HALLUCINATION AS MISCLASSIFICATION: A COMPOSITE ABSTENTION ARCHITECTURE FOR LANGUAGE MODEL OUTPUT CONTROL

Angelina (Davini) Hintsanen
NEXUS Laboratory
Finland
angelina@nexus-ailab.org

ABSTRACT

Large language models routinely produce unsupported claims—a failure termed hallucination. We propose a control-theoretic framing: hallucination is a misclassification error at the output boundary, where internally generated completions are emitted as if grounded in evidence. This framing motivates a composite intervention combining instruction-based refusal with a structural abstention gate. The gate computes a support deficit score S_t from three black-box signals—self-consistency (A_t), paraphrase stability (P_t), and citation coverage (C_t)—and blocks output when S_t exceeds a threshold. In a controlled evaluation across 50 items, five epistemic regimes, and three models (GPT-4o-mini, GPT-4o, GPT-3.5-turbo), neither mechanism alone was sufficient: instruction-only prompting reduced hallucination sharply, but exhibited over-cautious abstention on 10% of answerable items for GPT-4o-mini and GPT-4o, and residual hallucination for GPT-3.5-turbo (6% overall; driven primarily by conflicting-evidence items). The structural gate preserved 100% answerable accuracy across models but missed confident confabulation on conflicting-evidence items (70% hallucination for GPT-4o-mini and GPT-4o). The composite architecture achieved 96–98% overall accuracy with 0–4% hallucination, while inheriting the instruction component's 10% abstention on answerable items for GPT-4o-mini and GPT-4o. A supplementary 100-item no-context stress test derived from TruthfulQA confirmed that structural gating provides a capability-independent abstention floor: instruction-only abstention degraded to 62% for GPT-3.5-turbo, whereas the gate and composite conditions enforced 98–100% abstention across all models.These results are consistent across the tested autoregressive models, though architecture-level generality remains to be established. Overall, instruction-based refusal and structural gating exhibit complementary failure modes—instruction can over-abstain on answerable items, while the gate can miss confident confabulation under conflicting evidence—suggesting that effective hallucination control benefits from combining both mechanisms.

1 INTRODUCTION

Large language models (LLMs) generate text by predicting the next token conditional on preceding context (Brown et al., 2020; Touvron et al., 2023). This autoregressive process produces fluent output but routinely generates claims not supported by input evidence—termed hallucination (Ji et al., 2023; Huang et al., 2023; Maynez et al., 2020). Hallucination persists under retrieval augmentation (Shuster et al., 2021), instruction tuning, and reinforcement learning from human feedback (Ouyang et al., 2022).

1.1 THE GAP-FILLING PROBLEM

When a query requires information absent from the prompt, retrieved context, and reliable model parameters, the system faces an epistemic gap. The model continues generating because training

rewards fluent completion rather than epistemic caution (Lin et al., 2022; Kadavath et al., 2022). We propose that this gap-filling behavior constitutes a misclassification: the model's internal completion process generates candidate content, and when that content is emitted without distinguishing evidence-backed from prior-only generation, the system misclassifies prior-driven gap-filling as grounded output.

This framing is conceptually inspired by control-theoretic models of biological inference, where internally generated signals (e.g., interoceptive arousal during simulation) may be misclassified as confirmatory external evidence rather than regulated as internal states, sustaining problematic positive-feedback loops. The analogous loop in LLMs:

$$\text{query} \rightarrow \text{gap} \rightarrow \text{prior-only completion} \rightarrow \text{emitted as answer} \rightarrow \text{user accepts} \tag{1}$$

The intervention target is the classification at the output boundary.

1.2 From Post-Hoc Detection to Pre-Output Control

Dominant mitigation strategies operate after generation: checking output against sources (Manakul et al., 2023; Min et al., 2023), training verifiers (Lightman et al., 2024), or using self-consistency voting (Wang et al., 2023). These share a structural limitation: hallucinated content has already been produced. Selective prediction (El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017) offers an alternative: a pre-output gate that blocks generation when epistemic support is insufficient.

1.3 Contributions

1. A control-theoretic framing of hallucination as output-boundary misclassification.
2. A black-box mode score using three externally measurable signals.
3. Empirical evidence from 50 items $\times$ 3 models $\times$ 4 conditions demonstrating that a composite architecture (instruction + gate) achieves 96–98% accuracy with 0–4% hallucination, representing a favorable joint accuracy–coverage tradeoff.
4. A supplementary 100-item no-context stress test derived from TruthfulQA confirming that structural gating provides a capability-independent abstention floor when instruction-following degrades.
5. Identification of the specific failure mode of structural gating (confident confabulation) and the specific failure mode of instruction-only prompting (over-cautious abstention on answerable items for GPT-4o-mini/4o; residual hallucination for GPT-3.5).

2 Theoretical Framework

2.1 Two Modes of Generation

We distinguish evidence-backed generation (output supported by user-provided data, retrieved passages, explicit computation, or citations) from prior-only generation (output from the model's learned distribution without evidential grounding in the current context). Both modes are necessary; the failure occurs when prior-only generation is emitted as if evidence-backed.

2.2 The Misclassification Loop

In biological predictive-processing models (Clark, 2013; Friston, 2010), persistent error arises when internally generated signals are classified as externally grounded evidence. The analogous failure in LLMs is the absence of a mode classifier at the output boundary: the system generates internally but has no mechanism to gate output based on whether the generation is grounded. This framing predicts that: (a) instruction-based refusal will be unreliable because it depends on the model's own classification, which is the source of the error; (b) structural gating on external signals can catch cases where the model's self-assessment fails; and (c) the two mechanisms will have complementary failure modes, motivating a composite architecture.

3 METHOD

3.1 BLACK-BOX MODE SCORE

The mode score uses only signals computable without access to model internals.

Self-consistency ($A_t \in [0, 1]$). $K = 3$ independent responses are generated; A_t is the majority-vote agreement fraction. All generations used temperature $T = 0.7$ and top-p $= 0.9$.

Paraphrase stability ($P_t \in [0, 1]$). The query is rephrased and resubmitted; P_t measures semantic overlap between original and rephrased responses.

Citation coverage ($C_t \in [0, 1]$). Fraction of content words in the response traceable to provided context. $C_t = 0$ when no context is provided. Operationally, C_t is computed via keyword overlap between response content words and the provided context text (a proxy for attribution rather than entailment).

Support deficit:

$$S_t = 1 - \frac{A_t + P_t + C_t}{3} \tag{2}$$

3.2 ABSTENTION POLICY

$$\text{output}_t = \begin{cases} \text{answer} & \text{if } S_t \leq \tau, \\ \text{ABSTAIN} & \text{if } S_t > \tau, \end{cases} \tag{3}$$

where $\tau = 0.55$.

3.3 CONDITIONS

1. **Baseline.** Standard generation, no abstention instructions.
2. **Instruction-only.** System prompt instructs abstention when evidence is insufficient. No structural enforcement.
3. **Hard-gated.** Baseline prompt; structural gate blocks output when $S_t > \tau$.
4. **Composite.** Instruction prompt plus structural gate. Output is blocked if the model's own refusal fires or $S_t > \tau$. This tests whether combining the two mechanisms covers their individual failure modes.

3.4 EVALUATION DESIGN

3.4.1 CONTROLLED REGIME EVALUATION (50 ITEMS)

Fifty items across five regimes (10 per regime):

- **R1: Answerable** ($n = 10$). Factual queries with sufficient context.
- **R2: Unanswerable** ($n = 10$). Context is empty or does not contain the answer.
- **R3: Conflicting evidence** ($n = 10$). Two sources provide contradictory answers.
- **R4: Degraded retrieval** ($n = 10$). Topically related context missing the answer.
- **R5: Adversarial pressure** ($n = 10$). Unanswerable queries with instructions demanding confident answers.

Three models spanning a range of capability levels: GPT-3.5-turbo, GPT-4o-mini, and GPT-4o.

Table 1: Overall performance across 50 items per model. Bold indicates best performance per model.

Model	Condition	Accuracy	Hallucination	Abstention
GPT-4o-mini	1. Baseline	62%	38%	42%
	2. Instruction	98%	0%	82%
	3. Gate	80%	20%	60%
	4. Composite	**98%**	**0%**	82%
GPT-4o	1. Baseline	70%	30%	50%
	2. Instruction	98%	0%	82%
	3. Gate	82%	18%	62%
	4. Composite	**98%**	**0%**	82%
GPT-3.5-turbo	1. Baseline	50%	50%	30%
	2. Instruction	94%	6%	76%
	3. Gate	76%	24%	56%
	4. Composite	**96%**	**4%**	78%

3.4.2 TruthfulQA No-Context Abstention Stress Test (100 items)

To evaluate abstention enforcement at larger scale under a uniform epistemic condition, we constructed a no-context stress test from TruthfulQA (Lin et al., 2022). We sampled 100 items from the TruthfulQA "generation" validation split using a fixed random seed (seed=123) and presented each query with an empty context field. Under this configuration, no evidence is available, and the operationally correct behavior is to abstain; any non-abstaining response is counted as a hallucination. This stress test therefore measures abstention enforcement rather than factual accuracy on TruthfulQA.

The same three models (GPT-4o-mini, GPT-4o, GPT-3.5-turbo) were evaluated under the same four conditions. Gate parameters were $K = 3$ and $\tau = 0.55$. All generations used temperature $T = 0.7$ and top-$p = 0.9$.

Abstention detection. An output was classified as abstention if the normalized response was equal to or began with `abstain`, capturing punctuation and formatting variants (e.g., "ABSTAIN.").

3.5 Metrics

- **Accuracy:** correct answer OR correct abstention.
- **Hallucination rate:** answered when should have abstained.
- **Abstention rate:** declined to answer.

For the TruthfulQA stress test, because all 100 items have `should_abstain=True`, correct% equals abstain% and hallucination% equals the rate of answering despite absent evidence.

4 Results

We report results in two parts. First, we analyze complementary failure modes in a controlled 50-item evaluation across five epistemic regimes (R1–R5). Second, we evaluate abstention enforcement under absent evidence using a 100-item no-context stress test derived from TruthfulQA.

4.1 Overall Performance (50-item Regime Evaluation)

Table 1 reports overall accuracy, hallucination, and abstention across models and conditions.

The composite condition achieved 96–98% accuracy with 0–4% hallucination across all three models. Instruction-only already achieved 0% hallucination on GPT-4o-mini and GPT-4o; the composite

Table 2: TruthfulQA 100-item no-context abstention stress test. "Correct" denotes correct abstention (output begins with `ABSTAIN`); "Halluc" denotes answering despite absent evidence.

Model	Condition	Correct	Halluc	Abstain
GPT-4o-mini	Baseline	0%	100%	0%
	Instruction	100%	0%	100%
	Gate	98%	2%	98%
	Composite	98%	2%	98%
GPT-4o	Baseline	0%	100%	0%
	Instruction	100%	0%	100%
	Gate	99%	1%	99%
	Composite	100%	0%	100%
GPT-3.5-turbo	Baseline	0%	100%	0%
	Instruction	62%	38%	62%
	Gate	100%	0%	100%
	Composite	100%	0%	100%

reduced GPT-3.5-turbo hallucination from 6% to 4% while preserving the same abstention pattern. The composite inherits the instruction component's behavior: for GPT-4o-mini and GPT-4o, this includes 10% over-cautious abstention on answerable items; for GPT-3.5-turbo, it retains some vulnerability to instruction-following failures on conflicting evidence.

4.2 TruthfulQA-derived No-Context Abstention Stress Test (100 items)

Table 2 reports results on the 100-item TruthfulQA no-context stress test. In this setting, all items require abstention; "Correct" denotes correct abstention and "Halluc" denotes answering despite absent evidence.

Baseline generation produced 0% abstention across all three models. Instruction-only abstention was capability-sensitive: GPT-4o and GPT-4o-mini abstained on all 100 items, whereas GPT-3.5-turbo abstained on 62 of 100 items and answered 38. The hard gate enforced 98–100% abstention across models regardless of instruction-following capability. The composite achieved 98–100% abstention across models, with 2% leakage on GPT-4o-mini.

The GPT-3.5-turbo results illustrate the capability dependence of instruction-based abstention: the same verbal instruction that produced 100% abstention for GPT-4o and GPT-4o-mini produced only 62% for GPT-3.5-turbo. The structural gate compensated for this gap, enforcing 100% abstention for GPT-3.5-turbo in both the gate and composite conditions.

TruthfulQA setting. TruthfulQA was evaluated in an evidence-free (empty-context) stress test; these results quantify abstention enforcement rather than answer correctness on the benchmark.

4.3 Performance by Regime

Table 3 reports composite-condition performance by regime and model.

The composite achieved 100% accuracy on Regimes 2–5 for GPT-4o-mini and GPT-4o. GPT-3.5-turbo showed residual hallucination on R1 (10%, one item answered incorrectly) and R3 (10%, one undetected conflict), consistent with its weaker instruction-following capability.

4.4 Why Neither Mechanism Alone Suffices

The central empirical finding is that the gate and instruction conditions have complementary failure modes, and only their combination achieves near-zero hallucination. Table 4 illustrates this for GPT-4o-mini.

Table 3: Composite condition accuracy and hallucination by regime across models.

	GPT-4o-mini		GPT-4o		GPT-3.5-turbo	
Regime	Acc	Hall	Acc	Hall	Acc	Hall
R1: Answerable	90%	0%	90%	0%	90%	10%
R2: Unanswerable	100%	0%	100%	0%	100%	0%
R3: Conflicting	100%	0%	100%	0%	90%	10%
R4: Degraded	100%	0%	100%	0%	100%	0%
R5: Adversarial	100%	0%	100%	0%	100%	0%

Table 4: Complementary failure modes (GPT-4o-mini). Each mechanism fails where the other succeeds.

Regime	Baseline	Instruction	Gate	Composite
R1: Answerable (accuracy)	100%	90%	100%	90%
R3: Conflicting (halluc.)	100%	0%	70%	0%
R4: Degraded (halluc.)	30%	0%	20%	0%
R5: Adversarial (halluc.)	50%	0%	10%	0%

Gate failure mode: confident confabulation. On Regime 3 (conflicting evidence), the gate hallucinated on 70% of items across GPT-4o-mini and GPT-4o. Examination of mode scores reveals the mechanism: the model selected one source's answer and produced it with high self-consistency ($A_t \geq 0.67$), high paraphrase stability ($P_t \geq 0.46$), and high citation coverage ($C_t \geq 0.26$, since the chosen answer appeared in context). The support deficit S_t remained below threshold because the model was confidently wrong—consistent, stable, and grounded in one side of the conflict.

Instruction failure mode: over-cautious abstention and residual hallucination. Instruction-only produced 10% abstention on R1 for GPT-4o-mini and GPT-4o; GPT-3.5 instead produced 10% hallucination on one answerable item (answering when it should have abstained). The gate produced 0% bad abstentions on R1 across all models, because the mode score correctly identified high A_t, P_t, and C_t on answerable items.

Composite resolution. The composite combines both mechanisms via logical OR: output is blocked if the instruction-based model refuses or $S_t > \tau$. On R3, the instruction component catches the conflicts the gate misses. On R1, the gate answers all items correctly, but the instruction component abstains on one R1 item for GPT-4o-mini and GPT-4o; the composite inherits that abstention. For GPT-3.5-turbo, the composite improves upon instruction-only by catching some conflicting-evidence cases the instruction component missed. The net effect is elimination of hallucination for GPT-4o-mini and GPT-4o across all regimes, and residual 10% hallucination on R1 and R3 for GPT-3.5, with the tradeoff of inherited abstention patterns from the instruction component.

4.5 Cross-Model Consistency

The pattern is stable across models. All three models show: (a) baseline hallucination of 30–50%; (b) instruction-only reducing overall hallucination to 0% for GPT-4o-mini and GPT-4o and to 6% for GPT-3.5-turbo, with a 10% over-cautious abstention rate on answerable items for GPT-4o-mini/4o and one R1 answerable-item hallucination for GPT-3.5-turbo; (c) the gate alone reducing hallucination to 18–24% with 0% bad abstentions on answerable items; and (d) the composite achieving 0–4% hallucination overall. The consistency across this capability range (from GPT-3.5-turbo to GPT-4o) suggests the pattern is robust within the tested model family, though architecture-level generality remains to be established.

GPT-3.5-turbo showed slightly worse performance: instruction-only hallucinated on 6% of items (vs. 0% for the larger models), and the composite retained 4% residual hallucination. These failures

were concentrated in the conflicting-evidence regime, where weaker instruction-following allowed the model to bypass the verbal abstention instruction more often than GPT-4o-mini/4o.

The TruthfulQA stress test reinforces this pattern. Under a uniform no-context condition, instruction-only abstention degraded from 100% (GPT-4o, GPT-4o-mini) to 62% (GPT-3.5-turbo), while the structural gate maintained 98–100% abstention across all models. This confirms that instruction-based mechanisms degrade with model capability, while the structural gate provides a capability-independent abstention floor.

4.6 Hypothesis Evaluation

- **H1 (Hallucination reduction):** Supported. The composite reduced hallucination from 30–50% (baseline) to 0–4% across all models.
- **H2 (Accuracy preservation):** Partially supported. The gate alone achieved 100% R1 accuracy across all models (0% bad abstention). The composite achieved 90% for GPT-4o-mini/4o (due to instruction component's over-caution) and 90% for GPT-3.5 (due to residual hallucination).
- **H3 (Gate outperforms instruction on $\geq$1 regime):** Supported. On R1, the gate (100%) outperformed instruction (90%) on all three models.
- **H4 (Adversarial robustness):** Supported. The composite achieved 100% correct abstention on R5 across all models.
- **H5 (Capability-independent abstention floor):** Supported by the TruthfulQA stress test. The structural gate enforced 98–100% abstention across all three models, whereas instruction-only abstention was 62% for GPT-3.5-turbo.

5 Discussion

5.1 The Case for Composite Architectures

The central finding is that hallucination mitigation requires combining complementary mechanisms. Instruction-based refusal leverages the model's internal representations of evidential status—effective when those representations are accurate, but unreliable when the model is confidently wrong (R3, R4) or when instruction-following degrades (GPT-3.5-turbo). Structural gating bypasses the model's self-assessment using external signals—effective at detecting uncertainty and adversarial pressure, but blind to confident confabulation where the model produces consistent, stable, grounded-looking output from parametric memory.

The composite inherits the strengths of both: instruction-based classification catches confident confabulation (the model knows the sources conflict, even when its generation doesn't reflect this), while structural gating catches cases where instruction-following fails and provides a capability-independent safety floor.

5.2 The Confident-Confabulation Boundary

The gate's failure on Regime 3 identifies a principled boundary of black-box signal detection. Self-consistency measures internal agreement, not correspondence to external evidence. A model that consistently selects one side of a conflict produces high A_t and high P_t—indistinguishable, on these signals, from a model that correctly answers a well-grounded question.

This suggests that the black-box signal set should be extended with an explicit source-conflict detection signal: checking whether the context contains contradictory evidence, independent of the model's answer. This could be implemented as a lightweight entailment check or a structured prompt that asks the model to identify disagreements in the source material before answering.

5.3 Relation to Predictive Processing

The complementary failure modes map onto principles from predictive processing and control-theoretic models of biological inference (Clark, 2013; Friston, 2010). In these frameworks, the

misclassification loop is hardest to break when the internally generated signal closely resembles genuine external evidence. The gate fails for exactly this reason: confident confabulation mimics evidence-backed generation on all observable signals. This biological analogy is consistent with the finding that the most effective intervention combines a mode classifier (structural gate) with instruction-based source assessment—precisely the composite architecture tested here.

5.4 Practical Implications

The composite architecture requires approximately $6K + 4$ API calls per query (K consistency samples, 2 paraphrase probes, 1 gated generation, 1 instructed generation). At $K = 3$, this is 22 calls per query—a non-trivial cost that must be weighed against the application's tolerance for hallucination. For high-stakes domains (medical, legal, financial), the cost is justified. For casual conversation, it is not.

The 10% bad-abstention rate on answerable items for GPT-4o-mini and GPT-4o (inherited from the instruction component) represents a coverage-accuracy tradeoff. In deployment, this could be addressed by asymmetric thresholds: requiring both mechanisms to agree on abstention for answerable-type queries, or using the gate to override instruction-based refusal when S_t is very low.

5.5 Limitations

Scale. The controlled evaluation used 50 items per model, 10 per regime. The TruthfulQA stress test used 100 items per model under a single epistemic condition (no context). Larger-scale evaluation on established benchmarks (SQuAD 2.0, HaluEval) with greater sample sizes is needed for publication-grade statistical power.

TruthfulQA scope. TruthfulQA was used exclusively in an evidence-free (empty-context) stress test to evaluate abstention enforcement. Results quantify the rate at which each condition correctly withholds output when no evidence is available; they do not measure answer correctness on TruthfulQA and should not be interpreted as TruthfulQA benchmark performance.

Model family. All three models are from OpenAI. Cross-family evaluation (Llama, Claude, Mistral) would strengthen generalization claims. Claims about "autoregressive generation" are limited to this model family and may not generalize to other architectures (e.g., different RLHF stacks, safety tuning, or decoding defaults). Cross-family validation is required to establish architecture-level generality.

Simplified signals. Citation coverage uses keyword overlap rather than entailment. Paraphrase stability uses surface-level word overlap. Stronger signal implementations would likely improve performance. C_t is not a factuality verifier and should not be interpreted as such; it is one signal among three in a composite deficit score.

Equal weighting. The three signals are equally weighted in S_t. Learned weights calibrated on a development set could improve threshold discrimination.

Scope. Restricted to factual QA. Open-ended generation, reasoning chains, and multi-turn dialogue are not addressed.

Synthetic regime construction. All five epistemic regimes (R1–R5) were hand-constructed for controlled evaluation rather than sampled from natural query distributions. This enables precise failure-mode analysis but risks implicit prompt leakage or regime overfitting. We make no claim that these regimes approximate real-world query distributions; they are designed to isolate specific epistemic conditions.

Dependence on sampling stochasticity. Self-consistency (A_t) relies on stochastic sampling when temperature > 0; under deterministic decoding (temperature=0), A_t may collapse to 1.0 for all

items, eliminating the signal. We used $T = 0.7$ and top-$p = 0.9$ for all experiments but did not vary temperature, nucleus sampling, or decoding strategy systematically.

Cost and deployment feasibility. The composite architecture requires $\sim$22 API calls per query ($K = 3$). We did not analyze latency, conduct ablation studies on K, or evaluate deployment feasibility. This is presented as a research architecture demonstrating mechanism complementarity, not as a production-ready solution.

Instruction prompt sensitivity. Instruction-only performance depends on prompt wording, which we did not vary. The composite inherits potential fragility to instruction phrasing. No robustness analysis over paraphrased instructions or adversarial prompt engineering was conducted.

5.6 Disconfirmation Criteria

- If larger-scale replication shows the composite does not reduce hallucination below instruction-only, the gate component adds no value.
- If instruction-only achieves 0% bad abstention on answerable items across models, the gate's R1 advantage disappears.
- If the pattern does not hold across model families (Llama, Claude), the mechanism is OpenAI-specific.
- If temperature=0 baseline achieves comparable results, the multi-sample mode score is unnecessary overhead.
- If the pattern does not hold when varying instruction prompts (paraphrased, different framing), the composite's reliance on verbal instruction is fragile.
- If temperature=0 eliminates gate effectiveness (no variance for self-consistency), the multi-sample architecture is required.

6 Conclusion

We proposed that LLM hallucination is a misclassification error at the output boundary and evaluated four mitigation strategies across three models and five epistemic regimes. The central finding is that neither instruction-based refusal nor structural gating alone suffices: each has a complementary failure mode that the other covers. The composite architecture achieved 96–98% accuracy with 0–4% hallucination across models spanning a substantial capability range, suggesting the mechanism addresses a consistent pattern in autoregressive generation within the tested model family. A supplementary 100-item no-context stress test derived from TruthfulQA confirmed that the structural gate provides a capability-independent abstention floor: instruction-only abstention degraded to 62% for GPT-3.5-turbo, while the gate and composite maintained 98–100% abstention across all models. Whether this pattern persists across different architectures, training paradigms, and safety tuning regimes remains an open question.

The gate's specific failure—confident confabulation, where the model is consistently and stably wrong—identifies a principled boundary of black-box signal detection and motivates extension with source-conflict detection signals. The instruction component's specific failure—over-cautious abstention on answerable items for GPT-4o-mini/4o and residual hallucination for GPT-3.5—identifies the limits of verbal self-regulation and motivates structural override.

These complementary failure modes are consistent with the control-theoretic framing: a system that generates internal predictions requires both a mode classifier (structural gate) and instruction-based source assessment. Neither alone is a complete solution; together they approach one. We present these results as initial evidence for a composite approach within the OpenAI model family and invite replication at larger scale, across model families, and on established benchmarks.

Code and data are available in the supplementary material.

DATA AND CODE AVAILABILITY

All evaluation items for the controlled regime study (R1–R5), prompt templates, and the full run logs (CSV/JSON summaries) are available from the authors upon request during the review period. A self-contained reproducibility package (code, prompts, and logs) will be released in anonymized form upon acceptance.

REFERENCES

Tom B Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Lei Huang, Weijiang Yu, Weitao Ma, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Ziwei Ji, Nayeon Lee, Rita Frieske, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, et al. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252. Association for Computational Linguistics, 2022.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 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*. Association for Computational Linguistics, 2023.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, et al. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 2021.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

A APPENDIX

Additional experimental details, prompt templates, and item-level results are provided in the supplementary material.