Reducing Hallucinations in Language Models via Cross-Validation

Anonymous ACL submission

Abstract

Large language models frequently generate fluent but factually incorrect statements—hallucinations—that undermine their reliability in critical applications. We introduce a cross-validation method that reduces hallucinations by leveraging agreement between independent models. Our system employs two independent language models as generators to independently answer the same prompt, with a third language model serving as a judge to assess semantic agreement and determine the system's final output. The judge executes a binary decision rule, emitting as the system's response either the common answer when it determines semantic agreement or "I don't know" otherwise. This approach operationalizes the epistemic principle that agreement between independent intelligent agents provides evidence of truth. Across three diverse benchmarks—SimpleQA, TriviaQA, and GSM-8K Platinum—our method achieves hallucination reductions of 87%, 82%, and 74% respectively compared to the best individual model. The system requires no model modifications, fine-tuning, or access to internal states, making it immediately applicable to any combination of language models. While coverage is low when individual model accuracies are low, the method provides a principled way to expose uncertainty rather than fabricate content, particularly valuable for high-stakes deployments where avoiding false information outweighs maximizing response rate.

1 Introduction

Large language models (LLMs) have rapidly advanced general-purpose language capabilities—covering instruction following, multi-step reasoning, and tool use—and are now widely deployed across domains (Zhao et al., 2024). Despite their fluency and breadth, LLMs can produce confident but unsupported or incorrect statements—*hallucinations*—that limit reliability in factual QA, summarization, and decision support (Huang et al., 2023).

We propose a minimal, external ("black-box") hallucination reduction method grounded in a fundamental epistemic principle: *agreement between independent intelligent agents is evidence of truth*. Concretely, our system runs a prompt through two independent generators (e.g., models from different organizations, with different architectures, or with distinct training data), then passes their outputs to a third model—the *judge*—whose role is to test semantic agreement. The decision rule is binary: if the generators agree, the system returns their common answer; if they disagree, it returns "I don't know." Figure 1 illustrates this architecture, while Figure 2 previews one experimental result. For the full experimental results, see Table 1.

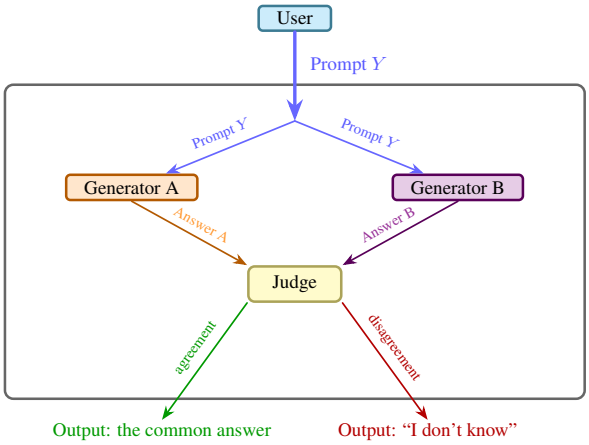

Figure 1: Cross-validation system.

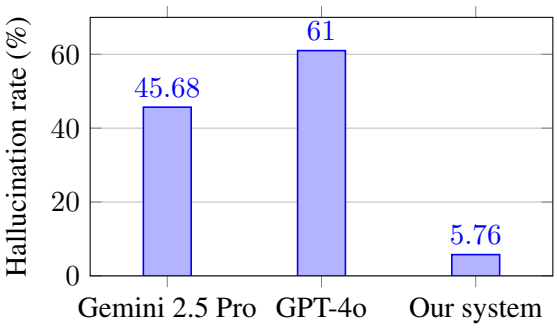

Figure 2: SimpleQA hallucination rate.

2 Related Work

Grounding with external knowledge. Retrieval-grounded generation reduces hallucinations by conditioning on evidence (RAG; REALM) (Lewis et al., 2020; Guu et al., 2020). Pipeline and editing frameworks orchestrate retrieval/feedback or rewrite unsupported spans (LLM-Augmenter; RARR) (Peng et al., 2023; Gao et al., 2023). Tool-use further externalizes skills without retraining (Toolformer) (Schick et al., 2023). These are complementary to ours: grounding lowers divergence; our agreement gate abstains when divergence remains.

Self-consistency. Unlike self-consistency methods that aggregate multiple samples from the same model (Wang et al., 2022; Manakul et al., 2023), our approach leverages model independence. This distinction is crucial: while self-consistency can reduce random errors, it cannot escape systematic biases inherent in a single model's training. By requiring agreement between models with different training data and architectures, our cross-validation approach offers a stronger guarantee against shared failure modes.

Multiagent debate (MAD). MAD has multiple LM instances propose answers and then argue over several rounds to reach a single final answer (Du et al., 2023). In the original experiments, agents are typically *copies of the same model* (e.g., multiple ChatGPT instances), so independence between generators is not required; a small heterogeneous demo (ChatGPT vs. Bard) is shown, but most settings remain intra-model. Because each round conditions on others' messages, debate further *increases dependence* among agents' final outputs. Moreover, MAD does not provide an explicit abstention rule: it produces a final answer even when debate does not converge. In contrast, we require cross-*model* independence and map disagreement to abstention.

Uncertainty and knowledge editing. Uncertainty signals (e.g., $P(\text{True})$ / "I know" estimates and semantic entropy) support selective refusal (Kadavath et al., 2022; Kuhn et al., 2023; Farquhar et al., 2024). Our criterion is orthogonal: inter-model disagreement $\Rightarrow$ abstain, optionally gated by a confidence threshold. Knowledge editors (ROME, MEND, SERAC, MEMIT) can reduce disagreement upstream; our judge still filters residual mismatches (Meng et al., 2022a; Mitchell et al., 2022a,b; Meng et al., 2022b).

3 Problem Formulation

Large language models (LLMs) frequently produce fluent but *unsupported* statements—*hallucinations*—that undermine reliability in factual QA, retrieval-grounded summarization, and multi-step reasoning. We take hallucination to mean generated content that is unfaithful to the provided context or false with respect to verifiable world knowledge, even when it appears locally coherent (Ji et al., 2023; Maynez et al., 2020; Bang et al., 2025; Lin et al., 2021). To reason across tasks, it is useful to distinguish *faithfulness* (consistency with the input and instructions) from *factuality* (correspondence to external facts). Recent taxonomies refine this into *intrinsic* hallucinations that contradict the input and *extrinsic* ones that introduce unverifiable content, with further subtypes capturing instruction, context, and logical inconsistencies (Maynez et al., 2020; Ji et al., 2023; Bang et al., 2025; Huang et al., 2023). In step-by-step settings, explanations can also be *unfaithful* to the model's own decision process, yielding inconsistent or post-hoc rationales (Turpin et al., 2023; Lanham et al., 2023).

Why do such errors arise? Theoretically, hallucinations reflect epistemic uncertainty and calibration failures in generative decoding: likelihood-trained models are rewarded for fluent continuations rather than calibrated refusals, so they often guess when outside their knowledge boundary (Ji et al., 2023; Maynez et al., 2020). This motivates a knowledge-relative framing that emphasizes justified abstention when evidence is insufficient and careful conditioning on available context at inference time (Kadavath et al., 2022; Ji et al., 2023). Empirically, models exhibit a limited ability to estimate their own correctness—suggesting a latent notion of "I know / I don't know"—but this self-knowledge is imperfect and task dependent (Kadavath et al., 2022). In retrieval-augmented systems, additional failure channels stem from retrieval noise, partial coverage, and attribution gaps that blur the line between grounded statements and plausible fabrications (Lewis et al., 2020; Ji et al., 2023)).

Measuring hallucination remains challenging. Truthfulness probes highlight susceptibility to widely held falsehoods (Lin et al., 2021); summarization suites target span-level faithfulness and report that even strong automatic judges disagree on difficult cases (Bao et al., 2025); dialogue-level datasets reveal spontaneous and induced halluci-

nations at scale and motivate specialized evaluators (Luo et al., 2024). Across settings, black-box studies show that sentence-level detection is far from solved, with detectors performing unevenly and failing on adversarial or long-context examples (Manakul et al., 2023; Bao et al., 2025).

The consequences are most acute in high-stakes domains. In clinical summarization, careful annotation still finds non-negligible hallucination and omission rates that require workflow controls to keep risk below human baselines (Asgari et al., 2025). Under adversarial prompting, general-purpose models readily incorporate fabricated clinical details; mitigation prompts reduce but do not eliminate the effect (Omar et al., 2025). These observations justify evaluating methods not only by average correctness but also by their *propensity to fabricate*, especially under distribution shift or hostile inputs.

To conclude, the problem we aim to mitigate is hallucinations, and our objective is to reduce them.

4 Proposed Method

Our approach is straightforward: when two independent language models agree on an answer, we trust it. When they disagree, we admit uncertainty. This method drives substantial reductions in hallucination rates without requiring any modifications to the underlying models.

4.1 How the System Works

The architecture uses three language models working together in a specific way:

Two Independent Generators. We run the same prompt through two language models—call them Generator A and Generator B. Ideally, the two generators would be perfectly independent. In practice, we aim for the strongest feasible independence by pairing models that differ in provider, architecture, or training data. Copies or close variants of the same base model should be avoided, as they are highly dependent.

The Judge. A third model serves as the judge (Hebrew: *matzliv*, "cross-checker"). It receives the prompt and the two outputs from the generators and performs a semantic agreement test. When it detects agreement on the facts required by the prompt, it emits, *as the system's final response*, the common answer; when it detects disagreement, it emits, *as the system's final response*, "I don't know".

4.2 The Core Principle

Agreement between independent intelligent agents is evidence of truth.

(See Appendix A for an extension to this principle.)

The method builds on the following epistemic principle: when independent intelligent agents converge on the same answer without coordinating, that agreement is evidence of truth.

This principle parallels established practice in expert consultation: when experts from distinct domains independently converge on the same answer to a factual question, their agreement constitutes stronger evidence than a single expert's response. We apply the same epistemic rationale to language models, treating independent inter-model agreement as evidence for correctness.

4.3 Why Independence Matters

We need the generators to be truly independent for this to work. If both models learned from the same training data, or share similar architectures, their agreement doesn't mean much—they might just be making the same systematic errors.

We ensure independence through:

- Using models from different organizations (OpenAI vs. Google vs. Anthropic)
- Different model architectures when possible
- Different training datasets and methodologies

Perfect independence is impossible—all models train on parts of the internet, after all—but we maximize it as much as current technology allows.

4.4 Inference Procedure (Pseudocode)

Algorithm: CROSS-VALIDATION

Input: `prompt_y`

```
answer_a ← Generator A(prompt_y)
answer_b ← Generator B(prompt_y)
judge_prompt ← the prompt from Section 4.5
return Judge(judge_prompt)
```

Notation. $x \leftarrow M(p)$ denotes running the language model M on the prompt p and assigning its response to the variable x.

4.5 The Judge Prompt

During the process of cross-validation, the judge is given the following prompt:

```
Prompt Y:
{prompt_y}
Answer A:
{answer_a}
Answer B:
{answer_b}
You are a cross-validator-judge.
You are given a prompt Y and two
answers to that prompt: Answer A
and Answer B.
Your response/output should look
like it's responding to the prompt Y,
and you shouldn't reference anything
that is said here. The user reading
your response shouldn't know that
behind the scene there are two
answers and you cross them.
If A and B agree, output their
common answer.
If A and B disagree, output
"I don't know".
You are not to determine which
answer is correct. You are not to
inject your own beliefs or additional
content. You don't break ties by
"choosing the better answer". You
don't validate domain facts.
To output the common answer, A and B
must agree exactly on the facts being
asked about.
```

Remark. The prompt above works well with GPT-4o (2024-08-06) as the judge in our experiments. Different judge models may benefit from adaptations. For a highly capable judge model, you may feed it this research paper—so it better understands its role as a judge in the system.

4.6 What Happens in Practice

We illustrate the system's operation with examples from our experiments:

Example 1: When the generators agree

- **Prompt:** "Who was Japan's Prime Minister after Ichirō Hatoyama left office on December 23, 1956?"
- **Generator A:** "The Prime Minister of Japan after Ichirō Hatoyama was Tanzan Ishibashi. He took office on December 23, 1956, but his tenure was very short. He suffered a stroke and resigned just two months later, on February 23, 1957."
- **Generator B:** "After Ichirō Hatoyama left office on December 23, 1956, Tanzan Ishibashi became the Prime Minister of Japan."
- **Judge Output:** "The Prime Minister of Japan after Ichirō Hatoyama was Tanzan Ishibashi."

Both models identify Tanzan Ishibashi, though with different levels of detail. The judge recognizes this agreement and outputs the consensus answer.

Example 2: When the generators disagree

- **Prompt:** "How many more votes did Freeman Freeman-Thomas win than George Sandys in the 1906 Bodmin by-election?"
- **Generator A:** "In the 1906 Bodmin by-election, Freeman Freeman-Thomas won 1,093 more votes than George Sandys."
- **Generator B:** "Freeman-Thomas won by a margin of 863 votes."
- **Judge Output:** "I don't know."

The models give different vote margins (1,093 vs. 863). Rather than guess which is correct or average them, the system admits uncertainty.

4.7 Why This Works

The method exploits a key asymmetry: when models are correct, they converge on the same answer (because there's typically one right answer). When models hallucinate, they tend to hallucinate differently (because there are many ways to be wrong).

Consider a factual question whose correct answer is "Paris." Both models, when correct, output "Paris." In contrast, when hallucinating, their responses typically diverge (e.g., "London" vs. "Berlin") because the space of incorrect answers is large, making accidental agreement unlikely.

This is why agreement between independent models provides such strong evidence. The probability of two models independently hallucinating the exact same false answer is much lower than the probability of both being correct.

4.8 Externality and Composability

Our method is purely external—it requires no modification to the underlying models, no additional training, and no access to internal states or confidence scores. It is immediately applicable to any combination of language models, and it composes with other hallucination-reduction methods: apply those upstream, then use our agreement gate as a final abstention filter to further reduce hallucinations.

4.9 Generalization Beyond Modern LLMs

The most essential feature of cross-validation is that it extends far beyond modern large language

models. It is not a technique tailored to deep learning or neural network architectures, nor is it confined to any current AI paradigm. Rather, it is a general epistemic framework for evaluating the reliability of information through the convergence of independent intelligent agents.

The structure is simple and principled: several intelligent agents, each capable of interpreting and responding to a task, are presented with the same input. These agents are independent—not just isolated during execution, but fundamentally distinct in origin, reasoning, or construction. This independence is what gives any convergence of their outputs epistemic significance. When such agents arrive at the same answer without influence or coordination, that agreement is taken as evidence pointing toward truth.

This framework is not bound to LLMs or modern AI tools. It generalizes to rule-based systems, human experts, committees, and entirely new forms of intelligence. It makes no assumptions about how the agents think—only that they think independently.

Cross-validation is thus a timeless and architecture-independent method. It makes no claims beyond what can be justified structurally. It is cautious where others are overconfident, and its restraint comes not from limitation, but from epistemic discipline.

By requiring agreement between distinct minds, cross-validation offers a principled way to surface what is most likely to be true—even when certainty remains out of reach.

5 Experiments

5.1 Experimental Setup

We evaluate our cross-validation hallucination reduction system using three diverse benchmarks that test different aspects of model reliability: SimpleQA, TriviaQA, and GSM-8K Platinum. These benchmarks were selected to assess factual accuracy, general knowledge, and mathematical reasoning respectively, providing comprehensive coverage of common LLM failure modes.

5.1.1 Model Configuration

We implement our system using two independent generators and a judge:

- **Generator A**: Gemini 2.5 Pro
- **Generator B**: GPT-4o (2024-08-06)
- **Judge**: GPT-4o (2024-08-06)

All models were run with default temperature settings (temperature = 1.0).

The generators were chosen for their architectural independence—developed by different organizations (Google and OpenAI) with distinct training methodologies, model architecture, and data curation practices. This independence is crucial for our epistemic framework, as agreement between models that share architecture or training data would provide weaker evidence of truth.

For the judge, we use the same model as one of the generators (GPT-4o). This does not violate the independence assumption since independence is only required between the generators themselves. The judge serves a fundamentally different role: it does not generate answers to the original prompt but instead evaluates the relationship between the two generator outputs, determining whether they agree or disagree. Since the judge only performs relational comparison rather than content generation, using the same model as one of the generators does not compromise the system's epistemic guarantees.

5.1.2 Evaluation Metrics

We measure three primary metrics to comprehensively assess system performance: **Hallucination/Error Rate**, **Attempted Rate**, and **Accuracy Given Attempted**.

Definitions and Symbols. Let $\mathcal{X}$ be the prompt space and $\mathcal{Y}$ the admissible answer space. Let $D = \{(x_i, y_i^*)\}_{i=1}^n$ be the evaluation dataset with $x_i \in \mathcal{X}$ and $y_i^* \in \mathcal{Y}$, where y_i^* denotes the gold (ground-truth) answer for x_i. Let $f : \mathcal{X} \to \mathcal{Y} \cup \{\perp\}$ be the system, where $\perp$ denotes abstention and $\perp \notin \mathcal{Y}$. For each $x \in \mathcal{X}$, let $\equiv_x$ be the equivalence relation on $\mathcal{Y}$ such that for any $y_1, y_2 \in \mathcal{Y}$, $y_1 \equiv_x y_2$ iff y_1 and y_2 agree on the facts being asked about by x. We use $\mathbf{1}\{\cdot\}$ for indicators and $\widehat{\cdot}$ for empirical estimates over D.

- **Hallucination/Error Rate**: Percentage of incorrect answers among all questions, measuring the system's propensity to generate false information
  Formal.

$$\widehat{H} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{ f(x_i) \neq \perp \ \wedge \ f(x_i) \not\equiv_{x_i} y_i^* \}.$$

- **Attempted Rate**: Percentage of questions where the system provided an answer rather than abstaining with "I don't know", measuring coverage

Formal.

$$\widehat{C} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{f(x_i) \neq \perp\}.$$

- **Accuracy Given Attempted**: Accuracy calculated only on questions where the system attempted an answer, measuring precision when confident
  Formal.

$$\widehat{\mathrm{Acc}}_{\mathrm{att}} = \frac{\sum_{i=1}^{n} \mathbf{1}\{f(x_i) \neq \perp \ \wedge \ f(x_i) \equiv_{x_i} y_i^*\}}{\sum_{i=1}^{n} \mathbf{1}\{f(x_i) \neq \perp\}}.$$

5.1.3 Benchmarks

SimpleQA consists of factual questions designed to elicit hallucinations, testing the model's ability to distinguish between what it knows and what it might confabulate. Questions span diverse domains including history, science, and current events.

TriviaQA evaluates reading comprehension and general knowledge retrieval across a broad spectrum of topics. This benchmark tests whether models can accurately recall and synthesize well-established facts. While models can abstain in the default TriviaQA setup, such abstentions are typically counted as errors. We augmented the evaluation to detect and properly handle abstentions, allowing us to distinguish between incorrect answers and acknowledged uncertainty.

GSM-8K Platinum comprises grade-school mathematics word problems requiring multi-step reasoning. This benchmark evaluates logical consistency and the ability to maintain coherent reasoning chains without computational errors. The standard GSM-8K Platinum format requires numerical answers and does not permit abstention. We added an abstention mechanism that allows models to output "I don't know" when uncertain, with detection logic to properly identify and count these abstentions separately from errors.

5.2 Results

Table 1 presents comprehensive results across all three benchmarks. The cross-validation system demonstrates consistent and substantial improvements across diverse task types, validating our theoretical framework.

5.2.1 SimpleQA: Factual Accuracy Under Uncertainty

On SimpleQA, designed to test resistance to factual hallucinations, the cross-validation system achieves a dramatic reduction in hallucinations—from 45.68% (best single model) to 5.76%. This represents an 87% reduction in error rate. While the attempted rate drops to 34.44%, the system maintains 83.29% accuracy on questions it does attempt, compared to just 54.11% for Gemini 2.5 Pro alone.

The conservative attempted rate reflects the system's epistemic discipline: when faced with factual questions where the generators disagree, the system correctly identifies this uncertainty and abstains rather than guessing. This behavior is particularly valuable for high-stakes applications where false information could have serious consequences.

5.2.2 TriviaQA: General Knowledge Retrieval

TriviaQA results demonstrate near-optimal performance, with hallucination rate dropping from 3.46% to 0.63%—an 82% reduction. Remarkably, the system maintains 92.14% coverage while achieving 99.32% accuracy on attempted questions. This high attempted rate follows directly from both models' strong individual performance on this benchmark.

The strong performance on TriviaQA validates our conjecture that agreement between independent intelligent agents provides evidence of truth, particularly when both models have high individual accuracy.

5.2.3 GSM-8K Platinum: Mathematical and Logical Reasoning

The system excels on GSM-8K Platinum mathematical tasks, reducing the error rate by 74% (from 0.66% to 0.17%) while maintaining 98.43% coverage. With 99.83% accuracy on attempted problems, the cross-validation approach nearly eliminates mathematical errors.

This exceptional performance demonstrates that when both generators achieve high individual accuracy (>98%), the system maintains near-complete coverage while further reducing errors through agreement-based filtering.

5.3 Analysis

5.3.1 Coverage and Individual Model Performance

Coverage (attempted rate) is driven mostly by the generators' individual accuracies; under perfect independence, it is well approximated by the product of those accuracies.

Model	Hallucinations (%)	Accuracy@Attempted (%)	Attempted (%)
SimpleQA			
Gemini 2.5 Pro	45.68	54.11	99.54
GPT-4o (2024-08-06)	61.00	37.79	98.06
Gemini 2.5 × GPT-4o → GPT-4o	**5.76**	**83.29**	34.44
TriviaQA			
Gemini 2.5 Pro	4.09	95.91	100.00
GPT-4o (2024-08-06)	3.46	96.54	100.00
Gemini 2.5 × GPT-4o → GPT-4o	**0.63**	**99.32**	92.14
GSM-8K Platinum			
Gemini 2.5 Pro	0.66	99.34	100.00
GPT-4o (2024-08-06)	1.57	98.43	100.00
Gemini 2.5 × GPT-4o → GPT-4o	**0.17**	**99.83**	98.43

Table 1: Performance comparison across three benchmarks. The notation $A \times B \rightarrow J$ denotes our system with generators A and B evaluated by judge J. Cross-validation (Gemini 2.5 × GPT-4o → GPT-4o) achieves hallucination reductions of 87% (SimpleQA), 82% (TriviaQA), and 74% (GSM-8K Platinum) compared to the best individual model.

More formally, let X be a random query with a gold answer $y^*(X)$. Let $\equiv_X$ denote semantic equivalence with respect to X: for $y_1, y_2 \in \mathcal{Y}$, $y_1 \equiv_X y_2$ iff they agree on the facts required by X. Let $\mathcal{Y}/\equiv_X$ denote the quotient of $\mathcal{Y}$ under $\equiv_X$.

For each query x, let $Y_A(x)$ and $Y_B(x)$ denote the responses produced by generators A and B, respectively. Define the individual (unconditional) accuracies $p_A := \Pr\big[Y_A(X) \equiv_X y^*(X)\big]$ and $p_B := \Pr\big[Y_B(X) \equiv_X y^*(X)\big]$. The system answers only when the generators *agree*, so the coverage (attempted rate) is

$$\begin{aligned} C &:= \Pr\big[Y_A(X) \equiv_X Y_B(X)\big] \\ &= \underbrace{\Pr\big[Y_A(X) \equiv_X Y_B(X) \equiv_X y^*(X)\big]}_{\text{both correct}} \\ &\quad + \underbrace{\mathbb{E}\left[\sum_{\substack{C \in \mathcal{Y}/\equiv_X \\ y^*(X) \notin C}} \mathbf{1}\{\, Y_A(X) \in C,\, Y_B(X) \in C \,\}\right]}_{\varepsilon \text{ (same-wrong mass)}}. \end{aligned}$$

Writing $\Pr\big[Y_A(X) \equiv_X Y_B(X) \equiv_X y^*(X)\big] = p_A p_B + \Delta_{cc}$, where Δ_{cc} captures correlation in being correct, we obtain

$$C \;=\; p_A p_B \;+\; \Delta_{cc} \;+\; \varepsilon.$$

Under the heuristic that the generators are (approximately) independent and same-wrong agreement is rare, this yields $C \approx p_A p_B$.

SimpleQA: With individual model accuracies of approximately 54% and 39%, the expected attempted rate under independence would be roughly 20%, though we observe 34.44%. This slight elevation suggests some correlation in correct answers and occasional agreement on non-gold answers.

Let $a_i = \Pr[\text{attempt}]$ and $r_i = \Pr[\text{correct} \mid \text{attempt}]$; then the unconditional accuracy is $p_i = a_i r_i$. From Table 1: $a_A = 0.9954$, $r_A = 0.5411$, so $p_A = a_A r_A \approx 0.5386$; $a_B = 0.9806$, $r_B = 0.3779$, so $p_B = a_B r_B \approx 0.3706$. The independence baseline is $p_A p_B \approx 0.1996$ (19.96%), whereas the observed coverage is $C = 0.3444$. Using $\mathrm{Acc}_{\text{att}} = 0.8329$,

$$\frac{p_A p_B + \Delta_{cc}}{C} = \mathrm{Acc}_{\text{att}}, \text{ so } \Delta_{cc} \approx 0.0873,$$

so $\varepsilon = C - (p_A p_B + \Delta_{cc}) \approx 0.0575$. Hence, excess agreement beyond independence arises from both correlated correctness ($\Delta_{cc} > 0$) and a nontrivial same-wrong mass ($\varepsilon > 0$).

TriviaQA. With individual accuracies of 95.91% (Gemini 2.5 Pro) and 96.54% (GPT-4o), the product baseline is 92.59%; observed coverage is 92.14%, a close match consistent with the product approximation for coverage.

GSM-8K Platinum. With individual accuracies of 99.34% and 98.43%, the product baseline is 97.78%; observed coverage is 98.43%, a close match consistent with the product approximation for coverage.

5.3.2 Empirical Validation of Theoretical Claims

The experimental results provide strong empirical support for our core epistemic principle: agreement between independent intelligent agents serves as evidence of truth. The relationship between individual model accuracy and system coverage follows the predicted pattern, with attempted rates approximating the product of individual accuracies, validating our independence assumption.

Importantly, the system's precision consistently exceeds both individual models' accuracies when it does attempt an answer. This demonstrates that agreement acts as an effective filter: when two independent models converge on the same answer, it's far more likely to be correct than either model's individual prediction.

6 Conclusion

We presented a cross-validation framework that reduces hallucinations in language models based on an epistemic principle: agreement between independent intelligent agents provides evidence of truth. By requiring consensus between two independent generators (e.g., GPT-4o and Gemini 2.5 Pro) and abstaining when they disagree, our system achieves hallucination reductions of 87% (SimpleQA), 82% (TriviaQA), and 74% (GSM-8K Platinum) compared to the best individual model.

The method's strength lies in its simplicity and generality. It requires no model modifications, fine-tuning, or access to internal states—just three models working together under a binary decision rule. When generators agree, output their answer; when they disagree, output "I don't know." This transforms model disagreement from a hidden failure mode into an explicit signal of epistemic uncertainty.

While coverage decreases when individual model accuracies are low, this trade-off is often desirable in high-stakes applications where avoiding false information outweighs maximizing response rates. The framework extends beyond current LLMs to any form of intelligent agent, making it a timeless strategy for reliability rather than a temporary patch for today's models.

Future work could explore extending the framework to more than two generators and integration with complementary approaches like retrieval augmentation. As language models become more prevalent in critical applications, the ability to reliably say "I don't know" becomes as important as accuracy itself—and cross-validation provides a principled, immediately deployable path toward this goal.

Limitations

While our cross-validation system demonstrates substantial improvements in hallucination reduction, several limitations constrain its applicability:

Low coverage under low individual accuracies. When individual generators perform poorly on a task, the system's coverage is severely limited. For instance, on SimpleQA where Gemini 2.5 Pro achieves $\approx 54\%$ accuracy and GPT-4o achieves $\approx 37\%$ accuracy, our cross-validation system only attempts to answer $\approx 34\%$ of questions. This occurs because the system requires both generators to agree, and when models frequently make errors, they rarely make the same error. The probability of agreement combines both cases where models are correct together and cases where they happen to be wrong in the same way.

Limited applicability to open-ended generation tasks. Our method is designed for prompts with well-defined correct answers where agreement can be meaningfully assessed. For creative writing, brainstorming, or other open-ended tasks where multiple diverse responses may be equally valid, requiring exact agreement between generators would inappropriately reject legitimate variations. The judge's binary agree/disagree decision framework does not naturally extend to evaluating the compatibility of creative or subjective outputs.

Computational overhead. The system requires running two independent generators and a judge for each query, increasing computational cost and latency beyond a single-model baseline (with monetary cost roughly additive across the three components). While this overhead may be acceptable for high-stakes applications where accuracy is paramount, it poses challenges for deployment at scale or in resource-constrained environments.

References

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: Llm hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. 2025. FaithBench: A diverse hallucination benchmark for summarization by Modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics:*

Human Language Technologies (Volume 2: Short Papers), pages 448–461, Albuquerque, New Mexico. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12). Also available as arXiv:2202.03629.

Saurav Kadavath, Ethan Perez, Tamera Lanham, Sam McCandlish, Tom Henighan, Jared Kaplan, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. In *Advances in Neural Information Processing Systems (NeurIPS)*. NeurIPS 2021; arXiv:2109.07958.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: Benchmarking hallucinations in information-seeking dialogues. *arXiv preprint arXiv:2406.07070*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Memit: Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*.

Mahmud Omar, Vera Sorin, Jeremy D. Collins, David Reich, Robert Freeman, Nicholas Gavin, Alexander Charney, Lisa Stump, Nicola Luigi Bragazzi, Girish N. Nadkarni, and 1 others. 2025. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Communications Medicine*, 5(1):330.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Michael Turpin, Thomas Stepleton, Jonathan Uesato, Po-Sen Huang, Sven Gowal, Timothy G. J. Rudner, and 1 others. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Jacob Devlin, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Cui, Yifan Li, Pengfei Liu, Zhicheng Dou, and Ji-Rong Wen. 2024. A survey of large language models. *arXiv preprint arXiv:2402.06196*.

A Extension: Converging Opinions Among Agents

Conjecture: On Converging Opinions Among Agents

Let multiple agents form the same opinion.
Then:

- *The more such agents there are,*
- *the more intelligent each agent is,*
- *and the more independent they are from one another,*

the stronger their agreement is as evidence for the truth of that opinion.