When Measures Become Targets: Construct Validity Failures in Uncertainty-Based Hallucination Detection

Anonymous authors
Paper under double-blind review

Abstract

Uncertainty-based hallucination metrics use token confidence, sampled semantic agreement, and representation geometry as proxies for factual reliability. We investigate when these signals diverge from factuality or semantic consistency under controlled perturbations. We develop metric-specific diagnostic tests for First Output Token Entropy, Semantic Entropy, and EigenScore, and evaluate them across four question-answering datasets and four model configurations from three model families. Targeted prompt edits lower first-token entropy while incorrectness is preserved; factual corruption remains below the NLI-based Semantic Entropy threshold; and meaning-preserving paraphrases can increase EigenScore enough to trigger hallucination flags. A human audit supports these discrepancies in the assessed examples and shows that low NLI-estimated entropy can coexist with reduced perceived semantic consistency. We also find related discrepancies under fixed generation, prompting, and rewriting conditions without metric-directed optimization. We therefore identify limitations of the evaluated operationalizations and motivate complementing predictive benchmarks with controlled tests of factuality and semantic preservation.

1 Introduction

Uncertainty-based hallucination metrics use token confidence, sampling-based semantic consistency, or embedding geometry to estimate whether generated content is factually reliable (Farquhar et al., 2024; Yao et al., 2024; Chen et al., 2024). These signals proxy a human-interpretable construct rather than directly measuring factuality. Strong benchmark performance does not ensure alignment once the proxy becomes an optimization target. Predictive validity under ordinary sampling and robustness of the operationalization under controlled variation are therefore distinct: a score may correlate with factuality on benchmark data yet respond to changes unrelated to the construct. Assessing this alignment is especially difficult for sampled outputs: semantic consistency requires judging whether multiple generations express the same meaning rather than checking a single answer against a reference.

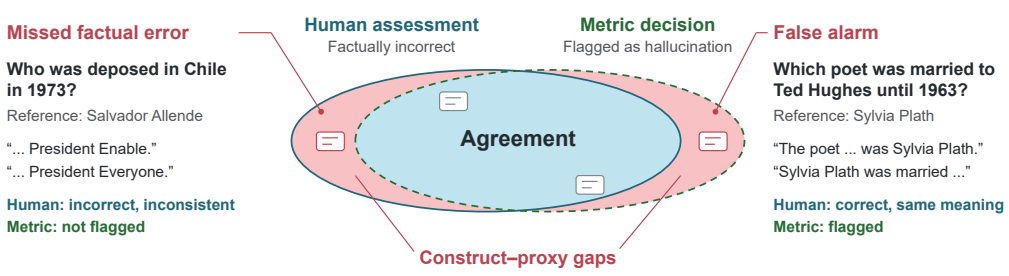

Figure 1: Construct–proxy gaps. The overlap represents correctly flagged factual errors; the non-overlapping regions show missed errors (Semantic Entropy) and false alarms (EigenScore). Response-set excerpts are shown in Tables 12, 13, and 14

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We examine uncertainty-based hallucination signals in relation to factual incorrectness and semantic consistency. We distinguish *confabulations* (Farquhar et al., 2024) — factually incorrect outputs whose content varies under stochastic sampling — from stable factual errors, and evaluate both settings. We study three detection metrics spanning token confidence (First Output Token Entropy), sampled semantic consistency (Semantic Entropy), and representation geometry (EigenScore).

For a concrete example, consider the correct response: "The poet who was married to Ted Hughes until her death in 1963 was Sylvia Plath." Rephrasing it as "Sylvia Plath was married to Ted Hughes until her death in 1963" preserves the answer's factual content and meaning. Yet, in our EigenScore experiments, a response set containing such meaning-preserving paraphrases crosses the metric's hallucination threshold, whereas the original response set does not (Table 14). The metric's decision changes even though the responses remain correct and semantically equivalent.

The gap between operational signals and the properties they are intended to measure reflects a broader problem in NLP evaluation: abstract properties are often approximated through operational correlates whose validity depends on design assumptions (Xiao et al., 2023). Goodhart's Law (Manheim & Garrabrant, 2019) captures this risk: optimizing a measure can make it less informative about the construct it is intended to represent. Targeted perturbations can expose weaknesses in this alignment, while ordinary generation and routine variation can reveal whether related failures recur without deliberate optimization. We therefore ask:

RQ1 (Targeted divergence). Can targeted perturbations expose gaps between uncertainty-based hallucination metrics and human judgements of factuality or semantic consistency?

RQ2 (Recurrence under ordinary variation). Do these gaps also arise across models and datasets under ordinary generation and routine changes in prompting or wording, without metric-directed optimization?

To investigate RQ1, we design metric-specific perturbations around the assumptions underlying each metric. We test whether First Output Token Entropy can fall below its threshold while incorrectness is preserved, whether NLI-estimated Semantic Entropy can remain low after factual corruption, and whether EigenScore can flag correct, meaning-preserving paraphrases. To investigate RQ2, we study fixed changes in generation, prompting, and expression across models and datasets without metric-directed search.

We use *attack* diagnostically to denote a constrained perturbation targeting a metric while preserving or changing factuality according to the mode. We do not study attacks on deployed systems; Attack Success Rate therefore measures proxy–construct divergence under the diagnostic protocol.

Our contributions are:

- We frame uncertainty-based hallucination metrics as operational proxies and examine their construct validity under targeted and ordinary variation.
- We propose a threat model linking metric-specific proxy–construct gaps to structural assumptions and guiding diagnostic perturbations.
- We evaluate three detection metrics across four model configurations and four QA datasets under targeted perturbations as well as ordinary-variation studies, examining proxy–construct divergence with and without metric-directed optimization.
- We show that EigenScore can flag factually correct and meaning-preserving paraphrases when their representational dispersion increases, illustrating a construct-validity failure distinct from factual incorrectness.

2 Related Work

Uncertainty-Based Hallucination Proxies Hallucination in LLMs has been studied extensively, with prior work distinguishing between *factuality hallucinations*, where outputs contradict verifiable world knowledge, and *faithfulness hallucinations*, where outputs diverge from the provided context or instructions (Huang et al., 2025). Our study concerns uncertainty-based detection of factual errors, with particular attention to the distinction between factual incorrectness and inconsistency across sampled responses (Farquhar et al., 2024).

Existing hallucination detection methods broadly fall into retrieval-based, model-based, and uncertainty-based approaches (Huang et al., 2025). The detection metrics studied here belong to the uncertainty-based family. **First Output Token Entropy** (Yao et al., 2024) estimates hallucination likelihood from uncertainty in the first generated token distribution. Sampling-based approaches such as **SelfCheckGPT** (Manakul et al., 2023) compare an original response with independently sampled responses, treating greater inconsistency as evidence of hallucination. **Semantic Entropy** (Farquhar et al., 2024) likewise uses multiple generations but groups them by meaning through NLI-based clustering. **EigenScore** (Chen et al., 2024) instead estimates hallucination likelihood from geometric dispersion in the model's internal embedding space. Although these methods operationalize hallucination differently, they all rely on measurable uncertainty proxies rather than direct access to the underlying human semantic construct.

Recent work has begun to question the reliability of such metrics even under natural conditions. Janiak et al. (2025) show that several hallucination detection metrics are often outperformed by simple length-based heuristics, while Kulkarni et al. (2025) find weak agreement between many hallucination metrics and human judgements.

Evaluation Validity and Metric Gaming Recent work argues that benchmark scores and evaluation metrics should not be interpreted as direct measurements of abstract constructs such as reasoning, factuality, or safety, but rather as operational proxies whose validity depends on the assumptions encoded in their design (Xiao et al., 2023; Bowman & Dahl, 2021; Jacobs & Wallach, 2021). Related concerns appear in work on shortcut learning (Geirhos et al., 2020), benchmark saturation, and specification gaming in ML systems (Amodei et al., 2016; Roth et al., 2024; Wiener, 1960).

Proxy–construct failures have been documented across NLP evaluation settings. Callison-Burch et al. (2006) showed that Machine Translation systems could improve BLEU scores through surface-level strategies that did not improve translation quality, while analogous fragility has been observed for summarization metrics (Fabbri et al., 2021). Recent work has further examined this fragility through diagnostic perturbations designed to expose mismatches between metric behaviour and human judgement, such as DEMETR for MT evaluation (Karpinska et al., 2022).

We extend this validity perspective to uncertainty-based factuality metrics under ordinary and targeted variation.

3 Operationalization Framework and Threat Model

Throughout the paper, we refer to First Output Token Entropy, Semantic Entropy, and EigenScore as hallucination detection metrics. Each produces an operational score that is thresholded to obtain a metric-based hallucination decision.

These metrics operationalize the target construct through measurable proxies. We formalize metric-specific proxy–construct gaps and study how targeted optimization exposes them, distinguishing the human-interpretable construct from the operational score used to approximate it.

Let z denote a metric input: a prompt–response pair or sampled response set. We distinguish between:

- a target factuality construct $C(z) \in \{0, 1\}$, representing whether z would be judged as factually incorrect under the operational criteria used in this work,
- and a metric-specific operational score $D_m(z) \in \mathbb{R}$, where larger values indicate greater estimated hallucination likelihood under detection metric m.

Detection metric m predicts hallucination according to threshold τ_m:

$$\hat{C}_m(z) = \mathbf{1}[D_m(z) > \tau_m].$$

We use the following perturbation patterns to identify candidate proxy–construct gaps:

$$C(z') = C(z) \quad \text{while} \quad D_m(z') \not\approx D_m(z),$$

or conversely,

$$D_m(z') \approx D_m(z) \quad \text{while} \quad C(z') \neq C(z).$$

These score patterns are diagnostic signatures rather than sufficient evidence of invalidity. Their interpretation depends on the mode-specific decision criteria and on separate assessments of factual correctness and semantic preservation.

Here, *threat model* denotes the objectives and constraints of each diagnostic perturbation, which tests metric–construct alignment under targeted optimization.

3.1 Diagnostic Objectives

Mode 1 operates on the first-step next-token distribution induced by prompt x, whereas Modes 2 and 3 operate on a stochastically sampled response set $\mathcal{Y}(x) = \{y_1, \ldots, y_k\}$. We study three attack modes:

- **Mode 1 (False Negative / Evasion).** Given x with incorrect $y(x)$ and $D_1(x) > \tau_1$, find x' such that $y(x')$ remains incorrect while $D_1(x') \leq \tau_1$. The attack exploits the fact that first-token entropy measures local decoding confidence rather than downstream factual correctness.
- **Mode 2 (Factual Corruption under Low NLI-Estimated Semantic Entropy).** Given an eligible response set $\mathcal{Y}(x)$ with $D_2(\mathcal{Y}(x)) \leq \tau_2$, construct a response set $\mathcal{Y}'$ whose responses are factually incorrect while still satisfying $D_2(\mathcal{Y}') \leq \tau_2$. This mode tests whether the implemented NLI-based Semantic Entropy pipeline can assign a low score to a factually corrupted response set. It does not by itself establish that the modified responses are semantically consistent; human evaluation provides separate evidence about perceived consistency.
- **Mode 3 (False Positive / Framing).** Given x with correct $\mathcal{Y}(x)$ and $D_3(\mathcal{Y}(x)) \leq \tau_3$, construct $\mathcal{Y}'$ that remains factually correct and semantically equivalent to $\mathcal{Y}(x)$, yet satisfies $D_3(\mathcal{Y}') > \tau_3$. The attack exploits the fact that representational dispersion is not identical to semantic divergence.

Modes 1 and 3 test threshold-based proxy–construct divergence, whereas Mode 2 tests whether factual corruption can remain associated with a low NLI-estimated semantic-entropy score. In practice, the target construct C is not directly observable, and we approximate it through operational criteria for factual correctness and semantic equivalence in Section 5. Under this interpretation, Attack Success Rate (ASR) measures how often optimization pressure induces proxy–construct divergence.

3.2 Access Assumptions for Diagnostic Evaluation

We assume an evaluator auditing a pipeline they control, with access already required to score the metric. ASR therefore measures proxy–construct divergence rather than external evasion risk.

4 Methodology: Targeted Tests of Proxy–Construct Alignment

To answer RQ1, we instantiate three metric-specific perturbation strategies corresponding to the diagnostic objectives defined in Section 3.1.

4.1 Mode 1: Reducing First Output Token Entropy While Preserving Incorrectness

Mode 1 optimizes first-token entropy while preserving factual incorrectness. Given a prompt x whose low-temperature reference answer is factually incorrect and whose metric score exceeds τ_1, we seek a perturbed prompt $\tilde{x}$ such that the model's answer remains incorrect while $D_1(\tilde{x}) \leq \tau_1$.

To test whether first-token confidence can be separated from factuality, we use constrained beam search over local prompt edits while preserving factual incorrectness. Candidates are retained only when they remain plausible and produce a lower FTE score, and the lowest-entropy valid candidate is selected. The edit operators, filtering rules, and beam-search settings are reported in Appendix A.6.

4.2 Mode 2: Factual Corruption under Low NLI-Estimated Semantic Entropy

Mode 2 constructs a modified response set $\mathcal{Y}'$ that is factually incorrect while satisfying $D_2(\mathcal{Y}') \leq \tau_2$, starting from an eligible clean set. Because the NLI-based pipeline contributes to candidate filtering and scoring, low Semantic Entropy is interpreted as a property of the operationalization rather than independent evidence of semantic equivalence. We assess human consistency separately.

To introduce factual errors while retaining a low NLI-estimated Semantic Entropy score, we construct locally perturbed response sets by replacing key answer tokens with embedding-neighbour candidates. Candidates are retained when they alter factual content and the resulting set remains below the calibrated Semantic Entropy threshold. Full candidate-pool, filtering, and decoding settings are reported in Appendix A.6.

4.3 Mode 3: Increasing Representational Dispersion Under Meaning Preservation with EigenScore

Mode 3 optimizes representational dispersion while preserving semantic equivalence and factual correctness. Starting from a correct response set below the EigenScore threshold, we construct semantically equivalent rewrites whose representational dispersion exceeds the calibrated threshold.

We filter candidate rewrites for factual correctness, semantic equivalence, plausibility, and semantic-cluster consistency, then select the highest-scoring valid response set. Rewriting and search details appear in Appendices A.5 and A.6.

5 Experiments

The experiments address RQ1 through targeted metric-specific perturbations and RQ2 through ordinary-variation studies. We interpret metric decisions alongside factual-correctness and, where required, semantic-preservation checks: a change in score alone does not establish a detection failure.

5.1 Datasets

We evaluate on four question-answering (QA) datasets: TriviaQA (Joshi et al., 2017), SQuAD (Rajpurkar et al., 2016), Natural Questions (NQ) (Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2023), using TriviaQA in SQuAD format. Their reference answers support factual-correctness assessment in free-form generation.

Sample sizes and eligibility vary by protocol; ordinary-variation studies use the same model–dataset combinations, with full coverage in Appendix A.7.

5.2 Models

We evaluate four instruction-tuned model configurations from three families: LLaMA-2-Chat (Touvron et al., 2023) at 7B and 13B parameters, Mistral-Instruct-v0.2 (Jiang et al., 2023) at 7B parameters, and Qwen2-Instruct (Yang et al., 2024) at 72B parameters. We evaluate each configuration with all metrics in both studies.

5.3 Prompting and Generation

Targeted experiments use a fixed few-shot prefix to elicit full-sentence answers under a consistent prompting format. Semantic Entropy and EigenScore use stochastic response sets. Correctness is assessed at the unit specified by each protocol: Mode 1 uses a separate low-temperature reference response, whereas Modes 2 and 3 apply the correctness criterion to the individual responses in the stochastic set. Prompts and decoding configurations appear in Appendix A.1 and Appendix A.2.

5.4 Detection Calibration

For each metric and model–dataset combination, we calibrate a threshold τ_m by maximizing balanced accuracy. First Output Token Entropy uses a separate calibration set; Semantic Entropy and EigenScore use the corresponding study's clean evaluation responses rather than a held-out split. Scores above τ_m are treated as hallucinations. Thresholds remain fixed across targeted and ordinary-variation conditions. Semantic Entropy uses DeBERTa-large-MNLI (He et al., 2021) for semantic clustering. Details and clean baselines appear in Appendix A.2 and Appendix A.10.

5.5 Evaluation Metrics

Attack Success Rate (ASR) is the fraction of eligible attempts satisfying the mode-specific criteria in Section 3.1. Modes 1 and 3 require threshold crossings under their respective preservation constraints; Mode 2 requires all modified responses to be incorrect while Semantic Entropy remains at or below its threshold. Factual correctness is assessed using lexical answer matching with an NLI-based equivalence check; Mode 3 additionally requires semantic-cluster consistency (Appendix A.4). We also report mean scores on clean and transformed examples, since ASR depends on eligibility, calibrated thresholds, and mode-specific correctness and preservation criteria.

5.6 Ordinary-Variation Studies: Generation and Rewriting

To answer RQ2, we study ordinary generation and routine changes in prompting and expression using the same models and datasets as the targeted experiments, without metric-directed search or candidate selection.

We evaluate ordinary-generation Semantic Entropy misses, extraction and paraphrasing for Semantic Entropy and EigenScore, and alternative answering instructions for FTE. Flagged-to-unflagged and unflagged-to-flagged transitions are reported separately against baselines using fixed calibrated thresholds. FTE uses a separate calibration set; SE and ES use clean evaluation responses rather than a held-out split. Transition rates are therefore protocol-specific, not held-out estimates. A transition alone does not establish a detection error: FTE requires unchanged correctness, ES additionally requires semantic preservation, and SE rewrite diagnostics concern the processed reference answer. Appendix A.7 details instructions, decoding conditions, sensitivity settings, transition denominators, and automated checks.

6 Human Evaluation

Because the target construct C is not directly observable, two annotators assessed response correctness and within-set semantic consistency on 150 targeted examples per metric. The full procedure is reported in Appendix A.12. The audit assesses whether metric decisions disagree with human judgements of correctness and within-set semantic consistency in the sampled targeted examples.

Results: Semantic Entropy. In the audited examples, correctness drops from approximately 92% on clean responses to under 1% on targeted responses, while mean consistency falls from 3.9 ± 0.3 to 1.5–2.2. Thus, the audited targeted sets were judged substantially less correct and less consistent than the clean sets. These results provide human-evaluation evidence of a discrepancy between low Semantic Entropy and human judgements, but do not establish that every successful attack produces the same pattern.

Results: EigenScore. In the audited examples, correctness rates are nearly identical between clean and targeted sets (92% vs. 91%), and mean consistency ratings differ by less than 0.2 points (clean: 3.9 ± 0.3; targeted: 3.7–3.8 ± 0.4). The audited targeted sets were therefore judged broadly comparable to the clean sets in correctness and consistency, despite the increase in EigenScore. This provides human-evaluation evidence of a false-positive discrepancy for the assessed examples, without implying that all successful attacks have identical effects. However, low inter-annotator agreement on EigenScore consistency limits the strength of conclusions about semantic preservation.

Model	Dataset	ASR FTE	ASR SE	ASR ES
LLaMA-2-Chat (7B)	TriviaQA	0.92	0.713	0.99
LLaMA-2-Chat (7B)	SQuAD	1.00	0.815	0.98
LLaMA-2-Chat (7B)	NQ-open	0.72	0.773	1.00
LLaMA-2-Chat (7B)	PopQA	0.64	0.715	0.90
LLaMA-2-Chat (13B)	TriviaQA	0.62	0.703	0.99
LLaMA-2-Chat (13B)	SQuAD	0.21	0.799	1.00
LLaMA-2-Chat (13B)	NQ-open	0.66	0.644	0.89
LLaMA-2-Chat (13B)	PopQA	0.70	0.766	0.99
Mistral-Instruct-v0.2 (7B)	TriviaQA	1.00	0.679	0.82
Mistral-Instruct-v0.2 (7B)	SQuAD	1.00	0.746	0.66
Mistral-Instruct-v0.2 (7B)	NQ-open	0.98	0.709	0.83
Mistral-Instruct-v0.2 (7B)	PopQA	1.00	0.700	0.98
Qwen2-Instruct (72B)	TriviaQA	1.00	0.691	0.86
Qwen2-Instruct (72B)	SQuAD	0.94	0.433	0.42
Qwen2-Instruct (72B)	NQ-open	0.92	0.546	0.88
Qwen2-Instruct (72B)	PopQA	1.00	0.265	0.97

Table 1: Attack Success Rate (ASR) across all model–dataset combinations. ASR is the fraction of eligible attempts satisfying the mode-specific criterion: FTE requires a threshold crossing to at most τ_1 while the low-temperature reference answer remains incorrect; SE requires all attacked responses to be incorrect while the response set remains at or below τ_2; and ES requires a score above τ_3 while correctness and semantic equivalence are preserved. Values are conditional on eligibility and are not directly comparable as a robustness ranking.

Results: First Output Token Entropy. In the audited stochastic responses, correctness remains low (13–17% clean; 10–12% targeted). Mode 1 eligibility and success instead depend on a separate low-temperature reference response. Mean intra-set consistency is also largely unchanged ($2.5 \pm$ 1.1–1.2), suggesting that the intervention does not substantially alter response-set coherence in the audited examples. The audit therefore supports persistent incorrectness among sampled responses, while the false-negative criterion is evaluated on the separate reference response.

7 Results

Across all three modes, targeted perturbations produce metric–factuality divergence under mode-specific preservation criteria. Table 1 reports ASRs across four model configurations and four datasets, with successful perturbations in every metric–model–dataset combination. The following subsections examine each mode, distinguishing score shifts from threshold-dependent success rates.

7.1 Mode 1: False Negatives from First Output Token Entropy

Targeted perturbations can lower first-output-token entropy while preserving an incorrect answer. Mode 1 achieves ASRs between 0.21 and 1.00 (Qwen TriviaQA: 1.00 from 13 eligible attempts), showing that first-token predictability can be increased while factual incorrectness is preserved. The result reflects FTE's local operationalization: prompt-level changes can make the first output token more predictable without correcting the factual content of the generated answer. Score summaries appear in Table 5; an illustration of the associated score reduction appears in Table 12.

These differences illustrate why score shifts and threshold crossings should be distinguished. For example, LLaMA-2-Chat (13B) on SQuAD has ASR 0.21 despite a decrease in mean FTE from 0.34 to 0.06. Its low calibrated threshold makes successful crossings difficult. Across configurations, prompt-local optimization increases first-token predictability without correcting the answer; ASR therefore reflects both the score reduction and the calibrated decision boundary, rather than how often the model is initially wrong.

7.2 Mode 2: Factual Corruption under Low Semantic Entropy

Targeted perturbations can preserve low Semantic Entropy while turning response sets factually incorrect. Mode 2 achieves ASRs between 0.265 and 0.815. For LLaMA and Mistral, success rates remain broadly comparable across datasets. Qwen shows lower attack success on SQuAD, NQ-open, and PopQA, with ASRs of 0.433, 0.546, and 0.265, respectively, compared with approximately 0.64–0.82 for the smaller models. Its TriviaQA ASR of 0.691 remains comparable to theirs.

As Mode 2 success requires every attacked response to be assessed as incorrect, lower ASR may reflect difficulty satisfying the corruption criterion and should not be interpreted directly as greater metric stability. Detailed per-run statistics are reported in Appendix A.8.

Across models and datasets, successful attacks show the same pattern: response sets can be factually incorrect while receiving low Semantic Entropy scores. Human evaluation qualifies this result: in the assessed examples, consistency ratings also decline, so the attack does not establish that the corrupted responses remain semantically consistent. This implementation-level failure is distinct from the separate possibility that genuinely consistent answers may still be factually incorrect. Score summaries appear in Table 6, and a representative example appears in Table 13.

7.3 Mode 3: Meaning-Preserving False Positives under EigenScore

Meaning-preserving responses can be flagged as hallucinations when their representational dispersion increases. Mode 3 achieves ASRs between 0.42 and 1.00. Attack success is generally high but varies across model–dataset combinations, and the corresponding score shifts must be interpreted together with each metric's calibrated threshold.

Variation in Mode 3 attack success is partly explained by calibrated thresholds. On Qwen SQuAD, mean EigenScore rises from -1.07 to 5.74, compared with a threshold of 5.82, while ASR is 0.42. This illustrates why mean score shifts and threshold-based success rates should be reported separately.

These results show that operationally correct, semantically preserved response sets can still cross the EigenScore threshold through representational change. They demonstrate that sensitivity to embedding geometry can expose meaning-preserving variation even when factual correctness and semantic equivalence are preserved. Score summaries appear in Table 7, and a representative example appears in Table 14.

7.4 Ordinary-Variation Studies: Proxy–Construct Gaps

Related metric–factuality discrepancies also arise under ordinary generation and routine rewriting. We address RQ2 with the protocols in Section 5.6. Table 2 reports pooled rates computed from summed counts; per-model–dataset results and ES temperature sensitivity appear in Appendix A.9.

First Output Token Entropy. Routine instruction changes can lower FTE while incorrectness is preserved. All four instruction changes produce flagged-to-unflagged transitions for answers that remain incorrect. The one-sentence instruction produces the largest transition rate (16.2%), while the other policies yield smaller effects between 6.2% and 9.0%.

Semantic Entropy. Semantic Entropy can miss incorrect answers when its NLI-based operationalization assigns low entropy to the corresponding response set. Under ordinary generation, 28.2% of incorrect reference answers remain unflagged. Routine extraction and paraphrasing also alter metric decisions, producing missed incorrect answers and false positives on correct answers.

EigenScore. Routine rewriting can change EigenScore decisions despite preserved response meaning. Paraphrasing produces meaning-preserving false positives in 7.8% of initially unflagged sets, while extraction produces the reverse failure in 5.6% of cases. The false-positive rate increases with paraphrasing temperature, indicating sensitivity to rewriting conditions.

Detection metric	Condition	Outcome	Rate
FTE	Neutral instruction	Flag removed; answer remains incorrect	0.062
FTE	Brief and direct	Flag removed; answer remains incorrect	0.090
FTE	Answer only	Flag removed; answer remains incorrect	0.065
FTE	One sentence	Flag removed; answer remains incorrect	0.162
SE	Ordinary generation	Incorrect reference answer remains unflagged	0.282
SE	Answer extraction	Flag removed; processed reference answer is incorrect	0.099
SE	Paraphrasing	Flag introduced; processed reference answer is correct	0.109
ES	Answer extraction	Flag removed; set remains not entirely correct and all rewrite meanings are preserved	0.056
ES	Paraphrasing ($T = 0.4$)	Flag introduced; set remains entirely correct and all rewrite meanings are preserved	0.078

Table 2: Pooled ordinary-variation results for First Output Token Entropy (FTE), Semantic Entropy (SE), and EigenScore (ES). Each rate reports either an ordinary-generation miss or a thresholded flag transition under the stated correctness and meaning-preservation condition. Flag-removal rates use initially flagged examples, flag-introduction rates use initially unflagged examples, and the ordinary-generation SE miss rate uses incorrect reference answers. Rates are pooled from summed numerators and denominators, are conditional on these baseline groups, and are not directly comparable across metrics or to targeted ASR.

8 Discussion

We answer RQ1 by showing that metric decisions can diverge from factuality in metric-specific ways. Low first-token entropy can accompany incorrect answers, low Semantic Entropy can accompany factually incorrect response sets, including cases where human consistency judgements are lower than the NLI-based score suggests, and high EigenScore can accompany meaning-preserving correct rewrites. These findings connect each discrepancy to the operational assumptions of the corresponding detection metric. Human evaluation supports this interpretation for the assessed examples.

We answer RQ2 by showing that related proxy–construct gaps arise under ordinary generation and routine rewriting, without metric-directed optimization. The reported rates are conditional on protocol-specific eligibility rules, automated correctness and semantic checks, and calibration procedures. Semantic Entropy and EigenScore thresholds use clean evaluation responses rather than independent held-out calibration data, and the automated correctness and semantic checks share an NLI component. Human evaluation covers a subset of targeted examples, with low inter-annotator agreement for EigenScore consistency. Our findings therefore provide diagnostic evidence for the evaluated pipelines, not estimates of deployment prevalence or a universal robustness ranking.

9 Conclusion

We examined whether uncertainty-based hallucination metrics remain aligned with factuality under targeted perturbations and ordinary variation. Targeted perturbations can induce proxy–construct divergence: incorrect answers can remain below the First Output Token Entropy or Semantic Entropy thresholds, while correct and meaning-preserving paraphrases can cross the EigenScore threshold. Related discrepancies also arise without metric-directed optimization under ordinary generation, answer extraction, paraphrasing, and alternative answering instructions. Human evaluation supports this interpretation for the assessed examples.

Uncertainty scores are operational proxies for factuality rather than factuality judgements themselves. Attack Success Rate therefore measures metric–construct divergence under the corresponding diagnostic protocol, not real-world evasion or deployment prevalence. Predictive benchmarks should consequently be complemented by direct tests of correctness and semantic preservation under controlled variation.

The reported rates are specific to the evaluated detection metrics, models, datasets, calibration procedures, and perturbation protocols. They do not establish a universal robustness ranking. Future work should extend this framework to other families of hallucination metrics and examine methods that better distinguish factual changes from harmless variation in wording and representation.

AI USE STATEMENT

In this work, we used generative AI tools for internal discussion of results, optimizing prompts later fed to LLMs during some experiments, refining of the wording of parts of our hypotheses, assistance in coding the pipelines used for the experiments, and cleaning and reformatting of datasets. We have not used generative AI tools for proposing hypotheses, assisting with translation, helping develop conceptual frameworks, and supporting qualitative and thematic data analysis. The rest of the required disclosure tasks are not applicable to this work.

Additionally, we use generative AI tools for suggestions of experimental parameters, editing code, summarizing existing literature, brainstorming, searching for information, editing the paper to improve readability, and identifying related literature.

We have reviewed all AI-assisted work. We verified that the code adjustments did not alter the results in unintended ways, we verified that the related work found is indeed relevant for our work, we verified that all optimized prompts are still true to the task at hand. We used the suggested experimental parameters as starting points and tested alternatives to assess their suitability. We made sure that after editing for readability, the core points of the paper were preserved and no contents were meaningfully altered, and we further tried to establish a consistent style throughout the paper. All LLM-assisted code was verified by the corresponding author, and the results were verified and sanity checked multiple times. All LLM-assisted re-formulations of sections of the paper were verified across many iterations of the paper drafts and final version of the paper by both authors.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work presents adversarial attacks on hallucination detection methods. We are aware that publishing attack methodology carries potential for misuse, and we wish to be explicit about intent and scope. Our attacks are not designed to circumvent deployed safety systems or to assist in spreading misinformation. The intended use is diagnostic: by identifying the structural conditions under which each metric's verdict decouples from human judgement, we aim to expose and characterise construct validity failures that would otherwise remain invisible under standard evaluation protocols.

We believe this kind of stress-testing is a necessary step toward more robust, reliable, and trustworthy hallucination detection. Detection metrics that appear to perform well under natural sampling conditions may still fail in ways that human reviewers would immediately catch; our framework makes those failures visible and quantifiable. Improving detection metric stability against the gaps we identify is a precondition for deploying such systems responsibly in downstream pipelines, where false assurance of factual correctness could propagate harmful or incorrect information to end users. The attacks described here are therefore offered as a diagnostic instrument rather than an offensive one, with the goal of raising the bar for what it means for a hallucination detection metric to genuinely measure the construct it is designed to track.

All datasets and models used in this work are publicly available for research purposes under permissive licenses (Apache 2.0, CC BY 4.0, CC BY-SA 3.0, MIT, Meta Research License); our use is consistent with their intended research use.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the experimental setup in the main paper and appendix. The main text specifies the datasets, model configurations, evaluation tasks, calibration procedure, metrics, and human-evaluation protocol. The appendix provides a representative excerpt of the few-shot prompt, decoding and generation settings, answer-matching criterion, metric-specific evaluation procedures, rewriting prompts, perturbation parameters, ordinary-variation protocols, sample sizes, and per-model–dataset results, as well as inter-annotator agreement statistics. Upon deanonymization, we will release the full source code, experiment scripts, configuration files, data-processing procedures, and materials required to reproduce the reported analyses in an accompanying repository. All datasets and model checkpoints used are publicly available, and the repository will docu-

ment the relevant software environment, random seeds, and commands for regenerating the tables and figures.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. URL https://arxiv.org/abs/1606.06565.

Samuel R. Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4843–4855, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385/.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In Diana McCarthy and Shuly Wintner (eds.), *11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 249–256, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1032/.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection, 2024. URL https://arxiv.org/abs/2402.03744.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 04 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00373. URL https://doi.org/10.1162/tacl_a_00373.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL http://dx.doi.org/10.1038/s42256-020-00257-z.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL https://arxiv.org/abs/2006.03654.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Fairness, Accountability, and Transparency (FAccT '21)*. ACM, March 2021. URL https://www.microsoft.com/en-us/research/publication/measurement-and-fairness/.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms, 2025. URL https://arxiv.org/abs/2508.08285.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. DEMETR: Diagnosing evaluation metrics for translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9540–9561, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.649. URL https://aclanthology.org/2022.emnlp-main.649/.

Atharva Kulkarni, Yuan Zhang, Joel Ruben Antony Moniz, Xiou Ge, Bo-Hsiang Tseng, Dhivya Piraviperumal, Swabha Swayamdipta, and Hong Yu. Evaluating evaluation metrics – the mirage of hallucination detection, 2025. URL https://arxiv.org/abs/2504.18114.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law, 2019. URL https://arxiv.org/abs/1803.04585.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Benjamin Roth, Pedro Henrique Luz de Araujo, Yuxi Xia, Saskia Kaltenbrunner, and Christoph Korab. Specification overfitting in artificial intelligence. *Artificial Intelligence Review*, 58(2):35, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh

Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Norbert Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410): 1355–1358, 1960.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10967–10982, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.676. URL https://aclanthology.org/2023.emnlp-main.676/.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples, 2024. URL https://arxiv.org/abs/2310.01469.

A Appendix

A.1 Few-Shot Prompt

The following is an excerpt from the prompt prefix prepended to all inputs across all three attack modes and all models. The implementation uses a 10-shot prefix; three representative demonstrations are shown here for brevity. Structural markers (`Question:` and `Answer:`) are frozen throughout all experiments.

```
Answer each question in one complete sentence
that rephrases the question naturally.

Question: What is the capital of France?
Answer: The capital of France is Paris.

Question: Who invented the telephone?
Answer: The telephone was invented by Alexander Graham Bell.

Question: Who were the first League Cup winners?
Answer: The first League Cup winners were Aston Villa.

Question: {question}
Answer:
```

A.2 Detection and Generation Settings

Shared protocol. For all three metrics, each prompt is evaluated with two distinct decoding regimes. A single *reference response* is generated at low temperature ($T = 0.1$, top-$k = 50$, top-$p = 0.9$) and is used for the Mode 1 correctness check and for clean-example eligibility where specified. A *stochastic response set* of $k = 10$ samples is drawn at high temperature ($T = 1.0$, top-$k = 50$, top-$p = 0.9$) and used as input to Semantic Entropy and EigenScore. For Modes 2 and 3, the

individual responses in this stochastic set are also evaluated using the correctness criterion required by the corresponding attack protocol.

First Output Token Entropy (Mode 1). FTE is computed via a single forward pass at the answer boundary; no sampling is required for the metric itself. The low-temperature reference ($T = 0.1$) is used to verify factual correctness of candidate prompt rewrites during beam search, with a maximum generation length of 128 tokens. Mode 1 attack success is defined using the low-temperature reference response described in Appendix A.4; additional $k = 10$ stochastic samples per attacked prompt are generated for reporting ($T = 1.0$, top-$k = 50$, top-$p = 0.9$, maximum 80 tokens).

Semantic Entropy (Mode 2). The $k = 10$ stochastic responses are NLI-clustered to form semantic equivalence classes, from which SE is computed. Maximum generation length is 128 tokens for both the reference and stochastic responses.

EigenScore (Mode 3). Response embeddings are extracted from the middle hidden layer of the target LLM. Maximum generation length is 60 tokens for both the reference and stochastic responses. The Mode 3 attack additionally generates paraphrase candidates using the same target model: 12 samples per rewrite-prompt family, with $T = 1.0$, top-$k = 50$, top-$p = 0.95$.

A.3 Answer-Matching Criterion

Factual correctness is assessed uniformly across all three attack modes using a two-stage criterion. Given a generated response $\hat{y}$ and a set of gold aliases $\mathcal{A}$ (the accepted answer strings for the question), a response is marked correct if either stage passes.

Stage 1 — Lexical overlap. Each alias $a \in \mathcal{A}$ is stripped of leading and trailing punctuation and matched against $\hat{y}$ case-insensitively. A response is correct if any alias appears as a substring of the response, or if the response appears as a substring of any alias. Aliases shorter than a minimum length are skipped.

Stage 2 — NLI fallback. If no lexical match is found, DeBERTa-large-MNLI (He et al., 2021) is used as a fallback. The model scores the hypothesis that $\hat{y}$ entails each alias $a \in \mathcal{A}$, optionally conditioned on the question. A response is marked correct if DeBERTa predicts mutual entailment for at least one alias — that is, both $\hat{y}$ entails a and a entails $\hat{y}$ must be predicted as the argmax class.

This procedure is an operational correctness criterion rather than an independent human truth oracle. In particular, DeBERTa-large-MNLI is also used for semantic clustering, so NLI-based correctness labels and semantic-cluster assignments are not independent validation signals.

A.4 Evaluation Procedure

Answer-matching criterion. The two-stage lexical-overlap and NLI-based criterion used to assess factual correctness across all three attack modes is described in Appendix A.3.

Mode 1 attack success. A Mode 1 attack on a given example is successful if (a) the low-temperature reference response generated from the attacked prompt fails the correctness criterion, and (b) the attacked prompt's first-output-token entropy H_1 is at or below the calibrated threshold τ_1.

Mode 2 attack success. The Mode 2 attack produces a modified set of $k = 10$ stochastic responses by substituting the key-answer token(s) in each original response. A Mode 2 attack is successful if (a) all k attacked responses fail the correctness criterion, and (b) the Semantic Entropy computed from the attacked response set is at or below τ_2. Candidate replacements are filtered by the NLI stage as part of the attack construction: a replacement is retained only if the attacked response does *not* mutually entail the original response. This filtering criterion helps induce factual corruption, but the resulting low Semantic Entropy score is not treated as independent evidence of semantic equivalence.

Mode 3 attack success. A Mode 3 attack on a given example is successful if all of the following constraints are jointly satisfied by the attacked response set:

- **Score constraint:** EigenScore exceeds the calibrated threshold τ_3.
- **Correctness constraint:** all k responses pass the correctness criterion.
- **Cluster constraint:** the responses form at most one semantic equivalence class under NLI-based clustering. Clusters are formed greedily: each response is assigned to the first existing cluster whose representative mutually entails it; otherwise a new singleton cluster is opened.
- **Form constraint:** all responses are well-formed full sentences (terminal punctuation, minimum five content tokens).
- **Size constraint:** the attacked set contains exactly $k = 10$ responses with at least $k_{\min}$ unique surface forms.

A.5 Mode 3 Rewriting Prompt Families

The Mode 3 candidate pool includes model-generated rewrites produced by querying the LLM under seven distinct rewriting prompts. All families share the following template structure:

```
[Family-specific instructions]
Question: {question}
Original answer: {answer}
Rewritten answer:
```

The seven families differ only in their instruction block, as listed in Table 3.

Family	Key instruction
General paraphrase	Use clearly different wording; keep factual meaning exactly the same
Active/passive	Switch between active and passive voice where possible
Clause reorder	Change word or clause order while staying natural
Shorter	Make it more compact; do not turn it into a fragment
Formal	Use a more formal writing style
Conversational	Use more natural conversational wording
Punctuation	Change wording lightly; use simple punctuation replacements

Table 3: Rewriting prompt families used in the Mode 3 candidate pool.

A.6 Full Perturbation Parameters

For Token-Entropy, we use a 10-shot prompt and optimize the full-vocabulary first-token entropy by default, while also reporting the top-10 renormalized variant. The attack uses beam width 24 and maximum depth 6, evaluates candidates with low-temperature decoding ($T = 0.1$, top-$k = 50$, top-$p = 0.9$), fully scores only the 48 most promising prescreened candidates per state, and keeps at most 2 candidates per diversity signature. On the question side, we retain at most 4 synonym candidates per step and allow at most 1 question edit; on the few-shot side, we retain up to 24 single-token variants and 8 two-token variants per step, and allow at most 7 few-shot edit steps. All substitutions are restricted to tokens of length at least 4, with a semantic similarity threshold of 0.9 and pruning slack of 0.05 relative to the clean prompt and 0.02 relative to the parent state. For reporting, we additionally sample 10 clean and 10 attacked responses with $T = 1.0$, top-$k = 50$, and top-$p = 0.9$. Edits are applied only to the question body or few-shot tokens; structural markers such as `Question:` and `Answer:` remain fixed. Candidates are ranked by full-vocabulary first-output-token entropy subject to the correctness and plausibility constraints, and the lowest-entropy valid candidate is selected.

Model	TriviaQA	SQuAD	NQ	PopQA
LLaMA-2 7B	400	400	400	600
LLaMA-2 13B	400	400	400	600
Mistral 7B	400	400	400	600
Qwen2 72B	400	800	800	800

Table 4: Evaluation examples per model–dataset combination for each ordinary-variation study. FTE calibration examples are excluded.

For Semantic Entropy, we again use a 10-shot prompt, obtaining one low-temperature answer for correctness scoring and 10 higher-temperature responses for entropy estimation. We calibrate a single semantic-entropy threshold by balanced accuracy and attack only examples that are correct, classified as non-hallucinating, and contain an identifiable answer token. The attack draws from up to 500 embedding neighbors, restricted to alphabetic words of length at least 5, and greedily selects NLI-compatible replacements that alter factual content; this filtering is part of the attack construction and does not establish semantic equivalence of the resulting set. All attacked response sets are then re-evaluated with the same semantic-entropy pipeline, and an attack is counted as successful only if the attacked responses are factually incorrect while the attacked set remains at or below the calibrated semantic-entropy threshold.

For EigenScore, we first generate 10 clean responses per example and identify eligible cases from those initially classified as non-hallucinating. For each such case, we construct a candidate pool of at most 80 responses, with at most 8 candidates from any single rewrite family, and search this pool with beam width 50 to maximize EigenScore. Candidate rewrites are generated with $T = 1.0$, top-$k = 50$, and top-$p = 0.95$, and are required to remain semantically equivalent, preserve at least half of the base-answer length, and maintain at least 40% lexical overlap with the original response. The final attacked set contains 10 responses, of which at least 5 must be unique.

A.7 Ordinary Generation and Rewriting Protocols

Coverage and sampling. Table 4 reports evaluation sample sizes for each detection metric. Each study comprises 16 model–dataset combinations and 8,200 evaluation examples. Runs use random seed 42 and $K = 10$ sampled responses per question, with baseline sampling temperature $T = 1.0$, top-$k = 50$, and top-$p = 0.9$. FTE additionally uses 100 separate calibration examples per run. The evaluation examples are reused across conditions within each study; conditions therefore do not constitute independent samples.

FTE instructions. Each policy is evaluated independently against the baseline, without few-shot demonstrations. The exact instructions are:

- **Baseline:** "Answer each question in one complete sentence that rephrases the question naturally."
- **Neutral:** "Answer the question."
- **Brief and direct:** "Answer briefly and directly."
- **Answer only:** "Give only the answer, without an explanation."
- **One sentence:** "Answer in one sentence."

The reference-answer temperature is $T = 0.01$. Correctness-conditioned transitions require the reference answer to remain incorrect or correct before and after the instruction change, but do not require identical answer text.

SE rewriting. Answer extraction and paraphrasing are evaluated as separate conditions, using temperature $T = 0.7$, top-$k = 50$, and top-$p = 0.95$. The extraction instruction is:

> Extract just the answer from the following response. Return one short natural sentence and do not add facts.

The paraphrasing instruction is:

> Rewrite the following answer while keeping the meaning the same. Keep it to one natural sentence. Do not add or remove facts.

Each prompt includes the question and the original answer. Both transition directions are recorded for each method. Correctness-qualified transitions assess the processed reference answer, without requiring its correctness status to match the baseline.

The baseline wrong-but-consistent criterion requires an incorrect reference answer whose semantic cluster contains at least 70% of the sampled responses and at least three responses. Cluster membership is determined by the NLI-based procedure; it does not require literal repetition or independently establish that every sampled response is incorrect.

ES rewriting. Each response is processed independently. Extraction uses $T = 0.2$ for LLaMA and Mistral and $T = 0.1$ for Qwen. Paraphrasing uses $T = 0.4$ as the primary condition, with $T = 0.2$ and $T = 0.8$ as sensitivity checks. All rewriting conditions use top-$k = 50$ and top-$p = 0.95$. The extraction instruction is:

> Extract the answer conveyed by the response and express it briefly and naturally. Do not add, remove, or correct facts.

The paraphrasing instruction is:

> Rewrite the response naturally while preserving exactly the same meaning. Use one natural sentence and do not add, remove, or correct facts.

Prompts include the question, original response, and context when available. The strict false-positive diagnostic requires all K responses to be correct before and after rewriting. The corresponding false-negative diagnostic requires the set to be not entirely correct before and after. Both additionally require every original–rewrite pair to pass the automated semantic-preservation check.

Calibration and aggregation. Thresholds maximize balanced accuracy on baseline data and remain fixed when measuring transitions. SE and ES use the clean evaluation responses for calibration; FTE uses its separate calibration set. Method-specific recalibrated error rates in the saved outputs are separate diagnostics and are not used for the reported transition rates.

Flag-removal rates divide by all initially flagged examples; flag-introduction rates divide by all initially unflagged examples. These denominators remain unchanged when the numerator additionally requires correctness or semantic preservation. Baseline SE miss rates instead divide by incorrect reference answers. Pooled rates are computed by summing numerators and denominators across runs, so larger runs receive greater weight. The different conditions can affect the same examples and must not be summed as unique failures.

A.8 Full Attack Performance

A.9 Ordinary-Variation Results by Model and Dataset

Tables 8–10 report ordinary-variation results by model and dataset. H denotes a detection metric flag and NH denotes no flag. Transition rates divide by all initially flagged examples for H→NH and all initially unflagged examples for NH→H, including when the numerator imposes correctness or semantic-preservation requirements. Pooled rates use summed numerators and denominators rather than averages of the displayed rates. Protocols and sample sizes are given in Appendix A.7.

A.10 Baseline Metric Performance

Table 11 reports clean detection performance for all three detection metrics across all model–dataset combinations, prior to any attack. Balanced accuracy is computed using the calibrated threshold τ_m, and AUROC is computed on the corresponding baseline data ($n = 200$ per combination). FTE uses a separate calibration set, whereas SE and ES use the clean evaluation responses specified by their protocols.

Model	Dataset	FTE Clean	FTE Targeted	ASR
LLaMA-2-Chat (7B)	TriviaQA	0.64	0.16	0.92
LLaMA-2-Chat (7B)	SQuAD	1.53	0.65	1.00
LLaMA-2-Chat (7B)	NQ-open	0.48	0.04	0.72
LLaMA-2-Chat (7B)	PopQA	0.38	0.02	0.64
LLaMA-2-Chat (13B)	TriviaQA	0.49	0.18	0.62
LLaMA-2-Chat (13B)	SQuAD	0.34	0.06	0.21
LLaMA-2-Chat (13B)	NQ-open	0.53	0.10	0.66
LLaMA-2-Chat (13B)	PopQA	0.49	0.20	0.70
Mistral-Instruct-v0.2 (7B)	TriviaQA	1.32	0.70	1.00
Mistral-Instruct-v0.2 (7B)	SQuAD	1.99	1.03	1.00
Mistral-Instruct-v0.2 (7B)	NQ-open	1.00	0.37	0.98
Mistral-Instruct-v0.2 (7B)	PopQA	1.49	0.80	1.00
Qwen2-Instruct (72B)	TriviaQA	1.19	0.67	1.00
Qwen2-Instruct (72B)	SQuAD	1.37	0.71	0.94
Qwen2-Instruct (72B)	NQ-open	0.19	2.0×10^{-4}	0.92
Qwen2-Instruct (72B)	PopQA	0.27	1.6×10^{-3}	1.00

Table 5: Mode 1 (FTE) attack results: mean full-vocabulary first-token entropy on clean and targeted examples, and Attack Success Rate (ASR). The Qwen TriviaQA row uses the $n = 400$, beam-width-36 run, with 13 eligible attack attempts.

Model	Dataset	SE Clean	SE Targeted	ASR
LLaMA-2-Chat (7B)	TriviaQA	0.008	0.148	0.713
LLaMA-2-Chat (7B)	SQuAD	0.128	0.217	0.815
LLaMA-2-Chat (7B)	NQ-open	0.166	0.242	0.773
LLaMA-2-Chat (7B)	PopQA	0.148	0.189	0.715
LLaMA-2-Chat (13B)	TriviaQA	0.000	0.127	0.703
LLaMA-2-Chat (13B)	SQuAD	0.189	0.320	0.799
LLaMA-2-Chat (13B)	NQ-open	0.179	0.211	0.644
LLaMA-2-Chat (13B)	PopQA	0.032	0.071	0.766
Mistral-Instruct-v0.2 (7B)	TriviaQA	0.021	0.201	0.679
Mistral-Instruct-v0.2 (7B)	SQuAD	0.193	0.354	0.746
Mistral-Instruct-v0.2 (7B)	NQ-open	0.347	0.482	0.709
Mistral-Instruct-v0.2 (7B)	PopQA	0.180	0.229	0.700
Qwen2-Instruct (72B)	TriviaQA	0.305	0.380	0.691
Qwen2-Instruct (72B)	SQuAD	0.841	0.942	0.433
Qwen2-Instruct (72B)	NQ-open	0.774	0.876	0.546
Qwen2-Instruct (72B)	PopQA	0.910	1.047	0.265

Table 6: Mode 2 (SE) attack results: mean Semantic Entropy on clean and targeted response sets, and Attack Success Rate (ASR). Qwen and PopQA entries use the SingleAtk results.

A.11 Qualitative Examples

Tables 12–14 illustrate score changes and response transformations under the three attack modes; their captions specify what each example establishes.

A.12 Human Evaluation: Inter-Annotator Agreement

The two annotators had different backgrounds. One works in software development and has limited familiarity with machine learning and deep learning through personal interest. The other has no formal background in computer science and interacts with LLMs primarily as a daily end user.

For each detection metric, we annotated 150 targeted examples: 50 each from TriviaQA, SQuAD, and NQ-open, drawn equally from two models. ES and SE used LLaMA-2-Chat (13B) and Mistral-Instruct-v0.2 (7B); FTE used LLaMA-2-Chat (7B) and Mistral-Instruct-v0.2 (7B). Each example contained the question, ground-truth answer, and two response sets of five responses, labelled Set A

Model	Dataset	ES Clean	ES Targeted	ASR
LLaMA-2-Chat (7B)	TriviaQA	-3.47	2.52	0.99
LLaMA-2-Chat (7B)	SQuAD	-3.53	2.04	0.98
LLaMA-2-Chat (7B)	NQ-open	-3.37	2.35	1.00
LLaMA-2-Chat (7B)	PopQA	-3.86	2.13	0.90
LLaMA-2-Chat (13B)	TriviaQA	-3.49	2.92	0.99
LLaMA-2-Chat (13B)	SQuAD	-3.49	3.09	1.00
LLaMA-2-Chat (13B)	NQ-open	-2.48	3.38	0.89
LLaMA-2-Chat (13B)	PopQA	-3.50	2.84	0.99
Mistral-Instruct-v0.2 (7B)	TriviaQA	-3.79	-0.25	0.82
Mistral-Instruct-v0.2 (7B)	SQuAD	-3.78	-0.48	0.66
Mistral-Instruct-v0.2 (7B)	NQ-open	-3.79	-0.43	0.83
Mistral-Instruct-v0.2 (7B)	PopQA	-4.17	-0.51	0.98
Qwen2-Instruct (72B)	TriviaQA	-0.90	5.99	0.86
Qwen2-Instruct (72B)	SQuAD	-1.07	5.74	0.42
Qwen2-Instruct (72B)	NQ-open	-1.21	5.76	0.88
Qwen2-Instruct (72B)	PopQA	-1.78	5.25	0.97

Table 7: Mode 3 (ES) attack results: mean EigenScore on clean and targeted response sets, and Attack Success Rate (ASR), across all model–dataset combinations.

Model	Dataset	Baseline miss rate	Extraction: H→NH	Paraphrase: NH→H
LLaMA-2-Chat (7B)	TriviaQA	0.169	0.000	0.248
LLaMA-2-Chat (7B)	SQuAD	0.195	0.028	0.078
LLaMA-2-Chat (7B)	NQ-open	0.311	0.058	0.046
LLaMA-2-Chat (7B)	PopQA	0.236	0.003	0.054
LLaMA-2-Chat (13B)	TriviaQA	0.321	0.031	0.435
LLaMA-2-Chat (13B)	SQuAD	0.361	0.149	0.061
LLaMA-2-Chat (13B)	NQ-open	0.191	0.048	0.349
LLaMA-2-Chat (13B)	PopQA	0.287	0.036	0.130
Mistral-Instruct-v0.2 (7B)	TriviaQA	0.327	0.045	0.064
Mistral-Instruct-v0.2 (7B)	SQuAD	0.283	0.108	0.046
Mistral-Instruct-v0.2 (7B)	NQ-open	0.389	0.093	0.029
Mistral-Instruct-v0.2 (7B)	PopQA	0.277	0.062	0.029
Qwen2-Instruct (72B)	TriviaQA	0.250	0.052	0.218
Qwen2-Instruct (72B)	SQuAD	0.248	0.211	0.033
Qwen2-Instruct (72B)	NQ-open	0.354	0.218	0.035
Qwen2-Instruct (72B)	PopQA	0.297	0.201	0.013
Pooled		0.282	0.099	0.109

Table 8: Ordinary-variation Semantic Entropy results. Baseline miss rate is the fraction of incorrect reference answers that remain unflagged. Extraction transitions require the processed reference answer to be incorrect; paraphrasing transitions require it to be correct. These rewrite diagnostics assess correctness after processing and do not require unchanged correctness.

and Set B; the clean and targeted sets were randomly shuffled between these labels for each example. Two annotators judged response correctness and rated within-set semantic consistency on a five-point scale.

Inter-annotator agreement is assessed via Cohen's κ and Krippendorff's α for correctness, and quadratic-weighted κ and ordinal α for consistency. For SE targeted correctness, the near-zero correct rate induces a floor effect, so the low κ should be interpreted alongside the near-perfect raw agreement and the highly imbalanced labels. Agreement on consistency is strong for FTE and SE, but low for ES: the ES consistency values are $\kappa_w = 0.099$ and $\alpha = 0.168$ for clean sets, and $\kappa_w = 0.103$ and $\alpha = 0.027$ for targeted sets. Accordingly, the ES consistency results should be interpreted cautiously. Full IAA statistics are given in Table 15.

The annotation instructions written into each sheet, and an excerpt of the annotation sheet itself, are provided below.

Model	Dataset	Extraction H→NH	Paraphrasing: NH→H $T = 0.2$	$T = 0.4$	$T = 0.8$
LLaMA-2-Chat (7B)	TriviaQA	0.113	0.071	0.159	0.282
LLaMA-2-Chat (7B)	SQuAD	0.068	0.050	0.059	0.109
LLaMA-2-Chat (7B)	NQ-open	0.042	0.025	0.037	0.124
LLaMA-2-Chat (7B)	PopQA	0.049	0.004	0.008	0.056
LLaMA-2-Chat (13B)	TriviaQA	0.146	0.048	0.079	0.096
LLaMA-2-Chat (13B)	SQuAD	0.059	0.046	0.061	0.099
LLaMA-2-Chat (13B)	NQ-open	0.141	0.040	0.075	0.069
LLaMA-2-Chat (13B)	PopQA	0.041	0.012	0.016	0.027
Mistral-Instruct-v0.2 (7B)	TriviaQA	0.057	0.115	0.202	0.385
Mistral-Instruct-v0.2 (7B)	SQuAD	0.047	0.036	0.096	0.181
Mistral-Instruct-v0.2 (7B)	NQ-open	0.084	0.062	0.093	0.160
Mistral-Instruct-v0.2 (7B)	PopQA	0.054	0.019	0.033	0.105
Qwen2-Instruct (72B)	TriviaQA	0.041	0.066	0.162	0.304
Qwen2-Instruct (72B)	SQuAD	0.021	0.033	0.066	0.114
Qwen2-Instruct (72B)	NQ-open	0.030	0.030	0.056	0.097
Qwen2-Instruct (72B)	PopQA	0.022	0.042	0.074	0.129
Pooled		0.056	0.042	0.078	0.142

Table 9: Ordinary-variation EigenScore results under strict correctness and semantic-preservation checks. Extraction requires the response set to remain not entirely correct before and after processing; paraphrasing requires all responses to remain correct. Every original–rewrite pair must pass the semantic-preservation check. The primary paraphrasing temperature is $T = 0.4$; the other temperatures are sensitivity checks.

Model	Dataset	H→NH: remains incorrect N	B	A	S	NH→H: remains correct N	B	A	S
LLaMA-2-Chat (7B)	TriviaQA	0.035	0.080	0.027	0.088	0.247	0.143	0.105	0.167
LLaMA-2-Chat (7B)	SQuAD	0.178	0.327	0.218	0.356	0.060	0.030	0.047	0.023
LLaMA-2-Chat (7B)	NQ-open	0.136	0.318	0.250	0.318	0.076	0.045	0.042	0.034
LLaMA-2-Chat (7B)	PopQA	0.000	0.023	0.013	0.087	0.217	0.130	0.116	0.072
LLaMA-2-Chat (13B)	TriviaQA	0.009	0.015	0.006	0.054	0.397	0.324	0.441	0.103
LLaMA-2-Chat (13B)	SQuAD	0.078	0.108	0.150	0.545	0.090	0.069	0.077	0.004
LLaMA-2-Chat (13B)	NQ-open	0.017	0.029	0.046	0.249	0.264	0.252	0.220	0.082
LLaMA-2-Chat (13B)	PopQA	0.167	0.200	0.133	0.500	0.072	0.047	0.053	0.011
Mistral-Instruct-v0.2 (7B)	TriviaQA	0.061	0.034	0.073	0.057	0.246	0.283	0.116	0.261
Mistral-Instruct-v0.2 (7B)	SQuAD	0.134	0.141	0.056	0.204	0.060	0.060	0.129	0.060
Mistral-Instruct-v0.2 (7B)	NQ-open	0.252	0.245	0.202	0.288	0.030	0.046	0.101	0.051
Mistral-Instruct-v0.2 (7B)	PopQA	0.158	0.158	0.211	0.137	0.026	0.028	0.034	0.022
Qwen2-Instruct (72B)	TriviaQA	0.000	0.143	0.143	0.143	0.005	0.046	0.013	0.020
Qwen2-Instruct (72B)	SQuAD	0.115	0.213	0.126	0.336	0.102	0.117	0.127	0.076
Qwen2-Instruct (72B)	NQ-open	0.095	0.165	0.087	0.190	0.163	0.244	0.201	0.172
Qwen2-Instruct (72B)	PopQA	0.000	0.003	0.000	0.010	0.300	0.300	0.200	0.300
Pooled		0.062	0.090	0.065	0.162	0.103	0.100	0.093	0.066

Table 10: Ordinary-variation FTE results for four instruction policies: N = neutral, B = brief and direct, A = answer only, S = one sentence. Each policy is compared independently with the same baseline. Correctness status must be preserved before and after the instruction change, but answer text need not be identical. Rates are conditional on the initial detection metric decision, not on initial correctness.

		FTE		SE		ES	
Model	**Dataset**	**Bal. Acc.**	**AUROC**	**Bal. Acc.**	**AUROC**	**Bal. Acc.**	**AUROC**
LLaMA-2-Chat (7B)	TriviaQA	0.656	0.655	0.749	0.789	0.689	0.742
LLaMA-2-Chat (7B)	SQuAD	0.551	0.490	0.758	0.809	0.779	0.789
LLaMA-2-Chat (7B)	NQ-open	0.614	0.592	0.682	0.712	0.732	0.730
LLaMA-2-Chat (7B)	PopQA	0.604	0.610	0.770	0.833	0.861	0.898
LLaMA-2-Chat (13B)	TriviaQA	0.615	0.617	0.727	0.745	0.677	0.708
LLaMA-2-Chat (13B)	SQuAD	0.533	0.486	0.747	0.779	0.850	0.875
LLaMA-2-Chat (13B)	NQ-open	0.550	0.520	0.691	0.698	0.732	0.765
LLaMA-2-Chat (13B)	PopQA	0.563	0.541	0.816	0.857	0.833	0.866
Mistral-Instruct-v0.2 (7B)	TriviaQA	0.620	0.631	0.750	0.761	0.679	0.686
Mistral-Instruct-v0.2 (7B)	SQuAD	0.527	0.474	0.703	0.757	0.754	0.778
Mistral-Instruct-v0.2 (7B)	NQ-open	0.551	0.546	0.700	0.737	0.704	0.738
Mistral-Instruct-v0.2 (7B)	PopQA	0.612	0.587	0.811	0.859	0.888	0.919
Qwen2-Instruct (72B)	TriviaQA	0.611	0.568	0.783	0.804	0.696	0.728
Qwen2-Instruct (72B)	SQuAD	0.541	0.509	0.729	0.759	0.734	0.747
Qwen2-Instruct (72B)	NQ-open	0.532	0.467	0.724	0.718	0.739	0.723
Qwen2-Instruct (72B)	PopQA	0.586	0.590	0.810	0.867	0.806	0.858

Table 11: Baseline performance for FTE, SE, and EigenScore across all model–dataset combinations ($n = 200$ per combination). FTE = First Output Token Entropy, SE = Semantic Entropy, ES = EigenScore.

Clean responses *(FTE = 0.800)*	**Targeted responses** *(FTE = 0.314)*
Animals use oxygen for respiration to extract energy from food.	Oxygen is used by animal life for respiration, or the process of breathing, to extract energy from food and survive.
Animals use oxygen for respiration, which is the process of breathing.	Oxygen is used by animal life for respiration, which is the process of taking in air and expelling carbon dioxide to maintain life.
Animals use oxygen for respiration to release energy from food and live.	Oxygen is used by animal life for respiration, which is the process of taking in air and releasing carbon dioxide.
Oxygen is used by animal life for respiration.	Oxygen is used by animal life for respiration, specifically for the metabolic process that releases energy and carbon dioxide.
Animals use oxygen for respiration, which is the process of taking in and using air to release energy from food.	Oxygen is used by animal life for respiration, which is the process of taking in air and expelling carbon dioxide.
Question: For what purpose is oxygen used by animal life?	
Ground truth: cellular respiration	

Table 12: Mode 1 qualitative example (5 of 10 responses shown). The example illustrates a reduction in first-token entropy under prompt editing. The displayed stochastic responses do not by themselves establish factual incorrectness; attack success is determined using the separate low-temperature reference response and criteria in Appendix A.4.

Clean responses *(SE = -0.000)*	**Targeted responses** *(SE = -0.000)*
An American aided coup in Chile in 1973 resulted in the deposition of President Salvador Allende.	An American aided coup in Chile in 1973 resulted in the deposition of President Enable.
An American aided coup in Chile in 1973 resulted in the deposition of President Salvador Allende.	An American aided coup in Chile in 1973 resulted in the deposition of President Enable.
An American aided coup in Chile in 1973 deposed President Salvador Allende.	An American aided coup in Chile in 1973 deposed President Everyone.
An American aided coup in Chile in 1973 deposed President Salvador Allende.	An American aided coup in Chile in 1973 deposed President Throughout.
An American aided coup in Chile in 1973 deposed President Salvador Allende.	An American aided coup in Chile in 1973 deposed President Entity.
Question: An American aided coup in Chile in 1973, deposed which President?	
Ground truth: salvador allende	

Table 13: Mode 2 qualitative example (5 of 10 responses shown). The attacked responses replace the key answer word with factually incorrect embedding neighbours. Semantic Entropy remains low because the NLI-based clustering procedure groups the modified responses together; the example illustrates that low NLI-based Semantic Entropy does not by itself establish semantic equivalence.

Clean responses *(ES = -4.189)*	**Targeted responses** *(ES = 5.657)*
The poet who was married to Ted Hughes until her death in 1963 was Sylvia Plath.	Sylvia Plath, until her death in 1963, was married to Ted Hughes.
The poet who was married to Ted Hughes until her death in 1963 was Sylvia Plath.	Sylvia Plath, who was married to Ted Hughes until her death in 1963, was the poet.
The poet who was married to Ted Hughes until her death in 1963 was Sylvia Plath.	The poet who was married to Ted Hughes until her death in 1963 was Sylvia Plath!
The poet who was married to Ted Hughes until her death in 1963 was Sylvia Plath.	Sylvia Plath was married to Ted Hughes until her death in 1963.
The poet who was married to Ted Hughes until her death in 1963 was Sylvia Plath.	Sylvia Plath was the poet who was married to Ted Hughes until her demise in 1963.
Question: Which poet was married to Ted Hughes until her death in 1963?	
Ground truth: sylvia plath	

Table 14: Mode 3 qualitative example (5 of 10 responses shown). The targeted responses are semantically equivalent paraphrases of the clean responses; EigenScore rises above τ_3 despite preserved factual correctness, illustrating sensitivity to representational dispersion under meaning-preserving rewriting.

		Correctness		**Consistency**	
Metric	**Condition**	κ	α	κ_w	α
FTE	Clean	0.771	0.771	0.628	0.644
	Targeted	0.677	0.677	0.783	0.791
SE	Clean	1.000	1.000	0.720	0.813
	Targeted	0.283	0.283	0.611	0.560
ES	Clean	0.772	0.772	0.099	0.168
	Targeted	0.686	0.686	0.103	0.027

Table 15: Inter-annotator agreement. Correctness is measured with unweighted Cohen's κ and nominal Krippendorff's α. Consistency is measured with quadratic-weighted Cohen's κ_w and ordinal Krippendorff's α. For SE targeted correctness, the very low positive rate induces a floor effect, so the low κ should be interpreted alongside the near-perfect raw agreement reported in the text.

Annotation instructions (as shown in-sheet).

Each example block shows a question, its ground truth answer, and two sets of model responses: **RESPONSE SET A** and **RESPONSE SET B**. Please fill in every yellow cell.

CORRECTNESS (one mark per response row):
For every response, decide whether it correctly answers the question given the ground truth. Write **Y** (yes, correct) or **N** (no, incorrect) in the yellow cell next to each response.

CONSISTENCY (0–4 scale — one score per set, at the bottom of each example block):
Rate how much the responses within each set mean the same thing, regardless of writing style or phrasing.

- 0 — not consistent
- 1 — rather not consistent
- 2 — somewhat consistent
- 3 — rather consistent
- 4 — consistent

Example 1 [llama13B]

Question: The English towns of Abingdon, Wallingford and Marlow lie on which river?

Ground truth: river thames

#	Response Set A	Correct? (Y / N)	Response Set B	Correct? (Y / N)
1	The English towns of Abingdon, Wallingford, and Marlow lie on the River Thames.		These towns lie on the Thames River.	
2	The English towns of Abingdon, Wallingford, and Marlow lie on the River Thames.		The River Thames flows through the English towns of Abingdon, Wallingford, and Marlow.	
3	The English towns of Abingdon, Wallingford, and Marlow lie on the River Thames.		The English towns of Abingdon, Wallingford, and Marlow lie on the River Thames!	
4	The English towns of Abingdon, Wallingford, and Marlow lie on the River Thames.		The towns of Abingdon, Wallingford, and Marlow sit on the River Thames.	
5	The English towns of Abingdon, Wallingford, and Marlow lie on the River Thames.		The English towns of Abingdon, Wallingford, and Marlow are situated on the River Thames.	

Consistency — SET A: How much do the Set A responses mean the same thing, regardless of writing style? [0 = not consistent ... 4 = consistent]

Consistency — SET B: How much do the Set B responses mean the same thing, regardless of writing style? [0 = not consistent ... 4 = consistent]

Example 2 [llama13B]

Question: "Who wrote the opera ""Dido and Aeneas""?"

Ground truth: henry purcell

#	Response Set A	Correct? (Y / N)	Response Set B	Correct? (Y / N)
1	Henry Purcell is credited with writing the opera ""Dido and Aeneas"".		The opera "Dido and Aeneas" was written by Henry Purcell.	
2	The opera Dido and Aeneas was penned by none other than Henry Purcell.		The opera "Dido and Aeneas" was written by Henry Purcell.	
3	The opera ""Dido and Aeneas"" was written by Henry Purcell!		The opera "Dido and Aeneas" was written by Henry Purcell.	
4	The opera ""Dido and Aeneas"" was composed by Henry Purcell.		The opera "Dido and Aeneas" was written by Henry Purcell.	
5	The opera ""Dido and Aeneas"" is written by Henry Purcell.		The opera ""Dido and Aeneas"" was written by Henry Purcell.	

trivia_qa | nq | squad

Figure 2: Excerpt of the annotation sheet.

A.13 GPU USAGE

All experiments were conducted on two high-performance computing clusters. The core results were produced in under 100 GPU hours in total, distributed across NVIDIA A40, A100, and H200 GPUs.