# When Do Instruction Perturbations Help Self-Consistency for Hallucination Detection?

Anonymous ACL submission

## Abstract

Self-consistency is a widely used black-box strategy for hallucination detection, but repeated same-instruction sampling can produce spurious agreement under small sampling budgets. We propose **Instruction-Perturbed Self-Consistency (IP)**, a sampling strategy for self-consistency-based hallucination detection that replaces repeated same-instruction sampling with a small set of controlled instruction variants. In this work, we instantiate IP within NLI-based SelfCheckGPT, keeping its contradiction-based scoring mechanism unchanged while varying how auxiliary responses are generated. On TriviaQA and POPQA across four LLMs, four instruction-perturbed samples consistently improve over four same-instruction samples and remain competitive with a ten-sample SelfCheckGPT baseline at roughly half of its sampling cost on average. Further analysis shows that IP is not simply prompt diversification: useful perturbations selectively increase contradictions for hallucinated outputs, whereas some models exhibit non-discriminative instruction sensitivity. These findings suggest that instruction perturbation can serve as a simple, strict black-box sampling intervention for self-consistency-based hallucination detection.

## 1 Introduction

Large language models (LLMs) remain prone to hallucinations, i.e., outputs that are fabricated or inconsistent with real-world facts (Zhang et al., 2025). A common black-box detection strategy is self-consistency: reliable factual content should remain stable across generations, whereas hallucinated content is more likely to vary. NLI-based SelfCheckGPT (Manakul et al., 2023) operationalizes this idea by comparing a primary response against multiple auxiliary responses from the same model and using contradiction probability as a hallucination score.

However, repeated same-instruction sampling can be inefficient and weakly informative. Under a small sampling budget, a model may follow the same answer path and reproduce the same plausible but false response, creating spurious consensus rather than uncertainty. Figure 1 illustrates this failure mode: same-instruction samples repeat the incorrect answer, while instruction variants elicit contradictory alternatives.

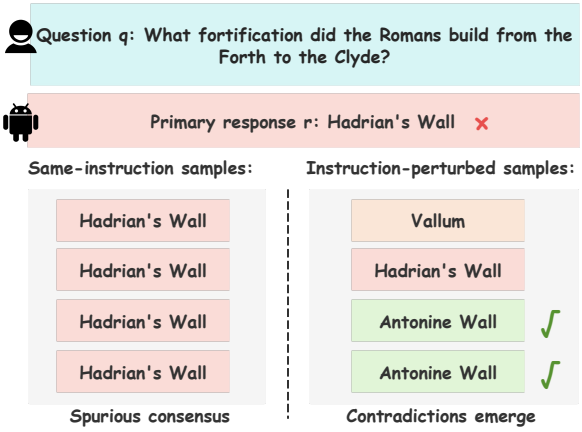


Figure 1: Motivating example of instruction-perturbed self-consistency from TriviaQA using Gemma3-4B-IT. Repeated same-instruction sampling reproduces the same incorrect answer and creates a spurious consensus. Instruction perturbations elicit a more diverse set of responses, revealing contradictions to the primary response.

We revisit self-consistency from the perspective of instruction sensitivity. Prior work shows that instructions can change LLM outputs (Sun et al., 2024; Jang et al., 2025). Instead of treating this sensitivity only as a robustness concern, we use it as a controlled sampling intervention. We propose **Instruction-Perturbed Self-Consistency** (IP): auxiliary responses are generated from task-equivalent instruction variants, while the SelfCheckGPT contradiction scorer is kept unchanged.

Our work makes three contributions. First, we propose Instruction-Perturbed Self-Consistency (IP), a black-box sampling strategy that replaces

repeated same-instruction sampling with controlled instruction variants for self-consistency-based hallucination detection. Second, we show on TriviaQA and POPQA across four LLMs that IP with four samples consistently improves over same-sample SelfCheckGPT and is competitive with a ten-sample baseline. Third, we analyze when perturbations help: output granularity and content constraints provide the most reliable gains, whereas surface mood changes and typographic noise provide weak or unstable signals. Our discriminative-gap analysis further shows that IP is model-dependent: perturbations help when they selectively increase contradictions for hallucinated outputs, but can fail under non-discriminative instruction sensitivity.

| Dataset | Model | $SC_4$ | $IP_4$ | $SC_{10}$ | Cost ratio |
|---|---|---|---|---|---|
| TriviaQA | Gemma3-4B | 0.772 | **0.840** | 0.801 | 0.41 |
| | Qwen3-4B | 0.786 | **0.823** | 0.788 | 0.91 |
| | Qwen3-8B | 0.815 | **0.849** | 0.845 | 0.45 |
| | Llama3.1-8B | 0.853 | **0.869** | 0.874 | 0.35 |
| POPQA | Gemma3-4B | 0.827 | **0.892** | 0.864 | 0.42 |
| | Qwen3-4B | 0.854 | **0.873** | 0.862 | 0.72 |
| | Qwen3-8B | 0.864 | **0.888** | 0.889 | 0.45 |
| | Llama3.1-8B | 0.831 | **0.855** | 0.858 | 0.35 |
| Average | – | 0.825 | **0.861** | 0.848 | 0.49 |

Table 1: Main AUROC comparison. $IP_4$ consistently improves over same-sample SelfCheckGPT ($SC_4$) and is competitive with the higher-sample $SC_{10}$ baseline. The last column (Cost ratio) reports the ratio of average auxiliary sampling tokens (input + output) between $IP_4$ and $SC_{10}$. Full costs and additional efficient baselines are in Appendix A.

## 2 Related Work

**Self-consistency for hallucination detection.** SelfCheckGPT (Manakul et al., 2023) detects hallucinations by comparing a primary response with sampled alternatives from the same model, including an NLI-based variant that uses contradiction as a black-box hallucination score. Semantic uncertainty methods instead aggregate generations at the meaning level, for example by clustering semantically equivalent responses and measuring uncertainty over semantic distributions (Farquhar et al., 2024; Lin et al., 2024). These methods motivate consistency as a hallucination signal, but typically obtain evidence by repeatedly sampling from the same instruction.

**Efficient consistency and instruction sensitivity.** Recent efficiency-oriented methods reduce redundancy in repeated generation, for example by reusing cached decoding states (Gao et al., 2025) or constructing beam-search-based candidate sets (Fadeeva et al., 2026). These approaches are complementary to IP, but often require access to model internals such as KV caches or token probabilities. In parallel, work on instruction sensitivity shows that different instructions can elicit different outputs and that cross-instruction consistency is related to task accuracy (Sun et al., 2024; Jang et al., 2025). We repurpose this sensitivity as a strict black-box sampling mechanism for hallucination detection.

## 3 Method

Let $q$ be a question, $p^*$ a reference instruction, and $\mathcal{M}$ the target LLM. The primary response is generated greedily as $r^0 = \mathcal{M}([p^*; q])$. Standard NLI-based SelfCheckGPT then samples $k$ auxiliary responses from the same instruction, $r_{\text{SC}}^{(j)} \sim \mathcal{M}([p^*; q])$, and compares them with $r^0$.

IP changes only the auxiliary sampling step. Let $\mathcal{P}_{\text{pool}}$ be a pool of task-equivalent instruction variants that preserve the factual QA task but vary phrasing, answer granularity, or output constraints. Given budget $k$, IP selects $\mathcal{P}_k = \{p^{(i_1)}, \ldots, p^{(i_k)}\} \subseteq \mathcal{P}_{\text{pool}}$ and samples

$$r_{\text{IP}}^{(j)} \sim \mathcal{M}([p^{(i_j)}; q]), \quad j = 1, \ldots, k. \tag{1}$$

The NLI scoring step is identical to SelfCheckGPT. We use the NLI model's contradiction score for each auxiliary–primary response pair, treating the auxiliary response as the premise and the primary response as the hypothesis. The final hallucination score is the average contradiction score:

$$h_{\text{IP}}(q, r^0) = \frac{1}{k} \sum_{j=1}^{k} s_{\text{NLI}}(r_{\text{IP}}^{(j)}, r^0), \tag{2}$$

where $s_{\text{NLI}}(a, b)$ denotes the contradiction score assigned by the NLI model to the pair $(a, b)$.

We construct the instruction pool along five axes: **Output Granularity**, which varies answer length; **Output Content Constraint**, which suppresses explanations and speculative content; **Pragmatic Form**, which changes surface mood; **Epistemic Pressure**, which warns that the question may be tricky; and **Misspelling**, which adds typographic noise without changing the task. Unless otherwise stated, the main four-sample IP setting uses the fixed subset $\{p^{(1)}, p^{(3)}, p^{(5)}, p^{(8)}\}$, covering

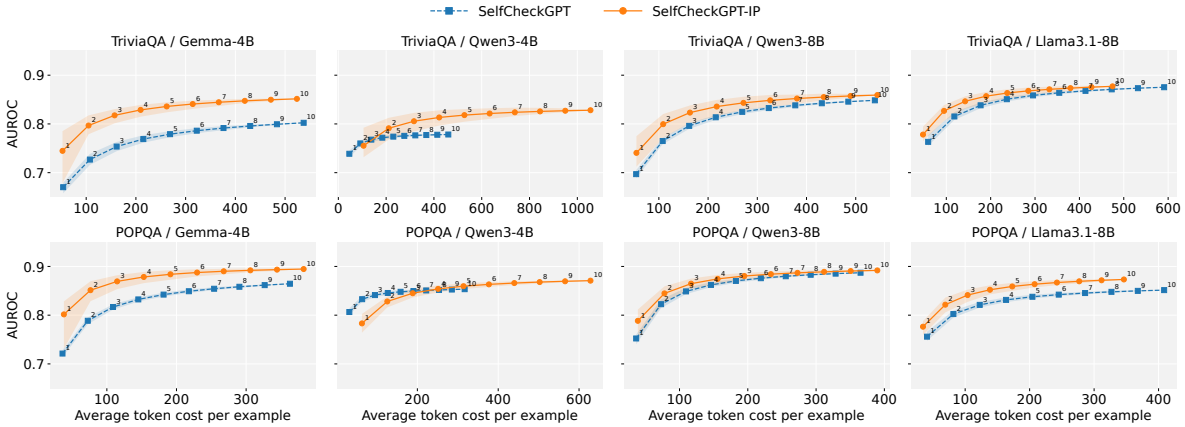


Figure 2: AUROC–cost trade-off across dataset–model pairs. Markers denote budget $k$. SELFCHECKGPT-IP averages over all $\binom{10}{k}$ subsets of ten Output Granularity and Output Content Constraint instructions; SELFCHECKGPT averages over $k$-subsets of ten same-instruction samples. Shading shows the 5th–95th percentile AUROC.

three granularity levels and one factual-answer constraint. The full instruction pool is in Appendix A.

## 4 Experimental Setup

**Datasets and models.** We evaluate on TriviaQA (Joshi et al., 2017) and POPQA (Mallen et al., 2023), randomly sampling 3,000 questions from each dataset. We test four instruction-tuned LLMs: Gemma3-4B-IT (Team, 2025a), Qwen3-4B-Instruct-2507 (Team, 2025b), Qwen3-8B (Team, 2025b), and Llama-3.1-8B-Instruct (Dubey et al., 2024).

**Labels and scoring.** We label the primary response by comparing it with reference answers using DeepSeek-V3-0324 as an LLM judge (DeepSeek-AI et al., 2024); a manual check estimates 97% labeling accuracy. For contradiction scoring, we use DeBERTa-v2-xlarge-mnli (He et al., 2021; Williams et al., 2018), following NLI-based SelfCheckGPT. We report AUROC for hallucination detection.

**Baselines and cost.** We compare $\mathrm{IP}_4$ against SelfCheckGPT with four and ten same-instruction samples ($\mathrm{SC}_4$, $\mathrm{SC}_{10}$). Full results for beam-search candidates and Decoding Memories (Fadeeva et al., 2026; Gao et al., 2025) are reported in Appendix A, as both methods require access to model internals and are therefore not strictly black-box. The primary response is generated with greedy decoding ($T = 0$); auxiliary responses use $T = 1.0$. Cost is the average number of input prompt tokens plus newly generated output tokens for auxiliary sampling.

## 5 Results

Table 1 shows that $\mathrm{IP}_4$ improves over $\mathrm{SC}_4$ in all eight dataset–model settings, with a mean gain of $+0.036$ AUROC. The gains are largest for Gemma-3-4B on both TriviaQA ($+0.068$) and POPQA ($+0.065$), suggesting that instruction perturbations are especially useful when same-instruction sampling produces spurious agreement. Figure 2 further shows that the improvement is most visible in the low-sample regime: the IP curve rises more quickly than same-instruction sampling for most dataset–model pairs.

Compared with the higher-sample $\mathrm{SC}_{10}$ baseline, $\mathrm{IP}_4$ is competitive despite using fewer samples. On average, $\mathrm{IP}_4$ improves over $\mathrm{SC}_{10}$ by $+0.013$ AUROC; in the settings where it is lower, the gap is at most $0.005$ AUROC. $\mathrm{IP}_4$ uses $205.0$ sampling tokens per example on average, compared with $419.7$ for $\mathrm{SC}_{10}$, or roughly 49% of the $\mathrm{SC}_{10}$ sampling cost. These results support a narrower claim than strict token-budget dominance: $\mathrm{IP}_4$ is a simple black-box way to obtain more informative low-sample consistency signals, though it is not always cheaper than $\mathrm{SC}_4$ because some perturbed instructions elicit longer responses.

Appendix A reports beam-search and Decoding Memories baselines; $\mathrm{IP}_4$ generally provides stronger AUROC while remaining strict black-box, although DMP is sometimes cheaper.

## 6 Analysis: When Do Perturbations Help?

Not all instruction variants are equally useful. Table 2 shows that Output Granularity and Output

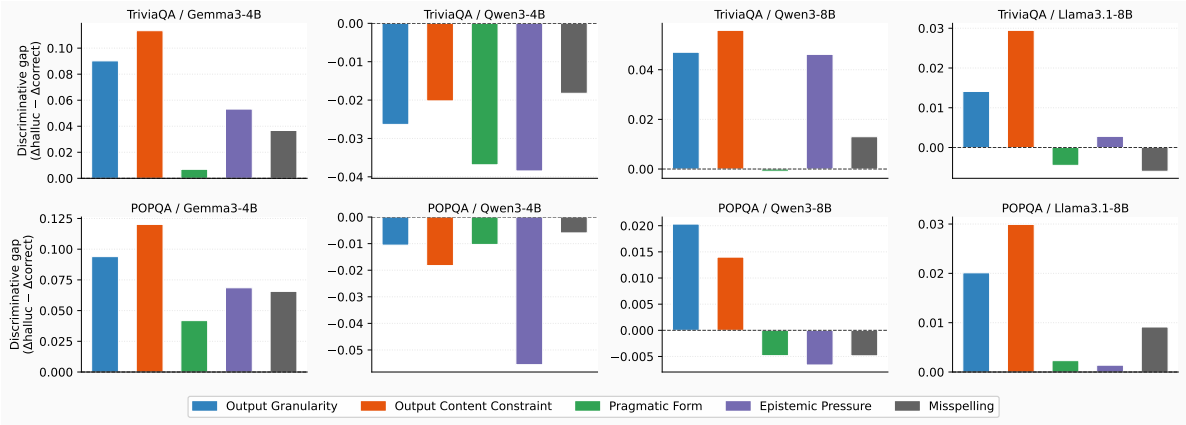


Figure 3: Category-level discriminative gap, defined as the contradiction-score increase on hallucinated examples minus the increase on correct examples, relative to same-instruction SelfCheckGPT. Positive values indicate hallucination-specific instability; negative values indicate non-discriminative sensitivity.

Content Constraint are the most reliable perturbation axes. Both encourage direct factual commitment: length constraints push the model toward concise answers, while content constraints suppress explanations and speculative continuations. In contrast, Pragmatic Form mainly changes surface mood and yields weak gains, while Misspelling and Epistemic Pressure are unstable.

To understand this pattern, we compute a discriminative gap for each perturbation category: the increase in average contradiction score on hallucinated examples minus the increase on correct examples, relative to same-instruction sampling. A useful perturbation should not simply create more disagreement; it should increase contradiction selectively for hallucinated outputs while preserving stability for correct answers.

Figure 3 supports this criterion. Output Granularity and Output Content Constraint usually produce positive gaps, indicating hallucination-specific instability. Pragmatic Form is near zero or negative, suggesting that changing declarative, imperative, interrogative, or exclamative mood does not meaningfully change factual commitment. A notable failure case is Qwen3-4B, where all perturbation categories have negative gaps on both datasets: the model appears sensitive to instruction surface changes in a non-discriminative way, increasing contradiction for both correct and hallucinated responses. Thus, IP is model-dependent: instruction perturbation helps when it induces hallucination-specific instability, but can fail under non-discriminative instruction sensitivity.

| Perturbation type | Mean ΔAUROC ↑ |
|---|---|
| Output Granularity | **+0.033** |
| Output Content Constraint | +0.027 |
| Misspelling | +0.012 |
| Pragmatic Form | +0.009 |
| Epistemic Pressure | +0.008 |

Table 2: Mean AUROC gain of pure-category pools over same-instruction SelfCheckGPT with $k = 3$, averaged over two datasets and four LLMs. Full results are in Appendix A.

This analysis also clarifies why prompt diversity alone is an insufficient objective. Perturbations that merely change surface form can create noisy disagreement, but a useful hallucination detector requires separation between correct and hallucinated primary responses. Effective variants therefore act like controlled stress tests of factual commitment: they should preserve stable answers while making unsupported answers easier to contradict.

## 7 Conclusion

We presented Instruction-Perturbed Self-Consistency, a black-box strategy replacing repeated same-instruction sampling in standard self-consistency methods with a small set of instruction variants. Across both datasets, four instruction-perturbed samples consistently outperform four same-instruction samples and are competitive with a ten-sample baseline at substantially lower sampling cost. Our analysis shows that useful perturbations are not arbitrary prompt variants: they selectively increase contradictions for hallucinated outputs while keeping correct answers stable. This makes instruction perturbation a controlled, model-sensitive sampling intervention rather than a general-purpose prompt-diversity heuristic.

## Limitations

This work has several limitations that point to directions for future research. First, our evaluation is limited to short-form open-domain QA, and it remains unclear whether instruction-perturbed self-consistency remains effective for long-form generation tasks. Second, the instruction variants are manually designed along five predefined axes, which may not generalize to other languages, domains, or task formats. Automatically constructing perturbation pools for specific target scenarios is a direction for future work.

## Ethical considerations

This work uses only publicly available datasets and open-source models, all under their respective licenses. We see no significant ethical risks, but recommend additional validation before deployment in high-stakes domains.

## Acknowledgments

## References


DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models.

Ekaterina Fadeeva, Maiya Goloburda, Aleksandr Rubashevskii, Roman Vashurin, Artem Shelmanov, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. 2026. Don't throw away your beams: Improving consistency-based uncertainties in LLMs via beam search. In *The Fourteenth International Conference on Learning Representations*.

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

Weizhi Gao, Xiaorui Liu, Feiyi Wang, Dan Lu, and Junqi Yin. 2025. Decoding memories: An efficient pipeline for self-consistency hallucination detection. *ArXiv*, abs/2508.21228.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dongjun Jang, Youngchae Ahn, and Hyopil Shin. 2025. RCScore: Quantifying response consistency in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5690–5708, Suzhou, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfcheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2024. Evaluating the zero-shot robustness of instruction-tuned language models. In *The Twelfth International Conference on Learning Representations*.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

# A Appendix

## A.1 Instruction Variants

Table 3 lists all instruction variants used in our instruction perturbation pool.

## A.2 Full Results

Table 4 reports full AUROC and token-cost results for all methods, including beam-search candidates (BS) and Decoding Memories (DMP). Table 5 reports per-category AUROC results under the pure category pool setting.

| ID | Instruction |
|---|---|
| $p^*$ | Answer the following question in a single brief but complete sentence. |
| | *Output Granularity* |
| $p^{(1)}$ | Answer the following question. |
| $p^{(2)}$ | Provide the answer to the following question. |
| $p^{(3)}$ | Answer the following question in a single brief but complete sentence. |
| $p^{(4)}$ | State the answer in one short complete sentence. |
| $p^{(5)}$ | Answer the question with a short phrase. |
| $p^{(6)}$ | Respond with the shortest answer you can. |
| | *Output Content Constraint* |
| $p^{(7)}$ | Identify the answer and state it briefly. |
| $p^{(8)}$ | Give a concise factual answer to the question. |
| $p^{(9)}$ | Provide only the final answer to the question. No explanation. |
| $p^{(10)}$ | Give just the answer, not the reasoning. |
| | *Pragmatic Form* |
| $p^{(11)}$ | The following question should be answered in a single brief but complete sentence. |
| $p^{(12)}$ | Could you answer the following question in a single brief but complete sentence? |
| $p^{(13)}$ | How important it is to answer the following question in a single brief but complete sentence! |
| | *Epistemic Pressure* |
| $p^{(14)}$ | Answer the following question in a single brief but complete sentence. Note that this question is commonly answered incorrectly. |
| $p^{(15)}$ | Answer the following question in a single brief but complete sentence. Your first instinct may be wrong. |
| $p^{(16)}$ | Answer the following question in a single brief but complete sentence. Think carefully before answering, as this question may be a trick. |
| | *Misspelling* |
| $p^{(17)}$ | Anwser the following question in a single brief but complete sentence. |
| $p^{(18)}$ | Answer the following question in a sinlge breif but compelte sentance. |
| $p^{(19)}$ | Answr the folowing queston in a singl breif but complet sentece. |

Table 3: The original instruction $p^*$ and the pool of perturbed instruction variants used in IP.

| Method | Gemma-3-4B | | Qwen3-4B | | Qwen3-8B | | Llama-3.1-8B | |
|---|---|---|---|---|---|---|---|---|
| | AUROC↑ | Cost↓ | AUROC↑ | Cost↓ | AUROC↑ | Cost↓ | AUROC↑ | Cost↓ |
| **TriviaQA** | | | | | | | | |
| $\text{SelfCheckGPT}_{10}$ | 0.801 | 508.50 | 0.788 | 378.18 | 0.845 | 519.87 | 0.874 | 556.77 |
| $\text{SelfCheckGPT}_{4}$ | 0.772 | 203.65 | 0.786 | 150.03 | 0.815 | 208.05 | 0.853 | 223.35 |
| $\text{SelfCheckGPT}_{4}$ + BS | 0.779 | 201.32 | 0.824 | 215.23 | 0.853 | 203.37 | 0.829 | 208.44 |
| $\text{SelfCheckGPT}_{4}$ + DMP | 0.800 | 137.08 | 0.771 | 162.07 | 0.783 | 142.48 | 0.840 | 137.88 |
| $\text{SelfCheckGPT}_{4}$ + IP | 0.840 | 210.20 | 0.823 | 345.65 | 0.849 | 233.61 | 0.869 | 194.95 |
| **POPQA** | | | | | | | | |
| $\text{SelfCheckGPT}_{10}$ | 0.864 | 348.48 | 0.862 | 294.87 | 0.889 | 355.94 | 0.858 | 394.64 |
| $\text{SelfCheckGPT}_{4}$ | 0.827 | 139.19 | 0.854 | 116.90 | 0.864 | 142.10 | 0.831 | 157.92 |
| $\text{SelfCheckGPT}_{4}$ + BS | 0.854 | 136.16 | 0.863 | 151.46 | 0.872 | 139.80 | 0.820 | 146.25 |
| $\text{SelfCheckGPT}_{4}$ + DMP | 0.843 | 96.44 | 0.850 | 115.42 | 0.841 | 102.83 | 0.840 | 100.66 |
| $\text{SelfCheckGPT}_{4}$ + IP | 0.892 | 147.70 | 0.873 | 211.00 | 0.888 | 159.00 | 0.855 | 138.10 |

Table 4: AUROC and token cost of self-consistency methods on TriviaQA and POPQA across different LLMs. Cost denotes the average total number of input and newly generated tokens per example.

| Perturbation Type | Gemma-3-4B | | Qwen3-4B | | Qwen3-8B | | Llama-3.1-8B | |
|---|---|---|---|---|---|---|---|---|
| | AUROC↑ | Δ ↑ | AUROC↑ | Δ ↑ | AUROC↑ | Δ ↑ | AUROC↑ | Δ ↑ |
| **TriviaQA** | | | | | | | | |
| **Output Granularity** | 0.828 | +0.066 | 0.810 | +0.039 | 0.829 | +0.031 | 0.847 | +0.005 |
| **Output Content Constraint** | 0.811 | +0.049 | 0.804 | +0.033 | 0.823 | +0.025 | 0.851 | +0.009 |
| **Pragmatic Form** | 0.767 | +0.005 | 0.800 | +0.029 | 0.800 | +0.002 | 0.839 | -0.003 |
| **Epistemic Pressure** | 0.785 | +0.023 | 0.794 | +0.023 | 0.824 | +0.026 | 0.834 | -0.008 |
| **Misspelling** | 0.786 | +0.024 | 0.806 | +0.035 | 0.805 | +0.007 | 0.828 | -0.014 |
| **POPQA** | | | | | | | | |
| **Output Granularity** | 0.871 | +0.060 | 0.852 | +0.012 | 0.870 | +0.017 | 0.845 | +0.030 |
| **Output Content Constraint** | 0.874 | +0.063 | 0.837 | -0.003 | 0.862 | +0.009 | 0.843 | +0.028 |
| **Pragmatic Form** | 0.842 | +0.031 | 0.842 | +0.002 | 0.851 | -0.002 | 0.820 | +0.005 |
| **Epistemic Pressure** | 0.845 | +0.034 | 0.821 | -0.019 | 0.842 | -0.011 | 0.819 | +0.004 |
| **Misspelling** | 0.849 | +0.038 | 0.842 | +0.002 | 0.846 | -0.007 | 0.824 | +0.010 |

Table 5: AUROC of each perturbation category under the Pure Category Pool setting (k = 3) on TriviaQA and POPQA across LLMs. Δ denotes the AUROC gain over the SelfCheckGPT baseline , with darker shades indicating larger Δ.