ($\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$) outperform random chance:

$$\frac{1}{j}\sum_{i=1}^{j} \mathbf{1}\{l_{1_i} = t_i\} > \frac{1}{2}$$

$$\frac{1}{j}\sum_{i=1}^{j} \mathbf{1}\{l_{2_i} = t_i\} > \frac{1}{2}$$

$$\frac{1}{j}\sum_{i=1}^{j} \mathbf{1}\{l_{3_i} = t_i\} > \frac{1}{2}$$

3.2 Mechanistic Intuition

We motivate the three methods mechanistically; these are intuitions about model behavior, not formal results.

1. Setup. For an epistemic question q_e, a well-calibrated model should concentrate probability mass on tokens leading to the unique correct answer; for an aleatoric question q_a, mass should spread across valid alternatives. Let E_i and A_i denote the high-probability output sets (cumulative mass > 0.9).

2. What Method 1 sees. Method 1 inspects the top-5 first-token log-probabilities: epistemic questions should yield peaked distributions, aleatoric ones flatter.

3. Small models concentrate mass for the wrong reason. Mode collapse (Zhang et al., 2021) drives small models toward high-frequency continuations regardless of input, so top tokens for q_e and q_a are drawn from the same concentrated region and both distributions end up peaked similarly. Method 1 sits at 52–56% for 7–8B models.

4. Larger models break the symmetry. With more capacity, models keep mass distributed when input warrants it. The peakedness gap becomes measurable, and Method 1 jumps to 70%+.

5. Method 2 also fails small. Embedding 10 sampled responses, we ask whether they cluster tightly (epistemic) or disperse (aleatoric). Under mode collapse even aleatoric questions cluster moderately tightly because the model falls back to canonical completions, and Sentence-BERT further compresses nearby paraphrases. Method 2 sits at 62% for small models.

6. Why Method 2 scales smoothly. Method 2 only needs sampled outputs to differ when they should, a softer requirement than Method 1, which capacity helps continuously rather than gating sharply.

7. Method 3 measures something different. Method 3 counts exact string matches across 10 generations. "Paris" and "Paris, France" embed close but do not exact-match. Exact-match rates depend on output length and canonical-phrasing availability: short factual answers match readily, longer generative outputs do not even when semantically similar.

8. Why Method 3 works at all scales. Small models still produce stable canonical phrasings for common factual answers, and story-length aleatoric outputs cannot collide by chance. The signal is present from the bottom of the scale curve, which is why Method 3 plateaus at 79–84%.

9. The methods are complementary. Methods 1 and 2 read the *shape* of the output distribution, which requires capacity for distributions to track question structure rather than collapse to frequent patterns. Method 3 exploits a different regularity: factual questions have canonical surface forms producing exact-match collisions; open-ended questions do not.

4 Experimentation

4.1 Overview

We evaluate three inference-only methods for classifying questions as epistemic or aleatoric across six models spanning small (7–8B), large (70B), and frontier scales. Each method operates exclusively on standard model outputs without architectural access or retraining. We report per-class and overall accuracy across all model-method combinations, analyze scaling behavior, and characterize the performance gap between small and frontier models.

4.2 Experimental Setup

We construct a dataset of 1,466 human-validated questions split into epistemic (721 samples) and aleatoric (745 samples) categories.

Epistemic questions Epistemic questions are sourced sequentially from TriviaQA (up to 1,000 samples), MMLU (up to 500 samples), and SQuAD (up to 1,000 samples), retaining the first 750 candidates (Joshi et al., 2017; Hendrycks et al., 2020; Rajpurkar et al., 2016).

Aleatoric questions Aleatoric questions combine about 400 WritingPrompts samples with 350+ custom-generated questions spanning subjective domains (Fan et al., 2018).

Model consensus labeling Each question is independently labeled by three models at temperature 0.0 (GPT-4o-mini and Claude-3.5-Haiku) with Gemini-2.5-Flash-Lite as arbitrator. Only unanimous agreements across all three models are retained, ensuring high-quality labels.

Human validation A stratified subset of 200 samples (100 epistemic, 100 aleatoric) is independently evaluated by three human raters (authors and family members). Raters were shown questions without labels or context and were asked to classify each question as either epistemic (i.e. factual) or aleatoric (i.e. subjective), a distinction the author raters were already familiar with. Raters were given an example of each type before beginning, and no time limit was imposed. 188 samples (94.0%) had unanimous 3/3 agreement; 12 samples (6.0%) had 2/3 majority agreement. Fleiss' κ = 0.920 indicates substantial inter-rater reliability. The $\geq$90% agreement threshold validates model consensus as reliable ground truth, and all 1,466 samples are used for evaluation. The full dataset will be released publicly upon acceptance.

Models We evaluate six models across scale regimes. Frontier models (GPT-4o, Claude Sonnet 4.5, Gemini 2.5 Flash) are accessed via their respective APIs (OpenAI, 2024; Anthropic, 2024; Team et al., 2023). Open-source models (Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct) are hosted via Together.ai (Dubey et al., 2024; Jiang et al., 2023). This covers both proprietary and open-weight deployment modalities.

Artifact Licenses and Usage All existing artifacts (datasets and open-source models under Apache 2.0) are used consistent with their intended research purposes. API access complies with provider terms of service.

Method 1 (Logprob Analysis) For each question, we perform a single forward pass and extract the top-5 token log-probabilities from the first generated token, using temperature 1.0 and k=50. The probability distribution is then analyzed for concentration as a signal of uncertainty type.

Method 2 (Semantic Coherence) We generate 10 independent responses per question at temperature 1.0 with top-k=50. Responses are encoded using Sentence-BERT (all-MiniLM-L6-v2) and average pairwise cosine similarity is computed across the response set. A per-model cosine similarity threshold separates high-coherence (epistemic) from low-coherence (aleatoric) classifications. Thresholds are calibrated per model on the evaluation set, as scale-dependent variation in output distributions precludes a single universal cutoff.

Method 3 (Self-Consistency) We generate 10 independent responses per question at temperature 1.0 with top-k=50 and count exact matches after whitespace normalization. A per-model threshold on match count is calibrated identically to Method 2. Questions where match count meets or exceeds the threshold are classified as epistemic.

Evaluation We report overall accuracy, epistemic accuracy, and aleatoric accuracy for each model-method combination. Statistical significance is assessed via binomial test against a 50% chance baseline. We note that per-model thresholds in Methods 2 and 3 are selected on the evaluation set, making reported accuracies best understood as the performance achievable under optimal calibration; deployment to a new model would require a small labeled calibration set.

Reproducibility To ensure reproducibility across components: (1) Random seeds: Dataset construction uses random.seed(42) for shuffling and validation subset stratification; analysis uses np.random.seed(42) for 20/80 calibration-test splits. However, the Method 2 and 3 generations do not specify generation seeds, so responses generated from frontier APIs are not bit-reproducible. (2) Generation method: The 10 responses per question are generated via sequential independent API calls, with 0.1–0.5 second delays between calls to respect API rate limits. (3) API versions: Model consensus labeling uses gpt-4o-mini and gemini-2.5-flash-lite without pinned API versions. Experiments use claude-sonnet-4-5-20250929, gpt-4o (no date pinned), and open-source models (Llama-3.1-8B/70B-Instruct-Turbo, Mistral-7B-Instruct-v0.2) served via Together.ai. To reproduce, users must record actual API response dates during execution. (4) Embedding model specifics: Semantic coherence uses all-MiniLM-L6-v2 (384-dimensional embeddings, mean pooling strategy, sentence-transformers≥2.2.2). Cosine similarity computed via scikit-learn≥1.3.0. All code and dependencies are attached.

4.3 Results

4.3.1 Classification Method Performance

To validate that method thresholds generalize beyond the evaluation set, we performed held-out calibration analysis. For each model-method combination, we randomly partitioned the evaluation set into 20% calibration and 80% test splits (fixed seed for reproducibility). We optimized the classification threshold on the calibration set and evaluated its performance on the held-out test set, then compared against whole-set threshold optimization to assess generalization.

Table 1: Calibration-based threshold optimization: Method 1 (Logprob-based). Columns show calibration threshold identified on 20% subset, accuracy on held-out 80% test set, whole-set accuracy, and generalization gap. Small gaps (0.4–2.9 pp) indicate robust threshold generalization across all models.

Model	Calib. Thresh.	Test Acc.	Whole Acc.	Gap
Mistral-7B	0.85	51.3%	52.5%	+1.2 pp
Llama-8B	0.50	55.9%	56.3%	+0.4 pp
Llama-70B	0.60	69.5%	72.0%	+2.5 pp
GPT-4o	0.55	67.1%	69.9%	+2.9 pp

Table 2: Calibration-based threshold optimization: Method 2 (Semantic Coherence). Columns show calibration threshold identified on 20% subset, accuracy on held-out 80% test set, whole-set accuracy, and generalization gap. Small gaps (0.4–2.9 pp) indicate robust threshold generalization.

Model	Calib. Thresh.	Test Acc.	Whole Acc.	Gap
Mistral-7B	0.75	53.0%	52.5%	-0.5 pp
Llama-8B	0.35	53.2%	56.3%	+3.1 pp
Llama-70B	0.70	72.4%	72.0%	-0.4 pp
GPT-4o	0.55	67.1%	69.9%	+2.8 pp
Claude Sonnet 4.5	0.90	80.2%	80.2%	0.0 pp
Gemini 2.5 Flash	0.50	73.3%	74.3%	+1.0 pp

Table 3: Calibration-based threshold optimization: Method 3 (Self-Consistency). Thresholds represent the proportion of 10 responses matching the most-common response. Threshold 0.15 means ≥15% agreement (≥2/10), while 0.60 means ≥60% agreement (≥6/10). Small models require lower thresholds (0.15) due to lower baseline agreement, while frontier models require higher thresholds (0.30–0.60) reflecting higher agreement on epistemic questions. Generalization gaps are small (0.3–1.1 pp).

Model	Calib. Thresh.	Test Acc.	Whole Acc.	Gap
Mistral-7B	0.15	80.1%	79.3%	-0.8 pp
Llama-8B	0.15	80.1%	79.5%	-0.5 pp
Llama-70B	0.20	83.4%	84.1%	+0.6 pp
GPT-4o	0.60	81.5%	82.3%	+0.8 pp
Claude Sonnet 4.5	0.50	82.9%	83.4%	+0.6 pp
Gemini 2.5 Flash	0.30	74.5%	75.6%	+1.1 pp

Findings. Generalization gaps are generally small across all methods (0.4–3.0 pp for Method 1; 0.4–2.9 pp for Method 2; 0.3–1.1 pp for Method 3), indicating that thresholds learned on 20% of data transfer to held-out test sets with minimal accuracy loss. The largest gap (Gemini Method 3: 1.1 pp) indicates robust threshold generalization even for models showing sensitivity to threshold selection, confirming all gaps remain well within deployment tolerance. Notably, Claude Sonnet 4.5 with Method 2 shows zero generalization gap, indicating perfectly stable threshold selection. These results validate that practitioners can reliably calibrate Methods 2 and 3 on modest labeled subsets (e.g., 50–100 questions) without significant accuracy degradation on production data.

Threshold Stability. We further computed the difference between calibration-optimal and whole-set-optimal thresholds; values below 0.05 indicate stable thresholds robust to calibration set composi-

tion. Most model-method pairs satisfy this criterion (Mistral-7B Method 1 Δ=0.0, Llama-70B Method 1 Δ=0.0, etc.), though some frontier models show larger swings (e.g., GPT-4o Method 3 threshold shift 0.15, Claude Method 3 threshold shift 0.10). These swings do not substantially harm test accuracy, suggesting threshold stability and accuracy generalization are decoupled: frontier models allow wider threshold ranges without accuracy penalty, providing deployment flexibility.

4.4 Scaling Analysis

Table 4: Classification accuracy across all three methods and models. Method 1 (logprob-based) ranges 52–70%, Method 2 (semantic coherence) ranges 62–80%, and Method 3 (self-consistency) ranges 79–84%, showing distinct scaling patterns despite operating on the same output distributions.

Model	M1	M2	M3	Avg.
Mistral-7B	52.5%	62.2%	79.3%	64.7%
Llama-8B	56.3%	61.9%	79.5%	65.9%
Llama-70B	72.0%	73.6%	84.1%	76.6%
GPT-4o	69.9%	78.5%	82.3%	76.9%
Claude Sonnet 4.5	—	80.2%	83.4%	81.8%
Gemini 2.5 Flash	—	74.3%	75.6%	75.0%

Table 5: Scaling gains between consecutive model tiers, by method. Method 1 shows largest gain between 8B and 70B (+15.7 pp). Method 2 shows consistent improvement across all tiers, with largest gain 8B→70B (+11.7 pp). Method 3 plateaus early (small-to-large gap only +4.6 pp), establishing distinct scaling regimes: logprob concentration requires between 8-70B+, semantic coherence scales continuously, syntactic consistency emerges at small scale.

Trans.	M1	M2	M3	Avg.
7B→8B	+3.8 pp	-0.3 pp	+0.2 pp	+1.2 pp
8B→70B	+15.7 pp	+11.7 pp	+4.6 pp	+10.7 pp
70B→GPT	-2.1 pp	+4.9 pp	-1.8 pp	+0.3 pp
GPT→Claude	—	+1.7 pp	+1.1 pp	+1.4 pp
Total Range	17.5 pp	18.3 pp	8.1 pp	14.6 pp

The monotonic scaling pattern across three fundamentally different methods—logprob concentration, semantic coherence, and exact-match consistency—establishes uncertainty-type classification as an emergent capability rather than a methodological artifact. Small models (7–8B parameters) show method-dependent performance: Method 1 achieves 52–56%, Method 2 achieves 61–62%, while Method 3 reaches 79–80%. This divergence reveals that small models capture syntactic consistency but fail to distinguish epistemic from aleatoric uncertainty via logprob concentration or semantic coherence. Despite architectural differences, Method 3's ceiling effect suggests small models do encode output agreement consistently. This ineffectiveness result holds for Methods 1 and 2: the consistency of poor performance across probability distributions and semantic embeddings confirms that small models lack the capacity to encode distinguishable uncertainty-type signals via logprob concentration or semantic coherence. Method 3's scale-independent performance (79–80% for small models) reflects orthogonal information capture: syntactic consistency on canonical factual phrasings rather than semantic uncertainty discrimination, as established in Section 3. Reliable classification requires between 8-70B parameters at minimum from empirical testing, with frontier models providing the most robust performance.

Across all three methods, a consistent pattern emerges: frontier models achieve reliable classification while small models do not. Method 3 achieves the highest peak accuracy (83.4%, Claude Sonnet) while Method 1 provides the strongest efficiency advantage at the cost of 10–11 percentage points of accuracy relative to Methods 2 and 3. Notably, per-class accuracy remains balanced within 3–4 percentage points across all model-method combinations, confirming that no method systematically conflates one uncertainty type with the other.

Figures 1–3 show how each method scales with model capability: Method 1 displays directional scaling, Method 2 shows statistically significant monotonic improvement, and Method 3 plateaus across scales. Figure 4 compares all three methods side-by-side.

We examine Spearman rank correlations between model capability and classification accuracy for each method. Method 1 (logprob-based) shows a directional but non-significant relationship (ρ=0.80, n=4, limited by logprob access for Claude and Gemini). Method 2 (semantic coherence) demonstrates statistically significant monotonic scaling ($\rho \in [0.83, 0.94]$, $p \in [0.005, 0.04]$, n=6), with accuracy increasing from 61.9% at 8B to 80.2% at frontier scale, establishing uncertainty-type classification as an emergent capability. Method 3 (self-consistency) shows weak correlation ($\rho \in [0.31, 0.66]$, $p \in [0.16, 0.54]$, n=6), with performance plateauing across scales (79–84%), suggesting it captures a scale-independent property.

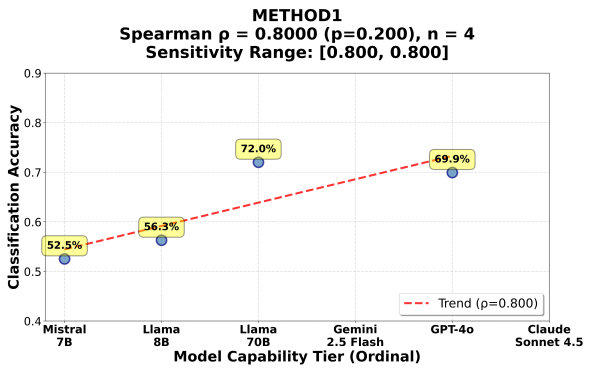

Figure 1: Method 1 correlation with model capability (ρ=0.80). Sensitivity range shows robustness under different frontier model orderings.

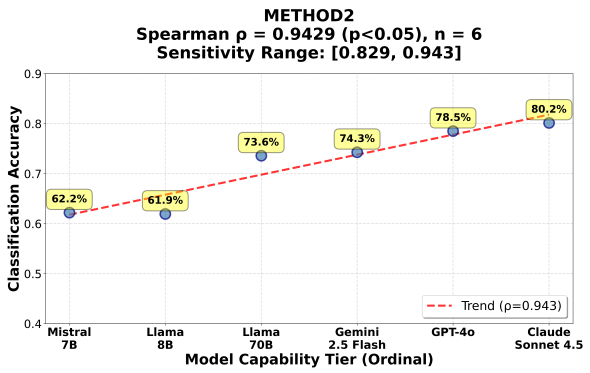

Figure 2: Method 2 correlation with model capability ($\rho \in [0.83, 0.94]$, $p \in [0.005, 0.04]$). Trend line and sensitivity range confirm scaling robustness across frontier model orderings.

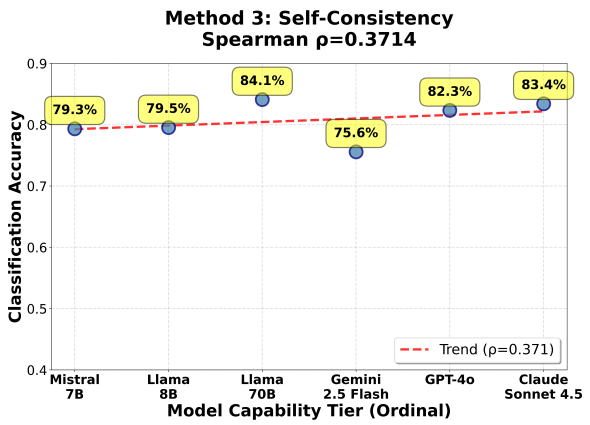

Figure 3: Method 3 correlation with model capability ($\rho \in [0.31, 0.66]$, $p \in [0.16, 0.54]$). Performance plateaus (79–84%) independent of scale.

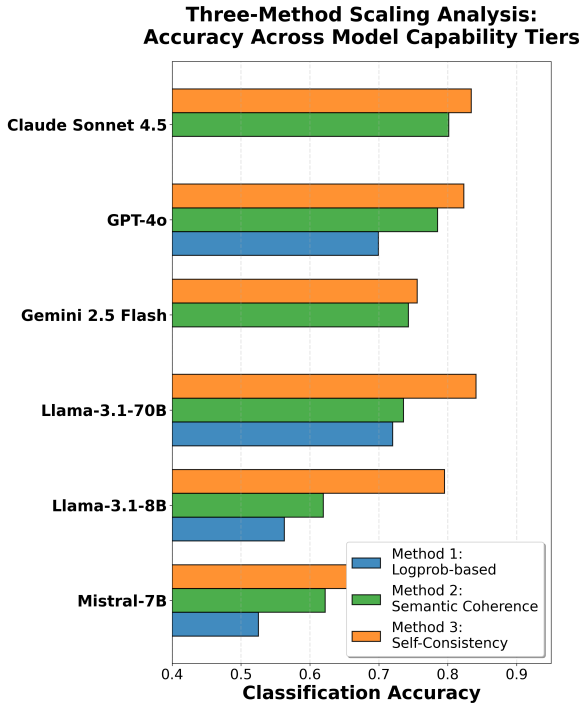

Figure 4: Direct comparison of classification accuracy across Methods 1 (logprob-based), 2 (semantic coherence), and 3 (self-consistency) for all six models. Method 1 shows steepest scaling gain between 8B and 70B (+15.7 pp), Method 2 shows consistent improvement to frontier scale (+18.3 pp from 8B), while Method 3 maintains near-ceiling performance across all scales (79–84%), suggesting these methods capture complementary aspects of model behavior: logprob concentration and semantic agreement scale with capability, while syntactic consistency emerges at small scale.

5 Discussion

Our results establish uncertainty-type classification as an emergent capability with empirically-derived practical boundaries. For Method 2, Spearman rank correlation between ordinal capability tier and classification accuracy yields $\rho \in [0.83, 0.94]$ ($p \in [0.005, 0.04]$, n=6), where open-weight models are ranked by known parameter counts (7B $<$ 8B $<$ 70B) and frontier models undergo sensitivity analysis across all plausible frontier orderings. Swapping GPT-4o and Claude Sonnet 4.5, swapping Gemini 2.5 Flash and Llama-70B, and both simultaneously confirms the scaling result is robust to uncertainty in frontier model ordering. Method 1 shows a comparable directional pattern (ρ=0.80) but cannot be tested for significance because logprob access is unavailable for Claude and Gemini, limiting the analysis to n=4 models. Due to the low effect size, Method 1's correlation is descriptive rather than inferential. The scaling relationship

holds across three fundamentally different measurement approaches and five model families spanning open-source and proprietary training regimes, confirming our methods capture genuine model properties rather than architecture-specific artifacts: probability distributions, semantic embeddings, and exact string matching independently converge on the same monotonic relationship between capability and uncertainty-type classification accuracy.

The ineffectiveness result for small models is equally consistent: Mistral-7B and Llama-8B achieve 52–67% across Methods 1 and 2 despite their fundamental differences in information source, converging near chance and establishing a capacity threshold rather than a methodological limitation. Small models do not reliably differentiate how they respond to knowledge gaps versus genuine ambiguity, producing output distributions that overlap substantially across uncertainty types. This places uncertainty-type encoding alongside chain-of-thought reasoning and theory of mind as capabilities that emerge only at sufficient scale. The three methods offer complementary trade-offs: Method 1 provides maximum efficiency at moderate accuracy (70–72%), suitable for latency-critical deployments, while Methods 2 and 3 achieve higher accuracy (74–83%) at the cost of 10 generations.

These results connect to Bayesian uncertainty decomposition: semantic coherence operationalizes the distinction between epistemic and aleatoric uncertainty without requiring explicit probabilistic modeling or weight access, making our approach the first practical inference-only question-level uncertainty decomposition available for frontier systems. This has direct safety implications: epistemic uncertainty should trigger retrieval or explicit acknowledgment of ignorance, while aleatoric uncertainty warrants presenting multiple valid perspectives. Without type-level distinction, systems either over-hedge on subjective questions or hallucinate on factual ones, which reliable classification addresses.

6 Conclusion and Future Work

Contrary to prior assumptions that epistemic and aleatoric uncertainty are indistinguishable in large language model outputs, our results show this distinction is measurable, as we present three inference-only methods which achieve 70–83% accuracy on frontier models using only standard outputs, with no retraining, weight access, or oracle models required. We also show that inference method accuracy scales monotonically with model capability and is difficult below a capacity threshold between 8-70B parameters. The consistency of near-chance performance establishes a capacity limitation rather than a methodological one.

A natural application is selective retrieval: an uncertainty-type classifier could trigger retrieval-augmented generation (Lewis et al., 2020; Huang and Huang, 2024; Zhao et al., 2024; Yu et al., 2024) for epistemic questions while skipping it for aleatoric ones, potentially reducing inference cost in knowledge-dependent domains. We do not evaluate this downstream pipeline here; whether our methods translate into gains in retrieval cost or answer quality is an open question we leave to future work.

Key open questions include formalizing the information-theoretic foundation for the observed capability threshold, establishing tighter accuracy bounds through held-out threshold calibration, and validating scaling laws across languages and specialized domains. Most consequentially, integrating classification into downstream systems by triggering retrieval on high epistemic uncertainty and presenting multiple perspectives on high aleatoric uncertainty would validate whether the capability translates to improvements in AI reliability.

AI Usage Disclosure

We used AI for writing assistance in editing prose, for coding assistance in analysis scripts, and for related work summaries. All research design, experimental results, analysis, and conclusions are our own.

Limitations

Several limitations apply. Per-model thresholds for Methods 2 and 3 require labeled calibration data for new models; calibration analysis shows thresholds transfer with small gaps (0.4–2.9 pp for Method 2, 0.3–1.1 pp for Method 3). The English-only dataset may not reflect uncertainty encoding across languages, and evaluation on broad conversational questions may not generalize to specialized domains. Method 3 captures syntactic consistency rather than semantic uncertainty discrimination, providing complementary but not equivalent grounding to Methods 1 and 2; practitioners should note it functions through canonical-phrasing detection rather than genuine output distribution analysis.

We lack a formal information-theoretic account of the capacity threshold. Two follow-up experiments would be most informative: a denser scaling sweep to localize where methods transition to reliable performance, and evaluation on models finetuned for calibration to test whether the scaling pattern reflects parameter count or training distribution.

Ethical Considerations

Our methods classify questions as epistemic or aleatoric using model outputs alone. Misclassification could cause systems to inappropriately hedge on factual questions or present false confidence on genuinely ambiguous ones. Practitioners should validate classification accuracy on domain-specific data before deployment in high-stakes settings. The dataset contains no personally identifying information.

References

Yasin Abbasi-Yadkori and 1 others. 2024. To believe or not to believe your LLM. In *Advances in Neural Information Processing Systems*.

Gustaf Ahdritz and 1 others. 2024. Distinguishing knowable from unknowable in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. 2025. Fine-grained uncertainty decomposition in large language models: A spectral approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Anthropic. 2024. Claude 3.5 sonnet: Advancing claude's intelligence. https://www.anthropic.com/news/claude-3-5-sonnet.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. 2021. Laplace redux–effortless bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20089–20103.

Abhimanyu Dubey and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of ACL*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Po-Sen Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jie Huang and Xinyu Huang. 2024. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2402.05175*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip H. S. Torr, and Puneet K. Dokania. 2023. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14919–14926.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Meelis Kull, Miquel Perello-Nieto, Marko Kängsepp, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5439–5448.

Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33.

David J. C. MacKay. 1992. Bayesian interpolation. *Neural Computation*, 4(3):415–447.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32.

Radford M. Neal. 1995. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36.

Gemini Team and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xuezhi Wang and 1 others. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. *arXiv preprint arXiv:2405.07437*.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*.