Adaptive Hallucination Detection in Large Language Models Using Multi-Stage Confidence Estimation and Semantic Consistency Analysis

Maud Estey
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
maud1984@binghamton.edu

Juhana Raty
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
juhanar@uc.edu

Egbert Haering
School of Computing, Clemson University, Clemson, SC, USA.
haering1982@clemson.edu

Abstract

Large language models can produce fluent responses that contain unsupported or contradictory claims, and single-signal detectors often mistake high token likelihood for factual reliability. This paper develops an adaptive hallucination detection framework that combines multi-stage confidence estimation with semantic consistency analysis. The framework treats hallucination as a latent relational property between a prompt, a generated response, and available evidence, rather than as a surface-level lexical defect. It integrates four complementary signals: sequence-level lexical confidence, meaning-cluster semantic entropy, sample-based semantic consistency, and claim-level evidence support. These signals are normalized into support scores and fused through reliability-weighted geometric pooling, so that unstable or poorly calibrated estimators receive lower influence. An ambiguity-sensitive threshold adjusts the acceptable risk according to retrieval volatility and semantic disagreement, and a three-way decision rule separates acceptance, flagging for verification, and rejection. The analytical contribution is a coherent decision architecture that links uncertainty estimation, semantic clustering, evidence alignment, and selective prediction within one risk-control logic. The discussion identifies boundary conditions that matter in practice: correlated samples can overstate consistency, semantic clustering can merge distinct claims, retrieval noise can mislead evidence scores, and calibration can drift under distribution shift. The framework provides a technically grounded basis for deploying hallucination detection as a calibrated, context-sensitive risk-control layer rather than as a static post hoc filter.

Keywords

hallucination detection; semantic consistency; confidence calibration; selective prediction; retrieval-augmented generation; risk control

1. Introduction

Large language models can generate text that is syntactically fluent, pragmatically plausible, and yet not supported by the prompt, the retrieved context, or verifiable world

knowledge. This failure mode is usually described as hallucination, and it spans summarization, question answering, dialogue, and retrieval-conditioned generation [1]. In abstractive settings, unsupported content can appear even when the output remains grammatical and internally coherent, which makes detection a relational problem rather than a simple fluency check [2]. A detector must therefore decide whether the claims in a response are licensed by the available evidence and whether the model's own confidence is trustworthy.

Retrieval-augmented generation reduces some factual errors by conditioning generation on external documents [3], but it does not eliminate the problem. Retrieval can return stale, incomplete, or contradictory evidence, and a model may still assign high probability to an incorrect continuation. Raw token likelihood is a weak proxy for factual correctness because modern neural models are often poorly calibrated [4], [5]. A high-probability sequence can be confidently wrong, while a correct but uncertain response can receive a lower likelihood score. This mismatch motivates detectors that combine likelihood with additional signals rather than relying on a single confidence threshold.

Sampling-based consistency offers a complementary perspective. When a model produces multiple responses to the same prompt, correct reasoning paths often converge on the same answer, whereas hallucinated content may vary across samples [6]. Semantic uncertainty extends this intuition by clustering generations according to meaning instead of surface string similarity [7]. Under retrieval conditions, uncertainty-aware generation and risk control become especially important because the meaning of a confidence score depends on the quality and stability of the evidence supplied to the model [8]. A response that is consistent across samples may still be consistently wrong if the retrieved context is misleading, and a response that varies across samples may reflect legitimate ambiguity rather than hallucination.

Meaning-level uncertainty provides a more principled basis for consistency analysis than token overlap. Semantic entropy quantifies uncertainty over meaning clusters and has been used to detect hallucinated content in large language models [9]. Black-box consistency checks can also identify unsupported statements without access to model internals or external labels [10]. Internal representations contain additional signals of factual correctness [11], [12], but these signals are not interchangeable with output-level confidence. Dropout-based uncertainty [13], deep ensembles [14], and distribution-shift evaluations [15] show that predictive confidence must be validated beyond the conditions in which it was estimated. Factuality evaluation [16], retrieval-augmented generation assessment [17], and verbalized uncertainty [18] provide complementary measurement perspectives, yet they do not by themselves specify how multiple heterogeneous signals should be combined into a deployable decision rule.

The central contribution of this paper is an adaptive hallucination detection framework that integrates multi-stage confidence estimation with semantic consistency analysis. The framework is organized around a latent hallucination state, four support signals, reliability-weighted fusion, an ambiguity-sensitive threshold, and a three-way decision rule. The design addresses a specific gap: existing detectors often use static thresholds and treat confidence as a scalar property of the model, whereas hallucination risk depends on evidence stability, meaning-level disagreement, and the cost of false acceptance. The analysis develops the estimator logic, decision properties, and boundary conditions of the framework without treating any numerical calibration as a fixed empirical result.

2. Problem Formulation and Related Work

Let x denote the prompt and any retrieved context, and let y be a generated response. The latent variable H indicates whether y contains at least one unsupported or contradicted claim, with H=1 denoting hallucination and H=0 denoting an acceptable response. The detector observes x, y, and a vector of signals s(x,y) that includes lexical confidence, semantic consistency, and evidence support. The task is to estimate a risk score R(x,y) and apply a decision rule that controls the trade-off between accepting unsupported content and rejecting valid content.

This formulation differs from binary factuality classification because the unit of analysis is a response containing multiple claims. A response can be partly supported and partly unsupported, and the evidence may cover some claims while leaving others unresolved. The detector therefore needs a claim-level view of support, not only a sequence-level score. Related work on factuality evaluation has emphasized atomic claim decomposition and fine-grained verification [16], while retrieval-augmented generation evaluation has highlighted the need to separate retrieval quality from generation quality [17]. The framework developed here uses these ideas at the level of signal construction: evidence support is computed over atomic claims rather than over the whole response as an undifferentiated string.

Uncertainty estimation for neural models provides the first family of signals. Bayesian approximations such as dropout [13] and deep ensembles [14] estimate predictive uncertainty by measuring variation across stochastic or independently trained models. These methods are useful when multiple models or stochastic passes are available, but they do not directly address semantic correctness. A model ensemble can agree on a hallucinated answer, and dropout variance can be small when the model is confidently wrong. Calibration research further shows that modern neural networks often produce probability estimates that do not match empirical frequencies [5], and distribution-shift studies show that calibration can deteriorate under changes in input distribution [15]. These findings imply that confidence signals should be treated as evidence about model behavior, not as direct estimates of factual truth.

Consistency-based methods provide a second family of signals. Self-consistency samples multiple reasoning paths and selects the most frequent answer [6]. Semantic uncertainty clusters sampled outputs by meaning and computes entropy over the resulting clusters [7]. Semantic entropy has been used to detect hallucinations by measuring uncertainty over meanings rather than token sequences [9]. SelfCheckGPT detects hallucination through consistency across sampled responses without requiring external resources [10]. These methods are attractive because they require no labeled hallucination data, but their reliability depends on the independence and diversity of samples. If samples are highly correlated, consistency can be artificially high; if the prompt is genuinely ambiguous, low consistency may reflect legitimate variation rather than hallucination.

Evidence-based methods provide a third family of signals. Retrieval-augmented generation conditions output on retrieved documents [3], and uncertainty-aware retrieval risk control addresses the possibility that retrieved evidence is unreliable or conflicting [8]. Claim-level factuality evaluation decomposes generated text into atomic statements and checks each against evidence [16]. Retrieval evaluation frameworks assess whether generated answers are grounded in retrieved context [17]. These approaches are necessary when the detector has access to external documents, but they introduce new dependencies: retrieval quality, claim decomposition accuracy, and entailment model reliability all affect the evidence score. The

proposed framework treats evidence support as one signal among several, with a reliability weight that can be reduced when retrieval ambiguity is high.

Internal-state methods provide a fourth family of signals. Probes trained on hidden representations can reveal whether a model's internal state is consistent with a truthful answer [11], [12]. These methods can be efficient at inference time, but they require access to model internals and may not transfer across architectures. Verbalized uncertainty provides a lower-cost alternative in which the model is prompted to express confidence in words [18]. However, verbalized confidence can be overconfident or poorly calibrated, and it is itself a generated output that may be hallucinated. The framework developed here does not assume that any single signal is authoritative; instead, it estimates signal reliability and uses that estimate to modulate fusion.

3. Multi-Stage Confidence Estimation Framework

The framework proceeds in four stages. The first stage computes sequence-level lexical confidence from token probabilities. The second stage samples multiple responses, clusters them by meaning, and computes semantic entropy and pairwise consistency. The third stage decomposes the response into atomic claims and estimates evidence support from retrieved documents. The fourth stage normalizes these signals, estimates their reliability, fuses them into a risk score, and applies an ambiguity-sensitive decision rule. The stages are sequential in computation but complementary in interpretation: lexical confidence captures model certainty, semantic consistency captures meaning-level stability, and evidence support captures external grounding.

3.1 Lexical confidence

The first signal is the geometric mean token probability of the generated sequence. For a response y of length |y|, the lexical confidence is defined as the exponentiated average log-probability:

$$C_lex(y|x) = exp((1/|y|)\sum_\{t = 1\}\^\{|y|\} log\ p\theta(y_t|x, y_ < t))$$

Here pθ denotes the language model's conditional distribution, y_t is the token at position t, and y_<t is the preceding token sequence. The geometric mean prevents long responses from being systematically penalized by sequence length, which is important when comparing responses of different lengths. C_lex lies in the interval (0,1] and is high when the model assigns high probability to each generated token. The signal is useful but insufficient: it measures fluency and model preference, not factual correctness. A model can assign high probability to a false claim that is common in its training distribution, and a correct but unusual claim can receive a lower score.

3.2 Semantic consistency analysis

The second stage samples m responses y_1,...,y_m from the model for the same prompt and groups them according to meaning. Let C(x) be the set of meaning clusters induced by an entailment-based equivalence relation, where two responses are placed in the same cluster if each entails the other under a specified entailment model. The cluster mass p(c|x) is the total probability or empirical frequency of responses assigned to cluster c. Semantic entropy is then:

$$H_sem(x) = -\sum_{c \in C(x)} p(c|x) \log p(c|x)$$

High semantic entropy indicates that the model distributes probability across multiple meanings, which is a warning sign for hallucination or ambiguity. Low semantic entropy indicates convergence on one meaning, but it does not guarantee correctness: the model may converge on the same false claim. For this reason, semantic entropy is paired with pairwise consistency. The consistency score is:

$$S_con(x) = \frac{2}{(m(m-1))} \sum_{1 \le i < j \le m} sim(y_i, y_j)$$

Here sim(y_i,y_j) is a normalized semantic similarity or bidirectional entailment score between two sampled responses, and the double sum averages over all unordered pairs. S_con is high when samples agree in meaning and low when they diverge. The combination of H_sem and S_con separates two cases that a single consistency score would conflate: convergence on one meaning versus dispersal across many meanings. The semantic clustering step is critical because surface-level string overlap can misclassify paraphrases as disagreement and can treat contradictory statements with similar wording as agreement.

3.3 Evidence support

The third stage evaluates whether the claims in y are supported by retrieved documents D(x). The response is decomposed into atomic claims A(y), and each claim a is compared with each document d using an entailment score e(a,d) in [0,1]. The evidence support score is:

$$S_ev(x, y) = 1/|A(y)|\sum_\{a \in A(y)\} max_\{d \in D(x)\} e(a, d)$$

The maximum over documents reflects the assumption that a claim needs at least one supporting passage, while the average over claims prevents a long response from being dominated by a single well-supported sentence. This formulation is deliberately simple; it exposes the dependence on claim decomposition and entailment quality rather than hiding them inside a learned end-to-end model. If A(y) is empty because decomposition fails, the score is undefined and the evidence signal should be down-weighted rather than treated as zero. If D(x) is empty or retrieval is unreliable, S_ev should also receive a lower reliability weight. This is where uncertainty-aware retrieval risk control becomes relevant: the meaning of evidence support depends on the credibility and consistency of the retrieved context [8].

3.4 Reliability-weighted fusion

The four support signals are normalized to the interval [0,1]. Lexical confidence is already in this range; semantic entropy is converted to a support score by exponentiating its negative value; consistency and evidence support are used directly after normalization. Let s_1,...,s_K denote these normalized support scores, with K=4 in the full framework. The fused risk score is a reliability-weighted geometric pool:

$$R(x, y) = 1 - \prod_{k=1}^{K} s_k(x, y)^{w_k(x)}$$

The weights w_k(x) are nonnegative and sum to one. A low support score on any signal drives the product toward zero and the risk score toward one, which is appropriate when hallucination can be triggered by a single unsupported claim or a single inconsistent meaning. The geometric form also prevents a high lexical confidence from simply overriding weak semantic consistency or weak evidence support. The risk score is bounded in [0,1], but it should be interpreted as a decision score rather than a calibrated probability unless it is explicitly calibrated on held-out data.

The weights are adapted to estimator reliability. Let u_k(x) be a normalized reliability penalty for signal k, combining estimator variance, calibration residual, or evidence volatility. The weight is:

$$w_k(x) = \exp\frac{\left(-u_k(x)\right)}{\sum_{j=1}^{K}\exp\left(-u_j(x)\right)}$$

When a signal is unstable or poorly calibrated, u_k increases and its weight decreases. This design allows the detector to shift influence toward more reliable signals without hard-coding a fixed hierarchy. For example, under heavy retrieval noise, the evidence reliability penalty rises and the fused risk depends more on semantic consistency and lexical confidence. Under high sample correlation, the consistency penalty rises and the evidence signal receives more influence. The weights are context-sensitive but not arbitrary: they are derived from explicit reliability estimates that can be monitored and audited.

4. Adaptive Decision Rule and Error Analysis

A static risk threshold is poorly matched to hallucination detection because the acceptable level of risk depends on context. A legal or medical response may require a lower tolerance for unsupported claims than a brainstorming response, and a prompt with ambiguous evidence may require a more conservative threshold. The framework therefore uses an ambiguity-sensitive acceptable risk:

$$\tau(x) = \tau 0 - \gamma\nu(x)$$

Here τ0 is a baseline acceptable risk, ν(x) is a normalized ambiguity or evidence-volatility score in [0,1], and γ is a nonnegative sensitivity parameter bounded so that τ(x) remains in [0,1]. Higher ambiguity lowers the acceptable risk, making acceptance harder. The ambiguity score can combine semantic entropy, retrieval disagreement, and claim coverage; the exact composition is a design choice that should be documented because it determines the operating point. This threshold is not a claim that the risk score is a true probability. It is a policy device that translates context-sensitive caution into a decision boundary.

The decision rule separates three outcomes: acceptance, flagging for verification, and rejection. Let τ_l(x) and τ_h(x) be lower and upper thresholds with τ_l(x)≤τ_h(x), and let $1\{\cdot\}$ denote the indicator function. The decision is:

$$D(x,y) = 1\{R(x,y) \le \tau_l(x)\} - 1\{R(x,y) \ge \tau_h(x)\}$$

A value of 1 indicates acceptance, 0 indicates flagging for review or additional evidence, and −1 indicates rejection or abstention. The intermediate region is important because many hallucination decisions are not best treated as binary. When evidence is incomplete, a flagging decision can trigger retrieval, claim verification, or human review without discarding a potentially correct response. The thresholds can be set from cost-sensitive validation, where the cost of false acceptance is compared with the cost of false rejection and the cost of deferral. In deployment, τ_l and τ_h should be recalibrated when the retrieval corpus, model version, or task distribution changes.

The error analysis follows from the signal structure. False acceptance occurs when R(x,y)≤τ_l(x) but H=1. This is most likely when lexical confidence is high, samples are correlated and agree on a false claim, and retrieval provides misleading support. False rejection occurs when R(x,y)≥τ_h(x) but H=0. This is most likely when the response is correct but unusual, retrieval is sparse, or semantic clustering incorrectly separates

paraphrases. The flagging region absorbs some of this ambiguity, but it also imposes a verification cost. The framework does not eliminate these trade-offs; it makes them explicit and adjustable.

The reliability weights introduce a second-order error mode. If u_k is misspecified, the fusion can over-trust a weak signal. For example, if consistency is estimated from highly correlated samples, S_con will be high even when the model is systematically wrong. The reliability penalty for consistency should therefore include a diversity measure, such as the effective number of independent samples or the entropy of sampling parameters. Similarly, if evidence support is computed with an entailment model that over-accepts, S_ev will be inflated. The framework's modularity is an advantage here: each signal can be validated separately before being fused, and each reliability penalty can be monitored as a diagnostic.

5. Technical Discussion and Boundary Conditions

The proposed framework is designed for settings in which the detector can observe generated text, sample multiple responses, and access retrieved context. Its main advantage is that it does not treat hallucination as a single scalar prediction problem. Lexical confidence, semantic consistency, and evidence support capture different failure mechanisms, and reliability-weighted fusion allows the detector to adapt when one mechanism is unreliable. The three-way decision rule further aligns detection with risk management: uncertain cases can be deferred rather than forced into a binary accept-reject decision.

Several boundary conditions limit the interpretation of the framework. First, semantic clustering depends on an entailment model or similarity function. If the clustering model merges distinct claims, semantic entropy will be underestimated and consistency will be overestimated. If it splits paraphrases, consistency will be underestimated and correct responses may be flagged. The choice of equivalence relation is therefore a substantive design decision, not a preprocessing detail. Bidirectional entailment is preferable to cosine similarity because it is more sensitive to contradiction, but it introduces its own errors. A practical implementation should include a clustering-confidence penalty in u_k for the consistency signal.

Second, sampled responses are not independent draws from a well-specified posterior. They are generated by the same model with the same weights and often with similar decoding parameters. Consistency across samples can therefore reflect shared training biases rather than truth. This is why the framework uses semantic entropy and evidence support in addition to consistency. The reliability weight for consistency should be reduced when sample diversity is low, when decoding temperature is near zero, or when the prompt is highly constrained. The framework's adaptive weighting is intended to handle this condition, but the diversity measure must be explicitly defined.

Third, retrieval adds a dependency on external evidence quality. A retrieved document can be outdated, irrelevant, or contradictory, and an entailment model can misread it. Uncertainty-aware retrieval risk control is relevant because it treats retrieval quality as part of the uncertainty calculation rather than as an external assumption [8]. In the present framework, retrieval ambiguity enters both the evidence support score and the adaptive threshold. When retrieval volatility is high, the threshold for acceptance should be lower and the evidence weight should be reduced. This prevents a confidently wrong retrieval from dominating the decision.

Fourth, calibration is distribution-dependent. A risk score that is well calibrated on one task may be poorly calibrated on another, and calibration can drift as the model or retrieval corpus changes [5], [15]. The framework should therefore be deployed with a calibration protocol: estimate reliability penalties and decision thresholds on held-out data from the target distribution, monitor false acceptance and false rejection rates after deployment, and recalibrate when drift is detected. This is not an empirical claim about a specific dataset; it is a design requirement that follows from the calibration literature.

Fifth, the framework has computational and latency costs. Sampling m responses and running semantic clustering and claim-level evidence checks can be expensive relative to a single forward pass. The multi-stage design allows cost control: lexical confidence is cheap, semantic consistency can be sampled at a lower rate for high-stakes queries, and evidence support can be computed only when the fused risk falls into the flagging region. This staged allocation is consistent with selective prediction: the detector spends more computation on cases where the decision is uncertain. The trade-off between detection quality and latency should be treated as an explicit operating parameter.

The framework also has implications for evaluation. A detector should be assessed not only by aggregate accuracy but by the risk it controls at a given acceptance rate. Metrics should separate false acceptance from false rejection, report performance in the flagging region, and evaluate calibration under distribution shift. Claim-level factuality metrics [16] and retrieval-augmented generation evaluation frameworks [17] can inform this assessment, while verbalized uncertainty [18] can serve as a baseline rather than as a ground truth. The key question is whether the detector's decisions are stable and interpretable when the underlying signals disagree.

6. Conclusion

This paper developed an adaptive hallucination detection framework for large language models that combines multi-stage confidence estimation with semantic consistency analysis. The framework treats hallucination as a latent relational property between a prompt, a generated response, and available evidence. It computes lexical confidence, meaning-cluster semantic entropy, sample-based semantic consistency, and claim-level evidence support; normalizes these signals; fuses them through reliability-weighted geometric pooling; and applies an ambiguity-sensitive three-way decision rule. The design addresses a practical weakness of single-signal detectors: high token likelihood, sample agreement, and retrieved evidence can each fail in characteristic ways, and their reliability varies with context.

The analytical contribution is the integration of uncertainty estimation, semantic clustering, evidence alignment, and selective prediction within one risk-control logic. The framework makes the trade-off between false acceptance, false rejection, and deferral explicit, and it provides a principled role for adaptive thresholds when retrieval ambiguity or semantic disagreement is high. Its boundary conditions are also clear: semantic clustering can misclassify meaning, sampled responses can be correlated, retrieval can be misleading, and calibration can drift. These conditions do not undermine the framework; they define the points at which reliability penalties and monitoring are needed. The result is a technically grounded architecture for hallucination detection as a context-sensitive risk-control layer rather than a static post hoc filter.

References

1. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

2. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1906–1919). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

3. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (Vol. 33, pp. 9459–9474). Curran Associates.

4. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., ... Kaplan, J. (2022). Language models (mostly) know what they know (arXiv:2207.05221). arXiv. https://arxiv.org/abs/2207.05221

5. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321–1330). PMLR.

6. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In International Conference on Learning Representations.

7. Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In International Conference on Learning Representations.

8. Zhang, R. (2025, December). Uncertainty-Aware RAG: Trustworthy Generation and Risk Control. In 2025 5th International Symposium on Artificial Intelligence and Big Data (AIBDF) (pp. 526-531). IEEE.

9. 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

10. Manakul, P., Liusie, A., & Gales, M. J. F. (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.

11. 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.

12. Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering latent knowledge in language models without supervision. In International Conference on Learning Representations.

13. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 1050–1059). PMLR.

14. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems (Vol. 30, pp. 6402–6413). Curran Associates.

15. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In Advances in Neural Information Processing Systems (Vol. 32, pp. 13991–14002). Curran Associates.

16. Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 12076–12100). Association for Computational Linguistics.

17. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (pp. 150–158). Association for Computational Linguistics.

18. Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. Transactions on Machine Learning Research. https://arxiv.org/abs/2205.14334