Epistemic Bundles: Diagnosing Reliability in LLM Ensembles

Adama Diallo
Independent AI Scientist
adamadatasci@hotmail.com

Abstract

Ensemble methods for Large Language Models (LLMs) are commonly framed as selection problems: given multiple candidate outputs, the system attempts to identify the most likely correct one via voting, reranking, or verification. These approaches implicitly rely on the *Selection Hypothesis*: the assumption that the correct answer is present, in its entirety, among the candidates. In reasoning-heavy tasks, this assumption frequently fails—truth may be fragmented across outputs, partially correct, or entangled with hallucinated steps.

We introduce a framework for **Epistemic Integration**, which characterizes ensemble behavior across five distinct epistemic states: *Redundant Correctness, Complementary Fragmentation, Reasoning–Result Divergence, Contradictory Hallucination,* and *False Consensus*. These states describe how truth and error are distributed across model generations, exposing structural failure modes of selection-based ensembling that persist regardless of model scale.

To operationalize this framework, we propose **Epistemic Bundles**, a general methodology for constructing diagnostic ensemble benchmarks by algorithmically perturbing reference solutions to control truth coverage and error contamination. We present **Epistemic-GSM**, a concrete instantiation derived from GSM8K, consisting of 8,500 epistemic bundles spanning all five states. We argue that true reliability requires systems that reason over partial, noisy, and contradictory outputs to reconstruct truth when possible, and crucially, to recognize epistemic insufficiency when abstention is warranted.

1 Introduction

Large Language Models (LLMs) exhibit inherent stochasticity in their reasoning paths, intermediate computations, and final answers—even when subjected to identical prompts and parameters. While such variability can theoretically increase solution diversity, it undermines reliability in multi-step reasoning tasks such as mathematical problem solving, symbolic logic, and multi-hop question answering.

To mitigate this instability, prior work has explored ensemble techniques including self-consistency sampling [1], verifier-based reranking [2], and fusion methods [3]. Despite architectural differences, these approaches largely operate under a shared assumption: that the correct answer exists largely intact within the generated candidate set and can be isolated through discriminative selection. We refer to this as the **Selection Hypothesis**.

In practice, the Selection Hypothesis is frequently violated. Different models—or different samples from the same model—often err at distinct reasoning steps. In such regimes, no single output is strictly "correct," yet the aggregate set of outputs contains sufficient information to reconstruct the solution. This reflects a state of distributed epistemic signal rather than pure noise.

We argue that advancing ensemble reliability requires shifting the paradigm from discriminative selection to *epistemic integration*: designing systems that synthesize truth from partial fragments and diagnose when the available evidence is insufficient.

Our contributions are:

- A taxonomy of epistemic states that explains structural ensemble failure modes.
- A general benchmark construction methodology, **Epistemic Bundles**.
- A concrete instantiation, **Epistemic-GSM**, demonstrating the feasibility of this approach.

2 Related Work

2.1 Selection-Based Ensembling

Self-consistency [1] samples multiple reasoning paths and selects the most frequent answer. Verifier-based methods [2] rank candidates using learned correctness models. These approaches are effective only when the correct answer appears among candidates; they are information-theoretically bounded by the quality of the raw generation.

2.2 Generative Fusion

Methods like LLM-Blender [3] combine ranking with generative fusion, conditioning a new generation on top-ranked candidates. However, existing benchmarks rarely enforce regimes where the correct answer is absent from all candidates, limiting the evaluation of true reconstruction behaviors versus simple copying.

2.3 Hallucination & Uncertainty

Work on hallucination detection [6], self-evaluation [7], and factual verification [8] typically operates on single outputs. These methods do not address the complex dynamics of distributed truth across a multi-generation ensemble.

3 A Taxonomy of Epistemic States

Let a query Q induce a set of outputs $\mathcal{A} = \{A_1, \dots, A_n\}$ with unknown ground truth T. We define five epistemic states describing the relationship between $\mathcal{A}$ and T:

Class I: Redundant Correctness At least one output contains the full correct answer. Selection methods suffice here.

Class II: Complementary Fragmentation No individual output is fully correct, but the union of correct segments across outputs suffices to reconstruct T. This represents a "known-unknown" regime prior to integration.

Class III: Reasoning–Result Divergence Outputs contain valid reasoning logic but arrive at incorrect final answers (e.g., calculation errors).

Class IV: Contradictory Hallucination Outputs are mutually inconsistent and factually incorrect. This corresponds to an "unknown-unknown" regime: truth is absent.

Class V: False Consensus Outputs converge on the same incorrect answer due to shared pre-training biases or artifacts.

4 Epistemic Bundles: Construction Framework

4.1 General Methodology

Given a benchmark with query Q and reference solution T, an *Epistemic Bundle* is constructed by generating a set of candidate outputs $\mathcal{A}$ such that the distribution of information matches a target epistemic state.

We employ algorithmic perturbations including:

- **Reasoning Erasure:** Removal or corruption of intermediate steps.
- **Answer Swapping:** Replacement of final answers while preserving reasoning traces.
- **Noise Injection:** Insertion of plausible but incorrect sub-steps.
- **Fragment Distribution:** Ensuring non-overlapping correct segments across candidates.

This methodology is task-agnostic and applies to any benchmark with executable traces or step-by-step reference solutions.

4.2 Epistemic-GSM

We apply this framework to GSM8K, producing **Epistemic-GSM**, a dataset of 8,500 bundles. This allows us to rigorously test if an ensemble model can reconstruct a valid proof when no single candidate in the input provides it.

5 Query Access and Epistemic Restraint

We evaluate refiners under two conditions: access to $\mathcal{A}$ alone, and access to $(Q, \mathcal{A})$. While conditioning on the query Q improves grounding, it introduces a "shortcutting" risk where the model solves the problem ab initio, ignoring the ensemble context.

This failure mode is critical when truth is absent. An unrestricted, query-conditioned model may confidently hallucinate an answer even when the provided evidence ($\mathcal{A}$) is contradictory (Class IV). To measure this, Epistemic Bundles include:

- **Empty-bundle control:** $R(Q, \varnothing)$
- **Mismatched-bundle control:** $R(Q, \pi(\mathcal{A}))$

Successful epistemic integration requires reconstructing truth when evidence exists (Class II/III) and exercising restraint when it does not (Class IV).

6 Theoretical Analysis

Proposition 1 (Selection Limitation). Selection-based ensembles cannot recover T when $T \notin \mathcal{A}$. The probability of correctness is bounded by $P(\exists A \in \mathcal{A} : A = T)$.

Proposition 2 (Integration Feasibility). If the Shannon information required to derive T is distributed across $\mathcal{A}$, epistemic integration is not information-theoretically blocked. The challenge becomes one of alignment and logic synthesis rather than generation.

7 Future Work: Refiner Architectures

Future work should explore refiner architectures that jointly encode candidates to detect inter-candidate contradictions, verify consistency without external tools, and express calibrated uncertainty.

8 Limitations

While this framework provides a robust diagnostic tool, it is subject to three primary limitations.

First, **Synthetic Validity**: Epistemic Bundles rely on synthetic perturbations to simulate model errors. While we attempted to model natural hallucinations (e.g., calculation slips, logic jumps), real-world model failures may exhibit more subtle semantic entanglements that algorithmic perturbation cannot fully capture.

Second, **Domain Scope**: The current instantiation, Epistemic-GSM, focuses on mathematical reasoning. Mathematical truth is generally binary and verifiable. Generalizing this framework to open-ended domains—such as creative writing or moral reasoning—where "truth" is subjective or multimodal remains an open challenge.

Third, **Computational Cost**: Constructing high-quality bundles requires multiple passes of generation and perturbation for every data point. This makes the creation of Epistemic Bundles significantly more computationally intensive than standard inference or simple majority voting benchmarks.

9 Broader Impact

This work supports the development of more reliable and epistemically grounded language systems. As LLMs are increasingly deployed in high-stakes environments—such as clinical decision support, legal analysis, and automated code generation—the cost of "confident hallucinations" becomes prohibitive.

By shifting the goal from consensus maximization to epistemic integration, we encourage the development of systems that can:

1. **Recover Truth:** Utilize the collective intelligence of smaller, less capable models to reconstruct solutions that no single model could produce alone.

2. **Signal Ignorance:** Crucially, this framework promotes "Epistemic Humility"—the ability of a system to recognize when the provided information is insufficient to answer a query, thereby reducing the rate of misleading falsehoods in production systems.

However, we acknowledge that better truth-reconstruction methods could theoretically be employed to synthesize coherent misinformation from fragmented sources. Mitigation of this risk requires robust alignment training alongside the architectural improvements proposed here.

10 Conclusion

The prevailing approach to LLM ensembling has relied heavily on the Selection Hypothesis, treating the problem as a search for a needle in a haystack. We have demonstrated that in many reasoning regimes, the needle does not exist; instead, we have only the scattered iron filings of the truth.

We introduced the **Epistemic Integration** framework and the **Epistemic Bundles** methodology to formalize this challenge. Our analysis suggests that reliability requires a fundamental architectural shift: from systems that force a choice among candidates to systems that reason over the relationships between them. From an epistemic perspective, the ultimate measure of intelligence is not merely the ability to reconstruct truth from fragments, but the wisdom to recognize when the evidence is insufficient and abstention is the only correct action.

References

[1] Wang, X. et al. Self-Consistency Improves Chain of Thought Reasoning. *ICLR*, 2023.

[2] Cobbe, K. et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint*, 2021.

[3] Jiang, D. et al. LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion. *ACL*, 2023.

[4] Du, Y. et al. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint*, 2023.

[5] Madaan, A. et al. Self-Refine: Iterative Refinement with Self-Feedback. *NeurIPS*, 2023.

[6] Manakul, P. et al. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection. *EMNLP*, 2023.

[7] Kadavath, S. et al. Language Models Know What They Know. *arXiv preprint*, 2022.

[8] Lin, S. et al. Teaching Models to Evaluate Facts. *EMNLP*, 2022.