Embedding Trust: Semantic Isotropy Predicts Nonfactuality in Long-Form Text Generation

Dhrupad Bhardwaj [1] **Julia Kempe** [*1] **Tim G. J. Rudner** [*2 3]

Website Code

Abstract

To deploy large language models (LLMs) in high-stakes application domains that require substantively accurate responses to open-ended prompts, we need reliable, computationally inexpensive methods that assess the trustworthiness of long-form responses generated by LLMs. However, existing approaches often rely on claim-by-claim fact-checking, which is computationally expensive and brittle in long-form responses to open-ended prompts. In this work, we introduce semantic isotropy—the degree of uniformity across normalized text embeddings on the unit sphere—and use it to assess the trustworthiness of long-form responses generated by LLMs. To do so, we generate several long-form responses, embed them, and estimate the level of semantic isotropy of these responses as the angular dispersion of the embeddings on the unit sphere. We find that higher semantic isotropy—that is, greater embedding dispersion—reliably signals lower factual consistency across samples. Our approach requires no labeled data, no fine-tuning, and no hyperparameter selection, and can be used with open- or closed-weight embedding models. Across multiple domains, our method consistently outperforms existing aggregate trust signals in predicting nonfactuality using only a handful of samples, offering a practical, low-cost first-pass signal that complements claim-level verification in real-world LLM workflows.

[*]Equal contribution [1]New York University [2]University of Toronto [3]Vijil. Correspondence to: Dhrupad Bhardwaj <db4045@nyu.edu>, Tim G. J. Rudner <tim.rudner@utoronto.ca>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Large language models (LLMs) increasingly serve as frontline knowledge workers (Mayer et al., 2025). Yet the application of LLMs to high-stakes settings that require long-form responses to open-ended prompts is hamstrung by the fact that reliably ascertaining the trustworthiness of long-form text generated by LLMs remains challenging.

Without scalable, computationally inexpensive methods that are able to reliably indicate the level of trustworthiness of long-form responses generated by a language model, deployment to high-stakes application domains that require substantively accurate responses to open-ended prompts will remain risky. Manual approaches are often infeasible at scale, and the standard stop-gaps—prompt engineering, system messages, and ensemble voting—remain brittle and expensive (Shorinwa et al., 2024). What is needed is a lightweight, model-agnostic, and data-agnostic method that is able to flag potentially untrustworthy text generations without needing to resort to slow, claim-by-claim fact-checking.

Prior work has attempted to gauge factuality in long-form natural language generation (NLG) by aligning atomic claims with a knowledge graph or by training auxiliary classifiers on annotated corpora (Wang et al., 2023). However, these approaches fall short in three important ways. First, they require structured references or costly ground-truth labels that may not exist in certain domains. Second, they struggle with open-ended, multi-sentence answers where relevant facts are implicit rather than explicit (Nikitin et al., 2024). Third, their computational footprint grows rapidly with the number of sentences or claims, making them ill-suited to real-time applications (Liu et al., 2024; Farquhar et al., 2024; Zhang et al., 2024; Manakul et al., 2023).

We address these limitations by introducing *semantic isotropy*—the degree of uniformity across normalized text embeddings on the unit sphere—and use it as a proxy for nonfactuality in long-form text generation. Intuitively, if a prompt admits a single, factually grounded explanation, independently sampled responses from a model should cluster tightly in embedding space (Qiu & Miikkulainen, 2024;

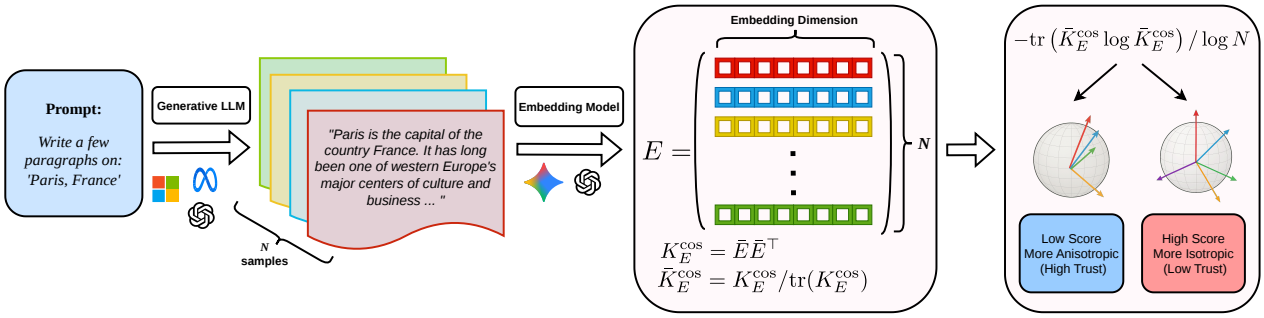


Figure 1. Diagrammatic Overview of the *Semantic Isotropy* Scoring Pipeline. This illustration presents the experimental process used to derive a *semantic isotropy* score. For a given input, we sample N i.i.d. responses using a **Generative LLM**. Each response is fed through an **Embedding Model** which produces a vector representation in R^D. The matrix of $N \times D$ response embedding vectors is denoted as E. E is transformed into a similarity matrix $K_E^{\cos}$ using the cosine kernel, which is used to compute the isotropy score as $\mathcal{I}(K_E^{\cos}) = \mathrm{vNE}(K_E^{\cos})/\log N$.

Wang & Holmes, 2024). Conversely, when an LLM hallucinates, subtle changes in semantics and content pull the embeddings apart, inflating angular dispersion.

To measure the semantic isotropy of a set of embeddings—and gauge the trustworthiness of the corresponding long-form text generations—we compute a *semantic isotropy score* according to the following simple and computationally inexpensive steps: We first generate a handful of responses, we then embed them with any off-the-shelf text encoder, and finally, we compute a semantic isotropy score by estimating the von Neumann entropy of the cosine kernel under the set of embeddings. No labels, fine-tuning, or hyperparameter search are required.

To enable broad evaluation of semantic isotropy as a proxy for nonfactuality—particularly across varying text lengths and at scale—we develop *Segment-Score*, a new factuality scoring method. *Segment-Score* is designed to address key limitations of existing approaches (Min et al., 2023): It is more efficient in terms of token usage, scales effectively to longer responses, and offers clearer, more consistent criteria for labeling statements as true or false. Using *Segment-Score*, we assess the reliability of *semantic isotropy scoring* in predicting nonfactuality in long-form text generation. Computing semantic isotropy scores is model-agnostic and can be done with off-the-shelf embedding models, making this approach inexpensive and easy to use in practice.

In our empirical evaluation, we find that *semantic isotropy scoring* achieves state-of-the-art performance across relevant benchmarks. It does so robustly across a wide range of models, response lengths, and evaluation settings, demonstrating its effectiveness and generalizability as a lightweight and scalable proxy for trustworthiness.

To summarize, our key contributions are:

1. We introduce *semantic isotropy* and describe a simple and computationally inexpensive method for *semantic isotropy scoring* as a means to assess nonfactuality in long-form natural language generation.
2. We develop the *Segment-Score* protocol to generate and score datasets of long-form LLM generations for open-ended prompts and create a dataset of $1,182$ unique entities along with $\approx 65,450$ scored responses by three different models, each containing between 25 and 60 distinct claims.
3. We demonstrate the efficacy of using *semantic isotropy scores* as a proxy for factual inconsistency in long-form natural language generation across a range of generative models, embeddings, and relevant experimental settings.

Code to reproduce our results is available at:
`github.com/dhrupadb/semantic_isotropy`

2. Related Work

2.1. Uncertainty Quantification in Language Models

Early work on uncertainty quantification in LLMs largely involved analyzing token-level logits and derived distributions. OpenAI (2023) demonstrated the differences in token-level calibration dynamics of pre-trained and RLHF post-trained LLMs with Kirk et al. (2024) inferring that the latter process led to model miscalibration.

On the other hand, Monte Carlo sampling-based approaches (Zhang, 2020; Malinin & Gales, 2021; Lin et al., 2024; Lakshminarayanan et al., 2017) have shown promise in estimating uncertainty in LLMs. However, a key challenge highlighted by these approaches is the need to standardize LLM responses to effectively compare them. In multiple-choice style discrete finite class output settings, studies have used a combination of Natural Language Inference (NLI) (Williams et al., 2018), oracle LLMs (Band et al., 2024), and embedding models (Qiu & Miikkulainen, 2024; Grewal et al., 2025) to standardize model responses for precise class

attribution. However, in NLI methods specifically, models have historically had limited context windows and several of the comparison algorithms grow quadratically in the number of inputs (Farquhar et al., 2024; Zhang et al., 2024; Manakul et al., 2023; Wang et al., 2024c; Chen et al., 2025). As a result, while the techniques used are in principle capable of addressing the semantic variations introduced by short-form responses, several limitations remain that hinder widespread adoption in practical settings.

A direction that addresses some of the pain points of NLI methods is the introspection of the model's internal state to estimate uncertainty. Kadavath et al. (2022) use the activations in the model's last hidden layer to train a linear classifier probe that indicates whether the model knows the answer to the prompt. Kossen et al. (2024) adopt a similar approach by developing a cheap predictive model on top of the LLM's internal state that is capable of estimating semantic entropy (Farquhar et al., 2024) without repeated sampling; Nikitin et al. (2024) extend this approach to analyze a graph representation and its Laplacian. Ji et al. (2024) use the state information to specifically predict the likelihood of hallucination while Ao et al. (2024) use latent representations in a multi-modal context to develop an uncertainty quantification method for both text and images simultaneously. While the success of these approaches justifies the value of leveraging latent representations for uncertainty quantification, these existing methods generally do not scale well with increases in response size and model complexity.

Recently, several studies have made meaningful advances in addressing uncertainty quantification, specifically for longer-format settings. Long Uncertainty Quantification (LUQ) (Zhang et al., 2024) and INSIDE (EigenScore) (Chen et al., 2024) are most relevant to our method and serve as key benchmarks in our analysis. While LUQ uses NLI methods to calculate a similarity matrix between samples, the LUQ-Atomic variant that we benchmark against averages normalized entailment at a sentence / statement level, providing a much more fine-grained and robust measure of similarity. EigenScore also samples a batch of responses and computes the eigen-distribution over the batch of latent representations in the generating model. Kernel Language Entropy (KLE; Nikitin et al., 2024) likewise uses von Neumann entropy of a kernel matrix, but its kernel is constructed from pairwise NLI entailment between sampled responses; in contrast, our cosine kernel operates directly on text embeddings, which removes the NLI bottleneck and is trivially vectorized at scale.

2.2. Factuality and Long-Form NLG

While several NLP datasets are designed for multiple choice-style tasks (Joshi et al., 2017; Rajpurkar et al., 2016; Lin et al., 2022; Reddy et al., 2019), the options for open-ended long-form responses are much more limited. A key challenge is defining a suitable metric that is a meaningful measure of quality but also tractable to compute. Factuality, or the degree to which a response is supported by a ground truth, has been a go-to metric for several studies (Guo et al., 2022). Min et al. (2023) developed FactScore, a popular scoring algorithm for generating factuality scores using a ground truth text. FactScore operates by decomposing a text into a set of atomic facts by using an oracle model (LLM) and scoring each of these facts with the help of the ground truth reference document. This method has been applied to create datasets such as FactScore-Bio (Zhang et al., 2024), which operates on biographies taken from Wikipedia.

FactScore has given rise to several derivative methods such as LongFact & SAFE (Wei et al., 2024) (which uses dynamic web searches to find supporting references), FELM (Chen et al., 2023) (applied to Math and Reasoning problems), Multi-FAct (Shafayat et al., 2024) (multi-lingual setting), FactAlign (Huang & Chen, 2024) and LoGU (Yang et al., 2024b) (post-training / RLHF to incorporate uncertainty phrasing into model generations), and VeriScore (Song et al., 2024) (better precision by scoring verifiable facts only) to name a few. But most of these methods share key limitations of the FactScore framework, scaling poorly with response length and requiring several LLM calls per sample. We address some of these drawbacks through our method, *Segment-Score* (see Algorithm 1), which utilizes capabilities of more modern LLMs such as longer context windows and the ability to respond with structured outputs.

3. Semantic Isotropy in Long-Form Generation

In this section, we introduce *semantic isotropy*, a measure of semantic variation in long-form text generations, and use it to motivate *semantic isotropy scoring*, a method that expresses the level of semantic isotropy in long-form responses and can serve as a proxy for nonfactuality.

3.1. Semantic Isotropy

We start with the conventional definition of isotropy. Intuitively, isotropy describes a state where a set of vectors or a distribution has no preferred direction—its properties are the same in all orientations. More formally, for a set of unit vectors, we can express this property as follows:

Definition (Isotropy). *The column vectors $x_1, ..., x_N \in \mathbb{R}^D$ with $\|x_i\|_2 = 1$ for all $i \in \{1, ..., N\}$ are said to be isotropic if for the matrix $X = [x_1, ..., x_N]$: $X^\top X = I_N$.*

This condition means that, on average, the transformation represented by X preserves directions and spreads uniformly across all directions in $\mathbb{R}^D$. When $X^\top X = I_N$, the linear transformation $x \to X^\top x$, for $x \in \mathbb{R}^D$, acts

Algorithm 1 SEGMENT-SCORE(t)

Require: Topic t, LLM response $Y(t)$, reference document $\mathcal{D}(t)$, oracle LLM $\mathcal{O}$
1: $S(t) \leftarrow \mathsf{Segmenter}_{\mathcal{O}}\big(Y(t)\big) = \langle s_1, s_2, \ldots, s_m \rangle$, where $m \leftarrow |S(t)|$ {Partition $Y(t)$ into m atomic segments}
2: Initialize $\mathbf{v} \leftarrow []$
3: **for** $i = 1$ **to** m **do**
4: $C_i \leftarrow \langle s_1, s_2, \ldots, s_{i-1} \rangle$ {Context before segment s_i}
5: $v_i \leftarrow \mathsf{Score}_{\mathcal{O}}\big(C_i, s_i, \mathcal{D}(t)\big)$ {Verify s_i against $\mathcal{D}(t)$ in context}
6: Append v_i to $\mathbf{v}$
7: **end for**
8: $\phi \leftarrow \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(v_i = \texttt{True})$ {Fraction of segments verified as True}
9: **return** ϕ

as an isometry on $\mathbb{R}^N$ embedded in $\mathbb{R}^D$ (with $N \ll D$ in our case). Geometrically, if the unit vectors $x_1, ..., x_N$ are isotropic, this implies that the unit vectors are uniformly dispersed across the unit sphere, giving us an intuitively meaningful measure of variation within $x_1, ..., x_N$.

We will now build on this intuition to define *semantic isotropy*. Let $\mathcal{R} = \{r_1, ..., r_N\} \in \mathcal{T}^N$ be a set of N long-form NLG responses given a prompt p and let $\Gamma_\theta(\cdot) : \mathcal{T} \to \mathbb{R}^D$ be an embedding model parameterized by neural network parameters θ. We define $E = \{e_1, ..., e_N\}$ to be the collection of embedding vectors, where $e_i = \Gamma_\theta(r_i) \in \mathbb{R}^D$ is the embedding vector of response r_i, $\bar{e}_i \in \mathbb{R}^D$ is the corresponding normalized embedding vector, $\bar{e}_i = e_i/||e_i||$, and $\bar{E}$ is the matrix of normalized embedding vectors, $\bar{E} = [\bar{e}_1, ..., \bar{e}_N]^\top \in \mathbb{R}^{N \times D}$. The cosine kernel function for two embedding vectors, $e_i, e_j \in \mathbb{R}^D$, is then defined as

$$k^{\cos}(e_i, e_j) = [\bar{E}\bar{E}^\top]_{i,j} = \bar{e}_i^\top \bar{e}_j. \tag{1}$$

We denote the cosine kernel matrix for E by $K_E^{\cos} \in \mathbb{R}^{N \times N}$.

With these preliminaries in place, we are now ready to formally define *semantic isotropy* to represent uniform dispersion of the normalized embedding vectors:

> **Definition (Semantic Isotropy).** *A set $\mathcal{R} = \{r_1, ..., r_N\}$ of long-form responses with corresponding embeddings $E = \{e_1, ..., e_N\}$ is semantically isotropic if $K_E^{\cos} = I_N$.*

3.2. Measuring Semantic Isotropy in Long-Form Language Generation

As defined, isotropy represents a strict condition that is either met or not met. However, to use semantic isotropy in practice as a means of ascertaining nonfactuality in long-form language generation, we need to gauge the *level* of isotropy of a given cosine kernel. We want to formalize the intuition that factuality corresponds to a high level of semantic alignment within a set of long-form generations (highly anisotropic), whereas uncertainty and nonfactuality correspond to a high level of dispersion (very isotropic).

To estimate the level of isotropy of a set of long-form text generations, we use the von Neumann entropy (vNE) (von Neumann & Beyer, 2018) of the cosine kernel $K_E^{\cos}$, normalized to have trace $= 1$, which allows us to view the eigenvalues as defining a discrete probability distribution. Let $\bar{K}_E^{\cos} = K_E^{\cos}/\text{trace}(K_E^{\cos})$ and define

$$\text{vNE}(K_E^{\cos}) = -\text{trace}(\bar{K}_E^{\cos} \log \bar{K}_E^{\cos}). \tag{2}$$

To be precise, we view the above definition in the eigenspace of the positive semidefinite matrix $K_E^{\cos}$, that is, $\text{vNE}(K_E^{\cos}) = -\sum_i^N \lambda_i \log \lambda_i$, where $\lambda_1, ..., \lambda_N$ are the eigenvalues of $\bar{K}_E^{\cos}$. Note that $\text{vNE}(K_E^{\cos})$ attains its maximum value of $\log N$ for semantically (perfectly) isotropic responses (i.e., isotropic normalized embedding vectors) and is minimized at 0 when all embedding vectors are aligned (parallel). It thus measures the degree of dispersion of $\{\bar{e}_i\}_{i=1}^N$ on the unit sphere and serves as a measure of the level of isotropy within the set of text embeddings.

We are now ready to define the *semantic isotropy score* $\mathcal{I}(\cdot)$ that will serve as our proxy for factuality. Consider N sampled long-form text generations in response to a prompt and define

$$\mathcal{I}(K_E^{\cos}) = \text{vNE}(K_E^{\cos})/\log N, \tag{3}$$

with $0 \leq \mathcal{I}(K_E^{\cos}) \leq 1$. Under this definition, (i) a lower degree of semantic isotropy (a semantic isotropy score closer to 0) is associated with stronger alignment/lower dispersion and consequently higher certainty and factuality, whereas (ii) a higher degree of semantic isotropy (a semantic isotropy score closer to 1) is associated with weaker alignment/higher dispersion and consequently lower certainty and factuality. In a nutshell: *The lower* the semantic isotropy score, *the more trustworthy* the generating model's long-form responses to the prompt are.

4. Empirical Evaluation

In this section, we conduct an empirical evaluation to assess whether semantic isotropy scores can serve as an effective

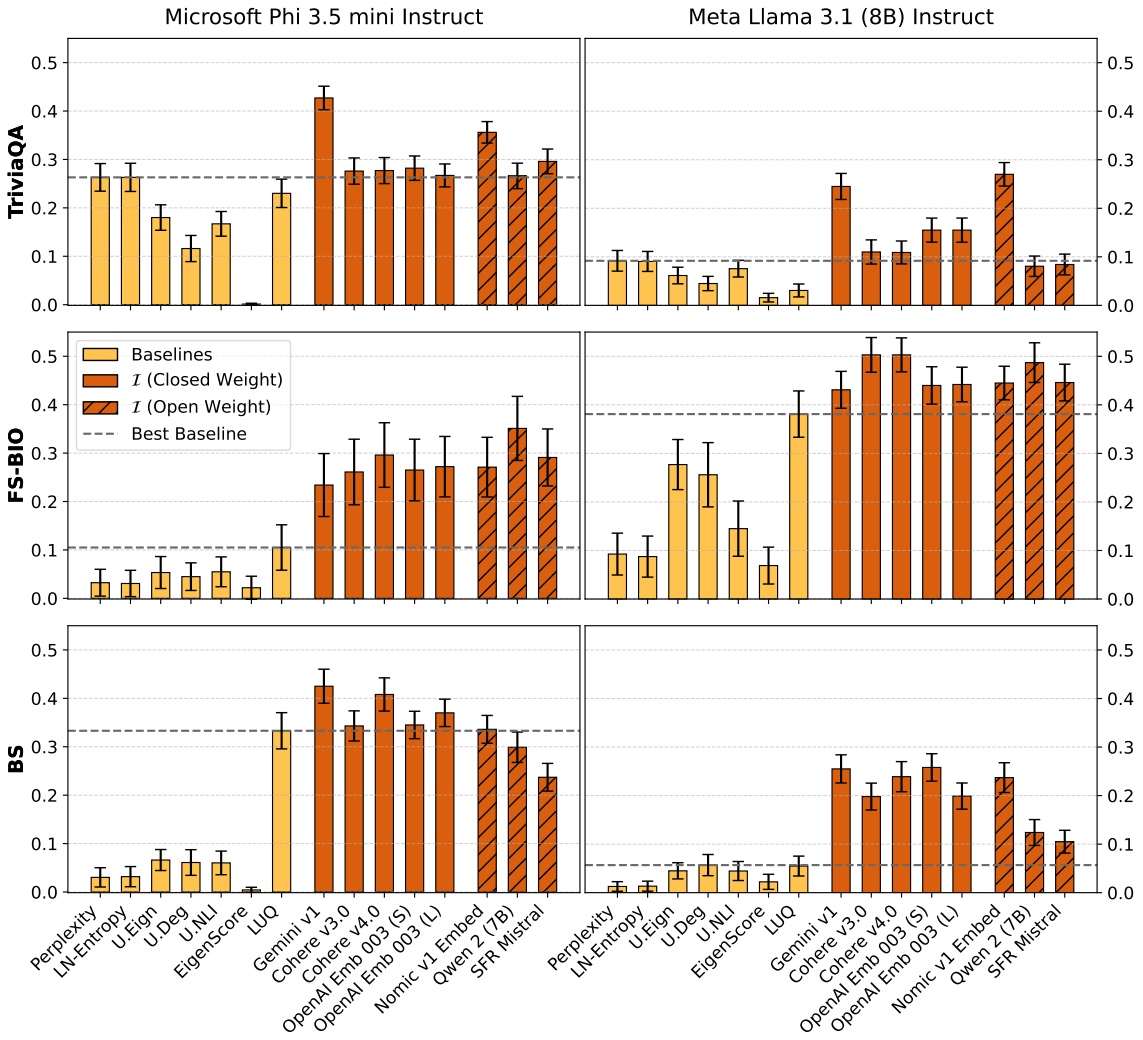


Figure 2. Main Experimental Results. Bar charts comparing the performance (as measured by the R^2 of a linear model of Factuality $\sim$ Semantic Isotropy); implemented using various embedding models and benchmark uncertainty metrics on three datasets: TriviaQA Entities **[TriviaQA]** (Top Row), FactScore-Biographies **[FS-BIO]** (Middle Row) and CMU Book Summaries **[BS]** (Bottom Row); scored using the *Segment-Score* **(SS)** algorithm. Responses are $\approx$500 words. 1,500 bootstrapped samples are used to generate 1-SD error bars. Semantic isotropy $\mathcal{I}$ (ours) outperforms all other baselines for all embedding models.

proxy for nonfactuality in long-form natural language generation with LLMs. First, we will describe the creation of a ground truth evaluation dataset using *Segment-Score* and present our evaluation metrics and benchmarks. We will then present our main findings about the efficacy of semantic isotropy scoring as a means to predict nonfactuality in long-form text generation.

4.1. Creation of a Dataset of Long-Form Responses to Open-Ended Prompts

To construct a dataset of long-form responses to open-ended prompts, we must select a corpus of entities such that each is associated with an underlying ground truth reference document. These entities are used to create a set of open-ended prompts that can be fed into an LLM to generate long-form responses. We use three sources—*FactScore-Bio* **[FS-BIO]** (182 unique entities) (Min et al., 2023; Zhang et al., 2024), **TriviaQA** (Joshi et al., 2017) (1,000 unique entities) and CMU Book Summaries **[BS]** (509 unique entities) (Bamman & Smith, 2013). All three datasets use Wikipedia as their ground truth to source reference documents. For **TriviaQA**, we exclude all entities that correspond to days, dates or numerical values and only select those that match the title of the underlying Wikipedia page. Of the 5,245 qualifying TriviaQA entities across the training and validation set, we randomly select a subset of 1,000 entities to create open-ended prompts for long-form text generation. For *Book*

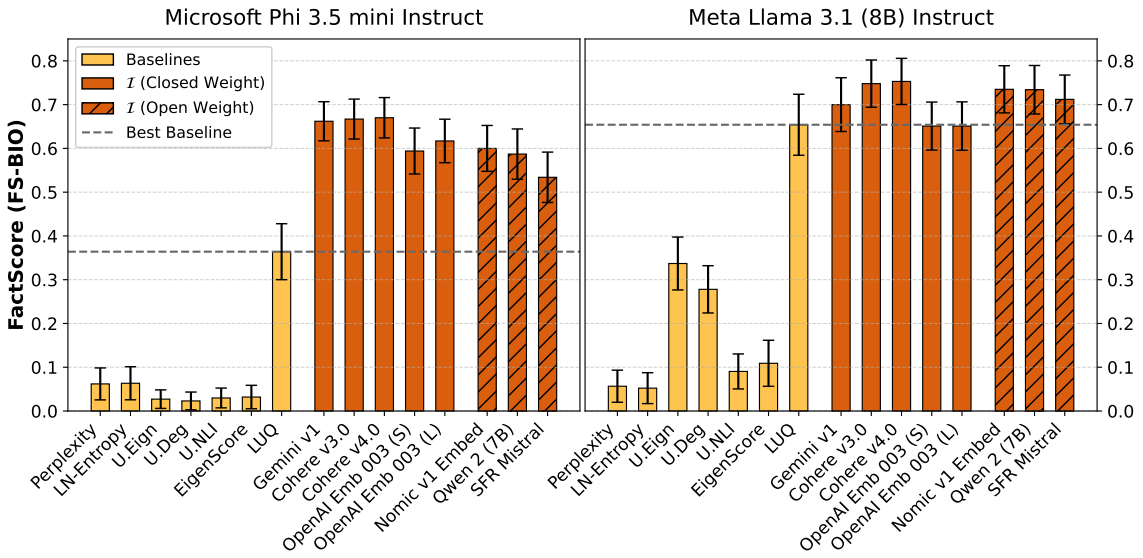


Figure 3. FactScore Experimental Results. Bar charts comparing the performance (as measured by the R^2 of a linear model of factuality with semantic isotropy as the explanatory variable); implemented using various embedding models and benchmark uncertainty metrics on the **FS-BIO** dataset using the *FactScore* **(FS)** scoring algorithm. Left hand side: Benchmark UQ metrics. Right Hand Side: semantic isotropy ($\mathcal{I}$) implemented using various embedding models. Same experimental setting as in Figure 2. We observe that our method's performance is robust to the scoring scheme.

Summaries (**BS**), we constrain the dataset to entities where the underlying reference text is at least 5,000 characters and the publish date is known, and is on or after Jan 1, 2000. Of a total of 16,559 entries, we select 509 entities that meet this criterion.

To generate long-form responses, we use Meta Llama 3.1 8B Instruct (Grattafiori et al., 2024), Microsoft Phi 3.5 Mini Instruct (Abdin et al., 2024), and OpenAI GPT 4.1 Mini (OpenAI, 2025) to generate responses for **TriviaQA** and **FS-BIO**. For **BS**, we only generate responses using Llama 3.1 and Phi 3.5 Mini. For **TriviaQA** we sample up to $k = 20$ responses while for **FS-BIO** and **BS** we sample $k = 10$ responses, targeting approximately 500 words for each response. For **TriviaQA** using Llama 3.1 8B Instruct, we also generate a longer dataset of approximately $1,000$ words and derived datasets of intermediate word counts of approximately 125, 250, 375, 500, and 750 words by truncating the $1,000$ word variant to the nearest sentence. Model inference is run with temperature $\tau = 0.7$ and FP16 quantization using the vLLM framework (Kwon et al., 2023). Finally, we compute factuality scores for the datasets using the *Segment-Score* **(SS)** procedure (Algorithm 1). Implementation details and prompts are presented in Appendix B.7. The dataset and Segment-Score algorithm are released to allow future benchmarking by the research community.

We also generate *FactScore* (**FS**; Min et al., 2023) scores for the **FS-BIO** datasets from each model. This allows us to benchmark our scoring algorithm **(SS)** against the existing state-of-the-art method, and we can show that our results hold robustly. See Appendix C.2 for a comparison of the scoring methods. We use GPT 4.1 Mini (OpenAI, 2025) as the oracle LLM in both algorithms.

4.2. Evaluation Metrics

To assess the performance of our methods and compare them to appropriate baselines from the literature, we measure the degree of the relationship between each individual measure and the average factuality of a given topic, aggregated across all responses. First, we selected a set of embedding models that rank highly on the MTEB Benchmark (Muennighoff et al., 2023) to use in computing our isotropy scores. We measure predictiveness of these scores using R^2, the explained variance of a linear model with factuality scores as the dependent variable and isotropy scores as the independent variable. To estimate error bounds, we perform bootstrap sampling and generate 1-SD error bars for each measure. We compare *Semantic Isotropy* to the following baseline measures: **Perplexity** - Exponential of the average negative log-likelihood of the generation (Ren et al., 2023); **LN-Entropy** - Length Normalized Entropy (Malinin & Gales, 2021); **U.Eign** - Sum of Eigenvalues of the Graph Laplacian, **U.Deg** - Average pairwise distance measured using an NLI model (Lin et al., 2024); **U.NLI** - NLI uncertainty derived using SelfCheckNLI (Manakul et al., 2023); **Semantic Entropy** - Entropy of the probability distribution of semantic clusters over the responses (Farquhar et al., 2024); **LUQ-Atomic** - Long Uncertainty Quantification using atomic facts via LLM based fact extraction (Zhang et al.,

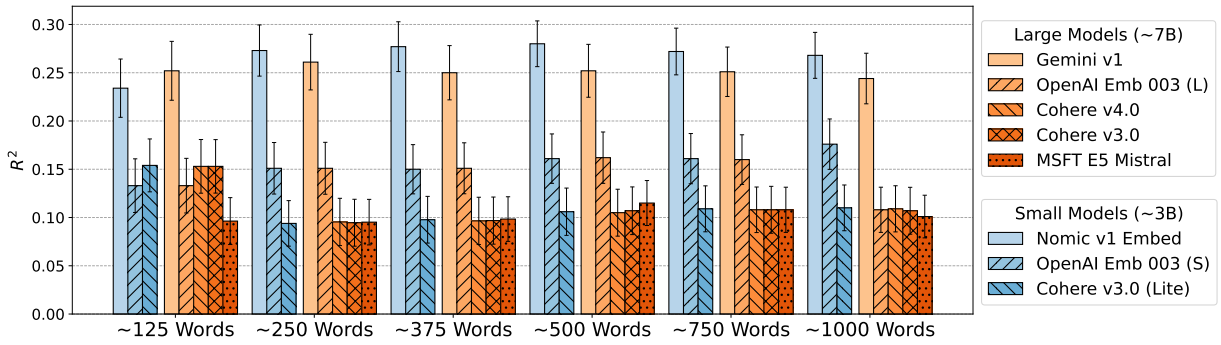


Figure 4. Performance by Response Length: Comparing *Semantic Isotropy* in terms of explained variance (R^2) of Factuality **(SS)** across various response lengths, contrasted by model size: Small (up to $3B$ parameters) and Large ($\approx 7B$ or more parameters). 1,500 bootstrapped samples are used to generate 1-SD error bars.

2024); **EigenScore** - the log-determinant of the centered embedding matrix of the generating model's internal state activations over a set of sampled responses (Chen et al., 2024). Each baseline method is implemented using the original backbone (i.e., DeBERTa-v3 Large (MNLI) for the NLI components of LUQ, SelfCheckNLI, and semantic entropy; internal-state activations for EigenScore); full numerical results, including semantic entropy, are reported in Tables 2 and 3 (Appendix A).

While relevant to our study, we do not include comparisons with **Kernel Language Entropy (KLE)** (Nikitin et al., 2024) and **Graph Longform Uncertainty** (Jiang et al., 2024). We found that the KLE computation was highly numerically unstable given the degree of semantic entailment overlap among the sampled long-form responses and did not produce meaningful results. For Graph Uncertainty, the distinct claim union algorithm proved computationally difficult to implement at the response lengths considered in our study. For context, a typical set of responses had over 600 distinct claims, which implied 6,000 LLM calls to measure claim to response entailment for one topic, assuming $N = 10$.

Sampling the appropriate embedding. Unlike closed-weight models that return just an embedding vector, for open-weight models we select the activations corresponding to the last token in the final hidden state, except for Nomic v1 (Nussbaum et al., 2025) where we take the average of the activations over the token dimension instead. For further details, see Section 5.

4.3. Main Result: Factuality in Long-Form Responses to Open-Ended Prompts

We present our main empirical results in Figure 2 (see also Table 3 in Appendix A). We find that semantic isotropy scores are superior to existing methods in predicting nonfactuality as per the **(SS)** algorithm on the **FS-BIO**, **TriviaQA** and **BS** datasets. Using R^2 (i.e., the explained variance) as our performance measure, $\mathcal{I}$ outperforms all existing metrics, in most cases by a wide margin. We see some variance in the performance under different embedding models, and discuss this finding in more detail in Section 5. While certain embedding models are better suited to specific tasks, semantic isotropy scoring outperforms existing state-of-the-art methods across the board. Notably, the Nomic v1 embedding (Nussbaum et al., 2025) demonstrates exceptional performance on all datasets, occasionally surpassing larger and closed-source models such as Gemini (Gemini Team, 2024) and OpenAI Embeddings (OpenAI, 2024).

4.4. FactScore Evaluation

To ensure that our results are not merely an artifact of the *Segment-Score* method, we also evaluate semantic isotropy scoring under the *FactScore* (**FS**) algorithm applied to the **FS-BIO** dataset (see Figure 3). We find that the results mirror those in Figure 2, even improving the predictive performance of semantic isotropy scoring. This consistency across scoring schemes highlights the robustness and generalizability of semantic isotropy as a measure of nonfactuality. A more detailed comparison of the two scoring methods can be found in Appendix C.2.

4.5. Discriminative Power at the Operating Point

To assess discriminative power at a concrete operating point, we discretize the factuality target: topics with average *Segment-Score* factuality above 50% are labeled 1 and the remainder 0, with $1 - \mathcal{I}$ serving as the confidence score. The resulting receiver operating characteristic curves are shown in Figure 6. On the entity-grounded datasets, the score discriminates strongly, with the AUC reaching a value of approximately 0.95 on Book Summaries (Phi 3.5 Mini) and a value of 0.90 on FS-BIO (Llama 3.1). High-isotropy prompts can therefore be routed to expensive claim-level verification while low-isotropy ones pass with minimal additional checking; threshold calibration is inherently deployment-specific, following standard practice for sampling-based uncertainty scores.

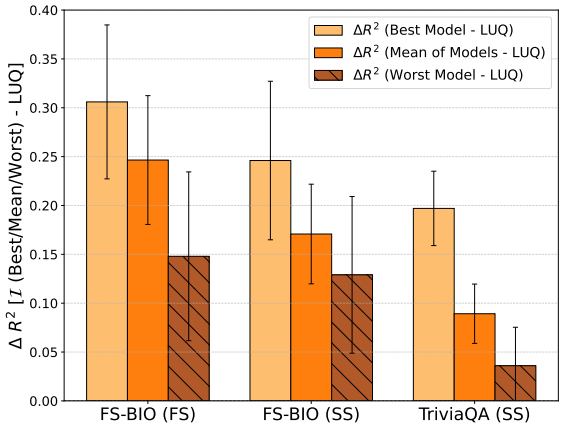


Figure 5. Model Agnostic Performance Comparison: Performance difference between current state-of-the-art **LUQ-Atomic** (Zhang et al., 2024) and semantic isotropy based on different embedding models across each dataset using Phi 3.5 Mini Instruct as the generator model. In each case, all versions of semantic isotropy are more predictive than LUQ. 1,500 bootstrapped samples are used to generate 1-SD error bars.

4.6. Computational Cost

LUQ and Graph Uncertainty scale as $\mathcal{O}(MN^2)$, where M is the average number of facts or segments per response and N is the number of sampled responses, since each pair of segments requires an NLI forward pass. Semantic entropy and related methods are $\mathcal{O}(N^2)$ in entailment computations and cannot be vectorized naively. Semantic isotropy is also $\mathcal{O}(N^2)$ in the kernel construction, but the cosine kernel is fully parallelizable and the cost of the vNE computation on an $N \times N$ matrix with $N \leq 20$ is negligible. In practice, computing isotropy scores for one batch of $N = 20$ responses, each comprised of approximately 500 words, requires 1.8 ± 0.05 seconds on a V100 (amortized, Nomic v1) compared to 302 ± 48 seconds for LUQ-Atomic using DeBERTa-v3 Large (MNLI) (Manakul et al., 2023) of comparable parameter count—approximately $170\times$ faster.

5. Sensitivity Studies

Effect of Response Length. One important study we undertake is to understand the relationship between measure performance and response length. While our main study focuses on ≈500 word responses, we explore the relationship between semantic isotropy and factuality for shorter and longer texts as well. We outline these results in Figure 4. While certain models perform better (Gemini and Nomic v1), performance is generally consistent across the board. OpenAI's Text Embedding 3 is a notable exception, with performance consistently increasing with response length, independent of model size.

Table 1. Segment-Score Coherence Test: Comparison of label agreement of *Segment-Score* classifications and those obtained using a majority voting scheme. For each dataset, we randomly sample 30 entities and score their response corpus using GPT 4.1, Claude 4 Sonnet and DeepSeek, taking 5 samples each. Scores are in percentage (%).

Dataset	Gen. Model	OpenAI (GPT 4.1)	Claude (4 Sonnet)	DeepSeek
BS	Llama 3.1	**89.24**	81.33	75.87
	Phi 3.5 Mini	**91.28**	86.13	80.38
FS-BIO	Llama 3.1	83.92	86.13	**86.39**
	Phi 3.5 Mini	76.98	86.67	**88.30**

Effect of Choice of Embedding Model. We analyze how the size and type of embedding model impact performance. As shown in Figure 5, *Semantic Isotropy* consistently outperforms LUQ across all three datasets/scoring methods and embedding models. Notably, even the least effective embedding model yields improvements over LUQ, while the best performs substantially better. These results highlight the robustness of *Semantic Isotropy*, demonstrating its effectiveness regardless of the embedding model used. We compare small and large variants of closed source models such as those provided by OpenAI and Cohere. In both cases, the performance difference is negligible and likely attributable to sampling noise. This finding demonstrates that semantic isotropy is robust even at small model sizes. Because the factual-vs.-nonfactual signal we exploit is a coarse, distribution-level effect, encoders with lower-dimensional representations (Kusupati et al., 2024) retain it well, often matching larger models whose additional capacity primarily captures finer-grained stylistic detail.

Effect of Choice of Isotropy Measure. In Table 4 (Appendix A) we ablate a variety of alternative measures of isotropy, among them the Frobenius norm $\|K_E^{\cos}\|$, $\log(\det(K_E^{\cos}))$ and $\text{trace}(K_E^{\cos})^{-1}$. We find that von Neumann entropy-based isotropy scoring performs best overall, outperforming other measures on average across different embedding model choices and dataset settings. However, most of these measures demonstrate nearly equivalent performance, further underscoring the robustness of semantic isotropy as a proxy for nonfactuality.

Coherence of Segment-Score. While *Segment-Score* offers several practical benefits, it remains critical to establish that it is ultimately a sound proxy for factuality. To calibrate FactScore, Min et al. (2023) used a cohort of human evaluators. Given the size and span of our datasets, we opted for an LLM-assisted workflow using several SOTA oracle models such as GPT 4.1 (OpenAI, 2025) or Claude 4 (Anthropic, 2025). For each response, we sample up to 5 True/False labels for each segment within, given the reference document. We use the majority of these labels as

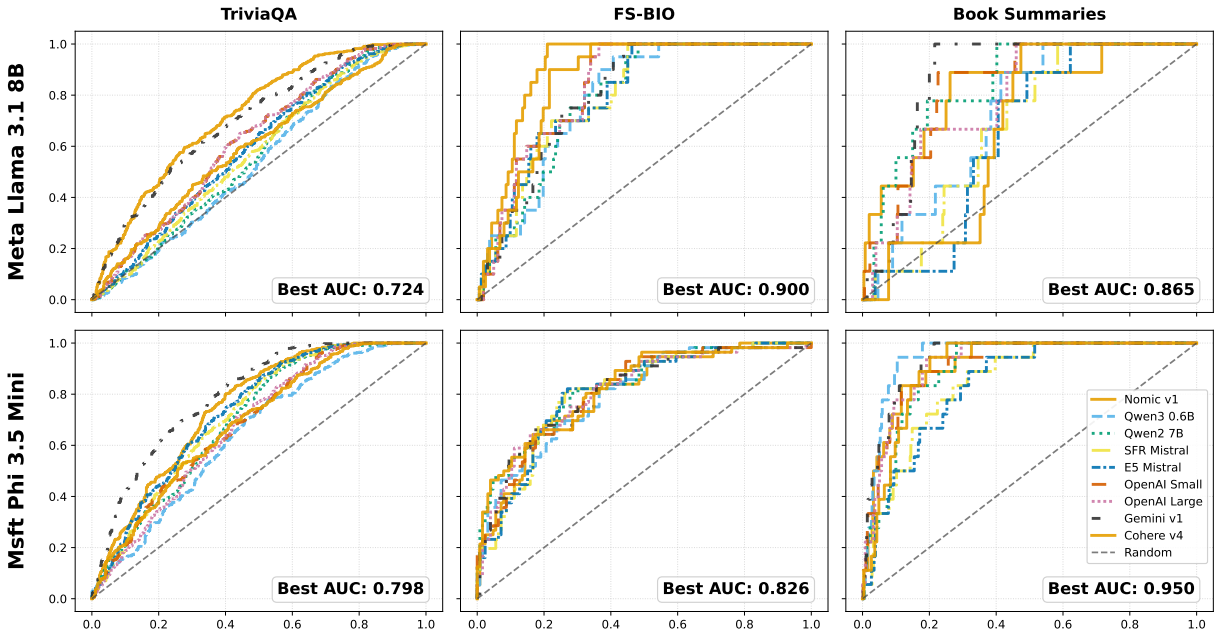


Figure 6. Isotropy vs. Factuality (ROC). ROC curves comparing *Semantic Isotropy* against $\{0, 1\}$ class labels obtained by thresholding *Segment-Score* at 50%. Curves are computed on Llama 3.1 and Phi 3.5 Mini responses across all three datasets.

our proxy for a true class label. In Table 1, we provide the matching accuracy against the one-shot labels generated by *Segment-Score*. The high degree of agreement indicates that the *Segment-Score* algorithm can be effectively utilized to proxy factuality. In Appendix C.2, we elaborate on the study relating the scores obtained by *Segment-Score* and *FactScore* algorithms on the **FS-BIO** dataset.

6. Limitations

Computing semantic isotropy scores requires repeated sampling, which incurs extra inference cost relative to single-pass probe-based uncertainty estimates, but this overhead is decreasing with modern high-throughput inference engines and reasoning-oriented LLMs. Furthermore, semantic isotropy scoring operates at the prompt level and does not localize errors to particular claims, so it is best paired with claim-level verification when fine-grained attributions are required. Semantic isotropy further relies on extrinsic divergence under uncertainty, that is, the empirical regularity that hallucinated responses tend to disagree with one another (Manakul et al., 2023; Ricco et al., 2025). As such, the nonfactuality signal obtainable through semantic isotropy scoring weakens for generators whose outputs remain stylistically uniform even when hallucinating, as we observe with GPT 4.1 Mini on TriviaQA (Figure 9, with a style analysis in Figures 10, 11a and 11b), and is uninformative for intrinsic hallucinations that remain self-consistent across samples. Across the broad range of settings considered in our empirical evaluation, however, semantic isotropy consistently identifies nonfactual long-form generations with high accuracy and provides a practical first-pass signal that can be paired with claim-level verification if needed.

7. Conclusion

In this work, we introduced *Semantic Isotropy*, proposed a computationally inexpensive method for semantic isotropy scoring, presented *Segment-Score* for cheap long-form ground truth data generation, and demonstrated that semantic isotropy scores yield state-of-the-art factuality prediction in long-form natural language generation. We found that semantic isotropy is highly predictive of nonfactuality in long-form LLM responses to open-ended prompts and that semantic-isotropy scores outperform alternative approaches—in most cases by a large margin. We also found that, in most of the cases considered in our analysis, semantic isotropy scoring is robust to the choice of embedding model, the choice of isotropy measure, the response length, and the number of samples used. We released code to reproduce the empirical evaluation and implement the *Segment-Score* method, and we will make the *Segment-Score*-annotated dataset publicly available to facilitate further research into assessing factuality in long-form text generation. Our findings suggest that semantic isotropy is a simple and yet effective predictor of nonfactuality and open the door to more reliable and cost-effective evaluation of long-form LLM text generation at scale.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning by providing methods to assess the trustworthiness of long-form text generated by large language models. The ability to reliably detect nonfactual content in LLM outputs has positive societal implications, as it can help prevent the spread of misinformation and enable safer deployment of LLMs in high-stakes domains such as healthcare, law, and education. We note that our method is designed as a complementary early-warning signal rather than a replacement for comprehensive fact-checking, and is not intended for use in adversarial scenarios where models may have been deliberately trained on faulty content.

Acknowledgments

JK thanks the Simons Foundation for support through the Collaborative Grant "The Physics of Learning and Neural Computation". DB and JK acknowledge support by the NSF through NRT Award 1922658. TGJR acknowledges support from the Foundational Research Grants program at Georgetown University's Center for Security and Emerging Technology. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A. D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., Jacobs, S. A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y. J., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C. C. T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

AI, C. Introducing embed v3. https://cohere.com/blog/introducing-embed-v3, 2023.

AI, C. Introducing embed 4: Multimodal search for business. https://cohere.com/blog/embed-4, 2025.

Anthropic. Introducing claude 4, 2025. URL https://www.anthropic.com/news/claude-4.

Ao, S., Rueger, S., and Siddharthan, A. Css: Contrastive semantic similarity for uncertainty quantification of llms, 2024. URL https://arxiv.org/abs/2406.03158.

Bamman, D. and Smith, N. A. New alignment methods for discriminative book summarization, 2013. URL https://arxiv.org/abs/1305.1319.

Band, N., Li, X., Ma, T., and Hashimoto, T. Linguistic calibration of long-form generations, 2024. URL https://arxiv.org/abs/2404.00474.

Chen, C., Liu, K., Chen, Z., Gu, Y., Wu, Y., Tao, M., Fu, Z., and Ye, J. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

Chen, S., Zhao, Y., Zhang, J., Chern, I.-C., Gao, S., Liu, P., and He, J. Felm: Benchmarking factuality evaluation of large language models, 2023. URL https://arxiv.org/abs/2310.00741.

Chen, T., Liu, X., Da, L., Chen, J., Papalexakis, V., and Wei, H. Uncertainty quantification of large language models through multi-dimensional responses, 2025. URL https://arxiv.org/abs/2502.16820.

DeepSeek-AI, Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Yang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Chen, J., Yuan, J., Qiu, J., Song, J., Dong, K., Gao, K., Guan, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Pan, R., Xu, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Zheng, S., Wang, T., Pei, T., Yuan, T., Sun, T., Xiao, W. L., Zeng, W., An, W., Liu, W., Liang, W., Gao, W., Zhang, W.,

Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Chen, X., Nie, X., Sun, X., Wang, X., Liu, X., Xie, X., Yu, X., Song, X., Zhou, X., Yang, X., Lu, X., Su, X., Wu, Y., Li, Y. K., Wei, Y. X., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Zheng, Y., Zhang, Y., Xiong, Y., Zhao, Y., He, Y., Tang, Y., Piao, Y., Dong, Y., Tan, Y., Liu, Y., Wang, Y., Guo, Y., Zhu, Y., Wang, Y., Zou, Y., Zha, Y., Ma, Y., Yan, Y., You, Y., Liu, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Huang, Z., Zhang, Z., Xie, Z., Hao, Z., Shao, Z., Wen, Z., Xu, Z., Zhang, Z., Li, Z., Wang, Z., Gu, Z., Li, Z., and Xie, Z. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Gemini Team. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg,

L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Grewal, Y. S., Bonilla, E. V., and Bui, T. D. Improving uncertainty quantification in large language models via semantic embeddings, 2025. URL https://openreview.net/forum?id=N4mb3MBV6J.

Guo, Z., Schlichtkrull, M., and Vlachos, A. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 02 2022. ISSN 2307-387X. doi: 10.1162/tacl_a_00454. URL https://doi.org/10.1162/tacl_a_00454.

Huang, C.-W. and Chen, Y.-N. Factalign: Long-form factuality alignment of large language models, 2024. URL https://arxiv.org/abs/2410.01691.

Ji, Z., Chen, D., Ishii, E., Cahyawijaya, S., Bang, Y., Wilie, B., and Fung, P. Llm internal states reveal hallucination risk faced with a query, 2024. URL https://arxiv.org/abs/2407.03282.

Jiang, M., Ruan, Y., Sattigeri, P., Roukos, S., and Hashimoto, T. Graph-based uncertainty metrics for long-form language model outputs, 2024. URL https://arxiv.org/abs/2410.20783.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

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., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL https://arxiv.org/abs/2310.06452.

Kossen, J., Han, J., Razzak, M., Schut, L., Malik, S., and Gal, Y. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024. URL https://arxiv.org/abs/2406.15927.

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., and Farhadi, A. Matryoshka representation learning, 2024. URL https://arxiv.org/abs/2205.13147.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

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

Liu, L., Pan, Y., Li, X., and Chen, G. Uncertainty estimation and quantification for llms: A simple supervised approach, 2024. URL https://arxiv.org/abs/2404.15993.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction, 2021. URL https://arxiv.org/abs/2002.07650.

Manakul, P., Liusie, A., and Gales, M. J. F. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023. URL https://arxiv.org/abs/2303.08896.

Mayer, H., Yee, L., Chui, M., and Roberts, R. Superagency in the workplace: Empowering people to unlock AI's full potential. Technical report, McKinsey, 2025.

Meng, R., Liu, Y., Joty, S. R., Xiong, C., Zhou, Y., and Yavuz, S. Sfr-embedding-mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog, 2024. URL https://www.salesforce.com/blog/sfr-embedding/.

Min, S., Krishna, K., Lyu, X., Lewis, M., tau Yih, W., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023. URL https://arxiv.org/abs/2305.14251.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. Mteb: Massive text embedding benchmark, 2023. URL https://arxiv.org/abs/2210.07316.

Nikitin, A., Kossen, J., Gal, Y., and Marttinen, P. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8901–8929. Curran Associates, Inc., 2024.

Nussbaum, Z., Morris, J. X., Duderstadt, B., and Mulyar, A. Nomic embed: Training a reproducible long context text embedder, 2025. URL https://arxiv.org/abs/2402.01613.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. New embedding models and api updates, 2024. URL https://openai.com/index/new-embedding-models-and-api-updates/.

OpenAI. Introducing gpt-4.1 in the api, 2025. URL https://openai.com/index/gpt-4-1/.

Qiu, X. and Miikkulainen, R. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space, 2024. URL https://arxiv.org/abs/2405.13845.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

Reddy, S., Chen, D., and Manning, C. D. Coqa: A conversational question answering challenge, 2019. URL https://arxiv.org/abs/1808.07042.

Ren, J., Luo, J., Zhao, Y., Krishna, K., Saleh, M., Lakshminarayanan, B., and Liu, P. J. Out-of-distribution detection and selective generation for conditional language models, 2023. URL https://arxiv.org/abs/2209.15558.

Ricco, E., Cima, L., and Pietro, R. D. Hallucination detection: A probabilistic framework using embeddings distance analysis, 2025. URL https://arxiv.org/abs/2502.08663.

Shafayat, S., Kim, E., Oh, J., and Oh, A. Multi-fact: Assessing factuality of multilingual llms using factscore, 2024. URL https://arxiv.org/abs/2402.18045.

Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., and Majumdar, A. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions, 2024. URL https://arxiv.org/abs/2412.05563.

Song, Y., Kim, Y., and Iyyer, M. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation, 2024. URL https://arxiv.org/abs/2406.19276.

von Neumann, J. and Beyer, R. T. *Mathematical Foundations of Quantum Mechanics: New Edition*. Princeton University Press, ned - new edition edition, 2018. ISBN 9780691178561. URL http://www.jstor.org/stable/j.ctt1wq8zhp.

Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Gao, W., Hu, X., Qi, Z., Wang, Y., Yang, L., Wang, J., Xie, X., Zhang, Z., and Zhang, Y. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity, 2023. URL https://arxiv.org/abs/2310.07521.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. Improving text embeddings with large language

models, 2024a. URL https://arxiv.org/abs/2401.00368.

Wang, X., Song, L., Tian, Y., Yu, D., Peng, B., Mi, H., Huang, F., and Yu, D. Towards self-improvement of llms via mcts: Leveraging stepwise knowledge with curriculum preference learning, 2024b. URL https://arxiv.org/abs/2410.06508.

Wang, Y.-H., Bai, A., Tsai, C.-P., and Hsieh, C.-J. Clue: Concept-level uncertainty estimation for large language models, 2024c. URL https://arxiv.org/abs/2409.03021.

Wang, Z. and Holmes, C. On subjective uncertainty quantification and calibration in natural language generation, 2024. URL https://arxiv.org/abs/2406.05213.

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., Du, C., and Le, Q. V. Long-form factuality in large language models, 2024. URL https://arxiv.org/abs/2403.18802.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1101.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., Yang, J., Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, X., Zhang, X., Wei, X., Ren, X., Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., and Fan, Z. Qwen2 technical report, 2024a. URL https://arxiv.org/abs/2407.10671.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yang, R., Zhang, C., Zhang, Z., Huang, X., Yang, S., Collier, N., Yu, D., and Yang, D. Logu: Long-form generation with uncertainty expressions, 2024b. URL https://arxiv.org/abs/2410.14309.

Zhang, C., Liu, F., Basaldella, M., and Collier, N. Luq: Long-text uncertainty quantification for llms, 2024. URL https://arxiv.org/abs/2403.20279.

Zhang, J. Modern monte carlo methods for efficient uncertainty quantification and propagation: A survey, 2020. URL https://arxiv.org/abs/2011.00680.

Zhao, Y., Yin, H., Zeng, B., Wang, H., Shi, T., Lyu, C., Wang, L., Luo, W., and Zhang, K. Marco-o1: Towards open reasoning models for open-ended solutions, 2024. URL https://arxiv.org/abs/2411.14405.

Appendix

A. Full Experimental Results

Table 2. Main Experimental Results: FactScore. Comparison of semantic isotropy with other uncertainty metrics on **FactScore-Bio** using FactScore. Each dataset column is subdivided by base model: Llama 3.1 8B, Phi 3.5 Mini, and GPT 4.1 Mini. Values are R^2 (explained variance) of a simple linear model of Factuality $\sim$ Isotropy score. 1,500 bootstrapped samples are used to generate 1-SD error bars. †denotes API-only models.

	FactScore-Bio (FactScore)		
Measure / Metric	**Llama 3.1 8B**	**Phi 3.5 Mini**	**GPT 4.1 Mini**
Perplexity (Ren et al., 2023)	0.0578 ± 0.04	0.0622 ± 0.04	0.0158 ± 0.02
LN Entropy (Malinin & Gales, 2021)	0.0535 ± 0.04	0.0608 ± 0.04	0.00625 ± 0.009
U.Eign (Lin et al., 2024)	0.344 ± 0.06	0.0279 ± 0.02	0.0185 ± 0.02
U.Deg (Lin et al., 2024)	0.28 ± 0.05	0.0224 ± 0.02	0.0255 ± 0.02
U.NLI (Manakul et al., 2023)	0.0916 ± 0.04	0.0299 ± 0.02	0.0263 ± 0.02
EigenScore (Chen et al., 2024)	0.107 ± 0.05	0.0309 ± 0.03	N/A
LUQ (Zhang et al., 2024)	0.657 ± 0.07	0.367 ± 0.06	0.192 ± 0.06
Semantic Entropy (Farquhar et al., 2024)	0.136 ± 0.06	0.00868 ± 0.01	0.0152 ± 0.02
$\mathcal{I}$: Gemini† (Gemini Team, 2024)	0.7 ± 0.06	0.661 ± 0.05	0.48 ± 0.08
$\mathcal{I}$: OpenAI Small† (OpenAI, 2024)	0.65 ± 0.06	0.593 ± 0.05	0.38 ± 0.08
$\mathcal{I}$: OpenAI Large† (OpenAI, 2024)	0.649 ± 0.06	0.614 ± 0.05	0.376 ± 0.08
$\mathcal{I}$: Nomic v1 (Nussbaum et al., 2025)	0.737 ± 0.05	0.6 ± 0.05	0.547 ± 0.08
$\mathcal{I}$: Qwen 2 (7B) (Yang et al., 2024a)	0.733 ± 0.05	0.589 ± 0.05	0.569 ± 0.08
$\mathcal{I}$: SF Mistral (Meng et al., 2024)	0.714 ± 0.05	0.533 ± 0.06	$\mathbf{0.607 \pm 0.08}$
$\mathcal{I}$: Mistral (E5) (Wang et al., 2024a)	0.714 ± 0.06	0.508 ± 0.06	0.604 ± 0.08
$\mathcal{I}$: Cohere v4.0 † (AI, 2025)	0.747 ± 0.06	$\mathbf{0.671 \pm 0.05}$	0.5 ± 0.08
$\mathcal{I}$: Cohere v3.0 † (AI, 2023)	$\mathbf{0.75 \pm 0.06}$	0.665 ± 0.05	0.503 ± 0.08
$\mathcal{I}$: Cohere v3.0 (Lite) † (AI, 2023)	0.749 ± 0.05	0.661 ± 0.05	0.503 ± 0.08
$\mathcal{I}$: Qwen 3 (0.6B) (Yang et al., 2025)	0.719 ± 0.06	0.525 ± 0.06	0.593 ± 0.08
$\mathcal{I}$: Qwen 3 (4B) (Yang et al., 2025)	0.713 ± 0.06	0.588 ± 0.05	0.554 ± 0.08
$\mathcal{I}$: Qwen 3 (8B) (Yang et al., 2025)	0.709 ± 0.06	0.561 ± 0.06	0.528 ± 0.07

Table 3. Main Experimental Results: Segment-Score. Comparison of semantic isotropy with other uncertainty metrics across **TriviaQA Entities**, **FactScore-Bio** and **Book Summaries** datasets using the *Segment-Score* algorithm. 1,500 bootstrapped samples are used to generate 1-SD error bars. †denotes API-only models. Note that EigenScore values are not generated for GPT 4.1 Mini.

	TriviaQA Entities			**FactScore-Bio**			**Book Summaries**	
Measure / Metric	**Llama 3.1**	**Phi 3.5**	**GPT 4.1**	**Llama 3.1**	**Phi 3.5**	**GPT 4.1**	**Llama 3.1**	**Phi 3.5**
Perplexity (Ren et al., 2023)	0.09 ± 0.02	0.26 ± 0.03	0.00 ± 0.00	0.09 ± 0.04	0.03 ± 0.03	0.01 ± 0.01	0.01 ± 0.01	0.03 ± 0.02
LN Entropy (Malinin & Gales, 2021)	0.09 ± 0.02	0.26 ± 0.03	0.03 ± 0.01	0.09 ± 0.04	0.03 ± 0.03	0.02 ± 0.02	0.01 ± 0.01	0.03 ± 0.02
U.Eign (Lin et al., 2024)	0.06 ± 0.02	0.18 ± 0.03	0.00 ± 0.00	0.28 ± 0.05	0.05 ± 0.03	0.02 ± 0.02	0.05 ± 0.02	0.07 ± 0.02
U.Deg (Lin et al., 2024)	0.05 ± 0.01	0.12 ± 0.03	0.00 ± 0.00	0.26 ± 0.07	0.05 ± 0.03	0.03 ± 0.02	0.06 ± 0.02	0.06 ± 0.03
U.NLI (Manakul et al., 2023)	0.08 ± 0.02	0.17 ± 0.03	0.00 ± 0.00	0.14 ± 0.06	0.06 ± 0.03	0.01 ± 0.01	0.04 ± 0.02	0.06 ± 0.02
EigenScore (Chen et al., 2024)	0.02 ± 0.01	0.00 ± 0.00	N/A	0.07 ± 0.04	0.02 ± 0.02	N/A	0.02 ± 0.02	0.00 ± 0.01
LUQ (Zhang et al., 2024)	0.03 ± 0.01	0.23 ± 0.03	0.10 ± 0.02	0.38 ± 0.05	0.10 ± 0.05	0.12 ± 0.04	0.05 ± 0.02	0.33 ± 0.04
Semantic Entropy (Farquhar et al., 2024)	0.03 ± 0.01	0.08 ± 0.02	0.00 ± 0.00	0.16 ± 0.07	0.03 ± 0.02	0.01 ± 0.01	0.02 ± 0.01	0.03 ± 0.02
$\mathcal{I}$: Gemini† (Gemini Team, 2024)	$\mathbf{0.24 \pm 0.03}$	$\mathbf{0.43 \pm 0.02}$	0.03 ± 0.01	0.44 ± 0.04	0.24 ± 0.07	0.34 ± 0.04	0.26 ± 0.03	$\mathbf{0.43 \pm 0.04}$
$\mathcal{I}$: OpenAI Small† (OpenAI, 2024)	0.16 ± 0.02	0.28 ± 0.02	0.00 ± 0.00	0.44 ± 0.04	0.27 ± 0.07	0.31 ± 0.04	0.26 ± 0.03	0.35 ± 0.03
$\mathcal{I}$: OpenAI Large† (OpenAI, 2024)	0.16 ± 0.02	0.27 ± 0.02	0.00 ± 0.00	0.44 ± 0.03	0.27 ± 0.06	0.32 ± 0.04	0.20 ± 0.03	0.37 ± 0.03
$\mathcal{I}$: Nomic v1 (Nussbaum et al., 2025)	0.27 ± 0.02	0.36 ± 0.02	$\mathbf{0.14 \pm 0.03}$	0.45 ± 0.04	0.27 ± 0.06	0.42 ± 0.04	0.24 ± 0.03	0.34 ± 0.03
$\mathcal{I}$: Qwen 2 (7B) (Yang et al., 2024a)	0.08 ± 0.02	0.27 ± 0.03	0.02 ± 0.01	0.49 ± 0.04	$\mathbf{0.36 \pm 0.06}$	0.42 ± 0.04	0.12 ± 0.03	0.30 ± 0.03
$\mathcal{I}$: SF Mistral (Meng et al., 2024)	0.08 ± 0.02	0.30 ± 0.03	0.01 ± 0.01	0.45 ± 0.04	0.29 ± 0.06	0.46 ± 0.04	0.11 ± 0.02	0.24 ± 0.03
$\mathcal{I}$: Mistral (E5) (Wang et al., 2024a)	0.11 ± 0.02	0.32 ± 0.02	0.00 ± 0.00	0.45 ± 0.04	0.28 ± 0.06	$\mathbf{0.47 \pm 0.04}$	0.09 ± 0.02	0.23 ± 0.03
$\mathcal{I}$: Cohere v4.0 † (AI, 2025)	0.11 ± 0.02	0.28 ± 0.03	0.00 ± 0.00	$\mathbf{0.50 \pm 0.03}$	0.30 ± 0.07	0.41 ± 0.04	0.24 ± 0.03	0.41 ± 0.03
$\mathcal{I}$: Cohere v3.0 † (AI, 2023)	0.11 ± 0.02	0.28 ± 0.03	0.00 ± 0.00	0.50 ± 0.04	0.27 ± 0.07	0.41 ± 0.04	0.20 ± 0.03	0.34 ± 0.03
$\mathcal{I}$: Cohere v3.0 (Lite) † (AI, 2023)	0.11 ± 0.02	0.28 ± 0.03	0.00 ± 0.00	0.50 ± 0.03	0.24 ± 0.06	0.41 ± 0.04	0.25 ± 0.03	0.36 ± 0.03
$\mathcal{I}$: Qwen 3 (0.6B) (Yang et al., 2025)	0.05 ± 0.02	0.17 ± 0.02	0.01 ± 0.01	0.47 ± 0.04	0.31 ± 0.06	0.42 ± 0.04	0.18 ± 0.03	0.40 ± 0.03
$\mathcal{I}$: Qwen 3 (4B) (Yang et al., 2025)	0.01 ± 0.01	0.17 ± 0.03	0.14 ± 0.03	0.44 ± 0.04	0.31 ± 0.06	0.37 ± 0.05	0.31 ± 0.03	0.40 ± 0.03
$\mathcal{I}$: Qwen 3 (8B) (Yang et al., 2025)	0.00 ± 0.00	0.12 ± 0.02	0.13 ± 0.03	0.43 ± 0.04	0.28 ± 0.06	0.33 ± 0.05	$\mathbf{0.33 \pm 0.03}$	0.36 ± 0.04

Table 4. Comparison of different isotropy measures: Analysis of various isotropy measures. Comparing explained variance (R^2) of Factuality on ≈500 word length responses generated using Phi 3.5 Mini Instruct on the **TriviaQA (SS)** dataset. 1,500 bootstrapped samples are used to generate 1-SD error bars.

Model	Frob.	Inv. Tr.	LogDet	vNE
Gemini	0.41 ±0.02	0.38 ±0.02	**0.43 ±0.02**	0.43 ±0.02
Nomic v1	0.33 ±0.02	0.33 ±0.02	**0.39 ±0.02**	0.35 ±0.02
OpenAI (L)	0.26 ±0.02	0.21 ±0.02	0.26 ±0.02	**0.27 ±0.02**
OpenAI (S)	0.27 ±0.02	0.17 ±0.02	0.24 ±0.02	**0.28 ±0.02**
Cohere v4	0.27 ±0.02	0.22 ±0.02	0.26 ±0.02	**0.28 ±0.02**
Cohere v3	0.27 ±0.02	0.22 ±0.02	0.26 ±0.02	**0.28 ±0.02**
Qwen 2	**0.27 ±0.02**	0.17 ±0.02	0.23 ±0.02	0.27 ±0.02
Mistral	0.29 ±0.02	0.19 ±0.02	0.27 ±0.02	**0.30 ±0.02**

B. Experimental Details

B.1. Effect of Number of Sampled Responses.

Figure 7 illustrates how performance scales with the number of sampled responses. Encouragingly, semantic isotropy scoring does not require many samples (≈6-8) to achieve comparable performance to its implied asymptotic level (at $N = 10$ samples), across both small and large models. While the method is generally scalable, this result confirms that it can deliver consistent truthfulness indicators with minimal overhead.

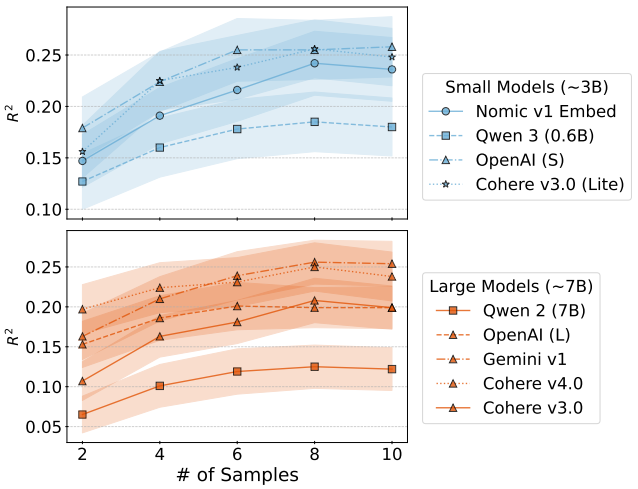


Figure 7. Performance by # of Samples: Line Charts comparing the performance of semantic isotropy by number of sampled responses on **FS-BIO (SS)** dataset of ≈500 words, contrasted by model size: Small (up to $3B$ parameters) and Large (≈$7B$ or more parameters). 1,500 bootstrapped samples are used to generate 1-SD error regions.

B.2. Effect of Generator Capability Level.

In addition to the results discussed in Section 4.3, where we used long-form responses generated by Phi 3.5 Mini Instruct and Llama 3.1 8B Instruct, we also considered long-form responses generated by a more capable, closed-weight model, GPT 4.1 Mini. Applying the *Segment-Score* method to long-form responses for FS-BIO and TriviaQA, we find that the absolute performance of semantic isotropy scoring as well as its relative performance compared to the baselines significantly increases on FS-BIO but decreases on TriviaQA. On TriviaQA, semantic isotropy scoring achieves a higher R^2 than all baselines using Nomic v1 embeddings, but the best absolute R^2 value is barely above 10%, implying that almost none of the variation in the factuality scores can be explained by any of the methods used. Since, for FS-BIO, semantic isotropy scoring performs better on long-form responses generated by GPT 4.1 Mini than on long-form responses generated by Phi 3.5 Mini Instruct and Llama 3.1 8B Instruct, the poor performance (across methods) on TriviaQA is unlikely to be due to increased model capability per se and more likely due to dataset idiosyncrasies, such as inherent entity ambiguities, or model pre-training artifacts. We present a detailed analysis and discussion in Appendix C.1.

B.3. Computational Considerations.

End-to-end timing for the proposed method versus the strongest NLI-based baseline is reported in Section 4.6. We summarize the asymptotic analysis here for completeness. Existing methods such as LUQ and Graph Uncertainty scale as $\mathbf{O}(MN^2)$, where M is the average number of facts or segments per response and N is the number of sampled responses, since each pair requires a separate NLI forward pass. Semantic entropy and derived methods are $\mathbf{O}(N^2)$ in entailment computations, which also cannot be naively vectorized because the NLI model consumes (premise, hypothesis) string pairs rather than tensors. In contrast, while semantic isotropy is naively $\mathbf{O}(N^2)$, the cosine kernel can be computed in a fully parallelized fashion, yielding substantial performance improvements.

B.4. Effect of Choice of Embedding Aggregation.

Our experimental results require an embedding vector $z_i \in \mathbb{R}^D$. Closed-weight models return a single vector per response and allow the user to configure embedding size (D) as an input. We use the model default for Gemini (768), OpenAI (1536) and Cohere (1024). However, Open-weight embedding models like Qwen 2 (Yang et al., 2024a) and Mistral (Meng et al., 2024) specify using the last token embedding in the final hidden state. This can be found in their reference implementations on HuggingFace. In contrast, for Nomic v1

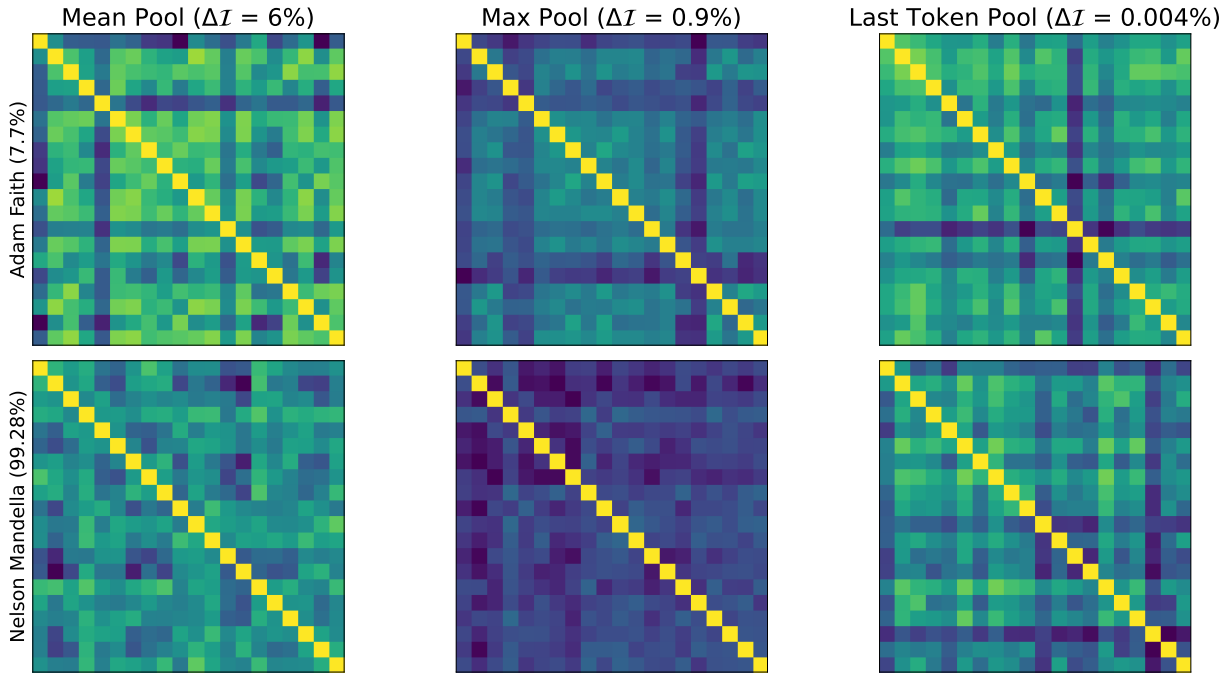


Figure 8. Nomic v1 Embedding Choice. K_E^{cos} for two sample topics: **Adam Faith** (low factuality, top row) and **Nelson Mandela** (high factuality, bottom row). The number in the parenthesis next to the topic name is the average factuality across sampled responses. $\Delta\mathcal{I}$ for each pooling method denotes the percentage difference in semantic isotropy scores between the two topics (as a percentage of $\mathcal{I}$ for **Nelson Mandela**). We infer that Mean Pooling for Nomic v1 yields the most informative matrices for an isotropy analysis.

(Nussbaum et al., 2025), authors do not recommend a pooling method and encourage users to find methods that work well for their use case. We find that the last token embedding method yields a weak semantic isotropy score. Instead, we use the mean pooling over the token dimension. Given model activations of the last hidden layer as a matrix $E \in \mathbb{R}^{L \times D}$, where L is the number of tokens in the response and D the embedding dimension, we produce $Z_{\texttt{nomic}} = \texttt{mean}(E, \texttt{dim=1})$ such that $Z_{\texttt{nomic}} \in \mathbb{R}^D$. In Figure 8, we show the difference between a highly trustworthy and a low trustworthy topic by mean, max and last token pooling for Nomic v1.

B.5. Infrastructure Requirements

Our work was performed using Google Cloud Platform compute to run inference on the white-box models using a single node with up to 4 NVIDIA Tesla V100 GPUs. GPT 4.1 Mini was used as the oracle LLM in Algorithm 1. OpenAI `text-embedding-3-small`, `text-embedding-3-large`, Gemini `gemini-embedding-001`, Cohere Embedding v4.0, v3.0 (AI, 2023; 2025) were also used to compute response embeddings for semantic isotropy computation. Additionally, OpenAI GPT 4.1, Anthropic Claude 4 Sonnet, and DeepSeek (DeepSeek-AI et al., 2024) were also used in verification runs to study robustness of the scoring pipeline. Overall, we utilize approximately 1,440 GPU hours (in addition to CPU compute time) to generate samples and score the responses.

B.6. Deploying Semantic Isotropy Score in a Task-Specific Context

The discriminative analysis is presented in the main body as Section 4.5 and Figure 6. As measures of utility or correctness over long-form generations could be real valued and non-binary, we discretize the experimental setup by classifying responses with average factuality $> 50\%$ as 1 and 0 otherwise, then use $1 - \mathcal{I}$ as the confidence measure. The resulting plots are consistent with our regression results, with performance varying across datasets while working well in most settings; in the more specific datasets such as **FS-BIO** and **BS**, the score performs particularly well at classifying responses.

B.7. Segment-Score Implementation Specifics

Unlike *FactScore*, the *Segment-Score* algorithm is achieved through a single prompt per response. The basic prompt template for our method can be found in Figure 13. The prompt is structured into three components. First, an instruction block that covers the overall objective of the scoring exercise, expected data, desired output structure, and specific handling for edge cases. Next, we provide two manually curated examples to guide the model's output. One such example can be found in Figure 14. Last, we provide the entity name, contents of the reference document and response to be scored. The model is prompted to respond using an XML tag structure which

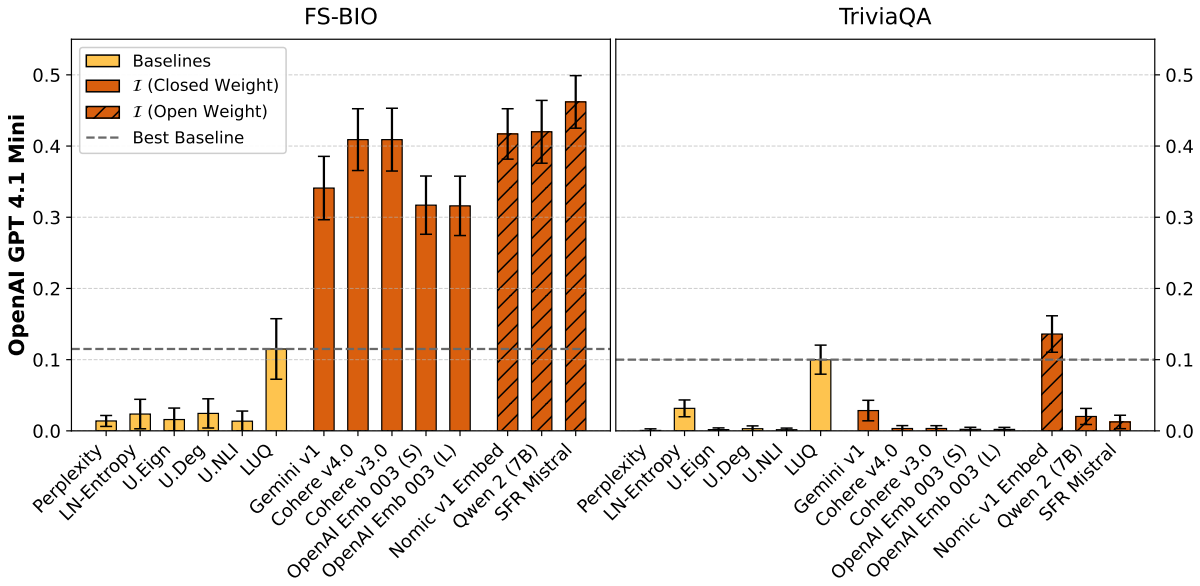


Figure 9. Experimental Results using OpenAI GPT 4.1 Mini. Bar charts comparing the performance (as measured by the R^2 of a linear model of factuality with semantic isotropy as the explanatory variable); implemented using various embedding models and benchmark uncertainty metrics on both the **FS-BIO** and **TriviaQA** datasets using the *Segment-Score* **(SS)** scoring algorithm. Left hand side: Benchmark UQ metrics. Right Hand Side: semantic isotropy ($\mathcal{I}$) implemented using various embedding models. Same experimental setting as in Figure 2. We observe that performance on generations using OpenAI GPT 4.1 Mini is significantly lower than other generation models; however, this extends to all benchmark metrics as well. Semantic isotropy continues to outperform all other baselines. Note that EigenScore values are not generated for GPT 4.1 Mini as it is a black-box model and the internal layer activations are not available.

simplifies the data extraction process. We also force the model to respond with only `'0'` or `'1'` for the output classes. For 99.8% of segments, the top 2 tokens are the '0' or '1' class labels, and we can use the log-probabilities from the API to generate a normalized probability score for each labeled segment.

C. Ablation Studies

C.1. Response Generation Using OpenAI GPT 4.1 Mini

To understand the impact of the generator model on the performance of the semantic isotropy score, we also generate responses using OpenAI GPT 4.1 Mini (OpenAI, 2025), scored using *Segment-Score*. Figure 9 shows the results of this experiment in the same format as our main results. For the **FS-BIO** dataset, we see semantic isotropy remains predictive of overall response factuality. However, we see a notable drop in performance in the case of **TriviaQA**. Notably, this extends to all benchmark metrics as well. We also observe much greater variance in the performance across embedding models, suggesting there are some settings where the choice of embedding model is crucial.

Consequently, it was important to understand the impact of the generation model on the performance of the semantic isotropy score. Zhang et al. (2024) note that the LUQ method does not perform well using GPT 4, a fact we were able to replicate.

> *"We also observe that LUQ is better suited for models with relatively lower factuality and a lack of self-expressiveness regarding uncertainty. For models with high factuality capabilities, such as GPT-4, LUQ only demonstrates a moderate correlation with factuality scores"* - (Zhang et al., 2024)

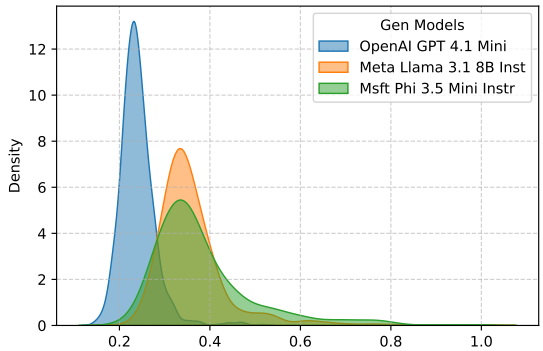
One would expect that GPT 4.1 Mini would perform better in terms of instruction following and maintaining consistent structure. This is anecdotally observed in the token count distribution (see Figure 11b), where GPT 4.1 Mini adheres much better to the 500 word guidance given in the prompt across different topic areas, as compared to the other generative models being considered.

When studying the distribution of semantic isotropy scores, in Figure 11a, we observe that OpenAI GPT 4.1 Mini's scores are smaller in magnitude and more tightly clustered. This highlighted that there may be some inherent structural properties of the OpenAI GPT 4.1 Mini generations that manifested in the embeddings in this specific way. To tackle this, we performed a comparative study between generations from GPT 4.1 Mini (OpenAI, 2025), Phi 3.5 Mini (Abdin et al., 2024) and Llama 3.1 8B (Grattafiori et al., 2024). Using an LLM-as-a-judge setup, we asked OpenAI GPT 4.1 to compare and contrast 10 generations from GPT 4.1 Mini vs 10 from either Phi 3.5 Mini or Llama 3.1 8B for 50 different topics where all of the models had relatively low factuality (i.e. $\leq 33\%$ accuracy on average). The LLM generated comparisons were summarized into a brief analysis (See Figure 10). This highlighted that properties such as

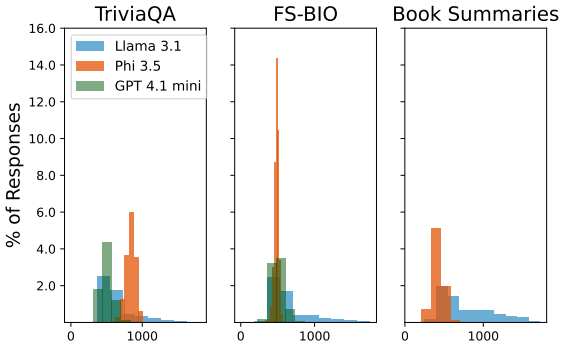
```
Across the comparisons, Model A consistently demonstrates formal, polished, and
    structured writing with high factual accuracy, depth, and analytical rigor. Its
    responses are typically essay-like, varied in perspective, and nuanced, often
    employing advanced vocabulary, rhetorical devices, and critical engagement with
    broader themes such as legacy, cultural context, and impact.

In contrast, Model B tends to be more informal, conversational, and formulaic,
    frequently exhibiting repetition, superficial coverage, and factual errors. Model
    B's diversity often arises from inconsistency or error rather than true creativity
    , and its outputs are generally less coherent, less organized, and less engaging
    than Model A's.
```

Figure 10. Summary Analysis of Low Factuality Generations by GPT 4.1 Mini (Model A) vs Llama 3.1 8B (Model B).



(a) Kernel Density Estimation plots of semantic isotropy scores for TriviaQA. Kernel Density Estimation plots of semantic isotropy scores for TriviaQA, comparing the distributions of scores for GPT 4.1 Mini, Phi 3.5 Mini and Llama 3.1 8B. The embedding used in Gemini v1 (Gemini Team, 2024). We observe that GPT 4.1 Mini's scores are smaller in magnitude and more tightly clustered.



(b) Word Count by Generation Model and Dataset: Density histograms comparing the response length (in tokens) for each generative model and dataset for a prompt with a guidance length of 500 words. We observe that GPT 4.1 Mini is much better at instruction following than Llama 3.1 and Phi 3.5 Mini, sticking fairly closely to the intended word limit.

Figure 11. Response Characteristics: Comparing the characteristics of generated responses by model type and dataset. Distribution of semantic isotropy scores for TriviaQA using Gemini [Left] and Histogram of response lengths in tokens by dataset and generative model [Right].

structure, tone and other qualitative aspects of a generation materially impacted the resulting embeddings. We expect that improvements in embedding models or models specifically tuned to disregard qualitative aspects of a generation could improve the performance of the semantic isotropy score, but leave this exploration for future work.

C.2. Comparing the Efficiency of the Two Scoring Methods

Segment-Score's overall process closely follows the FactScore algorithm. However, we make several key optimizations that improve runtime. First, unlike FactScore, we break the input text into atomic segments rather than extracting atomic facts. As a result, each sentence may be broken into 2-3 segments that need to be scored whereas FactScore may generate several atomic facts pertaining to each sentence segment. For **FS-BIO**, this translates to 28 ± 6 distinct segments for each response in *Segment-Score*, while *FactScore* produces 100 ± 20 atomic facts (2-σ interval).

Second, FactScore does not distinguish between verifiable and ambiguous facts which increases noise in the final results. We explicitly prompt the model to classify such examples as False. This generally biases our result lower, which can be seen in the middle plot of Figure 12.

Finally, our entire process is achieved using one single query to the oracle, while FactScore may use an order of magnitude more depending on the number of atomic facts extracted and the length of the ground truth document. Consequently, our method scales more favorably as the length of the response to be scored increases. Despite these differences, we see comparable results for **FS-BIO** in Table 2 and Table 3 for each scoring method when considering the various isotropy configurations and response generating models.

D. Extended Discussion of Limitations

A condensed version of this discussion appears in the main body (Section 6). The first drawback of our method is the need for repeated sampling, which can be expensive to generate. Approaches such as Kossen et al. (2024) have shown that it is possible to train cheap probes that can a priori estimate uncertainty scores without needing repeated sampling. With the advent of reasoning models with long

Chain-of-Thought reasoning contexts (Zhao et al., 2024) and Monte-Carlo Tree Search (Wang et al., 2024b), as well as fast inference engines with significantly higher throughput (Kwon et al., 2023), the relative overhead of repeated sampling has been greatly diminished, making methods like ours compelling.

Additionally, while studies such as Manakul et al. (2023) have observed that hallucinated responses tend to diverge, and generally are inconsistent with each other, Ricco et al. (2025) find that this effect is not universal and introduce the taxonomy of *extrinsic* (characterized by arbitrary and orthogonal content) vs *intrinsic* (contradictory or anti-parallel content) hallucination. The semantic isotropy framework would likely fail in cases of *intrinsic* hallucinations. However, in practice these intrinsic hallucinations would be exceedingly rare, especially in the long-form response format.

Finally, it is important to call out that our method is not designed to work in adversarial scenarios. In the event that one were to use *semantic isotropy* on a model that had been intentionally trained on faulty content, semantic consistency across samples would only indicate the model is faithfully inferring from its training corpora, and would not indicate factuality or trustworthiness. While *semantic isotropy* remains far more robust to naive manipulation techniques such as ballot stuffing, whereby facts are repeatedly stated to boost accuracy metrics, this failure case is beyond the scope of standard embedding models. In this case, we deem *semantic isotropy* as a complementary early-warning signal, rather than a full-fledged replacement for claim-by-claim verification.

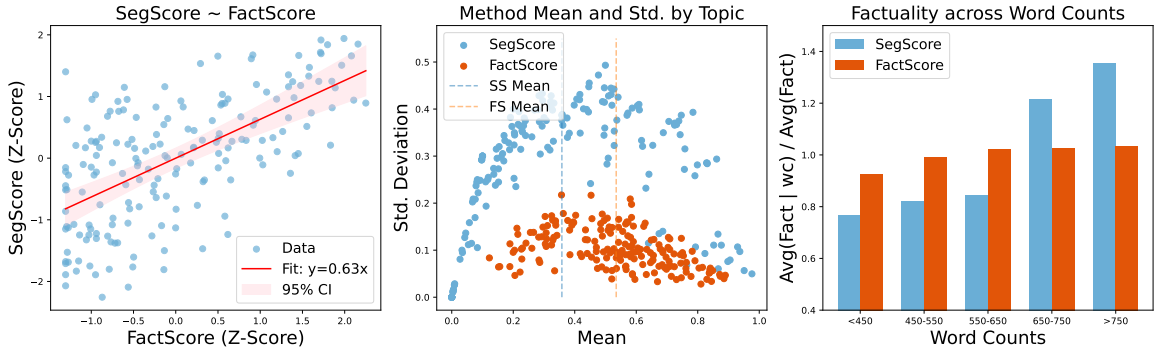


Figure 12. FS vs SS [FS-BIO]. Comparison of *FactScore* and *Segment-Score* on the **FS-BIO** dataset. *Segment-Score* yields a dataset with lower overall factuality and higher variance; however, the relationship between the two scores is robust and consistent.

E. Reproducibility

We have made a significant effort to ensure the reproducibility of our results. Code is provided at the GitHub repository linked in the introduction, which includes training, evaluation, and analysis scripts. The experimental setup—including hyperparameters, model configurations, and sampling parameters—is described in Section 4.1. All datasets used in our experiments are publicly available, and we additionally provide scripts for data preparation. Our experiments rely on open-weight language models and certain API-only models; instructions for reproducing results with both are included in the `README.md` file in the released code. Infrastructure and compute requirements are documented in Appendix B.5. The *Segment-Score*-annotated dataset is released alongside the code repository, enabling direct benchmarking of our results.

```
// INSTRUCTION PROMPT FOR SEGMENT SCORE
You are an NLP segmentation and evaluation engine.
Examine the scenario below. You are given:
1. The name of an entity/person/place/thing etc. in <entity> tags.
2. A reference document regarding the entity in <reference_doc> tags.
3. A response about the entity to evaluate in <response> tags.

### Your Tasks:
1. **Segmentation Task:**
  Segment the `<response>` into individual statements. Each statement can be a sentence,
      phrase or word and should convey a single, complete, and independent piece of
      information about the `<entity>`. Do not modify, rephrase, or paraphrase the
      original text. Ensure no semantic overlaps exist between statements. Individual
      proper nouns should be part of their own statement however when determining the
      appropriate classification, preceding context can be used when appropriate.
  Verify that the concatenated content of all statements exactly matches the original
      response.

  Format the segmented response as follows:
  ```
  <statements>
  <statement>Statement 1</statement> <class>Class 1</class>
  <statement>Statement 2</statement> <class>Class 2</class>
  ...
  </statements>
  ```

2. **Factual Classification Task:**
  For each segmented `<statement>`, classify it as 1 (True) or 0 (False) based solely on
      the information in the `<reference_doc>`. Follow these guidelines:
  - If a statement is factually accurate and supported by the `<reference_doc>`, classify
      it as '1'.
  - If a statement is inaccurate, unverifiable, or not supported by the `<reference_doc>`,
       classify it as '0'.
  - If a statement is partially true, but contains incorrect or unsupported information,
      classify it as '0'.
  - Do not rely on any external knowledge or context beyond the `<reference_doc>`.
  - Include only the classification for each statement. Do not provide any explanations or
       additional information.
  - Specify the class in <class> tags.
  - The ONLY valid class values are `1` and `0`. No other values or words should appear
      within the `<class>` tags.

3. **Error Handling:**
  - If the `<response>` contains unparseable text, incomplete sentences, or conflicting
      information that cannot be resolved using the `<reference_doc>`, include the flagged
       statement as is and classify it as 'False'.

Examples:
####### EXAMPLE 1 ######
{{example_one}}
########################
####### EXAMPLE 2 ######
{{example_two}}
########################

Entity:
<entity>
{entity}
</entity>

Reference Document:
<reference_doc>
{reference_doc}
</reference_doc>

Response to Evaluate:
<response>
{response}
```


Figure 13. Segment-Score Prompt Template with instructions

```
<entity>
London, UK
</entity>

Reference Document:
<reference_doc>
London, England's capital, boasts a rich history spanning millennia. Founded by the Romans
     as Londinium around 47 AD, it became a major port and trading center. After the Roman
     withdrawal, Anglo-Saxons established Lundenwic, which later fell to Viking raids. The
     Norman Conquest in 1066 led to the construction of the Tower of London, a symbol of
    royal power. London thrived during the medieval period, becoming a major center for
    trade, finance, and culture. It weathered plagues, fires, and civil wars, emerging as
    a global metropolis and the heart of the British Empire. Today, London remains a
    vibrant hub, blending its historical legacy with modern dynamism, home to over 9
    million people.
</reference_doc>

Response to Evaluate:
<response>
London, the capital city of England and the United Kingdom, is a vibrant metropolis
    steeped in history and brimming with modern energy. With a population of over 9
    million people, it stands as one of the world's most influential global cities, known
    for its diverse culture, iconic landmarks, and rich heritage.

The city's history stretches back over three millennia, founded by the Romans as Londinium
     in 43 AD. Throughout the centuries, London has played a pivotal role in world affairs,
     serving as the heart of the British Empire and surviving tumultuous events such as
    the Great Fire of 1666 and the Blitz during World War I.

Today, London is a melting pot of cultures, with over 300 languages spoken within its
    boundaries. This diversity is reflected in its neighborhoods, each with its own unique
     character and charm. From the trendy streets of Shoreditch to the upscale boutiques
    of Mayfair, there's something for everyone in this cosmopolitan city.
</response>

Segmented and classified response:
<statements>
<statement>London, the capital city of England and the United Kingdom</statement> <class
    >1</class>
<statement>is a vibrant metropolis steeped in history and brimming with modern energy</
    statement> <class>1</class>
<statement>With a population of over 9 million people</statement> <class>1</class>
<statement>it stands as one of the world's most influential global cities, known for its
    diverse culture, iconic landmarks, and rich heritage.</statement> <class>1</class>
<statement>The city's history stretches back over three millennia, founded by the Romans
    as Londinium in 43 AD</statement> <class>0</class>
<statement>Throughout the centuries, London has played a pivotal role in world affairs,
    serving as the heart of the British Empire</statement> <class>1</class>
<statement>and surviving tumultuous events such as the Great Fire of 1666</statement> <
    class>1</class>
<statement>and the Blitz during World War I</statement> <class>0</class>
<statement>Today, London is a melting pot of cultures, with over 300 languages spoken
    within its boundaries</statement> <class>1</class>
<statement>This diversity is reflected in its neighborhoods, each with its own unique
    character and charm.</statement> <class>1</class>
<statement>From the trendy streets of Shoreditch to the upscale boutiques of Mayfair,
    there's something for everyone in this cosmopolitan city</statement> <class>1</class>
</statements>
```

Figure 14. Segment-Score Example used as part of model in-context training for oracle LLM.