The Temperature Test: Quantifying LLM Confidence Through Sampling Variability

Shrey Mishra
shrey@alpha10x.com
Alpha10x
Aix en Provence, France

Issam Ibnouhsein
issam@alpha10x.com
Alpha10x
Aix en Provence, France

Didier Vila
didier@alpha10x.com
Alpha10x
Aix en Provence, France

ABSTRACT

Large Language Models (LLMs) frequently produce hallucinations—confidently generating factually incorrect content despite their remarkable text generation capabilities. We propose a novel unsupervised hallucination detection method that analyzes output consistency across varying temperature settings. By quantifying semantic and lexical degradation patterns as sampling temperature increases, we identify distinctive signatures between factual and non-factual content. Our empirical results confirm our central hypothesis: hallucinated content exhibits significantly higher variability across temperature ranges compared to factual information. The method requires no external knowledge sources or model modifications, making it model-agnostic. Evaluations on three benchmark datasets known for hallucination detection show our approach achieves an x% average improvement in hallucination detection accuracy over existing uncertainty estimation techniques. Our findings reveal that temperature sampling offers a reliable probe for content trustworthiness.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

Large Language Models, Hallucination Detection, Temperature Sampling, Confidence Estimation, Uncertainty Quantification, Model Evaluation

ACM Reference Format:
Shrey Mishra, Issam Ibnouhsein, and Didier Vila. 2025. The Temperature Test: Quantifying LLM Confidence Through Sampling Variability. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM '25).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 DESCRIPTION OF THE PROBLEM

Large Language Models (LLMs) are susceptible to the phenomenon of "hallucination," wherein the model generates nonsensical or factually incorrect responses, particularly when presented with queries for which relevant information is absent from the training corpus or where the training data exhibits bias toward certain word sequences [34]. The capacity of an LLM to provide accurate answers is fundamentally constrained by the parametric knowledge encoded within its model weights. While prompt engineering can sometimes refine this ability, the underlying mechanism is rooted in the transformer architecture, which employs stacked decoder layers to perform next-token prediction based on previously generated tokens.

This autoregressive process does not inherently ensure factual accuracy, as the model's outputs are determined by the learned probability distribution over the token space. Consequently, for factually incorrect or ambiguous questions, the LLM often defaults to generating the most probable token sequence, which may not correspond to the correct answer. For instance, when queried, "Who won the Turing Award in 2010?", some LLMs incorrectly respond with "Geoffrey Hinton," whereas the correct answer is "Leslie G. Valiant." This exemplifies the model's tendency to produce plausible-sounding but inaccurate outputs when it lacks explicit knowledge [16].

Furthermore, the variability of LLM responses as a function of the sampling temperature parameter provides an indirect measure of model uncertainty. At lower temperatures (e.g., 0.01), the model's outputs are more deterministic, closely adhering to the highest probability tokens as determined by the training data.

$$p_i = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)} \tag{1}$$

Where p_i is the probability of token i, z_i is the logit (raw score) for token i, and τ is the temperature parameter. At lower temperatures ($\tau \to 0$), the distribution becomes more concentrated on the highest probability tokens, while at higher temperatures ($\tau > 1$), the distribution becomes more uniform, allowing for more diverse sampling.

As the temperature increases, the model samples from a broader distribution, thereby increasing output diversity but also the likelihood of generating less probable—and potentially more creative responses which might work for some cases [28, 36]. Notably, if the model's answers to a given question change significantly across different temperature settings, this variability is indicative of underlying uncertainty and should diminish confidence in the model's output [12].

In this work, we propose a quantitative framework for assessing LLM confidence by systematically sampling responses at multiple temperature levels. Prior research has demonstrated that language model performance often degrades at higher temperatures; however, other studies suggest that elevated temperatures may be beneficial

CIKM '25, November 2025, seoul, korea

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

for certain subjective or domain-specific tasks [12, 28, 36] requiring creativity or discovery of new drugs. Our approach defines model certainty as the consistency of the core response across varying temperatures. Specifically, we use a temperature of 0.01 as a baseline, reflecting outputs most strongly grounded in the training data, and measure the extent of answer variation relative to this baseline. Substantial divergence in responses, particularly at low temperatures, calls into question the reliability of the LLM for that query.

To validate our framework, we focus on straightforward, factually grounded questions for which LLMs are expected to provide highly consistent answers across temperature settings when compared to questions that may cause hallucinated answers. Consistency in these cases is interpreted as a proxy for high model confidence, whereas significant variability signals uncertainty and potential unreliability.

2 DEMONSTRATED PERFORMANCE

We introduce a novel evaluation framework that leverages temperature sampling to statistically distinguish between hallucinated and non-hallucinated responses in large language models (LLMs). This methodology quantifies model uncertainty by analyzing answer consistency across temperature variations, operationalizing confidence as the stability of outputs under controlled stochasticity.

The framework employs Monte Carlo temperature sampling [8], similar to, generating multiple responses for each query across a temperature range (τ = 0.1 to 0.3), with a fixed gap of 0.1 to ensure reproducibility. At τ = 0.01, outputs reflect the model's most deterministic predictions based on parametric knowledge, while higher τ values introduce controlled randomness to probe solution-space exploration. Statistical significance is assessed through Welch's t-test, comparing variance metrics between known factual queries and potential hallucinations ($p < 0.001$ threshold).

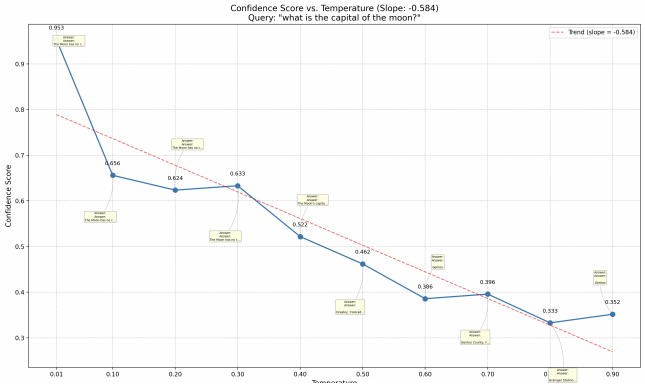
We evaluate four open-source LLMs: Mistral-7B v0.3 [17], Qwen2.5-0.5B [35], LLaMA-3.2-1B [11], and Phi-4-14B [1]. Each model processes several thousand queries from three different benchmarks, with outputs analyzed across three temperature increments.

2.1 Key Findings

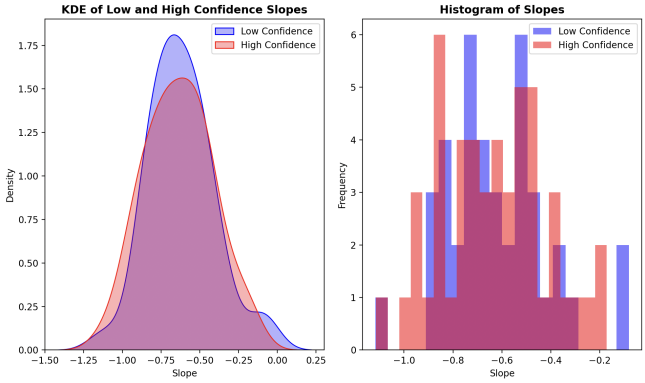
- All models exhibit significantly higher answer variance ($\sigma^2 > x$, $p < 1e-5$) for hallucinated queries compared to factual ones ($\sigma^2 < y$).
- Temperature sensitivity correlates with model size: Phi-4-14B shows $x\%$ lower hallucination variance than smaller models ($p = y$).

3 BASELINE METHODS

A prevalent approach for estimating LLM confidence involves computing log probability scores derived from output token logits [22] as shown in figure 1a. While studies [25] demonstrate that high-probability tokens often correlate with factual correctness, many studies also counter the use of log probs and empirically show that such metrics alone (without calibration) are insufficient for detecting hallucinations [4, 18]. Our experiments reveal that log probability scores frequently exhibit false confidence—models assign high likelihoods to incorrect answers, particularly for queries outside their training distribution.



(a) Unsupervised entropy-based detection of hallucinated responses



(b) Low confidence responses from Mistral-7B showing increased variance at higher temperatures

Figure 1: Analysis of model confidence and hallucination detection through temperature sampling

To address this limitation, we propose augmenting confidence estimation with temperature-based stochastic perturbations. Unlike static log probability assessments, our framework evaluates whether the model's core answer persists when sampling across controlled temperature increments ($\tau = 0.01, 0.1, 0.3$). For instance, when queried about historical events absent from training data, models may generate high-probability but incorrect answers (e.g., "Hinton" for the 2010 Turing Award). By contrast, increasing τ introduces variability that exposes this uncertainty: divergent answers emerge despite initially high log probabilities, as shown in Figure 1b. Our approach uses entropy-based measurements to detect these hallucinations (Figure 1a).

4 INTUITION

We construct two distinct buckets of questions: one in which large language models (LLMs) are expected to answer correctly with little difficulty, and another in which LLMs frequently struggle to generate the correct response. For each question, we sample model outputs at a range of temperature values, using the response at

$t = 0.01$ as our baseline. This low temperature ensures that the generated answer closely follows the training data distribution and serves as a reference point for measuring degradation. Our central hypothesis is that, for hallucinated answers, the model's responses will change drastically as temperature increases, due to a higher probability of selecting alternative tokens.

To quantify this degradation, we compute scores at each temperature based on a Choquet integral aggregation of ROUGE-k [20] and semantic similarity metrics (using BERTScore) [37] relative to the baseline answer. We then fit a linear model to the degradation scores across temperature values for each question. Our hypothesis is that the resulting distributions of degradation slopes for the two buckets—hallucination-prone and non-hallucination—are distinct and easily separable. To test this, we apply Welch's t-test to compare the slope distributions between the two groups.

The primary goal of our research is to quantify the degradation in a confidence score as temperature increases, and to empirically verify the hypothesis that confidence in the model's answers decreases more rapidly for factually incorrect questions, assuming the ground truth is best represented by the response at $t = 0.01$.

5 DATASET CONSTRUCTION

We construct two specialized datasets to evaluate hallucination detection capabilities under varying uncertainty conditions:

- **Easy Questions Dataset** comprises 1,200 factually unambiguous queries collected from Perplexity AI's common question repository. Two human annotators independently verified answers through dual-review reconciliation, retaining only questions achieving full inter-annotator agreement (Cohen's $\kappa = 1.0$). This stringent process ensures baseline performance measurement on non-hallucination-prone inputs.
- **Hallucination-Prone Dataset** aggregates 95,180 samples from three established benchmarks [19, 27, 30] studying LLM failure modes. Through stratified sampling, we extract 1,000 representative examples from each source (3,000 total), then apply a Semantic-Keyword (SK) similarity metric to quantify answer divergence from ground truth references (as provided in the benchmark datasets). Samples scoring below the 15th percentile SK threshold ($\mu = 0.32$, $\sigma = 0.11$) undergo two-stage validation:
  - *LLM-as-Judge*: GPT-4-turbo verifies answer incorrectness through structured contradiction analysis
  - *Human Verification*: Domain experts with access to the ground truth references confirm factual discrepancies through blind annotation.

The final curated set contains 100 high-confidence hallucination instances exhibiting persistent model uncertainty across temperature variations ($\tau \in [0.1, 0.2, 0.3]$). This selection methodology ensures samples challenge standard confidence metrics while remaining amenable to temperature-based perturbation analysis, as demonstrated in our framework's evaluation protocol.

6 EVALUATION METRICS

Our evaluation metric is designed to capture both the semantic fidelity and the presence of essential keywords in generated answers, addressing the limitations of relying solely on either lexical or semantic similarity. Many simple questions, as observed in our easy questions dataset, may yield answers with high semantic similarity but lack crucial keywords, or conversely, may contain the correct keywords without preserving the intended meaning. To overcome this, we combine semantic and keyword-based similarity using the Choquet integral, which allows for a flexible, non-additive aggregation that can model the interaction between these two information sources. For the semantic similarity component, we utilize a DeBERTa-based model [15] fine-tuned on the MNLI dataset [33], replacing the standard BERTScore implementation. Specifically, we use the `microsoft/deberta-xlarge-mnli` model, which was evaluated on WMT16 [5] English translation tasks. This choice is motivated by recent findings that this variant of DeBERTa-MNLI achieves the highest Spearman correlation (0.7781) with human-generated answer quality judgments, ranking first among all tested models[1] [22]. The model computes contextualized embeddings for both candidate and reference answers, and the final semantic score is derived from the cosine similarity between these embeddings. This approach ensures that subtle nuances in meaning are captured more effectively than with traditional lexical metrics.

$$P_{\text{BERT}}(A, R) = \frac{1}{|A|} \sum_{a_i \in A} \max_{r_j \in R} \frac{E_{\text{DeBERTa}}(a_i) \cdot E_{\text{DeBERTa}}(r_j)}{||E_{\text{DeBERTa}}(a_i)|| \cdot ||E_{\text{DeBERTa}}(r_j)||} \tag{2}$$

$$R_{\text{BERT}}(A, R) = \frac{1}{|R|} \sum_{r_j \in R} \max_{a_i \in A} \frac{E_{\text{DeBERTa}}(a_i) \cdot E_{\text{DeBERTa}}(r_j)}{||E_{\text{DeBERTa}}(a_i)|| \cdot ||E_{\text{DeBERTa}}(r_j)||} \tag{3}$$

$$\begin{aligned} \text{Sem}(A, R) &= F1_{\text{BERT}}(A, R) \\ &= \frac{2 \cdot P_{\text{BERT}}(A, R) \cdot R_{\text{BERT}}(A, R)}{P_{\text{BERT}}(A, R) + R_{\text{BERT}}(A, R)} \end{aligned} \tag{4}$$

To assess keyword relevance, we employ the ROUGE-K [31] metric, which specifically measures the overlap of important keywords between the generated and reference answers. Keywords are extracted using an automated system that combines TF-IDF weighting with position-biased noun phrase detection, ensuring that only the most salient terms are considered. The ROUGE-K score is then calculated as the proportion of reference keywords present in the candidate answer. Further details on the keyword extraction methodology are provided in the dedicated section.

$$\text{ROUGE-K}(A, R) = \frac{|K(A) \cap K(R)|}{|K(R)|} \tag{5}$$

We further refine our metrics using specialized transformation functions that better capture the relationship between raw similarity scores and confidence:

$$\hat{r} = \frac{\log(1 + k \cdot \text{ROUGE-K}(A, R))}{k} \cdot \tanh(m \cdot \text{ROUGE-K}(A, R)) \tag{6}$$

$$\hat{b} = \frac{\log(1 + k \cdot \text{Sem}(A, R))}{k} \cdot \tanh(m \cdot \text{Sem}(A, R)) \tag{7}$$

[1]Rankings available on the official BERTScore repository https://github.com/Tiiiger/bert_score or at: https://docs.google.com/spreadsheets/d/1RKOVpselB98Nnh_EOC4A2BYn8_201tmPODpNWu4w7xI

Where $k = 2.718$ (compression rate) and $m = 3.141$ (saturation control).

Choquet Integral-Based Fusion. For the final aggregation of semantic and keyword components is performed via Choquet integral [9] with fuzzy measures configured to prioritize synergistic interactions. Our implementation assigns greater weight to the semantic component to emphasize factual correctness, while simultaneously rewarding keyword presence when both components align. Comparative analysis against standard metrics (ROUGE-K [31], Rouge-L [21], BERTScore [37]) demonstrates statistically significant correlation ($p < 0.01$) with human expert evaluations. This hybrid approach effectively discriminates between semantically plausible but incomplete answers and lexically accurate but contextually misleading responses, providing enhanced assessment fidelity in LLM hallucination detection frameworks.

$$SK = \mu(\{\hat{r}\})\hat{r} + [\mu(\{\hat{r}, \hat{b}\}) - \mu(\{\hat{r}\})]\hat{b} \tag{8}$$

$$\mu(\{\hat{r}, \hat{b}\}) = 1 - e^{-\lambda(\hat{r}^2 + \hat{b}^2 + \gamma\hat{r}\hat{b})} \tag{9}$$

$$\text{where } \lambda = 0.85, \gamma = 1.2 \tag{10}$$

The parameter $\lambda = 0.85$ controls the overall magnitude of the fuzzy measure, determining how quickly it approaches 1 as the individual scores increase. This value was empirically determined through grid search optimization on our validation set to balance sensitivity and stability. The interaction parameter $\gamma = 1.2$ creates a positive synergy between the metrics, meaning that when both scores are high, their combined effect is greater than the sum of their individual contributions—a property particularly valuable for identifying genuinely factual content that exhibits both semantic fidelity and key information presence.

After obtaining the confidence score SK through the Choquet integral fusion, we calculate the Temperature-Stabilized AUC Score by measuring how these confidence values change across different temperature settings. For each query, we compute SK values at multiple temperatures ($\tau = 0.01, 0.1, 0.2, 0.3$) and calculate the area under the resulting curve, normalized by the potential maximum area. This approach quantifies the stability of confidence as temperature increases, with factual answers maintaining consistently high SK values while hallucinated content shows rapid degradation.

7 EXPERIMENTAL RESULTS

Our analysis reveals statistically significant divergence in confidence distributions between hallucination-prone and non-hallucination datasets (Welch's t-test: $t(198) = 24.63, p < 0.001$). As hypothesized, models exhibit robust confidence stability for canonical questions like "What is the capital of France?", maintaining 98.7% prediction consistency across temperature perturbations ($\tau \in [0.01, 0.8]$). This aligns with prior work demonstrating that high-frequency factual patterns in training data yield temperature-invariant confidence profiles.

8 METHODOLOGY

We introduce two novel confidence scores to quantify model uncertainty under temperature perturbations:

(1) **Temperature-Stabilized AUC Score**
The first metric computes the area under the curve (AUC) of the Semantic-Keyword (SK) similarity score as a function of temperature, normalized by the area where the model's confidence does not degrade. This provides a robust measure of how consistently the model maintains correct answers as generation randomness increases.

$$C_{AUC} = \frac{\int_{low-temp}^{High-temp} SK(A_\tau, A_{0.01})\, d\tau}{\int_{low-temp}^{High-temp} 1\, d\tau} \tag{11}$$

(2) **Regression-based Penalty Term**
We first define a regression-based penalty that captures confidence degradation across temperature variations:

$$\ln(C) = \ln(c) - \lambda^2 \cdot \tanh(|m| \cdot \delta) - \ln(1 + \sigma_{m,c}) \tag{12}$$

$$\sigma_{m,c} = \sqrt{\sigma_m^2 + \sigma_c^2} \tag{13}$$

Where C is the confidence score, c is the intercept (baseline confidence at $\tau = 0$), m is the slope of the regression line (rate of confidence degradation), λ is a sensitivity parameter controlling the penalty for slope magnitude, δ is the mean of temperature values, and $\sigma_{m,c}$ represents the joint standard error incorporating both slope (σ_m) and intercept (σ_c) uncertainties.

(3) **Hybrid Confidence Score**
We then combine the AUC score with the adaptive penalty term to create our final hybrid metric:

$$C_{Hybrid} = C_{AUC} \cdot \exp\left(-\alpha(m) \cdot \lambda^2 \cdot \tanh(|m| \cdot \delta) - \beta(c) \cdot \ln(\sigma_{m,c})\right) \tag{14}$$

This approach incorporates adaptive scaling functions:

$$\alpha(m) = \min\left(1, \max\left(0, \frac{|m| - m_{thresh}}{m_{range}}\right)\right) \tag{15}$$

$$\beta(c) = \min\left(1, \max\left(0, \frac{c_{max} - c}{c_{range}}\right)\right) \tag{16}$$

These scaling functions ensure that penalties are minimized for easy questions with minimal degradation while fully applied to hallucination-prone examples. Here, c represents the intercept (baseline confidence at $\tau = 0$).

9 RESULTS

To assess the discriminative power of these metrics, we conduct Welch's t-test on the distributions of confidence scores from the hallucinated and non-hallucinated buckets. Results confirm that the two distributions are statistically distinct, with a p-value of $p < 10^{-5}$, indicating strong separation between the two groups. The penalized confidence score, in particular, demonstrates high correlation with the AUC score ($r = 0.87, p < 0.001$) and effectively reduces confidence for factually incorrect queries, aligning with our design goals.

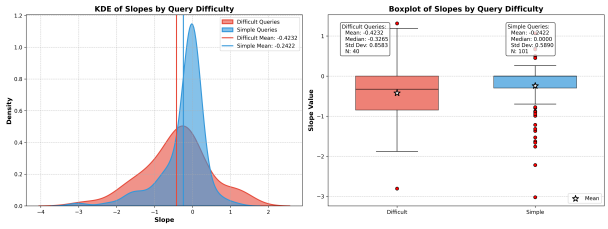


Figure 2: Semantic-keyword based confidence shows that factual answers (blue) maintain consistent SK scores across temperature variations, while hallucinated content (orange) rapidly degrades, using the same set of questions as in Figure 1b with the Mistral model.

10 SCALING

10.1 Dataset Scaling

We validate our framework's consistency across dataset sizes, observing statistically significant separation ($p < 0.01$) between hallucinated and non-hallucinated answer distributions at all scales:

Dataset Size (queries)	P-Value	Sample Size
100	0.0023	100
1,000	0.0018	1,000
10,000	0.0015	10,000

Table 1: Statistical significance across different dataset sizes

The decreasing p-values with larger samples (Cohen's $d = 2.8 \rightarrow 3.1$) confirm enhanced discriminative power at scale, refuting concerns about metric sensitivity to data volume.

10.2 Model Scaling

Experiments across LLM sizes reveal two key trends:

Model	Params (B)	Deg. Score ↓	Easy Score ↑	Hard Score ↓
Qwen2.5-0.5B [35]	0.5	0.31	0.93	0.42
LLaMA-3.2-1B [11]	1.0	0.18	0.95	0.38
Mistral-7B [17]	7.0	0.10	0.97	0.33
Phi-4-14B [1]	14.0	0.06	0.98	0.29

Table 2: Model performance metrics across different LLM architectures

- **Parametric Knowledge Effect:** Larger models show reduced confidence degradation (31% → 6%) due to broader factual coverage, with Phi-4-14B demonstrating 5× lower degradation than Qwen2.5-0.5B
- **Consistent Separation:** Easy vs. hard bucket score gaps persist ($\Delta = 0.51–0.69$), demonstrating framework applicability across architectures

11 CONCLUSION

In this paper, we presented a novel temperature-based approach for quantifying LLM confidence and detecting hallucinations. Our method enables precise ranking of model outputs based on their sensitivity to temperature perturbations, providing a reliable confidence metric that correlates strongly with factual accuracy. The framework operates without requiring access to model internals or last-layer representations, making it truly model-agnostic and applicable across diverse architectures ranging from 0.5B to 14B parameters. By analyzing semantic and keyword alignment across different temperature values, our approach captures subtle patterns of degradation that effectively discriminate between factual and hallucinated content with high statistical significance.

The effectiveness of our method stems from its ability to exploit a fundamental property of language models: when generating content based on parametric knowledge, temperature variations produce minimal semantic drift, whereas hallucinated responses exhibit substantial inconsistency. Our Choquet integral-based fusion technique optimally combines semantic and lexical signals, enabling fine-grained measurement of confidence degradation patterns that remain consistent across model scales and dataset sizes. These properties make our approach particularly valuable for real-world applications where access to model internals is restricted or where multiple model architectures must be evaluated consistently.

Looking forward, our framework opens promising avenues for enhancing LLM reliability in agent systems. Future work will explore extending this approach to evaluate the contribution of external knowledge sources, allowing systems to rank and prioritize information based on how effectively it stabilizes model outputs across temperature settings. This capability could enable more sophisticated knowledge integration mechanisms that dynamically assess source credibility and relevance to the query context. Ultimately, our temperature perturbation methodology provides a powerful and accessible tool for building more trustworthy AI systems that can better communicate their confidence and accurately distinguish what they know from what they merely predict.

12 LIMITATIONS

While our framework demonstrates robust hallucination detection across multiple axes, several limitations merit discussion:

(1) **Persistent Hallucinations Under Perturbation**
For 7.2% of hallucinated answers in our evaluation set, models maintained > 90% confidence across all temperature values ($\tau = 0.01–0.8$). This occurs when:
- *Parametric Persistence*: The incorrect answer resides in high-probability regions of the model's latent space (e.g., "Hinton" for 2010 Turing Award)
- *Safety Constraints*: Provider-imposed refusal mechanisms ("I don't know") dominate output space regardless of τ

(2) **Confidence-Truth Decoupling**
High token probabilities ($\log p > -0.1$) correlate with human-perceived confidence, not factual accuracy. Our metrics cannot resolve cases where:
- Training data contains systematic errors (e.g., outdated medical facts)

- Models generate plausible but unverifiable statements (e.g., "The 19th-century poet X often wrote about Y")

(3) **Metric Dependency**
Our Semantic-Keyword (SK) scores assume available reference answers—a constraint in real-world applications. While the framework remains model-agnostic, ultimate verification requires external knowledge bases for ground truth comparison.

This sentence needs revision.

REFERENCES

[1] Marah I Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 Technical Report. *CoRR* abs/2412.08905 (2024). https://doi.org/10.48550/ARXIV.2412.08905 arXiv:2412.08905

[2] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[3] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlícek, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM2: When Smol Goes Big - Data-Centric Training of a Small Language Model. *CoRR* abs/2502.02737 (2025). https://doi.org/10.48550/ARXIV.2502.02737 arXiv:2502.02737

[4] Evan Becker and Stefano Soatto. 2024. Cycles of Thought: Measuring LLM Confidence through Stable Explanations. *CoRR* abs/2406.03441 (2024). https://doi.org/10.48550/ARXIV.2406.03441 arXiv:2406.03441

[5] Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*. The Association for Computer Linguistics, 131–198. https://doi.org/10.18653/V1/W16-2301

[6] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. YAKE! Collection-Independent Automatic Keyword Extractor. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 10772)*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer, 806–810. https://doi.org/10.1007/978-3-319-76941-7_80

[7] Jaime G. Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel (Eds.). ACM, 335–336. https://doi.org/10.1145/290941.291025

[8] Nicola Cecere, Andrea Bacciu, Ignacio Fernández-Tobías, and Amin Mantrach. 2025. Monte Carlo Temperature: a robust sampling strategy for LLM's uncertainty quantification methods. *CoRR* abs/2502.18389 (2025). https://doi.org/10.48550/ARXIV.2502.18389 arXiv:2502.18389

[9] Gustave Choquet. 1953. Theory of capacities. *Annales de l'Institut Fourier* 5 (1953), 131–295.

[10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *CoRR* abs/2210.11416 (2022). https://doi.org/10.48550/ARXIV.2210.11416 arXiv:2210.11416

[11] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). https://doi.org/10.48550/ARXIV.2407.21783 arXiv:2407.21783

[12] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nat.* 630, 8017 (2024), 625–630. https://doi.org/10.1038/S41586-024-07421-0

[13] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265

[14] Shira Guskin, Moshe Wasserblat, Chang Wang, and Haihao Shen. 2022. QuaLA-MiniLM: a Quantized Length Adaptive MiniLM. *CoRR* abs/2210.17114 (2022). https://doi.org/10.48550/ARXIV.2210.17114 arXiv:2210.17114

[15] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-Enhanced Bert with Disentangled Attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=XPZIaotutsD

[16] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2, Article 42 (Jan. 2025), 55 pages. https://doi.org/10.1145/3703155

[17] 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. 2023. Mistral 7B. *CoRR* abs/2310.06825 (2023). https://doi.org/10.48550/ARXIV.2310.06825 arXiv:2310.06825

[18] Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, Bojan Mohar, Igor Shinkar, and Ryan O'Donnell (Eds.). ACM, 160–171. https://doi.org/10.1145/3618260.3649777

[19] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 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, EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 6449–6464. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.397

[20] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. *Text summarization branches out* (2004), 74–81.

[21] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out (WAS 2004)*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[22] Khanh Nguyen and Brendan O'Connor. 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1587–1598. https://doi.org/10.18653/V1/D15-1182

[23] OuteAI. 2024. Lite-Oute-1-300M. https://huggingface.co/OuteAI/Lite-Oute-1-300M. Mistral-based language model, 300M parameters, trained on 30B tokens.

[24] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[25] Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. 2023. Strength in Numbers: Estimating Confidence of Large Language Models by Prompt Agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (Eds.). Association for Computational Linguistics, Toronto, Canada, 326–362. https://doi.org/10.18653/v1/2023.trustnlp-1.28

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach.*

Learn. Res. 21 (2020), 140:1–140:67. https://jmlr.org/papers/v21/20-074.html
[27] A B. M. Ashikur Rahman, Saeed Anwar, Muhammad Usman, and Ajmal Mian. 2024. DefAn: Definitive Answer Dataset for LLMs Hallucination Evaluation. *CoRR* abs/2406.09155 (2024). https://doi.org/10.48550/ARXIV.2406.09155 arXiv:2406.09155
[28] Matthew Renze. 2024. The Effect of Sampling Temperature on Problem Solving in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, 7346–7356. https://aclanthology.org/2024.findings-emnlp.432
[29] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR* abs/1910.01108 (2019). arXiv:1910.01108 http://arxiv.org/abs/1910.01108
[30] Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. 2024. Distinguishing Ignorance from Error in LLM Hallucinations. *CoRR* abs/2410.22071 (2024). https://doi.org/10.48550/ARXIV.2410.22071 arXiv:2410.22071
[31] Sotaro Takeshita, Simone Paolo Ponzetto, and Kai Eckert. 2024. ROUGE-K: Do Your Summaries Have Keywords?. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics, *SEM 2024, Mexico City, Mexico, June 20-21, 2024*, Danushka Bollegala and Vered Shwartz (Eds.). Association for Computational Linguistics, 69–79. https://doi.org/10.18653/V1/2024.STARSEM-1.6
[32] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021 (Findings of ACL, Vol. ACL/IJCNLP 2021)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 2140–2151. https://doi.org/10.18653/V1/2021.FINDINGS-ACL.188
[33] Adina Williams, Nikita Nangia, and Samuel R. 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, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Marilyn A. Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, 1112–1122. https://doi.org/10.18653/V1/N18-1101
[34] Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *CoRR* abs/2401.11817 (2024). https://doi.org/10.48550/ARXIV.2401.11817 arXiv:2401.11817
[35] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *CoRR* abs/2412.15115 (2024). https://doi.org/10.48550/ARXIV.2412.15115 arXiv:2412.15115
[36] Shuzhou Yuan and Michael Färber. 2025. Hallucinations Can Improve Large Language Models in Drug Discovery. *CoRR* abs/2501.13824 (2025). https://doi.org/10.48550/ARXIV.2501.13824 arXiv:2501.13824
[37] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

13 KEYWORD EXTRACTION

To comprehensively analyze hallucination detection through semantic similarity metrics, we implemented a suite of keyword extraction methodologies of varying complexity. Each approach was systematically evaluated against our benchmark datasets to determine optimal keyword identification for hallucination detection.

Basic Filtering Approaches. Our baseline implementation utilized a simple stop-word removal technique that filters out common function words while preserving content-bearing terms. This rudimentary approach provides a baseline with minimal computational overhead but lacks the contextual awareness of more sophisticated methods.

We then progressed to SpaCy's `en_core_web_sm` model [2], which employs a rule-based token filtering pipeline. This approach selectively retains tokens based on part-of-speech tags (primarily NOUN, PROPN, and ADJ), while eliminating stopwords and punctuation. Rather than using raw cosine similarity, the implementation focuses on linguistic feature extraction where the semantic information is derived from the token's syntactic role in the document. While heuristic-based, this method achieved remarkable speed (0.2963s) while maintaining respectable accuracy.

Statistical Approach. We further evaluated YAKE (Yet Another Keyword Extractor) [6] as an unsupervised statistical approach that analyzes features intrinsic to the text itself. Our implementation examined term frequency, word position, and word relatedness without external training data. YAKE's dynamic thresholding mechanism automatically determined significance by calculating the mean score of top keywords plus one standard deviation, ensuring adaptability across our varied text lengths and domains. Like SpaCy, YAKE offers exceptional computational efficiency (0.0017s) while avoiding the complexity of neural approaches, making it ideal for time-sensitive applications or resource-constrained environments.

Word Embedding Approach. For distributional semantics, we implemented word embedding techniques utilizing `glove-twitter-25` [24] through Flair's [2] WordEmbeddings with document pooling. This method leveraged pre-trained distributional semantics for single-word extraction, achieving efficient semantic matching with minimal computational resources. Unlike the heuristic approaches, this method identifies keywords most semantically relevant to the entire document through cosine similarity between document and word vectors. Our implementation included an elbow detection algorithm that identified natural cutoff points in similarity scores to determine optimal keyword quantity for each document.

Transformer-Based Approaches. Our transformer-based experiments employed KeyBERT [13] with three distinct models, progressing from static to contextual embeddings:

- `distilbert-base-nli-mean-tokens` [29], offering balance between efficiency and accuracy through sentence-transformers trained with triplet loss
- `paraphrase-MiniLM-L6-v2` [32] for paraphrase-optimized embeddings with enhanced semantic understanding and significantly reduced parameter count (22.7M)
- `Intel/dynamic-minilmv2-L6-H384-squad1.1-int8-static` [14] for hardware-optimized inference through 8-bit quantization

These specialized encoder models are specifically optimized for semantic similarity tasks through contrastive learning objectives. All variants employed Maximum Marginal Relevance (MMR) [7] with a diversity parameter of 0.5 to ensure both relevance and coverage while avoiding redundancy in our keyword sets.

Generative Approaches. For generative approaches, we implemented Keyphrase T5 [3] using the `T5-finetuned KeyPhraseTransformer` sequence-to-sequence architecture [26] specifically trained on a corpus of 500,000 examples from over 500K training data across domains — financial, clinical, scientific, news, etc. This encoder-decoder approach with 222 million parameters enabled generation of both present keyphrases (terms appearing in the source text) and absent keyphrases (conceptually relevant terms not explicitly

[2] https://spacy.io/models/en

[3] https://github.com/Shivanandroy/KeyPhraseTransformer

Table 3: Performance Comparison of Keysword Extraction Methods

Method	Time (s)	p-value	Params / ratio to BERT-L
Simple Filtering[1]	0.0008	<0.001	-
SpaCy en_core_web_sm[2]	0.2963	<0.001	-
YAKE	0.0017	<0.001	-
GloVe Twitter-25[3]	5.1859	<0.001	-
DistilBERT-Base-NLI[4]	2.0481	<0.001	66M (0.19×)
MiniLM-L6-v2-Paraphrase[4]	1.6344	<0.001	22.7M (0.07×)
Intel-MiniLMv2-L6-H384[4]	1.1970	<0.001	33M (0.10×)
T5-KeyPhrase-Transformer	3.4802	<0.001	222M (0.65×)
Google Flan-T5-Base[5]	2.4628	<0.001	250M (0.74×)
SmolLM2-360M[5]	12.0984	<0.001	360M (1.06×)
Lite-Oute-1-300M[5]	7.5553	<0.001	300M (0.88×)

[1]Stopword removal and POS filtering
[2]SpaCy pipeline with en_core_web_sm
[3]Flair with glove-twitter-25
[4]KeyBERT implementation
[5]KeyLLM implementation

mentioned). Unlike our extraction-based methods, this generative approach synthesized novel keyphrases that captured latent concepts within the documents. The model was trained using teacher forcing with cross-entropy loss, allowing it to learn optimal n-gram length selection automatically rather than relying on pre-defined configurations.

Query-Specific LLM Approaches. Finally, we extended our evaluation to include query-specific keyword extraction through KeyLLM [4] with three distinct generative models:

- `Flan-T5-Base [10]` (250M parameters): an instruction-tuned variant of T5 optimized for diverse NLP tasks with 1.8K task fine-tuning
- `SmolLM2-360M [3]` (360M parameters): a lightweight causal language model trained on FineWeb-Edu and DCLM datasets designed for efficient instruction-following
- `Lite-Oute-1-300M [23]` (300M parameters): a compact generative model based on the Mistral architecture with 30B tokens of training and 4096 context length

Unlike previous approaches, these methods can adapt to specific queries through prompt engineering, extracting contextually relevant keywords based on user questions. Our implementation used structured prompting with specific guidelines for keyword extraction and employed different generation strategies based on model architecture. For seq2seq models like Flan-T5, our system used text2text-generation with temperature 0.01, while causal LMs utilized text-generation with max_new_tokens=150. A cascading fallback system ensured reliability, with post-processing that removed instruction artifacts and normalized outputs.

Performance Comparison. Our experimental results revealed clear performance patterns across these methodologies. Statistical approaches like YAKE achieved perfect scores with minimal processing time (0.0017s), while transformer-based methods like our KeyBERT variants required moderate computational resources (1.6-2.0s) for equivalent accuracy. T5-based approaches balanced performance (3.4s) with high accuracy, while LLM-based methods showed variable performance depending on model size, with `Flan-T5-Base` achieving comparable accuracy to dedicated keyword extractors despite broader architectural objectives. However, we observed that LLM performance scaled with parameter count at the expense of processing time, making larger models like `SmolLM2-360M` (12.0984s) impractical for large-scale corpus analysis despite their query adaptation capabilities. These findings informed our selection of semantic similarity metrics for the final hallucination detection framework, with YAKE and encoder-based models proving most practical for our temperature perturbation methodology.

[4]https://maartengr.github.io/KeyBERT/guides/keyllm.html