ESTIMATING SEMANTIC ALPHABET SIZE FOR LLM UNCERTAINTY QUANTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many black-box techniques for quantifying the uncertainty of large language models (LLMs) rely on repeated LLM sampling, which can be computationally expensive. Therefore, practical applicability demands reliable estimation from few samples. Semantic entropy (SE) is a popular sample-based uncertainty estimator with a discrete formulation attractive for the black-box setting. Recent extensions of semantic entropy exhibit improved LLM hallucination detection, but do so with less interpretable methods that admit additional hyperparameters. For this reason, we revisit the canonical discrete semantic entropy estimator, finding that it underestimates the "true" semantic entropy, as expected from theory. We propose a modified semantic alphabet size estimator, and illustrate that using it to adjust discrete semantic entropy for sample coverage results in more accurate semantic entropy estimation in our setting of interest. Furthermore, our proposed alphabet size estimator flags incorrect LLM responses as well or better than recent top-performing approaches, with the added benefit of remaining highly interpretable.

1 INTRODUCTION

Large language models (LLMs) are not fact engines. They have been shown to forget provided context (Liu et al., 2024), fabricate records (Lee et al., 2023), and otherwise hallucinate (Ji et al., 2023). LLMs' underlying training data may be of mixed factual reliability, but models can also hallucinate when they have ample knowledge to respond adequately (Simhi et al., 2024). In risk-sensitive settings, it may be prudent for systems to abstain when uncertainty is high (or, alternatively, confidence is low) (Murphy, 2022; Hasan et al., 2025). For these reasons and others, it is prudent to estimate LLMs' intrinsic uncertainty.

Uncertainty quantification (UQ) in LLMs is particularly challenging, however, in part due to their computational scale (Liu et al., 2025). Extensive sampling can be financially prohibitive or computationally intractable, and these concerns are magnified if the UQ method is computationally complex with respect to the number of samples. Furthermore, internal activations and sequence log-probabilities are not always available from commercial inference providers (Farquhar et al., 2024), potentially disqualifying so-called "white-box" methods. Therefore, sample-efficient UQ in the black-box setting - where LLM internals are assumed inaccessible - is a crucial area of research for deploying trustworthy AI systems.

We provide empirical evidence that canonical discrete semantic entropy underestimates the "true" SE for typical sample sizes, on average (Figure 2). To address this limitation, we first suggest a straightforward modification to existing alphabet size estimators, where we draw, in part, from population ecology. Then, we use it to adjust entropy estimates for sample coverage, resulting in reduced bias (Figure 2) and more accurate estimation of the "true" semantic entropy, compared to other black-box semantic entropy estimators (Table 1).

Finally, we evaluate our suggested estimators for inaccuracy classification in sentence-length LLM responses. In our experiments, the aforementioned coverage-adjusted estimator outperforms other explicit discrete estimators of semantic entropy. More strikingly, the alphabet size estimator alone outranks all other considered UQ methods in overall strength - except for one, from which we do not statistically distinguish performance, after accounting for uncertainty in establishing an overall

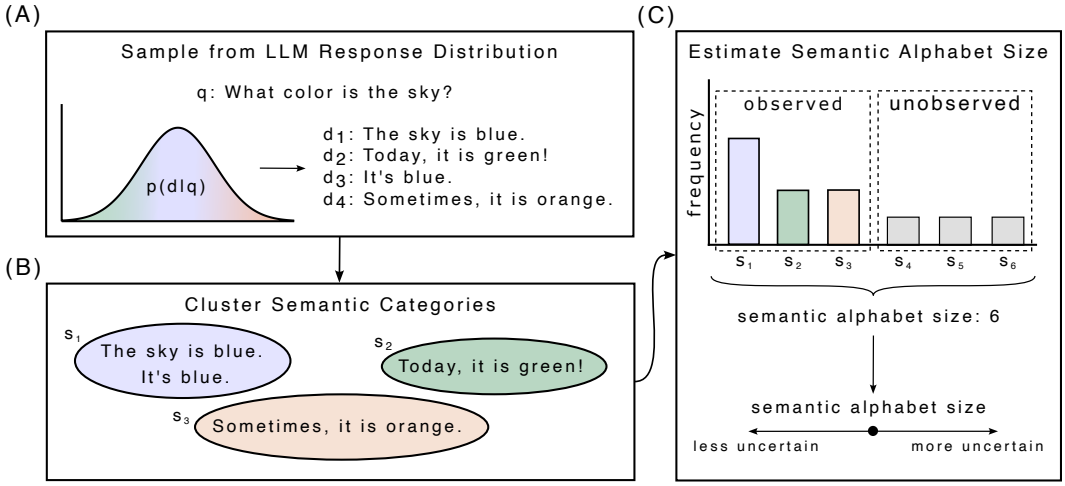

Figure 1: High-level schematic of semantic alphabet size estimation for LLM uncertainty quantification (Section 3). **(A)** Generate LLM responses to a query. **(B)** Assign responses to categories of shared meaning. **(C)** Estimate semantic alphabet size, accounting for semantic classes unobserved in the sample (Equation 9). LLM response examples are hypothetical for illustrative purposes.

rank of methods. Our results indicate that semantic alphabet size estimation, which is highly interpretable, can perform as well or better than state-of-the art methods in LLM inaccuracy detection.

2 Preliminaries

Information entropy. Information entropy quantifies the expected surprise, or uncertainty, of a random variable. For a discrete random variable X with finite alphabet S, it is given by:

$$\mathbb{H}(X) = -\sum_{s \in S} p(s) \log p(s), \tag{1}$$

where $p(s)$ depicts the probability of symbol s (Shannon, 1948). When a random variable's probability density is not known analytically, entropy estimators are employed. A well-studied approach in the finite-sample regime is the plugin approximation:

$$\hat{\mathbb{H}}_{plugin}(X) = -\sum_{i=1}^{k} \hat{p}(s_i) \log \hat{p}(s_i), \tag{2}$$

where k is the number of distinct observations in a sample, and $\hat{p}(s_i)$ conveys the empirical frequency of s_i among the observations.

Semantic classification. Lexically distinct text sequences may belong to the same *semantic equivalence class* (equivalently, "semantic category" or "semantic set") - i.e., a set of text sequences with mutually shared meaning. Sequence d_1 entails sequence d_2 if d_1 implies d_2 (Dagan et al., 2005). Kuhn et al. (2023) assign sequences d_1 and d_2 to a shared semantic equivalence class under so-called "strict entailment" if their textual entailment is bidirectional.

Semantic entropy. Semantic entropy (SE), introduced by Kuhn et al. (2023), aims to quantify intrinsic LLM uncertainty with a three-step procedure:

1. **Sampling**: Given a query q, generate n passages $d_1, d_2, \ldots, d_n$.
2. **Semantic Clustering**: Iterating over pairs (d_i, d_j), determine if both sequences belong to the same semantic equivalence class. Greedily assign passages to classes based on the pairwise semantic classifications.

3. **Estimation**: Semantic equivalence classes are treated as symbols in an alphabet S. SE is the entropy calculated over observed semantic equivalence classes:

$$SE(q|\theta) = -\sum_{s \in S} p(s|q,\theta) \log p(s|q,\theta) \tag{3}$$

where $p(s|q,\theta)$ represents the probability that an LLM parameterized by θ generates a passage belonging to semantic equivalence class s in response to a query q.

Because the distribution over semantic sets is not known, it is approximated from response probabilities $p(d|q,\theta)$ using so-called "Rao–Blackwellized Monte Carlo integration":

$$p(s_i|q,\theta) \approx \frac{\sum_d \mathbb{1}_{d \in s_i} p(d|q,\theta)}{\sum_d p(d|q,\theta)}, \tag{4}$$

where $d \in s$ indicates that response d belongs to semantic equivalence class s, and $p(d|q,\theta)$ are response probabilities returned by the LLM, for each of the k observed semantic categories $s_1, s_2, \ldots, s_k$ (Farquhar et al., 2024).

The black box setting. Response probabilities are not always available. Farquhar et al. (2024) consider a discrete formulation of semantic entropy (DSE), where aggregated document probabilities for each semantic equivalence class are replaced with empirical class frequencies:

$$DSE(q|\theta) = -\sum_{i=1}^{k} \left(\frac{\sum_d \mathbb{1}_{d \in s_i}}{n} \right) \log \left(\frac{\sum_d \mathbb{1}_{d \in s_i}}{n} \right) \tag{5}$$

with n sampled passages and k observed semantic categories. This is the plugin estimator (Equation 2) applied to semantic entropy (Equation 3). Both SE and DSE correspond with hallucination rate in question-answering problems (Farquhar et al., 2024).

Generalizations. Recent work is said to have generalized semantic entropy. Two pertinent examples are Kernel Language Entropy (KLE) and Semantic Nearest Neighbor Entropy (SNNE), which reported state-of-the-art performance for incorrectness classification (Nikitin et al., 2024; Nguyen et al., 2025). These stronger estimators may come at the expense of interpretability, in part owing to implementation complexity (e.g., applying a kernel to embed graph nodes in KLE) and introduction of additional hyperparameters and design choices (e.g., similarity function and scale parameter in SNNE). We elaborate on KLE and SNNE definitions and implementation details in Appendix A.5.

3 METHODS

The total number of semantic equivalence classes (i.e., the *semantic alphabet size* $|S|$) that may be ellicited from an LLM in response to a prompt is generally unknown, and not all categories are necessarily observed in the sample (i.e., $k < |S|$). For instance, under a simple Zipfian model of semantic category-abundance, we expect at least one category to be unobserved for sample size $n = 10$ if $|S| > 4$ (Appendix B). This situation is known as the *under-sampled regime*, where the empirical distribution over semantic categories can be less surprising than than the true one. In other words, the plugin method for DSE may underestimate LLM uncertainty, which we illustrate in Figure 2. For this reason, we are interested in methods for estimating semantic alphabet size and adjusting for it when estimating semantic entropy.

Semantic alphabet size. Because plugin DSE does not directly account for unobserved semantic categories, the implicit alphabet size used by the plugin estimator is k, called "NumSets" by Lin et al. (2024). In the under-sampled regime, NumSets underestimates $|S|$, so a more accurate semantic alphabet size estimator for small n is desirable. Parallels exist with the so-called "unseen species" problem in population ecology: given a sample of n observations belonging to one or more species, estimate the number of yet unseen species that would be discovered by collecting additional observations (Fisher et al., 1943). In this setting, the sample coverage C, the fraction of all possible categories observed in a sample (i.e., $C = \frac{k}{|S|}$), is also of interest (Chao & Shen, 2003).

The so-called "Good-Turing" sample coverage estimator is $\hat{C}_{GT} = 1 - \frac{f_1}{n}$, where f_1 is the number of singletons, or semantic categories observed only once (Good, 1953). Modest arithmetic converts the Good-Turing sample coverage estimator into an alphabet size estimator:

$$\widehat{|S|}_{GT} = \frac{kn}{n - f_1}. \tag{6}$$

More recently, Lin et al. (2024) suggested a "continuous" NumSets analogue: responses are interpreted as nodes of a fully-connected graph G with weights $w_{ij} = \frac{a_{ij}+a_{ji}}{2}$, where a_{ij} is the entailment probability for response pair d_i, d_j, via an NLI model. Given the eigenvalues ($\lambda_1 < \cdots < \lambda_n$) of G's normalized Laplacian, the estimator is given by:

$$U_{EigV} = \sum_{i=1}^{n} \max(0, 1 - \lambda_i). \tag{7}$$

Coverage-adjusted entropy. To quantify a population's ecological diversity, Chao & Shen (2003) provide a coverage-adjusted discrete entropy estimator that scales empirical category frequencies by estimated sample coverage:

$$\hat{\mathbb{H}}_{CS} = -\sum_{i=1}^{k} \frac{\hat{C}_{GT}\hat{p}_i \log(\hat{C}_{GT}\hat{p}_i)}{1 - (1 - (\hat{C}_{GT}\hat{p}_i))^n}. \tag{8}$$

This so-called "Chao-Shen" estimator is consistent and less biased than many empirical alternatives (Vu et al., 2007; Pinchas et al., 2024).

Hybrid estimators. First, we adapt the two aforementioned semantic alphabet size estimators to address shortcomings of each. When the number of singletons is zero, $\widehat{|S|}_{GT}$ reduces to NumSets, and it is undefined when all samples belong to distinct semantic categories. On the other hand, U_{EigV} can be less than k, which is a lower bound for $|S|$. To remediate the above limitations, we propose an alternative "hybrid" semantic alphabet size estimator:

$$\widehat{|S|}_{Hybrid} = \begin{cases} U_{EigV}, & \text{if } f_1 = n \\ max\left(\widehat{|S|}_{GT}, U_{EigV}\right), & \text{otherwise.} \end{cases} \tag{9}$$

Additionally, we propose a Chao-Shen-like DSE estimator that takes the form of Equation 8, but invokes the hybrid semantic alphabet size estimator for coverage adjustment:

$$\hat{\mathbb{H}}_{Hybrid} = -\sum_{i=1}^{k} \frac{\frac{k\hat{p}_i}{\widehat{|S|}_{Hybrid}} \log\left(\frac{k\hat{p}_i}{\widehat{|S|}_{Hybrid}}\right)}{1 - \left(1 - \frac{k\hat{p}_i}{\widehat{|S|}_{Hybrid}}\right)^n}. \tag{10}$$

4 Experiments

4.1 Experimental settings

Models. Following Farquhar et al. (2024), we focus on fine-tuned models - in our case, the instruction-tuned models of Gemma-2-9B (Team et al., 2024), Llama-3.1-8B (Grattafiori et al., 2024), Mistral-v0.3-7B, and Phi-3.5-3.8B (Abdin et al., 2024). We perform text generation at two temperatures, for distinct purposes. Following related prior work, we sample at temperature $\tau = 1.0$ to calculate uncertainty scores and again at $\tau = 0.1$ to obtain a "best guess" response for assessing the correctness of LLM responses (Farquhar et al., 2024; Nikitin et al., 2024; Nguyen et al., 2025). With the exception of Figure 2, which iterates over several sample sizes, the results shown in the main body of this work use a sample size of $n = 10$, which is also found in prior work on semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024; Nikitin et al., 2024; Nguyen et al., 2025). Due to its reduced computational burden compared to larger sample sizes and prevalence in the relevant literature, we will refer to $n = 10$ as a "practical" or "typical" sample size.

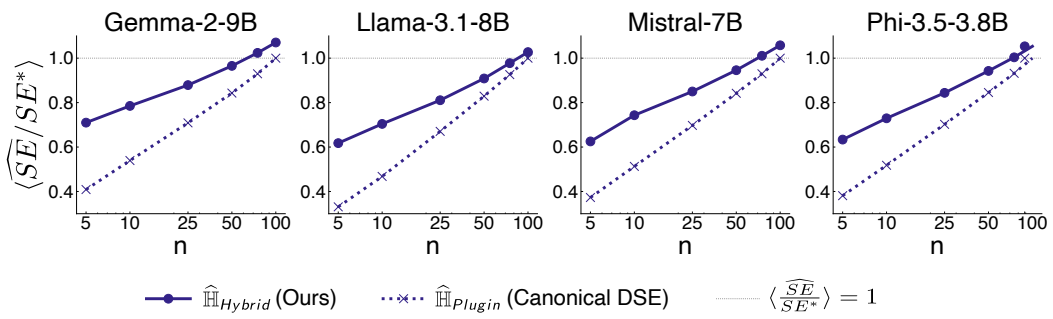

Figure 2: Illustrating underestimation in discrete semantic entropy calculation with typical sample sizes. The ratios of semantic entropy estimators with varying sample size ($n = 5, 10, 25, 50, 75, 100$) to white-box semantic entropy with $n = 100$ (denoted SE^*) are shown, with values below 1 suggesting underestimation (dotted grey line). The estimators displayed are the plugin estimator of canonical discrete semantic entropy (i.e., Equation 5, dotted indigo line) and the "hybrid" semantic entropy estimator of Equation 10 (solid indigo line). Results are averaged across queries within each dataset, then uniformly averaged across datasets. Log scale is used on the x-axis to highlight differences between estimators with smaller sample sizes. Instances with only one observed semantic category at $n = 100$, resulting in a denominator of 0, are ignored.

Datasets. We consider question-answering datasets wherein each query has either one or multiple correct answer(s). For the former, we use the validation sets of HotpotQA (Yang et al., 2018) and SQuAD 2.0 (Rajpurkar et al., 2018; 2016), which contain 1852 and 11864 exemplars, respectively. Our experiments require extensive sampling of LLM responses, and BEC demands combinatorially many language model calls. For these reasons, and because SQuAD 2.0 is much larger than the other data sets, our experiments using it are performed on a random 20% subset. We also consider queries with a wider variety of possible correct answers. The first such dataset is BioASQ, a biomedical QA benchmark with 4719 exemplars. Each question in BioASQ has between 1 and 15 reference answers. We also prepare a small (131 exemplars) supplementary dataset called Plethora Of accepTable cATegOries (POTATO), for which the number of possible correct semantic categories has an even larger range (up to 722). We discuss POTATO in greater detail in Section A.2.

4.2 Metrics and Baselines

Semantic entropy estimator evaluation. For a given prompt-LLM pair, $|S|$ and the true probability distribution over semantic equivalence classes are unknown, so semantic alphabet size and semantic entropy have no ground truths. Instead, we assume that white-box semantic entropy with a large number of samples ($n = 100$), denoted SE^*, is interchangeable with the true estimand (i.e., SE^* well-represents the true entropy over all possible semantic categories). When illustrating the degree to which SE^* is underestimated (i.e., Figure 2), we report the ratio of the estimator of interest to SE^*. Since information entropy is non-negative, values below 1 indicate underestimation. Our primary evaluation metric for entropy estimators, however, is mean-squared error (MSE): we assess the MSE between an estimate using the small sample size ($n = 10$) and SE^*.

Incorrectness evaluation. We also assess the ability of estimators to classify LLM responses as "correct" or "incorrect," sometimes referred to as "hallucination" or "confabulation" detection (Farquhar et al., 2024). We calculate the area under the receiver operating characterstic curve (AUROC), corresponding to the probability that a randomly selected incorrect LLM response has a higher uncertainty score than a randomly selected correct response. Empirical AUROC values are point estimates, which themselves have uncertainty. Nikitin et al. (2024) point out that such uncertainty is strongly driven by the model and dataset, rather than UQ method, motivating head-to-head comparisons and evaluation by win rate. We advance a similar approach, but we rely on Bradley-Terry latent strength scores (Zermelo, 1929; Bradley & Terry, 1952), calculated via minorization-maximization (Caron & Doucet, 2012; Hunter, 2004), allowing us to establish an overall rank of methods.

Dataset	Estimator	Gemma-2-9B	Llama-3.1-8B	Mistral-7B	Phi-3.5-3.8B
HotpotQA	$\widehat{\mathbb{H}}_{Plugin}$	0.47 ± 0.02	0.72 ± 0.03	0.60 ± 0.03	0.62 ± 0.03
	$\widehat{\mathbb{H}}_{CS}$	0.39 ± 0.02	0.57 ± 0.03	0.46 ± 0.03	0.47 ± 0.03
	$\widehat{\mathbb{H}}_{Hybrid}$	**0.30** ± 0.02	**0.45** ± 0.02	**0.39** ± 0.02	**0.39** ± 0.02
SQuAD 2.0	$\widehat{\mathbb{H}}_{Plugin}$	0.69 ± 0.03	0.84 ± 0.03	1.40 ± 0.04	1.46 ± 0.04
	$\widehat{\mathbb{H}}_{CS}$	0.51 ± 0.03	0.60 ± 0.03	0.86 ± 0.06	0.91 ± 0.06
	$\widehat{\mathbb{H}}_{Hybrid}$	**0.43** ± 0.02	**0.50** ± 0.02	**0.68** ± 0.03	**0.74** ± 0.03
POTATO	$\widehat{\mathbb{H}}_{Plugin}$	0.42 ± 0.08	0.71 ± 0.12	1.75 ± 0.15	1.96 ± 0.16
	$\widehat{\mathbb{H}}_{CS}$	0.30 ± 0.07	0.52 ± 0.11	0.94 ± 0.29	1.57 ± 0.55
	$\widehat{\mathbb{H}}_{Hybrid}$	**0.27** ± 0.06	**0.42** ± 0.09	**0.70** ± 0.11	**0.72** ± 0.11
BioASQ	$\widehat{\mathbb{H}}_{Plugin}$	1.66 ± 0.03	1.81 ± 0.03	2.06 ± 0.03	1.85 ± 0.02
	$\widehat{\mathbb{H}}_{CS}$	0.96 ± 0.04	1.05 ± 0.05	1.31 ± 0.08	0.98 ± 0.04
	$\widehat{\mathbb{H}}_{Hybrid}$	**0.78** ± 0.02	**0.78** ± 0.02	**0.72** ± 0.02	**0.82** ± 0.02

Table 1: Empirically evaluating the accuracy of explicit discrete semantic entropy estimators. Values reflect MSE between the estimated value using $n = 10$ samples and white-box semantic entropy with $n = 100$ (i.e., SE^*). The lowest MSE value for each model-dataset pair is shown in bold. Intervals, shown in grey, reflect 95% CIs via the standard error of the mean.

We also employ a Monte Carlo procedure that attempts to account for uncertainty in estimating both AUROC and strength scores. First, we fit Gaussian uncertainty distributions about AUROC point estimates from the 95% confidence intervals (CIs) obtained via DeLong's method (DeLong et al., 1988; Sun & Xu, 2014), a modeling assumption we justify by the approximate normality of the U statistic (Mann & Whitney, 1947).[1] For each model-dataset pair, we simulate $L = 100$ matches between each pair of methods by comparing AUROC values sampled from their corresponding uncertainty distributions. After obtaining strength scores from the simulated matches, we calculate CIs about the Bradley-Terry scores and establish 95% CIs about the ranks of latent strengths using the methods of Gao et al. (2023). We provide additional details in Appendix A.6.

Baselines. Because our priority is black-box uncertainty estimation, we compare with other black-box approaches. For semantic entropy estimation, we compare our presented hybrid approach (i.e., Equation 10) to the other explicit discrete semantic entropy estimators: the canonical discrete approach of Farquhar et al. (2024) (i.e., the plugin estimator, Equation 5) and the coverage-adjusted estimator of Chao & Shen (2003) (i.e., Equation 8).

For incorrectness classification, we consider the three aforementioned explicit semantic entropy estimators, four alphabet size estimators, and three other uncertainty methods. The alphabet size estimators are the number of semantic categories (i.e., NumSets) (Kuhn et al., 2023), the alphabet size estimator converted from the Good-Turing sample coverage estimator (Good, 1953), the spectral estimator of Lin et al. (2024) (i.e., U_{EigV}), and our presented hybrid alphabet size estimator (i.e., Equation 9). The other uncertainty methods are Predictive Entropy (PE, see Appendix A.5) (Kadavath et al., 2022), SNNE (Nguyen et al., 2025), and KLE (Nikitin et al., 2024).

4.3 Results

Canonical DSE underestimates "true" semantic entropy. The plugin estimator for information entropy $\hat{\mathbb{H}}_{plugin}$ has a theoretically-established negative bias (Basharin, 1959; Harris, 1975), which largely governs its mean-squared error (Antos & Kontoyiannis, 2001). We are interested in observing this phenomenon empirically, where we do not have access to the true distribution over semantic equivalence classes.

[1] In principle, samples from these distributions may exist outside the $[0, 1]$ range, but this is unlikely in our case, given the AUROC CI bounds (Figure 7).

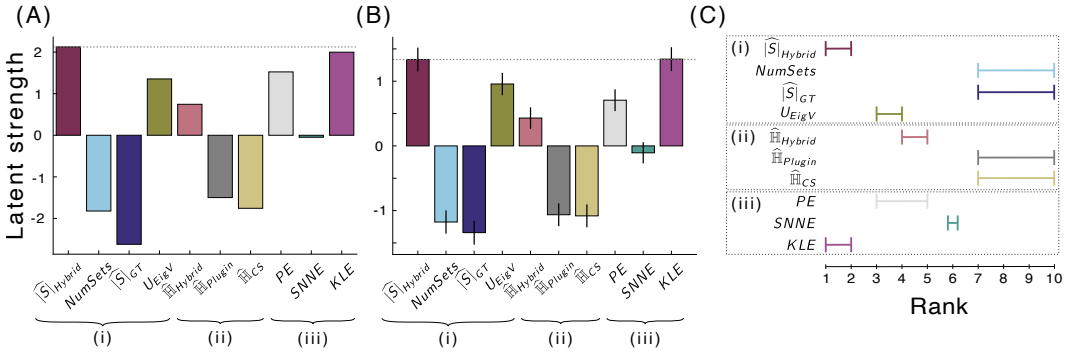

Figure 3: Establishing overall performance of ten UQ methods on incorrectness detection. **(A)** Bradley-Terry latent strength scores from pairwise comparison of AUROC point estimates. **(B)** Bradley-Terry latent strength scores after accounting for uncertainty in estimating AUROC. Error bars are "conservative" CIs about strength scores, which may be slightly stricter than 95%; see Section A.6 for details. **(C)** For each method, we establish 95% CIs about the rank of Bradley-Terry latent strength MLEs (Gao et al., 2023) for the incorrectness detection task; see Section A.6 for details. We highlight **i)** semantic alphabet size estimators, **ii)** explicit discrete semantic entropy estimators, and **iii)** other uncertainty estimators. The interval [a, b] denotes all integers from a to b, inclusively. The CI for SNNE reflects [6, 6] and is extended for readability.

In Figure 2, we compare plugin DSE estimates to SE^* (dotted lines), indicating that the canonical plugin DSE approach underestimates its quantity of interest for practical sample sizes. Such underestimation can be problematic for reliable UQ, because drawing a large number of samples from an LLM can be costly and time-consuming at scale.

Our proposed estimator improves the accuracy of discrete semantic entropy estimation. In Figure 2, we observed that $\hat{\mathbb{H}}_{Hybrid}$ underestimated its target quantity by less than canonical plugin DSE for a range of sample sizes. Detailed results, broken out by dataset, are shown in Appendix C (Figure 6), where similar patterns are exhibited. We perform a more granular comparison of semantic entropy estimator accuracy with $n = 10$ in Table 1, where $\hat{\mathbb{H}}_{Hybrid}$ consistently achieves the lowest MSE among explicit discrete semantic entropy estimators across four models and four datasets.

We improve upon prior work on incorrectness detection. LLM uncertainty estimation is often employed as a proxy for hallucination, confabulation, or incorrectness detection among LLM responses (Kuhn et al., 2023; Farquhar et al., 2024). Following Nikitin et al. (2024), we calculate pairwise AUROC win rates between models – for each model-dataset pair, a win is recorded in method i's favor over method j if the corresponding AUROC for method i is greater than that of method j (Fig-

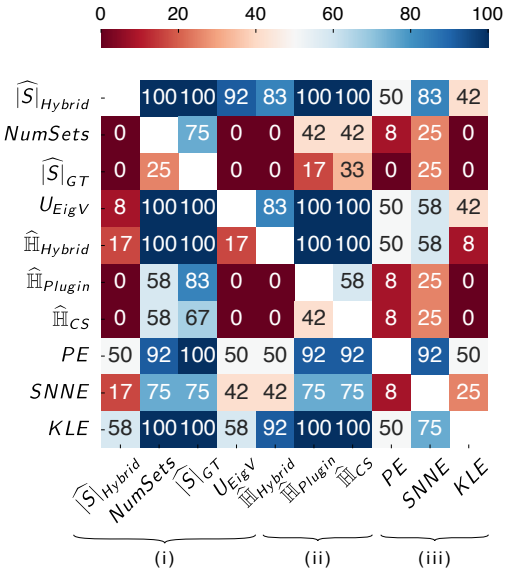

Figure 4: Heatmap illustrating the portion of model-dataset pairs, rounded to the nearest integer, for which a row's method achieved a larger AUROC point estimate than a column's method. Uncertainty methods are organized into three groups: **(i)** semantic alphabet size estimators, **(ii)** semantic entropy estimators, and **(iii)** other uncertainty methods. The hybrid discrete semantic entropy estimator consistently outperforms other DSE estimators, and the hybrid semantic alphabet size estimator consistently outperforms other alphabet size estimators.

ure 4). We characterize overall performance of each uncertainty method via Bradley-Terry latent strength scores. Figure 3A displays maximum likelihood estimates of the methods' strength scores from pairwise AUROC comparisons. Here, $\widehat{|S|}_{Hybrid}$ achieves the highest strength score, followed closely by KLE.

We highlight the uncertainty in both AUROC and Bradley-Terry estimation in Figure 3B (CIs about Bradley-Terry estimates, which may be slightly stricter than 95%) and Figure 3C (95% CIs around each method's rank by Bradley-Terry score). We observe that the hybrid estimators presented herein outperform estimators of the same type - i.e., $\widehat{|S|}_{Hybrid}$ achieves the highest strength score among semantic alphabet size estimators, and $\widehat{|S|}_{Hybrid}$ does the same among explicit semantic entropy estimators. Finally, we find that $\widehat{|S|}_{Hybrid}$ shares the [1, 2] CI with KLE (Figure 3C), a conceptually complex approach that may be more difficult to interpret. We do not statistically distinguish the incorrectness detection performance of $\widehat{|S|}_{Hybrid}$ from that of KLE, with both methods holding top performance, after accounting for uncertainty in our rank-generating procedure. Our results are consistent with Kuhn et al. (2023)'s observation that the number of observed semantic categories is itself "a reasonable uncertainty measure."

5 Related work

Herein, we briefly review related contributions not discussed elsewhere in this work.

Earlier LLM uncertainty quantification methods. Linguistic Confidence assesses whether the LLM articulates its confidence in its response (Mielke et al., 2022). The P(True) method of Kadavath et al. (2022) similarly relies on an LLM's self-perceived uncertainty, asking a language model if a response is "True" or "False," with the response probability of "False" ultimately reported. SelfCheckGPT draws $n + 1$ responses from an LLM and assesses the consistency of each sentence of the first response with each of the n following responses (Manakul et al., 2023). In prior work, such approaches have been consistently superseded by the methods otherwise described herein (Kuhn et al., 2023; Lin et al., 2024; Farquhar et al., 2024; Nikitin et al., 2024; Nguyen et al., 2025).

Uncertainty and internal representations. In the white-box setting, it may be desirable to ascertain LLM uncertainty from internal representations, rather than repeated sampling. Han et al. (2024b) and Han et al. (2024a), for example, approach this from an interpretability point of view, building on earlier work on semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024). Other analyses, however, contend that so-called "truthfulness encodings" are difficult to generalize (Orgad et al., 2025), warranting further examination.

Alphabet size estimation. The Good-Turing estimator discussed herein accounts for singletons in the sample (Chao & Shen, 2003). Alternative alphabet size estimators accounting for doubletons (i.e., semantic categories appearing exactly twice in the sample) (Chao, 1987) and tripletons (i.e., categories appearing exactly three times) (Lanumteang & Böhning, 2011) exist, but they are undefined when the number of doubletons is zero.

Unbiased entropy estimation. The entropy estimator of Montgomery-Smith & Schürmann (2014) is unbiased, but it is incompatible with the possibility of unobserved semantic equivalence classes. Otherwise, no unbiased estimator exists, to the best of our knowledge (Paninski, 2003).

6 Discussion

6.1 Conclusion

Several approaches for estimating LLM uncertainty have emerged in recent years, but UQ performance can be constrained absent white-box LLM access. Furthermore, the practicality of sampling-based UQ methods is limited by the computational costs associated with repeated text generation. We illustrate the importance of semantic alphabet size in LLM uncertainty estimation with small sample sizes. Adjusting discrete semantic entropy for sample coverage using our proposed semantic

alphabet size estimator results in more accurate semantic entropy estimation and improved LLM incorrectness classification, compared to other discrete semantic entropy estimators. Further, the aforementioned semantic alphabet size estimator achieves as good or better performance on incorrectness classification than all nine other black-box UQ methods considered in our QA experiments.

Though out of scope for the present study, we underscore that semantic alphabet size estimation - and $\widehat{|S|}_{Hybrid}$, in particular - may have broader application than merely entropy estimation. For instance, concurrent work by Li et al. (2025) applied a Good-Turing method to estimate the extent of LLMs' unexpressed factual knowledge (e.g., mathematical theorems and diseases).

6.2 Limitations

We highlight several limitations in the present work. First, computational constraints limit our ability to run the experiments herein with models larger than 9 billion parameters. That said, our analysis is extensive, spanning four distinct model families, four datasets, and ten uncertainty estimators. Though our results are empirical, and experimental results may vary across studies, we make efforts to express our statistical uncertainty, and our final rank-generating procedure aims to account for it.

Second, our work aims to improve upon plugin discrete semantic entropy, whose estimator is negatively biased. For this reason, we take semantic cluster labels as fixed, without ablating across alternative clustering strategies. End-to-end semantic entropy calculation involves several steps, however, and a biased estimator does not necessitate that the entire process results in a negatively biased estimate. For instance, a high false negative rate in the assignment of responses to semantic equivalence classes (e.g., due to minor lexical alterations inducing false negatives) may positively bias the final estimate (Grewal et al., 2024).

Finally, the uncertainty estimation methods considered herein do not directly measure factual inaccuracy. Instead, they may be better understood as indicators of semantic diversity. Of course, high semantic diversity may correspond to a high factual error rate, but there are scenarios wherein semantic diversity does not necessarily imply incorrectness (Ilia & Aziz, 2024), and, conversely, models can be confidently wrong.

6.3 Reproducibility

We outline our major experimental settings in Section 4.1, as well as evaluation metrics and baseline methods in Section 4.2. In Appendix A, we elaborate extensively on additional implementation details needed to reproduce our results, including algorithms, models, and prompt templates. Interested readers will also find a brief overview of implementation variations found in other work in Section E. Finally, we offer our code in supplementary materials, with hope to support scientific accessibility and reproducibility in future work.

6.4 LLM Usage

During this project, LLMs were employed at times for coding assistance.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*, 2024.

András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

Georgij P Basharin. On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory of Probability & Its Applications*, 4(3):333–336, 1959.

Ralph Allan Bradley and Milton E Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Francois Caron and Arnaud Doucet. Efficient bayesian inference for generalized bradley–terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

Anne Chao. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics*, pp. 783–791, 1987.

Anne Chao and Tsung-Jen Shen. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10:429–443, 2003.

Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, and Alex Hall. Pydantic, June 2024. URL https://github.com/pydantic/pydantic.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges Workshop*, pp. 177–190. Springer, 2005.

Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, pp. 837–845, 1988.

Sebastian Farquhar, Joel Kossen, Lukas Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Ronald A Fisher, A Steven Corbet, and Carrington B Williams. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *The Journal of Animal Ecology*, pp. 42–58, 1943.

Edward B Fowlkes and Colin L Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

Chao Gao, Yandi Shen, and Anderson Y Zhang. Uncertainty quantification in the Bradley–Terry–Luce model. *Information and Inference: A Journal of the IMA*, 12(2):1073–1140, 2023.

Irving J Good. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, 40(3-4):237–264, 1953.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

Yashvir S Grewal, Edwin V Bonilla, and Thang D Bui. Improving Uncertainty Quantification in Large Language Models via Semantic Embeddings. *arXiv preprint arXiv:2410.22685*, 2024.

Jiatong Han, Jannik Kossen, Muhammed Razzak, and Yarin Gal. Semantic Entropy Neurons: Encoding Semantic Uncertainty in the Latent Space of LLMs. In *NeurIPS 2024 - Workshop on Foundation Model Interventions (MINT)*, 2024a.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024b.

Bernard Harris. *The Statistical Estimation of Entropy in the Non-Parametric Case*. University of Wisconsin-Madison, Mathematics Research Center, 1975.

Md Mehedi Hasan, Moloud Abdar, Abbas Khosravi, Uwe Aickelin, Pietro Lio, Ibrahim Hossain, Ashikur Rahman, and Saeid Nahavandi. Survey on Leveraging Uncertainty Estimation Toward Trustworthy Deep Neural Networks: The Case of Reject Option and Post-Training Processing. *ACM Computing Surveys*, 57(9):1–35, 2025.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*, 2021.

David R Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32 (1):384–406, 2004.

Evgenia Ilia and Wilker Aziz. Variability Need Not Imply Error: The Case of Adequate but Semantically Distinct Responses. *arXiv preprint arXiv:2412.15683*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

George Kingsley Zipf. *Selected studies of the principle of relative frequency in language*. Harvard University Press, 1932.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

Krisana Lanumteang and Dankmar Böhning. An extension of Chao's estimator of population size based on the first three capture frequency counts. *Computational Statistics & Data Analysis*, 55 (7):2302–2311, 2011.

Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023.

Xiang Li, Jiayi Xin, Qi Long, and Weijie J Su. Evaluating the Unseen Capabilities: How Many Theorems Do LLMs Know? *arXiv preprint arXiv:2506.02058*, 2025.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Chin-Yew Lin and Franz Josef Och. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 605–612, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1219032. URL https://aclanthology.org/P04-1077/.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research*, 2024, 2024. ISSN 2835-8856.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. *arXiv preprint arXiv:2503.15850*, 2025.

Andrey Malinin and Mark Gales. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Henry B Mann and Donald R Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, pp. 50–60, 1947.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50/.

Giorgia Minello, Luca Rossi, and Andrea Torsello. On the von Neumann entropy of graphs. *Journal of Complex Networks*, 7(4):491–514, 2019.

Stephen Montgomery-Smith and Thomas Schürmann. Unbiased estimators for entropy and class number. *arXiv preprint arXiv:1410.5002*, 2014.

Kevin P. Murphy. Classification with the "reject" option. In *Probabilistic Machine Learning: An Introduction*, pp. 166. The MIT Press, 2022.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond Semantic Entropy: Boosting LLM Uncertainty Quantification with Pairwise Semantic Similarity. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4530–4540, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. volume 37, pp. 8901–8929, 2024.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, 2003.

Dénes Petz. Entropy, von Neumann and the von Neumann Entropy: Dedicated to the memory of Alfred Wehrl. In *John von Neumann and the Foundations of Quantum Physics*, pp. 83–96. Springer, 2001.

Assaf Pinchas, Irad Ben-Gal, and Amichai Painsky. A Comparative Analysis of Discrete Entropy Estimators for Large-Alphabet Problems. *Entropy*, 26(5):369, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124.

Claude Elwood Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. Distinguishing Ignorance from Error in LLM Hallucinations. *arXiv preprint arXiv:2410.22071*, 2024.

Xu Sun and Weichao Xu. Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *IEEE Signal Processing Letters*, 21 (11):1389–1393, 2014.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*, 2024.

Werner Ulrich, Marcin Ollik, and Karl Inne Ugland. A meta-analysis of species–abundance distributions. *Oikos*, 119(7):1149–1155, 2010.

Vincent Q Vu, Bin Yu, and Robert E Kass. Coverage-adjusted entropy estimation. *Statistics in Medicine*, 26(21):4039–4060, 2007.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259.

Ernst Zermelo. Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, 1929.

A Additional implementation details

A.1 Text generation

Throughout this work, we generate a maximum of 100 tokens for each response, with early stopping enabled if the EOS token is generated. All otherwise unspecified hyperparameters are set to their defaults (e.g., we do not invoke top-k sampling, top-p sampling, beam search, repetition penalty, etc.).

For incorrectness classification, we follow prior work that elicits concise responses by pre-pending QA queries with the following "pre-prompt" (Farquhar et al., 2024; Nikitin et al., 2024):

Pre-Prompt for Single-Sentence QA

Answer the following question in a single brief but complete sentence.

A.2 Synthetic data generation

Plethora Of accepTable cATegOries (POTATO) is a small synthetic dataset wherein every question is, in principle, answerable with a one-word or few-word response, and each question has more than one correct answer. We repeatedly invoke the following prompt against OpenAI's *GPT-4-turbo*:

Prompt for POTATO Question Generation

Using the available function, generate 10 questions from diverse topic areas. Each question should request only a single answer, but there should be exactly [NUM_ANSWERS] possible semantically distinct correct answer(s) to the question. For example, 'Name a continent on Earth' has seven possible correct answers, because Earth has seven continents, but 'Name all the continents on Earth' only has one possible semantically distinct correct answer (a list of all seven continents).

	$\lvert S\rvert > 1$			$\lvert S\rvert \geq 1$		
Classifier	FMI	NMI	PA	FMI	NMI	PA
LLM	**0.84**	**0.72**	0.84	**0.86**	0.63	0.83
NLI-NQ	0.84	0.64	**0.85**	0.85	**0.63**	**0.83**

Table 2: Evaluating semantic classification methods against human-annotated ground truth. Results are averaged across questions from the POTATO dataset, either excluding or including questions for which all responses were semantically equivalent ($|S| > 1$ and $|S| \geq 1$, respectively), according to the human rater. Metrics considered are the Fowlkes–Mallows index (FMI), normalized mutual information (NMI), and pairwise agreement (PA), assessed between the semantic equivalence classes assigned by the provided method and the human-annotated ground truth. Values are rounded to two decimal places. The best-performing BEC method by each metric (unrounded) is in bold. Instances where a Semantic Embedding method outperforms a BEC method are underlined. Responses are generated by *GPT-4o-mini*.

Above, [NUM_ANSWERS] is an integer between 1 and 50, and the available function is a Pydantic object enforcing structured generation (Colvin et al., 2024). A human annotator assessed the number of semantic categories for possible correct answers, discarded those with only one possible correct semantic category, and removed duplicate questions. The resulting dataset has 131 unique questions. We generate 100 responses to each question from the POTATO dataset using *GPT-4o-mini*. A human annotator manually assigned the responses to semantic categories.

One query ("Name a piece of classical music composed by Ludwig van Beethoven.") has a particularly large number of possible correct semantic categories, owing to Beethoven's prolific output of several hundred compositions. Figure 5A suggests that the distribution of semantic alphabet sizes for model-query pairs is generally dominated by the distribution of the numbers of possible correct semantic categories for those queries. On a per-query basis, however, these quantities exhibit low correlation (Figure 5B).

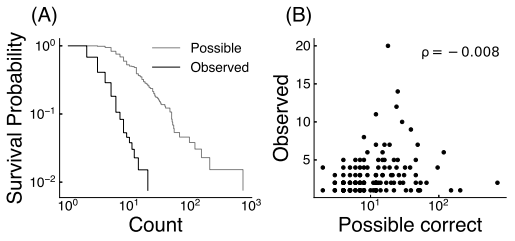

Figure 5: **(A)** Empirical survival function of the number of possible correct semantic categories and the number of observed semantic categories in the responses generated by GPT-4o-mini, according to a human annotator. **(B)** Scattergram of the number of observed semantic categories ("Observed"), according to the human annotator, against the number of possible correct semantic categories ("Possible"). One question is excluded ("Identify a programming language designed by Microsoft."), because the total number of possible correct semantic categories was not known by the authors of this work.

A.3 Bidirectional entailment clustering

Semantic classification. The semantic clustering step of semantic entropy calculation is performed using a Bidirectional Entailment Clustering (BEC), which classifies pairs of passages (d_i, d_j) bidirectionally as "entailment," "neutral," or "contradiction," and uses these labels to greedily assign passages to semantic equivalence classes. Like previous work, our implementation focuses on strict entailment, where both unidirectional relations must be considered "entailment" or equivalent (i.e., not "neutral"). We refer the reader to Kuhn et al. (2023); Farquhar et al. (2024) for further detail on the BEC algorithm. The classifier used varies by implementation, either an LLM- or NLI-based method:

Large Language Model (LLM). Farquhar et al. (2024) classify passage pairs (T1, T2) by invoking OpenAI's *GPT-3.5* endpoint with the following prompt:

Prompt for Entailment Classification

We are evaluating answers to the question {question}
Here are two possible answers:
Possible Answer 1: {T1}
Possible Answer 2: {T2}
Does Possible Answer 1 semantically entail Possible Answer 2? Respond with entailment, contradiction, or neutral.

Our implementation invokes OpenAI's *GPT-4o-mini* model, due to the sunsetting of *GPT-3.5*.

Natural Language Inference (NLI). The BEC approach of Kuhn et al. (2023) classifies (d_i, d_j) using a *DeBERTa* model fine-tuned for NLI tasks.[2] Our implementation, used throughout this work, updates the classifier to a fine-tuned NLI model based on Microsoft's newer *DeBERTaV3* (He et al., 2021).[3] The NLI classification can be performed with or without the source questions (NLI-Q and NLI-NQ, respectively); in the former, the query is prepended to both d_i and d_j before passing through the NLI model.

Comparison. To compare the alignment between each clustering method and the human-annotated ground-truth, we measure the Fowlkes–Mallows index (FMI) (Fowlkes & Mallows, 1983) and normalized mutual information (NMI) (Danon et al., 2005; Lancichinetti et al., 2009). We also consider pairwise agreement (PA), wherein we iterate over pairs of passages, classify each pair as "entailment" or "contradiction," based on the clustering results, and report for each method the portion of pairs whose entailment label agrees with that of the human annotator (Farquhar et al., 2024). Across metrics, we observe that the NLI method without the inclusion of questions performs roughly similarly to LLM (Table 2). For its comparable performance while limiting costs, we use NLI-NQ throughout this work.

A.4 LLM-AS-JUDGE

Prior work labeled an LLM's response as correct if the ROUGE-L score (Lin, 2004; Lin & Och, 2004) between the response and a reference answer was above 0.3 (Kuhn et al., 2023). Appropriate thresholds may vary by model and dataset - for instance, the reference answers in HotPotQA appear rather brief, and models can vary in their verbosity, without it necessarily impacting the correctness of their responses. Additionally, such methods may not capture semantically equivalent, but lexically distinct, paraphrasings (e.g., using an acronym).

Instead, Lin et al. (2024) prompt a commercial LLM to provide a numerical rating of consistency between LLM responses and reference answers, with ratings above 70 indicating an accurate response. They observe, however, that a small portion of LLM judgements do not have an easily-parseable rating. For this reason, we modify their prompt to request ratings within XML-style tags and include an in-context example:

Prompt for Consistency-Based Judge

Rate the level of consistency between the answer to the question and the reference answer, from 0 to 100. Output the float rating inside <rating>< /rating> tags.
Here is an example output:
Question: What is the capital of France? Reference: Paris is the capital of France. Answer: The capital of France is Paris. <rating>100< /rating>
Now rate the following:
Question: {question}
Reference: {groundtruth}
Answer: {pred}

[2] https://huggingface.co/microsoft/deberta-large-mnli
[3] https://huggingface.co/cross-encoder/nli-deberta-v3-base

When a query has multiple reference answers (e.g., those in the BioASQ), we invoke the above prompt to compare the LLM response to each reference answer, then report the maximum LLM-as-judge rating. Our choice of commercial model for LLM-as-judge is OpenAI's *GPT-4o-mini*.

A.5 Uncertainty estimators

Predictive entropy. The so-called "total uncertainty" associated with a prediction is the information entropy of the corresponding predictive posterior distribution (Malinin & Gales, 2021). In UQ for natural language generation, an analogue is predictive entropy (PE), the entropy of the distribution of generable sequences by an LLM in response to a prompt (Kuhn et al., 2023; Lin et al., 2024; Farquhar et al., 2024). Since not all generable sequences are typically known, PE is calculated using the above plugin method to estimate the entropy of a prompt-model pair's "answer distribution" (Kadavath et al., 2022). A shortcoming of this approach is that it treats all lexically distinct sequences as unique elements of an alphabet, even if they are semantically similar.

SNNE. SNNE aims to account for both "intra-and inter-cluster similarity." Provided a scale factor τ and a similarity function f for assessing passage pairs (d_i, d_j), SNNE is defined as:

$$SNNE(q) = -\frac{1}{n}\sum_{i=1}^{n}\log\sum_{j=1}^{n}\exp\left(\frac{f(d_i,d_j)}{\tau}\right) \tag{11}$$

(Nguyen et al., 2025). Our implementation uses ROUGE-L for f and scale factor $\tau = 1$, which the authors report as best-performing (Nguyen et al., 2025).

KLE. Like U_{EigV}, KLE constructs a graph over LLM responses to a query, where nodes are responses and edge weights are prescribed by the results of an NLI model. Instead of using the model's categorical scores, however, KLE's semantic graph has weights $w_{i,j} = g(d_i, d_j) + g(d_j, d_i)$, where g is 1 if the NLI model predicts "entailment" for the provided response pair, 0.5 if it predicts "neutral," and 0 otherwise. The standard graph Laplacian (i.e., the difference between the degree matrix and the weight matrix) is then taken over the resultant semantic graph (Nikitin et al., 2024).

KLE admits a choice of kernel to apply to the Laplacian, resulting in a "density" matrix K. Given the eigenvalues $(\lambda_1 < \cdots < \lambda_n)$ of K, the von Neumann entropy is calculated:

$$VNE(K) = \sum_{i=1}^{n}\lambda_i \log \lambda_i. \tag{12}$$

Although the concept of von Neumann entropy has its roots in quantum physics (Petz, 2001), it has also been used more recently in network science literature to quantify graph complexity (Minello et al., 2019).

Nikitin et al. (2024) report the heat kernel as best-performing overall:

$$K_{heat} = e^{-tL}. \tag{13}$$

Our implementation of KLE invokes K_{heat} with hyperparameter $t = 0.3$, which the authors consider a "reasonable default" that outperforms prior approaches without additional hyperparameter optimization (Nikitin et al., 2024).

A.6 Bradley-Terry confidence intervals

Gao et al. (2023) offers a procedure for ascertaining CIs about ranks of Bradley-Terry strength estimates, which we cursorily summarize as follows: Assuming a Bradley-Terry model, let $\beta = \{\beta_1, \ldots, \beta_m\}$ be the true latent strengths of m methods, and let $\widehat{\beta} = \{\widehat{\beta}_1, \ldots, \widehat{\beta}_m\}$ be MLEs of the same. Given a provided level α and a "target" method i, we may establish a $(1-\alpha)$ CI about $\widehat{\beta}_i$, as well as "slightly more conservative" intervals about the remaining strength estimates. Let n_1 be the number of resultant intervals whose lower bounds are greater than the upper bound of method i's. Similarly, let n_2 be the number of intervals whose upper bounds are less than the lower bound of method i's. The true rank of method i's Bradley-Terry strength is in the integer interval $[n_1 + 1, n - n_2]$ with approximate probability $1 - \alpha$. Because Proposition 4.1 of Gao et al. (2023) only holds for fixed number of methods/agents/players, we illustrate the "conservative" CIs in Figure 3C. We repeat this procedure for each UQ method to establish CIs for all strength ranks.

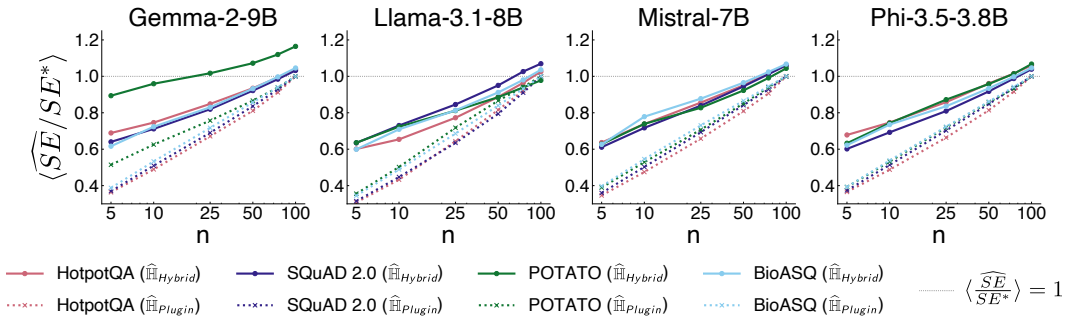

Figure 6: Ratio of semantic entropy estimator with varying sample size ($n = 5, 10, 25, 50, 75, 100$) to white-box semantic entropy with $n = 100$ (denoted SE^*), with values below 1 suggesting underestimation (dotted grey line). Estimators shown are the plugin estimator of canonical discrete semantic entropy (dotted rose, indigo, green, and cyan lines) and the presented semantic entropy estimator of Equation 10 (solid rose, indigo, green, and cyan lines). Results are averaged across queries for each dataset. Log scale is used on the x-axis to highlight differences between estimators with smaller sample sizes. Instances with only one observed semantic category at $n = 100$, resulting in a denominator of 0, are ignored.

We refer interested readers to Gao et al. (2023) for further details.

The algorithm we use to estimate strength scores employs a regularization parameter, which we take as $a = 0.1$ in the main body of the paper. Table 3, which repeats the analysis of Figure 3C for varying a, indicates that the rank order of our Bradley-Terry results is not sensitive to this choice.

	a			
Method	0	0.01	0.1	1
$\widehat{\lvert S\rvert}_{Hybrid}$	[1, 2]	[1, 2]	[1, 2]	[1, 2]
NumSets	[7, 10]	[7, 10]	[7, 10]	[7, 10]
Good-Turing	[7, 10]	[7, 10]	[7, 10]	[7, 10]
U_{EigV}	[3, 4]	[3, 4]	[3, 4]	[3, 4]
$\widehat{\mathbb{H}}_{Plugin}$	[7, 10]	[7, 10]	[7, 10]	[7, 10]
$\widehat{\mathbb{H}}_{CS}$	[7, 10]	[7, 10]	[7, 10]	[7, 10]
$\widehat{\mathbb{H}}_{Hybrid}$	[4, 5]	[4, 5]	[4, 5]	[4, 5]
PE	[3, 5]	[3, 5]	[3, 5]	[3, 5]
SNNE	[6, 6]	[6, 6]	[6, 6]	[6, 6]
KLE	[1, 2]	[1, 2]	[1, 2]	[1, 2]

Table 3: Employing distinct values of regularization parameter a, including zero regularization, we compute MLE Bradley-Terry latent strength scores and establish 95% CIs of each method's latent strength rank via the methods of Gao et al. (2023). Results are identical for all a values considered.

B THE UNDER-SAMPLED REGIME

Consider a simple model of semantic category-abundance, where the probability that a model parameterized by θ generates a response belonging to the r^{th}-most prevalent semantic category s_r in response to a query q follows a Zipfian distribution, similar to some models of species-abundance (Ulrich et al., 2010):

$$p(s_r|q, \theta) = \frac{1}{rH_{|S|}}, \tag{14}$$

where H_j is the j^{th} harmonic number (Kingsley Zipf, 1932). If the expected number of occurrences of s_r in a sample of n responses is less than 1, then $|S| > \frac{n}{H_{|S|}}$. For sample size $n = 10$, we expect at least one semantic category to be unobserved with just $|S| > 4$.

C ADDITIONAL UNDERESTIMATION RESULTS

Detailed results for SE underestimation, broken out across four datasets, are shown in Figure 6. Results are qualitatively similar to those in Figure 2, with $\hat{\mathbb{H}}_{Hybrid}$ consistently underestimating SE^* less than plugin DSE for practical sample sizes.

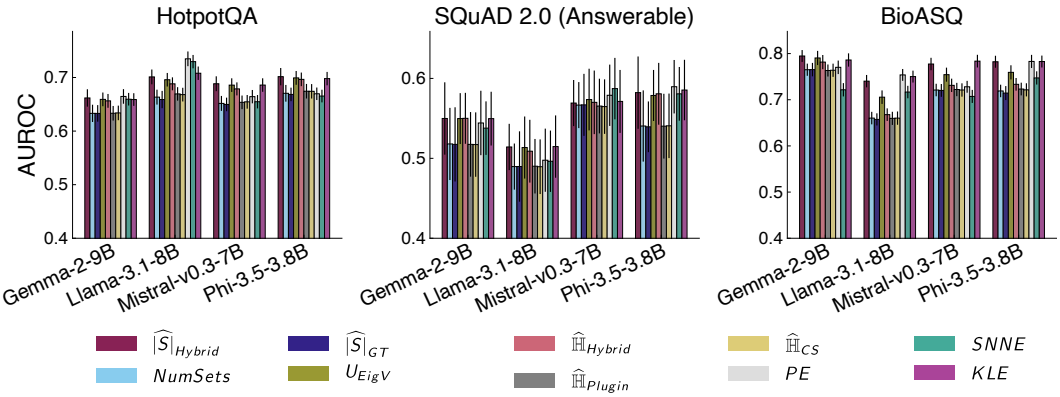

Figure 7: Performance of incorrectness classification on single-answer QA for uncertainty measures using $n = 10$ samples. "Plugin," "CS-GT," and "CS-H" are estimators of semantic entropy. "NumSets," "U_{EigV}," and "$\widehat{|S|}_{Hybrid}$" are estimators of semantic alphabet size. Error bars reflect 95% CIs about the AUROC.

D Additional incorrectness detection results

Detailed results for incorrectness classification across three datasets, four LLMs, and ten UQ methods are shown in Figure 7. Error bars are 95% CIs about empirical AUROC values, calculated via DeLong's method (DeLong et al., 1988; Sun & Xu, 2014).

E Implementation variations

The relevant LLM UQ literature is not always consistent in experimental setup. For instance, the plurality of considered works use $n = 10$ sampled responses per query (Kuhn et al., 2023; Farquhar et al., 2024; Nikitin et al., 2024; Nguyen et al., 2025), but Lin et al. (2024) use $n = 20$. Some works vary prompting templates across datasets (e.g., zero-shot vs. multi-shot prompting) (Kuhn et al., 2023; Lin et al., 2024). A variety of methods for automated LLM incorrectness evaluation have been employed, including thresholded ROUGE-L (Kuhn et al., 2023), binary LLM-as-judge (Farquhar et al., 2024), thresholded consistency-based LLM-as-judge Lin et al. (2024), and BERTScore (Nguyen et al., 2025). Although it is out of this work's scope to exhaustively ablate all such choices, we attempt to be explicit in describing our selections to support future work.