Available Online: 16 November 2026 DOI: 10.54254/2977-3903/2026.39108

Conformal selective answering with hidden-state probes controls FDR on HaluEval but fails under domain shift in small language models

***Fan Si*[1,*], *Tingzhang Yang*[2], *Yulin Tu*[3], *Xing Li*[3]**

[1]School of Economics, Peking University, Beijing, China
[2]School of Management Science and Engineering, Chongqing Technology and Business University, Chongqing, China
[3]Graduate School of Global Studies, Yonsei University, Seoul, Republic of Korea

*Corresponding Author. Email: sifan@pku.org.cn

Abstract. Whether the hidden state probe can provide a False Discovery Rate (FDR) guarantee for the answer release of the small language model still lacks a cross-domain shift test. This article uses a multi-layer probe to identify the hallucinated answer, decides to release it with conformal selection, and tests it on HaluEval and TruthfulQA. The probe AUROC (area under the ROC curve) on HaluEval is 0.991, which is higher than 0.974 of the answer-length baseline; the FDR at the target level of 0.10 is 0.050, and the power is 0.992, which is higher than the baseline of 0.883. The exploratory AUROC with a single feature of punctuation at the end of the sentence reached 0.969, and the difference in the length and sentence style of HaluEval's answer contributed most of the AUROC. When cross-domain, AUROC dropped to 0.491, and FDR rose to 0.409, exceeding the target level; after recalibration, 0 answers were released. Exploratory in-domain retraining brings AUROC back to 0.904, FDR is 0.045, and power is 0.502. High AUROC cannot be directly interpreted as the model grasping the true and false information, and the probe must be retrained with the data of the target domain.

Keywords: hallucination detection, small language models, hidden-state probes, conformal selection, false discovery rate

1. Introduction

The hallucinated answer generated by the small language model [1] will cause errors if given directly to the user, so it needs to be controlled before release. Li et al. [2] used HaluEval to evaluate hallucination recognition and found that some queries of ChatGPT produced hallucination content. The identification results do not stipulate which answers to release, and the release decision also needs to control the False Discovery Rate (FDR) within the target level.

The difference in the length and sentence style of HaluEval's answer contributes to most of the AUROC, and the high score of the hidden state linear probe mainly reflects the benchmark artifact. The question of

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

TruthfulQA [3] is aimed at common human misunderstandings, and the length distribution of the answer is also different from HaluEval. RQ1 tests whether the ability of the hidden state linear probe to recognize the hallucinated answer exceeds the answer length, log-likelihood and text feature baseline. RQ2 tests whether conformal selection can control the FDR of the release set within the target level on a finite sample and retain how many correct answers. RQ3 checks whether this control is still valid on TruthfulQA, and whether recalibration or retraining can restore it.

2. Literature review

2.1. Hallucination detection method

The sampling method of hallucination detection samples the same prompt many times. SelfCheckGPT of Manakul et al. [4] uses the consistency between responses to judge factuality without the need for an external database.

Uncertainty-class methods judge hallucinations by response uncertainty [5]. Farquhar et al. [6] used entropy-based uncertainty to estimate and detect confabulation without task-specific data.

2.2. Hidden state probe and its cross-domain generalization

Azaria and Mitchell [7] use hidden layer activations to train the classifier to judge the truth of the statement. It is found that the probability of the sentence given by the model is also affected by the sentence length and word frequency, and the classifier is more reliable. Liu et al. [8] pointed out that methods based on internal representation are often difficult to generalize to out-of-distribution data, and found that the diversity of the training set is more important than the sample size. Zhang et al. [9] pointed out that it is often difficult to generalize supervised detectors to other domains after training in specific domains, and use prompts to guide the internal state to improve cross-domain performance. Liu et al. [8] and Zhang et al. [9] did not compare recalibration and retraining.

2.3. Conformal method and false discovery rate

Benjamini and Hochberg [10] proposed that controlling the false discovery rate and its sequential procedure (Benjamini-Hochberg procedure, BH procedure) is more powerful than controlling FWER.

The conformal selection of Jin and Candes [11] is model-free. It uses conformal p-value and multiple test thresholds to pick out candidates whose results may exceed the user's given value, and controls the proportion of incorrect selection, and the guarantee is based on the assumption of mild exchangeability. Jin and Candes [12] proposed a weighted conformal p-value to control the false positive proportion when there is a covariate shift in training and testing.

The literature cited in this article does not put hidden state probe, FDR guarantee and cross-domain test together for testing.

3. Methodology

3.1. Data set and hidden state extraction

The experiment uses two data sets, the QA part of HaluEval [2] and TruthfulQA [3]. HaluEval QA contains 10,000 questions, each with 1 correct answer and 1 hallucinated answer, a total of 20,000 answers; the experiment adopts a closed-book setting and does not provide knowledge segments to the model. Lin et al. [3]

reported that TruthfulQA has 817 questions. This article uses the current CSV of the official repository, with a total of 790 questions and 1,580 answers. HaluEval QA is divided into 5,000 training questions, 2,000 calibration questions and 3,000 test questions according to questions. The candidate answer to the same question does not cross sets; the calibration set and the test set are re-partitioned 200 times.

The main model is Qwen2.5-3B-Instruct (36 layers, hidden dimension 2048), and the secondary model is SmolLM2-1.7B-Instruct (24 layers, hidden dimension 2048). Both are small language models [1], both of which use nf4 quantization and run on RTX 5080. Figure 1 gives the answer release process.

3.2. Scorers

This article compares five scorers, of which B1, B2 and B3 are the baselines, and P1 and P2 are the probes. B1 is characterized by the number of words in the answer to fit logistic regression as the answer-length baseline. B2 takes the negative value of the average log-likelihood of the answer token. B3 is characterized by TF-IDF (1–2 gram) of questions plus answers to fit logistic regression.

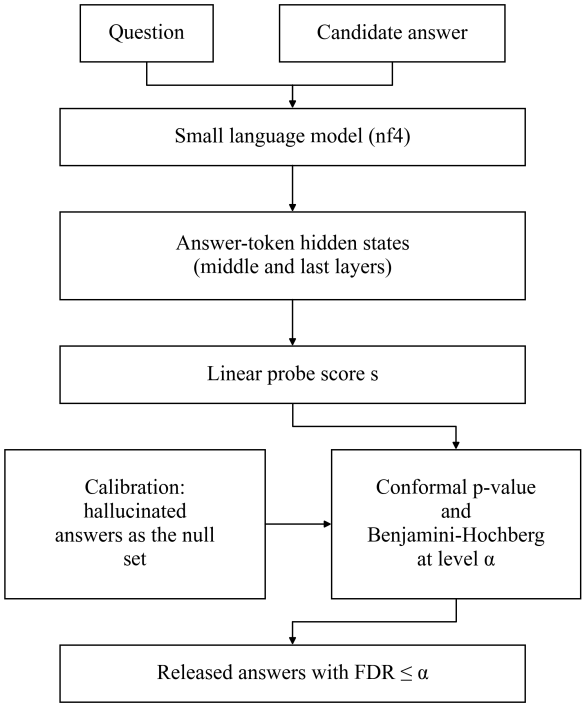


Figure 1. Workflow for releasing answers with a hidden-state probe and conformal selection

Both probes take the hidden state as input, following the idea of Azaria and Mitchell [7] to train the classifier with hidden layer activations. P1 takes the 36th layer, the last layer, and the answer token is pooled by mean pooling and then fitting logistic regression. P2 is the scorer proposed in this article, which concatenates the middle 18th layer and the last 36th layer and fits the logistic regression. The probe score $s(x)$ is given by formula (1).

$$s(x) = w^{\top} z(x) + b \tag{1}$$

Among them, x is a question and candidate answer, $z(x)$ is its hidden state feature, w is the weight vector, and b is the intercept. The parameter C of logistic regression is selected by 5-fold cross-validation in the training set, and the values are 0.01, 0.1, 1.

3.3. Conformal selection

This article takes the hallucinated answer in the calibration set as a reference to calculate the conformal p-value of the test answer j.

$$p_j = \frac{1+\#\{i \in D^1_{cal} : s_i \leq s_j\}}{n_1+1} \tag{2}$$

Where D^1_{cal} is the hallucinated answer in the calibration set, n_1 is the number of these answers, and s is the score of the answer.

The sequential procedure (BH procedure) of Benjamini and Hochberg [10] decides the release set. The p-value of the m test answers is arranged in ascending order as $p_{(1)} \leq \ldots \leq p_{(m)}$, and the first k^* are released.

$$k^* = max\{k : p_{(k)} \leq \alpha k/m\} \tag{3}$$

Where α is the target level.

$$FDR = E[\frac{V}{max(R,1)}], \quad Power = \frac{Numberofcorrectanswersreleased}{Numberofcorrectanswersinthetestset} \tag{4}$$

Where V is the number of hallucinated answers in the release, and R is the total number of releases.

3.4. Hypothesis and test

H1 to H3 are registered before the experiment. H1 expects that the AUROC of P2 probe is higher than the best baseline, and the question-level paired bootstrap 2，000 times, one-sided α = 0.05. H2 expects α = 0.05, 0.10, 0.20, the average measured FDR of 200 re-partitions does not exceed α. H3 expects when α = 0.10, the P2 power is higher than the best baseline, and the lower bound of the 95% interval of the paired difference is greater than 0. The robustness is predetermined as R1 seed replacement (7, 123), R2 training ratio 20%, R3 secondary model, R4 four-layer, R5 cross-domain, R6 cross-domain recalibration, all reported.

4. Results

4.1. Comparison of scorers

P2 is 0.991, the best baseline B1 is 0.974, the difference is 0.016, and the 95% interval of the question-level paired bootstrap 2000 times is [0.013, 0.020], the one-sided p = 0.0005, and H1 is established (Table 1). B2 is 0.152, which is lower than 0.5, and the direction is opposite to expectations.

Table 1. Detection and selection results on the HaluEval test split (Qwen2.5-3B-Instruct)

Scorer	AUROC	FDR α = 0.05	Power α = 0.05	FDR α = 0.10	Power α = 0.10	FDR α = 0.20	Power α = 0.20
B1 Length	0.974	0.016	0.309	0.042	0.883	0.060	0.946
B2 Log-likelihood	0.152	0.000	0.000	0.005	0.000	0.060	0.002
B3 TF-IDF	0.842	0.013	0.023	0.048	0.097	0.098	0.254
P1 Last-layer	0.987	0.024	0.755	0.050	0.979	0.100	0.993
P2 Multi-layer	0.991	0.025	0.887	0.050	0.992	0.099	0.998

The following readings are all exploratory readings added after reading the results. The average correct answer is 2.2 words (median 2), and the average hallucinated answer is 11.0 words (median 9). The punctuation at the end of the sentence also distinguishes two types of answers: 95.2% of the hallucinated

answers end with the punctuation at the end of the sentence, and the correct answer is 1.4%. Only with the feature of punctuation at the end of the sentence, AUROC is 0.969. The two categories on TruthfulQA are 0.6% and 0.5% respectively, and the same feature AUROC is 0.501, which is almost random.

P2's discrimination power decreases in the shortest answer stratum and equal length pairing, and this group of readings is also exploratory. The test set 1–3 word stratum AUROC is 0.788, and the hallucinated answers in this stratum account for 3.8%; the 4–6 word stratum is 0.990. There are only 75 pairs of two answers of equal length, accounting for 2.5% of the test pairing, and the P2 consistency rate is 0.853. Therefore, the high AUROC of P2 partly comes from the difference in the length of the answer, which cannot be directly interpreted as the model grasping the true and false information.

4.2. False discovery rate and power

The average measured FDR of P2 does not exceed α at all three target levels, and H2 is valid (Figure 2(a)). After 200 calibration and test re-partitions, the FDRs at α = 0.05, 0.10 and 0.20 are 0.025, 0.050 and 0.099 respectively, and the 95% intervals are [0.015, 0.034], [0.039, 0.063], [0.084, 0.117], and the upper limit of the interval is lower than the corresponding α. P2 releases more correct answers at α = 0.10, and H3 is established (Figure 2(b)). At this level, the power of P2 is 0.992, B1 is 0.883, the paired difference is 0.110, and the lower bound of the 95% interval [0.098, 0.116] is greater than 0.

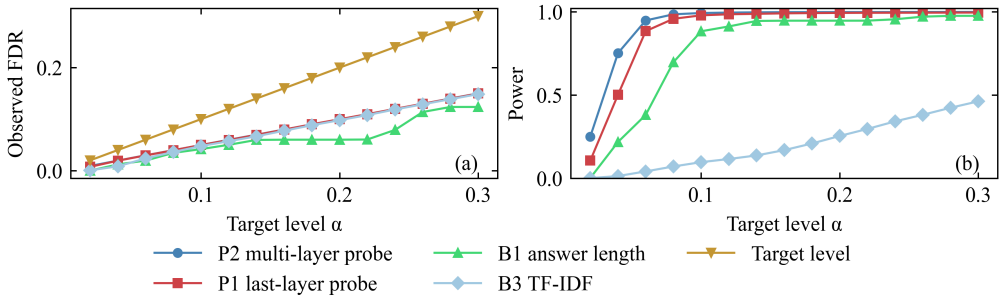


Figure 2. Observed FDR (a) and power (b) against the target level on HaluEval over 200 calibration and test re-partitions

4.3. Robustness and cross-domain results

The pre-registered robustness test changes the seed, training ratio, model and number of layers (Table 2). In R1, R2 and R3, the one-sided p of H1 is 0.0005, the 95% interval of H3 does not contain 0, and H2 is valid at all three target levels. R4 only checked H2, and all three levels are valid.

R5 trained and calibrated P2 on HaluEval, and tested on TruthfulQA [3]. This pre-registered test does not hold. The AUROC of P2 is 0.491, which is close to random. When α = 0.05, 0.10 and 0.20, the actual measured FDR is 0.400, 0.409 and 0.421, all of which exceed α. 20, 22 and 38 answers are released, and the power is 0.015, 0.016 and 0.028. FDR control did not extend to cross-domain, and the first question of RQ3 received a negative answer.

After the same probe was recalibrated (R6) with half of the TruthfulQA questions, the release almost disappeared. At α = 0.05 and 0.10, the average release of 200 re-partitions is 0; the average release is 0.81 at α = 0.20, the average FDR is 0.025, and the power is 0.001.

In the exploratory reading, the average correct answer of TruthfulQA is 9.4 words, the average hallucinated answer is 8.7 words, and the length AUROC is 0.447. The exploratory length matching training only uses pairs

with a length difference of no more than 1 word in the HaluEval training set (558 answers). The HaluEval test AUROC is 0.948, and the TruthfulQA is 0.552.

Table 2. Preregistered robustness checks, cross-domain tests and exploratory in-domain training (P2 probe, α = 0.10)

Setting	Test data	AUROC	FDR	Power
Main setting	HaluEval	0.991	0.050	0.992
R1 Seed 7	HaluEval	0.989	0.050	0.994
R1 Seed 123	HaluEval	0.990	0.050	0.993
R2 20% training share	HaluEval	0.990	0.049	0.993
R3 SmolLM2-1.7B	HaluEval	0.991	0.050	0.993
R4 Four layers	HaluEval	0.992	0.050	0.995
R5 Cross-domain, no recalibration	TruthfulQA	0.491	0.409	0.016
R6 Cross-domain, recalibrated	TruthfulQA	0.491	0.000	0.000
Exploratory: in-domain retraining	TruthfulQA	0.904	0.045	0.502

After seeing the results, the additional exploratory in-domain retraining has restored some of the power. The out-of-fold AUROC of P2 in TruthfulQA is 0.912. With 395 training questions, 395 calibration and testing questions, AUROC is 0.904, and the average FDR at α = 0.05, 0.10 and 0.20 is 0.019, 0.045, 0.095, which are lower than α, and the power is 0.409, 0.502, 0.647.

5. Discussion

The AUROC of the answer length baseline B1 has reached 0.974, and the 0.991 of P2 is only 0.016 higher. The hidden state contains a learnable signal, but this signal does not transfer. In the exploratory reading, the out-of-fold AUROC of P2 in TruthfulQA is 0.912, the length AUROC is 0.447, and the punctuation feature AUROC at the end of the sentence is 0.501. The cross-domain test is 0.491, which is almost random. This is consistent with the fact that the supervised detectors described by Zhang et al. [9] are often difficult to generalize to other domains, and the internal representation described by Liu et al. [8] is often difficult to generalize to out-of-distribution data.

The conformal guarantee relies on the exchangeability of the calibration answer and the test answer, and this condition does not apply to cross-domain. The weighted conformal *p*-value of Jin and Candes [12] is for the covariate shift, and the conformal p-value in this article is not weighted.

There are only 75 pairs (accounting for 2.5%) of equal length pairs, and the next step is to pre-register and test after expansion. Only one direction from HaluEval to TruthfulQA was measured across the domain, and the next step is to make up for the reverse evaluation. Retraining in the domain is exploratory, and the power is 0.502 when α = 0.10. The next step is to pre-register and re-check in a new domain.

6. Conclusion

The AUROC of the probe is higher than the length, log-likelihood and text feature baseline, and H1 is valid. This high AUROC mainly reflects the difference in the length of the answer and the sentence style. On the in-distribution HaluEval, the conformal selection controls the FDR of the released answer within the target level, the power of the probe is higher than the baseline, and H2 and H3 are established. After changing to

TruthfulQA, the FDR control failed, and the recalibration could not release the answer. The exploratory results show that control can only be restored by retraining with the data of the target domain.

References

[1] Wang, F., Lin, M., Ma, Y., Liu, H., He, Q., Tang, X., Tang, J., Pei, J., & Wang, S. (2025). A survey on small language models in the era of large language models: Architecture, capabilities, and trustworthiness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Volume 2)* (pp. 6173–6183). Association for Computing Machinery. https://doi.org/10.1145/3711896.3736563

[2] Li, J., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6449–6464). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.397

[3] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

[4] Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9004–9017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.557

[5] Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., & Majumdar, A. (2026). A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys, 58*(3), 1–38. https://doi.org/10.1145/3744238

[6] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature, 630*(8017), 625–630. https://doi.org/10.1038/s41586-024-07421-0

[7] Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 967–976). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.68

[8] Liu, J., Chen, S., Cheng, Y., & He, J. (2024). On the universal truthfulness hyperplane inside LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 18199–18224). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1012

[9] Zhang, F., Yu, P., Yi, B., Zhang, B., Li, T., & Liu, Z. (2025). Prompt-guided internal states for hallucination detection of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (pp. 21806–21818). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.1058

[10] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[11] Jin, Y., & Candès, E. J. (2023). Selection by prediction with conformal p-values. *Journal of Machine Learning Research, 24*(244), 1–41. https://www.jmlr.org/papers/v24/22-1176.html

[12] Jin, Y., & Candès, E. J. (2025). Model-free selective inference under covariate shift via weighted conformal p-values. *Biometrika, 113*(1), Article asaf066. https://doi.org/10.1093/biomet/asaf066