LOGITSCOPE: A FRAMEWORK FOR ANALYZING LLM UNCERTAINTY THROUGH INFORMATION METRICS

Farhan Ahmed
IBM Research
farhan.ahmed@ibm.com

Yuya Jeremy Ong
Plastic Labs
yuya@plasticlabs.ai

Chad DeLuca
IBM Research
delucac@us.ibm.com

ABSTRACT

Understanding and quantifying uncertainty in large language model (LLM) outputs is critical for reliable deployment. However, traditional evaluation approaches provide limited insight into model confidence at individual token positions during generation. To address this issue, we introduce **LogitScope**, a lightweight framework for analyzing LLM uncertainty through token-level information metrics computed from probability distributions. By measuring metrics such as entropy and varentropy at each generation step, LogitScope reveals patterns in model confidence, identifies potential hallucinations, and exposes decision points where models exhibit high uncertainty, all without requiring labeled data or semantic interpretation. We demonstrate LogitScope's utility across diverse applications including uncertainty quantification, model behavior analysis, and production monitoring. The framework is model-agnostic, computationally efficient through lazy evaluation, and compatible with any HuggingFace model, enabling both researchers and practitioners to inspect LLM behavior during inference.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable capabilities in text generation (Vaswani et al., 2017; Brown et al., 2020; Allal et al., 2025), yet understanding when and why they produce uncertain, incorrect, or unexpected outputs remains challenging (Liang et al., 2023; Ji et al., 2023). Unlike traditional machine learning systems with well-defined prediction tasks and confidence scores, LLMs generate sequences autoregressively where uncertainty manifests across multiple dimensions: individual token probabilities, distribution shapes, and temporal patterns across generation steps. Quantifying this uncertainty is essential for reliable deployment (Naveed et al., 2025; Paleyes et al., 2022), debugging model failures, and understanding the boundaries of model knowledge.

Current approaches to uncertainty quantification in LLMs fall into two broad categories. First, aggregate evaluation metrics computed over benchmark datasets (Gao et al., 2024; Liang et al., 2023) provide limited insight into token-level behavior during individual inference runs. Second, semantic-level approaches such as self-consistency (Wang et al., 2023) or external verification models (Manakul et al., 2023; Min et al., 2023) require multiple forward passes or additional models, introducing computational overhead and their own sources of error. Neither approach provides direct, interpretable access to the model's internal uncertainty at each generation step.

We introduce **LogitScope**, a lightweight framework that addresses these limitations by analyzing token probability distributions during inference to quantify uncertainty through information metrics. At each generation step, language models produce a distribution over their vocabulary. LogitScope computes entropy, varentropy, and surprisal from these distributions to reveal meaningful patterns in model behavior. High entropy indicates broad uncertainty across many tokens, high varentropy suggests multimodal distributions where the model considers distinct alternatives, and high surprisal on selected tokens flags statistically unexpected outputs. These metrics require no labeled data, no additional model calls, and can be computed efficiently in real-time, making them applicable to both research and production settings. We release LogitScope as open-source software to enable the broader community to analyze LLM uncertainty in their applications[1].

[1]Code available at: `https://github.com/IBM/logitscope`

Our contributions are as follows:

- We present LogitScope, an open-source framework that quantifies token-level LLM uncertainty by computing information metrics such as entropy and varentropy from probability distributions. This requires no labeled data, additional models, or multiple forward passes.
- We demonstrate how these information metrics provide interpretable signals about model confidence, revealing patterns associated with hallucinations, decision points, and unexpected outputs across diverse applications including model analysis, debugging, and production monitoring.

2 Background and Related Work

Uncertainty Quantification in LLMs Recent work has explored various approaches to quantifying uncertainty in language models. Perplexity and token probability have long been used as confidence measures (Jelinek et al., 2005; Brown et al., 2020), but provide limited insight into distribution characteristics. Self-consistency methods (Wang et al., 2023) generate multiple outputs and measure agreement, but require multiple forward passes. Semantic uncertainty approaches (Kuhn et al., 2023) cluster model outputs in semantic space, but depend on external models and are computationally expensive. Entropy-based metrics have been explored in specific contexts (Malinin & Gales, 2018; Lin et al., 2024), but typically focus on individual metrics in isolation for specific tasks. Hallucination detection methods (Manakul et al., 2023; Ji et al., 2023; Min et al., 2023) similarly rely on multiple generations or external knowledge bases. LogitScope provides a unified framework for computing and analyzing multiple information metrics across diverse uncertainty quantification and model analysis applications.

Model Monitoring and Drift Detection The ML monitoring community has developed extensive tooling for supervised learning systems (Paleyes et al., 2022; Naveed et al., 2025), focusing on input drift, prediction drift, and performance degradation. However, these approaches assume access to labels and well-defined prediction tasks. For generative language models, defining appropriate monitoring signals is more challenging due to open-ended outputs and subjective quality assessment (Liang et al., 2023; Gao et al., 2024). Recent work has explored self-supervised signals for LLM monitoring, including consistency checks (Manakul et al., 2023; Wang et al., 2023), factuality scoring (Min et al., 2023), and attention pattern analysis (Darcet et al., 2024). These methods often require multiple forward passes (Wang et al., 2023) or external verification systems (Min et al., 2023; Manakul et al., 2023). LogitScope complements these approaches by providing lightweight, interpretable metrics that require no additional model calls or external knowledge bases.

3 Method

3.1 Information Metrics for Uncertainty Analysis

LogitScope computes metrics from the probability distribution $p_t = p(x_t|x_{<t})$ at each token position t (formal definitions in Appendix A):

- **Probability**: Direct confidence measure; the model's assigned probability for the selected token.
- **Surprisal**: Negative log-probability of the selected token; quantifies how unexpected the choice was given the context.
- **Entropy**: Sum of weighted surprisals across all tokens; measures overall uncertainty in the distribution.
- **Varentropy**: Variance of surprisal values; high varentropy with high entropy indicates multimodal distributions where the model considers distinct alternatives.
- **Skewentropy**: Distribution asymmetry; reveals whether probability mass is concentrated or dispersed.
- **Perplexity**: Exponential of average surprisal; provides cumulative sequence-level quality measure.

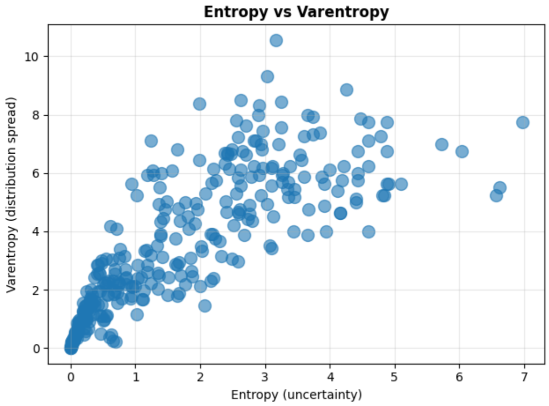

Figure 1: Entropy vs varentropy scatter plot. Each point represents a token position, with lower-left indicating confident predictions and upper-right indicating multimodal uncertainty.

Table 1: Average metrics for the Declaration of Independence preamble with original vs word-reversed order. Reversing the word order destroys coherence, dramatically increasing uncertainty.

Metric	Original	Reversed
Tokens	320	320
Characters	1628	1628
Entropy	1.70	6.33
Varentropy	3.46	8.63
Skewentropy	7.50	0.48
Perplexity	11.75	1833.76
Probability	0.55	0.09
Log Probability	-1.05	-2.70

3.2 Implementation

LogitScope is implemented as a lightweight wrapper around HuggingFace Transformers (Wolf et al., 2020). The core `LogitScope` class takes a tokenizer and model, performs inference on input text, and returns a `Results` object containing the probability distributions and lazy-evaluated metrics.

The framework uses two key design principles:

1. **Lazy evaluation**: Metrics are computed on-demand and cached, avoiding unnecessary computation when only specific metrics are needed.
2. **Zero-copy access**: Raw logits and probability distributions are accessible for custom analysis without copying data.

The framework supports CPU, CUDA, and Apple Silicon (MPS) acceleration and works with any HuggingFace model without modification.

4 Analysis and Applications

4.1 Uncertainty Pattern Analysis

We demonstrate LogitScope's analytical capabilities through two complementary views of the Declaration of Independence preamble (320 tokens; see Appendix B) processed by SmolLM2-135M-Instruct (Allal et al., 2025), a compact 135M parameter model. Figure 1 reveals token-level uncertainty patterns through entropy and varentropy, while Table 1 quantifies how destroying linguistic structure by reversing word order affects aggregate model confidence.

Token-level patterns Figure 1 shows that tokens cluster into distinct regions. *Low entropy and low varentropy* (bottom-left) represents confident predictions where the model assigns high probability to a single token. Common words and grammatically constrained positions fall into this region, indicating strong prior expectations. *High entropy and high varentropy* (top-right) indicates multimodal uncertainty, where the model distributes probability mass across multiple plausible alternatives. This pattern often appears at semantic decision points when multiple valid continuations exist. *Intermediate regions* reveal gradations of confidence between these extremes.

Aggregate patterns Table 1 compares the original and word-reversed text which reveals how linguistic structure affects model confidence. Reversing word order destroys both semantic coherence and syntactic dependencies, forcing the model to process grammatically invalid sequences. The reversed text shows significantly higher values for entropy (6.33 vs. 1.70), varentropy (8.63 vs

3.46), perplexity (1833.76 vs. 11.75), while average token probability drops from 0.55 to 0.09. This demonstrates LogitScope's ability to quantify how language structure influences model uncertainty across entire sequences.

4.2 Applications

Beyond understanding uncertainty patterns, LogitScope's information metrics support diverse practical workflows:

Hallucination detection: High entropy and varentropy regions often correlate with hallucinated content, as models exhibit uncertainty when generating facts beyond their training data. By flagging tokens with unusual metric patterns, practitioners can identify outputs requiring verification.

Model debugging: When models produce unexpected outputs, LogitScope reveals whether the issue stems from low-confidence predictions (high entropy), competition between alternatives (high varentropy), or statistically unlikely selections (high surprisal). This diagnostic information guides debugging efforts.

Prompt engineering: Comparing metric distributions across different prompt formulations reveals which prompts elicit more confident predictions on correct outputs. Effective prompts typically reduce entropy while maintaining high probability on expected tokens.

Production monitoring: Aggregate statistics (mean entropy, median surprisal, etc.) provide real-time signals about model behavior. Sudden shifts in these distributions can indicate input drift, adversarial inputs, or edge cases requiring attention, all without accessing ground truth labels.

Model comparison: LogitScope enables quantitative comparison of different models on the same inputs. Beyond accuracy metrics, practitioners can assess whether models differ in their uncertainty patterns, confidence calibration, or decision-making strategies.

5 Limitations and Future Work

Limitations: Information metrics provide signals about distribution characteristics and model uncertainty but do not directly measure semantic correctness or factual accuracy. High confidence (low entropy) does not guarantee correct outputs; models can be confidently wrong. Similarly, high entropy does not always indicate errors; it may reflect genuine ambiguity. LogitScope is best viewed as an analysis tool that reveals uncertainty patterns for human inspection, rather than an automated correctness verifier. The framework currently focuses on token-level analysis and does not capture longer-range semantic patterns or factual consistency.

Broader Impact: By providing interpretable signals about model uncertainty, LogitScope can help practitioners identify when models are operating outside their reliable range, potentially reducing the deployment of overconfident but incorrect outputs. However, the metrics can also be gamed through careful prompt engineering to artificially reduce entropy without improving actual correctness.

Future Directions: The framework can be extended with additional metrics tailored to specific failure modes, integrated into MLOps pipelines for automated anomaly detection, or combined with semantic clustering approaches for comprehensive uncertainty quantification. Large-scale empirical studies correlating metric patterns with human quality judgments across diverse domains and model architectures would further validate the utility of information uncertainty measures.

6 Conclusion

We presented LogitScope, a lightweight framework for analyzing LLM uncertainty through information metrics computed from token probability distributions. By measuring entropy, varentropy, and surprisal at each generation step, LogitScope enables researchers and practitioners to quantify model confidence, identify uncertainty patterns, and detect potential issues in real-time without requiring labeled data or additional model calls. We demonstrated the framework's utility across diverse applications and released it as open-source software to support the community in understanding and improving LLM reliability.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, 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. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 08 2005. ISSN 0001-4966. doi: 10.1121/1.2016299. URL https://doi.org/10.1121/1.2016299.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL https://arxiv.org/abs/2302.09664.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023. URL https://arxiv.org/abs/2211.09110.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024. URL https://arxiv.org/abs/2305.19187.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 7047–7058, Red Hook, NY, USA, 2018. Curran Associates Inc.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Hira Naveed, John Grundy, Chetan Arora, Hourieh Khalajzadeh, and Omar Haggag. Understanding Practitioners' Perspectives on Monitoring Machine Learning Systems . In *2025 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 743–754, Los Alamitos, CA, USA, sep 2025. IEEE Computer Society. doi: 10.1109/ICSME64153.2025.00074. URL https://doi.ieeecomputersociety.org/10.1109/ICSME64153.2025.00074.

Andrei Paleyes, Raoul-Gabriel Urma, and Neil D. Lawrence. Challenges in deploying machine learning: A survey of case studies. *ACM Comput. Surv.*, 55(6), December 2022. ISSN 0360-0300. doi: 10.1145/3533378. URL https://doi.org/10.1145/3533378.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

A Metric Definitions

At each generation step t, the model produces logits $\mathbf{z}_t \in \mathbb{R}^{|V|}$ over vocabulary V, which are transformed into a probability distribution via softmax:

$$p_t(v) = p(x_t = v|x_{<t}) = \frac{\exp(z_t^{(v)})}{\sum_{v' \in V} \exp(z_t^{(v')})} \tag{1}$$

LogitScope computes the following metrics from this probability distribution p_t at each token position t:

Probability is the direct confidence measure for the selected token:

$$P(x_t) = p_t(x_t) \tag{2}$$

High probability indicates confident selection, while low probability suggests the token was unlikely given the context.

Surprisal quantifies how unexpected the selected token x_t was:

$$I(x_t) = -\log p_t(x_t) \tag{3}$$

Surprisal is the fundamental building block for other metrics. High surprisal indicates the model assigned low probability to the token it ultimately selected, signaling statistically unlikely outputs.

Entropy measures the overall uncertainty in the distribution:

$$H(p_t) = -\sum_{v \in V} p_t(v) \log p_t(v) = \sum_{v \in V} p_t(v) \cdot I(v) \tag{4}$$

Entropy is the expected surprisal across all tokens. High entropy indicates broad uncertainty across many possible tokens, while low entropy suggests confident predictions concentrated on few tokens.

Varentropy measures the variance of surprisal values across the distribution:

$$\text{Var}(p_t) = \sum_{v \in V} p_t(v) \left(\log p_t(v)\right)^2 - H(p_t)^2 \tag{5}$$

High varentropy combined with high entropy indicates a multimodal distribution where the model is torn between multiple distinct options, which often occurs at decision points or when hallucinating.

Skewentropy measures the asymmetry of the surprisal distribution:

$$\text{Skew}(p_t) = \frac{\sum_{v \in V} p_t(v) \left(\log p_t(v) + H(p_t)\right)^3}{\text{Var}(p_t)^{3/2}} \tag{6}$$

Skewentropy reveals whether probability mass is concentrated (high absolute skew) or uniformly distributed (low skew), providing insight into distribution shape beyond entropy and varentropy.

Perplexity provides a cumulative measure of model performance over the sequence:

$$\text{PPL} = \exp\left(\frac{1}{T} \sum_{t=1}^{T} I(x_t)\right) \tag{7}$$

Perplexity exponentiates the average surprisal, yielding an interpretable measure of predictive quality. Lower perplexity indicates better model performance.

B Evaluation Text

All examples in this paper use the preamble of the United States Declaration of Independence as the evaluation text, processed by SmolLM2-135M-Instruct (Allal et al., 2025). The complete text (320 tokens when tokenized by SmolLM2) is:

> We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness.–That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, –That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness. Prudence, indeed, will dictate that Governments long established should not be changed for light and transient causes; and accordingly all experience hath shewn, that mankind are more disposed to suffer, while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, pursuing invariably the same Object evinces a design to reduce them under absolute Despotism, it is their right, it is their duty, to throw off such Government, and to provide new Guards for their future security.–Such has been the patient sufferance of these Colonies; and such is now the necessity which constrains them to alter their former Systems of Government. The history of the present King of Great Britain is a history of repeated injuries and usurpations, all having in direct object the establishment of an absolute Tyranny over these States. To prove this, let Facts be submitted to a candid world.

This text was chosen for its historical significance, formal register, complex syntactic structure, and moderate length suitable for demonstration purposes. The word-reversed version used in Table 1 reverses the order of words while preserving individual word spellings, creating grammatically invalid but tokenizable sequences.

C Interactive Web Interface

In addition to the Python library, LogitScope provides an interactive web-based interface for visual exploration of model uncertainty. The UI enables real-time analysis during text generation, displaying token-level metrics with color-coded visualizations and interactive controls. Users can switch between different metrics, inspect top-k alternative tokens at each position, and observe temporal patterns across the generation sequence.

Figures 2, 3, 4, and 5 show the interface analyzing the Declaration of Independence preamble using SmolLM2-135M-Instruct (Allal et al., 2025). The interface highlights tokens by their metric values, with color intensity indicating magnitude. The sidebar displays running statistics (mean, median, min, max) and allows users to toggle between different metrics. Clicking on individual tokens reveals the top-k alternatives the model considered at that position, along with their probabilities.

The web interface is particularly useful for:

- **Rapid prototyping**: Quickly test different prompts and observe their effect on model uncertainty without writing code.
- **Educational purposes**: Demonstrate model behavior to students or stakeholders through intuitive visual feedback.
- **Debugging**: Identify specific tokens or regions where models exhibit unexpected uncertainty patterns.
- **Comparative analysis**: Switch between metrics to understand different aspects of the same generation.

The interface is a component of the LogitScope framework and is launched directly from it. All models compatible with HuggingFace Transformers are supported with automatic device detection for CPU, CUDA, and Apple Silicon acceleration.

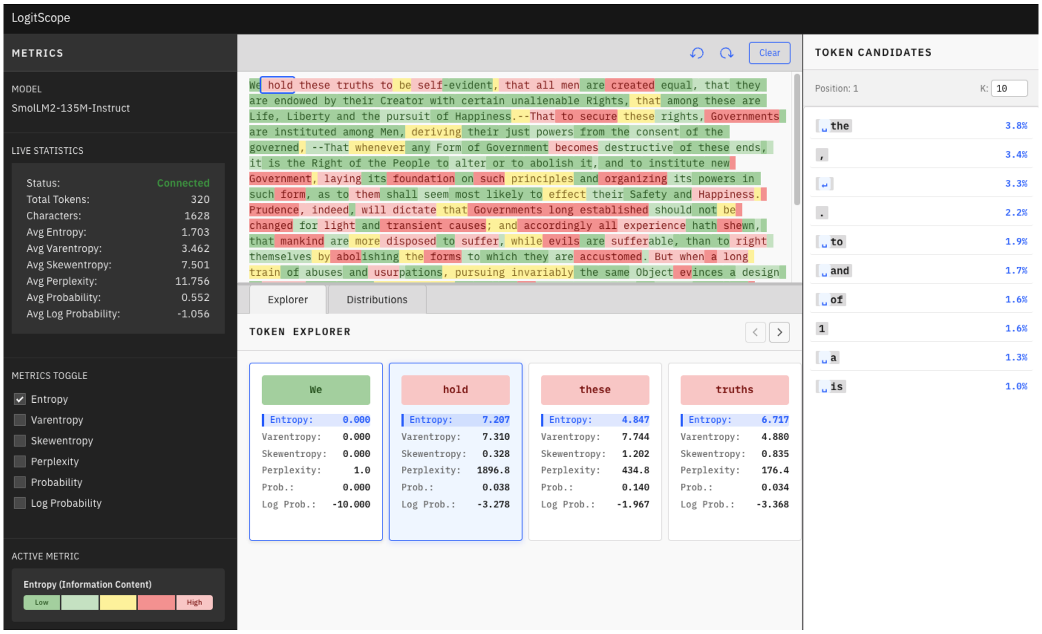

Figure 2: The web interface displaying the entropy metric. Tokens are color-coded by magnitude, with brighter colors indicating higher uncertainty. The sidebar shows aggregate statistics for quick assessment of overall model confidence.

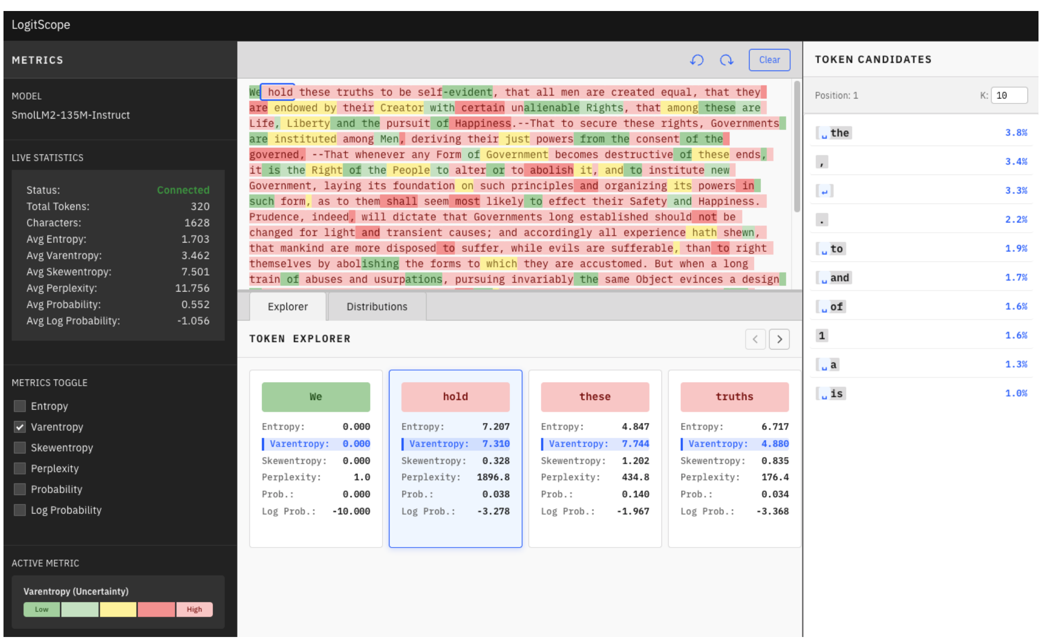

Figure 3: The web interface displaying the varentropy metric. Varentropy reveals multimodal distributions where the model considers multiple alternatives. High varentropy regions indicate decision points with competing continuations.

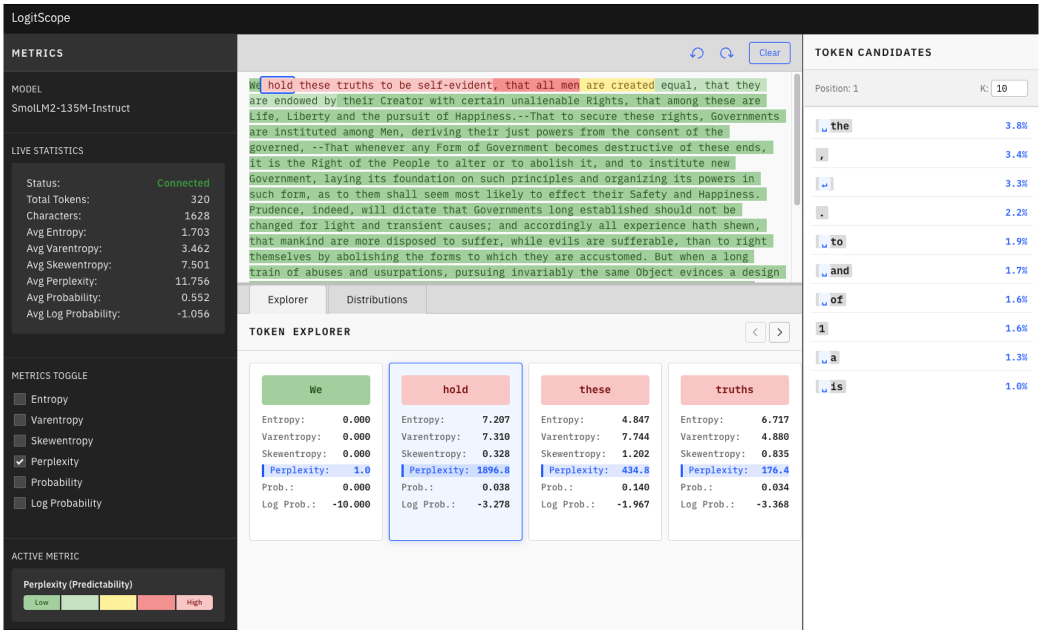

Figure 4: The web interface displaying the perplexity metric. Lower values indicate better predictive confidence. The interface enables identification of regions where the model exhibits uncertainty.

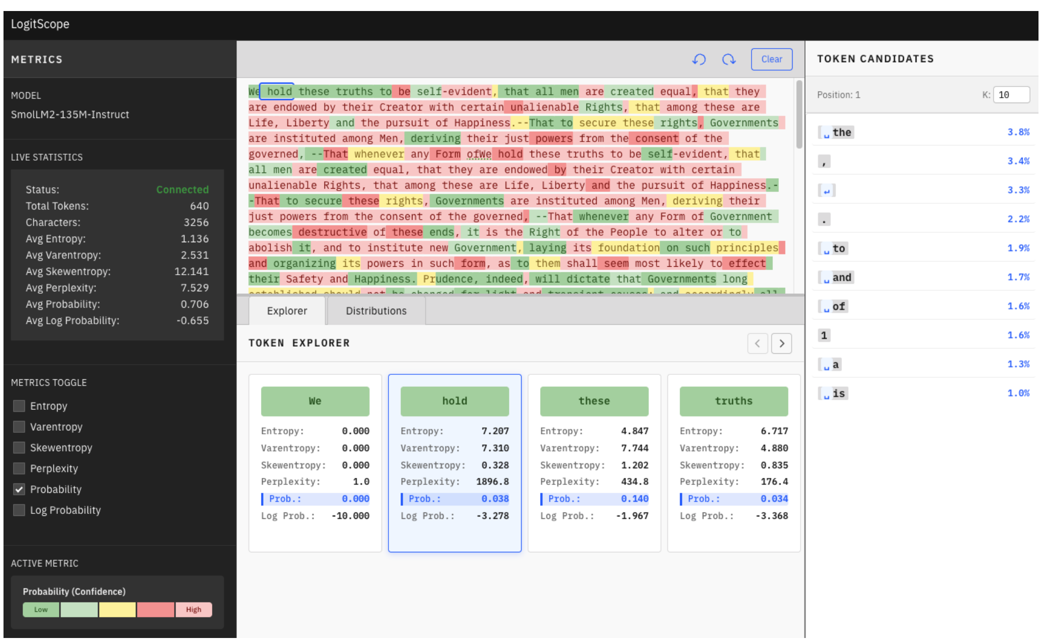

Figure 5: The web interface displaying the probability metric. Higher values indicate more confident token selections. The direct probability measure provides intuitive assessment of model certainty.