GLIB-SE: LOW-BUDGET SEMANTIC UNCERTAINTY QUANTIFICATION VIA RESIDUAL SEMANTIC MASS MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable deployment of large language models (LLMs) requires uncertainty estimates that remain informative under tight inference budgets. Existing semantic uncertainty estimators often rely on a limited set of sampled responses to characterize the semantic distribution, which can understate uncertainty when semantically plausible but unobserved alternatives still carry non-negligible probability mass. We propose GLIB-SE, which augments observed semantic classes with a Ghost Cluster to represent residual semantic mass. Semantic counts and energy derived from generated-token scores parameterize an augmented Dirichlet proposal, while response likelihoods constrain Monte Carlo entropy scoring. By incorporating model-internal information into the residual prior, GLIB-SE can distinguish response sets with identical semantic count distributions. We evaluate GLIB-SE using ten scoring responses per input generated by nine language models across six datasets. GLIB-SE achieves higher mean AUROC than likelihood-weighted and count-based semantic entropy in all 18 dataset–model combinations and the highest mean among seven methods in 16 combinations. Support-stratified analysis shows that most of the gain over count-based entropy arises when all sampled answers fall into one observed semantic class, where counts alone provide no ranking information. These findings support combining residual semantic support with model-internal signals to improve error detection under limited sampling.

1 INTRODUCTION

Large language models (LLMs) perform strongly across question answering, summarization, and reasoning, but fluent answers can still contain factual errors. Deployment requires uncertainty estimates that remain useful under limited generation budgets to guide abstention, retrieval, and human review (Ji et al., 2023; Zhang et al., 2023; Kadavath et al., 2022; Murphy, 2022; Yadkori et al., 2024).

Semantic Entropy (SE) groups sampled responses by meaning to estimate semantic uncertainty (Farquhar et al., 2024; Kuhn et al., 2023). A finite response set, however, can miss plausible semantic alternatives. Normalizing over observed classes implicitly assigns zero mass to classes that have not appeared, introducing a closed-world bias under finite sampling. When sampled answers agree, the estimate depends particularly strongly on whether sampling has covered the relevant meanings. An estimator therefore needs to account for both the distribution of observed answers and semantic support beyond the sample.

We address this problem through *residual semantic mass*. GLIB-SE introduces a **Ghost Cluster** that aggregates unobserved semantic classes into one residual component. This construction follows the unseen-mass problem in finite-sample distribution estimation: a class absent from the sample can still have positive generation probability (Good, 1953). Semantic alphabet-size estimation studies this problem through coverage and the number of classes (McCabe et al., 2026). We extend the Dirichlet representation used in Bayesian semantic estimation (Ciosek et al., 2025) with a residual dimension, reserving mass for unobserved support through a positive prior concentration.

The central question is how much prior strength to assign to this component. We use generated-token energy to map model-internal confidence information to a query-dependent Ghost concentration.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Semantic counts describe variation among observed answers, while energy adjusts the residual component using information beyond those counts. Likelihood constraints connect draws from the augmented distribution to response likelihoods, and their mean entropy gives the final score.

Consider two response sets with the same semantic histogram but different token scores. A count-only estimator gives them the same result, whereas GLIB-SE can distinguish them through their residual priors. This combines finite semantic samples and model-internal signals within one estimator. Comparisons with published baselines, dataset–model analyses, and fixed-budget component experiments assess how this construction affects error ranking.

We focus on semantic uncertainty with only ten sampled answers. The output can prioritize retrieval, abstention, or human review. The calculation also returns entropy variance for use in subsequent adaptive sampling.

Our contributions are:

- **Residual semantic mass modeling:** We augment observed semantic classes with a single Ghost Cluster to construct a coarse-grained semantic distribution that includes unobserved support.
- **Energy-conditioned semantic uncertainty:** We use generated-token energy to adjust a Dirichlet residual prior and compute semantic uncertainty subject to likelihood constraints.
- **Empirical analysis at a small budget:** We compare six published baselines across six datasets, 18 dataset–model combinations, and 54 runs, then examine the roles of priors, likelihoods, and semantic counts through component controls at the same budget.

2 Related Work

From token confidence to semantic uncertainty. Token likelihood can change between equivalent phrasings. G-NLL estimates uncertainty from the negative log-likelihood of a single greedily decoded response (Aichberger et al., 2026). Length-normalized predictive entropy (LNPE) summarizes response likelihoods without grouping answers; Semantic Entropy (SE) instead groups equivalent responses and aggregates their likelihoods (Kuhn et al., 2023; Farquhar et al., 2024). Its count-based counterpart, Discrete Semantic Entropy (DSE), replaces likelihood weights with observed class frequencies (Farquhar et al., 2024). Our SE baseline uses length-normalized response likelihoods, with its exact definition given in Appendix A.3. Semantic density and black-box confidence estimators also use relationships among sampled answers (Qiu & Miikkulainen, 2024; Lin et al., 2024). These approaches motivate a distinction between uncertainty within observed support and uncertainty about support that has not been sampled.

Unseen semantic support. Ciosek et al. study efficient Bayesian estimation of semantic entropy from sampled responses (Ciosek et al., 2025). UEigV is a spectral graph statistic introduced by Lin et al. (2024). McCabe et al. combine it with Good–Turing coverage to estimate semantic alphabet size; their hybrid estimate supplies a Chao–Shen-style entropy correction (McCabe et al., 2026; Chao & Shen, 2003). The alphabet-size scores can also be used directly for error detection. We compare these outputs separately. GLIB-SE addresses the related missing-support problem using a finite residual bucket whose prior depends on model-internal energy. It models the allocation of probability between observed support and the residual component.

Energy and continuous semantic relationships. Internal model states provide confidence signals beyond output agreement (Azaria & Mitchell, 2023). Semantic Energy combines logit information with semantic grouping (Ma et al., 2025), while Kernel Language Entropy retains continuous relationships between responses (Nikitin et al., 2024). In GLIB-SE, energy determines the residual prior rather than class assignments. Appendix A.5 examines local direct-energy and energy-plus-partition controls to assess how different uses of the signal affect estimation.

Consistency and generation budgets. SelfCheckGPT detects inconsistency among sampled outputs (Manakul et al., 2023); consistent errors remain difficult for agreement-based signals. Wang et al. use Bayesian sequential decisions to determine whether to retrieve additional external evidence for hallucination detection (Wang et al., 2023). Adaptive semantic-entropy methods decide whether

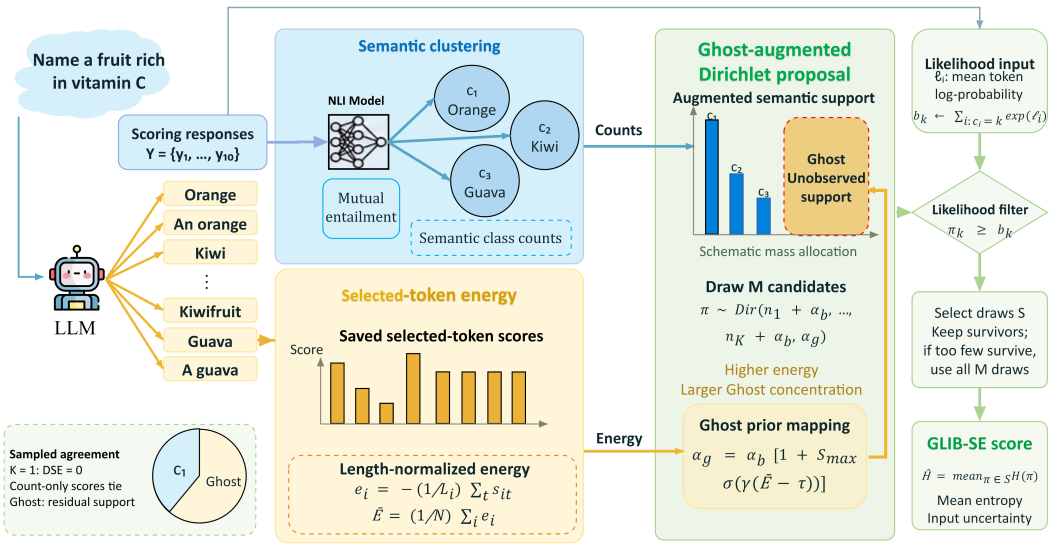


Figure 1: GLIB-SE computes an uncertainty score from $N = 10$ scoring responses. Semantic counts and selected-token energy parameterize a Ghost-augmented Dirichlet proposal. Response likelihoods constrain the draws, with finite-draw fallback before mean-entropy scoring. AUROC evaluates scores across inputs against the correctness of separately stored target answers.

to sample further responses (Sun et al., 2026). We hold the response budget fixed for each input and compare how uncertainty estimators use the same answers.

3 Methodology

GLIB-SE separates semantic grouping from uncertainty estimation. Grouping determines which sampled answers express the same meaning. The estimator combines observed class counts with a query-dependent reserve for meanings absent from the sample. The output is a semantic uncertainty score for the response set.

3.1 Observed support and the estimation target

For input $\mathbf{x}$, let $Y = \{\mathbf{y}_i\}_{i=1}^{N}$ be sampled scoring responses. A grouping function assigns each answer to one of K observed classes with counts $n_1, \ldots, n_K$. The count-based estimator is

$$\widehat{H}_{\mathrm{DSE}}(\mathbf{x}) = -\sum_{k=1}^{K} \frac{n_k}{N} \log \frac{n_k}{N}. \tag{1}$$

3.2 A residual component with an energy-dependent prior

The Ghost Cluster c_{g} represents aggregate mass outside the observed classes. Given base concentration $\alpha_{\mathrm{b}} > 0$ and residual concentration $\alpha_{\mathrm{g}} > 0$, we form the Dirichlet proposal

$$\boldsymbol{\pi} \mid Y \sim \mathrm{Dir}(n_1 + \alpha_{\mathrm{b}}, \ldots, n_K + \alpha_{\mathrm{b}}, \alpha_{\mathrm{g}}). \tag{2}$$

Before applying likelihood constraints, the mean residual mass is

$$\mathbb{E}[\pi_{\mathrm{g}} \mid Y] = \frac{\alpha_{\mathrm{g}}}{N + K\alpha_{\mathrm{b}} + \alpha_{\mathrm{g}}}. \tag{3}$$

For fixed concentrations, the reserve decreases as additional answers are observed. Taking $\alpha_{\mathrm{g}} \to 0$ removes the residual component and leaves a smoothed distribution over observed classes.

Energy definition. Following the negative-logit construction of Ma et al. (2025), we compute a length-normalized selected-token energy from saved generation scores. Let s_{it} denote the saved score of the generated token at position t in response i, and let L_i be the length of its saved token-score sequence. We use

$$e_i = -\frac{1}{L_i}\sum_{t=1}^{L_i} s_{it}, \qquad \bar{E}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{4}$$

This energy differs from token surprisal and vocabulary-level log-partition energy. The mapping

$$\alpha_{\mathrm{g}}(\bar{E}) = \alpha_{\mathrm{b}}\left[1 + S_{\max}\sigma\big(\gamma(\bar{E} - \tau)\big)\right] \tag{5}$$

uses sensitivity γ, offset τ, and maximum scale $S_{\max}$. Higher energy increases the residual concentration for a fixed configuration. The offset τ is automatically calibrated from each run's unlabeled response-energy distribution using a common rule; the other main hyperparameters are shared across runs. Configuration and numerical details appear in Appendix A.1.

Likelihood-constrained scoring. We use exponentiated mean token log-probabilities as length-normalized compatibility weights. Summing these weights within each observed class gives an empirical threshold for proposal draws. The filter retains vectors from Eq. (2) that allocate at least this threshold to each observed class, linking the count-and-energy proposal to response likelihoods. If too few draws satisfy the thresholds, the implementation uses the original proposal draws. For the retained set $\mathcal{S}$, the score and its variance are

$$\widehat{H}_{\mathrm{GLIB}} = \frac{1}{|\mathcal{S}|}\sum_{\boldsymbol{\pi}\in\mathcal{S}} H(\boldsymbol{\pi}), \qquad \widehat{V}_H = \mathrm{Var}_{\boldsymbol{\pi}\in\mathcal{S}}[H(\boldsymbol{\pi})]. \tag{6}$$

Here H sums over the K observed components and the Ghost component. The score averages entropy, rather than taking entropy of the mean vector. Constraints and finite-draw fallback distinguish this score from the unconstrained closed-form Dirichlet expectation; Appendix A.1 specifies both rules.

Interpretation of the Ghost component. Let r be the true mass of unobserved meanings, and let q and u be normalized distributions within observed and unobserved support. Their entropy decomposes as

$$H_{\mathrm{full}} = h(r) + (1-r)H(q) + rH(u), \tag{7}$$

where h is binary entropy. A single residual bucket omits the within-tail term $rH(u)$. Reserving mass prevents observed support from being exhaustive by construction.

Response budget and optional stopping. The main comparison uses all ten scoring responses. An adaptive implementation can repeat the calculation after each new response and stop once a minimum prefix has been collected and $\widehat{V}_H$ falls below a threshold.

4 Experiments

We evaluate three questions: how GLIB-SE compares with published uncertainty estimators at a small response budget; which response patterns account for its gains; and how the Ghost prior interacts with semantic counts and likelihood constraints.

4.1 Evaluation setup

We evaluate BioASQ (Tsatsaronis et al., 2015), CoQA (Reddy et al., 2019), GSM8K (Cobbe et al., 2021), NQ-Open (Kwiatkowski et al., 2019), SVAMP (Patel et al., 2021), and TriviaQA (Joshi et al., 2017) using nine models across 18 dataset–model combinations. Seeds 10, 20, and 30 give 54 runs. Each run contains 500 examples for CoQA, 300 for SVAMP, and 1,000 for the other datasets. For each input, ten sampled responses supply the semantic groups, energy, and likelihoods used to compute one uncertainty score. AUROC evaluates this score against the saved error label of a separately generated target answer. $N = 10$ specifies the scoring budget. Excluding six records with missing token-score inputs from all methods leaves 45,294 example-run records. Every run contains both target classes, with at least 49 minority-class examples. Appendix A.2 specifies generation and labeling details.

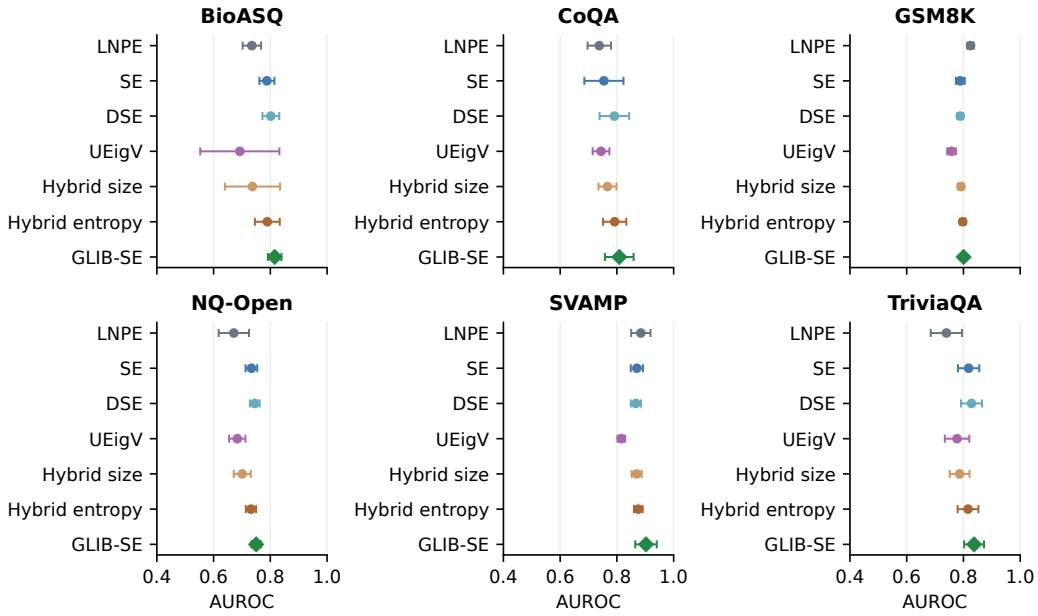


Figure 2: Seven-method comparison using ten scoring responses per input and the same 45,294 records. Points show dataset mean AUROC; bars show one sample standard deviation across runs. Paired difference intervals appear in Table 4.

The six published baselines are length-normalized predictive entropy (LNPE), Semantic Entropy (SE), Discrete Semantic Entropy (DSE), UEigV, Hybrid alphabet size, and Hybrid entropy. LNPE averages negative mean-token log-probabilities; SE aggregates length-normalized response likelihoods within semantic classes (Kuhn et al., 2023; Farquhar et al., 2024). DSE uses class frequencies (Farquhar et al., 2024). UEigV follows Lin et al. (2024), and the Hybrid estimators follow McCabe et al. (2026). Graph methods use local DeBERTa probabilities as specified in Appendix A.3. Local energy controls appear in Appendix A.5.

We estimate the macro-average of within-run AUROC differences over the fixed model and seed roster. For each dataset, 10,000 paired bootstrap resamples draw question/context clusters, sharing each cluster's multiplicity across all runs and methods. The 2.5th and 97.5th percentiles yield marginal 95% intervals. Model-level comparisons average over their three fixed seeds. Configuration and cluster definitions appear in Appendix A.1.

4.2 Performance across datasets and models

GLIB-SE exceeds both SE and DSE in all 18 model-level means (Table 1) and has the highest mean among the seven methods in 16 combinations. Its six dataset means range from 0.750 to 0.903. The dataset-level gains over SE range from 0.0116 to 0.0542, and those over DSE from 0.0049 to 0.0358; all twelve paired intervals are above zero (Table 4). Seed variation is reported in Table 3.

The comparator matters when interpreting the size of the gain. On CoQA, GLIB-SE gains 0.0542 over SE but 0.0172 over DSE. The stronger count-based baseline is consistent with response-likelihood weighting contributing to SE's lower performance in this setting. Gains over DSE are smaller on NQ-Open (0.0049) than on SVAMP (0.0358), motivating an analysis of where semantic counts lose ranking information. Relative to Hybrid entropy, the five non-GSM8K dataset intervals are above zero; the GSM8K difference is 0.0031 with interval $[-0.0018, 0.0079]$. The residual representation therefore complements the coverage-based correction of McCabe et al. (2026), with different ranking behavior across tasks.

The results are heterogeneous across comparators and datasets. GSM8K favors the likelihood signal: LNPE achieves 0.825 mean AUROC versus 0.801 for GLIB-SE, a difference of -0.0244 for

Table 1: Comparison with published baselines for every dataset–model combination. Values are mean AUROC over seeds 10, 20, and 30 (higher is better). All methods use the same 45,294 records across 54 runs. LNPE is length-normalized predictive entropy; SE uses length-normalized response likelihoods.

Dataset	Model	LNPE	SE	DSE	UEigV	Hybrid size	Hybrid entropy	GLIB-SE (ours)
BioASQ	Llama-3.1-70B-Instruct	0.738	0.816	0.833	0.775	0.809	0.833	**0.841**
	Qwen3-30B-A3B	0.744	0.759	0.759	0.741	0.760	0.768	**0.787**
	Qwen3-32B	0.749	0.805	0.810	0.788	0.799	0.821	**0.826**
	Seed-Coder-8B-Instruct	0.709	0.771	0.805	0.466	0.580	0.735	**0.807**
CoQA	Llama-3.1-70B-Instruct	0.746	0.678	0.756	0.748	0.735	0.763	**0.764**
	Qwen3-30B-A3B	0.730	0.764	0.777	0.736	0.786	0.787	**0.812**
	Qwen3-32B	0.738	0.821	0.841	0.748	0.780	0.827	**0.849**
GSM8K	Llama-3.1-70B-Instruct	**0.817**	0.799	0.794	0.765	0.787	0.799	0.804
	Qwen3-32B	**0.833**	0.779	0.783	0.751	0.796	0.796	0.797
NQ-Open	Llama-2-13B-chat-hf	0.607	0.723	0.735	0.697	0.712	0.731	**0.742**
	Llama-3.1-8B-Instruct	0.685	0.727	0.745	0.654	0.671	0.721	**0.749**
	Qwen3-8B	0.722	0.750	0.755	0.700	0.720	0.743	**0.759**
SVAMP	Llama-2-70B-chat-hf	0.859	0.854	0.855	0.811	0.856	0.863	**0.871**
	Llama-3.1-70B-Instruct	0.910	0.887	0.879	0.820	0.884	0.888	**0.934**
TriviaQA	Llama-2-13B-chat-hf	0.667	0.799	0.806	0.763	0.777	0.798	**0.818**
	Llama-3.1-8B-Instruct	0.772	0.829	0.845	0.791	0.788	0.829	**0.848**
	Qwen3-4B-Instruct	0.778	0.834	0.839	0.778	0.793	0.823	**0.847**
	Qwen3-8B	0.742	0.811	0.824	0.778	0.788	0.814	**0.837**

Table 2: Single-class response sets ($K = 1$). Percentages are per-run proportions averaged within each dataset: the share of all examples and the share of all errors.

Dataset	$K = 1$ (%)	Error share (%)	DSE	GLIB-SE	LNPE
BioASQ	41.4	16.3	0.500	0.610	0.640
CoQA	21.2	12.5	0.500	0.649	0.682
GSM8K	30.1	15.2	0.500	0.646	0.840
NQ-Open	23.0	13.1	0.500	0.554	0.551
SVAMP	65.0	17.9	0.500	0.750	0.752
TriviaQA	38.4	12.2	0.500	0.627	0.569

GLIB-SE with interval $[-0.0349, -0.0139]$. This occurs for both models: LNPE versus GLIB-SE is 0.817 versus 0.804 for Llama-3.1-70B-Instruct and 0.833 versus 0.797 for Qwen3-32B. Their answer accuracies are 30.9% and 34.1%, respectively. By contrast, GLIB-SE exceeds LNPE on the four factual-QA datasets and SVAMP, with intervals above zero. The contrasting arithmetic results motivate an analysis of their response patterns.

4.3 Sources of performance gains

We stratify inputs by the number of semantic classes among their ten scoring responses: $K = 1$, $K = 2$, and $K \geq 3$. Table 2 examines $K = 1$, where all scoring responses share one meaning, using the original scores and common evaluation set. GLIB-SE retains token-level information through the residual dimension and reaches dataset mean AUROCs of 0.554–0.750 for target-error detection. These inputs account for approximately 12–18% of each dataset's target errors when the per-run proportions are averaged. Sampled agreement therefore leaves a material set of target errors to distinguish, and the augmented estimator supplies ranking information within that set.

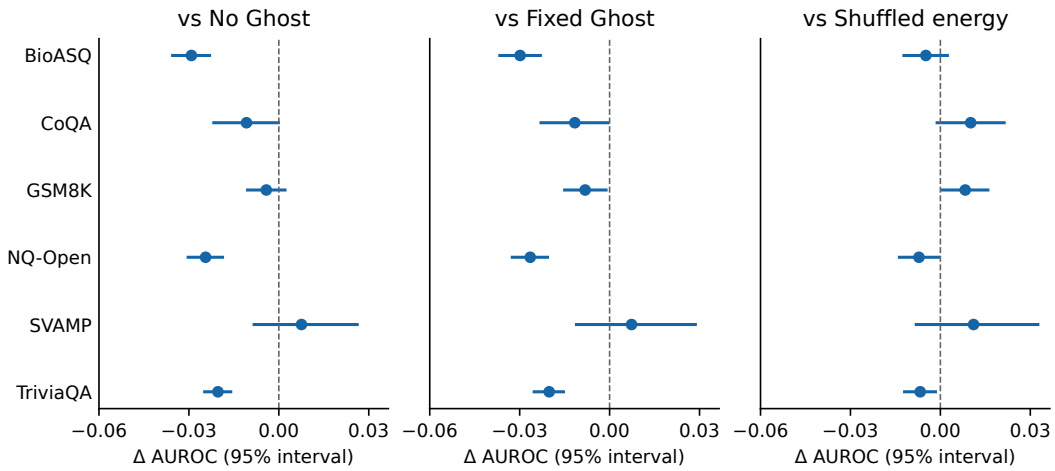


Figure 3: Separate component diagnostics at a fixed ten-response scoring budget. Points show Energy Ghost minus each control; bars are 95% question/context-cluster intervals. The arms are internally matched under the diagnostic configuration ($\gamma = 1$).

To connect this observation to total AUROC, we partition all incorrect–correct example pairs by the two examples' K groups. Pairs in which both examples have $K = 1$ account for 77–104% of the GLIB-SE improvement over DSE across the six datasets. The contributions exceed the net gain on BioASQ and SVAMP because the remaining pairs contribute a small net loss. The principal improvement thus comes from resolving a tie in the count-based representation, rather than uniformly reordering every response pattern. This supports using model-internal information alongside a semantic histogram when the observed support is narrow.

The same decomposition locates the GSM8K result more precisely. On its $K = 1$ inputs, LNPE reaches 0.840 AUROC and GLIB-SE 0.646. Pairs within this stratum contribute -0.02285 to the overall GLIB-SE–LNPE difference of -0.02443. Most of the deficit is therefore localized to cases with complete sampled agreement, where response likelihoods retain useful information. On SVAMP, the two methods are close in the same stratum: 0.750 for GLIB-SE and 0.752 for LNPE. The two arithmetic datasets differ in how their score constructions rank target errors conditional on sampled agreement. Resolving count-only ties is useful, but the resulting ranking also depends on the informativeness of the internal signal.

GLIB-SE also exceeds direct mean-energy scoring in all six dataset means, while energy-plus-partition remains competitive on GSM8K and TriviaQA (Appendix A.5).

4.4 Interaction between the prior and observed diversity

A separate component diagnostic compares No Ghost, Fixed Ghost, Energy Ghost, and within-run Shuffled Energy Ghost. Its four arms share ten scoring responses, saved partitions, the common example set, $M = 100$ Monte Carlo draws, and likelihood filter/fallback rules. They use $\gamma = 1$ and $S_{\max} = 100$, with $\alpha_{\mathrm{b}} = 0.5$ in 53 runs and 0.1 in one, and share observed Gamma draws. Details appear in Appendix A.6.

Energy Ghost is below No Ghost on BioASQ, NQ-Open, and TriviaQA, with intervals below zero (Table 6). Compared with Shuffled Energy, its interval is above zero on GSM8K and below zero on NQ-Open and TriviaQA; the other three include zero. The marginal intervals therefore show task-dependent effects under this diagnostic configuration. Stratifying by support clarifies the pattern. With $K = 1$, removing the residual dimension collapses the distribution to one point, while adding it permits a varying entropy. With $K \geq 3$, Energy Ghost is below all three controls in every dataset mean, and 98.6–99.9% of examples use fallback proposal draws.

The paired decomposition separates these gains and losses. For Energy Ghost versus No Ghost on TriviaQA, pairs within $K = 1$ contribute $+0.0076$, whereas pairs within $K \geq 3$ contribute

-0.0275, yielding an overall difference of -0.0203 after all pair types are combined. On SVAMP, the corresponding contributions are $+0.0202$ and -0.0034, with an overall gain of $+0.0076$. The balance between narrow-support benefits and diverse-support losses explains why the same component change has different aggregate effects.

A possible mechanism follows from entropy geometry. For fixed observed-component proportions q and residual mass r, the coarse-grained entropy is $h(r) + (1 - r)H(q)$, with derivative $\log((1 - r)/r) - H(q)$. It peaks at $r = 1/(1 + \exp H(q))$ and decreases beyond that point. More observed diversity lowers this peak location; a dominant residual component can suppress entropy. This identity concerns a probability vector, while the estimator averages over Dirichlet draws. Under the component mapping, the energy–score rank correlation is negative in every dataset's $K \geq 3$ stratum.

A new paired scoring diagnostic uses the main configuration and identical proposal draws to isolate likelihood filtering (Appendix A.7). Full minus no-filter AUROC ranges from -0.0053 to $+0.0002$, with all six intervals including zero. The comparison does not establish a consistent independent gain from the filter.

Together, the main experiments and gain decomposition show that GLIB-SE recovers informative error rankings from semantically consistent response sets and improves overall detection across datasets and models. With only ten scoring responses, GLIB-SE achieves higher mean AUROC than SE and DSE in all 18 dataset–model combinations and the highest mean among seven methods in 16, demonstrating the effectiveness of combining residual semantic mass modeling with model-internal signals under limited sampling.

5 Conclusion

GLIB-SE augments a finite semantic sample with an energy-dependent Ghost Cluster and scores the resulting distribution under likelihood constraints. With ten scoring responses per input, it improves on both likelihood-weighted and count-based semantic entropy across all 18 dataset–model means. Pairwise analysis locates most of the gain over count-based entropy in response sets with one observed meaning, where counts alone yield tied scores. The GSM8K comparison and separate component diagnostics show that the useful contribution depends on the confidence signal and its interaction with observed diversity. These findings motivate residual support as a way to retain ranking information under sampled agreement, with future work on direct missing-mass estimation and adaptive generation budgets.

Limitations

The single Ghost component aggregates the unobserved tail. The support-stratified analysis is descriptive, with few minority-class examples in some BioASQ subsets. Generation and semantic-grouping backends vary across runs. Direct tail-mass measurements and online budget experiments would extend the present results.

AI-use statement

AI tools assisted code development, manuscript translation, review, and semantic grouping. The authors are responsible for the methods, interpretations, citations, and final manuscript.

Ethics statement

Uncertainty scores can prioritize retrieval, abstention, and human review. High-stakes applications also require factual verification and assessment of biases in semantic grouping and model-internal signals. The aggregate supplement excludes raw responses and model weights; distribution follows the recorded rights and privacy decisions.

REPRODUCIBILITY STATEMENT

The supplement provides manuscript source, supplied-input numerical methods with tests, and aggregate results. The core code accepts semantic IDs, token scores, probabilities, and fixed parameters. It covers method computation; generation, parameter search, and plotting are separate research workflows. A separate statistical audit supplies de-identified fixed scores, canonical labels, cluster mappings and code to reconstruct the reported AUROCs and intervals. The filter diagnostic supplies numerical proposal inputs and a deterministic replay program. Appendix A specifies graph construction, labels, and component settings.

REFERENCES

Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. Rethinking uncertainty estimation in llms: A principled single-sequence measure, 2026. URL https://arxiv.org/abs/2412.15176.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

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(4):429–443, 2003. doi: 10.1023/A:1026096204727. URL https://doi.org/10.1023/A:1026096204727.

Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. Hallucination detection on a budget: Efficient bayesian estimation of semantic entropy, 2025. URL https://arxiv.org/abs/2504.03579.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4):237–264, 1953. doi: 10.1093/biomet/40.3-4.237. URL https://doi.org/10.1093/biomet/40.3-4.237.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

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.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *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/.

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

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

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

Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. Semantic energy: Detecting llm hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*, 2025. URL https://arxiv.org/abs/2508.14496.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Lucas H. McCabe, Rimon Melamed, Thomas Hartvigsen, and H. Howie Huang. Estimating semantic alphabet size for llm uncertainty quantification, 2026. URL https://arxiv.org/abs/2509.14478.

K.P. Murphy. *Probabilistic Machine Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 2022. ISBN 9780262046824. URL https://books.google.com.tw/books?id=NNOMEAAAQBAJ.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. *arXiv preprint arXiv:2405.20003*, 2024.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

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

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019. doi: 10.1162/tacl_a_00266. URL https://aclanthology.org/Q19-1016/.

Qiyao Sun, Xingming Li, Xixiang He, Ao Cheng, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Efficient hallucination detection: Adaptive bayesian estimation of semantic entropy with guided semantic exploration, 2026. URL https://arxiv.org/abs/2603.22812.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, 2015. doi: 10.1186/s12859-015-0564-6. URL https://doi.org/10.1186/s12859-015-0564-6.

Xiaohua Wang, Yuliang Yan, Longtao Huang, Xiaoqing Zheng, and Xuanjing Huang. Hallucination detection for generative large language models by Bayesian sequential estimation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15361–15371, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.949. URL https://aclanthology.org/2023.emnlp-main.949/.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your llm, 2024. URL https://arxiv.org/abs/2406.02543.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

A Implementation and additional results

A.1 Numerical implementation and configuration

The energy inputs s_{it} are selected entries of saved Hugging Face generation scores and may reflect decoding-time processing. The length L_i in Eq. (4) counts the saved token-score sequence. For each response, the likelihood input is its mean token log-probability ℓ_i. Its exponential is the geometric mean of token probabilities, used here as a length-normalized compatibility weight. The empirical class thresholds are

$$b_k = \sum_{i:c_i=k} \exp(\ell_i), \qquad b_{\mathrm{g}} = 0. \tag{8}$$

When $\sum_k b_k > 0.99$, the implementation rescales all thresholds by $0.99/(\sum_k b_k + 10^{-9})$. A proposal draw is retained when $\pi_k \geq b_k$ for every component. If fewer than $\max(5, \lfloor M/20 \rfloor)$ of the M draws survive, all proposal draws are used. Probabilities are clipped to $[10^{-12}, 1]$ before evaluating entropy. The Ghost concentration has a numerical floor of 10^{-4}.

The numerical core uses SciPy's Dirichlet sampler with NumPy's shared random state. The component study uses $M = 100$ draws for every arm. The number of draws affects the frequency of finite-draw fallback.

Parameter settings and statistical inference. The main experiments use $\alpha_{\mathrm{b}} = 0.1$, $\gamma = 10$ (equivalently mapping temperature $T = 0.1$), $S_{\max} = 100$, and $M = 200$ Monte Carlo draws across dataset–model combinations. At the start of each analysis run, the pipeline automatically calibrates τ and passes the resulting offset to the scorer. It pools per-response energies from the run's saved scoring-response set, which is also used for evaluation, and retains the 1st–99th percentile range. This is run-level, label-free batch calibration. The current experiments retain the resulting saved offsets; the supplementary configuration provides all 54 values. It combines a two-component Gaussian-mixture boundary with an isotonic boundary based on local energy variation, without correctness labels. The automatic rule averages both boundaries when the GMM returns its regular boundary, prefers isotonic when the GMM falls back, and uses the GMM result if isotonic fitting is unavailable. The supplement specifies the threshold and fallback rules. The component diagnostic uses the separate settings in Appendix A.6.

Confidence intervals hold the listed models and seeds fixed. Within each dataset, we sample question/context clusters with replacement and share each cluster's multiplicity across all its model–seed records and methods. CoQA groups questions by story; TriviaQA groups the same question across passage IDs; the other datasets use question and context. Canonicalization merges matching saved IDs or normalized content. The resulting cluster counts are 2,231, 499, 1,308, 2,394, 300, and 2,468 for BioASQ, CoQA, GSM8K, NQ-Open, SVAMP, and TriviaQA. Each replicate computes weighted AUROC within each run and then averages paired differences equally across runs. We report marginal 95% percentile intervals from 10,000 replicates per dataset. This estimates uncertainty from question sampling for the fixed model/seed roster.

For an unconstrained Dirichlet distribution with parameters $a_1, \ldots, a_J$, the expected entropy is

$$\mathbb{E}[H] = \psi(A+1) - \sum_{j=1}^{J} \frac{a_j}{A} \psi(a_j + 1), \qquad A = \sum_j a_j. \tag{9}$$

This expectation describes the unconstrained distribution; Eq. (6) instead averages the draws selected by the likelihood filter and fallback rule.

A.2 Evaluation setup

The evaluation contains 18 dataset–model combinations and 54 runs. All retained runs contain both correctness classes, with at least 49 examples in the smaller class after the common mask. GSM8K answer accuracies are 30.9% for Llama-3.1-70B-Instruct and 34.1% for Qwen3-32B. Supplementary aggregate data provide correct/error counts for all 18 combinations on the original 45,300 example-run records; questions may recur across seeds and models. Table 3 reports four-decimal means and seed standard deviations on the 45,294-record common evaluation set.

Each record contains ten scoring responses and a separately stored target answer. The scoring responses provide the semantic groups, token scores, and likelihoods; the target is not appended to this set. The retained generation implementation samples the target at temperature 0.1. Recorded generation arguments specify temperature 1.0 for scoring responses and ten few-shot examples. Maximum new-token limits are 1,024 for BioASQ, 1,536 for CoQA, 2,048 for GSM8K and SVAMP, and 128 for NQ-Open and TriviaQA.

The recorded correctness metric is `llm_judge`. The retained judging implementation uses Llama-3.1-70B-Instruct to compare the question, reference answers, and prediction for semantic agreement, treating refusals and unparseable judgments as incorrect; historical judging prompts are not fully recorded. AUROC uses canonical error labels from the saved uncertainty artifacts, with 1 denoting an incorrect target answer. Three GSM8K/Qwen3-32B records, involving two questions, disagree with the target-accuracy field. Using that alternative field for these records, or excluding them jointly across methods, preserves the 18/18 improvements over SE/DSE and the 16/18 highest-mean count. These checks leave the primary labels unchanged.

A.3 Baseline definitions

For answer y_i of length L_i, write its mean token log-probability as

$$\ell_i = \frac{1}{L_i} \sum_{t=1}^{L_i} \log p(y_{it} \mid x, y_{i,<t}). \tag{10}$$

LNPE is $-N^{-1} \sum_i \ell_i$ (Kuhn et al., 2023). Our SE baseline assigns class mass

$$p_k = \frac{\sum_{i:c_i=k} \exp(\ell_i)}{\sum_{j=1}^{N} \exp(\ell_j)}, \qquad H_{\mathrm{SE}} = -\sum_k p_k \log p_k. \tag{11}$$

This is semantic entropy with length-normalized response likelihoods (Kuhn et al., 2023; Farquhar et al., 2024). DSE uses $-\sum_k (n_k/N) \log(n_k/N)$ (Farquhar et al., 2024). All three scores use the same retained response sets as GLIB-SE.

Graph baselines use `microsoft/deberta-v2-xlarge-mnli` (He et al., 2021) to score every ordered answer pair with answer-only input. The directed matrix P has unit diagonal and finite entries in $[0, 1]$, with dimensions matching the ten saved responses. Define

$$W_{ij} = \frac{P_{ij} + P_{ji}}{2} \ (i \neq j), \qquad W_{ii} = 0, \qquad D_{ii} = \sum_j W_{ij}, \tag{12}$$

and the normalized Laplacian $L = I - D^{-1/2} W D^{-1/2}$. For isolated nodes, inverse degrees and the corresponding rows and columns of L are set to zero. The UEigV score is

$$S_{\mathrm{UEigV}} = \sum_j \max(0, 1 - \lambda_j(L)). \tag{13}$$

Writing f_1 for singleton count, $S_{\mathrm{GT}} = KN/(N - f_1)$ when $f_1 < N$. The hybrid estimate is $S_{\mathrm{hyb}} = \max(S_{\mathrm{GT}}, S_{\mathrm{UEigV}})$; when every class is a singleton, it uses S_{UEigV}. With $\widetilde{p}_k = K(n_k/N)/S_{\mathrm{hyb}}$, Hybrid entropy is

$$-\sum_k \frac{\widetilde{p}_k \log \widetilde{p}_k}{1 - (1 - \widetilde{p}_k)^N}, \tag{14}$$

Terms with nonpositive $\widetilde{p}_k$ or denominator are omitted; an empty sum returns zero. UEigV follows Lin et al. (2024); Hybrid alphabet size and Hybrid entropy follow McCabe et al. (2026).

Table 3: Detailed AUROC for the same common evaluation set as Table 1. Each cell gives the three-seed mean and sample standard deviation to four decimals. Bold marks the highest unrounded mean, not statistical significance.

Dataset	Model	LNPE	SE	DSE	UEigV	Hybrid size	Hybrid entropy	GLIB-SE (ours)
BioASQ	Llama-3.1-70B -Instruct	0.7377 ±0.0056	0.8164 ±0.0116	0.8330 ±0.0082	0.7753 ±0.0111	0.8092 ±0.0131	0.8335 ±0.0096	**0.8410** ±0.0109
	Qwen3-30B-A3B	0.7440 ±0.0058	0.7591 ±0.0139	0.7590 ±0.0087	0.7407 ±0.0164	0.7603 ±0.0050	0.7684 ±0.0079	**0.7871** ±0.0071
	Qwen3-32B	0.7488 ±0.0039	0.8048 ±0.0069	0.8095 ±0.0055	0.7879 ±0.0070	0.7991 ±0.0078	0.8206 ±0.0077	**0.8265** ±0.0074
	Seed-Coder-8B -Instruct	0.7085 ±0.0650	0.7713 ±0.0127	0.8055 ±0.0192	0.4658 ±0.0471	0.5795 ±0.0131	0.7348 ±0.0294	**0.8070** ±0.0205
CoQA	Llama-3.1-70B -Instruct	0.7459 ±0.0360	0.6777 ±0.0469	0.7563 ±0.0681	0.7482 ±0.0390	0.7345 ±0.0291	0.7627 ±0.0572	**0.7640** ±0.0648
	Qwen3-30B-A3B	0.7296 ±0.0440	0.7635 ±0.0333	0.7765 ±0.0203	0.7359 ±0.0079	0.7859 ±0.0172	0.7865 ±0.0130	**0.8125** ±0.0246
	Qwen3-32B	0.7381 ±0.0580	0.8214 ±0.0083	0.8406 ±0.0039	0.7482 ±0.0420	0.7803 ±0.0224	0.8272 ±0.0085	**0.8487** ±0.0017
GSM8K	Llama-3.1-70B -Instruct	**0.8170** ±0.0095	0.7986 ±0.0128	0.7940 ±0.0112	0.7649 ±0.0186	0.7868 ±0.0101	0.7991 ±0.0134	0.8044 ±0.0099
	Qwen3-32B	**0.8330** ±0.0077	0.7795 ±0.0130	0.7834 ±0.0070	0.7511 ±0.0099	0.7957 ±0.0118	0.7958 ±0.0075	0.7967 ±0.0061
NQ-Open	Llama-2-13B -chat-hf	0.6072 ±0.0160	0.7231 ±0.0166	0.7351 ±0.0154	0.6969 ±0.0121	0.7120 ±0.0101	0.7315 ±0.0128	**0.7417** ±0.0140
	Llama-3.1-8B -Instruct	0.6847 ±0.0253	0.7266 ±0.0108	0.7453 ±0.0087	0.6540 ±0.0155	0.6709 ±0.0227	0.7208 ±0.0087	**0.7491** ±0.0086
	Qwen3-8B	0.7220 ±0.0161	0.7503 ±0.0255	0.7551 ±0.0240	0.6999 ±0.0305	0.7199 ±0.0296	0.7427 ±0.0265	**0.7593** ±0.0253
SVAMP	Llama-2-70B -chat-hf	0.8591 ±0.0297	0.8538 ±0.0023	0.8550 ±0.0068	0.8111 ±0.0106	0.8557 ±0.0073	0.8627 ±0.0063	**0.8708** ±0.0145
	Llama-3.1-70B -Instruct	0.9097 ±0.0100	0.8871 ±0.0184	0.8786 ±0.0181	0.8202 ±0.0168	0.8842 ±0.0112	0.8883 ±0.0081	**0.9345** ±0.0184
TriviaQA	Llama-2-13B -chat-hf	0.6674 ±0.0161	0.7990 ±0.0529	0.8062 ±0.0485	0.7628 ±0.0608	0.7771 ±0.0401	0.7985 ±0.0462	**0.8180** ±0.0460
	Llama-3.1-8B -Instruct	0.7721 ±0.0358	0.8295 ±0.0514	0.8447 ±0.0517	0.7911 ±0.0588	0.7880 ±0.0492	0.8293 ±0.0538	**0.8485** ±0.0500
	Qwen3-4B -Instruct	0.7781 ±0.0570	0.8336 ±0.0180	0.8389 ±0.0230	0.7784 ±0.0396	0.7934 ±0.0423	0.8229 ±0.0270	**0.8468** ±0.0264
	Qwen3-8B	0.7420 ±0.0190	0.8112 ±0.0289	0.8235 ±0.0253	0.7776 ±0.0305	0.7879 ±0.0262	0.8139 ±0.0282	**0.8367** ±0.0233

A.4 Paired comparisons on the common evaluation set

Six examples have empty token-score inputs: two in BioASQ, one in CoQA, and three in SVAMP. Saved SE and LNPE are nonfinite there; saved energy uses a zero fallback. Applying the same complete-case mask to every method retains 45,294 example-run records and all 54 runs. GSM8K and TriviaQA retain all their original examples. Table 4 reports GLIB-SE differences against SE, DSE, LNPE, and Hybrid entropy under the inference settings in Appendix A.1.

A.5 Local signal controls

Two local controls test direct aggregation of the saved energy signal. Simple energy is $\bar{E}$ from Eq. (4). Energy plus partition (E+P) is $\min_k N^{-1} \sum_{i:c_i=k} e_i$, normalized by the total response count N. These local definitions distinguish the controls from the published Semantic Energy detector (Ma et al., 2025). Table 5 compares them and NumSets, the observed class count K, with GLIB-SE on the same 45,294 records as the main table.

Table 4: Mean paired AUROC differences over the fixed model and seed roster. The 95% intervals use 10,000 question/context-cluster bootstrap resamples with shared cluster weights across runs and methods.

Dataset	Comparator	GLIB-SE minus comparator	95% interval
BioASQ	SE	+0.0275	[+0.0220, +0.0329]
BioASQ	DSE	+0.0136	[+0.0098, +0.0175]
BioASQ	LNPE	+0.0807	[+0.0717, +0.0897]
BioASQ	Hybrid entropy	+0.0261	[+0.0208, +0.0315]
CoQA	SE	+0.0542	[+0.0454, +0.0630]
CoQA	DSE	+0.0172	[+0.0120, +0.0225]
CoQA	LNPE	+0.0705	[+0.0513, +0.0899]
CoQA	Hybrid entropy	+0.0162	[+0.0096, +0.0231]
GSM8K	SE	+0.0116	[+0.0059, +0.0175]
GSM8K	DSE	+0.0119	[+0.0079, +0.0160]
GSM8K	LNPE	-0.0244	[-0.0349, -0.0139]
GSM8K	Hybrid entropy	+0.0031	[-0.0018, +0.0079]
NQ-Open	SE	+0.0167	[+0.0118, +0.0215]
NQ-Open	DSE	+0.0049	[+0.0028, +0.0070]
NQ-Open	LNPE	+0.0787	[+0.0679, +0.0893]
NQ-Open	Hybrid entropy	+0.0184	[+0.0139, +0.0229]
SVAMP	SE	+0.0322	[+0.0143, +0.0518]
SVAMP	DSE	+0.0358	[+0.0206, +0.0524]
SVAMP	LNPE	+0.0182	[+0.0022, +0.0360]
SVAMP	Hybrid entropy	+0.0272	[+0.0107, +0.0449]
TriviaQA	SE	+0.0192	[+0.0153, +0.0230]
TriviaQA	DSE	+0.0091	[+0.0070, +0.0114]
TriviaQA	LNPE	+0.0976	[+0.0902, +0.1051]
TriviaQA	Hybrid entropy	+0.0213	[+0.0178, +0.0247]

Table 5: Additional count baseline and local signal controls on the same 45,294 records as the main comparison. Definitions appear in Appendix A.5.

Dataset	GLIB-SE	NumSets	Simple energy	E+P
BioASQ	0.8154	0.7998	0.7462	0.8116
CoQA	0.8084	0.7955	0.7599	0.7936
GSM8K	0.8006	0.7841	0.7172	0.8006
NQ-Open	0.7500	0.7433	0.6866	0.7465
SVAMP	0.9026	0.8645	0.8213	0.8845
TriviaQA	0.8375	0.8249	0.7968	0.8382

GLIB-SE exceeds Simple energy in all six dataset means, while E+P is close on GSM8K and slightly higher on TriviaQA. GLIB-SE minus E+P is approximately -0.000075 on GSM8K and -0.000730 on TriviaQA, with question-cluster intervals $[-0.0041, 0.0040]$ and $[-0.0030, 0.0015]$.

The near tie on GSM8K combines opposite model-level differences: GLIB-SE minus E+P is -0.0064 for Llama-3.1-70B-Instruct, with interval $[-0.0115, -0.0013]$, and $+0.0062$ for Qwen3-32B, with interval $[-0.0001, 0.0126]$. All four model-level intervals on TriviaQA include zero. Direct energy aggregation with the partition is therefore already competitive in these settings; the separate component diagnostic examines prior effects under internally matched conditions.

A.6 Component diagnostic at a separate fixed configuration

The four arms share the full saved partition, ten scoring responses, common valid-example mask, $M = 100$ draws, $\gamma = 1$, and $S_{\max} = 100$. Recorded base concentrations are 0.5 in 53 runs and 0.1 in CoQA/Qwen3-32B/Seed30; offsets are read without refitting. Thus the arms are internally matched, but differ from the main configuration's $\alpha_{\mathrm{b}} = 0.1$ and $\gamma = 10$. Fixed Ghost sets $\alpha_{\mathrm{g}} = 51\alpha_{\mathrm{b}}$. Energy Ghost uses Eq. (5); shuffled energy uses one deterministic permutation within each run, independent of correctness labels. No Ghost removes the residual dimension and applies the same observed-class likelihood thresholds.

Table 6: Energy Ghost minus each internally matched control under the separate diagnostic configuration. The 95% intervals use 10,000 question/context-cluster bootstrap resamples over fixed models and seeds. Interval endpoints retain five decimals to resolve values near zero.

Dataset	vs. No Ghost	vs. Fixed Ghost	vs. Shuffled Energy
BioASQ	-0.0292 [-0.03594, -0.02260]	-0.0299 [-0.03713, -0.02259]	-0.0048 [-0.01266, +0.00285]
CoQA	-0.0108 [-0.02221, +0.00039]	-0.0116 [-0.02339, -0.00002]	+0.0101 [-0.00159, +0.02178]
GSM8K	-0.0042 [-0.01093, +0.00257]	-0.0082 [-0.01550, -0.00067]	+0.0083 [+0.00025, +0.01636]
NQ-Open	-0.0244 [-0.03076, -0.01828]	-0.0265 [-0.03297, -0.02021]	-0.0071 [-0.01417, -0.00007]
SVAMP	+0.0076 [-0.00875, +0.02664]	+0.0073 [-0.01155, +0.02911]	+0.0111 [-0.00855, +0.03302]
TriviaQA	-0.0203 [-0.02525, -0.01554]	-0.0201 [-0.02566, -0.01489]	-0.0067 [-0.01244, -0.00114]

Table 7: Main-configuration filter diagnostic with shared proposal draws ($M = 200$). Accepted counts, fallback percentages and Spearman correlations between full and no-filter scores are averaged across runs. Differences are full minus no-filter AUROC with 95% question/context-cluster intervals.

Dataset	Accepted/200	Fallback (%)	ρ	ΔAUROC [95% CI]
BioASQ	67.2	57.8	0.900	-0.0008 [-0.0046, +0.0029]
CoQA	38.2	73.0	0.985	-0.0024 [-0.0065, +0.0015]
GSM8K	50.6	66.2	0.972	-0.0007 [-0.0044, +0.0031]
NQ-Open	38.6	73.6	0.986	-0.0004 [-0.0026, +0.0018]
SVAMP	105.7	33.0	0.756	-0.0053 [-0.0181, +0.0070]
TriviaQA	64.7	57.2	0.943	+0.0002 [-0.0024, +0.0030]

Common uniform variates and Gamma inverse-CDF transforms generate coupled Dirichlet draws across arms, sharing the observed Gamma variates. Sampling and shuffling seeds are fixed, and per-example acceptance counts and fallback indicators are recorded. Energy Ghost is below No Ghost on BioASQ, NQ-Open, and TriviaQA, with intervals below zero. Against shuffled energy, its interval is above zero on GSM8K, below zero on NQ-Open and TriviaQA, and includes zero on the other three datasets. The GSM8K and NQ-Open bounds lie close to zero. These marginal comparisons describe this diagnostic configuration, rather than identifying an independent energy contribution in the main setting. Fallback is frequent: Fixed Ghost uses unfiltered proposal draws for about 99.2–99.9% of examples across datasets, and Energy Ghost for about 60.9–91.7%.

The configuration difference also changes prior mass allocation. For $N = 10$ and $K = 1$, the mean Ghost mass before likelihood filtering approaches the range $[0.0098, 0.5]$ under the main base concentration, compared with $[0.0455, 0.8279]$ when $\alpha_{\mathrm{b}} = 0.5$. Binary entropy peaks at residual mass 0.5. This illustrates why the two configurations can behave differently; it does not equate entropy of a mean vector with the mean entropy score.

A.7 Likelihood filtering at the main configuration

We run a paired scoring diagnostic on the same 54 runs and 45,294 records, using $\alpha_{\mathrm{b}} = 0.1$, $\gamma = 10$, $S_{\max} = 100$, $M = 200$ and the 54 saved offsets. For each example, both arms share the same 200 Dirichlet draws. The full arm applies the likelihood filter and falls back to all draws when fewer than ten survive; the no-filter arm always uses all draws. Table 7 reports the results using the question/context-cluster inference protocol in Appendix A.1.

The filter changes the score for examples with sufficiently many accepted draws; fallback makes the two arms identical for the remaining examples. Fallback ranges from 33.0% on SVAMP to 73.6% on NQ-Open. Full minus no-filter AUROC ranges from -0.0053 to $+0.0002$, and all six intervals include zero. Thus this diagnostic shows no consistent aggregate benefit from likelihood filtering. Its role is to condition proposal draws on response compatibility, with the fallback retaining the proposal when acceptance is scarce.

This diagnostic is a separate recomputation and does not replace the current manuscript results. Its full-arm dataset AUROCs are 0.0072–0.0200 below the current manuscript scores, and two additional fixed Monte Carlo streams retain the gap. The diagnostic's recomputation path remains to be reconciled with the current experimental implementation.