WHAT DOES SEMANTIC UNCERTAINTY REVEAL AND WHAT DOES IT MISS? A LARGE-SCALE STUDY OF SEMANTIC UNCERTAINTY IN LLM NATURAL LANGUAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic uncertainty quantification (SUQ) has become a central paradigm for detecting hallucinations in free-form language generation, as it measures uncertainty over meanings rather than surface forms. However, existing studies typically evaluate SUQ under limited tasks, models, and generation settings, making it difficult to understand when semantic uncertainty is reliable and where it fundamentally fails. We investigate both questions through SURE, a large-scale empirical study treating SUQ as both an object of evaluation and a lens on LLM behavior. We compare 19 measures and variants from six studies across four tasks, 19 English-output datasets, and 13 LLMs, complemented by 12-model translation experiments across 14 non-English-target directions. First, we analyze Semantic Uncertainty itself: method rankings vary with task and simple consistency measures remain competitive in several evaluated settings. We identify SURE-OC, diagnostic subsets of low-quality generations assigned confidence, to characterize where semantic agreement and quality diverge. Second, we use semantic uncertainty to analyze LLMs: higher-performing models tend to exhibit lower uncertainty, including on low-quality outputs, while QA temperature sweeps show that better uncertainty separability can accompany worse answer accuracy. Together, the two perspectives distinguish the effectiveness of an uncertainty estimator from the behavior of the generator, revealing why lower semantic uncertainty alone does not establish greater reliability.

1 INTRODUCTION

Large language models (LLMs) have become ubiquitous across a wide range of applications. Despite this progress, LLMs remain prone to hallucinations: they may generate fluent and plausible responses that are factually incorrect, unsupported, or inconsistent with the input. This reliability issue is particularly challenging in free-form generation, where the output space is open-ended. Semantic Entropy (Kuhn et al., 2023) and its black-box variant, Discrete Semantic Entropy (Farquhar et al., 2024), motivate a growing family of semantic uncertainty quantification (SUQ) methods (Lin et al., 2023; Nguyen et al., 2025; Qiu & Miikkulainen, 2024; Nikitin et al., 2024).

Existing studies demonstrate the promise of SUQ, but evaluate on different datasets and LLMs under varying quality metrics and generation protocols (Farquhar et al., 2024; Qiu & Miikkulainen, 2024; Nguyen et al., 2025). Their results are therefore difficult to compare directly. Broader UQ studies provide valuable comparisons (Vashurin et al., 2025a), but important questions about SUQ remain open:

- Is the effectiveness of SUQ consistent across **diverse tasks**?
- What are the **limitations and blind spots** of black-box SUQ methods?
- As LLM capabilities improve, does expressed uncertainty exhibit **scaling-law-like patterns**?

Addressing these questions requires a shared, multi-task protocol that compares estimators on matched generations and examines how uncertainty varies across models and decoding settings. This provides

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.

a common basis for two complementary perspectives: evaluating **SUQ effectiveness and its blind spots**, and understanding **the LLM behavior reflected in uncertainty**.

In this work, we investigate these questions through **S**emantic **U**nce**R**tainty **E**valuation (**SURE**), a large-scale empirical study. We call its curated evaluation data suite SUREBench. Shared generations and NLI evidence support matched comparisons of 19 measures and variants from six studies on 19 English-output datasets and 13 LLMs, spanning question answering (QA), mathematics, summarization, and translation. A multilingual extension covers **1**4 non-English-target translation directions with 12 LLMs. Overall, the study includes **more than 10 million generated responses, and over 24 million semantic uncertainty estimates**. This breadth supports both evaluating SUQ itself and using SUQ to understand LLMs. Our main findings are:

SUQ is not always better than simple consistency measures. In our summarization and translation comparisons, BERT similarity and cosine similarity can outperform semantic estimators under the evaluated quality metrics. Comparisons between continuous NLI scores and discrete labels further highlight the importance of semantic granularity. These observations motivate representations that capture partial semantic overlap in longer responses, rather than treating all agreement as strict equivalence.

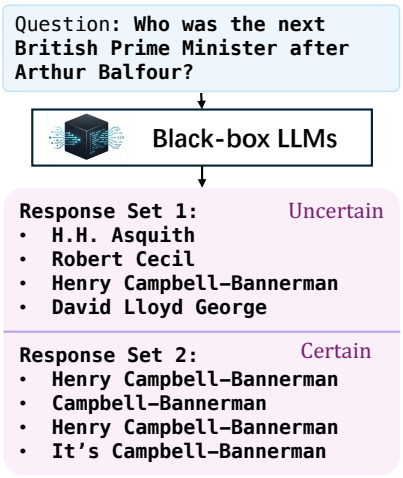


Figure 1: Semantic Uncertainty Quantification in confabulation detection.

Semantic agreement can hide overconfident errors. When sampled responses are semantically identical but wrong, agreement alone cannot distinguish them from consistently correct responses. We identify **SURE-OC**, model-specific diagnostic subsets of SUREBench combining low quality with low uncertainty. These subsets connect the two perspectives: they expose missed failures of an estimator and support comparison of the generators' errors. Within-family models show moderately greater overlap in these failures, without establishing a shared causal mechanism.

More capable LLMs are generally more confident. Across the evaluated models, higher task performance is associated with lower average semantic uncertainty, indicating greater agreement among sampled responses. However, the relative uncertainty gap between low- and high-quality outputs shows a marginal decreasing trend. QA temperature sweeps further reveal that better uncertainty-based error separation can accompany worse answer accuracy. These observations distinguish semantic consistency from generation quality and motivate examining both when interpreting uncertainty.

Our contributions are threefold: (1) a large-scale study with SUREBench and a unified, reusable evaluation pipeline for future SUQ research; (2) SURE-OC, diagnostic subsets for investigating low-uncertainty failures; and (3) empirical analyses linking semantic uncertainty to model capability and decoding behavior, informing how uncertainty signals should be interpreted alongside generation quality. We provide the code and an interactive results explorer [1] to support reproducibility and further analysis.

2 Related Work

Semantic uncertainty estimators. LLM uncertainty can be estimated from token-level scores, response similarity, or self-evaluation (Kadavath et al., 2022; Fomicheva et al., 2020; Lin et al., 2022; Malinin & Gales, 2020; Manakul et al., 2023). SUQ accounts for the fact that different response strings can express the same answer. Semantic Entropy (SE) (Kuhn et al., 2023) and its black-box variant Discrete Semantic Entropy (DSE) (Farquhar et al., 2024) estimate uncertainty over semantic clusters rather than raw generations. Graph-based methods (Lin et al., 2023) derive uncertainty from eigenvalues and degree statistics of a semantic graph, while KLE (Nikitin et al., 2024) encodes similarity through semantic kernels and von Neumann entropy. Semantic Density (Qiu & Miikkulainen, 2024) estimates confidence in semantic space; our comparison includes its black-box

[1] https://anonymous.4open.science/r/SUREBench-CA71/index.html

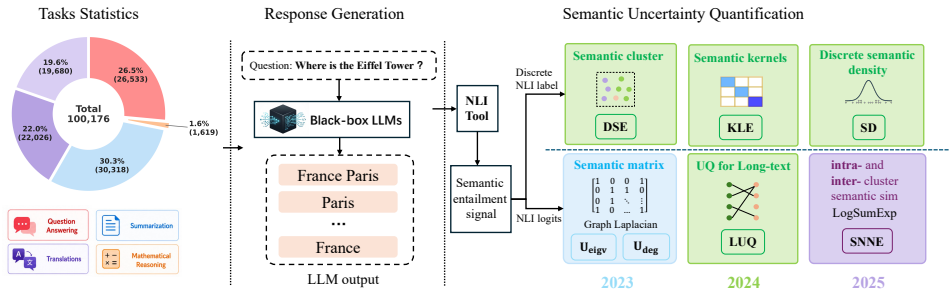


Figure 2: Shared experimental basis for **SURE**. Reusing sampled responses and NLI evidence supports matched comparisons of estimators. The resulting uncertainty scores also support analyses of overconfident failures and LLM behavior.

variant. LUQ (Zhang et al., 2024) targets longer, multi-sentence responses, and SNNE (Nguyen et al., 2025) replaces hard clustering with aggregation of intra-cluster and inter-cluster similarities.

Other approaches use embedding similarity (Grewal et al., 2024), learned estimators (Li et al., 2025b), or distillation (Phillips et al., 2026), with task-specific extensions for multiple valid references (Yang et al., 2025) and reasoning-intensive generation (Pandey et al., 2026). We distinguish output-only black-box estimation from white-box approaches using generator logits or hidden states (Kossen et al., 2024; Scalena et al., 2025; Farr et al., 2025; Kunitomo-Jacquin et al., 2026; Sun et al., 2026; Park & Cho, 2025). Our comparison focuses on black-box semantic estimators under shared generations and semantic evidence.

Evaluating uncertainty and understanding model behavior. Prior studies provide complementary protocols and implementations. Ye et al. (2024) study uncertainty through conformal prediction, while LM-Polygraph (Vashurin et al., 2025a) compares black-box and white-box methods across generation tasks. UQLM (Bouchard et al., 2026) provides implementations spanning black-box, white-box, and LLM-as-a-judge approaches. Overconfident errors are also recognized in prior work (Farquhar et al., 2024; Ye et al., 2024; Simhi et al., 2025). SURE connects estimator evaluation with analysis of the generators: matched comparisons characterize SUQ effectiveness, SURE-OC records low-quality outputs assigned low uncertainty, and capability and decoding analyses examine the model behavior reflected in uncertainty. This separates an estimator's effectiveness from the quality and semantic consistency of the model responses it evaluates.

Multilingual uncertainty. Prior studies examine multilingual QA through BiG-SURE (Bhattacharya et al., 2026) and translation gender bias (Staliunaite et al., 2026). Our 14-direction translation study extends evaluation beyond English outputs, complementing analyses of task representations and model behavior.

3 Study Design and Evaluation Protocol

SURE supports both perspectives using SUREBench, an evaluation data suite assembled from established datasets: 19 English-output datasets spanning four task families and 14 non-English-target translation directions. Section 4 compares estimators on shared generations and semantic evidence, then identifies SURE-OC. Section 5 uses uncertainty summaries to examine model capability and decoding behavior in the English-output suite. Comparisons are conditional on the evaluated tasks, models, backends, and metrics; model-level correlations do not isolate causal effects. Appendix B details coverage and aggregation.

3.1 Tasks and Datasets

The four task families contrast factual recall, exact-answer reasoning, abstraction, and cross-lingual meaning preservation, allowing us to examine SUQ across different output structures (Table 1).

Question Answering (QA). Short factual answers test agreement at the entity or phrase level. We use TriviaQA (Joshi et al., 2017), SQuAD (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019), CoQA (Reddy et al., 2019), and BioASQ (Tsatsaronis et al., 2015).

Mathematical Reasoning. Reasoning traces may differ while reaching the same final answer. SVAMP (Patel et al., 2021) and GSM8K (Cobbe et al., 2021) support comparison of full-response and final-answer representations.

Summarization. Summaries can overlap on some claims and differ on others, motivating comparisons of semantic granularity. We use XSum (Narayan et al., 2018), AESLC (Zhang & Tetreault, 2019), CNN/DailyMail (See et al., 2017), and Multi-News (Fabbri et al., 2019).

Translation. Translation tests uncertainty in meaning-preserving generation across output languages. We evaluate eight into-English directions from WMT19 (Barrault et al., 2019), six non-English-to-non-English directions from OPUS-100 (Zhang et al., 2020), and eight English-to-multilingual directions from WMT24++ (Deutsch et al., 2025), using multilingual mDeBERTa for non-English targets.

Details of each task and dataset in Appendix B.3.

3.2 Black-box Semantic Uncertainty Quantification Methods

We choose publicly reproducible SUQ methods that use generated outputs without prompt modification or access to generator internals. Six studies yield 19 measures and variants. Shared generations and NLI evidence enable matched comparisons under a common protocol, not estimates of independently tuned optima. Existing SUQ methods differ in their formulations, but they share a common high-level pipeline as shown in Fig 2. Given an input query q, a black-box language model $\mathcal{M}$ generates multiple responses

$$r_i \sim \mathcal{M}(q), \quad i = 1, \ldots, n, \tag{1}$$

where n is the number of sampled outputs. Then a natural language inference (NLI) tool is employed to evaluate the semantic relations among these responses, such as *entailment*, *neutral*, or *contradict*. We use a DeBERTa NLI comparator for English outputs and multilingual mDeBERTa for non-English targets. The NLI results are aggregated into an uncertainty score $u(q)$. Here, certainty refers to consistency among sampled outputs, not a calibrated probability of correctness or a model's self-reported belief. We consider black box SUQ methods from (Lin et al., 2023; Farquhar et al., 2024; Nguyen et al., 2025; Qiu & Miikkulainen, 2024; Zhang et al., 2024; Nikitin et al., 2024), with implementation details in Appendix B.4. Methods reuse NLI evidence, consuming either discrete labels or continuous class scores (Figure 2). Access to this external comparator remains compatible with black-box access to the generator.

3.3 Evaluation Metrics

Following (Kuhn et al., 2023), we evaluate whether higher uncertainty ranks lower-quality generations ahead of higher-quality ones. We evaluate QA and math with accuracy. For summarization and translation, ROUGE-L provides a lexical-overlap reference following the SUQ evaluation precedent of SNNE (Nguyen et al., 2025). We retain ROUGE-L for the main analyses to maintain continuity with prior work, and also evaluate translation with COMET (Rei et al., 2020) and summarization with SEAHORSE (Clark et al., 2023), retaining the same PRR computation. Appendix F presents the metric rationale, comparisons, and task-specific PRR results; metric definitions are in Appendix B.6.

We report **AUROC** (Farquhar et al., 2024) for QA/math and the **normalized-quality retention area** (denoted **PRR**; Appendix B.6) for summarization/translation. Higher values indicate better uncertainty-based quality ranking under the same quality target; neither metric measures probabilistic calibration.

4 Evaluating Semantic Uncertainty Methods with SUREBench

With SUREBench, we systematically examine the effectiveness of SUQ methods across diverse tasks and LLMs. We focus on two aspects: (1) **SUQ methods**, where we examine their performance; and (2) **Blind spots of SUQ**, where we discuss the phenomenon of overconfidence and introduce SURE-OC, model-specific diagnostic subsets of SUREBench formed by overconfident instances.

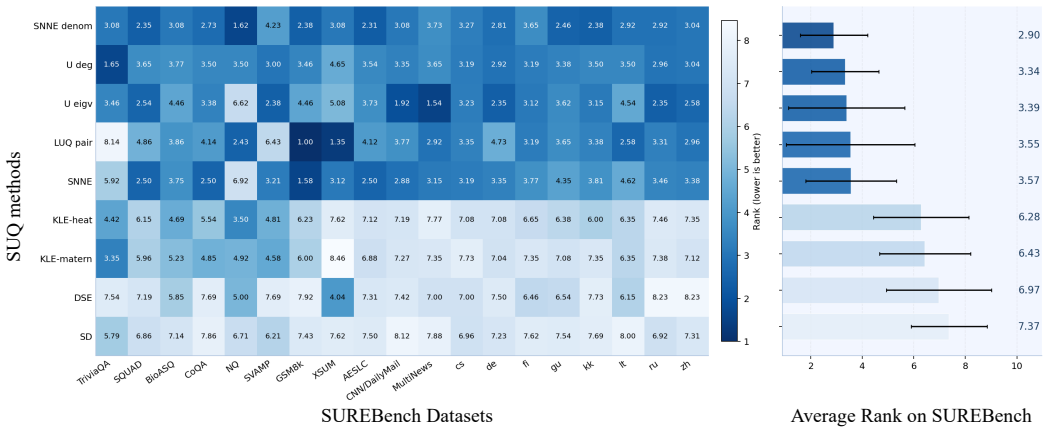
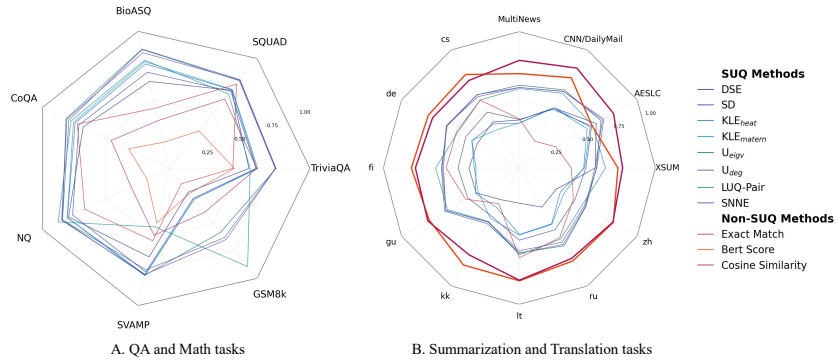


Figure 3: The heatmap shows the **relative ranking of SUQ methods**. Lower rank means higher performance on AUROC (QA/Math) and PRR (Summarization/Translation). The methods are ordered by their mean rank across 13 LLMs. The bar chart summarizes the overall mean rank of each method.



A. QA and Math tasks B. Summarization and Translation tasks

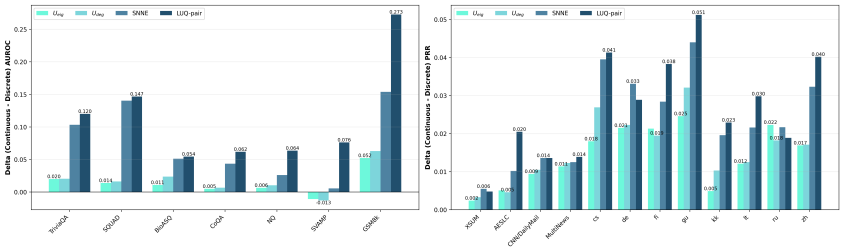
Figure 4: SUQ and consistency baselines on LLAMA-3.1-405B. SUQ performs better on QA/math; BERT similarity and Cosine Similarity are stronger on summarization/translation under ROUGE-L. "Bert Score" in the legend denotes mean-pooled BERT similarity, not token-level BERTScore.

4.1 EVALUATING SUQ METHODS VIA SURE

We generate outputs using 13 LLMs across 19 datasets in SUREBench and implement each SUQ method. The task performance of each model is provided in the Appendix C. As shown in Figure 3, SNNE denom achieves the best average performance. Although the KLE family performs well on QA tasks, its performance is weaker on summarization and translation. Notably, SNNE Nguyen et al. (2025) is a relatively recent method, suggesting that SUQ methods have continued to improve over time, and that newer methods can, on average, outperform the earlier ones.

We also compare three consistency baselines: *Exact Match* Cole et al. (2023); Chen & Mueller (2024), mean-pooled *BERT similarity*, and *Cosine Similarity* Shorinwa et al. (2025); Bouchard et al. (2026). Implementation details are in Appendix B.4. We test their uncertainty quantification performance on SUREBench with LLAMA 3.1-405B as the generator. As shown in Figure 4, although SUQ methods outperform these baselines on QA and math tasks, **on summarization and translation, *BERT similarity* and *Cosine Similarity* are stronger under ROUGE-L in this comparison**.

To understand this phenomenon, we inspect representative cases from QA and summarization in Figure 5. These cases suggest that **hard semantic-equivalence decisions can obscure partial overlap**, a possible limitation of the evaluated representation rather than an established mechanism shared by all SUQ methods. Summarization and translation produce longer outputs, where two responses may be partially aligned without being strictly equivalent. In such cases, reducing a response pair to an NLI label is often too coarse to represent its real semantic relation. By comparison, continuous similarity signals such as Cosine Similarity can better preserve graded relations between



QA and Math task | Summarization and Translation task

Figure 6: Original continuous SUQ vs. corresponding discrete variants on SUREBench, LLAMA3-405B generation. Each bar shows the improvement of the continuous method over the discrete variant. **Continuous methods generally outperform their discrete counterparts**, with the most pronounced gains on summarization and translation tasks.

long-form outputs, which may explain their stronger empirical performance on summarization and translation. Our empirical analysis reveals that:

On summarization and translation under ROUGE-L, SUQ is not uniformly stronger than simple consistency measures in the evaluated settings, motivating task-aware semantic representations.

4.1.1 MULTILINGUAL TRANSLATION EVALUATION.

Across the six non-English-to-non-English and eight English-to-multilingual directions, Discrete SNNE (only-denom) has the best macro PRR among the 19 SUQ measures under ROUGE-L. Simple consistency measures remain competitive. Appendix D reports the direction-level results; full 19-measure COMET-based multilingual UQ is available for Llama 3.1-8B only (Appendix F).

4.1.2 IMPACT OF SEMANTIC SIGNAL GRANULARITY.

As shown in Figure 3, the methods with stronger average performance make use of DeBERTa's corresponding logits, such as SNNE, U_{deg}, U_{eigv}, and LUQ-Pair. To further examine the effect of NLI output granularity, we modify SUQ methods that originally rely on fine-grained NLI signals by degrading their logits into binary signals, and then compare their performance against the original versions. We implement both original SUQs and their discrete version on SUREBench and show the comparison in Fig 6, while the full experimental results are deferred to the Appendix C. The results show that **fine-grained NLI signals are effective for uncertainty estimation and bring more expressive uncertainty representations.**

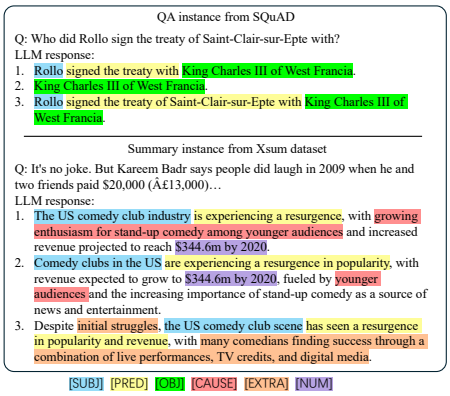
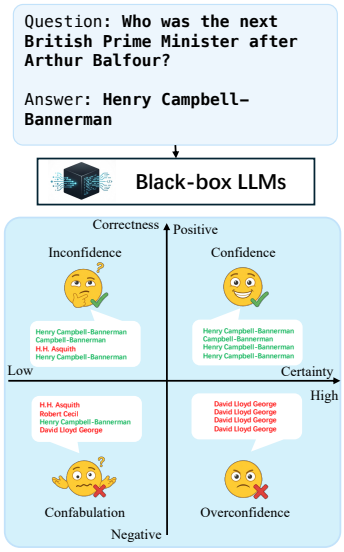
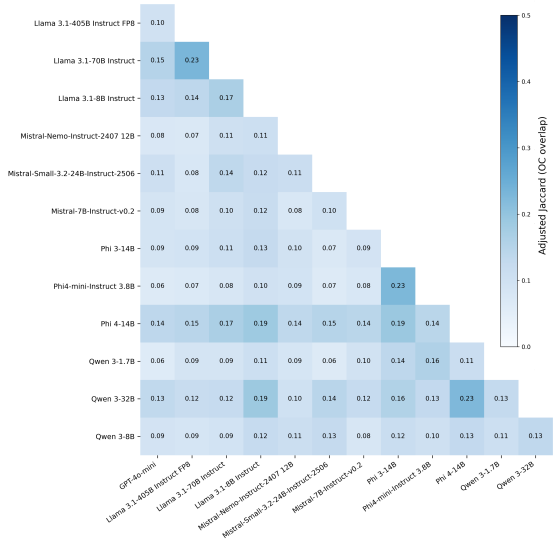


Figure 5: Semantic alignment in QA and summarization. Summarization outputs exhibit partial overlap in subject, cause, and details. Such relations are difficult to represent with NLI signals.

Additionally, we study how the choice of the NLI tools affects SUQ. In particular, we replace DeBERTa with an LLM for semantic equivalence judgment and further evaluate both NLI tools on standard NLI benchmark datasets. The detailed analysis, experimental results, and corresponding figures are presented in the Appendix C.2. Furthermore, we analyze the correlation between LLM output length and SUQ performance, and find a statistically significant correlation on QA and math datasets. Detailed results are provided in the Appendix C.3.



A. Four types of LLM generations categorized by correctness and certainty



B. Pairwise overlap of overconfidence of different LLMs on NQ

Figure 7: A: a schematic illustration of four generation types defined by correctness and certainty. B: pairwise overlap of overconfident errors across different LLMs on NQ, where darker colors indicate two LLMs are more likely to be overconfident on the same question.

4.2 Overconfidence: The Blind Spot of Uncertainty Quantification

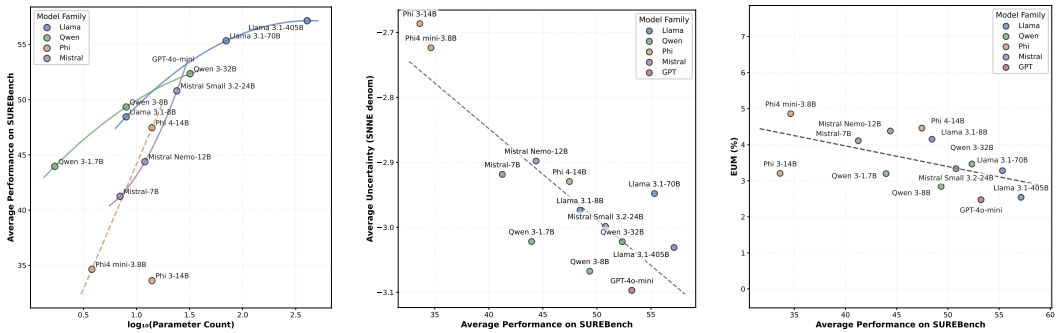
We have evaluated a range of SUQ methods and examined several key factors that influence their performance. In this section, we move beyond specific SUQ methods and explore a more fundamental question: **What is the fundamental limitation of SUQ methods?** According to the correctness and certainty of LLM outputs in NLG, they can be roughly categorized into four types, as illustrated in Figure 7. A. When the LLM produces an incorrect answer with high certainty, while its sampled outputs remain nearly identical, in this scenario, UQ can no longer reliably indicate whether the response is correct. This phenomenon has been noticed in prior work Farquhar et al. (2024); Ye et al. (2024); Simhi et al. (2025). It has been described as **overconfidence** Ye et al. (2024) or hallucination with high certainty Simhi et al. (2025), and closely relates to cases where LLMs remain systematically wrong despite exhibiting low uncertainty.

To identify overconfident instances, we build **SURE-OC, model-specific diagnostic subsets of SUREBench, by extracting the data points that exhibit low quality with high certainty**. SURE-OC identifies low-SNNE-denom failures under this protocol, rather than failures shared by all uncertainty estimators. These diagnostic subsets expose missed errors and can support future research on black-box hallucination detection and LLM training.

First, we define overconfident instances as low-quality generations that are assigned low uncertainty by an SUQ method. Formally, for a dataset $\mathcal{D}$ and an uncertainty score u_i for instance i, we define:

$$\mathcal{O}_\alpha = \{i \in \mathcal{D} \mid \mathcal{E}_t(i) = 1 \ \wedge \ u_i \leq Q_\alpha(\{u_j\}_{j \in \mathcal{D}})\}\,, \tag{2}$$

where $\mathcal{E}_t(i)$ is a low quality indicator, and $Q_\alpha(\cdot)$ denotes the α-quantile of uncertainty scores within the dataset. In the following analysis, we adopt SNNE denom as the uncertainty metric, as it achieves the best average performance in SUREBench. We set $\alpha = 0.2$, corresponding to selecting generations in the lowest 20% uncertainty percentile of the dataset. We additionally report results of $\alpha \in \{0.05, 0.1, 0.3, 0.4\}$ in the Appendix C.4. For QA and math tasks, $\mathcal{E}_t(i) = 1$ if the generated answer is incorrect. For summarization and translation, where quality is evaluated by Rouge-L, we define $\mathcal{E}_t(i) = 1$ if Rouge-L falls within the bottom 25% of the dataset. Following this procedure, we construct SURE-OC for each LLM, recording its low-quality, low-uncertainty outputs under the stated selection rule. The statistics of SURE-OC QA and Math are provided in Appendix H.6 (Table 27).



A **B** **C**

Figure 8: Exploratory associations between (A) model size and performance, (B) performance and average uncertainty, and (C) performance and EUM. Higher performance is associated with lower average uncertainty ($p < 0.001$), while its negative association with EUM is marginal ($p \approx 0.05$).

Based on SURE-OC, we further analyze the overconfidence overlap across different LLMs. We compute this correlation for all LLM pairs, organize the results, and visualize it as a heatmap as Figure 7. B. The overlap within the same LLM family is slightly higher than that between LLMs from different families ($p < 0.005$). This indicates a moderate within-family association: **Same-family LLMs share more overconfident instances under this selection rule.** The analysis does not isolate the roles of shared training data, architectures, or alignment strategies. More detailed discussion about how to define and compute overlap is included in Appendix C.4.

5 Understanding LLMs Through the Lens of Semantic Uncertainty

5.1 Exploratory Scaling Trends in Uncertainty Quantification

The scaling law in LLMs suggests that model performance improves predictably as a function of compute, data volume, and parameter count Kaplan et al. (2020). However, **as model size increases and generation performance improves, how does the uncertainty in model responses change?** We first examine model size and performance on SUREBench (Figure 8. A). Within the evaluated families, log-parameter count correlates positively with average task performance; these observational comparisons do not isolate the effect of model size.

Model Performance and Average Uncertainty. Given these performance differences, we investigate whether higher-performing models exhibit different global uncertainty patterns. We analyze the performance of all models on SUREBench and their corresponding average uncertainty across these tasks. As shown in Figure 8. B, we observe a strong negative correlation between a model's average performance and its average uncertainty. This indicates that:

> **More powerful LLMs are generally more confident in their generations. They showed more consistent semantics in responses on average.**

Uncertainty on Low Quality Generations. To better understand uncertainty under errors, we analyze the uncertainty of low-quality generations. Following the construction of SURE-OC, we treat incorrect answers in QA and math, and bottom-25% ROUGE-L samples in summarization and translation, as low-quality generations. We use SNNE denom as the uncertainty score for this analysis. Given that different LLMs exhibit inherently different uncertainty scales (as illustrated in Figure 8.B), we define an Error Uncertainty Margin (EUM) to estimate the relative uncertainty on low-quality generations: EUM $= (\bar{u}_{low} - \bar{u}_{high})/|\bar{u}_{high}|$, $\bar{u}_{low}$ denotes the average uncertainty of the LLM on low-quality generations, while $\bar{u}_{high}$ represents the average uncertainty on high-quality generations. For QA and math, high-quality generations are defined as correct responses; for summarization and translation, as samples within the top 25% of ROUGE-L scores. A positive EUM indicates higher mean uncertainty on low-quality than high-quality generations. EUM measures relative mean

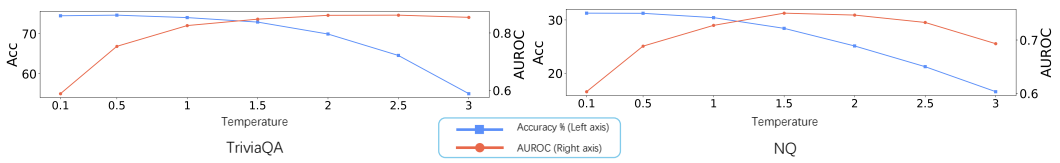


Figure 9: Joint effect of generation temperature on answer accuracy and uncertainty quality in TriviaQA and NQ. We plot Acc and DSE-based AUROC versus temperature for LLAMA-3.1 8B, enabling direct comparison of performance-uncertainty trade-offs under different decoding regimes.

separation; it does not by itself establish error-discrimination performance, probabilistic calibration, or the frequency of overconfident errors.

We plot the average performance and EUM of each LLM on SUREBench and compute the Pearson correlation coefficient, as shown in Figure 8. C. We observe that as model performance improves, EUM exhibits a decreasing trend ($p = 0.05$). We observe a marginal trend that:

Model performance is marginally negatively associated with the relative uncertainty gap between low- and high-quality outputs.

5.2 How does decoding temperature affect the observation of semantic uncertainty?

The sampling temperature controls the randomness of black-box LLM outputs. The choice of temperature has been widely discussed on LLM performance Li et al. (2025a); Renze (2024), creativity Peeperkorn et al. (2024), and clinical applications Patel et al. (2024). Specifically, we use LLAMA3.1-8B and LLAMA3.1-70B as the generation models on SUREBench with sampling temperatures $T \in \{0.1, 0.5, 1, 1.5, 2, 2.5, 3\}$. We adopt DSE Farquhar et al. (2024) as the SUQ method for this analysis. Figure 9 plots the generation performance and AUROC of Discrete Semantic Entropy on SUREBench QA under different temperature settings. As shown in Figure 9, a clear trade-off emerges as the temperature increases.

Increasing temperature can improve uncertainty-based error separation in QA while reducing answer accuracy; the trade-off is task- and model-dependent.

Based on this observation, we formulate temperature selection in UQ as a simple constrained optimization problem: given a retention threshold ρ, we seek the temperature τ^* that maximizes the AUROC of a UQ method while keeping the task accuracy above ρ fraction of its peak value, i.e.,

$$\tau^* = \arg\max_{\tau} \text{AUROC}(\tau) \quad \text{s.t.} \frac{\text{Acc}(\tau)}{\max_{\tau'} \text{Acc}(\tau')} \geq \rho. \tag{3}$$

For LLAMA-3.1-8B on the evaluated temperature grid, under a strict constraint ($\rho = 0.99$), **the optimal temperature is $\tau^* = 1.0$ for TriviaQA, SQUAD, and BioASQ, and $\tau^* = 0.5$ for CoQA and NQ**. These are descriptive optima within this sweep, not universal deployment recommendations. The experimental results and discussion about other SUREBench tasks are provided in Appendix C.5.

6 Conclusion

SURE characterizes the strengths and limitations of black-box semantic uncertainty across tasks, metrics, and output languages. Method rankings are task-dependent, and simple consistency measures remain competitive in several settings. SURE-OC reveals that stronger models can assign lower uncertainty to remaining errors, while QA temperature sweeps expose a trade-off between accuracy and uncertainty separability. Reliable evaluation must therefore account for task representation and quality criteria.

AI USE STATEMENT

Generative AI tools were used for writing, editing, and visual presentation. AI Agents assisted with manuscript formatting, organizing supplementary experiments, and drafting summaries of experimental results. All numerical results are taken from the experimental materials. The authors are responsible for the final manuscript and its scientific claims.

REPRODUCIBILITY STATEMENT

The anonymous repository linked in the introduction provides the code for reproducing this study. Appendix B.3 describes all evaluated datasets and translation directions, Appendix B.4 describes the UQ methods, and Table 4 lists the prompts. Appendices D–F document multilingual results and task-specific quality evaluation. Appendix B specifies the selection criteria, reusable interfaces, and ranking procedure; Appendices E and G report task-adapted and additional-model checks.

ETHICS STATEMENT

This work is an offline empirical study of existing language models on established datasets. To the best of our knowledge, it introduces no additional study-specific ethical concerns beyond those associated with the underlying datasets and models. Existing data and generated outputs may contain biases or inaccurate or harmful content. Our results concern uncertainty estimation under the evaluated protocols and should not be interpreted as a guarantee of correctness or safety in downstream applications.

REFERENCES

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, 2019. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301/.

Debarpan Bhattacharya, Malay Phadke, and Sriram Ganapathy. BiG-SURE: Bipartite graph for semantic uncertainty and reliability estimation of LLMs. *arXiv preprint arXiv:2608.30646*, 2026. URL https://arxiv.org/abs/2608.30646.

Dylan Bouchard, Mohit Singh Chauhan, David Skarbrevik, Ho-Kyeong Ra, Viren Bajaj, and Zeya Ahmad. Uqlm: A python package for uncertainty quantification in large language models. *Journal of Machine Learning Research*, 27(13):1–10, 2026.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5186–5200, 2024.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9397–9413, 2023. doi: 10.18653/v1/2023.emnlp-main.584. URL https://aclanthology.org/2023.emnlp-main.584/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 530–543, 2023.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*, 2025. URL https://arxiv.org/abs/2502.12404.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model, 2019.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

David Farr, Nico Manzonelli, Iain Cruickshank, and Jevin West. Red-ct: A systems design methodology for using llm-labeled data to train and deploy edge linguistic classifiers. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 58–67, 2025.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

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.

Mandar Joshi, Eunsol Choi, Daniel S 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, 2017.

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.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

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.

Lucie Kunitomo-Jacquin, Edison Marrese-Taylor, Ken Fukuda, and Masahiro Hamasaki. Evidential semantic entropy for llm uncertainty quantification. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7107–7122, 2026.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Lujun Li, Lama Sleem, Geoffrey Nichil, Radu State, et al. Exploring the impact of temperature on large language models: Hot or cold? *Procedia Computer Science*, 264:242–251, 2025a.

Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, and Zhifang Sui. Towards harmonized uncertainty estimation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22938–22953, 2025b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 4160–4173, 2022.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329, 2023.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Potsawee Manakul, Adian Liusie, and Mark 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.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1797–1807, 2018.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4530–4540, 2025.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Satwik Pandey, Suresh Raghu, and Shashwat Pandey. Selfdoubt: Uncertainty quantification for reasoning llms via the hedge-to-verify ratio. *arXiv preprint arXiv:2604.06389*, 2026.

Ji Won Park and Kyunghyun Cho. Efficient semantic uncertainty quantification in language models via diversity-steered sampling. *arXiv preprint arXiv:2510.21310*, 2025.

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, 2021.

Dhavalkumar Patel, Prem Timsina, Ganesh Raut, Robert Freeman, Matthew A levin, Girish N Nadkarni, Benjamin S Glicksberg, and Eyal Klang. Exploring temperature effects on large language models across various clinical tasks. *medRxiv*, pp. 2024–07, 2024.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*, 2024.

Edward Phillips, Sean Wu, Fredrik K Gustafsson, Boyan Gao, and David A Clifton. Semantic self-distillation for language model uncertainty. *arXiv preprint arXiv:2602.04577*, 2026.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 2685–2702, 2020. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213/.

Matthew Renze. The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 7346–7356, 2024.

Daniel Scalena, Leonidas Zotos, Elisabetta Fersini, Malvina Nissim, and Ahmet Üstün. Eager: Entropy-aware generation for adaptive inference-time scaling. *arXiv preprint arXiv:2510.11170*, 2025.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1099. URL https://www.aclweb.org/anthology/P17-1099.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38, 2025.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, i'm wrong: Llms hallucinate with certainty despite knowing the answer. *arXiv preprint arXiv:2502.12964*, 2025.

Ieva Staliunaite, Julius Cheng, and Andreas Vlachos. Uncertainty quantification for evaluating gender bias in machine translation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2204–2225. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.116. URL https://aclanthology.org/2026.findings-eacl.116/.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33117–33125, 2026.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138, 2015.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, et al. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025a.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Cocoa: A minimum bayes risk framework bridging confidence and consistency for uncertainty quantification in llms. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

Yongjin Yang, Haneul Yoo, and Hwaran Lee. Maqa: Evaluating uncertainty quantification in llms regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5846–5863, 2025.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385, 2024.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, 2020. doi: 10.18653/v1/2020.acl-main.148. URL https://aclanthology.org/2020.acl-main.148/.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. Luq: Long-text uncertainty quantification for llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5244–5262, 2024.

Rui Zhang and Joel Tetreault. This email could save your life: Introducing the task of email subject line generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 446–456, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1043. URL https://aclanthology.org/P19-1043.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

APPENDIX CONTENTS

A Limitations **16**

B Study Design, Reproducibility, and Task-Level Comparison **16**

B.1 Selection Principles and Scope 16

B.2 Prior-Work Task Coverage 17

B.3 Datasets Details 17

B.4 Methods Introduction 18

B.5 Controlled Evaluation and Reusable Interfaces 19

B.6 Details of Generation and Evaluation 20

C Experiments Results and Additional Discussion **21**

C.1 LLMs Performance and SUQ Evaluation on SUREBench 22

C.2 Experiments with different NLI backbone 23

C.3 Output Length and UQ Performance. 25

C.4 Detailed Discussion of LLM overconfidence. 25

C.5 Experiments with different temperatures during generation. 26

C.6 The relationship of Thinking and Uncertainty? 27

D Multilingual Translation Evaluation **27**

E Task-Adaptive Final-Answer Grouping for Mathematics **29**

F Rationale for ROUGE-L and Task-Specific Quality Evaluation **30**

F.1 Agreement of Model-Level Translation Quality Metrics 32

G Additional Model Evaluation on Sampled Inputs **32**

H Complete Experimental Results **35**

Reading guide. From page 35 onward, the appendix consists entirely of experimental results; we strongly recommend browsing them interactively through the anonymous repository at `https://anonymous.4open.science/r/SUREBench-CA71/index.html`. For multilingual comparisons, see Appendix D; for quality-metric checks, Appendix F; for task-level summaries and study scope, Appendix B; and for overconfidence analysis, Appendix C.4. Figures and standalone comparisons accompany the analysis. Study scope, prior-work coverage, datasets, methods, and generation/evaluation details are grouped in Appendix B. Complete model-by-dataset tables are collected in Appendix H, with a linked index. Section titles and page numbers above are clickable.

A LIMITATIONS

Despite covering four representative natural language generation tasks, 19 English-output datasets, and 14 translation directions with non-English targets, SURE still represents only a snapshot of the broader landscape of open-ended generation tasks. Other natural language generation settings, such as multi-hop reasoning, remain unexplored. We hope future work can extend SURE to investigate the effectiveness of UQ methods on a broader and more diverse range of open-ended generation scenarios.

The 13-model English-output evaluation is selected to study scale, within-family comparisons, and generation settings. Appendix G adds targeted checks with GPT-5.3-Chat and DeepSeek-V4-Flash using 100 sampled inputs per dataset; this does not provide exhaustive model coverage or full-suite evaluation of the additional models. The study evaluates 19 English-output datasets with 13 models and 14 non-English-target translation directions with 12 open-weight models. Non-English QA, reasoning, and summarization are not evaluated. Multilingual translation is evaluated at sentence level; document-level translation remains future work. Direction-level COMET quality is reported for six models, while the complete 19-measure COMET-based multilingual UQ tables cover Llama 3.1-8B; these scopes should not be conflated.

Finally, as discussed throughout the paper, many UQ methods require additional assumptions or interventions, such as white-box model access or prompt modifications. To ensure fair comparisons and control experimental variables, our study focuses on uncertainty estimation under a fixed-generation setting. The results do not establish rankings for white-box methods, which may use token probabilities or internal representations. Such methods require a separate protocol with explicit information-access assumptions. Task selection and task weights can also affect overall rankings; dataset-level and task-family results should accompany the aggregate score.

B STUDY DESIGN, REPRODUCIBILITY, AND TASK-LEVEL COMPARISON

Selection principles, prior-work coverage, datasets, implemented methods, reproducibility, and aggregation are presented together below.

B.1 SELECTION PRINCIPLES AND SCOPE

Tasks. SURE evaluates factual QA, exact-answer reasoning, abstractive summarization, and translation. Translation comprises eight into-English directions from WMT19, six non-English-to-non-English directions from OPUS-100, and eight English-to-multilingual directions from WMT24++. This design examines both task–representation mismatch and semantic uncertainty across output languages. Table 1 places these settings alongside prior studies; Appendix B.3 specifies the datasets and sample counts. Comparisons share generation evidence within each model–dataset setting. Failure of full-response NLI on numerical answers or partial semantic overlap does not establish failure of every semantic uncertainty representation.

Methods. The evaluated variants satisfy four criteria: they require no prompt modification; estimate uncertainty from generated outputs without generator internals; have a public implementation and sufficient information for reproduction; and follow the output-semantic uncertainty paradigm. These criteria define an eligible comparison set, not an exhaustive survey or a performance-based selection. The 19 reported measures include variants from six papers; access to an external NLI model's scores remains compatible with black-box access to the generator.

Models. The English-output evaluation spans 1.7B–405B open-weight models plus an API model; non-English-target translation evaluates twelve open-weight models over the same size range. Multiple sizes within Llama-3.1, Qwen3, Phi, and Mistral support within-family comparisons, while the selected models also support analyses of temperature and generation behavior. This selection serves the research questions rather than exhaustive coverage. Appendix G supplies targeted checks with two additional API models.

B.2 Prior-Work Task Coverage

Table 1: Representative task and language coverage in prior work and SURE. A dash indicates no evaluation in the listed category in that study. Non-English coverage refers to output languages; language sets can differ across datasets. This is a scope comparison, not the eligibility list for our fixed-generation method comparison.

Study	QA	Math	Summ.	Translation	Non-English coverage
Graph-based UQ (Lin et al., 2023)	TriviaQA, CoQA, NQ	–	–	–	–
DSE (Farquhar et al., 2024)	BioASQ, SQuAD, TriviaQA, NQ	SVAMP	–	–	–
KLE (Nikitin et al., 2024)	BioASQ, SQuAD, TriviaQA, NQ	SVAMP	–	–	–
Semantic Density (Qiu & Miikkulainen, 2024)	CoQA, TriviaQA, NQ, SciQ	–	DUC-2004	–	–
LM-Polygraph (Vashurin et al., 2025a)	TriviaQA, CoQA, MMLU	GSM8K	XSum	WMT14 fr–en; WMT19 de–en	–
SNNE (Nguyen et al., 2025)	SQuAD, TriviaQA, NQ, BioASQ	SVAMP	XSum, AESLC	WMT14 de–en, fr–en	–
CoCoA (Vashurin et al., 2025b)	TriviaQA, CoQA, MMLU	GSM8K	XSum	WMT14 fr–en; WMT19 de–en	–
BiG-SURE (Bhattacharya et al., 2026)	TriviaQA; SciQ; OK-VQA (vision)	SVAMP	–	–	SciQ in fr, ja, zh (also en)
Gender-bias UQ (Staliunaite et al., 2026)	–	–	–	WINOMT; MGENTE	en→es, fr, uk, ru, it, de, el; gender ambiguity
SURE (Ours)	TriviaQA, SQuAD, BioASQ, CoQA, Natural Questions (NQ)	SVAMP, GSM8K	XSum, AESLC, CNN/ DailyMail, Multi-News	WMT19; OPUS-100; WMT24++	6 non-English→non-English and 8 English→multilingual directions

BiG-SURE studies multilingual QA, while gender-bias UQ uses semantic uncertainty to analyze gender ambiguity in translation. SURE compares 19 existing semantic uncertainty measures and variants for generation-quality ranking across translation directions, alongside QA, mathematics, and summarization.

B.3 Datasets Details

Table 2 lists the evaluated inputs across QA, mathematics, summarization, and translation. Translation covers into-English, non-English-to-non-English, and English-to-multilingual generation.

Question answering. We use TriviaQA's Wikipedia-dev split (Joshi et al., 2017), SQuAD's dev-v2.0 split (Rajpurkar et al., 2016), Natural Questions validation data (Kwiatkowski et al., 2019), CoQA's coqa-dev-v1.0 split (Reddy et al., 2019), and BioASQ Task 12B golden data (12B1–12B4) (Tsatsaronis et al., 2015). These datasets cover general knowledge, conversational questions, and biomedical questions; inputs follow the closed-book QA formulation in Section 3.1.

Mathematical reasoning. SVAMP (Patel et al., 2021) tests arithmetic word problems, while GSM8K (Cobbe et al., 2021) tests multi-step numerical reasoning. We use their test splits and the evaluated counts in Table 2.

Summarization. We use the test splits of XSum (Narayan et al., 2018), AESLC (Zhang & Tetreault, 2019), CNN/DailyMail v3.0.0 (See et al., 2017), and Multi-News (Fabbri et al., 2019). They cover short news summaries, email subject generation, multi-sentence news summaries, and multi-document summaries, respectively.

Translation into English. WMT19 (Barrault et al., 2019) supplies eight validation subsets with Czech, German, Finnish, Gujarati, Kazakh, Lithuanian, Russian, and Chinese sources and English targets.

Non-English to non-English translation. OPUS-100 (Zhang et al., 2020) supplies six directions: ar–de, ar–fr, ar–zh, de–fr, de–zh, and fr–zh, with 2,000 inputs each. These directions compare non-English generations across Arabic, German, French, and Chinese source/target combinations.

English to multilingual translation. WMT24++ (Deutsch et al., 2025) supplies eight English-source directions with 960 valid inputs each. Targets are Egyptian Arabic (ar_EG), German (de_DE), Mexican Spanish (es_MX), French (fr_FR), Hindi (hi_IN), Japanese (ja_JP), Russian (ru_RU), and Chinese (zh_CN). The 14 OPUS-100/WMT24++ directions contain 19,680 inputs and 196,800 sampled translations per model.

Table 2: SUREBench dataset statistics. Counts denote evaluated inputs; each translation direction is listed separately.

Dataset / direction	Task	Inputs	Dataset / direction	Task	Inputs
TriviaQA	Factoid QA	7,993	OPUS-100 Arabic→German	Translation	2,000
Squad	Factoid QA	5,928	OPUS-100 Arabic→French	Translation	2,000
Natural Question	Factoid QA	4,289	OPUS-100 Arabic→Chinese	Translation	2,000
CoQA	Factoid QA	7,983	OPUS-100 German→French	Translation	2,000
BioASQ	Factoid QA	340	OPUS-100 German→Chinese	Translation	2,000
SVAMP	Math	300	OPUS-100 French→Chinese	Translation	2,000
GSM8K	Math	1,319	WMT24++ English→Egyptian Arabic	Translation	960
XSUM	Summarization	11,300	WMT24++ English→German (Germany)	Translation	960
AESLC	Summarization	1,906	WMT24++ English→Spanish (Mexico)	Translation	960
CNN/DailyMail	Summarization	11,490	WMT24++ English→French (France)	Translation	960
MultiNews	Summarization	5,622	WMT24++ English→Hindi (India)	Translation	960
WMT-19 Czech	Translation	2,983	WMT24++ English→Japanese (Japan)	Translation	960
WMT-19 German	Translation	2,998	WMT24++ English→Russian (Russia)	Translation	960
WMT-19 Finnish	Translation	3,000	WMT24++ English→Chinese (China)	Translation	960
WMT-19 Gujarati	Translation	1,998			
WMT-19 Kazakh	Translation	2,066			
WMT-19 Lithuanian	Translation	2,000			
WMT-19 Russian	Translation	3,000			
WMT-19 Chinese	Translation	3,981			

B.4 Methods Introduction

The six SUQ studies and their core ideas are summarized in Table 3. Methods share sampled responses and cached NLI evidence. English-output experiments use DeBERTa-large-MNLI; multilingual translation uses mDeBERTa. Continuous-score and discrete-label variants isolate the effect of semantic-signal granularity.

Implementation details. DSE uses `strict_entailment=False`. Graph-based UQ includes eigenvalue, degree, and eccentricity measures; the eccentricity threshold is 0.9. KLE uses heat kernels with $t = 0.1$ and Matérn kernels with $\kappa = \nu = 1$. SNNE uses softmax temperature 1.0, with full,

numerator-only, and denominator-only scores. LUQ-Pair averages $1 - \max_{j \neq i} s_{ij}$ over responses. For Semantic Density, we use the authors' black-box variant. Reported continuous/discrete variants follow the corresponding result-table labels.

Consistency baselines. We compare Exact Match after text normalization (Cole et al., 2023; Chen & Mueller, 2024), BERT similarity, and Sentence-BERT cosine similarity (Shorinwa et al., 2025; Bouchard et al., 2026). BERT similarity uses mean-pooled, L2-normalized `bert-base-uncased` sentence embeddings, not the token-level matching in standard BERTScore (Zhang et al., 2019). Sentence-BERT cosine similarity uses `all-MiniLM-L6-v2` embeddings. Both embedding scores are rescaled to $[0, 1]$. These are confidence scores; we convert them to uncertainty as $1 - \text{confidence}$ for evaluation.

Table 3: Comparison of covered black-box semantic uncertainty quantification (SUQ) methods in SURE. We summarize each method by its core idea.

Paper	Black-box SUQ	Year	Venue	# LLMs	# Datasets	Implementation signal	Core idea
(Lin et al., 2023)	U_{eigv}, U_{deg}	2023	TMLR	7	3	Continuous logits	Build a semantic graph over sampled answers and quantify uncertainty from its spectral structure.
(Farquhar et al., 2024)	Discrete Semantic Entropy	2024	Nature	6	5	Discrete label	Treat semantically equivalent responses as one meaning cluster, then measure how spread the generations are across distinct meanings.
(Nikitin et al., 2024)	KLE_{heat}, KLE_{matern}	2024	NeurIPS	7	5	Discrete label	Map the semantic graph into a kernel space and quantify how diffuse the graph is through kernel entropy. Different kernels emphasize different notions of smoothness and locality.
(Qiu & Mikkulainen, 2024)	Discrete Semantic Density	2024	NeurIPS	7	4	Discrete label	Measure how densely the probability mass is concentrated in semantic space. If likely responses are mutually similar, semantic density is high and uncertainty is low.
(Zhang et al., 2024)	LUQ-pair	2024	EMNLP	6	2	Continuous logits	For each sampled answer, look for its nearest semantic neighbor. If many answers have no close semantic counterpart, the response set is diverse, and uncertainty is high.
(Nguyen et al., 2025)	SNNE (full/only num / only denom)	2025	ACL Findings	2	9	Continuous logits	Compare within-cluster attraction against global similarity. A sample is more certain when it is semantically close to responses from the same meaning cluster and relatively separated from other clusters.

B.5 Controlled Evaluation and Reusable Interfaces

What is fixed. Within a model–dataset comparison, methods share prompts, sampled responses, response counts, and semantic evidence where applicable. This controls the evidence available to each estimator and isolates differences in uncertainty computation. It measures performance under a common protocol, not each method's independently tuned optimum. Backbone and signal-granularity ablations change an identified component while retaining the associated comparison inputs; generation-temperature experiments examine a different axis of variation.

Multilingual generation and scoring. For each of the twelve models, the 14 non-English-target directions use ten responses per input and shared multilingual mDeBERTa evidence. Fifteen UQ configurations yield 19 measures because graph and kernel methods expose multiple scores. The Llama-3.1-405B and other five-model comparisons also include three consistency baselines. Generators use local Hugging Face weights through vLLM. For ROUGE-L, we normalize all ten responses, select the unique majority response when present and otherwise the first response, and score against the reference. We use character tokens for Chinese/Japanese targets and whitespace tokens otherwise; whitespace-only scores remain available in the artifacts.

How to reuse the pipeline. The implementation separates answer generation, uncertainty computation, and evaluation. Answer artifacts contain a stable question identifier, the input, and its

sampled responses; cached NLI evidence can be reused by compatible estimators. A new estimator can consume these artifacts and return uncertainty scores aligned by identifier, allowing the evaluator to retain the same quality targets and sample set. A backbone study instead recomputes semantic evidence for the same responses. Dataset acquisition remains subject to the original access conditions, including the separately obtained BioASQ data. SURE-OC adds model-dependent, low-quality and low-uncertainty subsets for diagnosing failure modes without reconstructing the selection procedure for each study.

B.6 Details of Generation and Evaluation

For generations, we have used sampling-based decoding with a default temperature of 1.0 for all tested LLMs. The maximum number of newly generated tokens is set to 256 for each response. To support uncertainty estimation, we generate 10 responses for each question. We further adopt sampling with top_p=0.9, and enable sampling by default rather than greedy decoding. The prompts used in SURE and the prompts we used during NLI are shown in Table 4. For LLM generation, we conduct experiments on open-source models using a server equipped with 8×A100 GPUs, which is sufficient to support the computational requirements of our largest evaluated model, LLAMA-3.1 405B INSTRUCT. For closed-source models, we conduct experiments through API access.

As described in the main paper, we use Accuracy for QA and Math tasks, ROUGE-L (Lin, 2004) for Summarization and Translation tasks. About the UQ methods evaluation, we choose to report AUROC, AUARC, and PRR. For tasks such as QA and math, where each sample is associated with a short answer and a binary correctness label, we evaluate uncertainty estimation using **AUROC** and **AUARC**. These metrics assess whether uncertainty scores rank incorrect generations ahead of correct ones. In particular, **AUROC** measures how well an uncertainty score distinguishes correct from incorrect generations, treating UQ evaluation as a binary discrimination problem in which erroneous outputs should receive higher uncertainty. A higher AUROC indicates that the method is more effective at assigning larger uncertainty to incorrect generations.

Let the uncertainty score for sample i be u_i, and let the binary correctness label be $y_i \in \{0, 1\}$. Then, AUROC is defined as the area under the ROC curve:

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(r)\, d\,\mathrm{FPR}(r). \tag{4}$$

AUARC is defined as the area under the *accuracy-rejection curve* (ARC). It evaluates the average accuracy of the retained samples after rejecting those with the highest uncertainty, and summarizes this behavior across rejection thresholds. Therefore, a higher AUARC indicates that the uncertainty score is more useful for selective prediction, as it more effectively filters out unreliable outputs and improves the quality of the accepted ones (Farquhar et al., 2024; Lin et al., 2023).

Concretely, after sorting samples by uncertainty in ascending order, we compute the cumulative average accuracy of the top-k retained samples and integrate it over the retention coverage:

$$\mathrm{AUARC} = \int_0^1 \bar{y}(c)\, dc, \tag{5}$$

where $\bar{y}(c)$ denotes the average correctness of the retained samples at coverage level c.

For summarization and translation, each sample has a continuous quality score, such as **ROUGE-L**, rather than a binary correctness label. Following the rejection-based evaluation framework of prior work (Vashurin et al., 2025a; Nguyen et al., 2025), we measure the **area under the normalized-quality retention curve**, denoted PRR throughout our tables. Quality scores are min-max normalized before integration; we do not apply the random/oracle area normalization used in the standard prediction rejection ratio.

Let q_i denote the continuous quality score for sample i. We first apply min-max normalization:

$$\tilde{q}_i = \frac{q_i - \min(q)}{\max(q) - \min(q)}. \tag{6}$$

Table 4: Prompts used in SURE.

`QA Task`
System:
[QA task] Output a short answer with minimum words
User:
[TriviaQA/NQ/BioASQ] *{question}*
[CoQA] Story: *{story}*
Question: *{question}*
[SQuAD]: Context: *{context}*
Question: *{question}*
`Math task`
System:
[SVAMP] Only output a number answer
[GSM8k] Solve this math problem step by step and provide the final numerical answer after ####.
User:
[SVAMP] *{body}*
{question}
[GSM8k] *{question}*
`Summarization task`
System:
[XSUM] Summarize the given article in one short sentence.
[AESLC] Generate a concise email subject line.
[CNN-DailyMail] Summarize the given news article in 2-3 concise sentences.
[MultiNews] Summarize the following multi-document news in 2-3 concise sentences.
User:
[XSUM/MultiNews] *{document}*
[AESLC *{email body}*
[CNN-DailyMail] *{email body}*
`Translation task`
System:
Translate from {source lang} to {target lang}. Output only the translated text.
User:
{source text}
`NLI prompt in SUQ methods`
User:
We are evaluating answers to the question *{question}*
Possible Answer 1: *{text1}*
Possible Answer 2: *{text2}*
Does Possible Answer 1 semantically entail Possible Answer 2?
Respond with ONLY entailment, contradiction, or neutral.
`NLI in MNLI/SNLI/ANLI dataset`
User: Here are two statements:
Statement 1: *{text1}*
Statement 2: *{text2}*
Does Statement 1 semantically entail Statement 2? Respond with ONLY entailment, contradiction, or neutral.

We then sort samples by uncertainty u_i in ascending order, and denote the reordered normalized quality scores by $\tilde{q}_{(1)}, \tilde{q}_{(2)}, \ldots, \tilde{q}_{(n)}$. The reported retention-curve area is computed as:

$$\mathrm{PRR} = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{k} \tilde{q}_{(i)}. \tag{7}$$

Higher values indicate higher normalized quality among retained outputs. Comparisons between uncertainty methods use the same model, dataset, and quality target; values across different quality targets are not directly comparable.

C Experiments Results and Additional Discussion

This appendix keeps the analysis, figures, representative results, and standalone comparisons together. The complete model-by-dataset tables are collected by topic in Appendix H. Table 5 illustrates the QA/math comparison; it is an excerpt rather than the full model evaluation.

C.1 LLMs Performance and SUQ Evaluation on SUREBench

Since SUREBench covers a wide range of tasks, and uncertainty estimation requires LLMs to first generate responses before uncertainty scores can be computed, we first summarize the generation performance of all models in Table 16.

We present the complete SUQ results on SUREBench in six multipart tables in Appendix H.2 and Appendix H.3. Specifically, we use Tables 17, 18, and19 to report the performance of SUQ methods on QA and math tasks. Then, we report the performance of all evaluated methods on the summarization and translation tasks in Tables 20, 21, and 22. In these tables, we mark the methods implemented with discrete NLI signals using the prefix **Discrete**. Meanwhile, Tables 23 and 24 report lexical and embedding-based consistency baselines for the Llama-3.1 and Qwen3 models listed in those tables. And we plot the difference of switching to discrete NLI in Figure 10. The English-output comparisons broadly favor continuous variants, consistent with the main-text example. This pattern is conditional on the evaluated settings: the non-English-target suite instead has a discrete variant as its best macro-ranked SUQ measure.

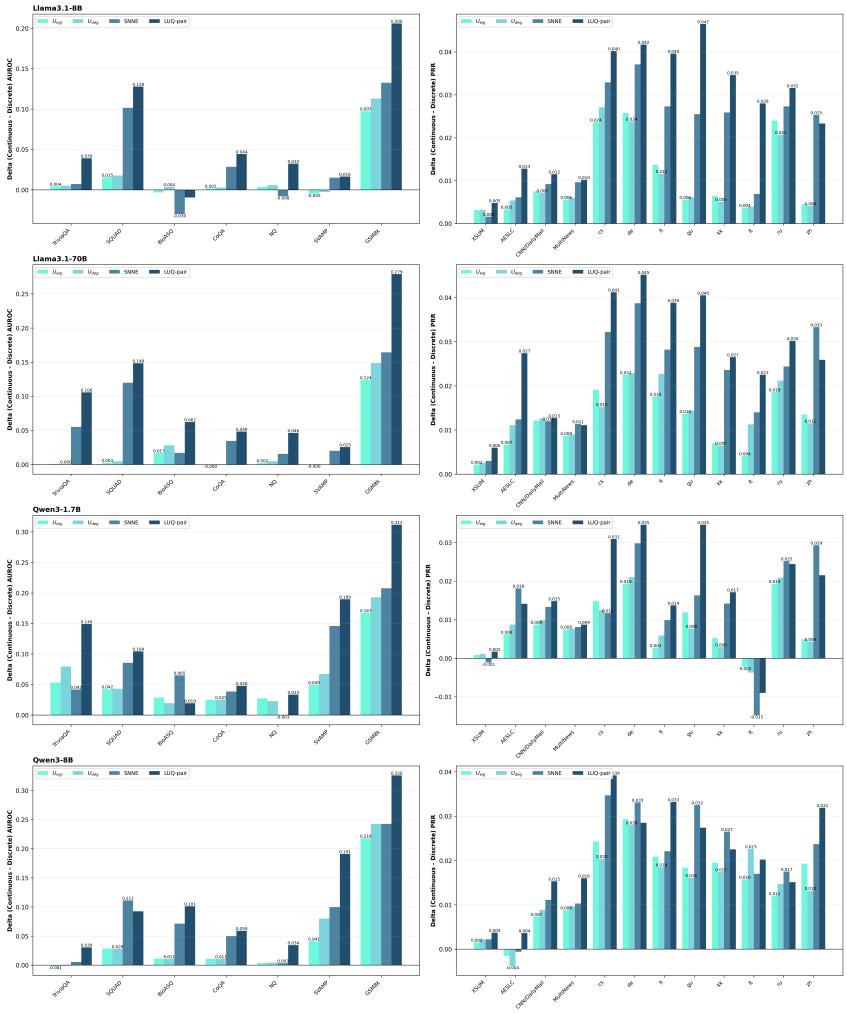


Figure 10: Original continuous vs. discrete SUQ on SUREBench with LLAMA3-8B/70B and QWEN3-1.7B/32B generation. Each bar shows the improvement of the continuous variant over the discrete variant.

Table 5: SUQ experimental results on QA and Math (Part I). We report AUROC/AUARC for QA and Math tasks. Representative excerpt: the first two generation models. Complete QA/math results are in Appendix H.2.

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT							
Discrete semantic entropy	0.8259/0.8914	0.7197/0.9350	0.7472/0.6916	0.6725/0.7502	0.7274/0.4611	0.7088/0.6376	0.5325/0.8281
Discrete semantic density	0.8462/0.8993	0.7254/0.9345	0.7081/0.6818	0.6844/0.7489	0.7248/0.4567	0.7287/0.6808	0.4456/0.7772
KLE-heat$_{t=0.1}$	0.8483/0.9019	0.7290/0.9366	0.7375/0.7039	0.6960/0.7579	0.7314/0.4614	0.7342/0.6818	0.5567/0.8276
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8487/0.8980	0.7297/0.9372	0.7181/0.6937	0.6999/0.7614	0.7295/0.4618	0.7300/0.6397	0.5587/0.8211
Discrete U_{eig}	0.8446/0.9019	0.7336/0.9386	0.6975/0.6849	0.7028/0.7617	0.7156/0.4526	0.7453/0.6558	0.5661/0.8370
Discrete U_{deg}	0.8494/0.9015	0.7294/0.9352	0.7181/0.6939	0.6991/0.7594	0.7253/0.4585	0.7405/0.6559	0.5577/0.8207
Discrete C_{ecc}	0.6387/0.7979	0.6933/0.9312	0.5428/0.5763	0.6172/0.7222	0.6151/0.3826	0.5325/0.5001	0.5580/ 0.8262
Discrete LUQ-pair	0.7540/0.8517	0.6103/0.8964	0.7900/0.7440	0.6443/0.7230	0.6986/0.4241	0.6711/0.6425	0.4988/0.8002
Discrete SNNE	0.8302/0.8897	0.6456/0.9193	0.7489/0.7022	0.6754/0.7396	0.7239/0.4404	0.7191/0.6292	0.5415/0.8140
Discrete SNNE (only_num)	0.8478/0.9044	0.7384/0.9421	0.7333/0.6647	0.6947/0.7638	0.7338/0.4548	0.7238/0.6305	0.7005/0.8932
Discrete SNNE (only_denom)	0.8247/0.8953	0.7196/0.9405	0.5978/0.5972	0.6621/0.7489	0.6797/0.4273	0.7138/0.6459	0.7005/0.8914
Semantic density	0.7989/0.8904	0.7601/0.9480	0.7333/0.6872	0.6898/0.7648	0.7108/0.4527	0.6798/0.6484	0.4925/0.7977
U_{eig}	0.8488/0.9028	0.7490/0.9432	0.6944/0.6754	0.7046/0.7695	0.7191/0.4574	0.7403/0.6599	0.6634/0.8891
U_{deg}	0.8544/0.9055	0.7470/0.9435	0.7217/0.6901	0.7019/0.7648	0.7310/0.4606	0.7376/0.6513	0.6706/ 0.8916
C_{ecc}	0.5996/0.7881	0.7127/0.9353	0.5314/0.5662	0.6359/0.7424	0.6041/0.3706	0.5000/0.4886	0.6182/0.8735
LUQ-pair	0.7930/0.8789	0.7381/0.9417	0.7806/0.7273	0.6886/0.7620	0.7310/0.4551	0.6875/0.6273	0.7047/0.8954
SNNE	0.8373/0.9022	0.7472/0.9421	0.7189/0.6885	0.7040/0.7672	0.7160/0.4547	0.7342/0.6627	0.6740/0.8925
SNNE (only_num)	0.8389/0.9024	0.7470/0.9433	0.7211/0.6895	0.7035/0.7689	0.7172/0.4555	0.7376/0.6637	0.6734/0.8923
SNNE (only_denom)	0.8535/0.9059	0.7473/0.9445	0.7300/0.6939	0.7017/0.7662	0.7345/0.4635	0.7338/0.6534	0.6727/0.8922
LLAMA 3.1-70B INSTRUCT							
Discrete semantic entropy	0.8073/0.9592	0.7654/0.9492	0.6335/0.7226	0.6774/0.7721	0.6971/0.5562	0.8194/0.9287	0.5085/0.9074
Discrete semantic density	0.8215/0.9620	0.7666/0.9491	0.6413/0.7245	0.6906/0.7798	0.6892/0.5597	0.8406/0.9304	0.5367/0.9258
Discrete U_{eig}	0.8218/0.9601	0.7686/0.9447	0.6961/0.7555	0.7040/0.7826	0.6884/0.5567	0.8550/0.9301	0.5490/0.9303
Discrete U_{deg}	0.8229/0.9603	0.7651/0.9454	0.6606/0.7340	0.7009/0.7818	0.6932/0.5604	0.8525/0.9301	0.5409/0.9281
Discrete C_{ecc}	0.7367/0.9477	0.7042/0.9428	0.6543/0.7089	0.6460/0.7557	0.6322/0.5244	0.7162/0.9003	0.5504/0.9301
KLE-heat$_{t=0.1}$	0.8216/0.9608	0.7655/0.9451	0.6497/0.7315	0.6989/0.7827	0.6969/0.5559	0.8467/0.9287	0.5404/0.9303
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8230/0.9592	0.7653/0.9482	0.6569/0.7329	0.7010/0.7831	0.6951/0.5612	0.8513/0.9300	0.5415/0.9238
Discrete LUQ-pair	0.6950/0.9420	0.6150/0.9160	0.6202/0.6823	0.6473/0.7415	0.6526/0.5187	0.7973/0.8993	0.4992/0.9103
Discrete SNNE	0.7647/0.9533	0.6521/0.9248	0.6825/0.7284	0.6675/0.7570	0.6767/0.5391	0.8321/0.9309	0.5320/0.9185
Discrete SNNE (only_num)	0.8021/0.9581	0.7573/0.9502	0.6719/0.7211	0.6808/0.7694	0.6903/0.5537	0.8445/0.9329	0.7211/0.9601
Discrete SNNE (only_denom)	0.7699/0.9521	0.7259/0.9447	0.5554/0.6585	0.6488/0.7566	0.6591/0.5372	0.8292/0.9294	0.7395/0.9621
Semantic density	0.7638/0.9504	0.8032/0.9608	0.6672/0.7308	0.6953/0.7806	0.6858/0.5552	0.8108/0.9267	0.5000/0.9246
U_{eig}	0.8228/0.9611	0.7713/0.9534	0.7128/0.7644	0.7037/0.7851	0.6906/0.5585	0.8549/0.9355	0.6728/0.9534
U_{deg}	0.8232/0.9610	0.7697/0.9519	0.6886/0.7550	0.7013/0.7842	0.6979/0.5590	0.8525/0.9350	0.6894/0.9568
C_{ecc}	0.6857/0.9311	0.7155/0.9433	0.6477/0.7170	0.6478/0.7606	0.6275/0.5151	0.6353/0.8221	0.6379/0.9463
LUQ-pair	0.8007/0.9585	0.7634/0.9528	0.6822/0.7481	0.6955/0.7848	0.6988/0.5651	0.8228/0.9115	0.7780/0.9696
SNNE	0.8198/0.9605	0.7718/0.9518	0.6995/0.7593	0.7020/0.7849	0.6921/0.5580	0.8521/0.9349	0.6962/0.9578
SNNE (only_num)	0.8203/0.9606	0.7714/0.9516	0.6961/0.7575	0.7019/0.7835	0.6924/0.5579	0.8520/0.9348	0.6948/0.9576
SNNE (only_denom)	0.8227/0.9609	0.7710/0.9529	0.6857/0.7527	0.7011/0.7843	0.6999/0.5599	0.8500/0.9327	0.6932/0.9573

C.2 EXPERIMENTS WITH DIFFERENT NLI BACKBONE

The complete numerical tables for this analysis are in Appendix H.5.

SUQ methods rely on semantic entailment judgment tools. We are interested in whether replacing DeBERTa with LLMs for this role would affect their performance. To this end, we conduct an additional experiment in which we use LLAMA3-8B, rather than DeBERTa, as the NLI classifier. Specifically, we replace DeBERTa with LLM-based NLI and implement a subset of SUQ methods under this setting. Since the LLM outputs discrete labels, we test discrete semantic entropy and discrete variants of SUQ methods that originally require fine-grained NLI signals. In addition, we evaluate the classification performance of the two NLI tools on NLI tasks using MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), and ANLI (Nie et al., 2020). We report their performance comparison, together with the results of the two DSE variants on SUREBench QA+Math, in Figure 11. We can find that (1) DeBERTa NLI outperforms Llama3 8B NLI across six NLI benchmarks; (2) On SUREBench QA/Math tasks, replacing the weaker NLI module with the stronger one yields upward DSE AUROC changes for 100% of tasks. (3) The weak-to-strong transition increases both average NLI accuracy and average downstream DSE AUROC, supporting a positive capability-performance association. This indicates that the **effectiveness of SUQ methods is closely tied to the capability of the underlying NLI tool.**

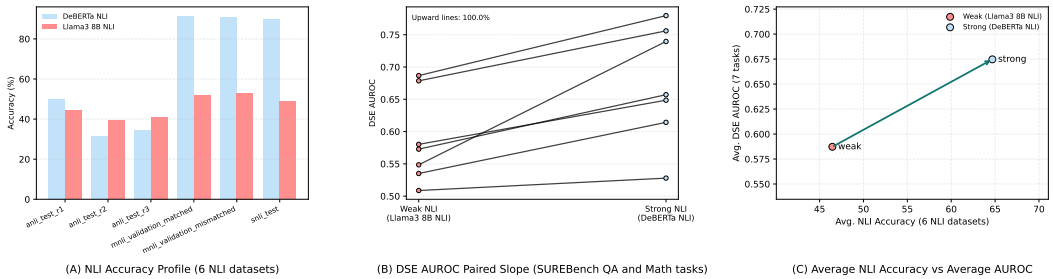


Figure 11: Stronger NLI capability is associated with improved downstream uncertainty estimation.

We show the complete experimental results of using LLAMA3.1-8B INSTRUCT as the NLI backbone in Table 25and 26. Figure 12 compares DeBERTa and LLAMA 3.1-8B as NLI tools on three generators. The observed differences are larger on QA/math than on summarization/translation. Partial semantic overlap in longer outputs is one candidate explanation discussed in the main text; this comparison does not establish that mechanism or a general ordering of NLI backends.

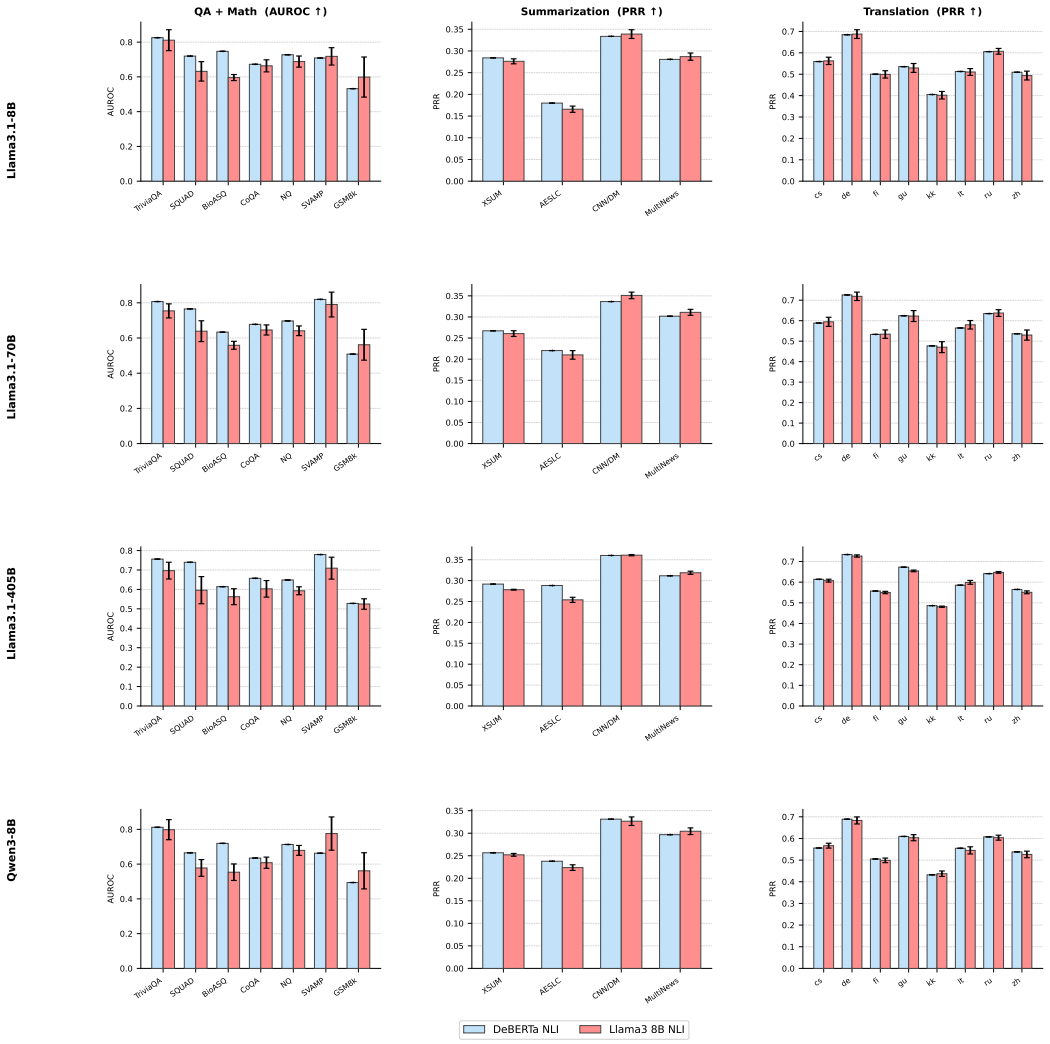


Figure 12: Comparison of using different NLI tools on SUREBench.

C.3 Output Length and UQ Performance.

We further analyze the relationship between average output length and the effectiveness of SUQ, using U_{deg} on generation from LLAMA 3.1-405B INSTRUCT FP8. Following our evaluation setup, we study this relationship separately for two groups of tasks on SUREBench: QA and math tasks, where UQ quality is measured by AUROC, and summarization and translation tasks, where it is measured by PRR.

For QA and math, we observe a clear cross-dataset negative correlation between average output length and AUROC. Specifically, **the Pearson correlation is -0.755 (p=0.017) and the Spearman correlation is -0.964**, indicating that datasets with longer average responses tend to exhibit worse UQ performance. In other words, SUQ is generally more effective on short-answer datasets such as TriviaQA and SQuAD, but becomes less reliable on long-generation reasoning datasets such as GSM8K. For summarization and translation, the overall mixed correlation between average output length and PRR is also negative, but much weaker and not statistically significant. On both summarization and translation tasks, the trend is not significant. This suggests that, unlike QA/math, output length alone is insufficient to explain SUQ performance on these long-form generation tasks. At the same time, within individual datasets, longer responses are usually associated with higher uncertainty values. This pattern holds for most QA/math datasets and also appears in the majority of summarization/translation datasets. Taken together, these results suggest a two-level phenomenon: within a task, longer outputs tend to be assigned higher uncertainty; however, across tasks, the overall effectiveness of UQ is jointly shaped by output length, task type, and answer format.

C.4 Detailed Discussion of LLM overconfidence.

The complete numerical tables for this analysis are in Appendix H.6.

We show the statistics of SURE-OC Summarization and Translation in Table 28. And the SURE-OC sets with $\alpha \in \{0.05, 0.10, 0.15, 0.30, 0.40\}$ are shown in Table 29, 30, 31, 32, 33, 34, 35, 36, 37, 38.

And we provide a more detailed discussion of the overlap of overconfidence below. To verify whether overconfident errors are more likely to be shared within the same model family, we conduct an additional significance analysis based on the adjusted Jaccard (AJ) similarity. Compared with the overlap coefficient used in the main text, this metric explicitly corrects for the overlap expected by random chance. For each dataset d, let S_i and S_j denote the overconfident error sets of two models i and j, respectively. We define the adjusted Jaccard similarity as:

$$\mathrm{AJ}_{ij}^{(d)} = \frac{|S_i \cap S_j| - \mathbb{E}[|S_i \cap S_j|]}{|S_i \cup S_j| - \mathbb{E}[|S_i \cap S_j|]}, \tag{8}$$

where the expected overlap under a random baseline is:

$$\mathbb{E}[|S_i \cap S_j|] = \frac{|S_i| \cdot |S_j|}{N_d}, \tag{9}$$

and N_d is the total number of samples in dataset d. Following the main paper, the overconfident set is defined as the subset of incorrect responses whose uncertainty lies within the lowest 20% quantile under SNNE denom. All adjusted Jaccard values are clipped into the range $[0, 1]$.

We then compute the family gap for each dataset:

$$\Delta_d = \overline{\mathrm{AJ}}_{\mathrm{intra},d} - \overline{\mathrm{AJ}}_{\mathrm{inter},d}, \tag{10}$$

where $\overline{\mathrm{AJ}}_{\mathrm{intra},d}$ and $\overline{\mathrm{AJ}}_{\mathrm{inter},d}$ denote the average adjusted Jaccard similarity for intra-family and inter-family model pairs, respectively. Table 6 reports the per-dataset statistics. Across all 19 datasets, **the average family gap is $+0.0147$, with 17 out of 19 datasets showing a positive gap**, indicating that overconfident errors overlap more strongly within model families than across different families. To assess statistical significance, we perform three complementary tests: a paired t-test, a one-sided sign test, and a sign-flip permutation test. The family gap is statistically significant ($t = 6.13$, sign test $p = 3.64 \times 10^{-4}$, permutation test $p = 2.0 \times 10^{-5}$). In addition, bootstrap estimation yields a 95% confidence interval of $[+0.0102, +0.0193]$, which excludes zero. The positive gap indicates moderately greater overlap within families under the stated selection rule. It does not identify whether architecture, training data, alignment, or other shared factors account for that association.

Table 6: Per-dataset LLM overconfidence overlap analysis using adjusted Jaccard (AJ) similarity. Positive Δ indicates higher overlap within model families than across different families.

Dataset	Intra-family AJ	Inter-family AJ	Δ
TriviaQA	0.0523	0.0444	+0.0079
SQUAD	0.0804	0.0573	+0.0231
BioASQ	0.0633	0.0375	+0.0258
CoQA	0.1252	0.0918	+0.0335
NQ	0.1468	0.1116	+0.0352
SVAMP	0.0212	0.0215	-0.0002
GSM8k	0.0503	0.0283	+0.0221
XSUM	0.1102	0.0890	+0.0211
AESLC	0.1027	0.0940	+0.0086
CNN/DailyMail	0.0854	0.0748	+0.0106
MultiNews	0.1152	0.0947	+0.0205
wmt19 cs	0.1387	0.1269	+0.0118
wmt19 de	0.1130	0.1038	+0.0092
wmt19 fi	0.1130	0.1017	+0.0112
wmt19 gu	0.0569	0.0494	+0.0075
wmt19 kk	0.0996	0.0862	+0.0134
wmt19 lt	0.0969	0.1028	-0.0059
wmt19 ru	0.1789	0.1665	+0.0124
wmt19 zh	0.1678	0.1566	+0.0113

C.5 Experiments with different temperatures during generation.

The complete numerical tables for this analysis are in Appendix H.7.

We use Llama 3.1-8B Instruct and Llama 3.1-70B Instruct to generate responses on SUREBench under different sampling temperatures and compute the corresponding SUQ scores. The complete experimental results are shown in Table 39 and Table 41. Figure 13 shows how generation performance, as well as AUROC and PRR of SUQ, change with increasing temperature on Llama 3.1-70B Instruct. We observe that once the temperature exceeds 1, the generation quality is significantly affected across almost all tasks. The impact of temperature on SUQ effectiveness varies by task type. We have already discussed its effect on QA and math in the main text. For summarization and translation tasks, SUQ performance also shows a declining trend as the temperature increases.

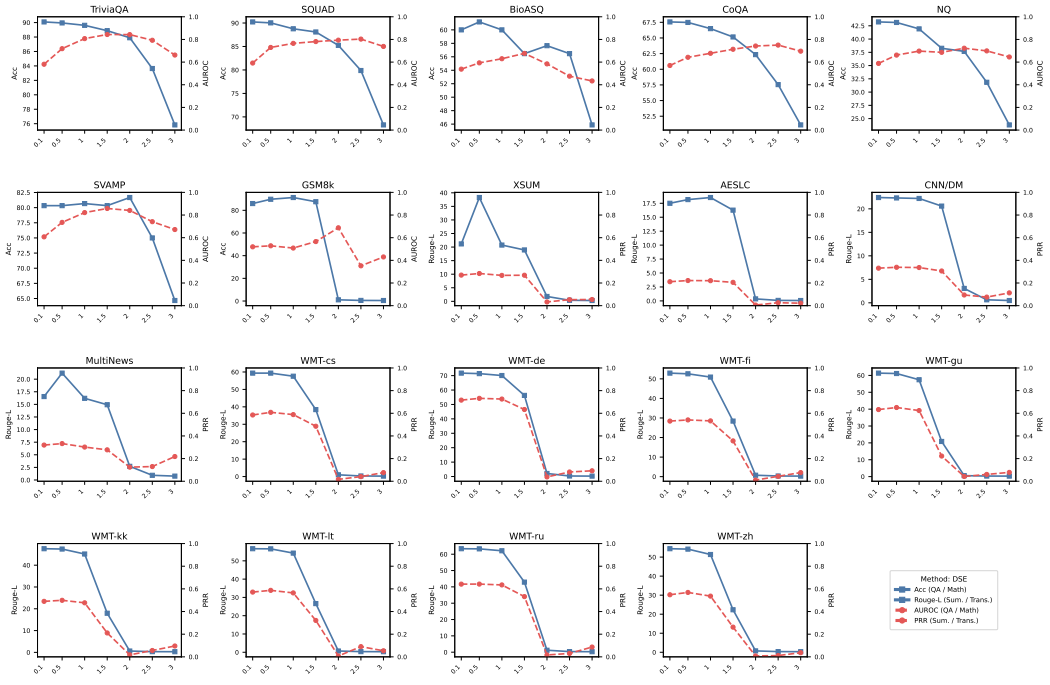


Figure 13: Joint effect of generation temperature on answer accuracy and uncertainty quality on SUREBench. All generations from Llama3.1-70B Instruct.

C.6 THE RELATIONSHIP OF THINKING AND UNCERTAINTY?

The complete numerical tables for this analysis are in Appendix H.8.

The introduction of thinking in LLM generation enables the model to explicitly reason before producing a final answer, which has been shown to improve performance on reasoning tasks (Wei et al., 2022). Prior work has also suggested that models with an explicit thinking process may produce final answers with fewer hallucinations (Lyu et al., 2023). Motivated by this, we investigate whether the thinking mechanism also affects uncertainty quantification in LLMs.

To this end, we conduct a group of experiments on SUREBench with QWEN3-8B, a model with native thinking capability, and evaluate it under both thinking-enabled and thinking-disabled settings. Figure 14 compares average AUROC and PRR across SUQ methods for thinking-enabled and thinking-disabled QWEN3-8B. The differences are limited on QA, summarization, and translation, but larger on math. Under the stated overconfidence criterion, the math comparison records 145 selected samples with thinking enabled versus 15 with thinking disabled, without a commensurate reduction in total errors. Thus, the thinking-enabled configuration has more low-uncertainty errors in this experiment. This is a model- and protocol-specific association; it does not establish a general effect of reasoning or explain the mechanism behind the difference.

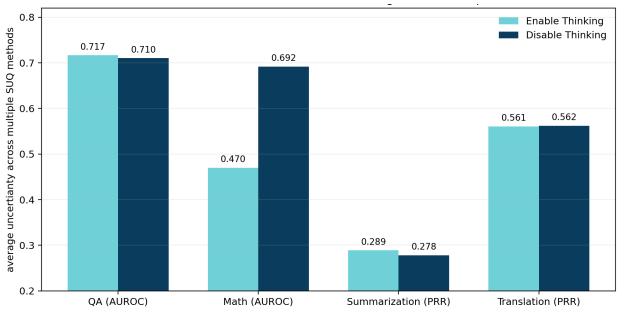


Figure 14: Uncertainty-based quality ranking for thinking-enabled and thinking-disabled QWEN3-8B. The largest differences in this comparison occur on mathematics.

D MULTILINGUAL TRANSLATION EVALUATION

Table 8 provides one model-level example near the discussion. The standalone model and consistency comparisons follow the discussion below; complete direction-level generation and SUQ tables are retained in Appendix H.9.

Task definitions and sample counts are in Appendix B.3; shared generation and scoring procedures are in Appendix B.

All 14 evaluated directions have non-constant per-example ROUGE-L, so their PRR values are informative. Tables 43 and 51 report model-by-direction generation results for all twelve models. Tables 44–50 and 52–56 report the complete model-by-method-by-direction PRR results for the same 19 uncertainty measures listed in Table 20. In particular, weighted heat and Matérn kernel entropy are read from the KLE outputs; continuous U_{eig} and U_{ecc} are read from the continuous-GWC outputs. Table 7 separately reports Exact Match, BERT similarity, and Cosine Similarity for LLAMA 3.1-405B INSTRUCT FP8. These methods produce confidence-style consistency scores, which we convert to uncertainty as $1 - \text{confidence}$ before applying the same continuous-quality PRR evaluation.

For the LLAMA 3.1-8B INSTRUCT baseline, the sample-weighted ROUGE-L is 34.21 (33.25 on OPUS-100 and 35.71 on WMT24++). Performance varies substantially across directions, from 10.33 on `en-ar_EG` to 55.58 on `en-es_MX`. Across the seven models in Table 43, overall ROUGE-L ranges from 34.21 to 46.83, while the best macro PRR among the 19 SUQ measures ranges from 0.4019 to 0.5222; LLAMA 3.1-405B INSTRUCT FP8 performs best on both measures. For UQ, we use the same continuous-quality PRR definition as in Appendix B.6. Among the 19 SUQ measures, Discrete SNNE (only-denom) achieves the best overall macro PRR for every model. On LLAMA

3.1-405B INSTRUCT FP8, Exact Match, BERT similarity, and Cosine Similarity obtain overall macro PRRs of 0.5001, 0.5184, and 0.5260, respectively; Cosine Similarity is numerically 0.0038 above the strongest SUQ result (0.5222). These multilingual results therefore support the main observation that UQ effectiveness is task- and direction-dependent, while providing comparative evidence on semantic uncertainty beyond English-output settings.

Table 7: Per-direction multilingual PRR of three traditional non-SUQ consistency baselines for LLAMA 3.1-405B INSTRUCT FP8. Each confidence score is converted to uncertainty as $1 - \text{confidence}$ and evaluated against the same per-example ROUGE-L used in Table 46. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Non-SUQ Baseline	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Exact Match	0.4567	0.5090	0.5232	0.5028	0.4040	0.5161	0.4853
BERT similarity	0.4589	0.5310	0.5453	0.5163	0.4297	0.5461	0.5045
Cosine Similarity	0.4712	0.5386	0.5547	0.5203	0.4194	0.5563	0.5101

WMT24++: English→multilingual									
Non-SUQ Baseline	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Exact Match	0.2627	0.5675	0.6553	0.6515	0.4042	0.5037	0.4337	0.6112	0.5112
BERT similarity	0.2537	0.5788	0.6985	0.6637	0.3984	0.5267	0.4756	0.6356	0.5289
Cosine Similarity	0.2683	0.5964	0.6953	0.6718	0.4091	0.5378	0.4837	0.6416	0.5380

Table 8: Per-direction multilingual PRR for LLAMA 3.1-8B INSTRUCT. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus. Representative model shown here; the complete multilingual record is in Appendix H.9.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.3499	0.4017	0.3516	0.4087	0.2918	0.3769	0.3634
Discrete semantic density	0.3707	0.4098	0.3616	0.4323	0.3012	0.3815	0.3762
Discrete U_{eig}	0.3673	0.4040	0.3548	0.4198	0.3024	0.3750	0.3706
Discrete U_{deg}	0.3639	0.4029	0.3584	0.4177	0.3061	0.3771	0.3710
Discrete C_{ecc}	0.3695	0.3993	0.3460	0.4304	0.3063	0.3744	0.3710
KLE-heat$_{t=0.1}$	0.3627	0.4003	0.3586	0.4220	0.3024	0.3786	0.3708
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.3647	0.4023	0.3580	0.4192	0.3037	0.3796	0.3712
Discrete LUQ-pair	0.3275	0.3804	0.3387	0.4059	0.2702	0.3325	0.3425
Discrete SNNE	0.3485	0.3972	0.3524	0.4058	0.2931	0.3616	0.3598
Discrete SNNE (only-num)	0.3947	0.4378	0.3695	0.4545	0.3104	0.3867	0.3922
Discrete SNNE (only-denom)	0.4010	0.4501	0.3600	0.4646	0.3082	0.3637	0.3913
Semantic density	0.3734	0.4154	0.3556	0.4339	0.3005	0.3769	0.3760
U_{eig}	0.3763	0.4197	0.3582	0.4383	0.3033	0.3789	0.3791
U_{deg}	0.3766	0.4173	0.3575	0.4304	0.3017	0.3784	0.3770
U_{ecc}	0.2719	0.3456	0.2849	0.3886	0.2198	0.2785	0.2982
LUQ	0.3747	0.4005	0.3573	0.4326	0.3066	0.3809	0.3754
SNNE	0.3754	0.4175	0.3576	0.4310	0.3004	0.3781	0.3767
SNNE (only-num)	0.3761	0.4168	0.3568	0.4331	0.3002	0.3779	0.3768
SNNE (only-denom)	0.3794	0.4179	0.3575	0.4352	0.3013	0.3785	0.3783

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.1122	0.4591	0.5474	0.5179	0.2974	0.2436	0.3332	0.4526	0.3704
Discrete semantic density	0.1163	0.4598	0.5613	0.5227	0.3029	0.2537	0.3452	0.4561	0.3772
Discrete U_{eig}	0.1174	0.4616	0.5605	0.5136	0.2988	0.2528	0.3428	0.4588	0.3758
Discrete U_{deg}	0.1167	0.4604	0.5600	0.5122	0.3017	0.2544	0.3421	0.4549	0.3753
Discrete C_{ecc}	0.1158	0.4640	0.5655	0.5089	0.3023	0.2374	0.3418	0.4650	0.3751
KLE-heat$_{t=0.1}$	0.1174	0.4676	0.5587	0.5099	0.2994	0.2561	0.3387	0.4555	0.3754
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1171	0.4648	0.5587	0.5169	0.3005	0.2548	0.3392	0.4565	0.3760
Discrete LUQ-pair	0.0976	0.4392	0.5351	0.5200	0.2782	0.1982	0.3192	0.4378	0.3532
Discrete SNNE	0.1086	0.4558	0.5480	0.5146	0.2868	0.2504	0.3212	0.4331	0.3648
Discrete SNNE (only-num)	0.1203	0.4944	0.5921	0.5469	0.3130	0.2432	0.3680	0.4594	0.3922
Discrete SNNE (only-denom)	0.1572	0.5288	0.6340	0.5861	0.3323	0.1681	0.4065	0.4667	0.4100
Semantic density	0.1221	0.4666	0.5642	0.5133	0.2989	0.2646	0.3512	0.4546	0.3795
U_{eig}	0.1223	0.4707	0.5622	0.5157	0.2984	0.2669	0.3534	0.4563	0.3807
U_{deg}	0.1238	0.4672	0.5633	0.5127	0.2983	0.2696	0.3508	0.4548	0.3801
U_{ecc}	0.1046	0.4422	0.5681	0.5318	0.2586	0.1301	0.3434	0.4439	0.3528
LUQ	0.1232	0.4579	0.5619	0.5229	0.2958	0.2658	0.3462	0.4579	0.3790
SNNE	0.1228	0.4679	0.5636	0.5137	0.2975	0.2697	0.3516	0.4549	0.3802
SNNE (only-num)	0.1231	0.4668	0.5633	0.5128	0.2975	0.2697	0.3508	0.4546	0.3798
SNNE (only-denom)	0.1243	0.4664	0.5628	0.5141	0.2984	0.2698	0.3502	0.4546	0.3801

E TASK-ADAPTIVE FINAL-ANSWER GROUPING FOR MATHEMATICS

Mathematical reasoning offers a direct test of representation choice: two different explanations can give the same numerical answer, while superficially similar explanations can end in different answers. We therefore extract and normalize final answers from the same sampled generations and group responses whose final answers match. This comparison evaluates final-answer agreement instead of exact matching of complete response strings, the baseline used in the main consistency comparison.

Table 9: AUROC on the same Llama-3.1-405B generations under two uncertainty formulations. Scores are shown to three decimal places. Final-answer grouping substantially improves GSM8K, but is lower on SVAMP.

Uncertainty formulation	GSM8K	SVAMP
SNNE denom	0.681	0.794
Final-answer exact matching	0.886	0.778

The result supports adapting semantic representations to the structure of the task, rather than assuming that NLI over full reasoning traces is universally appropriate. It does not show uniform gains: the direction of the effect differs between GSM8K and SVAMP. The comparison concerns one generator and does not establish a universal ranking across models.

Implications for SUQ design. The answer-grouping experiment complements the signal-granularity ablation in Section 4.1. Together they motivate preserving continuous semantic evidence while selecting the unit of comparison to match the task. For summaries and translations, partial overlap between long responses may matter more than strict equivalence; for exact-answer reasoning, the final answer is a natural unit. The length analysis in Appendix C.3 finds a significant association for QA/math but not summarization/translation, so it does not justify a universal length penalty. Likewise, the Llama-based NLI comparison in Appendix C.2 supports DeBERTa only within the evaluated settings. Multilingual outputs additionally require semantic backbones appropriate to their languages.

F Rationale for ROUGE-L and Task-Specific Quality Evaluation

Correlation and method-rank comparisons are kept with the metric rationale. Full quality scores and recalculated PRRs are collected in Appendix H.10.

Established use of ROUGE-L. Our choice of ROUGE-L follows established practice in uncertainty evaluation for language generation. In particular, SNNE (Nguyen et al., 2025) evaluates PRR using ROUGE-L as a quality target for both summarization and translation, directly motivating the same choice in SURE. LM-Polygraph (Vashurin et al., 2025a) also reports ROUGE-L for summarization, while Semantic Density (Qiu & Miikkulainen, 2024) uses a ROUGE-L-based correctness criterion for QA. These precedents make ROUGE-L a well-motivated reference metric for comparison with prior SUQ studies. Its use here maintains a consistent quality target across the base evaluation and subsequent analyses.

Task-specific neural quality metrics. We complement this lexical-overlap metric with two learned evaluators: COMET (Rei et al., 2020) uses multilingual representations of the source, translation, and reference to predict translation quality, while SEAHORSE-based evaluators (Clark et al., 2023) are trained on multilingual human ratings of summary quality. For summarization, we use the per-example arithmetic mean of SEAHORSE Q4/Q5/Q6, covering attribution, main ideas, and conciseness:

$$\text{SEAHORSE Average} = (Q4 + Q5 + Q6)/3.$$

To test whether our UQ comparisons depend on the quality metric, we evaluate all four summarization datasets with SEAHORSE Average and all eight WMT19 translation directions with reference-based COMET for the 13-model English-output evaluation. We retain the same generated outputs and uncertainty scores, align examples by sample ID, and recompute PRR using the original implementation. Thus the quality target changes while the uncertainty estimator and PRR computation remain fixed. Table 57 reports generation quality; Tables 61–63 provide the corresponding PRRs for all 19 SUQ measures, matching Tables 20–22. Table 64 likewise covers all five populated model blocks and three consistency baselines from Table 24, and Table 65 places all 22 measures together for Llama 3.1-405B.

The paired model-level analysis in Table 12 shows **strong agreement between ROUGE-L and COMET on WMT19**: each direction contains 13 observations, with Pearson correlations of 0.9671–

0.9956 and Spearman correlations of 0.9396–0.9945. The pooled Pearson correlation is 0.9678. This supports consistent model-level translation-quality comparisons despite the different numerical scales of the metrics. To examine the effect on uncertainty evaluation directly, Tables 10 and 11 compare method rankings under the old and new quality targets, holding the 13 models, 12 task columns, and nine original methods fixed. SNNE denom remains best on average under SEAHORSE/COMET, with a mean rank of 3.39 versus 2.97 under ROUGE-L, supporting the main aggregate comparison.

Table 10: Per-dataset mean ranks of the nine original SUQ methods on the four summarization datasets. Within each model–dataset cell, the nine methods are ranked by PRR (higher is better; ties receive average ranks), after which ranks are averaged over the same 13 models. Each dataset pairs its original ROUGE-L rank with its new SEAHORSE rank. Lower is better.

Method	XSUM		AESLC		CNN/DM		Multi-News	
	ROUGE-L	SEAHORSE	ROUGE-L	SEAHORSE	ROUGE-L	SEAHORSE	ROUGE-L	SEAHORSE
SNNE denom	3.08	4.38	2.31	4.46	3.08	4.69	3.73	5.08
U_{deg}	4.65	2.92	3.54	3.92	3.35	3.69	3.65	4.23
U_{eigv}	5.08	1.62	3.73	3.38	1.92	2.69	1.54	2.85
LUQ-pair	1.35	7.15	4.12	7.00	3.77	6.85	2.92	5.62
SNNE	3.12	4.62	2.50	4.38	2.88	5.69	3.15	4.62
KLE-heat	7.62	6.46	7.12	2.77	7.19	2.15	7.77	2.69
KLE-Matérn	8.46	5.46	6.88	2.92	7.27	2.23	7.35	2.92
DSE	4.04	8.46	7.31	8.69	7.42	8.92	7.00	8.38
SD	7.62	3.92	7.50	7.46	8.12	8.08	7.88	8.62

Table 11: Per-dataset mean ranks of the nine original SUQ methods on the eight WMT19 translation directions. Within each model–dataset cell, the nine methods are ranked by PRR (higher is better; ties receive average ranks), after which ranks are averaged over the same 13 models. Each dataset pairs its original ROUGE-L rank with its new COMET rank. Lower is better.

Method	cs		de		fi		gu	
	ROUGE-L	COMET	ROUGE-L	COMET	ROUGE-L	COMET	ROUGE-L	COMET
SNNE denom	3.27	2.85	2.81	2.38	3.65	2.46	2.46	2.54
U_{deg}	3.19	3.54	2.92	3.23	3.19	2.62	3.38	3.62
U_{eigv}	3.23	3.38	2.35	3.08	3.12	3.77	3.62	4.15
LUQ-pair	3.35	5.85	4.73	6.54	3.19	5.38	3.65	4.69
SNNE	3.19	3.00	3.35	3.69	3.77	3.69	4.35	4.23
KLE-heat	7.08	4.77	7.08	4.15	6.65	5.69	6.38	4.69
KLE-Matérn	7.73	4.77	7.04	4.92	7.35	6.08	7.08	5.00
DSE	7.00	9.00	7.50	8.92	6.46	7.62	6.54	8.38
SD	6.96	7.85	7.23	8.08	7.62	7.69	7.54	7.69
Method	kk		lt		ru		zh	
	ROUGE-L	COMET	ROUGE-L	COMET	ROUGE-L	COMET	ROUGE-L	COMET
SNNE denom	2.38	2.23	2.92	2.62	2.92	3.15	3.04	3.85
U_{deg}	3.50	3.31	3.50	3.46	2.96	3.62	3.04	4.92
U_{eigv}	3.15	4.08	4.54	4.69	2.35	2.54	2.58	5.54
LUQ-pair	3.38	4.46	2.58	2.23	3.31	4.92	2.96	4.46
SNNE	3.81	3.15	4.62	4.77	3.46	4.31	3.38	4.69
KLE-heat	6.00	5.46	6.35	5.85	7.46	5.00	7.35	1.77
KLE-Matérn	7.35	5.92	6.35	6.08	7.38	4.46	7.12	2.77
DSE	7.73	8.77	6.15	7.77	8.23	9.00	8.23	8.85
SD	7.69	7.62	8.00	7.54	6.92	8.00	7.31	8.15

We retain ROUGE-L for the primary analyses to preserve continuity with prior SUQ work and a common evaluation target across our experiments. Its established use, the WMT19 quality correlations, and the matched method comparisons provide empirical support for this choice. We retain the SEAHORSE- and COMET-based PRRs in Appendix H.10 so that readers can also assess the methods under task-specific quality targets and inspect metric-dependent differences.

Scope of the additional result tables. Beyond the 13-model comparison, Table 59 averages over all available results: 74 summarization model–dataset pairs (550,922 examples) and 144 WMT19 model–direction pairs. The translation inventory contains 158 pairs (416,148 examples) in total; the remaining 14 are the six OPUS-100 and eight WMT24++ directions for Llama 3.1-8B, whose complete 19-measure COMET-based PRRs appear in Table 60. Table 58 reports COMET generation quality for six multilingual models; these quality scores do not expand the model coverage of the multilingual PRR table. The paired ranking tables average within-model, within-dataset ranks over the same 13 models, whereas Table 59 averages PRR over the broader available inventory.

F.1 Agreement of Model-Level Translation Quality Metrics

Table 12 gives the direction-level correlations for the paired 13-model WMT19 evaluation discussed above. All reported correlations have $p < 0.001$; the aggregated within-direction rank correlation is 0.9732. These are correlations between model-level quality scores, distinct from sample-level quality agreement and from rankings of uncertainty estimators.

Table 12: Correlation between paired model-level ROUGE-L and COMET scores. Each direction contains 13 observations; all reported correlations have $p < 0.001$.

Direction	Pearson r	Spearman ρ
cs–en	0.9956	0.9890
de–en	0.9925	0.9396
fi–en	0.9890	0.9890
gu–en	0.9826	0.9670
kk–en	0.9814	0.9945
lt–en	0.9900	0.9835
ru–en	0.9782	0.9725
zh–en	0.9671	0.9505

G Additional Model Evaluation on Sampled Inputs

Per-dataset quality, the matched comparison with confidence intervals, and the quality–uncertainty correlations are shown with their corresponding analyses below.

Scope. We evaluate GPT-5.3-Chat and DeepSeek-V4-Flash with a sampling budget of 100 uniformly sampled inputs per dataset. Table 13 reports quality on all 19 base tasks, using the base quality measures: accuracy for QA/math and ROUGE-L for summarization/translation, on a 0–100 scale. These are targeted checks of model coverage and quality–uncertainty trends, not full-suite replacements for the 13-model evaluation.

Table 13: Additional-model quality with 100 sampled inputs per dataset. QA/math use accuracy ($\times 100$); summarization/translation use ROUGE-L ($\times 100$). WMT19 directions translate into English.

Task family	Dataset/direction	GPT-5.3-Chat	DeepSeek-V4-Flash
QA	TriviaQA	96.00	95.00
	SQuAD	84.75	89.83
	BioASQ	62.00	60.00
	CoQA	64.00	64.00
	NQ	59.65	50.88
Math	SVAMP	97.00	97.00
	GSM8K	99.00	99.00
Summarization	XSum	21.13	20.30
	AESLC	23.12	27.00
	CNN/DM	18.89	18.87
	Multi-News	15.20	14.14
Translation	cs–en	61.66	62.00
	de–en	76.74	76.52
	fi–en	52.17	52.51
	gu–en	62.88	61.67
	kk–en	48.12	46.75
	lt–en	49.97	52.46
	ru–en	65.90	66.49
	zh–en	56.05	57.23

Matched quality comparison. Table 14 compares GPT-5.3-Chat with GPT-4o-mini on matched inputs. GPT-5.3-Chat has a higher mean in each task family, with the largest difference on mathematics. The QA confidence interval includes zero, so the data do not establish an improvement for every task family. The 19-task macro is an equal-weight average of dataset quality scores on their displayed scales; it is a descriptive summary of this fixed suite, not a common unit of quality across accuracy and ROUGE-L.

Table 14: Matched quality comparison and paired-bootstrap 95% confidence intervals for the difference. Means and differences are displayed at the reported precision; subtracting displayed means can differ from the reported difference by 0.01.

Task group	GPT-5.3-Chat	GPT-4o-mini	Δ	95% CI
QA	73.28	71.03	$+2.25$	$[-1.03, +5.53]$
Math	98.00	65.00	$+33.00$	$[+26.50, +39.50]$
Summarization	19.59	18.40	$+1.18$	$[+0.29, +2.06]$
Translation	59.18	57.42	$+1.77$	$[+1.04, +2.51]$
19-task macro	58.64	53.58	$+5.06$	$[+3.89, +6.23]$

Quality–uncertainty association. We also extend the model-level correlation analysis to include both additional models. Table 15 distinguishes global mean uncertainty from mean uncertainty restricted to low-quality generations. Both correlations with model quality remain negative and increase in magnitude. The low-quality association changes from $r = -0.549$ ($p = 0.0519$) to $r = -0.666$ ($p = 0.0067$). These estimates concern absolute mean uncertainty on the selected outputs; they are not the EUM statistic, which normalizes against the high-quality group in Section 5.1. The small-sample extension supports the observed association without establishing a universal scaling law or a causal effect of model size.

Table 15: Pearson correlation between model quality and mean uncertainty, before and after including the two additional models. The low-quality row uses the mean over low-quality generations, not EUM.

	Base model set		Base + two models	
Uncertainty summary	r	p	r	p
Global mean	-0.838	0.00035	-0.871	0.000024
Mean on low-quality outputs	-0.549	0.0519	-0.666	0.0067

H COMPLETE EXPERIMENTAL RESULTS

This appendix collects the detailed model-by-dataset tables. Figures and standalone comparisons accompany their discussions in the preceding appendices. The links below lead to each group of records.

Results	**Location**
Base Generation Quality	H.1 (p. 36)
Complete SUQ Results on QA and Mathematics	H.2 (p. 37)
Complete SUQ Results on Summarization and Translation	H.3 (p. 42)
Consistency Baselines	H.4 (p. 47)
NLI-Backbone Comparisons	H.5 (p. 48)
Overconfidence Statistics	H.6 (p. 50)
Generation-Temperature Experiments	H.7 (p. 60)
Thinking-Enabled and Thinking-Disabled Generation	H.8 (p. 68)
Complete Multilingual Translation Results	H.9 (p. 70)
Task-Specific Quality Scores and Recomputed PRRs	H.10 (p. 84)

H.1 Base Generation Quality

Discussion: Appendix C.

Table 16: LLMs generation performance on SUREBench. For QA and Math tasks, we report the accuracy. For Summarization and Translation tasks, we report ROUGE-L. Panel: QA and math.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
Llama 3.1-8B Instruct	74.03	87.63	52.94	64.70	30.43	49.00	85.60
Llama 3.1-70B Instruct	89.62	88.77	60.00	66.49	41.92	80.67	91.28
Llama 3.1-405B Instruct FP8	92.04	89.02	58.24	68.87	46.63	85.33	93.93
Qwen 3-1.7B	45.90	82.64	46.47	58.88	14.83	90.33	61.49
Qwen 3-8B	58.40	84.36	49.41	62.57	22.15	92.33	62.47
Qwen 3-32B	72.30	87.58	54.12	66.60	30.08	91.33	70.05
Phi 3-14B	67.37	76.89	53.24	59.85	26.58	87.33	62.62
Phi4-mini-Instruct 3.8B	46.95	77.19	48.53	58.42	18.05	76.67	62.09
Phi 4-14B	74.54	84.72	55.59	65.75	31.90	62.63	55.50
Mistral-7B-Instruct-v0.2	74.13	82.51	52.06	62.53	29.96	46.67	47.99
Mistral-Nemo-Instruct-2407 12B	72.89	73.89	40.59	56.02	22.06	51.33	68.99
Mistral-Small-3.2-24B-Instruct-2506	81.16	81.11	53.24	60.60	27.89	72.33	65.50
GPT-4o-mini	85.07	85.58	58.82	65.01	39.78	81.67	52.16

Table 16: LLMs generation performance on SUREBench. For QA and Math tasks, we report the accuracy. For Summarization and Translation tasks, we report ROUGE-L. Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
Llama 3.1-8B Instruct	18.93	15.13	21.70	15.81
Llama 3.1-70B Instruct	20.74	18.53	22.21	22.21
Llama 3.1-405B Instruct FP8	22.12	16.13	22.80	16.11
Qwen 3-1.7B	14.87	22.17	21.22	14.04
Qwen 3-8B	16.74	19.70	21.35	14.10
Qwen 3-32B	17.63	19.87	21.22	14.56
Phi 3-14B	14.09	13.88	19.97	11.41
Phi4-mini-Instruct 3.8B	11.45	11.50	16.89	10.42
Phi 4-14B	14.55	8.26	19.94	14.38
Mistral-7B-Instruct-v0.2	15.11	15.13	21.41	15.76
Mistral-Nemo-Instruct-2407 12B	17.81	19.39	21.47	12.74
Mistral-Small-3.2-24B-Instruct-2506	19.10	21.08	22.19	15.32
GPT-4o-mini	17.49	22.30	21.03	14.89

Table 16: LLMs generation performance on SUREBench. For QA and Math tasks, we report the accuracy. For Summarization and Translation tasks, we report ROUGE-L. Panel: Translation.

Model / method	cs	de	fi	gu	kk	lt	ru	zh
Llama 3.1-8B Instruct	54.01	66.58	46.90	46.90	36.61	46.76	58.90	47.91
Llama 3.1-70B Instruct	57.58	70.07	50.91	57.43	45.13	54.23	62.11	51.36
Llama 3.1-405B Instruct FP8	59.88	71.95	53.95	63.27	46.90	59.37	64.03	55.14
Qwen 3-1.7B	48.17	61.16	37.65	43.99	30.83	38.76	54.37	47.47
Qwen 3-8B	55.28	66.91	48.19	57.30	41.44	53.06	59.48	52.17
Qwen 3-32B	57.00	68.25	50.53	60.72	43.78	54.92	60.77	53.31
Phi 3-14B	47.57	62.99	33.28	7.04	16.59	28.07	57.48	45.01
Phi4-mini-Instruct 3.8B	22.94	33.12	31.02	6.49	24.10	18.42	47.55	36.48
Phi 4-14B	54.23	66.55	48.41	49.05	37.28	49.02	59.58	49.85
Mistral-7B-Instruct-v0.2	50.73	62.51	37.79	18.25	19.09	33.16	55.50	43.48
Mistral-Nemo-Instruct-2407 12B	51.25	64.47	43.96	38.37	37.58	44.74	58.21	47.46
Mistral-Small-3.2-24B-Instruct-2506	56.87	69.57	50.38	57.30	44.23	52.91	61.99	52.39
GPT-4o-mini	58.54	70.89	53.14	62.15	47.80	58.47	62.94	53.69

H.2 COMPLETE SUQ RESULTS ON QA AND MATHEMATICS

Discussion: Appendix C.

Table 17: SUQ experimental results on QA and Math (Part I). We report AUROC/AUARC for QA and Math tasks.

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT							
Discrete semantic entropy	0.8259/0.8914	0.7197/0.9350	0.7472/0.6916	0.6725/0.7502	0.7274/0.4611	0.7088/0.6376	0.5325/0.8281
Discrete semantic density	0.8462/0.8993	0.7254/0.9345	0.7081/0.6818	0.6844/0.7489	0.7248/0.4567	0.7287/0.6808	0.4456/0.7772
KLE-heat$_{t=0.1}$	0.8483/0.9019	0.7290/0.9366	0.7375/0.7039	0.6960/0.7579	0.7314/0.4614	0.7342/0.6818	0.5567/0.8276
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8487/0.8980	0.7297/0.9372	0.7181/0.6937	0.6999/0.7614	0.7295/0.4618	0.7300/0.6397	0.5587/0.8211
Discrete U_{eig}	0.8446/0.9019	0.7336/0.9386	0.6975/0.6849	0.7028/0.7617	0.7156/0.4526	0.7453/0.6558	0.5661/0.8370
Discrete U_{deg}	0.8494/0.9015	0.7294/0.9352	0.7181/0.6939	0.6991/0.7594	0.7253/0.4585	0.7405/0.6559	0.5577/0.8207
Discrete C_{ecc}	0.6387/0.7979	0.6933/0.9312	0.5428/0.5763	0.6172/0.7222	0.6151/0.3826	0.5325/0.5001	0.5580/ 0.8262
Discrete LUQ-pair	0.7540/0.8517	0.6103/0.8964	0.7900/0.7440	0.6443/0.7230	0.6986/0.4241	0.6711/0.6425	0.4988/0.8002
Discrete SNNE	0.8302/0.8897	0.6456/0.9193	0.7489/0.7022	0.6754/0.7396	0.7239/0.4404	0.7191/0.6292	0.5415/0.8140
Discrete SNNE (only_num)	0.8478/0.9044	0.7384/0.9421	0.7333/0.6647	0.6947/0.7638	0.7338/0.4548	0.7238/0.6305	0.7005/0.8932
Discrete SNNE (only_denom)	0.8247/0.8953	0.7196/0.9405	0.5978/0.5972	0.6621/0.7489	0.6797/0.4273	0.7138/0.6459	0.7005/0.8914
Semantic density	0.7989/0.8904	0.7601/0.9480	0.7333/0.6872	0.6898/0.7648	0.7108/0.4527	0.6798/0.6484	0.4925/0.7977
U_{eig}	0.8488/0.9028	0.7490/0.9432	0.6944/0.6754	0.7046/0.7695	0.7191/0.4574	0.7403/0.6599	0.6634/0.8891
U_{deg}	0.8544/0.9055	0.7470/0.9435	0.7217/0.6901	0.7019/0.7648	0.7310/0.4606	0.7376/0.6513	0.6706/ 0.8916
C_{ecc}	0.5996/0.7881	0.7127/0.9353	0.5314/0.5662	0.6359/0.7424	0.6041/0.3706	0.5000/0.4886	0.6182/0.8735
LUQ-pair	0.7930/0.8789	0.7381/0.9417	0.7806/0.7273	0.6886/0.7620	0.7310/0.4551	0.6875/0.6273	0.7047/0.8954
SNNE	0.8373/0.9022	0.7472/0.9421	0.7189/0.6885	0.7040/0.7672	0.7160/0.4547	0.7342/0.6627	0.6740/0.8925
SNNE (only_num)	0.8389/0.9024	0.7470/0.9433	0.7211/0.6895	0.7035/0.7689	0.7172/0.4555	0.7376/0.6637	0.6734/0.8923
SNNE (only_denom)	0.8535/0.9059	0.7473/0.9445	0.7300/0.6939	0.7017/0.7662	0.7345/0.4635	0.7338/0.6534	0.6727/0.8922
LLAMA 3.1-70B INSTRUCT							
Discrete semantic entropy	0.8073/0.9592	0.7654/0.9492	0.6335/0.7226	0.6774/0.7721	0.6971/0.5562	0.8194/0.9287	0.5085/0.9074
Discrete semantic density	0.8215/0.9620	0.7666/0.9491	0.6413/0.7245	0.6906/0.7798	0.6892/0.5597	0.8406/0.9304	0.5367/0.9258
Discrete U_{eig}	0.8218/0.9601	0.7686/0.9447	0.6961/0.7555	0.7040/0.7826	0.6884/0.5567	0.8550/0.9301	0.5490/0.9303
Discrete U_{deg}	0.8229/0.9603	0.7651/0.9454	0.6606/0.7340	0.7009/0.7818	0.6932/0.5604	0.8525/0.9301	0.5409/0.9281
Discrete C_{ecc}	0.7367/0.9477	0.7042/0.9428	0.6543/0.7089	0.6460/0.7557	0.6322/0.5244	0.7162/0.9003	0.5504/0.9301
KLE-heat$_{t=0.1}$	0.8216/0.9608	0.7655/0.9451	0.6497/0.7315	0.6989/0.7827	0.6969/0.5559	0.8467/0.9287	0.5404/0.9303
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8230/0.9592	0.7653/0.9482	0.6569/0.7329	0.7010/0.7831	0.6951/0.5612	0.8513/0.9300	0.5415/0.9238
Discrete LUQ-pair	0.6950/0.9420	0.6150/0.9160	0.6202/0.6823	0.6473/0.7415	0.6526/0.5187	0.7973/0.8993	0.4992/0.9103
Discrete SNNE	0.7647/0.9533	0.6521/0.9248	0.6825/0.7284	0.6675/0.7570	0.6767/0.5391	0.8321/0.9309	0.5320/0.9185
Discrete SNNE (only_num)	0.8021/0.9581	0.7573/0.9502	0.6719/0.7211	0.6808/0.7694	0.6903/0.5537	0.8445/0.9329	0.7211/0.9601
Discrete SNNE (only_denom)	0.7699/0.9521	0.7259/0.9447	0.5554/0.6585	0.6488/0.7566	0.6591/0.5372	0.8292/0.9294	0.7395/0.9621
Semantic density	0.7638/0.9504	0.8032/0.9608	0.6672/0.7308	0.6953/0.7806	0.6858/0.5552	0.8108/0.9267	0.5000/0.9246
U_{eig}	0.8228/0.9611	0.7713/0.9534	0.7128/0.7644	0.7037/0.7851	0.6906/0.5585	0.8549/0.9355	0.6728/0.9534
U_{deg}	0.8232/0.9610	0.7697/0.9519	0.6886/0.7550	0.7013/0.7842	0.6979/0.5590	0.8525/0.9350	0.6894/0.9568
C_{ecc}	0.6857/0.9311	0.7155/0.9433	0.6477/0.7170	0.6478/0.7606	0.6275/0.5151	0.6353/0.8221	0.6379/0.9463
LUQ-pair	0.8007/0.9585	0.7634/0.9528	0.6822/0.7481	0.6955/0.7848	0.6988/0.5651	0.8228/0.9115	0.7780/0.9696
SNNE	0.8198/0.9605	0.7718/0.9518	0.6995/0.7593	0.7020/0.7849	0.6921/0.5580	0.8521/0.9349	0.6962/0.9578
SNNE (only_num)	0.8203/0.9606	0.7714/0.9516	0.6961/0.7575	0.7019/0.7835	0.6924/0.5579	0.8520/0.9348	0.6948/0.9576
SNNE (only_denom)	0.8227/0.9609	0.7710/0.9529	0.6857/0.7527	0.7011/0.7843	0.6999/0.5599	0.8500/0.9327	0.6932/0.9573

Table 17: SUQ experimental results on QA and Math (Part I). We report AUROC/AUARC for QA and Math tasks. (Continued.)

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
LLAMA 3.1-405B INSTRUCT FP8							
Discrete semantic entropy	0.7560/0.9638	0.7396/0.9381	0.6143/0.6860	0.6571/0.7742	0.6486/0.5695	0.7796/0.9243	0.5279/0.9207
Discrete semantic density	0.7643/0.9657	0.7377/0.9397	0.6318/0.6997	0.6634/0.7801	0.6401/0.5666	0.7999/0.9470	0.5481/0.9329
Discrete U_{eig}	0.7675/0.9661	0.7446/0.9407	0.6652/0.7202	0.6772/0.7841	0.6557/0.5716	0.8150/0.9490	0.5536/0.9240
Discrete U_{deg}	0.7661/0.9649	0.7405/0.9396	0.6514/0.7091	0.6713/0.7810	0.6515/0.5684	0.8101/0.9457	0.5502/0.9279
Discrete C_{ecc}	0.7155/0.9549	0.6840/0.9351	0.5353/0.6278	0.6179/0.7602	0.5920/0.5417	0.6622/0.9044	0.5536/0.9317
KLE-heat$_{t=0.1}$	0.7650/0.9648	0.7403/0.9389	0.6479/0.7072	0.6688/0.7812	0.6513/0.5686	0.8036/0.9409	0.5514/0.9290
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7663/0.9649	0.7411/0.9386	0.6520/0.7092	0.6712/0.7845	0.6518/0.5695	0.8049/0.9467	0.5512/0.9306
Discrete LUQ-pair	0.6373/0.9460	0.5863/0.9161	0.6005/0.6346	0.6131/0.7412	0.6077/0.5325	0.6632/0.9065	0.4992/0.9079
Discrete SNNE	0.6837/0.9531	0.6174/0.9213	0.6179/0.6694	0.6373/0.7516	0.6377/0.5520	0.7976/0.9281	0.5294/0.9100
Discrete SNNE (only_num)	0.7742/0.9656	0.7360/0.9470	0.6411/0.7059	0.6686/0.7866	0.6495/0.5738	0.7908/0.9294	0.5978/0.9361
Discrete SNNE (only_denom)	0.7646/0.9661	0.7198/0.9470	0.6488/0.6990	0.6476/0.7787	0.6081/0.5534	0.7689/0.9183	0.5800/0.9329
Semantic density	0.7688/0.9646	0.7735/0.9567	0.6065/0.6852	0.6943/0.7999	0.6511/0.5797	0.7579/0.9163	0.5000/0.9464
U_{eig}	0.7875/0.9665	0.7584/0.9536	0.6760/0.7286	0.6820/0.7919	0.6620/0.5851	0.8042/0.9340	0.6601/0.9676
U_{deg}	0.7865/0.9683	0.7567/0.9503	0.6750/0.7271	0.6779/0.7889	0.6618/0.5835	0.7971/0.9132	0.6768/0.9695
C_{ecc}	0.7241/0.9591	0.7099/0.9395	0.5398/0.6221	0.6341/0.7764	0.5990/0.5570	0.6229/0.8920	0.6254/0.9623
LUQ-pair	0.7574/0.9646	0.7330/0.9356	0.6548/0.7079	0.6749/0.7826	0.6712/0.5882	0.7393/0.8951	0.7720/0.9786
SNNE	0.7869/0.9687	0.7578/0.9521	0.6690/0.7244	0.6807/0.7933	0.6638/0.5815	0.8031/0.9254	0.6833/0.9701
SNNE (only_num)	0.7867/0.9654	0.7574/0.9504	0.6695/0.7246	0.6803/0.7920	0.6633/0.5816	0.8041/0.9387	0.6820/0.9699
SNNE (only_denom)	0.7862/0.9684	0.7572/0.9497	0.6737/0.7262	0.6779/0.7913	0.6627/0.5827	0.7943/0.9056	0.6806/0.9698
QWEN 3-1.7B							
Discrete semantic entropy	0.5515/0.4871	0.7051/0.9018	0.6509/0.6319	0.6424/0.6977	0.6978/0.2946	0.6314/0.9312	0.4940/0.6051
Discrete semantic density	0.5635/0.5185	0.7085/0.8895	0.7131/0.6585	0.6441/0.7001	0.7161/0.2951	0.7023/0.9555	0.5127/0.6369
Discrete U_{eig}	0.6045/0.5213	0.7085/0.8892	0.7112/0.6617	0.6479/0.7095	0.7048/0.2833	0.7252/0.9513	0.5220/0.6375
Discrete U_{deg}	0.5858/0.5065	0.7069/0.8886	0.7062/0.6565	0.6476/0.7060	0.7175/0.2919	0.7000/0.9414	0.5148/0.6127
Discrete C_{ecc}	0.6368/0.5382	0.6684/0.8900	0.6145/0.5703	0.6060/0.6803	0.5474/0.2318	0.7227/0.9537	0.5230/0.6450
KLE-heat$_{t=0.1}$	0.5810/0.5054	0.7067/0.8886	0.7013/0.6531	0.6474/0.7092	0.7244/0.2978	0.6963/0.9475	0.5144/0.6128
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5918/0.5130	0.7071/0.8893	0.7079/0.6574	0.6487/0.7096	0.7226/0.2958	0.7034/0.9488	0.5155/0.6461
Discrete LUQ-pair	0.5564/0.4835	0.6281/0.8740	0.6673/0.6117	0.6068/0.6807	0.7088/0.2829	0.5879/0.9425	0.4985/0.6218
Discrete SNNE	0.6171/0.5379	0.6649/0.8866	0.6652/0.6365	0.6343/0.7008	0.7253/0.2946	0.6218/0.9376	0.5067/0.6228
Discrete SNNE (only_num)	0.6727/0.6245	0.7302/0.9092	0.6785/0.6420	0.6579/0.7171	0.7186/0.2994	0.7369/0.9550	0.7492/0.7830
Discrete SNNE (only_denom)	0.6851/0.6332	0.7122/0.9055	0.6464/0.6213	0.6487/0.7126	0.5818/0.2446	0.6907/0.9472	0.7536/0.7841
Semantic density	0.7024/0.6583	0.7605/0.9179	0.5394/0.5051	0.6557/0.6847	0.6688/0.2403	0.6218/0.9305	0.5020/0.6208
U_{eig}	0.6578/0.5999	0.7507/0.9131	0.7397/0.6490	0.6727/0.6918	0.7321/0.2533	0.7744/0.9647	0.6892/0.7730
U_{deg}	0.6652/0.6033	0.7500/0.9114	0.7256/0.6391	0.6722/0.6970	0.7402/0.2566	0.7671/0.9639	0.7075/0.7822
C_{ecc}	0.6638/0.5884	0.7020/0.9018	0.6196/0.5350	0.6134/0.6616	0.5363/0.1762	0.7454/0.9572	0.6454/0.7435
LUQ-pair	0.7054/0.6415	0.7324/0.9009	0.6864/0.6131	0.6544/0.6759	0.7421/0.2566	0.7773/0.9662	0.8101/0.8274
SNNE	0.6588/0.6030	0.7504/0.9115	0.7300/0.6426	0.6727/0.6929	0.7247/0.2499	0.7674/0.9640	0.7143/0.7851
SNNE (only_num)	0.6587/0.6028	0.7502/0.9116	0.7289/0.6420	0.6726/0.6941	0.7263/0.2515	0.7675/0.9640	0.7129/0.7845
SNNE (only_denom)	0.6698/0.6060	0.7502/0.9094	0.7196/0.6356	0.6720/0.6941	0.7429/0.2569	0.7673/0.9639	0.7112/0.7838

Table 17: SUQ experimental results on QA and Math (Part I). We report AUROC/AUARC for QA and Math tasks. (Continued.)

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
QWEN 3-8B							
Discrete semantic entropy	0.8118/0.7746	0.6650/0.9001	0.7193/0.6844	0.6349/0.6975	0.7129/0.3381	0.6627/0.9511	0.4941/0.6101
Discrete semantic density	0.8214/0.7889	0.6732/0.9038	0.6991/0.6672	0.6392/0.6994	0.7131/0.3352	0.7145/0.9579	0.4944/0.6158
Discrete U_{eig}	0.8219/0.7832	0.6701/0.9020	0.6866/0.6676	0.6419/0.7070	0.7062/0.3324	0.7615/0.9716	0.4961/0.6154
Discrete U_{deg}	0.8234/0.7883	0.6698/0.9017	0.6852/0.6573	0.6414/0.7067	0.7134/0.3401	0.7237/0.9615	0.4945/0.6160
Discrete C_{ecc}	0.6347/0.6546	0.6365/0.8912	0.7016/0.6442	0.5997/0.6794	0.5979/0.2747	0.7062/0.9531	0.4952/0.6159
KLE-heat$_{t=0.1}$	0.8225/0.7842	0.6695/0.9018	0.6885/0.6581	0.6407/0.7043	0.7167/0.3403	0.7227/0.9539	0.4943/0.6154
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8234/0.7852	0.6698/0.9003	0.6899/0.6585	0.6417/0.7051	0.7155/0.3377	0.7254/0.9629	0.4941/0.6153
Discrete LUQ-pair	0.7472/0.7269	0.5579/0.8712	0.6373/0.6028	0.5679/0.6624	0.6690/0.2989	0.5611/0.9460	0.5000/0.6546
Discrete SNNE	0.8092/0.7722	0.5868/0.8775	0.6315/0.6113	0.6027/0.6817	0.7026/0.3244	0.7065/0.9601	0.5008/0.6128
Discrete SNNE (only_num)	0.8151/0.7791	0.6799/0.8957	0.6661/0.6341	0.6426/0.7103	0.7164/0.3360	0.8522/0.9738	0.7715/0.8049
Discrete SNNE (only_denom)	0.7849/0.7654	0.6661/0.8924	0.6503/0.6342	0.6249/0.7036	0.6785/0.3261	0.7949/0.9674	0.7797/0.8076
Semantic density	0.7807/0.7668	0.6919/0.8969	0.6755/0.6218	0.6494/0.7114	0.7036/0.3394	0.6404/0.9373	0.5000/0.6546
U_{eig}	0.8205/0.7827	0.6981/0.8931	0.6977/0.6602	0.6526/0.7101	0.7100/0.3393	0.8025/0.9615	0.7139/0.7918
U_{deg}	0.8223/0.7842	0.6976/0.8931	0.6960/0.6482	0.6520/0.7098	0.7177/0.3450	0.8037/0.9573	0.7369/0.8028
C_{ecc}	0.5604/0.6221	0.6543/0.8852	0.7392/0.6661	0.6135/0.6894	0.5816/0.2680	0.7231/0.9507	0.6500/0.7553
LUQ-pair	0.7776/0.7521	0.6504/0.8947	0.7384/0.6916	0.6269/0.6967	0.7034/0.3240	0.7518/0.9520	0.8258/0.8467
SNNE	0.8146/0.7774	0.6980/0.8931	0.7027/0.6524	0.6526/0.7063	0.7059/0.3391	0.8061/0.9576	0.7434/0.8055
SNNE (only_num)	0.8152/0.7805	0.6979/0.8931	0.7016/0.6516	0.6522/0.7062	0.7064/0.3395	0.8058/0.9559	0.7421/0.8050
SNNE (only_denom)	0.8220/0.7824	0.6976/0.8931	0.7010/0.6609	0.6522/0.7061	0.7193/0.3407	0.8062/0.9576	0.7407/0.8044

Table 18: SUQ experimental results on QA and Math (Part II). We report AUROC/AUARC for the QA and Math tasks.

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
QWEN 3-32B							
Discrete semantic entropy	0.8227/0.8816	0.6820/0.9196	0.7637/0.7406	0.6860/0.7664	0.7325/0.4513	0.6755/0.9381	0.4915/0.6772
Discrete semantic density	0.8451/0.8907	0.6805/0.9181	0.7715/0.7306	0.6905/0.7723	0.7293/0.4507	0.7283/0.9493	0.5093/0.6940
Discrete U_{eig}	0.8504/0.8909	0.6918/0.9205	0.7480/0.7162	0.6957/0.7664	0.7206/0.4431	0.7480/0.9411	0.4966/0.6932
Discrete U_{deg}	0.8518/0.8933	0.6846/0.9176	0.7681/0.7248	0.6949/0.7704	0.7307/0.4490	0.7482/0.9553	0.4922/0.6887
Discrete C_{ecc}	0.6193/0.7832	0.6358/0.9101	0.6761/0.6771	0.6281/0.7376	0.6032/0.3751	0.6390/0.9432	0.4932/0.7106
KLE-heat$_{t=0.1}$	0.8481/0.8906	0.6831/0.9170	0.7798/0.7290	0.6939/0.7749	0.7364/0.4537	0.7431/0.9615	0.4918/0.7020
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8511/0.8935	0.6859/0.9200	0.7709/0.7245	0.6958/0.7735	0.7341/0.4522	0.7462/0.9389	0.4915/0.7071
Discrete LUQ-pair	0.7874/0.8564	0.5969/0.9034	0.6945/0.6453	0.6227/0.7167	0.6995/0.4161	0.6496/0.9448	0.5014/0.6974
Discrete SNNE	0.8370/0.8892	0.6417/0.9210	0.7099/0.7072	0.6562/0.7399	0.7308/0.4367	0.7322/0.9545	0.5161/0.7078
Discrete SNNE (only_num)	0.8389/0.8864	0.6686/0.9217	0.7288/0.7112	0.6964/0.7798	0.7341/0.4525	0.7508/0.9538	0.7280/0.8287
Discrete SNNE (only_denom)	0.7422/0.8471	0.6257/0.9159	0.6446/0.6700	0.6762/0.7712	0.6411/0.4069	0.6988/0.9504	0.7253/0.8274
Semantic density	0.7868/0.8688	0.7319/0.9359	0.7324/0.7258	0.7051/0.7815	0.7142/0.4445	0.6221/0.9502	0.5000/0.7095
U_{eig}	0.8496/0.8905	0.7009/0.9277	0.7550/0.7281	0.7036/0.7784	0.7230/0.4514	0.7678/0.9651	0.6857/0.8284
U_{deg}	0.8502/0.8910	0.6971/0.9298	0.7790/0.7389	0.7031/0.7764	0.7350/0.4533	0.7677/0.9629	0.7122/0.8403
C_{ecc}	0.5459/0.7460	0.6614/0.9260	0.6977/0.7155	0.6404/0.7556	0.5873/0.3587	0.5966/0.9370	0.6247/0.7924
LUQ-pair	0.8243/0.8785	0.7095/0.9256	0.7570/0.7291	0.6955/0.7726	0.7356/0.4525	0.7362/0.9553	0.8208/0.8845
SNNE	0.8420/0.8878	0.6986/0.9256	0.7617/0.7321	0.7044/0.7769	0.7196/0.4503	0.7629/0.9549	0.7196/0.8431
SNNE (only_num)	0.8431/0.8880	0.6978/0.9278	0.7645/0.7329	0.7037/0.7758	0.7207/0.4500	0.7633/0.9548	0.7183/0.8426
SNNE (only_denom)	0.8489/0.8911	0.6974/0.9269	0.7829/0.7401	0.7037/0.7802	0.7381/0.4541	0.7647/0.9601	0.7168/0.8420
MISTRAL-7B-INSTRUCT-V0.2							
Discrete semantic entropy	0.7491/0.8655	0.7084/0.9021	0.6987/0.6748	0.6797/0.7357	0.6932/0.4283	0.7579/0.6603	0.5925/0.5418
Discrete semantic density	0.7987/0.8830	0.7033/0.9003	0.7352/0.6905	0.6773/0.7357	0.6946/0.4373	0.7902/0.6147	0.6192/0.5696
Discrete U_{eig}	0.8126/0.8874	0.7128/0.9032	0.7543/0.7006	0.6858/0.7346	0.6813/0.4287	0.7985/0.6741	0.6366/0.5801
Discrete U_{deg}	0.8071/0.8856	0.7075/0.8983	0.7514/0.6944	0.6848/0.7349	0.6953/0.4358	0.7934/0.6895	0.6225/0.5832
Discrete C_{ecc}	0.6273/0.8042	0.6316/0.8806	0.5855/0.5783	0.6029/0.7019	0.5900/0.3609	0.6288/0.5511	0.6276/0.5808
KLE-heat$_{t=0.1}$	0.7989/0.8816	0.7073/0.8987	0.7511/0.6959	0.6851/0.7340	0.7023/0.4387	0.7897/0.6757	0.6225/0.5692
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8027/0.8829	0.7076/0.9015	0.7539/0.6955	0.6852/0.7362	0.6985/0.4380	0.7914/0.6779	0.6236/0.5785
Discrete LUQ-pair	0.7388/0.8460	0.6325/0.8760	0.7215/0.6620	0.6464/0.7040	0.6690/0.3990	0.7158/0.5971	0.5010/0.4744
Discrete SNNE	0.7773/0.8725	0.6599/0.8656	0.6985/0.6608	0.6673/0.7257	0.6834/0.4142	0.7664/0.6741	0.5694/0.5331
Discrete SNNE (only_num)	0.7832/0.8804	0.7089/0.9065	0.7044/0.6713	0.6888/0.7450	0.6909/0.4281	0.7754/0.6920	0.6932/0.6376
Discrete SNNE (only_denom)	0.7154/0.8582	0.6871/0.9023	0.6642/0.6538	0.6821/0.7387	0.6510/0.4072	0.7644/0.6620	0.7112/0.6396
U_{eig}	0.8125/0.8899	0.7200/0.9100	0.7557/0.7030	0.6891/0.7413	0.6858/0.4335	0.7981/0.6874	0.7011/0.6407
U_{deg}	0.8052/0.8867	0.7184/0.9090	0.7584/0.7056	0.6898/0.7442	0.7017/0.4391	0.7916/0.6803	0.7111/0.6477
C_{ecc}	0.6110/0.8043	0.6504/0.8958	0.5683/0.5872	0.6061/0.7099	0.5771/0.3566	0.5983/0.5121	0.6497/0.6098
LUQ	0.7716/0.8670	0.7263/0.9103	0.7356/0.6907	0.6960/0.7533	0.7025/0.4307	0.7291/0.6318	0.7564/0.6800
Semantic density	0.7179/0.8621	0.7374/0.9161	0.5862/0.6294	0.6846/0.7446	0.6645/0.4194	0.7501/0.6195	0.5008/0.4958
SNNE	0.8049/0.8876	0.7210/0.9096	0.7614/0.7066	0.6919/0.7442	0.6867/0.4337	0.7927/0.6937	0.7146/0.6498
SNNE (only_num)	0.8059/0.8885	0.7206/0.9105	0.7624/0.7072	0.6915/0.7432	0.6883/0.4337	0.7939/0.6738	0.7144/0.6497
SNNE (only_denom)	0.8021/0.8863	0.7198/0.9106	0.7593/0.7064	0.6912/0.7397	0.7058/0.4397	0.7902/0.6798	0.7143/0.6495

Table 18: SUQ experimental results on QA and Math (Part II). We report AUROC/AUARC for the QA and Math tasks. (Continued.)

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
MISTRAL-NEMO-INSTRUCT-2407 12B							
Discrete semantic entropy	0.8249/0.8893	0.7427/0.8484	0.7357/0.5706	0.7506/0.7273	0.7453/0.3642	0.6214/0.6193	0.5281/0.6973
Discrete semantic density	0.8421/0.8931	0.7460/0.8597	0.7116/0.5767	0.7524/0.7296	0.7331/0.3611	0.7337/0.6887	0.5516/0.7389
Discrete U_{eig}	0.8404/0.8958	0.7472/0.8551	0.6746/0.5514	0.7493/0.7310	0.7179/0.3441	0.7545/0.7095	0.5612/0.7234
Discrete U_{deg}	0.8442/0.8967	0.7452/0.8530	0.7033/0.5621	0.7544/0.7254	0.7276/0.3530	0.7555/0.6949	0.5535/0.7024
Discrete C_{ecc}	0.6536/0.7939	0.6953/0.8313	0.6757/0.5500	0.6812/0.6848	0.6339/0.2820	0.4380/0.4763	0.5521/0.7421
KLE-heat$_{t=0.1}$	0.8429/0.8919	0.7457/0.8545	0.7229/0.5734	0.7572/0.7323	0.7378/0.3600	0.7418/0.6851	0.5532/0.7224
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8438/0.8948	0.7459/0.8504	0.7181/0.5630	0.7577/0.7307	0.7378/0.3578	0.7284/0.6872	0.5539/0.7225
LUQ-pair	0.7489/0.8444	0.6206/0.7924	0.7029/0.5451	0.6841/0.6642	0.7110/0.3310	0.6772/0.6496	0.4968/0.6840
SNNE	0.8256/0.8837	0.6761/0.8152	0.7164/0.5673	0.7173/0.6971	0.7327/0.3381	0.6610/0.6580	0.5226/0.6928
SNNE (only_num)	0.8451/0.8991	0.7841/0.8867	0.7175/0.5458	0.7470/0.7222	0.7546/0.3699	0.6502/0.6441	0.6909/0.8059
SNNE (only_denom)	0.8280/0.8930	0.7815/0.8872	0.6737/0.5308	0.7174/0.7165	0.7579/0.3660	0.6207/0.6205	0.6996/0.8078
U_{eig}	0.8461/0.8994	0.7633/0.8803	0.6818/0.5506	0.7508/0.7308	0.7241/0.3560	0.7503/0.7101	0.6307/0.7877
U_{deg}	0.8497/0.9014	0.7646/0.8783	0.7114/0.5686	0.7567/0.7331	0.7364/0.3598	0.7476/0.6947	0.6426/0.7955
C_{ecc}	0.6273/0.7876	0.7172/0.8623	0.6614/0.5296	0.6889/0.6909	0.6122/0.2751	0.4101/0.4649	0.5921/0.7645
LUQ	0.7954/0.8741	0.7529/0.8647	0.7480/0.5933	0.7468/0.7256	0.7465/0.3512	0.6924/0.6478	0.7158/0.8307
Semantic density	0.8032/0.8853	0.7903/0.8896	0.7326/0.5688	0.7496/0.7270	0.7362/0.3601	0.5018/0.5624	0.5022/0.7188
SNNE	0.8340/0.8971	0.7652/0.8782	0.6910/0.5582	0.7539/0.7331	0.7150/0.3543	0.7615/0.7126	0.6459/0.7970
SNNE (only_num)	0.8357/0.8980	0.7652/0.8794	0.6938/0.5582	0.7535/0.7313	0.7170/0.3534	0.7651/0.7164	0.6455/0.7968
SNNE (only_denom)	0.8490/0.9004	0.7662/0.8814	0.7221/0.5726	0.7593/0.7379	0.7424/0.3625	0.7343/0.6816	0.6451/0.7966
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506							
Discrete semantic entropy	0.7969/0.9171	0.7274/0.9023	0.6122/0.6305	0.6690/0.7135	0.7248/0.4250	0.7831/0.8718	0.4785/0.6669
Discrete semantic density	0.8129/0.9201	0.7379/0.8945	0.6255/0.6306	0.6783/0.7181	0.7195/0.4134	0.8011/0.8512	0.4946/0.6887
Discrete U_{eig}	0.8145/0.9230	0.7326/0.8841	0.6296/0.6436	0.6878/0.7177	0.7093/0.4116	0.8043/0.8750	0.4966/0.6754
Discrete U_{deg}	0.8150/0.9217	0.7299/0.8836	0.6303/0.6360	0.6859/0.7193	0.7177/0.4138	0.8109/0.8821	0.4931/0.6890
Discrete C_{ecc}	0.6612/0.8589	0.6829/0.8734	0.5627/0.5703	0.6242/0.6863	0.6398/0.3591	0.6121/0.7858	0.4941/0.6863
KLE-heat$_{t=0.1}$	0.8129/0.9226	0.7307/0.8835	0.6296/0.6190	0.6849/0.7169	0.7229/0.4245	0.8067/0.8538	0.4930/0.6881
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8146/0.9220	0.7303/0.8829	0.6334/0.6286	0.6866/0.7167	0.7216/0.4186	0.8011/0.8792	0.4935/0.6743
Discrete LUQ-pair	0.7099/0.8787	0.5857/0.8496	0.6069/0.6041	0.6271/0.6726	0.6658/0.3622	0.7249/0.8178	0.4978/0.6825
Discrete SNNE	0.7940/0.9131	0.6297/0.8627	0.6057/0.5833	0.6622/0.7047	0.7068/0.3894	0.7973/0.8575	0.5057/0.6860
Discrete SNNE (only_num)	0.8076/0.9151	0.7194/0.8928	0.5996/0.5855	0.6721/0.7088	0.7249/0.4172	0.7824/0.8691	0.5712/0.7171
Discrete SNNE (only_denom)	0.7690/0.9066	0.7027/0.8891	0.5691/0.5632	0.6376/0.6955	0.7154/0.4115	0.7635/0.8702	0.5730/0.7176
U_{eig}	0.8149/0.9185	0.7370/0.8996	0.6263/0.6228	0.6860/0.7177	0.7119/0.4152	0.7979/0.8735	0.5858/0.7689
U_{deg}	0.8158/0.9180	0.7367/0.9006	0.6268/0.6209	0.6855/0.7186	0.7221/0.4160	0.8018/0.8793	0.6054/0.7791
C_{ecc}	0.5898/0.8381	0.6807/0.8832	0.5647/0.5633	0.6204/0.6846	0.6086/0.3335	0.5750/0.7663	0.5524/0.7429
LUQ	0.7787/0.9110	0.7249/0.8938	0.6223/0.6233	0.6782/0.7197	0.7164/0.4083	0.7313/0.8266	0.7170/0.8247
Semantic density	0.7598/0.9049	0.7757/0.9137	0.6219/0.6019	0.6987/0.7309	0.7097/0.4114	0.7110/0.8512	0.4956/0.6740
SNNE	0.8110/0.9181	0.7368/0.9007	0.6279/0.6227	0.6862/0.7177	0.7058/0.4131	0.7843/0.8690	0.6101/0.7810
SNNE (only_num)	0.8116/0.9179	0.7366/0.9012	0.6301/0.6220	0.6860/0.7165	0.7064/0.4134	0.7878/0.8701	0.6092/0.7806
SNNE (only_denom)	0.8151/0.9172	0.7376/0.9001	0.6267/0.6224	0.6859/0.7188	0.7249/0.4195	0.7972/0.8781	0.6081/0.7802

Table 18: SUQ experimental results on QA and Math (Part II). We report AUROC/AUARC for the QA and Math tasks. (Continued.)

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
GPT-4O-MINI							
Discrete semantic entropy	0.7829/0.9323	0.7185/0.9185	0.7424/0.7682	0.6786/0.7519	0.7080/0.5347	0.7420/0.9196	0.4916/0.5218
Discrete semantic density	0.7917/0.9330	0.7239/0.9191	0.7159/0.7604	0.6790/0.7471	0.6964/0.5384	0.7507/0.9237	0.5020/0.5393
Discrete U_{eig}	0.7941/0.9347	0.7252/0.9208	0.7274/0.7562	0.6855/0.7572	0.6900/0.5385	0.7540/0.9244	0.5027/0.5410
Discrete U_{deg}	0.7943/0.9340	0.7201/0.9201	0.7233/0.7653	0.6823/0.7572	0.6957/0.5377	0.7540/0.9244	0.5022/0.5406
Discrete C_{ecc}	0.6418/0.8850	0.6917/0.9043	0.6470/0.6906	0.6349/0.7283	0.6351/0.4905	0.6281/0.8769	0.5029/0.5408
KLE-heat$_{t=0.1}$	0.7928/0.9340	0.7200/0.9201	0.7251/0.7614	0.6817/0.7535	0.7008/0.5414	0.7518/0.9240	0.5024/0.5407
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7941/0.9330	0.7205/0.9201	0.7241/0.7688	0.6831/0.7533	0.6998/0.5391	0.7541/0.9243	0.5021/0.5406
Discrete LUQ-pair	0.6973/0.9077	0.5967/0.8873	0.6463/0.6842	0.6030/0.6979	0.6358/0.4722	0.6509/0.8810	0.5016/0.5335
Discrete SNNE	0.7665/0.9272	0.6146/0.8922	0.7120/0.7121	0.6311/0.7102	0.6780/0.5111	0.7558/0.9242	0.5020/0.5319
Discrete SNNE (only_num)	0.7964/0.9322	0.7129/0.9175	0.7386/0.7761	0.6830/0.7585	0.7109/0.5525	0.7529/0.9238	0.6649/0.6289
Discrete SNNE (only_denom)	0.7703/0.9273	0.6987/0.9147	0.6967/0.7549	0.6732/0.7545	0.6778/0.5333	0.7407/0.9225	0.6679/0.6307
Semantic density	0.7727/0.9281	0.7634/0.9329	0.7716/0.7840	0.7073/0.7724	0.7296/0.5596	0.7368/0.9217	0.5000/0.5350
U_{eig}	0.8031/0.9347	0.7431/0.9228	0.7408/0.7780	0.6969/0.7656	0.7002/0.5516	0.7538/0.9245	0.7304/0.7176
U_{deg}	0.8034/0.9320	0.7411/0.9221	0.7407/0.7779	0.6952/0.7599	0.7096/0.5507	0.7512/0.9239	0.7563/0.7317
C_{ecc}	0.6226/0.8791	0.7094/0.9196	0.6694/0.7117	0.6533/0.7481	0.6472/0.5068	0.5910/0.8526	0.6713/0.6691
LUQ	0.7534/0.9238	0.7259/0.9219	0.7579/0.7785	0.6892/0.7561	0.7012/0.5342	0.6647/0.8900	0.8448/0.7812
SNNE	0.8016/0.9334	0.7422/0.9222	0.7446/0.7787	0.6969/0.7638	0.7019/0.5529	0.7462/0.9234	0.7619/0.7342
SNNE (only_num)	0.8018/0.9335	0.7419/0.9223	0.7442/0.7777	0.6964/0.7621	0.7018/0.5474	0.7462/0.9234	0.7607/0.7337
SNNE (only_denom)	0.8029/0.9348	0.7415/0.9222	0.7437/0.7775	0.6957/0.7630	0.7124/0.5504	0.7507/0.9239	0.7595/0.7332

Table 19: SUQ experimental results on QA and Math (Part III). We report AUROC/AUARC for the QA and Math tasks.

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
PHI 4-14B							
Discrete semantic entropy	0.8433/0.9120	0.7397/0.9241	0.6348/0.6903	0.7029/0.7804	0.7303/0.4858	0.8367/0.8388	0.4736/0.5399
Discrete semantic density	0.8732/0.9204	0.7292/0.9226	0.6922/0.7112	0.6956/0.7752	0.7257/0.4834	0.8390/0.8394	0.4850/0.5465
Discrete U_{eig}	0.8727/0.9204	0.7540/0.9299	0.6902/0.7178	0.7119/0.7821	0.7167/0.4802	0.8450/0.8330	0.4856/0.5466
Discrete U_{deg}	0.8781/0.9222	0.7387/0.9251	0.6843/0.7101	0.7085/0.7810	0.7282/0.4817	0.8447/0.8451	0.4850/0.5465
Discrete C_{ecc}	0.6546/0.8214	0.6360/0.8998	0.6174/0.6423	0.6054/0.7353	0.6158/0.4222	0.6299/0.6811	0.4854/0.5466
KLE-heat$_{t=0.1}$	0.8762/0.9208	0.7382/0.9234	0.6758/0.7067	0.7092/0.7803	0.7373/0.4893	0.8421/0.8393	0.4845/0.5463
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8766/0.9207	0.7397/0.9236	0.6812/0.7096	0.7091/0.7812	0.7320/0.4862	0.8406/0.8235	0.4851/0.5465
Discrete LUQ-pair	0.8100/0.8804	0.6331/0.8898	0.6425/0.6751	0.6592/0.7484	0.7226/0.4528	0.7388/0.7783	0.5000/0.5653
Discrete SNNE	0.8552/0.9034	0.6773/0.9039	0.6491/0.6690	0.6855/0.7638	0.7270/0.4678	0.8333/0.8086	0.5022/0.5536
Discrete SNNE (only_num)	0.8676/0.9171	0.7450/0.9293	0.6534/0.6978	0.7109/0.7858	0.7408/0.4923	0.8458/0.8479	0.6725/0.6689
Discrete SNNE (only_denom)	0.8400/0.9094	0.7210/0.9243	0.6123/0.6698	0.7057/0.7822	0.6984/0.4677	0.8361/0.8424	0.6713/0.6681
U_{eig}	0.8745/0.9198	0.7665/0.9332	0.7044/0.7270	0.7154/0.7873	0.7231/0.4858	0.8442/0.8453	0.7495/0.7541
U_{deg}	0.8797/0.9200	0.7622/0.9317	0.6956/0.7224	0.7163/0.7851	0.7384/0.4949	0.8426/0.8309	0.7797/0.7705
C_{ecc}	0.6072/0.8062	0.6541/0.9113	0.6042/0.6404	0.6104/0.7420	0.6007/0.4132	0.5608/0.6662	0.6603/0.6977
LUQ	0.8363/0.9006	0.7527/0.9312	0.6704/0.7036	0.7194/0.7899	0.7458/0.4914	0.7985/0.7951	0.8552/0.8134
Semantic density	0.8140/0.9016	0.7893/0.9431	0.5992/0.6649	0.7063/0.7905	0.7072/0.4868	0.7845/0.8253	0.4991/0.5622
SNNE	0.8599/0.9160	0.7630/0.9317	0.6953/0.7214	0.7177/0.7863	0.7235/0.4887	0.8429/0.8393	0.7850/0.7729
SNNE (only_num)	0.8628/0.9159	0.7627/0.9326	0.6949/0.7213	0.7173/0.7883	0.7254/0.4883	0.8426/0.8459	0.7840/0.7725
SNNE (only_denom)	0.8782/0.9208	0.7632/0.9316	0.6910/0.7198	0.7176/0.7868	0.7428/0.4952	0.8403/0.8390	0.7830/0.7720
PHI4-MINI-INSTRUCT 3.8B							
Discrete semantic entropy	0.8097/0.7144	0.7737/0.8951	0.6236/0.5825	0.7041/0.7292	0.6967/0.3015	0.5634/0.7845	0.4660/0.5939
Discrete semantic density	0.8407/0.7281	0.7711/0.8979	0.6651/0.5927	0.7043/0.7277	0.7140/0.3011	0.6291/0.8370	0.4761/0.6053
Discrete U_{eig}	0.8315/0.7273	0.7690/0.8978	0.6644/0.5973	0.7088/0.7324	0.7057/0.2953	0.6660/0.8471	0.4936/0.6128
Discrete U_{deg}	0.8438/0.7321	0.7745/0.8971	0.6757/0.6045	0.7153/0.7329	0.7225/0.3065	0.6404/0.8418	0.4853/0.6081
Discrete C_{ecc}	0.6693/0.5975	0.6835/0.8640	0.6419/0.5670	0.6303/0.6887	0.6227/0.2522	0.6171/0.8189	0.5019/0.6223
KLE-heat$_{t=0.1}$	0.8463/0.7289	0.7787/0.8983	0.6844/0.6093	0.7185/0.7340	0.7376/0.3123	0.6363/0.8383	0.4825/0.6159
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8457/0.7301	0.7757/0.8970	0.6797/0.6056	0.7163/0.7363	0.7296/0.3075	0.6419/0.8399	0.4874/0.6098
Discrete LUQ-pair	0.8040/0.6810	0.6875/0.8435	0.6811/0.5941	0.6879/0.7039	0.7386/0.2997	0.5653/0.7896	0.5009/0.6273
Discrete SNNE	0.8249/0.7140	0.7379/0.8730	0.6772/0.5987	0.6981/0.7205	0.7248/0.3084	0.6317/0.8263	0.5184/0.6487
Discrete SNNE (only_num)	0.8325/0.7234	0.7863/0.9053	0.6869/0.6154	0.7166/0.7363	0.7462/0.3170	0.6783/0.8543	0.6794/0.7742
Discrete SNNE (only_denom)	0.8187/0.7131	0.7803/0.9022	0.6212/0.5771	0.7190/0.7341	0.7176/0.3020	0.7420/0.8768	0.7779/0.8114
U_{eig}	0.8348/0.7271	0.7734/0.9003	0.6709/0.6008	0.7126/0.7358	0.7114/0.3001	0.7217/0.8806	0.6164/0.7248
U_{deg}	0.8457/0.7318	0.7823/0.9028	0.6837/0.6069	0.7195/0.7370	0.7289/0.3103	0.7445/0.8903	0.6508/0.7466
C_{ecc}	0.5961/0.5354	0.6960/0.8764	0.6290/0.5645	0.6313/0.6893	0.6031/0.2430	0.6310/0.8099	0.5480/0.6709
LUQ	0.8187/0.7095	0.7746/0.8976	0.6967/0.6202	0.7183/0.7350	0.7486/0.3166	0.7618/0.8775	0.7603/0.8067
Semantic density	0.7717/0.6876	0.8057/0.9083	0.5725/0.5695	0.6906/0.7286	0.6572/0.2905	0.5390/0.7945	0.5000/0.6195
SNNE	0.8001/0.7111	0.7785/0.9021	0.6726/0.6027	0.7179/0.7365	0.7106/0.3053	0.7391/0.8891	0.6601/0.7522
SNNE (only_num)	0.8078/0.7149	0.7790/0.9029	0.6735/0.6036	0.7184/0.7383	0.7139/0.3068	0.7425/0.8902	0.6595/0.7516
SNNE (only_denom)	0.8453/0.7313	0.7852/0.9034	0.6899/0.6110	0.7216/0.7377	0.7354/0.3138	0.7494/0.8924	0.6589/0.7510

Table 19: SUQ experimental results on QA and Math (Part III). We report AUROC/AUARC for the QA and Math tasks. (Continued.)

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
PHI 3-14B							
Discrete semantic entropy	0.8181/0.8666	0.7388/0.8838	0.6035/0.6178	0.6711/0.7258	0.6964/0.4065	0.6143/0.9099	0.6017/0.7004
Discrete semantic density	0.8623/0.8826	0.7284/0.8813	0.5915/0.5833	0.6657/0.7254	0.6800/0.3933	0.6301/0.9098	0.6201/0.7085
Discrete U_{eig}	0.8472/0.8793	0.7443/0.8894	0.5935/0.5914	0.6830/0.7369	0.6627/0.3861	0.6248/0.9110	0.6340/0.7115
Discrete U_{deg}	0.8642/0.8854	0.7467/0.8873	0.6014/0.5934	0.6905/0.7368	0.6868/0.3981	0.6339/0.9115	0.6239/0.7184
Discrete C_{ecc}	0.6608/0.7669	0.6177/0.8357	0.6036/0.5901	0.5680/0.6509	0.5809/0.3124	0.6035/0.8997	0.6195/0.7075
KLE-heat$_{t=0.1}$	0.8668/0.8849	0.7522/0.8887	0.6121/0.6014	0.6959/0.7375	0.7051/0.4069	0.6447/0.9097	0.6246/0.7110
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8648/0.8847	0.7478/0.8879	0.6028/0.5946	0.6917/0.7366	0.6957/0.4020	0.6374/0.9096	0.6231/0.7084
Discrete LUQ-pair	0.8082/0.8442	0.6695/0.8451	0.5915/0.5703	0.6697/0.7009	0.7104/0.3948	0.5599/0.8541	0.5085/0.6266
Discrete SNNE	0.8417/0.8710	0.7228/0.8682	0.6198/0.6001	0.6741/0.7205	0.7093/0.4028	0.6469/0.9085	0.5730/0.6762
Discrete SNNE (only_num)	0.8487/0.8782	0.7573/0.8944	0.6274/0.6064	0.6948/0.7387	0.7245/0.4164	0.6819/0.9259	0.6756/0.7561
Discrete SNNE (only_denom)	0.8111/0.8601	0.7343/0.8857	0.6212/0.6167	0.7134/0.7435	0.6741/0.3858	0.7004/0.9255	0.7316/0.7693
U_{eig}	0.8507/0.8796	0.7518/0.8928	0.6118/0.6025	0.6904/0.7404	0.6697/0.3907	0.6572/0.9195	0.6826/0.7567
U_{deg}	0.8670/0.8857	0.7633/0.8942	0.6233/0.6093	0.7012/0.7430	0.6957/0.4055	0.6973/0.9265	0.6929/0.7633
C_{ecc}	0.5914/0.7071	0.6388/0.8507	0.5772/0.5679	0.5711/0.6607	0.5729/0.3106	0.6009/0.8987	0.6179/0.7256
LUQ	0.8286/0.8665	0.7559/0.8912	0.6381/0.6242	0.7032/0.7432	0.7187/0.4163	0.7149/0.9346	0.7133/0.7777
Semantic density	0.7652/0.8450	0.7504/0.8956	0.5257/0.5534	0.6504/0.7269	0.6265/0.3817	0.5499/0.8988	0.5000/0.6279
SNNE	0.8178/0.8697	0.7571/0.8934	0.6240/0.6098	0.6987/0.7429	0.6592/0.3923	0.6833/0.9239	0.6937/0.7641
SNNE (only_num)	0.8252/0.8722	0.7583/0.8937	0.6275/0.6118	0.6995/0.7431	0.6628/0.3937	0.6891/0.9248	0.6942/0.7643
SNNE (only_denom)	0.8662/0.8849	0.7665/0.8954	0.6287/0.6128	0.7040/0.7439	0.7026/0.4096	0.7031/0.9273	0.6954/0.7646

H.3 Complete SUQ Results on Summarization and Translation

Discussion: Appendix C.

Table 20: SUQ experimental results on Summarization and Translation (Part I). We report PRR for summarization and translation.

Uncertainty Method	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT												
Discrete semantic entropy	0.2841	0.1799	0.3336	0.2806	0.5594	0.6848	0.5001	0.5356	0.4052	0.5128	0.6058	0.5093
Discrete semantic density	0.2830	0.1351	0.3281	0.2650	0.5134	0.6206	0.4327	0.4099	0.3346	0.4184	0.5627	0.4522
Discrete U_{eig}	0.2777	0.1806	0.3317	0.2798	0.5545	0.6981	0.5031	0.5408	0.4083	0.5070	0.6033	0.5124
Discrete U_{deg}	0.2780	0.1786	0.3317	0.2795	0.5504	0.6979	0.5056	0.5416	0.4090	0.5081	0.6050	0.5114
Discrete C_{ecc}	0.2682	0.1324	0.3337	0.2705	0.5143	0.6222	0.4314	0.4056	0.3322	0.4155	0.5646	0.5107
KLE-heat$_{t=0.1}$	0.2787	0.1788	0.3317	0.2794	0.5554	0.6971	0.5032	0.5405	0.4090	0.5090	0.6076	0.5101
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2780	0.1789	0.3317	0.2794	0.5529	0.6984	0.5041	0.5395	0.3418	0.4300	0.6035	0.5101
Discrete LUQ-pair	0.2811	0.1725	0.3272	0.2762	0.5388	0.6742	0.4798	0.5075	0.3790	0.4839	0.5965	0.4950
Discrete SNNE	0.2801	0.1786	0.3297	0.2762	0.5440	0.6844	0.4884	0.5213	0.3884	0.5036	0.5985	0.4916
Discrete SNNE (only_num)	0.2827	0.1831	0.3358	0.2822	0.5976	0.7345	0.5296	0.5628	0.4289	0.5287	0.6380	0.5343
Discrete SNNE (only_denom)	0.2910	0.1799	0.3641	0.3079	0.6037	0.7384	0.5411	0.5803	0.4444	0.5414	0.6408	0.5450
Semantic density	0.2741	0.1325	0.3357	0.2708	0.5072	0.6178	0.4364	0.4263	0.3529	0.4465	0.5610	0.4650
U_{eig}	0.2808	0.1839	0.3392	0.2854	0.5781	0.7239	0.5168	0.5464	0.4147	0.5107	0.6273	0.5169
U_{deg}	0.2812	0.1840	0.3388	0.2855	0.5775	0.7217	0.5172	0.5477	0.4141	0.5120	0.6257	0.5155
C_{ecc}	0.2699	0.1690	0.3343	0.2729	0.5759	0.7165	0.5160	0.5389	0.4111	0.5035	0.6246	0.5150
LUQ	0.2859	0.1853	0.3387	0.2864	0.5790	0.7159	0.5194	0.5540	0.4136	0.5119	0.6281	0.5183
SNNE	0.2817	0.1847	0.3389	0.2858	0.5769	0.7215	0.5157	0.5468	0.4143	0.5105	0.6258	0.5169
SNNE (only_num)	0.2817	0.1845	0.3389	0.2858	0.5779	0.7221	0.5162	0.5474	0.4138	0.5106	0.6268	0.5158
SNNE (only_denom)	0.2817	0.1845	0.3388	0.2857	0.5787	0.7221	0.5164	0.5485	0.4142	0.5123	0.6252	0.5168
LLAMA 3.1-70B INSTRUCT												
Discrete semantic entropy	0.2672	0.2204	0.3365	0.3018	0.5884	0.7257	0.5329	0.6241	0.4770	0.5644	0.6348	0.5361
Discrete semantic density	0.2533	0.1602	0.3412	0.2968	0.5544	0.6595	0.4811	0.5188	0.4217	0.5051	0.5958	0.4916
Discrete U_{eig}	0.2648	0.2187	0.3387	0.3024	0.5908	0.7285	0.5347	0.6274	0.4833	0.5770	0.6381	0.5406
Discrete U_{deg}	0.2645	0.2147	0.3370	0.3015	0.5941	0.7272	0.5288	0.6280	0.4823	0.5685	0.6351	0.5410
Discrete C_{ecc}	0.2508	0.2144	0.3421	0.3023	0.5901	0.7292	0.5308	0.6264	0.4794	0.5718	0.6359	0.5397
KLE-heat$_{t=0.1}$	0.2649	0.2146	0.3370	0.3015	0.5900	0.7281	0.5283	0.6292	0.4821	0.5716	0.6363	0.5418
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2644	0.2146	0.3371	0.3015	0.5878	0.7275	0.5281	0.6261	0.4818	0.5719	0.6358	0.5395
Discrete LUQ-pair	0.2626	0.2066	0.3350	0.2987	0.5695	0.7026	0.5126	0.6001	0.4622	0.5577	0.6267	0.5246
Discrete SNNE	0.2644	0.2154	0.3373	0.2990	0.5767	0.7116	0.5229	0.6118	0.4643	0.5654	0.6318	0.5193
Discrete SNNE (only_num)	0.2687	0.2173	0.3431	0.3037	0.6427	0.7705	0.5763	0.6692	0.5162	0.6113	0.6760	0.5782
Discrete SNNE (only_denom)	0.2784	0.2221	0.3714	0.3291	0.6489	0.7720	0.5863	0.6889	0.5305	0.6259	0.6796	0.5932
Semantic density	0.2545	0.1567	0.3235	0.3022	0.5485	0.6739	0.4835	0.5420	0.4409	0.5231	0.5995	0.4901
U_{eig}	0.2670	0.2254	0.3508	0.3111	0.6099	0.7510	0.5523	0.6411	0.4903	0.5811	0.6567	0.5541
U_{deg}	0.2671	0.2258	0.3496	0.3105	0.6094	0.7501	0.5515	0.6424	0.4888	0.5798	0.6563	0.5525
C_{ecc}	0.2509	0.2099	0.3413	0.3021	0.5976	0.7470	0.5474	0.6366	0.4894	0.5802	0.6560	0.5520
LUQ	0.2686	0.2340	0.3477	0.3098	0.6107	0.7478	0.5514	0.6406	0.4887	0.5802	0.6569	0.5505
SNNE	0.2674	0.2278	0.3493	0.3104	0.6089	0.7503	0.5511	0.6406	0.4879	0.5794	0.6562	0.5526
SNNE (only_num)	0.2674	0.2275	0.3494	0.3104	0.6099	0.7497	0.5515	0.6405	0.4883	0.5795	0.6556	0.5520
SNNE (only_denom)	0.2674	0.2270	0.3494	0.3104	0.6085	0.7497	0.5513	0.6426	0.4888	0.5799	0.6564	0.5525

Table 20: SUQ experimental results on Summarization and Translation (Part I). We report PRR for summarization and translation. (Continued.)

Uncertainty Method	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
LLAMA 3.1-405B INSTRUCT FP8												
Discrete semantic entropy	0.2917	0.2883	0.3602	0.3116	0.6137	0.7339	0.5570	0.6727	0.4864	0.5853	0.6416	0.5646
Discrete semantic density	0.2891	0.2866	0.3593	0.3118	0.6215	0.7440	0.5639	0.6796	0.4990	0.6160	0.6602	0.5789
Discrete U_{eig}	0.2896	0.2850	0.3622	0.3123	0.6172	0.7388	0.5532	0.6738	0.4999	0.6144	0.6510	0.5705
Discrete U_{deg}	0.2883	0.2861	0.3602	0.3106	0.6085	0.7386	0.5545	0.6652	0.4933	0.6143	0.6539	0.5700
Discrete C_{ecc}	0.2756	0.2649	0.3537	0.3099	0.6072	0.7314	0.5550	0.6715	0.4913	0.6112	0.6522	0.5701
KLE-heat$_{t=0.1}$	0.2887	0.2861	0.3602	0.3105	0.6123	0.7386	0.5525	0.6732	0.4923	0.6113	0.6553	0.5679
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2885	0.2849	0.3602	0.3106	0.6076	0.7387	0.5536	0.6711	0.5006	0.6117	0.6558	0.5704
Discrete LUQ-pair	0.2897	0.2665	0.3559	0.3086	0.5898	0.7220	0.5399	0.6449	0.4802	0.5983	0.6490	0.5446
Discrete SNNE	0.2866	0.2810	0.3570	0.3104	0.5943	0.7278	0.5448	0.6534	0.4839	0.6032	0.6486	0.5540
Discrete SNNE (only_num)	0.2935	0.2888	0.3637	0.3157	0.6492	0.7745	0.5938	0.7221	0.5273	0.6484	0.6863	0.6069
Discrete SNNE (only_denom)	0.3043	0.2914	0.3897	0.3395	0.6525	0.7766	0.5961	0.7271	0.5368	0.6523	0.6884	0.6126
continuous semantic density	0.2928	0.2858	0.3396	0.3120	0.6354	0.7537	0.5773	0.6925	0.4954	0.6312	0.6710	0.5860
U_{eig}	0.2920	0.2901	0.3716	0.3237	0.6352	0.7603	0.5745	0.6984	0.5048	0.6265	0.6733	0.5874
U_{deg}	0.2916	0.2907	0.3707	0.3229	0.6354	0.7609	0.5740	0.6973	0.5036	0.6268	0.6721	0.5871
U_{ecc}	0.2649	0.1985	0.3389	0.2959	0.5774	0.6789	0.5127	0.5728	0.4382	0.5601	0.6168	0.5233
LUQ	0.2945	0.2870	0.3695	0.3225	0.6311	0.7509	0.5782	0.6961	0.5031	0.6281	0.6679	0.5848
SNNE	0.2921	0.2912	0.3706	0.3229	0.6338	0.7609	0.5732	0.6974	0.5035	0.6248	0.6703	0.5863
SNNE (only_num)	0.2921	0.2912	0.3706	0.3229	0.6338	0.7609	0.5743	0.6966	0.5023	0.6246	0.6709	0.5856
SNNE (only_denom)	0.2920	0.2912	0.3707	0.3229	0.6340	0.7609	0.5721	0.6977	0.5046	0.6246	0.6712	0.5861
QWEN 3-1.7B												
Discrete semantic entropy	0.2523	0.2597	0.3457	0.3092	0.5036	0.6340	0.4059	0.5026	0.3434	0.4032	0.5607	0.4989
Discrete semantic density	0.2502	0.2602	0.3427	0.3085	0.5022	0.6498	0.3999	0.5007	0.3426	0.3997	0.5694	0.5009
Discrete U_{eig}	0.2496	0.2610	0.3432	0.3067	0.4995	0.6389	0.3991	0.4958	0.3414	0.4020	0.5611	0.5034
Discrete U_{deg}	0.2496	0.2594	0.3418	0.3063	0.5021	0.6389	0.3980	0.5001	0.3448	0.4047	0.5596	0.5042
Discrete C_{ecc}	0.2478	0.2460	0.3407	0.3141	0.4962	0.6375	0.4009	0.4940	0.3363	0.4040	0.5620	0.5031
KLE-heat$_{t=0.1}$	0.2501	0.2603	0.3420	0.3062	0.5056	0.6387	0.4045	0.5003	0.3455	0.4080	0.5601	0.5019
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2498	0.2606	0.3417	0.3062	0.4961	0.6388	0.4015	0.5001	0.3420	0.4065	0.5615	0.5029
Discrete LUQ-pair	0.2513	0.2446	0.3378	0.3053	0.4804	0.6190	0.3872	0.4711	0.3298	0.4086	0.5521	0.4869
Discrete SNNE	0.2519	0.2469	0.3387	0.3058	0.5020	0.6297	0.3937	0.4892	0.3335	0.4170	0.5547	0.4791
Discrete SNNE (only_num)	0.2521	0.2614	0.3461	0.3099	0.5211	0.6694	0.4107	0.5146	0.3526	0.4022	0.5835	0.5169
Discrete SNNE (only_denom)	0.2533	0.2520	0.3713	0.3312	0.5241	0.6736	0.4244	0.5239	0.3644	0.4099	0.5879	0.5258
Semantic density	0.2545	0.2569	0.3319	0.3113	0.5059	0.6536	0.3961	0.5044	0.3409	0.3718	0.5779	0.5061
U_{eig}	0.2504	0.2672	0.3519	0.3142	0.5143	0.6582	0.4018	0.5077	0.3466	0.3998	0.5804	0.5083
U_{deg}	0.2507	0.2681	0.3517	0.3139	0.5146	0.6599	0.4039	0.5078	0.3477	0.4010	0.5805	0.5085
U_{ecc}	0.2523	0.1860	0.3281	0.3038	0.4550	0.5654	0.3482	0.3822	0.2781	0.3628	0.5136	0.4507
LUQ	0.2530	0.2587	0.3526	0.3140	0.5113	0.6536	0.4009	0.5057	0.3469	0.3996	0.5765	0.5084
SNNE	0.2508	0.2650	0.3520	0.3139	0.5137	0.6595	0.4036	0.5055	0.3477	0.4022	0.5799	0.5084
SNNE (only_num)	0.2508	0.2670	0.3519	0.3139	0.5151	0.6595	0.4024	0.5066	0.3474	0.3990	0.5785	0.5089
SNNE (only_denom)	0.2509	0.2669	0.3519	0.3139	0.5137	0.6597	0.4029	0.5082	0.3481	0.4002	0.5794	0.5078

Table 20: SUQ experimental results on Summarization and Translation (Part I). We report PRR for summarization and translation. (Continued.)

Uncertainty Method	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
QWEN 3-8B												
Discrete semantic entropy	0.2566	0.2378	0.3312	0.2968	0.5559	0.6890	0.5052	0.6093	0.4321	0.5547	0.6075	0.5380
Discrete semantic density	0.2530	0.1875	0.3156	0.2913	0.5310	0.6328	0.4557	0.5254	0.3857	0.5032	0.5765	0.5033
Discrete U_{eig}	0.2534	0.2394	0.3297	0.2926	0.5607	0.6821	0.4997	0.6141	0.4374	0.5462	0.6096	0.5325
Discrete U_{deg}	0.2531	0.2409	0.3273	0.2911	0.5617	0.6820	0.4999	0.6165	0.4389	0.5405	0.6090	0.5382
Discrete C_{ecc}	0.2515	0.2270	0.3211	0.2962	0.5537	0.6909	0.5035	0.6046	0.4327	0.5391	0.6117	0.5337
KLE-heat$_{t=0.1}$	0.2540	0.2375	0.3276	0.2913	0.5593	0.6827	0.4985	0.6121	0.4378	0.5467	0.6100	0.5368
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2537	0.2403	0.3272	0.2909	0.5584	0.6821	0.5069	0.6110	0.4391	0.5405	0.6092	0.5377
Discrete LUQ-pair	0.2538	0.2367	0.3219	0.2865	0.5427	0.6724	0.4849	0.5964	0.4328	0.5403	0.6033	0.5185
Discrete SNNE	0.2534	0.2395	0.3254	0.2907	0.5485	0.6769	0.4975	0.5996	0.4296	0.5460	0.6039	0.5278
Discrete SNNE (only_num)	0.2557	0.2358	0.3320	0.2980	0.5860	0.7178	0.5271	0.6366	0.4643	0.5714	0.6269	0.5611
Discrete SNNE (only_denom)	0.2600	0.2375	0.3517	0.3209	0.5922	0.7195	0.5247	0.6398	0.4701	0.5749	0.6322	0.5653
Semantic density	0.2435	0.1979	0.3205	0.2891	0.5246	0.6272	0.4534	0.5255	0.3884	0.5083	0.5698	0.4930
U_{eig}	0.2549	0.2379	0.3370	0.3014	0.5850	0.7114	0.5206	0.6325	0.4569	0.5619	0.6216	0.5518
U_{deg}	0.2553	0.2371	0.3361	0.3007	0.5821	0.7100	0.5184	0.6326	0.4563	0.5631	0.6237	0.5514
C_{ecc}	0.2529	0.2280	0.3210	0.2962	0.5824	0.7064	0.5132	0.6244	0.4557	0.5580	0.6229	0.5492
LUQ	0.2575	0.2403	0.3372	0.3025	0.5819	0.7009	0.5181	0.6238	0.4553	0.5605	0.6184	0.5504
SNNE	0.2556	0.2389	0.3365	0.3010	0.5832	0.7100	0.5196	0.6321	0.4561	0.5630	0.6214	0.5515
SNNE (only_num)	0.2555	0.2390	0.3365	0.3010	0.5832	0.7100	0.5196	0.6322	0.4561	0.5626	0.6214	0.5514
SNNE (only_denom)	0.2556	0.2384	0.3364	0.3009	0.5833	0.7100	0.5202	0.6308	0.4573	0.5626	0.6216	0.5514

Table 21: SUQ experimental results on Summarization and Translation (Part II). We report PRR for summarization and translation.

Uncertainty Method	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
QWEN 3-32B												
Discrete semantic entropy	0.2403	0.2719	0.3262	0.2543	0.5752	0.7020	0.5257	0.6464	0.4692	0.5819	0.6250	0.5419
Discrete semantic density	0.2379	0.2050	0.3214	0.2526	0.5489	0.6425	0.4763	0.5611	0.4063	0.5200	0.5825	0.5108
Discrete U_{eig}	0.2373	0.2738	0.3269	0.2541	0.5784	0.7088	0.5260	0.6479	0.4711	0.5778	0.6217	0.5497
Discrete U_{deg}	0.2368	0.2719	0.3247	0.2527	0.5885	0.7091	0.5354	0.6514	0.4698	0.5653	0.6213	0.5493
Discrete C_{ecc}	0.2366	0.2671	0.3245	0.2555	0.5811	0.6980	0.5202	0.6460	0.4687	0.5804	0.6163	0.5476
KLE-heat$_{t=0.1}$	0.2373	0.2733	0.3248	0.2527	0.5782	0.7091	0.5350	0.6462	0.4699	0.5807	0.6206	0.5487
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2369	0.2725	0.3247	0.2528	0.5785	0.7095	0.5257	0.6454	0.4685	0.5812	0.6232	0.5497
Discrete LUQ-pair	0.2379	0.2629	0.3214	0.2507	0.5640	0.6834	0.5129	0.6153	0.4540	0.5572	0.6157	0.5275
Discrete SNNE	0.2378	0.2731	0.3216	0.2520	0.5639	0.6915	0.5219	0.6293	0.4613	0.5689	0.6209	0.5388
Discrete SNNE (only_num)	0.2405	0.2755	0.3304	0.2590	0.6205	0.7469	0.5579	0.6885	0.5083	0.5984	0.6582	0.5840
Discrete SNNE (only_denom)	0.2482	0.2809	0.3612	0.2867	0.6258	0.7466	0.5627	0.6994	0.5177	0.6047	0.6603	0.5915
Semantic density	0.2310	0.2198	0.3229	0.2512	0.5418	0.6407	0.4710	0.5612	0.4174	0.5255	0.5812	0.5019
U_{eig}	0.2400	0.2741	0.3363	0.2629	0.6066	0.7306	0.5483	0.6698	0.4845	0.5828	0.6405	0.5702
U_{deg}	0.2400	0.2736	0.3353	0.2623	0.6036	0.7290	0.5466	0.6685	0.4833	0.5829	0.6400	0.5683
C_{ecc}	0.2370	0.2685	0.3251	0.2563	0.6000	0.7252	0.5435	0.6620	0.4842	0.5796	0.6355	0.5692
LUQ	0.2421	0.2737	0.3365	0.2635	0.6013	0.7263	0.5473	0.6675	0.4809	0.5847	0.6350	0.5653
SNNE	0.2404	0.2741	0.3356	0.2625	0.6050	0.7286	0.5473	0.6687	0.4823	0.5823	0.6403	0.5682
SNNE (only_num)	0.2403	0.2741	0.3355	0.2625	0.6042	0.7284	0.5468	0.6686	0.4822	0.5827	0.6391	0.5680
SNNE (only_denom)	0.2403	0.2742	0.3354	0.2624	0.6056	0.7287	0.5469	0.6694	0.4832	0.5841	0.6404	0.5696
MISTRAL-7B-INSTRUCT-V0.2												
Discrete semantic entropy	0.2031	0.2304	0.3328	0.3275	0.5264	0.6521	0.4165	0.2518	0.2313	0.3846	0.5791	0.4665
Discrete semantic density	0.2017	0.2288	0.3324	0.3279	0.5347	0.6601	0.4230	0.2417	0.2251	0.3828	0.5829	0.4680
Discrete U_{eig}	0.2010	0.2303	0.3365	0.3292	0.5274	0.6538	0.4185	0.2391	0.2243	0.3797	0.5767	0.4696
Discrete U_{deg}	0.2010	0.2277	0.3350	0.3283	0.5322	0.6519	0.4180	0.2443	0.2284	0.3804	0.5735	0.4682
Discrete C_{ecc}	0.1957	0.2240	0.3327	0.3299	0.5348	0.6505	0.4135	0.2295	0.2062	0.3743	0.5750	0.4704
KLE-heat$_{t=0.1}$	0.2013	0.2281	0.3348	0.3282	0.5293	0.6520	0.4165	0.2475	0.2316	0.3817	0.5744	0.4717
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2010	0.2280	0.3350	0.3283	0.5318	0.6531	0.4200	0.2457	0.2296	0.3831	0.5752	0.4683
Discrete LUQ-pair	0.2030	0.2310	0.3316	0.3267	0.5034	0.6299	0.3925	0.2320	0.2187	0.3574	0.5634	0.4447
Discrete SNNE	0.2006	0.2161	0.3328	0.3273	0.5125	0.6372	0.4056	0.2455	0.2270	0.3733	0.5696	0.4516
Discrete SNNE (only_num)	0.2036	0.2276	0.3385	0.3328	0.5554	0.6899	0.4386	0.2564	0.2364	0.3958	0.6051	0.4873
Discrete SNNE (only_denom)	0.2128	0.2598	0.3660	0.3620	0.5603	0.6909	0.4524	0.2685	0.2511	0.4151	0.6074	0.5012
Semantic density	0.1991	0.2453	0.3189	0.3318	0.5385	0.6701	0.4163	0.2302	0.2147	0.3644	0.5936	0.4682
U_{eig}	0.2042	0.2390	0.3462	0.3372	0.5427	0.6741	0.4249	0.2418	0.2258	0.3802	0.5941	0.4748
U_{deg}	0.2045	0.2404	0.3457	0.3368	0.5457	0.6740	0.4258	0.2467	0.2299	0.3833	0.5949	0.4765
U_{ecc}	0.1924	0.1825	0.3202	0.3206	0.4683	0.5706	0.3360	0.1721	0.1687	0.2849	0.5178	0.3962
LUQ	0.2054	0.2452	0.3438	0.3355	0.5441	0.6714	0.4231	0.2534	0.2359	0.3869	0.5884	0.4770
SNNE	0.2044	0.2417	0.3455	0.3367	0.5453	0.6739	0.4252	0.2445	0.2306	0.3813	0.5948	0.4761
SNNE (only_num)	0.2044	0.2415	0.3455	0.3367	0.5431	0.6739	0.4245	0.2448	0.2307	0.3814	0.5949	0.4756
SNNE (only_denom)	0.2047	0.2413	0.3455	0.3367	0.5441	0.6740	0.4259	0.2484	0.2315	0.3843	0.5946	0.4769

Table 21: SUQ experimental results on Summarization and Translation (Part II). We report PRR for summarization and translation. (Continued.)

Uncertainty Method	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
MISTRAL-NEMO-INSTRUCT-2407 12B												
Discrete semantic entropy	0.2557	0.2297	0.3069	0.2283	0.5390	0.6773	0.4784	0.4661	0.4139	0.5010	0.5890	0.5028
Discrete semantic density	0.2505	0.2334	0.3069	0.2284	0.5479	0.6858	0.4874	0.4615	0.4182	0.5036	0.6132	0.5064
Discrete U_{eig}	0.2520	0.2340	0.3093	0.2311	0.5422	0.6814	0.4770	0.4571	0.4132	0.5019	0.6031	0.5070
Discrete U_{deg}	0.2517	0.2334	0.3087	0.2305	0.5368	0.6757	0.4789	0.4593	0.4163	0.4988	0.6017	0.5005
Discrete C_{ecc}	0.2433	0.2122	0.2981	0.2267	0.5379	0.6752	0.4799	0.4451	0.4111	0.4968	0.5995	0.5037
KLE-heat$_{t=0.1}$	0.2522	0.2334	0.3086	0.2303	0.5412	0.6773	0.4859	0.4619	0.4165	0.5044	0.6021	0.5015
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2518	0.2334	0.3087	0.2303	0.5374	0.6762	0.4792	0.4627	0.4139	0.5045	0.6011	0.5018
Discrete LUQ-pair	0.2554	0.2258	0.3040	0.2279	0.5125	0.6533	0.4536	0.4337	0.3927	0.4684	0.5909	0.4802
Discrete SNNE	0.2521	0.2220	0.3057	0.2268	0.5220	0.6612	0.4633	0.4475	0.4029	0.4935	0.5895	0.4871
Discrete SNNE (only_num)	0.2573	0.2314	0.3116	0.2309	0.5706	0.7199	0.5052	0.4740	0.4391	0.5208	0.6360	0.5362
Discrete SNNE (only_denom)	0.2720	0.2451	0.3300	0.2459	0.5803	0.7229	0.5164	0.4919	0.4540	0.5285	0.6398	0.5473
Semantic density	0.2575	0.2226	0.2970	0.2286	0.5530	0.7029	0.4852	0.4391	0.4105	0.4887	0.6234	0.5135
U_{eig}	0.2538	0.2367	0.3149	0.2355	0.5529	0.7021	0.4922	0.4612	0.4229	0.5062	0.6187	0.5151
U_{deg}	0.2538	0.2373	0.3150	0.2353	0.5524	0.7007	0.4929	0.4650	0.4223	0.5063	0.6183	0.5144
U_{ecc}	0.2489	0.1593	0.2922	0.2184	0.4782	0.5956	0.3892	0.3216	0.3363	0.3878	0.5508	0.4474
LUQ	0.2576	0.2351	0.3140	0.2331	0.5513	0.7018	0.4950	0.4681	0.4227	0.5067	0.6210	0.5123
SNNE	0.2542	0.2375	0.3151	0.2352	0.5522	0.6997	0.4922	0.4637	0.4221	0.5061	0.6182	0.5138
SNNE (only_num)	0.2541	0.2375	0.3151	0.2353	0.5523	0.7014	0.4922	0.4639	0.4224	0.5059	0.6185	0.5145
SNNE (only_denom)	0.2541	0.2375	0.3151	0.2353	0.5528	0.7006	0.4933	0.4665	0.4235	0.5067	0.6182	0.5146
MISTRAL-SMALL 3.2 24B												
Discrete semantic entropy	0.2281	0.2870	0.3265	0.2640	0.5892	0.7180	0.5277	0.6144	0.4640	0.5541	0.6277	0.5421
Discrete semantic density	0.2242	0.2869	0.3264	0.2638	0.5880	0.7308	0.5409	0.6287	0.4755	0.5639	0.6368	0.5558
Discrete U_{eig}	0.2246	0.2864	0.3293	0.2649	0.5877	0.7211	0.5224	0.6167	0.4620	0.5673	0.6298	0.5459
Discrete U_{deg}	0.2243	0.2850	0.3269	0.2639	0.5871	0.7210	0.5261	0.6214	0.4637	0.5683	0.6301	0.5443
Discrete C_{ecc}	0.2120	0.2749	0.3273	0.2649	0.5781	0.7202	0.5211	0.6172	0.4647	0.5618	0.6297	0.5481
KLE-heat$_{t=0.1}$	0.2250	0.2862	0.3269	0.2639	0.5847	0.7210	0.5237	0.6145	0.4665	0.5669	0.6302	0.5410
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2243	0.2849	0.3270	0.2639	0.5874	0.7210	0.5226	0.6179	0.4636	0.5667	0.6303	0.5436
Discrete LUQ-pair	0.2237	0.2670	0.3235	0.2615	0.5591	0.6997	0.5122	0.5829	0.4561	0.5373	0.6266	0.5184
Discrete SNNE	0.2237	0.2696	0.3253	0.2635	0.5646	0.7059	0.5138	0.6037	0.4629	0.5514	0.6249	0.5328
Discrete SNNE (only_num)	0.2274	0.2857	0.3335	0.2685	0.6189	0.7584	0.5593	0.6530	0.4993	0.5876	0.6611	0.5787
Discrete SNNE (only_denom)	0.2332	0.2995	0.3610	0.2885	0.6218	0.7607	0.5654	0.6656	0.5008	0.5912	0.6631	0.5848
Semantic density	0.2288	0.2769	0.3147	0.2685	0.6047	0.7428	0.5505	0.6389	0.4774	0.5691	0.6500	0.5654
U_{eig}	0.2267	0.2892	0.3395	0.2728	0.6029	0.7476	0.5494	0.6360	0.4877	0.5748	0.6511	0.5627
U_{deg}	0.2270	0.2898	0.3387	0.2724	0.6035	0.7439	0.5492	0.6366	0.4861	0.5741	0.6492	0.5605
C_{ecc}	0.2116	0.2639	0.3263	0.2657	0.6006	0.7401	0.5473	0.6306	0.4824	0.5647	0.6484	0.5606
LUQ	0.2304	0.2841	0.3379	0.2725	0.6067	0.7413	0.5510	0.6391	0.4839	0.5763	0.6496	0.5631
SNNE	0.2276	0.2895	0.3387	0.2726	0.6054	0.7461	0.5493	0.6353	0.4874	0.5742	0.6497	0.5612
SNNE (only_num)	0.2275	0.2896	0.3387	0.2725	0.6045	0.7461	0.5493	0.6367	0.4869	0.5752	0.6502	0.5607
SNNE (only_denom)	0.2275	0.2902	0.3387	0.2724	0.6045	0.7461	0.5484	0.6375	0.4869	0.5741	0.6502	0.5605

Table 21: SUQ experimental results on Summarization and Translation (Part II). We report PRR for summarization and translation. (Continued.)

Uncertainty Method	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
GPT-4O-MINI												
Discrete semantic entropy	0.1960	0.2723	0.3136	0.2602	0.5819	0.7203	0.5420	0.6503	0.4880	0.5942	0.6397	0.5487
Discrete semantic density	0.1943	0.2736	0.3131	0.2599	0.5970	0.7328	0.5523	0.6612	0.5011	0.5842	0.6449	0.5602
Discrete U_{eig}	0.1950	0.2759	0.3141	0.2611	0.5787	0.7244	0.5447	0.6507	0.4916	0.5968	0.6407	0.5478
Discrete U_{deg}	0.1940	0.2742	0.3129	0.2599	0.5785	0.7259	0.5435	0.6504	0.4915	0.5968	0.6407	0.5484
Discrete C_{ecc}	0.1914	0.2687	0.3167	0.2681	0.5805	0.7249	0.5427	0.6482	0.4916	0.5970	0.6403	0.5497
KLE-heat$_{t=0.1}$	0.1942	0.2750	0.3129	0.2599	0.5785	0.7258	0.5436	0.6506	0.4918	0.5969	0.6407	0.5486
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1940	0.2736	0.3129	0.2599	0.5785	0.7258	0.5431	0.6502	0.4915	0.5969	0.6407	0.5501
Discrete LUQ-pair	0.1934	0.2656	0.3067	0.2570	0.5762	0.7100	0.5359	0.6391	0.4862	0.5888	0.6372	0.5309
Discrete SNNE	0.1947	0.2696	0.3118	0.2597	0.5764	0.7130	0.5348	0.6394	0.4885	0.5869	0.6375	0.5342
Discrete SNNE (only_num)	0.1975	0.2814	0.3199	0.2669	0.6181	0.7591	0.5698	0.6939	0.5238	0.6158	0.6690	0.5877
Discrete SNNE (only_denom)	0.2027	0.2820	0.3482	0.3012	0.6182	0.7640	0.5740	0.6945	0.5267	0.6190	0.6704	0.5917
Semantic density	0.1975	0.2703	0.3017	0.2669	0.6081	0.7521	0.5662	0.6842	0.5075	0.6175	0.6617	0.5723
U_{eig}	0.1971	0.2786	0.3197	0.2670	0.6075	0.7512	0.5666	0.6848	0.5132	0.6124	0.6590	0.5732
U_{deg}	0.1966	0.2764	0.3192	0.2666	0.6073	0.7509	0.5667	0.6821	0.5103	0.6114	0.6623	0.5712
U_{ecc}	0.1852	0.2013	0.3033	0.2594	0.5661	0.6682	0.5049	0.5674	0.4526	0.5635	0.6054	0.5115
LUQ	0.1971	0.2758	0.3191	0.2669	0.6087	0.7437	0.5649	0.6822	0.5126	0.6098	0.6607	0.5672
SNNE	0.1967	0.2792	0.3192	0.2666	0.6073	0.7493	0.5655	0.6856	0.5122	0.6099	0.6624	0.5705
SNNE (only_num)	0.1967	0.2788	0.3192	0.2666	0.6073	0.7517	0.5655	0.6856	0.5080	0.6130	0.6624	0.5711
SNNE (only_denom)	0.1968	0.2781	0.3192	0.2666	0.6073	0.7527	0.5656	0.6857	0.5093	0.6130	0.6624	0.5712

Table 22: SUQ experimental results on Summarization and Translation (Part III). We report PRR for summarization and translation.

Uncertainty Method	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
PHI 4-14B												
Discrete semantic entropy	0.1942	0.1322	0.3281	0.2677	0.5675	0.6937	0.5121	0.5486	0.4155	0.5287	0.6082	0.5168
Discrete semantic density	0.1929	0.1342	0.3289	0.2655	0.5793	0.7075	0.5239	0.5563	0.4168	0.5283	0.6257	0.5299
Discrete U_{eig}	0.1916	0.1361	0.3303	0.2677	0.5696	0.7012	0.5071	0.5554	0.4183	0.5255	0.6143	0.5250
Discrete U_{deg}	0.1915	0.1352	0.3290	0.2666	0.5641	0.7013	0.5060	0.5537	0.4169	0.5299	0.6132	0.5263
Discrete C_{ecc}	0.1943	0.1232	0.3385	0.2793	0.5598	0.6981	0.5124	0.5484	0.4189	0.5251	0.6149	0.5240
KLE-heat$_{t=0.1}$	0.1917	0.1355	0.3291	0.2665	0.5643	0.7014	0.5047	0.5557	0.4151	0.5260	0.6140	0.5234
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1915	0.1351	0.3291	0.2667	0.5695	0.7014	0.5106	0.5524	0.4166	0.5272	0.6128	0.5227
Discrete LUQ-pair	0.1921	0.1266	0.3223	0.2662	0.5512	0.6698	0.4925	0.5237	0.3824	0.4904	0.5904	0.4919
Discrete SNNE	0.1917	0.1349	0.3291	0.2665	0.5440	0.6796	0.4939	0.5435	0.3999	0.5088	0.6059	0.5027
Discrete SNNE (only_num)	0.1948	0.1401	0.3373	0.2742	0.5986	0.7393	0.5454	0.5809	0.4393	0.5502	0.6484	0.5598
Discrete SNNE (only_denom)	0.2065	0.1412	0.3768	0.3172	0.6065	0.7424	0.5541	0.5960	0.4583	0.5565	0.6538	0.5700
Semantic density	0.1955	0.1343	0.3188	0.2777	0.5859	0.7238	0.5366	0.5594	0.4084	0.5234	0.6400	0.5395
U_{eig}	0.1949	0.1397	0.3384	0.2763	0.5820	0.7219	0.5320	0.5663	0.4240	0.5334	0.6330	0.5358
U_{deg}	0.1950	0.1403	0.3383	0.2760	0.5826	0.7194	0.5314	0.5666	0.4251	0.5332	0.6318	0.5362
U_{ecc}	0.1931	0.1100	0.3275	0.2679	0.5088	0.6181	0.4507	0.4283	0.3282	0.4452	0.5653	0.4717
LUQ	0.1957	0.1408	0.3385	0.2761	0.5816	0.7144	0.5288	0.5636	0.4245	0.5362	0.6344	0.5386
SNNE	0.1950	0.1404	0.3384	0.2760	0.5827	0.7207	0.5296	0.5658	0.4239	0.5336	0.6315	0.5356
SNNE (only_num)	0.1951	0.1405	0.3383	0.2760	0.5826	0.7213	0.5303	0.5661	0.4240	0.5326	0.6330	0.5351
SNNE (only_denom)	0.1951	0.1406	0.3383	0.2760	0.5823	0.7193	0.5313	0.5673	0.4251	0.5337	0.6327	0.5362
PHI4-MINI-INSTRUCT 3.8B												
Discrete semantic entropy	0.2266	0.2017	0.3442	0.3011	0.3101	0.4527	0.3703	0.0662	0.2873	0.2272	0.5275	0.4155
Discrete semantic density	0.2262	0.2100	0.3421	0.3366	0.3187	0.4492	0.3666	0.0705	0.2847	0.2422	0.5316	0.4176
Discrete U_{eig}	0.2269	0.2102	0.3444	0.3365	0.3167	0.4466	0.3641	0.0754	0.2841	0.2423	0.5285	0.4169
Discrete U_{deg}	0.2265	0.2085	0.3444	0.3363	0.3148	0.4465	0.3675	0.0711	0.2889	0.2421	0.5292	0.4178
Discrete C_{ecc}	0.2282	0.2113	0.3449	0.3435	0.3078	0.4272	0.3579	0.0799	0.2718	0.2346	0.5127	0.4158
KLE-heat$_{t=0.1}$	0.2264	0.2083	0.3447	0.3361	0.3170	0.4478	0.3705	0.0695	0.2914	0.2428	0.5286	0.4191
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2265	0.2088	0.3444	0.3362	0.3169	0.4464	0.3676	0.0714	0.2889	0.2438	0.5298	0.4170
Discrete LUQ-pair	0.2275	0.2048	0.3425	0.3089	0.2711	0.3770	0.3487	0.0687	0.2658	0.2213	0.4984	0.3971
Discrete SNNE	0.2259	0.2008	0.3405	0.3141	0.3137	0.4434	0.3637	0.0673	0.2824	0.2350	0.5124	0.4076
Discrete SNNE (only_num)	0.2283	0.2104	0.3473	0.3335	0.3324	0.4685	0.3838	0.0753	0.2997	0.2503	0.5510	0.4375
Discrete SNNE (only_denom)	0.2401	0.2410	0.3848	0.3723	0.3501	0.4720	0.4054	0.1207	0.3244	0.2743	0.5669	0.4587
Semantic density	0.2282	0.1985	0.3325	0.3493	0.3100	0.4382	0.3579	0.0844	0.2685	0.2050	0.5382	0.4096
U_{eig}	0.2300	0.2208	0.3523	0.3402	0.3284	0.4573	0.3694	0.0820	0.2896	0.2487	0.5386	0.4222
U_{deg}	0.2301	0.2214	0.3525	0.3400	0.3339	0.4630	0.3739	0.0800	0.2946	0.2515	0.5403	0.4241
U_{ecc}	0.2295	0.1704	0.3322	0.1995	0.1407	0.2192	0.2607	0.0842	0.1927	0.1387	0.4185	0.3183
LUQ	0.2296	0.2175	0.3510	0.3425	0.3371	0.4636	0.3768	0.0805	0.2976	0.2564	0.5431	0.4264
SNNE	0.2301	0.2210	0.3524	0.3401	0.3310	0.4612	0.3700	0.0804	0.2916	0.2499	0.5405	0.4229
SNNE (only_num)	0.2301	0.2211	0.3525	0.3401	0.3316	0.4620	0.3705	0.0803	0.2920	0.2499	0.5403	0.4230
SNNE (only_denom)	0.2301	0.2215	0.3526	0.3400	0.3356	0.4650	0.3750	0.0798	0.2961	0.2527	0.5411	0.4244

Table 22: SUQ experimental results on Summarization and Translation (Part III). We report PRR for summarization and translation. (Continued.)

Uncertainty Method	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
PHI 3-14B												
Discrete semantic entropy	0.2572	0.1976	0.2969	0.2823	0.5023	0.6530	0.3333	0.0871	0.1418	0.2798	0.5906	0.4588
Discrete semantic density	0.2542	0.2095	0.2979	0.3087	0.5120	0.6855	0.3533	0.0910	0.1459	0.2882	0.6062	0.4795
Discrete U_{eig}	0.2555	0.2083	0.2970	0.3100	0.5057	0.6678	0.3477	0.0922	0.1461	0.2836	0.5988	0.4725
Discrete U_{deg}	0.2554	0.2052	0.2970	0.3105	0.5081	0.6651	0.3469	0.0928	0.1467	0.2854	0.5968	0.4722
Discrete C_{ecc}	0.2542	0.2154	0.3009	0.3134	0.5132	0.6713	0.3608	0.0935	0.1520	0.3015	0.5924	0.4732
KLE-heat$_{t=0.1}$	0.2557	0.2045	0.2971	0.3084	0.5091	0.6663	0.3472	0.0917	0.1459	0.2869	0.6003	0.4708
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2553	0.2054	0.2970	0.3095	0.5069	0.6663	0.3463	0.0921	0.1458	0.2858	0.5974	0.4721
Discrete LUQ-pair	0.2598	0.2035	0.2940	0.2984	0.4677	0.6275	0.3267	0.1101	0.1646	0.2747	0.5798	0.4429
Discrete SNNE	0.2569	0.1927	0.2946	0.2931	0.4850	0.6376	0.3177	0.0883	0.1471	0.2606	0.5823	0.4458
Discrete SNNE (only_num)	0.2605	0.2003	0.2992	0.3048	0.5354	0.7182	0.3888	0.1022	0.1750	0.3112	0.6349	0.5130
Discrete SNNE (only_denom)	0.2764	0.2439	0.3322	0.3282	0.5549	0.7234	0.4388	0.1577	0.2296	0.3791	0.6414	0.5305
Semantic density	0.2588	0.2030	0.2836	0.2692	0.5163	0.7011	0.3662	0.0927	0.1491	0.2945	0.6214	0.4876
U_{eig}	0.2616	0.2138	0.3058	0.3165	0.5164	0.7010	0.3661	0.0932	0.1499	0.2936	0.6219	0.4893
U_{deg}	0.2620	0.2129	0.3057	0.3154	0.5172	0.7012	0.3673	0.0933	0.1499	0.2958	0.6214	0.4894
U_{ecc}	0.2584	0.1783	0.2883	0.1599	0.4391	0.5695	0.3162	0.1273	0.1834	0.2697	0.5404	0.4291
LUQ	0.2632	0.2102	0.3047	0.3123	0.5260	0.7058	0.3815	0.0971	0.1559	0.3112	0.6201	0.4979
SNNE	0.2622	0.2121	0.3057	0.3160	0.5164	0.7009	0.3677	0.0935	0.1501	0.2957	0.6208	0.4901
SNNE (only_num)	0.2622	0.2123	0.3057	0.3160	0.5159	0.7011	0.3679	0.0935	0.1500	0.2956	0.6214	0.4891
SNNE (only_denom)	0.2622	0.2125	0.3057	0.3147	0.5170	0.7007	0.3684	0.0934	0.1502	0.2967	0.6215	0.4891

H.4 CONSISTENCY BASELINES

Discussion: Appendix C.

Table 23: Non-semantic UQ methods on QA and Math. We report AUROC/AUARC.

Uncertainty Method	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT							
exact match	0.7688/0.8799	0.7316/0.9414	0.6961/0.6717	0.6573/0.7544	0.6655/0.4306	0.6718/0.6287	0.5023/0.8161
BERT similarity	0.7875/0.8855	0.6456/0.9290	0.3644/0.4174	0.5757/0.7179	0.5827/0.3784	0.5852/0.5870	0.5676/0.8438
cosine similarity	0.8294/0.8960	0.6896/0.9358	0.3672/0.4095	0.5979/0.7268	0.6394/0.4091	0.6124/0.6057	0.6348/0.8727
LLAMA 3.1-70B INSTRUCT							
exact match	0.7137/0.9434	0.7523/0.9531	0.6629/0.7640	0.6475/0.7630	0.6385/0.5351	0.7999/0.9382	0.5000/0.9246
BERT similarity	0.7639/0.9524	0.6635/0.9370	0.3812/0.5462	0.5956/0.7383	0.5532/0.4779	0.7852/0.9278	0.5050/0.9168
cosine similarity	0.8109/0.9609	0.7398/0.9523	0.4423/0.6076	0.6261/0.7594	0.6188/0.5232	0.7871/0.9346	0.5808/0.937
LLAMA 3.1-405B							
exact match	0.7373/0.9609	0.7506/0.9514	0.5635/0.6512	0.6656/0.7941	0.6141/0.5609	0.7579/0.9150	0.5000/0.9464
BERT similarity	0.7390/0.9626	0.6671/0.9409	0.4984/0.6077	0.5911/0.7631	0.4796/0.4762	0.7338/0.9316	0.5322/0.9473
cosine similarity	0.7652/0.9664	0.7241/0.9499	0.5399/0.6392	0.6164/0.7738	0.5139/0.4980	0.7505/0.9288	0.5943/0.9560
QWEN 3-1.7B							
exact match	0.6311/0.6131	0.7284/0.9056	0.5395/0.5413	0.6381/0.7332	0.6283/0.2709	0.5499/0.9223	0.5000/0.6186
BERT similarity	0.5421/0.5096	0.6578/0.8924	0.6138/0.5881	0.6384/0.7260	0.5958/0.2468	0.5725/0.9194	0.5495/0.6735
cosine similarity	0.6796/0.6281	0.6826/0.9009	0.6720/0.6327	0.6809/0.7447	0.6256/0.2607	0.6876/0.9483	0.6606/0.7410
QWEN 3-8B							
exact match	0.7532/0.7570	0.6758/0.8927	0.6434/0.6350	0.6429/0.7052	0.6851/0.3264	0.6111/0.9390	0.5012/0.6573
BERT similarity	0.7746/0.7666	0.6501/0.9023	0.5919/0.6047	0.6153/0.7019	0.6541/0.3130	0.6052/0.9457	0.5801/0.6910
cosine similarity	0.8063/0.7776	0.6850/0.9064	0.5861/0.6043	0.6312/0.7067	0.6825/0.3236	0.7143/0.9583	0.6894/0.7599

Table 24: Non semantical UQ on Summarization + Translation. We report PRR.

Uncertainty Method/ PRR	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT												
exact match	0.2496	0.1682	0.3149	0.2764	0.5804	0.7125	0.5059	0.5275	0.3850	0.4959	0.6193	0.5083
BERT similarity	0.2533	0.1738	0.3542	0.3014	0.5963	0.7313	0.5316	0.5613	0.4393	0.5367	0.6328	0.5423
cosine similarity	0.2610	0.1836	0.3557	0.3005	0.5918	0.7311	0.5308	0.5619	0.4371	0.5355	0.6328	0.5392
LLAMA 3.1-70B INSTRUCT												
exact match	0.2592	0.2110	0.3235	0.3022	0.6090	0.7387	0.5415	0.6192	0.4650	0.5651	0.6503	0.5423
BERT similarity	0.2702	0.1861	0.3516	0.3149	0.6375	0.7636	0.5719	0.6579	0.5253	0.6118	0.6727	0.5823
cosine similarity	0.2754	0.2198	0.3606	0.3212	0.6334	0.7639	0.5733	0.6663	0.5224	0.6123	0.6685	0.5819
LLAMA 3.1-405B INSTRUCT FP8												
exact match	0.2851	0.2732	0.3391	0.3120	0.6310	0.7520	0.5715	0.6814	0.4894	0.6306	0.6668	0.5778
BERT similarity	0.2979	0.2868	0.3793	0.3277	0.6513	0.7767	0.5937	0.7109	0.5364	0.6497	0.6857	0.6084
cosine similarity	0.2992	0.2948	0.3854	0.3323	0.6466	0.7729	0.5911	0.7121	0.5287	0.6492	0.6837	0.6081
QWEN 3-1.7B												
exact match	0.2511	0.2502	0.3305	0.3110	0.5010	0.6461	0.3916	0.4901	0.3333	0.3744	0.5717	0.4998
BERT similarity	0.2526	0.2465	0.3599	0.3236	0.5226	0.6735	0.4230	0.5150	0.3637	0.4120	0.5869	0.5270
cosine similarity	0.2539	0.2591	0.3631	0.3263	0.5205	0.6695	0.4214	0.5140	0.3629	0.4103	0.5863	0.5230
QWEN 3-8B												
exact match	0.2566	0.2285	0.3132	0.2958	0.5819	0.7057	0.5142	0.6274	0.4453	0.5625	0.6214	0.5463
BERT similarity	0.2589	0.2348	0.3428	0.3123	0.5920	0.7151	0.5290	0.6368	0.4716	0.5737	0.6271	0.5641
cosine similarity	0.2598	0.2348	0.3462	0.3166	0.5916	0.7168	0.5268	0.6365	0.4686	0.5717	0.6258	0.5598

H.5 NLI-Backbone Comparisons

Discussion: Appendix C.2.

Table 25: UQ using Llama3-8B as an NLI tool on QA and Math. We report AUROC/AUARC

Generation Model	QA					Math	
	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
Llama 3.1-8B Instruct							
Discrete semantic entropy	0.8344/0.8874	0.6054/0.9114	0.5703/0.5983	0.6741/0.7429	0.7018/0.4303	0.7436/0.6731	0.5167/0.8658
Discrete semantic density	0.8452/0.8966	0.6773/0.9208	0.6164/0.6357	0.6858/0.7533	0.7176/0.4535	0.7473/0.6517	0.3792/0.7869
Discrete U_{eig}	0.8385/0.8923	0.6290/0.9179	0.6106/0.6142	0.6900/0.7571	0.7071/0.4287	0.7439/0.6735	0.6203/0.8677
Discrete U_{deg}	0.8388/0.8911	0.6275/0.9177	0.6089/0.6121	0.6859/0.7524	0.7050/0.4202	0.7461/0.6711	0.6197/0.8951
Discrete C_{ecc}	0.6392/0.7734	0.6165/0.9146	0.5936/0.5702	0.6279/0.7321	0.6097/0.3584	0.5743/0.5327	0.6191/0.8769
KLE-heat$_{t=0.1}$	0.8391/0.8933	0.6275/0.9178	0.6100/0.6124	0.6855/0.7543	0.7053/0.4218	0.7463/0.6757	0.6201/0.8729
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8386/0.8913	0.6275/0.9178	0.6078/0.6119	0.6861/0.7517	0.7048/0.4271	0.7455/0.6591	0.6195/0.8952
Discrete LUQ-pair	0.7495/0.8510	0.5249/0.8956	0.5608/0.4974	0.5703/0.6812	0.6421/0.3733	0.6748/0.5338	0.4996/0.8566
Discrete SNNE	0.8245/0.8828	0.5709/0.9056	0.5800/0.5882	0.6496/0.7253	0.6826/0.4048	0.7274/0.6491	0.5167/0.8657
Discrete SNNE (only_denom)	0.8247/0.8953	0.7196/0.9405	0.5978/0.5972	0.6621/0.7489	0.6797/0.4273	0.7138/0.6459	0.7914/0.9492
Llama 3.1-70B Instruct							
Discrete semantic entropy	0.7598/0.9460	0.5980/0.9151	0.5551/0.6356	0.6533/0.7524	0.6387/0.4931	0.8104/0.9266	0.5240/0.9124
Discrete semantic density	0.7909/0.9518	0.6912/0.9319	0.6093/0.7015	0.6665/0.7647	0.6716/0.5455	0.8428/0.9422	0.4570/0.8955
Discrete U_{eig}	0.7711/0.9495	0.6394/0.9265	0.5594/0.6547	0.6703/0.7581	0.6502/0.5084	0.8311/0.9304	0.5430/0.9162
Discrete U_{deg}	0.7708/0.9482	0.6368/0.9262	0.5585/0.6531	0.6661/0.7607	0.6492/0.5173	0.8308/0.9314	0.5430/0.9162
Discrete C_{ecc}	0.6702/0.9277	0.6409/0.9273	0.5554/0.6711	0.5995/0.7388	0.6194/0.4999	0.7483/0.8644	0.5430/0.9162
KLE-heat$_{t=0.1}$	0.7708/0.9484	0.6368/0.9263	0.5574/0.6540	0.6657/0.7573	0.6492/0.5182	0.8318/0.9256	0.5431/0.9162
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7708/0.9481	0.6368/0.9262	0.5591/0.6530	0.6663/0.7560	0.6491/0.5180	0.8303/0.9299	0.5430/0.9162
Discrete LUQ-pair	0.6789/0.9385	0.5206/0.9050	0.5245/0.5584	0.5801/0.7145	0.5747/0.4756	0.6102/0.8439	0.5000/0.9246
Discrete SNNE	0.7433/0.9500	0.5718/0.9138	0.5260/0.5826	0.6313/0.7366	0.6144/0.4855	0.7003/0.8618	0.5039/0.9106
Discrete SNNE (only_denom)	0.7699/0.9521	0.7259/0.9447	0.5554/0.6585	0.6488/0.7566	0.6591/0.5372	0.8292/0.9294	0.7395/0.9621
Llama 3.1-405B Instruct FP8							
Discrete semantic entropy	0.6786/0.9553	0.5485/0.9109	0.5351/0.5962	0.5728/0.7440	0.5802/0.5217	0.6867/0.9227	0.5087/0.9108
Discrete semantic density	0.7254/0.9570	0.6614/0.9263	0.5762/0.6731	0.6533/0.7719	0.6259/0.5615	0.7784/0.9483	0.5172/0.9104
Discrete U_{eig}	0.6759/0.9441	0.5926/0.9129	0.5677/0.6541	0.6657/0.7794	0.5806/0.5070	0.7544/0.9369	0.5149/0.9058
Discrete U_{deg}	0.7001/0.9573	0.5706/0.9146	0.5476/0.6262	0.5878/0.7490	0.6004/0.5272	0.6989/0.9187	0.5192/0.9113
Discrete C_{ecc}	0.6687/0.9517	0.5789/0.9158	0.5242/0.5868	0.5778/0.7476	0.5827/0.5183	0.7081/0.9215	0.5193/0.9113
KLE-heat$_{t=0.1}$	0.6990/0.9572	0.5698/0.9175	0.5537/0.6118	0.5893/0.7487	0.5973/0.5304	0.6995/0.9187	0.5171/0.9104
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.6990/0.9572	0.5697/0.9174	0.5520/0.6127	0.5892/0.7480	0.5976/0.5240	0.6993/0.9131	0.5171/0.9104
Discrete LUQ-pair	0.6206/0.9455	0.5090/0.9042	0.5186/0.6149	0.5425/0.7065	0.5542/0.5105	0.5862/0.8876	0.5000/0.9268
Discrete SNNE	0.6563/0.9513	0.5154/0.9016	0.5282/0.5947	0.5492/0.7244	0.5730/0.5141	0.6455/0.9160	0.4996/0.9078
Discrete SNNE (only_denom)	0.7646/0.9661	0.7198/0.9470	0.6488/0.6990	0.6476/0.7787	0.6081/0.5534	0.7689/0.9183	0.5800/0.9329
Qwen 3-8B							
Discrete semantic entropy	0.8247/0.7818	0.5597/0.8719	0.5169/0.4914	0.6156/0.6838	0.6892/0.3135	0.7504/0.9671	0.5095/0.6119
Discrete semantic density	0.8277/0.7813	0.5982/0.8741	0.5770/0.5324	0.6271/0.6911	0.7025/0.3204	0.8335/0.9755	0.4738/0.6060
Discrete U_{eig}	0.8267/0.7833	0.5674/0.8735	0.5476/0.5484	0.6218/0.6852	0.6942/0.3213	0.8371/0.9735	0.5262/0.6243
Discrete U_{deg}	0.8267/0.7820	0.5671/0.8735	0.5479/0.5483	0.6205/0.6937	0.6928/0.3220	0.8359/0.9734	0.5262/0.6243
Discrete C_{ecc}	0.6248/0.6272	0.5606/0.8681	0.5122/0.4813	0.5883/0.6692	0.6045/0.2658	0.7911/0.9556	0.5262/0.6242
KLE-heat$_{t=0.1}$	0.8266/0.7864	0.5671/0.8735	0.5476/0.5484	0.6205/0.6901	0.6930/0.3159	0.8360/0.9735	0.5262/0.6242
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8268/0.7834	0.5671/0.8735	0.5476/0.5484	0.6206/0.6889	0.6928/0.3194	0.8359/0.9735	0.5262/0.6242
Discrete LUQ-pair	0.7709/0.7477	0.4991/0.8574	0.4745/0.4522	0.5162/0.6394	0.6446/0.2831	0.5416/0.9476	0.5000/0.6546
Discrete SNNE	0.8174/0.7720	0.5359/0.8645	0.5393/0.4975	0.5925/0.6733	0.6721/0.2959	0.6291/0.9543	0.5014/0.6183
Discrete SNNE (only_denom)	0.7849/0.7654	0.6661/0.8924	0.6503/0.6342	0.6249/0.7036	0.6785/0.3261	0.7949/0.9674	0.7797/0.8076

Table 26: UQ using Llama3-8B as an NLI tool on Summarization and Translation. We report PRR.

Uncertainty Method/ PRR	Summarization				Translation							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT												
Discrete semantic entropy	0.2722	0.1645	0.3374	0.2826	0.5612	0.6741	0.4902	0.5229	0.3891	0.5145	0.5928	0.4935
Discrete U_{eig}	0.2732	0.1596	0.3382	0.2833	0.5679	0.6952	0.4949	0.5345	0.4016	0.5141	0.6066	0.4922
Discrete U_{deg}	0.2724	0.1679	0.3296	0.2860	0.5584	0.6773	0.4986	0.5236	0.4009	0.5082	0.6049	0.4964
Discrete C_{ecc}	0.2753	0.1530	0.3381	0.2853	0.5574	0.6891	0.5001	0.5159	0.3981	0.5000	0.6068	0.4864
KLE-heat$_{t=0.1}$	0.2746	0.1675	0.3380	0.2859	0.5538	0.6779	0.4971	0.5253	0.3977	0.5136	0.6062	0.4820
Weighted matern$_{\kappa=1.0,\nu=1.0}$ kernel entropy	0.2743	0.1678	0.3299	0.2865	0.5561	0.6751	0.4943	0.5268	0.4029	0.5117	0.6012	0.4821
Discrete LUQ-pair	0.2761	0.1661	0.3356	0.2777	0.5410	0.6741	0.4776	0.5051	0.3797	0.4803	0.5980	0.4724
Discrete SNNE	0.2740	0.1700	0.3361	0.2814	0.5558	0.6776	0.4902	0.5188	0.3853	0.4992	0.6013	0.4752
Discrete SNNE (only_denom)	0.2910	0.1799	0.3641	0.3079	0.6037	0.7384	0.5411	0.5803	0.4444	0.5414	0.6408	0.5450
LLAMA 3.1-70B INSTRUCT												
Discrete semantic entropy	0.2577	0.2138	0.3474	0.3086	0.5799	0.7071	0.5315	0.6142	0.4664	0.5652	0.6261	0.5183
Discrete U_{eig}	0.2563	0.1982	0.3497	0.3063	0.5941	0.7208	0.5265	0.6215	0.4726	0.5746	0.6381	0.5253
Discrete U_{deg}	0.2594	0.2169	0.3486	0.3112	0.5983	0.7124	0.5284	0.6163	0.4677	0.5841	0.6372	0.5329
Discrete C_{ecc}	0.2590	0.1901	0.3486	0.3090	0.5902	0.7146	0.5260	0.6184	0.4707	0.5813	0.6286	0.5231
KLE-heat$_{t=0.1}$	0.2583	0.2162	0.3488	0.3092	0.5899	0.7094	0.5320	0.6161	0.4720	0.5857	0.6302	0.5182
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2583	0.2161	0.3489	0.3093	0.5854	0.7070	0.5304	0.6155	0.4260	0.5746	0.6302	0.5183
Discrete LUQ-pair	0.2570	0.2073	0.3449	0.3046	0.5686	0.7078	0.5128	0.5896	0.4598	0.5483	0.6284	0.5088
Discrete SNNE	0.2575	0.2072	0.3459	0.3078	0.5827	0.6992	0.5294	0.6006	0.4660	0.5600	0.6301	0.5122
Discrete SNNE (only_denom)	0.2784	0.2221	0.3714	0.3291	0.6489	0.7720	0.5863	0.6889	0.5305	0.6259	0.6796	0.5932
LLAMA 3.1-405B INSTRUCT FP8												
Discrete semantic entropy	0.2777	0.2514	0.3600	0.3217	0.6035	0.7198	0.5462	0.6498	0.4837	0.5800	0.6428	0.5492
Discrete U_{eig}	0.2772	0.2622	0.3634	0.3132	0.6181	0.7385	0.5501	0.6607	0.4867	0.6079	0.6570	0.5579
Discrete U_{deg}	0.2795	0.2575	0.3616	0.3225	0.6145	0.7268	0.5536	0.6566	0.4799	0.6109	0.6480	0.5652
Discrete C_{ecc}	0.2794	0.2414	0.3613	0.3172	0.6107	0.7299	0.5556	0.6565	0.4808	0.5907	0.6489	0.5507
KLE-heat$_{t=0.1}$	0.2791	0.2566	0.3618	0.3214	0.6054	0.7272	0.5554	0.6554	0.4808	0.6044	0.6478	0.5483
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2791	0.2572	0.3618	0.3237	0.6048	0.7268	0.5550	0.6577	0.4807	0.6034	0.6433	0.5491
Discrete LUQ-pair	0.2770	0.2573	0.3584	0.3138	0.5936	0.7235	0.5403	0.6500	0.4755	0.5965	0.6452	0.5419
Discrete SNNE	0.2772	0.2484	0.3590	0.3169	0.6076	0.7209	0.5422	0.6495	0.4802	0.5985	0.6458	0.5445
Discrete SNNE (only_denom)	0.3043	0.2913	0.3896	0.3395	0.6525	0.7766	0.5961	0.7271	0.5368	0.6523	0.6884	0.6126
QWEN 3-8B												
Discrete semantic entropy	0.2506	0.2178	0.3232	0.3055	0.5655	0.6689	0.4938	0.5964	0.4307	0.5558	0.5995	0.5216
Discrete U_{eig}	0.2492	0.2169	0.3233	0.2946	0.5604	0.6952	0.4897	0.6000	0.4353	0.5399	0.6086	0.5273
Discrete U_{deg}	0.2516	0.2219	0.3231	0.3050	0.5660	0.6723	0.4911	0.5969	0.4308	0.5132	0.5958	0.5194
Discrete C_{ecc}	0.2515	0.2280	0.3233	0.3010	0.5702	0.6892	0.5009	0.6013	0.4355	0.5467	0.6007	0.5236
KLE-heat$_{t=0.1}$	0.2504	0.2204	0.3234	0.3057	0.5677	0.6739	0.5003	0.5951	0.4326	0.5483	0.5967	0.5182
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2512	0.2214	0.3234	0.3055	0.5623	0.6738	0.4992	0.6033	0.4325	0.5474	0.5961	0.5194
Discrete LUQ-pair	0.2516	0.2256	0.3219	0.2971	0.5494	0.6724	0.4909	0.5946	0.4300	0.5332	0.5991	0.5147
Discrete SNNE	0.2520	0.2261	0.3222	0.2999	0.5529	0.6862	0.5005	0.5896	0.4270	0.5439	0.6065	0.5170
Discrete SNNE (only_denom)	0.2600	0.2375	0.3517	0.3209	0.5922	0.7195	0.5247	0.6398	0.4701	0.5749	0.6322	0.5653

H.6 OVERCONFIDENCE STATISTICS

Discussion: Appendix C.4.

Table 27: Statistics of SURE-OC of QA and Math tasks. Each cell shows an overconfident count and the ratio of overconfidence among all low-quality generations. The question IDs are provided with the code repository. Panel: QA.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ
LLAMA 3.1-8B INSTRUCT	119 (5.73%)	47 (6.41%)	3 (1.88%)	313 (11.11%)	346 (11.60%)
LLAMA 3.1-70B INSTRUCT	56 (6.75%)	45 (6.76%)	3 (2.21%)	258 (9.64%)	305 (12.24%)
LLAMA 3.1-405B INSTRUCT FP8	66 (10.38%)	47 (7.22%)	10 (7.04%)	261 (10.50%)	318 (13.89%)
QWEN 3-1.7B	535 (12.37%)	144 (13.99%)	19 (10.44%)	749 (22.81%)	581 (15.90%)
QWEN 3-8B	368 (11.07%)	295 (31.82%)	3 (1.74%)	1066 (35.68%)	487 (14.59%)
QWEN 3-32B	126 (5.69%)	61 (8.29%)	3 (1.92%)	385 (14.44%)	374 (12.47%)
PHI 3-14B	46 (1.76%)	63 (4.60%)	25 (15.72%)	268 (8.36%)	439 (13.94%)
PHI4-MINI-INSTRUCT 3.8B	137 (3.23%)	53 (3.92%)	22 (12.57%)	292 (8.80%)	530 (15.08%)
PHI 4-14B	61 (3.00%)	48 (5.30%)	9 (5.96%)	271 (9.91%)	321 (10.99%)
MISTRAL-7B-INSTRUCT-V0.2	120 (5.80%)	68 (6.56%)	15 (9.20%)	363 (12.14%)	380 (12.65%)
MISTRAL-NEMO-INSTRUCT-2407 12B	135 (6.23%)	48 (3.10%)	22 (10.89%)	304 (8.66%)	451 (13.49%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	115 (7.64%)	97 (8.66%)	19 (11.95%)	371 (11.80%)	395 (12.77%)
GPT-4O-MINI	213 (17.85%)	144 (16.84%)	6 (4.29%)	395 (14.14%)	277 (10.72%)

Table 27: Statistics of SURE-OC of QA and Math tasks. Each cell shows an overconfident count and the ratio of overconfidence among all low-quality generations. The question IDs are provided with the code repository. Panel: Math.

Model / method	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT	20 (13.07%)	6 (3.16%)
LLAMA 3.1-70B INSTRUCT	5 (8.62%)	5 (4.35%)
LLAMA 3.1-405B INSTRUCT FP8	11 (25.00%)	4 (5.00%)
QWEN 3-1.7B	1 (3.45%)	24 (4.72%)
QWEN 3-8B	2 (8.70%)	13 (2.63%)
QWEN 3-32B	2 (7.69%)	13 (3.29%)
PHI 3-14B	3 (7.89%)	51 (10.34%)
PHI4-MINI-INSTRUCT 3.8B	1 (1.43%)	38 (7.60%)
PHI 4-14B	8 (7.14%)	43 (7.33%)
MISTRAL-7B-INSTRUCT-V0.2	32 (20.00%)	61 (8.89%)
MISTRAL-NEMO-INSTRUCT-2407 12B	15 (10.27%)	40 (9.80%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	14 (16.87%)	36 (7.91%)
GPT-4O-MINI	22 (40.00%)	30 (4.75%)

Table 28: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
LLAMA 3.1-8B INSTRUCT	541 (18.78%)	84 (14.69%)	531 (18.48%)	219 (15.58%)
LLAMA 3.1-70B INSTRUCT	524 (18.19%)	83 (15.54%)	507 (17.56%)	212 (15.08%)
LLAMA 3.1-405B INSTRUCT FP8	499 (17.32%)	73 (13.70%)	452 (15.67%)	232 (16.49%)
QWEN 3-1.7B	590 (20.82%)	94 (14.18%)	469 (16.21%)	274 (19.49%)
QWEN 3-8B	568 (19.74%)	109 (18.83%)	454 (15.51%)	282 (20.06%)
QWEN 3-32B	559 (19.49%)	104 (18.06%)	490 (17.03%)	259 (18.41%)
PHI 3-14B	569 (19.99%)	104 (17.60%)	506 (17.61%)	24 (1.71%)
PHI4-MINI-INSTRUCT 3.8B	578 (20.37%)	83 (14.49%)	510 (17.73%)	128 (9.10%)
PHI 4-14B	584 (20.40%)	72 (15.09%)	563 (19.60%)	297 (21.12%)
MISTRAL-7B-INSTRUCT-V0.2	514 (18.02%)	83 (15.57%)	498 (17.03%)	277 (19.70%)
MISTRAL-NEMO-INSTRUCT-2407 12B	550 (19.38%)	99 (15.35%)	481 (16.66%)	241 (17.14%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	472 (16.63%)	88 (15.33%)	463 (16.08%)	262 (18.63%)
GPT-4O-MINI	578 (19.76%)	92 (15.38%)	535 (18.62%)	301 (21.41%)

Table 28: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). Panel: Translation: cs, de, fi, gu.

Model / method	cs	de	fi	gu
LLAMA 3.1-8B INSTRUCT	128 (17.14%)	71 (9.39%)	102 (13.58%)	42 (8.14%)
LLAMA 3.1-70B INSTRUCT	135 (17.90%)	95 (12.52%)	105 (13.83%)	54 (10.67%)
LLAMA 3.1-405B INSTRUCT FP8	127 (16.98%)	113 (14.83%)	139 (18.19%)	55 (10.56%)
QWEN 3-1.7B	125 (16.71%)	102 (13.11%)	118 (15.69%)	51 (10.06%)
QWEN 3-8B	133 (17.69%)	100 (13.33%)	114 (15.06%)	62 (12.25%)
QWEN 3-32B	121 (16.03%)	93 (12.38%)	97 (12.90%)	47 (9.40%)
PHI 3-14B	281 (21.70%)	272 (20.96%)	289 (21.49%)	180 (20.29%)
PHI4-MINI-INSTRUCT 3.8B	45 (5.45%)	28 (3.72%)	75 (9.89%)	196 (24.17%)
PHI 4-14B	111 (14.74%)	88 (11.72%)	98 (13.07%)	45 (9.00%)
MISTRAL-7B-INSTRUCT-V0.2	113 (15.09%)	96 (12.80%)	86 (11.45%)	71 (14.20%)
MISTRAL-NEMO-INSTRUCT-2407 12B	115 (15.42%)	68 (8.94%)	89 (11.77%)	36 (7.03%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	117 (15.68%)	87 (11.58%)	96 (12.61%)	43 (8.58%)
GPT-4O-MINI	137 (18.27%)	92 (12.25%)	131 (17.12%)	62 (11.57%)

Table 28: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). Panel: Translation: kk, lt, ru, zh.

Model / method	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT	67 (12.57%)	81 (16.20%)	123 (16.40%)	142 (14.24%)
LLAMA 3.1-70B INSTRUCT	81 (15.67%)	69 (13.75%)	120 (16.00%)	162 (16.27%)
LLAMA 3.1-405B INSTRUCT FP8	83 (16.05%)	78 (15.60%)	130 (17.26%)	144 (14.46%)
QWEN 3-1.7B	76 (14.42%)	105 (21.00%)	108 (14.38%)	155 (15.32%)
QWEN 3-8B	62 (11.88%)	85 (16.80%)	140 (18.57%)	165 (16.53%)
QWEN 3-32B	67 (12.76%)	78 (15.54%)	125 (16.60%)	150 (15.02%)
PHI 3-14B	31 (5.93%)	54 (10.76%)	103 (13.72%)	161 (16.05%)
PHI4-MINI-INSTRUCT 3.8B	46 (8.90%)	59 (11.75%)	79 (10.51%)	109 (10.94%)
PHI 4-14B	55 (10.64%)	76 (15.11%)	120 (15.85%)	145 (14.38%)
MISTRAL-7B-INSTRUCT-V0.2	56 (10.75%)	57 (11.29%)	107 (14.10%)	130 (13.04%)
MISTRAL-NEMO-INSTRUCT-2407 12B	75 (14.20%)	52 (10.38%)	121 (16.07%)	136 (13.65%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	64 (12.38%)	81 (15.37%)	132 (17.44%)	141 (14.16%)
GPT-4O-MINI	89 (17.12%)	98 (19.41%)	121 (15.84%)	171 (17.15%)

Table 29: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.05$) Panel: QA.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ
LLAMA 3.1-8B INSTRUCT	119 (5.73%)	47 (6.41%)	0 (0.00%)	142 (5.04%)	84 (2.82%)
LLAMA 3.1-70B INSTRUCT	15 (1.81%)	9 (1.35%)	1 (0.74%)	51 (1.91%)	85 (3.41%)
LLAMA 3.1-405B INSTRUCT FP8	66 (10.38%)	47 (7.22%)	1 (0.70%)	108 (4.35%)	69 (3.01%)
QWEN 3-1.7B	22 (0.51%)	144 (13.99%)	2 (1.10%)	749 (22.81%)	121 (3.31%)
QWEN 3-8B	368 (11.07%)	295 (31.82%)	2 (1.16%)	1066 (35.68%)	315 (9.43%)
QWEN 3-32B	55 (2.48%)	60 (8.15%)	1 (0.64%)	385 (14.44%)	87 (2.90%)
PHI 3-14B	11 (0.42%)	11 (0.80%)	6 (3.77%)	58 (1.81%)	58 (1.84%)
PHI4-MINI-INSTRUCT 3.8B	28 (0.66%)	23 (1.70%)	6 (3.43%)	69 (2.08%)	87 (2.48%)
PHI 4-14B	61 (3.00%)	6 (0.66%)	1 (0.66%)	60 (2.19%)	57 (1.95%)
MISTRAL-7B-INSTRUCT-V0.2	55 (2.66%)	13 (1.25%)	2 (1.23%)	102 (3.41%)	84 (2.80%)
MISTRAL-NEMO-INSTRUCT-2407 12B	135 (6.23%)	48 (3.10%)	5 (2.48%)	228 (6.49%)	148 (4.43%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	39 (2.59%)	26 (2.32%)	7 (4.40%)	103 (3.28%)	110 (3.56%)
GPT-4O-MINI	213 (17.85%)	144 (16.84%)	1 (0.71%)	395 (14.14%)	244 (9.45%)

Table 29: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.05$) Panel: Math.

Model / method	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT	20 (13.07%)	0 (0.00%)
LLAMA 3.1-70B INSTRUCT	5 (8.62%)	0 (0.00%)
LLAMA 3.1-405B INSTRUCT FP8	11 (25.00%)	0 (0.00%)
QWEN 3-1.7B	0 (0.00%)	1 (0.20%)
QWEN 3-8B	2 (8.70%)	1 (0.20%)
QWEN 3-32B	2 (7.69%)	1 (0.25%)
PHI 3-14B	1 (2.63%)	3 (0.61%)
PHI4-MINI-INSTRUCT 3.8B	0 (0.00%)	3 (0.60%)
PHI 4-14B	8 (7.14%)	0 (0.00%)
MISTRAL-7B-INSTRUCT-V0.2	32 (20.00%)	7 (1.02%)
MISTRAL-NEMO-INSTRUCT-2407 12B	3 (2.05%)	3 (0.73%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	14 (16.87%)	1 (0.22%)
GPT-4O-MINI	22 (40.00%)	2 (0.32%)

Table 30: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.05$) Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
LLAMA 3.1-8B INSTRUCT	132 (4.58%)	17 (2.97%)	128 (4.45%)	56 (3.98%)
LLAMA 3.1-70B INSTRUCT	129 (4.48%)	19 (3.56%)	116 (4.02%)	56 (3.98%)
LLAMA 3.1-405B INSTRUCT FP8	124 (4.30%)	11 (2.06%)	78 (2.70%)	51 (3.62%)
QWEN 3-1.7B	172 (6.07%)	26 (3.92%)	115 (3.97%)	81 (5.76%)
QWEN 3-8B	151 (5.25%)	27 (4.66%)	99 (3.38%)	57 (4.05%)
QWEN 3-32B	141 (4.92%)	20 (3.47%)	106 (3.68%)	68 (4.83%)
PHI 3-14B	142 (4.99%)	26 (4.40%)	128 (4.45%)	14 (1.00%)
PHI4-MINI-INSTRUCT 3.8B	136 (4.79%)	24 (4.19%)	121 (4.21%)	105 (7.47%)
PHI 4-14B	163 (5.69%)	16 (3.35%)	131 (4.56%)	93 (6.61%)
MISTRAL-7B-INSTRUCT-V0.2	131 (4.59%)	21 (3.94%)	108 (3.69%)	81 (5.76%)
MISTRAL-NEMO-INSTRUCT-2407 12B	133 (4.69%)	23 (3.57%)	108 (3.74%)	69 (4.91%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	105 (3.70%)	23 (4.01%)	124 (4.31%)	75 (5.33%)
GPT-4O-MINI	140 (4.79%)	20 (3.34%)	120 (4.18%)	88 (6.26%)

Table 30: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.05$) Panel: Translation: cs, de, fi, gu.

Model / method	cs	de	fi	gu
LLAMA 3.1-8B INSTRUCT	31 (4.15%)	16 (2.12%)	22 (2.93%)	6 (1.16%)
LLAMA 3.1-70B INSTRUCT	29 (3.85%)	20 (2.64%)	18 (2.37%)	6 (1.19%)
LLAMA 3.1-405B INSTRUCT FP8	19 (2.54%)	25 (3.28%)	34 (4.45%)	6 (1.15%)
QWEN 3-1.7B	36 (4.81%)	38 (4.88%)	45 (5.98%)	9 (1.78%)
QWEN 3-8B	75 (9.97%)	67 (8.93%)	45 (5.94%)	21 (4.15%)
QWEN 3-32B	25 (3.31%)	30 (3.99%)	23 (3.06%)	7 (1.40%)
PHI 3-14B	74 (5.71%)	73 (5.62%)	76 (5.65%)	48 (5.41%)
PHI4-MINI-INSTRUCT 3.8B	8 (0.97%)	2 (0.27%)	11 (1.45%)	62 (7.64%)
PHI 4-14B	24 (3.19%)	17 (2.26%)	15 (2.00%)	6 (1.20%)
MISTRAL-7B-INSTRUCT-V0.2	26 (3.47%)	24 (3.20%)	19 (2.53%)	12 (2.40%)
MISTRAL-NEMO-INSTRUCT-2407 12B	23 (3.08%)	18 (2.37%)	18 (2.38%)	5 (0.98%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	32 (4.29%)	24 (3.20%)	21 (2.76%)	7 (1.40%)
GPT-4O-MINI	79 (10.53%)	58 (7.72%)	61 (7.97%)	15 (2.80%)

Table 30: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.05$) Panel: Translation: kk, lt, ru, zh.

Model / method	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT	15 (2.81%)	24 (4.80%)	30 (4.00%)	39 (3.91%)
LLAMA 3.1-70B INSTRUCT	19 (3.68%)	16 (3.19%)	30 (4.00%)	32 (3.21%)
LLAMA 3.1-405B INSTRUCT FP8	22 (4.26%)	17 (3.40%)	27 (3.59%)	33 (3.31%)
QWEN 3-1.7B	17 (3.23%)	49 (9.80%)	34 (4.53%)	38 (3.75%)
QWEN 3-8B	15 (2.87%)	23 (4.55%)	71 (9.42%)	49 (4.91%)
QWEN 3-32B	15 (2.86%)	21 (4.18%)	29 (3.85%)	38 (3.80%)
PHI 3-14B	9 (1.72%)	11 (2.19%)	18 (2.40%)	26 (2.59%)
PHI4-MINI-INSTRUCT 3.8B	4 (0.77%)	10 (1.99%)	13 (1.73%)	20 (2.01%)
PHI 4-14B	9 (1.74%)	20 (3.98%)	29 (3.83%)	37 (3.67%)
MISTRAL-7B-INSTRUCT-V0.2	7 (1.34%)	12 (2.38%)	26 (3.43%)	38 (3.81%)
MISTRAL-NEMO-INSTRUCT-2407 12B	12 (2.27%)	17 (3.39%)	31 (4.12%)	30 (3.01%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	16 (3.09%)	21 (3.98%)	28 (3.70%)	29 (2.91%)
GPT-4O-MINI	22 (4.23%)	23 (4.55%)	53 (6.94%)	41 (4.11%)

Table 31: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.1$) Panel: QA.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ
LLAMA 3.1-8B INSTRUCT	119 (5.73%)	47 (6.41%)	0 (0.00%)	142 (5.04%)	167 (5.60%)
LLAMA 3.1-70B INSTRUCT	28 (3.37%)	14 (2.10%)	2 (1.47%)	122 (4.56%)	145 (5.82%)
LLAMA 3.1-405B INSTRUCT FP8	66 (10.38%)	47 (7.22%)	5 (3.52%)	108 (4.35%)	152 (6.64%)
QWEN 3-1.7B	178 (4.12%)	144 (13.99%)	5 (2.75%)	749 (22.81%)	260 (7.12%)
QWEN 3-8B	368 (11.07%)	295 (31.82%)	2 (1.16%)	1066 (35.68%)	315 (9.43%)
QWEN 3-32B	55 (2.48%)	60 (8.15%)	1 (0.64%)	385 (14.44%)	185 (6.17%)
PHI 3-14B	14 (0.54%)	28 (2.04%)	11 (6.92%)	127 (3.96%)	167 (5.30%)
PHI4-MINI-INSTRUCT 3.8B	59 (1.39%)	23 (1.70%)	11 (6.29%)	124 (3.74%)	222 (6.32%)
PHI 4-14B	61 (3.00%)	22 (2.43%)	1 (0.66%)	126 (4.61%)	144 (4.93%)
MISTRAL-7B-INSTRUCT-V0.2	55 (2.66%)	21 (2.03%)	7 (4.29%)	148 (4.95%)	182 (6.06%)
MISTRAL-NEMO-INSTRUCT-2407 12B	135 (6.23%)	48 (3.10%)	7 (3.47%)	228 (6.49%)	197 (5.89%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	59 (3.92%)	50 (4.46%)	9 (5.66%)	200 (6.36%)	210 (6.79%)
GPT-4O-MINI	213 (17.85%)	144 (16.84%)	1 (0.71%)	395 (14.14%)	244 (9.45%)

Table 31: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.1$) Panel: Math.

Model / method	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT	20 (13.07%)	0 (0.00%)
LLAMA 3.1-70B INSTRUCT	5 (8.62%)	2 (1.74%)
LLAMA 3.1-405B INSTRUCT FP8	11 (25.00%)	2 (2.50%)
QWEN 3-1.7B	0 (0.00%)	5 (0.98%)
QWEN 3-8B	2 (8.70%)	5 (1.01%)
QWEN 3-32B	2 (7.69%)	5 (1.27%)
PHI 3-14B	1 (2.63%)	15 (3.04%)
PHI4-MINI-INSTRUCT 3.8B	1 (1.43%)	12 (2.40%)
PHI 4-14B	8 (7.14%)	4 (0.68%)
MISTRAL-7B-INSTRUCT-V0.2	32 (20.00%)	24 (3.50%)
MISTRAL-NEMO-INSTRUCT-2407 12B	3 (2.05%)	9 (2.20%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	14 (16.87%)	6 (1.32%)
GPT-4O-MINI	22 (40.00%)	6 (0.95%)

Table 32: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.1$) Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
LLAMA 3.1-8B INSTRUCT	255 (8.85%)	35 (6.12%)	268 (9.32%)	100 (7.11%)
LLAMA 3.1-70B INSTRUCT	253 (8.78%)	40 (7.49%)	230 (7.96%)	110 (7.82%)
LLAMA 3.1-405B INSTRUCT FP8	244 (8.47%)	30 (5.63%)	206 (7.14%)	111 (7.89%)
QWEN 3-1.7B	308 (10.87%)	47 (7.09%)	231 (7.98%)	139 (9.89%)
QWEN 3-8B	280 (9.73%)	54 (9.33%)	212 (7.24%)	131 (9.32%)
QWEN 3-32B	283 (9.87%)	46 (7.99%)	219 (7.61%)	144 (10.23%)
PHI 3-14B	277 (9.73%)	53 (8.97%)	249 (8.66%)	20 (1.42%)
PHI4-MINI-INSTRUCT 3.8B	284 (10.01%)	45 (7.85%)	250 (8.69%)	112 (7.97%)
PHI 4-14B	301 (10.51%)	30 (6.29%)	270 (9.40%)	156 (11.10%)
MISTRAL-7B-INSTRUCT-V0.2	252 (8.84%)	42 (7.88%)	223 (7.63%)	141 (10.03%)
MISTRAL-NEMO-INSTRUCT-2407 12B	277 (9.76%)	41 (6.36%)	229 (7.93%)	118 (8.39%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	220 (7.75%)	46 (8.01%)	228 (7.92%)	131 (9.32%)
GPT-4O-MINI	280 (9.57%)	40 (6.69%)	257 (8.95%)	165 (11.74%)

Table 32: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.1$) Panel: Translation: cs, de, fi, gu.

Model / method	cs	de	fi	gu
LLAMA 3.1-8B INSTRUCT	61 (8.17%)	31 (4.10%)	50 (6.66%)	15 (2.91%)
LLAMA 3.1-70B INSTRUCT	64 (8.49%)	42 (5.53%)	45 (5.93%)	19 (3.75%)
LLAMA 3.1-405B INSTRUCT FP8	52 (6.95%)	49 (6.43%)	56 (7.33%)	18 (3.45%)
QWEN 3-1.7B	56 (7.49%)	48 (6.17%)	69 (9.18%)	22 (4.34%)
QWEN 3-8B	75 (9.97%)	67 (8.93%)	56 (7.40%)	25 (4.94%)
QWEN 3-32B	62 (8.21%)	46 (6.13%)	44 (5.85%)	16 (3.20%)
PHI 3-14B	150 (11.58%)	142 (10.94%)	145 (10.78%)	103 (11.61%)
PHI4-MINI-INSTRUCT 3.8B	18 (2.18%)	4 (0.53%)	30 (3.96%)	114 (14.06%)
PHI 4-14B	55 (7.30%)	36 (4.79%)	40 (5.33%)	17 (3.40%)
MISTRAL-7B-INSTRUCT-V0.2	54 (7.21%)	43 (5.73%)	36 (4.79%)	30 (6.00%)
MISTRAL-NEMO-INSTRUCT-2407 12B	50 (6.70%)	34 (4.47%)	35 (4.63%)	11 (2.15%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	57 (7.64%)	34 (4.53%)	46 (6.04%)	22 (4.39%)
GPT-4O-MINI	79 (10.53%)	58 (7.72%)	62 (8.10%)	22 (4.10%)

Table 32: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.1$) Panel: Translation: kk, lt, ru, zh.

Model / method	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT	31 (5.82%)	43 (8.60%)	59 (7.87%)	64 (6.42%)
LLAMA 3.1-70B INSTRUCT	39 (7.54%)	32 (6.37%)	55 (7.33%)	74 (7.43%)
LLAMA 3.1-405B INSTRUCT FP8	35 (6.77%)	40 (8.00%)	58 (7.70%)	71 (7.13%)
QWEN 3-1.7B	32 (6.07%)	66 (13.20%)	61 (8.12%)	74 (7.31%)
QWEN 3-8B	26 (4.98%)	45 (8.89%)	71 (9.42%)	83 (8.32%)
QWEN 3-32B	30 (5.71%)	39 (7.77%)	62 (8.23%)	82 (8.21%)
PHI 3-14B	19 (3.63%)	27 (5.38%)	40 (5.33%)	73 (7.28%)
PHI4-MINI-INSTRUCT 3.8B	15 (2.90%)	22 (4.38%)	28 (3.72%)	44 (4.42%)
PHI 4-14B	23 (4.45%)	37 (7.36%)	61 (8.06%)	69 (6.85%)
MISTRAL-7B-INSTRUCT-V0.2	21 (4.03%)	24 (4.75%)	48 (6.32%)	64 (6.42%)
MISTRAL-NEMO-INSTRUCT-2407 12B	36 (6.82%)	24 (4.79%)	70 (9.30%)	54 (5.42%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	25 (4.84%)	42 (7.97%)	64 (8.45%)	65 (6.53%)
GPT-4O-MINI	44 (8.46%)	40 (7.92%)	57 (7.46%)	82 (8.22%)

Table 33: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.15$) Panel: QA.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ
LLAMA 3.1-8B INSTRUCT	119 (5.73%)	47 (6.41%)	2 (1.25%)	213 (7.56%)	260 (8.71%)
LLAMA 3.1-70B INSTRUCT	41 (4.94%)	27 (4.05%)	2 (1.47%)	176 (6.58%)	212 (8.51%)
LLAMA 3.1-405B INSTRUCT FP8	66 (10.38%)	47 (7.22%)	7 (4.93%)	199 (8.01%)	237 (10.35%)
QWEN 3-1.7B	352 (8.14%)	144 (13.99%)	11 (6.04%)	749 (22.81%)	416 (11.39%)
QWEN 3-8B	368 (11.07%)	295 (31.82%)	2 (1.16%)	1066 (35.68%)	357 (10.69%)
QWEN 3-32B	95 (4.29%)	60 (8.15%)	2 (1.28%)	385 (14.44%)	279 (9.30%)
PHI 3-14B	29 (1.11%)	43 (3.14%)	17 (10.69%)	191 (5.96%)	299 (9.50%)
PHI4-MINI-INSTRUCT 3.8B	80 (1.89%)	32 (2.37%)	17 (9.71%)	210 (6.33%)	364 (10.36%)
PHI 4-14B	61 (3.00%)	42 (4.64%)	4 (2.65%)	198 (7.24%)	220 (7.53%)
MISTRAL-7B-INSTRUCT-V0.2	81 (3.92%)	42 (4.05%)	11 (6.75%)	265 (8.86%)	277 (9.22%)
MISTRAL-NEMO-INSTRUCT-2407 12B	135 (6.23%)	48 (3.10%)	16 (7.92%)	228 (6.49%)	311 (9.30%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	80 (5.31%)	71 (6.34%)	16 (10.06%)	289 (9.19%)	300 (9.70%)
GPT-4O-MINI	213 (17.85%)	144 (16.84%)	2 (1.43%)	395 (14.14%)	244 (9.45%)

Table 33: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.15$) Panel: Math.

Model / method	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT	20 (13.07%)	3 (1.58%)
LLAMA 3.1-70B INSTRUCT	5 (8.62%)	3 (2.61%)
LLAMA 3.1-405B INSTRUCT FP8	11 (25.00%)	3 (3.75%)
QWEN 3-1.7B	1 (3.45%)	11 (2.17%)
QWEN 3-8B	2 (8.70%)	10 (2.02%)
QWEN 3-32B	2 (7.69%)	8 (2.03%)
PHI 3-14B	3 (7.89%)	30 (6.09%)
PHI4-MINI-INSTRUCT 3.8B	1 (1.43%)	26 (5.20%)
PHI 4-14B	8 (7.14%)	13 (2.21%)
MISTRAL-7B-INSTRUCT-V0.2	32 (20.00%)	41 (5.98%)
MISTRAL-NEMO-INSTRUCT-2407 12B	8 (5.48%)	17 (4.16%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	14 (16.87%)	17 (3.74%)
GPT-4O-MINI	22 (40.00%)	14 (2.22%)

Table 34: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.15$) Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
LLAMA 3.1-8B INSTRUCT	394 (13.68%)	59 (10.31%)	397 (13.81%)	158 (11.24%)
LLAMA 3.1-70B INSTRUCT	395 (13.72%)	61 (11.42%)	360 (12.47%)	163 (11.59%)
LLAMA 3.1-405B INSTRUCT FP8	360 (12.50%)	46 (8.63%)	326 (11.30%)	165 (11.73%)
QWEN 3-1.7B	452 (15.95%)	73 (11.01%)	353 (12.20%)	212 (15.08%)
QWEN 3-8B	427 (14.84%)	79 (13.64%)	323 (11.04%)	204 (14.51%)
QWEN 3-32B	415 (14.47%)	73 (12.67%)	364 (12.65%)	199 (14.14%)
PHI 3-14B	423 (14.86%)	76 (12.86%)	376 (13.08%)	22 (1.56%)
PHI4-MINI-INSTRUCT 3.8B	430 (15.15%)	63 (10.99%)	378 (13.14%)	117 (8.32%)
PHI 4-14B	434 (15.16%)	50 (10.48%)	419 (14.58%)	231 (16.43%)
MISTRAL-7B-INSTRUCT-V0.2	372 (13.04%)	65 (12.20%)	371 (12.69%)	204 (14.51%)
MISTRAL-NEMO-INSTRUCT-2407 12B	422 (14.87%)	66 (10.23%)	356 (12.33%)	173 (12.30%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	344 (12.12%)	68 (11.85%)	347 (12.05%)	206 (14.65%)
GPT-4O-MINI	424 (14.50%)	64 (10.70%)	390 (13.57%)	231 (16.43%)

Table 34: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.15$) Panel: Translation: cs, de, fi, gu.

Model / method	cs	de	fi	gu
LLAMA 3.1-8B INSTRUCT	100 (13.39%)	53 (7.01%)	68 (9.05%)	25 (4.84%)
LLAMA 3.1-70B INSTRUCT	94 (12.47%)	67 (8.83%)	74 (9.75%)	36 (7.11%)
LLAMA 3.1-405B INSTRUCT FP8	90 (12.03%)	85 (11.15%)	100 (13.09%)	38 (7.29%)
QWEN 3-1.7B	88 (11.76%)	74 (9.51%)	88 (11.70%)	38 (7.50%)
QWEN 3-8B	97 (12.90%)	73 (9.73%)	82 (10.83%)	43 (8.50%)
QWEN 3-32B	91 (12.05%)	64 (8.52%)	75 (9.97%)	31 (6.20%)
PHI 3-14B	225 (17.37%)	204 (15.72%)	219 (16.28%)	139 (15.67%)
PHI4-MINI-INSTRUCT 3.8B	29 (3.52%)	19 (2.52%)	48 (6.33%)	154 (18.99%)
PHI 4-14B	82 (10.89%)	56 (7.46%)	71 (9.47%)	33 (6.60%)
MISTRAL-7B-INSTRUCT-V0.2	88 (11.75%)	69 (9.20%)	60 (7.99%)	48 (9.60%)
MISTRAL-NEMO-INSTRUCT-2407 12B	84 (11.26%)	50 (6.57%)	64 (8.47%)	23 (4.49%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	87 (11.66%)	58 (7.72%)	68 (8.94%)	32 (6.39%)
GPT-4O-MINI	96 (12.80%)	66 (8.79%)	90 (11.76%)	41 (7.65%)

Table 34: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.15$) Panel: Translation: kk, lt, ru, zh.

Model / method	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT	48 (9.01%)	64 (12.80%)	89 (11.87%)	103 (10.33%)
LLAMA 3.1-70B INSTRUCT	55 (10.64%)	47 (9.36%)	85 (11.33%)	117 (11.75%)
LLAMA 3.1-405B INSTRUCT FP8	55 (10.64%)	64 (12.80%)	95 (12.62%)	115 (11.55%)
QWEN 3-1.7B	58 (11.01%)	89 (17.80%)	82 (10.92%)	116 (11.46%)
QWEN 3-8B	45 (8.62%)	67 (13.24%)	102 (13.53%)	125 (12.53%)
QWEN 3-32B	49 (9.33%)	58 (11.55%)	96 (12.75%)	112 (11.21%)
PHI 3-14B	27 (5.16%)	43 (8.57%)	77 (10.25%)	109 (10.87%)
PHI4-MINI-INSTRUCT 3.8B	30 (5.80%)	46 (9.16%)	54 (7.18%)	70 (7.03%)
PHI 4-14B	37 (7.16%)	57 (11.33%)	91 (12.02%)	108 (10.71%)
MISTRAL-7B-INSTRUCT-V0.2	33 (6.33%)	34 (6.73%)	77 (10.14%)	105 (10.53%)
MISTRAL-NEMO-INSTRUCT-2407 12B	57 (10.80%)	39 (7.78%)	91 (12.08%)	95 (9.54%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	42 (8.12%)	64 (12.14%)	89 (11.76%)	106 (10.64%)
GPT-4O-MINI	69 (13.27%)	75 (14.85%)	86 (11.26%)	125 (12.54%)

Table 35: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.3$) Panel: QA.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ
LLAMA 3.1-8B INSTRUCT	128 (6.17%)	56 (7.64%)	5 (3.13%)	504 (17.89%)	578 (19.37%)
LLAMA 3.1-70B INSTRUCT	73 (8.80%)	69 (10.36%)	5 (3.68%)	462 (17.27%)	483 (19.39%)
LLAMA 3.1-405B INSTRUCT FP8	66 (10.38%)	53 (8.14%)	23 (16.20%)	431 (17.34%)	485 (21.19%)
QWEN 3-1.7B	902 (20.86%)	144 (13.99%)	37 (20.33%)	749 (22.81%)	929 (25.43%)
QWEN 3-8B	377 (11.34%)	295 (31.82%)	5 (2.91%)	1066 (35.68%)	770 (23.06%)
QWEN 3-32B	164 (7.41%)	117 (15.90%)	3 (1.92%)	428 (16.05%)	579 (19.31%)
PHI 3-14B	99 (3.80%)	116 (8.47%)	42 (26.42%)	471 (14.70%)	743 (23.59%)
PHI4-MINI-INSTRUCT 3.8B	334 (7.88%)	100 (7.40%)	36 (20.57%)	469 (14.13%)	854 (24.30%)
PHI 4-14B	76 (3.73%)	81 (8.94%)	24 (15.89%)	407 (14.89%)	557 (19.07%)
MISTRAL-7B-INSTRUCT-V0.2	206 (9.96%)	132 (12.73%)	23 (14.11%)	544 (18.19%)	632 (21.04%)
MISTRAL-NEMO-INSTRUCT-2407 12B	135 (6.23%)	114 (7.36%)	40 (19.80%)	451 (12.85%)	752 (22.49%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	164 (10.89%)	137 (12.23%)	30 (18.87%)	569 (18.09%)	652 (21.08%)
GPT-4O-MINI	213 (17.85%)	144 (16.84%)	16 (11.43%)	496 (17.76%)	486 (18.82%)

Table 35: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.3$) Panel: Math.

Model / method	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT	26 (16.99%)	13 (6.84%)
LLAMA 3.1-70B INSTRUCT	5 (8.62%)	11 (9.57%)
LLAMA 3.1-405B INSTRUCT FP8	11 (25.00%)	5 (6.25%)
QWEN 3-1.7B	2 (6.90%)	56 (11.02%)
QWEN 3-8B	2 (8.70%)	44 (8.89%)
QWEN 3-32B	2 (7.69%)	33 (8.35%)
PHI 3-14B	4 (10.53%)	78 (15.82%)
PHI4-MINI-INSTRUCT 3.8B	5 (7.14%)	83 (16.60%)
PHI 4-14B	9 (8.04%)	48 (8.18%)
MISTRAL-7B-INSTRUCT-V0.2	32 (20.00%)	113 (16.47%)
MISTRAL-NEMO-INSTRUCT-2407 12B	27 (18.49%)	64 (15.65%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	14 (16.87%)	74 (16.26%)
GPT-4O-MINI	22 (40.00%)	71 (11.25%)

Table 36: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.3$) Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
LLAMA 3.1-8B INSTRUCT	809 (28.09%)	129 (22.55%)	786 (27.35%)	334 (23.76%)
LLAMA 3.1-70B INSTRUCT	792 (27.50%)	118 (22.10%)	794 (27.49%)	347 (24.68%)
LLAMA 3.1-405B INSTRUCT FP8	737 (25.58%)	122 (22.89%)	697 (24.16%)	333 (23.67%)
QWEN 3-1.7B	881 (31.09%)	151 (22.78%)	736 (25.43%)	416 (29.59%)
QWEN 3-8B	852 (29.60%)	145 (25.04%)	730 (24.94%)	410 (29.16%)
QWEN 3-32B	820 (28.59%)	146 (25.35%)	766 (26.62%)	390 (27.72%)
PHI 3-14B	853 (29.97%)	161 (27.24%)	783 (27.24%)	29 (2.06%)
PHI4-MINI-INSTRUCT 3.8B	874 (30.80%)	141 (24.61%)	747 (25.97%)	141 (10.03%)
PHI 4-14B	879 (30.70%)	119 (24.95%)	866 (30.14%)	436 (31.01%)
MISTRAL-7B-INSTRUCT-V0.2	801 (28.09%)	141 (26.45%)	752 (25.72%)	392 (27.88%)
MISTRAL-NEMO-INSTRUCT-2407 12B	834 (29.39%)	157 (24.34%)	693 (24.00%)	366 (26.03%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	735 (25.89%)	135 (23.52%)	731 (25.38%)	388 (27.60%)
GPT-4O-MINI	840 (28.72%)	143 (23.91%)	829 (28.85%)	422 (30.01%)

Table 36: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.3$) Panel: Translation: cs, de, fi, gu.

Model / method	cs	de	fi	gu
LLAMA 3.1-8B INSTRUCT	199 (26.64%)	122 (16.14%)	164 (21.84%)	76 (14.73%)
LLAMA 3.1-70B INSTRUCT	196 (25.99%)	135 (17.79%)	163 (21.48%)	86 (17.00%)
LLAMA 3.1-405B INSTRUCT FP8	214 (28.61%)	170 (22.31%)	206 (26.96%)	92 (17.66%)
QWEN 3-1.7B	187 (25.00%)	156 (20.05%)	169 (22.47%)	90 (17.75%)
QWEN 3-8B	200 (26.60%)	148 (19.73%)	181 (23.91%)	95 (18.77%)
QWEN 3-32B	197 (26.09%)	151 (20.11%)	163 (21.68%)	75 (15.00%)
PHI 3-14B	421 (32.51%)	404 (31.12%)	431 (32.04%)	272 (30.67%)
PHI4-MINI-INSTRUCT 3.8B	78 (9.45%)	62 (8.23%)	121 (15.96%)	288 (35.51%)
PHI 4-14B	182 (24.17%)	134 (17.84%)	164 (21.87%)	78 (15.60%)
MISTRAL-7B-INSTRUCT-V0.2	170 (22.70%)	141 (18.80%)	151 (20.11%)	122 (24.40%)
MISTRAL-NEMO-INSTRUCT-2407 12B	172 (23.06%)	117 (15.37%)	138 (18.25%)	77 (15.04%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	193 (25.87%)	136 (18.11%)	158 (20.76%)	84 (16.77%)
GPT-4O-MINI	217 (28.93%)	147 (19.57%)	198 (25.88%)	98 (18.28%)

Table 36: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.3$) Panel: Translation: kk, lt, ru, zh.

Model / method	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT	112 (21.01%)	111 (22.20%)	178 (23.73%)	228 (22.87%)
LLAMA 3.1-70B INSTRUCT	124 (23.98%)	111 (22.11%)	181 (24.13%)	250 (25.10%)
LLAMA 3.1-405B INSTRUCT FP8	124 (23.98%)	134 (26.80%)	203 (26.96%)	226 (22.69%)
QWEN 3-1.7B	110 (20.87%)	141 (28.20%)	181 (24.10%)	240 (23.72%)
QWEN 3-8B	105 (20.11%)	124 (24.51%)	206 (27.32%)	241 (24.15%)
QWEN 3-32B	115 (21.90%)	118 (23.51%)	190 (25.23%)	222 (22.22%)
PHI 3-14B	51 (9.75%)	102 (20.32%)	168 (22.37%)	235 (23.43%)
PHI4-MINI-INSTRUCT 3.8B	77 (14.89%)	92 (18.33%)	122 (16.22%)	173 (17.37%)
PHI 4-14B	97 (18.76%)	111 (22.07%)	175 (23.12%)	233 (23.12%)
MISTRAL-7B-INSTRUCT-V0.2	107 (20.54%)	93 (18.42%)	170 (22.40%)	196 (19.66%)
MISTRAL-NEMO-INSTRUCT-2407 12B	114 (21.59%)	77 (15.37%)	182 (24.17%)	230 (23.09%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	108 (20.89%)	111 (21.06%)	202 (26.68%)	229 (22.99%)
GPT-4O-MINI	128 (24.62%)	146 (28.91%)	201 (26.31%)	263 (26.38%)

Table 37: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.4$) Panel: QA.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ
LLAMA 3.1-8B INSTRUCT	178 (8.57%)	98 (13.37%)	8 (5.00%)	696 (24.70%)	879 (29.46%)
LLAMA 3.1-70B INSTRUCT	96 (11.57%)	84 (12.61%)	8 (5.88%)	644 (24.07%)	691 (27.74%)
LLAMA 3.1-405B INSTRUCT FP8	70 (11.01%)	91 (13.98%)	36 (25.35%)	636 (25.59%)	667 (29.14%)
QWEN 3-1.7B	1285 (29.72%)	146 (14.19%)	46 (25.27%)	871 (26.53%)	1290 (35.31%)
QWEN 3-8B	497 (14.95%)	295 (31.82%)	8 (4.65%)	1066 (35.68%)	1098 (32.88%)
QWEN 3-32B	231 (10.43%)	180 (24.46%)	5 (3.21%)	672 (25.21%)	876 (29.21%)
PHI 3-14B	185 (7.09%)	172 (12.55%)	54 (33.96%)	717 (22.37%)	1031 (32.74%)
PHI4-MINI-INSTRUCT 3.8B	639 (15.07%)	173 (12.80%)	50 (28.57%)	697 (21.00%)	1195 (34.00%)
PHI 4-14B	121 (5.95%)	127 (14.02%)	43 (28.48%)	579 (21.18%)	836 (28.62%)
MISTRAL-7B-INSTRUCT-V0.2	296 (14.31%)	196 (18.90%)	32 (19.63%)	752 (25.14%)	911 (30.33%)
MISTRAL-NEMO-INSTRUCT-2407 12B	177 (8.17%)	229 (14.79%)	54 (26.73%)	694 (19.77%)	1096 (32.78%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	221 (14.67%)	181 (16.16%)	49 (30.82%)	827 (26.30%)	925 (29.91%)
GPT-4O-MINI	213 (17.85%)	144 (16.84%)	25 (17.86%)	695 (24.88%)	708 (27.41%)

Table 37: Overconfidence on QA and Math datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.4$) Panel: Math.

Model / method	SVAMP	GSM8k
LLAMA 3.1-8B INSTRUCT	35 (22.88%)	25 (13.16%)
LLAMA 3.1-70B INSTRUCT	5 (8.62%)	18 (15.65%)
LLAMA 3.1-405B INSTRUCT FP8	11 (25.00%)	10 (12.50%)
QWEN 3-1.7B	3 (10.34%)	88 (17.32%)
QWEN 3-8B	3 (13.04%)	82 (16.57%)
QWEN 3-32B	3 (11.54%)	72 (18.23%)
PHI 3-14B	5 (13.16%)	111 (22.52%)
PHI4-MINI-INSTRUCT 3.8B	12 (17.14%)	132 (26.40%)
PHI 4-14B	11 (9.82%)	102 (17.38%)
MISTRAL-7B-INSTRUCT-V0.2	32 (20.00%)	186 (27.11%)
MISTRAL-NEMO-INSTRUCT-2407 12B	37 (25.34%)	107 (26.16%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	14 (16.87%)	122 (26.81%)
GPT-4O-MINI	22 (40.00%)	125 (19.81%)

Table 38: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.4$) Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
LLAMA 3.1-8B INSTRUCT	1096 (38.06%)	181 (31.64%)	1051 (36.57%)	455 (32.36%)
LLAMA 3.1-70B INSTRUCT	1046 (36.32%)	169 (31.65%)	1033 (35.77%)	484 (34.42%)
LLAMA 3.1-405B INSTRUCT FP8	1009 (35.02%)	161 (30.21%)	952 (33.00%)	469 (33.33%)
QWEN 3-1.7B	1166 (41.14%)	209 (31.52%)	1003 (34.66%)	548 (38.98%)
QWEN 3-8B	1133 (39.37%)	195 (33.68%)	1009 (34.47%)	531 (37.77%)
QWEN 3-32B	1095 (38.18%)	205 (35.59%)	1040 (36.15%)	512 (36.39%)
PHI 3-14B	1137 (39.95%)	216 (36.55%)	1059 (36.85%)	37 (2.63%)
PHI4-MINI-INSTRUCT 3.8B	1155 (40.70%)	200 (34.90%)	1041 (36.20%)	153 (10.88%)
PHI 4-14B	1170 (40.87%)	169 (35.43%)	1137 (39.58%)	564 (40.11%)
MISTRAL-7B-INSTRUCT-V0.2	1091 (38.25%)	190 (35.65%)	1030 (35.23%)	523 (37.20%)
MISTRAL-NEMO-INSTRUCT-2407 12B	1138 (40.10%)	214 (33.18%)	959 (33.21%)	494 (35.14%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	1017 (35.82%)	192 (33.45%)	995 (34.55%)	525 (37.34%)
GPT-4O-MINI	1130 (38.63%)	197 (32.94%)	1122 (39.05%)	559 (39.76%)

Table 38: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.4$) Panel: Translation: cs, de, fi, gu.

Model / method	cs	de	fi	gu
LLAMA 3.1-8B INSTRUCT	270 (36.14%)	180 (23.81%)	236 (31.42%)	111 (21.51%)
LLAMA 3.1-70B INSTRUCT	261 (34.62%)	187 (24.64%)	240 (31.62%)	129 (25.49%)
LLAMA 3.1-405B INSTRUCT FP8	276 (36.90%)	226 (29.66%)	276 (36.13%)	143 (27.45%)
QWEN 3-1.7B	261 (34.89%)	213 (27.38%)	237 (31.52%)	128 (25.25%)
QWEN 3-8B	270 (35.90%)	200 (26.67%)	247 (32.63%)	142 (28.06%)
QWEN 3-32B	257 (34.04%)	201 (26.76%)	238 (31.65%)	122 (24.40%)
PHI 3-14B	557 (43.01%)	536 (41.29%)	576 (42.83%)	360 (40.59%)
PHI4-MINI-INSTRUCT 3.8B	133 (16.12%)	114 (15.14%)	190 (25.07%)	371 (45.75%)
PHI 4-14B	249 (33.07%)	191 (25.43%)	232 (30.93%)	120 (24.00%)
MISTRAL-7B-INSTRUCT-V0.2	228 (30.44%)	194 (25.87%)	209 (27.83%)	177 (35.40%)
MISTRAL-NEMO-INSTRUCT-2407 12B	228 (30.56%)	175 (23.00%)	203 (26.85%)	108 (21.09%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	263 (35.25%)	206 (27.43%)	234 (30.75%)	130 (25.95%)
GPT-4O-MINI	285 (38.00%)	207 (27.56%)	263 (34.38%)	151 (28.17%)

Table 38: Overconfidence on Summarization and Translation datasets. Each cell shows overconfident count (OC ratio %). ($\alpha = 0.4$) Panel: Translation: kk, lt, ru, zh.

Model / method	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT	151 (28.33%)	150 (30.00%)	259 (34.53%)	315 (31.59%)
LLAMA 3.1-70B INSTRUCT	164 (31.72%)	155 (30.88%)	240 (32.00%)	347 (34.84%)
LLAMA 3.1-405B INSTRUCT FP8	167 (32.30%)	184 (36.80%)	282 (37.45%)	328 (32.93%)
QWEN 3-1.7B	161 (30.55%)	171 (34.20%)	246 (32.76%)	344 (33.99%)
QWEN 3-8B	152 (29.12%)	173 (34.19%)	277 (36.74%)	329 (32.97%)
QWEN 3-32B	155 (29.52%)	169 (33.67%)	262 (34.79%)	324 (32.43%)
PHI 3-14B	67 (12.81%)	159 (31.67%)	237 (31.56%)	324 (32.30%)
PHI4-MINI-INSTRUCT 3.8B	108 (20.89%)	134 (26.69%)	187 (24.87%)	255 (25.60%)
PHI 4-14B	149 (28.82%)	147 (29.22%)	240 (31.70%)	328 (32.54%)
MISTRAL-7B-INSTRUCT-V0.2	154 (29.56%)	139 (27.52%)	243 (32.02%)	278 (27.88%)
MISTRAL-NEMO-INSTRUCT-2407 12B	151 (28.60%)	113 (22.55%)	258 (34.26%)	312 (31.33%)
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	154 (29.79%)	162 (30.74%)	286 (37.78%)	333 (33.43%)
GPT-4O-MINI	164 (31.54%)	192 (38.02%)	268 (35.08%)	345 (34.60%)

H.7 GENERATION-TEMPERATURE EXPERIMENTS

Discussion: Appendix C.5.

Table 39: Generation performance with different temperatures. We use LLAMA 3.1-8B INSTRUCT and LLAMA 3.1-70B INSTRUCT as the generation model Panel: QA and math.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
Temperature = 0.1							
LLAMA 3.1-8B INSTRUCT	74.49	87.72	51.76	65.61	31.27	49.33	78.32
LLAMA 3.1-70B INSTRUCT	90.09	90.23	60.00	67.56	43.23	80.33	85.97
Temperature = 0.5							
LLAMA 3.1-8B INSTRUCT	74.64	87.58	50.59	65.51	31.24	50.00	83.02
LLAMA 3.1-70B INSTRUCT	89.95	90.03	61.18	67.48	43.13	80.33	89.76
Temperature = 1							
LLAMA 3.1-8B INSTRUCT	74.03	87.63	52.94	64.70	30.43	49.00	85.60
LLAMA 3.1-70B INSTRUCT	89.62	88.77	60.00	66.49	41.92	80.67	91.28
Temperature = 1.5							
LLAMA 3.1-8B INSTRUCT	72.91	86.29	47.06	64.00	28.42	48.33	83.70
LLAMA 3.1-70B INSTRUCT	88.88	88.09	56.47	65.16	38.26	80.33	87.57
Temperature = 2							
LLAMA 3.1-8B INSTRUCT	69.92	84.18	48.24	62.22	25.11	46.33	0.53
LLAMA 3.1-70B INSTRUCT	87.90	85.24	57.65	62.34	37.68	81.67	1.06
Temperature = 2.5							
LLAMA 3.1-8B INSTRUCT	64.56	80.68	43.53	58.61	21.22	38.00	0.08
LLAMA 3.1-70B INSTRUCT	83.64	79.88	56.47	57.55	31.85	75.00	0.53
Temperature = 3							
LLAMA 3.1-8B INSTRUCT	55.01	73.48	38.82	52.65	16.53	29.67	0.08
LLAMA 3.1-70B INSTRUCT	75.83	68.29	45.88	51.11	23.81	64.67	0.45

Table 39: Generation performance with different temperatures. We use LLAMA 3.1-8B INSTRUCT and LLAMA 3.1-70B INSTRUCT as the generation model Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DailyMail	MultiNews
Temperature = 0.1				
LLAMA 3.1-8B INSTRUCT	19.14	15.60	22.02	16.02
LLAMA 3.1-70B INSTRUCT	21.16	17.50	22.41	16.55
Temperature = 0.5				
LLAMA 3.1-8B INSTRUCT	19.26	15.88	21.97	15.97
LLAMA 3.1-70B INSTRUCT	38.26	18.16	22.32	21.20
Temperature = 1				
LLAMA 3.1-8B INSTRUCT	18.92	15.13	21.70	15.81
LLAMA 3.1-70B INSTRUCT	20.74	18.53	22.21	16.20
Temperature = 1.5				
LLAMA 3.1-8B INSTRUCT	18.06	13.32	20.64	15.02
LLAMA 3.1-70B INSTRUCT	18.94	16.27	20.59	14.93
Temperature = 2				
LLAMA 3.1-8B INSTRUCT	2.90	0.47	5.08	4.30
LLAMA 3.1-70B INSTRUCT	1.79	0.36	3.06	2.71
Temperature = 2.5				
LLAMA 3.1-8B INSTRUCT	0.55	0.13	1.03	1.19
LLAMA 3.1-70B INSTRUCT	0.39	0.09	0.64	0.93
Temperature = 3				
LLAMA 3.1-8B INSTRUCT	0.37	0.08	0.61	0.89
LLAMA 3.1-70B INSTRUCT	0.30	0.07	0.49	0.77

Table 39: Generation performance with different temperatures. We use LLAMA 3.1-8B INSTRUCT and LLAMA 3.1-70B INSTRUCT as the generation model Panel: Translation.

Model / method	cs	de	fi	gu	kk	lt	ru	zh
Temperature = 0.1								
LLAMA 3.1-8B INSTRUCT	54.43	66.92	47.84	48.64	38.18	47.90	59.35	48.96
LLAMA 3.1-70B INSTRUCT	59.35	71.74	52.93	61.36	47.56	56.63	63.37	54.40
Temperature = 0.5								
LLAMA 3.1-8B INSTRUCT	54.45	66.94	47.68	48.65	38.05	47.77	59.37	48.92
LLAMA 3.1-70B INSTRUCT	59.35	71.39	52.58	61.12	47.41	56.55	63.28	54.19
Temperature = 1								
LLAMA 3.1-8B INSTRUCT	54.01	66.58	46.90	46.90	36.61	46.76	58.90	47.91
LLAMA 3.1-70B INSTRUCT	57.58	70.07	50.91	57.43	45.13	54.23	62.11	51.36
Temperature = 1.5								
LLAMA 3.1-8B INSTRUCT	51.17	64.43	43.35	33.87	29.82	39.22	56.34	42.13
LLAMA 3.1-70B INSTRUCT	38.41	56.28	28.41	20.84	17.96	26.61	42.91	22.33
Temperature = 2								
LLAMA 3.1-8B INSTRUCT	14.41	26.44	9.43	3.83	3.02	4.32	20.09	6.61
LLAMA 3.1-70B INSTRUCT	1.00	2.11	0.73	0.58	0.62	0.69	1.20	0.72
Temperature = 2.5								
LLAMA 3.1-8B INSTRUCT	0.91	1.90	0.79	0.59	0.60	0.64	1.30	0.78
LLAMA 3.1-70B INSTRUCT	0.33	0.36	0.33	0.33	0.33	0.36	0.37	0.36
Temperature = 3								
LLAMA 3.1-8B INSTRUCT	0.42	0.50	0.39	0.36	0.44	0.39	0.50	0.45
LLAMA 3.1-70B INSTRUCT	0.26	0.29	0.28	0.30	0.31	0.31	0.30	0.30

Table 40: SUQ experimental result with LLAMA 3.1-8B INSTRUCT with different generation temperature. Panel: QA and math.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
Temperature = 0.1							
Discrete semantic entropy	0.5885/0.7852	0.5731/0.8997	0.6012/0.5531	0.5649/0.6874	0.6032/0.3692	0.5607/0.5498	0.5345/0.8006
Discrete U_{eig}	0.5911/0.7856	0.5739/0.8998	0.6095/0.5747	0.5681/0.6859	0.6086/0.3698	0.5597/0.5493	0.5481/0.8058
Discrete U_{deg}	0.5910/0.7856	0.5737/0.8997	0.6073/0.6143	0.5680/0.6855	0.6079/0.3696	0.5606/0.5496	0.5466/0.8054
KLE-heat$_{t=0.1}$	0.5905/0.7855	0.5736/0.8997	0.6064/0.6135	0.5676/0.6862	0.6072/0.3700	0.5607/0.5496	0.5465/0.8054
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5910/0.7856	0.5737/0.8997	0.6125/0.6160	0.5680/0.6862	0.6077/0.3695	0.5609/0.5497	0.5467/0.8057
Temperature = 0.5							
Discrete semantic entropy	0.7535/0.8546	0.6757/0.9140	0.6980/0.6321	0.6232/0.7239	0.6885/0.4357	0.6761/0.6036	0.5667/0.8546
Discrete U_{eig}	0.7718/0.8649	0.6846/0.9149	0.7132/0.6530	0.6438/0.7315	0.6916/0.4315	0.7003/0.6160	0.5895/0.8596
Discrete U_{deg}	0.7715/0.8639	0.6830/0.9155	0.7029/0.6531	0.6422/0.7276	0.6951/0.4336	0.6984/0.6117	0.5785/0.8577
KLE-heat$_{t=0.1}$	0.7690/0.8655	0.6822/0.9139	0.6891/0.6391	0.6402/0.7292	0.6968/0.4366	0.6942/0.6173	0.5782/0.8570
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7711/0.8695	0.6830/0.9152	0.6991/0.6297	0.6423/0.7273	0.6974/0.4368	0.6995/0.6185	0.5789/0.8573
Temperature = 1							
Discrete semantic entropy	0.8259/0.8914	0.7197/0.9350	0.7417/0.6832	0.6727/0.7501	0.7274/0.4611	0.7079/0.6436	0.5283/0.8678
Discrete U_{eig}	0.8446/0.9019	0.7336/0.9386	0.6914/0.6745	0.7043/0.7627	0.7156/0.4526	0.7449/0.6654	0.5867/0.8805
Discrete U_{deg}	0.8494/0.9015	0.7294/0.9352	0.7139/0.6876	0.7006/0.7584	0.7253/0.4585	0.7404/0.6592	0.5730/0.8670
KLE-heat$_{t=0.1}$	0.8483/0.9019	0.7290/0.9366	0.7364/0.6969	0.6975/0.7611	0.7314/0.4614	0.7330/0.6576	0.5712/0.8781
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8487/0.8980	0.7297/0.9372	0.7181/0.6889	0.7014/0.7604	0.7295/0.4618	0.7293/0.6539	0.5748/0.8675
Temperature = 1.5							
Discrete semantic entropy	0.8482/0.9015	0.7615/0.9414	0.7219/0.6560	0.7077/0.7662	0.7503/0.4565	0.6606/0.6108	0.5511/0.8524
Discrete U_{eig}	0.8503/0.9032	0.7766/0.9416	0.6889/0.6451	0.7369/0.7771	0.7012/0.4309	0.6901/0.6269	0.6242/0.8735
Discrete U_{deg}	0.8630/0.9079	0.7709/0.9425	0.7164/0.6587	0.7363/0.7782	0.7266/0.4422	0.7056/0.6304	0.6028/0.8732
KLE-heat$_{t=0.1}$	0.8658/0.9090	0.7700/0.9416	0.7283/0.6671	0.7346/0.7774	0.7423/0.4503	0.6943/0.6287	0.6001/0.8673
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8641/0.9079	0.7716/0.9398	0.7206/0.6614	0.7369/0.7788	0.7359/0.4472	0.6873/0.6269	0.6062/0.8760
Temperature = 2							
Discrete semantic entropy	0.8616/0.8962	0.7926/0.9367	0.7622/0.7088	0.7132/0.7585	0.7467/0.4212	0.7263/0.6390	0.4898/0.0118
Discrete U_{eig}	0.8470/0.8912	0.7933/0.9387	0.7439/0.6875	0.7329/0.7705	0.6866/0.3841	0.7473/0.6656	0.4561/0.0103
Discrete U_{deg}	0.8704/0.8990	0.7920/0.9365	0.7539/0.6930	0.7379/0.7716	0.7196/0.4016	0.7679/0.6740	0.4598/0.0115
KLE-heat$_{t=0.1}$	0.8783/0.9021	0.7948/0.9387	0.7677/0.6997	0.7394/0.7717	0.7428/0.4133	0.7651/0.6638	0.4539/0.0102
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8750/0.9001	0.7933/0.9363	0.7605/0.6965	0.7391/0.7727	0.7348/0.4084	0.7608/0.6618	0.4544/0.0102
Temperature = 2.5							
Discrete semantic entropy	0.8622/0.8663	0.7898/0.9205	0.7041/0.6458	0.7332/0.7487	0.7329/0.3656	0.7722/0.6009	0.5463/0.0076
Discrete U_{eig}	0.8243/0.8486	0.7554/0.9157	0.6593/0.5817	0.7201/0.7470	0.6458/0.3048	0.7797/0.6131	0.4606/0.0043
Discrete U_{deg}	0.8566/0.8635	0.7739/0.9184	0.6847/0.5996	0.7403/0.7542	0.6842/0.3272	0.8008/0.6224	0.5629/0.0067
KLE-heat$_{t=0.1}$	0.8713/0.8701	0.7849/0.9214	0.7033/0.6102	0.7495/0.7568	0.7155/0.3458	0.8067/0.6249	0.5148/0.0054
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8663/0.8675	0.7749/0.9186	0.6965/0.6093	0.7422/0.7540	0.7016/0.3349	0.8054/0.6237	0.5174/0.0056
Temperature = 3							
Discrete semantic entropy	0.8546/0.8033	0.7959/0.8893	0.8033/0.6029	0.7296/0.7022	0.6931/0.2830	0.7346/0.4601	0.4733/0.0025
Discrete U_{eig}	0.8025/0.7695	0.6784/0.8522	0.7477/0.5775	0.6696/0.6704	0.6273/0.2260	0.7228/0.4487	0.4905/0.0039
Discrete U_{deg}	0.8419/0.7936	0.7370/0.8703	0.7695/0.5936	0.7076/0.6897	0.6647/0.2496	0.7522/0.4765	0.3361/0.0019
KLE-heat$_{t=0.1}$	0.8645/0.8061	0.7669/0.8802	0.7890/0.6058	0.7250/0.6987	0.6959/0.2676	0.7657/0.4858	0.4014/0.0024
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8587/0.8017	0.7417/0.8713	0.7716/0.5940	0.7100/0.6904	0.6816/0.2565	0.7725/0.4941	0.4014/0.0024

Table 40: SUQ experimental result with LLAMA 3.1-8B INSTRUCT with different generation temperature. Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
Temperature = 0.1				
Discrete semantic entropy	0.2427	0.2322	0.3574	0.3237
Discrete U_{eig}	0.2411	0.2325	0.3582	0.3237
Discrete U_{deg}	0.2418	0.2325	0.3579	0.3233
KLE-heat$_{t=0.1}$	0.2421	0.2323	0.3584	0.3231
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2418	0.2324	0.3575	0.3233
Temperature = 0.5				
Discrete semantic entropy	0.2479	0.1746	0.3607	0.3108
Discrete U_{eig}	0.2437	0.1739	0.3599	0.3109
Discrete U_{deg}	0.2436	0.1743	0.3582	0.3089
KLE-heat$_{t=0.1}$	0.2442	0.1748	0.3584	0.3091
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2439	0.1736	0.3581	0.3088
Temperature = 1				
Discrete semantic entropy	0.2539	0.1799	0.3316	0.2806
Discrete U_{eig}	0.2478	0.1806	0.3326	0.2798
Discrete U_{deg}	0.2480	0.1786	0.3317	0.2795
KLE-heat$_{t=0.1}$	0.2489	0.1788	0.3317	0.2794
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2480	0.1789	0.3317	0.2794
Temperature = 1.5				
Discrete semantic entropy	0.2473	0.1567	0.2568	0.2770
Discrete U_{eig}	0.2441	0.1585	0.2611	0.2842
Discrete U_{deg}	0.2443	0.1592	0.2597	0.2822
KLE-heat$_{t=0.1}$	0.2450	0.1602	0.2595	0.2819
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2443	0.1592	0.2598	0.2824
Temperature = 2				
Discrete semantic entropy	0.0613	0.0094	0.0949	0.1458
Discrete U_{eig}	0.0799	0.0088	0.0958	0.1469
Discrete U_{deg}	0.0800	0.0094	0.0962	0.1484
KLE-heat$_{t=0.1}$	0.0790	0.0095	0.0961	0.1480
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0800	0.0097	0.0961	0.1482
Temperature = 2.5				
Discrete semantic entropy	0.0237	0.0408	0.0961	0.1808
Discrete U_{eig}	0.0236	0.0392	0.0988	0.1785
Discrete U_{deg}	0.0237	0.0412	0.0980	0.1812
KLE-heat$_{t=0.1}$	0.0237	0.0401	0.0987	0.1790
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0237	0.0399	0.0987	0.1786
Temperature = 3				
Discrete semantic entropy	0.0517	0.0341	0.0589	0.2240
Discrete U_{eig}	0.0500	0.0272	0.0606	0.2248
Discrete U_{deg}	0.0510	0.0287	0.0600	0.2240
KLE-heat$_{t=0.1}$	0.0504	0.0270	0.0601	0.2241
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0505	0.0275	0.0601	0.2235

Table 40: SUQ experimental result with LLAMA 3.1-8B INSTRUCT with different generation temperature. Panel: Translation.

Model / method	cs	de	fi	gu	kk	lt	ru	zh
Temperature = 0.1								
Discrete semantic entropy	0.5382	0.6759	0.4876	0.5151	0.4031	0.4935	0.6042	0.4942
Discrete U_{eig}	0.5387	0.6756	0.4840	0.5149	0.4036	0.4975	0.6052	0.4964
Discrete U_{deg}	0.5387	0.6756	0.4839	0.5175	0.4033	0.4971	0.6052	0.4953
KLE-heat$_{t=0.1}$	0.5387	0.6756	0.4840	0.5183	0.4036	0.4955	0.6052	0.4952
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5648	0.6756	0.4839	0.5148	0.4034	0.5003	0.6052	0.4968
Temperature = 0.5								
Discrete semantic entropy	0.5532	0.6888	0.5019	0.5455	0.4154	0.5165	0.6036	0.5153
Discrete U_{eig}	0.5562	0.6879	0.5055	0.5502	0.4200	0.5146	0.6097	0.5131
Discrete U_{deg}	0.5588	0.6872	0.5041	0.5458	0.4213	0.5181	0.6139	0.5125
KLE-heat$_{t=0.1}$	0.5600	0.6884	0.5083	0.5460	0.4206	0.5152	0.6128	0.5110
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5579	0.6866	0.5073	0.5468	0.4142	0.5131	0.6161	0.5111
Temperature = 1								
Discrete semantic entropy	0.5594	0.6848	0.5001	0.5356	0.4052	0.5128	0.6058	0.5093
Discrete U_{eig}	0.5561	0.6978	0.5020	0.5400	0.4110	0.5076	0.6058	0.5124
Discrete U_{deg}	0.5504	0.6979	0.5056	0.5416	0.4090	0.5081	0.6050	0.5114
KLE-heat$_{t=0.1}$	0.5554	0.6971	0.5032	0.5405	0.4090	0.5090	0.6076	0.5101
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5529	0.6984	0.5040	0.5394	0.4092	0.5093	0.6035	0.5101
Temperature = 1.5								
Discrete semantic entropy	0.5511	0.6848	0.4873	0.4545	0.3629	0.4669	0.5961	0.4756
Discrete U_{eig}	0.5592	0.6919	0.4879	0.4538	0.3702	0.4665	0.5974	0.4813
Discrete U_{deg}	0.5555	0.6905	0.4876	0.4581	0.3714	0.4692	0.5945	0.4793
KLE-heat$_{t=0.1}$	0.5611	0.6921	0.4906	0.4596	0.3730	0.4713	0.5983	0.4798
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5555	0.6906	0.4889	0.4580	0.3702	0.4699	0.5977	0.4823
Temperature = 2								
Discrete semantic entropy	0.2904	0.4652	0.2063	0.0946	0.0557	0.0982	0.3733	0.1442
Discrete U_{eig}	0.2892	0.4621	0.2062	0.0969	0.0551	0.0985	0.3709	0.1443
Discrete U_{deg}	0.2903	0.4647	0.2068	0.0978	0.0557	0.0992	0.3736	0.1460
KLE-heat$_{t=0.1}$	0.2922	0.4653	0.2066	0.0975	0.0556	0.0997	0.3752	0.1460
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2917	0.4655	0.2062	0.0980	0.0556	0.1002	0.3744	0.1465
Temperature = 2.5								
Discrete semantic entropy	0.0426	0.0307	0.0106	0.0570	0.0453	0.0136	0.0177	0.0096
Discrete U_{eig}	0.0441	0.0290	0.0092	0.0537	0.0459	0.0126	0.0159	0.0094
Discrete U_{deg}	0.0451	0.0305	0.0093	0.0535	0.0466	0.0132	0.0174	0.0096
KLE-heat$_{t=0.1}$	0.0450	0.0307	0.0096	0.0540	0.0464	0.0132	0.0184	0.0085
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0449	0.0315	0.0097	0.0538	0.0463	0.0130	0.0185	0.0097
Temperature = 3								
Discrete semantic entropy	0.0760	0.0560	0.0923	0.0925	0.1122	0.0547	0.0659	0.0557
Discrete U_{eig}	0.0741	0.0581	0.1006	0.0956	0.1074	0.0502	0.0691	0.0587
Discrete U_{deg}	0.0753	0.0582	0.1023	0.0950	0.1084	0.0505	0.0690	0.0593
KLE-heat$_{t=0.1}$	0.0757	0.0583	0.1015	0.0951	0.1082	0.0507	0.0690	0.0587
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0752	0.0593	0.1011	0.0949	0.1081	0.0507	0.0691	0.0584

Table 41: SUQ experimental result with LLAMA 3.1-70B INSTRUCT with different generation temperature. Panel: QA and math.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
Temperature = 0.1							
Discrete semantic entropy	0.5811/0.9205	0.5917/0.9245	0.5378/0.6101	0.5699/0.7066	0.5884/0.4852	0.6074/0.8229	0.5194/0.8602
Discrete U_{eig}	0.5828/0.9207	0.5920/0.9245	0.5444/0.6074	0.5726/0.7034	0.5911/0.4844	0.6073/0.8226	0.5345/0.8678
Discrete U_{deg}	0.5828/0.9207	0.5920/0.9245	0.5384/0.6055	0.5729/0.7075	0.5911/0.4851	0.6074/0.8226	0.5335/0.8677
KLE-heat$_{t=0.1}$	0.5826/0.9207	0.5917/0.9245	0.5343/0.6108	0.5730/0.7064	0.5903/0.4870	0.6077/0.8228	0.5333/0.8677
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5828/0.9207	0.5920/0.9245	0.5369/0.6054	0.5730/0.7078	0.5910/0.4841	0.6077/0.8228	0.5335/0.8677
Temperature = 0.5							
Discrete semantic entropy	0.7186/0.9448	0.7294/0.9503	0.5932/0.6941	0.6424/0.7471	0.6619/0.5480	0.7352/0.9107	0.5283/0.8977
Discrete U_{eig}	0.7273/0.9458	0.7331/0.9416	0.6413/0.7015	0.6579/0.7568	0.6673/0.5365	0.7539/0.9180	0.5631/0.9121
Discrete U_{deg}	0.7273/0.9458	0.7318/0.9434	0.6334/0.6963	0.6555/0.7614	0.6659/0.5407	0.7501/0.9178	0.5519/0.9119
KLE-heat$_{t=0.1}$	0.7263/0.9458	0.7315/0.9420	0.6291/0.6975	0.6539/0.7584	0.6653/0.5402	0.7461/0.9189	0.5512/0.9134
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7270/0.9458	0.7319/0.9427	0.6332/0.6968	0.6554/0.7564	0.6660/0.5390	0.7478/0.9162	0.5534/0.9121
Temperature = 1							
Discrete semantic entropy	0.8073/0.9592	0.7654/0.9492	0.6298/0.7105	0.6777/0.7712	0.6971/0.5562	0.8215/0.9399	0.5085/0.9074
Discrete U_{eig}	0.8218/0.9601	0.7686/0.9447	0.6990/0.7570	0.7032/0.7818	0.6884/0.5567	0.8522/0.9480	0.5490/0.9303
Discrete U_{deg}	0.8229/0.9603	0.7651/0.9454	0.6635/0.7397	0.7003/0.7821	0.6932/0.5604	0.8521/0.9486	0.5409/0.9281
KLE-heat$_{t=0.1}$	0.8216/0.9608	0.7655/0.9451	0.6540/0.7336	0.6983/0.7796	0.6969/0.5559	0.8444/0.9470	0.5404/0.9303
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8230/0.9592	0.7653/0.9482	0.6621/0.7374	0.7002/0.7797	0.6951/0.5612	0.8492/0.9480	0.5415/0.9238
Temperature = 1.5							
Discrete semantic entropy	0.8435/0.9674	0.7811/0.9517	0.6768/0.6977	0.7132/0.7869	0.6862/0.5267	0.8578/0.9470	0.5657/0.8898
Discrete U_{eig}	0.8513/0.9679	0.7885/0.9560	0.6920/0.7094	0.7267/0.7939	0.6707/0.5308	0.8741/0.9516	0.6257/0.9022
Discrete U_{deg}	0.8570/0.9676	0.7843/0.9546	0.6827/0.6952	0.7260/0.7929	0.6842/0.5337	0.8780/0.9524	0.6164/0.9116
KLE-heat$_{t=0.1}$	0.8590/0.9689	0.7848/0.9553	0.6886/0.6983	0.7267/0.7917	0.6947/0.5408	0.8761/0.9521	0.6136/0.9115
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8584/0.9691	0.7846/0.9540	0.6824/0.6942	0.7268/0.7934	0.6919/0.5384	0.8767/0.9523	0.6186/0.9029
Temperature = 2							
Discrete semantic entropy	0.8427/0.9659	0.7937/0.9430	0.5842/0.6424	0.7420/0.7850	0.7226/0.5463	0.8407/0.9467	0.6874/0.0175
Discrete U_{eig}	0.8353/0.9659	0.7902/0.9445	0.6689/0.7115	0.7397/0.7863	0.6601/0.5212	0.8631/0.9539	0.8064/0.0183
Discrete U_{deg}	0.8492/0.9663	0.7899/0.9430	0.6497/0.6994	0.7442/0.7858	0.6893/0.5327	0.8725/0.9561	0.7673/0.0178
KLE-heat$_{t=0.1}$	0.8551/0.9673	0.7944/0.9439	0.6378/0.6913	0.7471/0.7872	0.7064/0.5411	0.8725/0.9563	0.7599/0.0188
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8517/0.9670	0.7896/0.9427	0.6463/0.6969	0.7440/0.7868	0.6945/0.5356	0.8790/0.9577	0.7737/0.0179
Temperature = 2.5							
Discrete semantic entropy	0.7928/0.9374	0.8031/0.9202	0.4758/0.5826	0.7501/0.7521	0.6989/0.4618	0.7413/0.8536	0.3517/0.0021
Discrete U_{eig}	0.8133/0.9448	0.7619/0.9138	0.5709/0.6418	0.7184/0.7385	0.6355/0.4328	0.7394/0.8851	0.2645/0.0010
Discrete U_{deg}	0.8278/0.9472	0.7796/0.9164	0.5481/0.6282	0.7337/0.7441	0.6733/0.4511	0.7562/0.8894	0.4389/0.0020
KLE-heat$_{t=0.1}$	0.8335/0.9488	0.7916/0.9191	0.5405/0.6246	0.7418/0.7479	0.6950/0.4641	0.7643/0.8929	0.3823/0.0017
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8320/0.9475	0.7787/0.9164	0.5501/0.6282	0.7334/0.7434	0.6787/0.4544	0.7751/0.8956	0.3910/0.0018
Temperature = 3							
Discrete semantic entropy	0.6623/0.8385	0.7383/0.8173	0.4337/0.4452	0.6964/0.6568	0.6452/0.3187	0.6720/0.7266	0.4300/0.0009
Discrete U_{eig}	0.7413/0.8833	0.6964/0.8211	0.5178/0.4683	0.6627/0.6509	0.6363/0.3335	0.6751/0.7857	0.4364/0.0023
Discrete U_{deg}	0.7385/0.8787	0.7190/0.8239	0.5025/0.4433	0.6805/0.6555	0.6599/0.3419	0.6998/0.7927	0.4324/0.0029
KLE-heat$_{t=0.1}$	0.7357/0.8778	0.7344/0.8299	0.4922/0.4431	0.6910/0.6608	0.6710/0.3460	0.7181/0.8016	0.4264/0.0026
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7438/0.8806	0.7191/0.8240	0.5022/0.4494	0.6810/0.6555	0.6617/0.3429	0.7219/0.8029	0.4216/0.0025

Table 41: SUQ experimental result with LLAMA 3.1-70B INSTRUCT with different generation temperature. Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
Temperature = 0.1				
Discrete semantic entropy	0.2698	0.2116	0.3311	0.3195
Discrete U_{eig}	0.2704	0.2116	0.3318	0.3204
Discrete U_{deg}	0.2702	0.2116	0.3311	0.3193
KLE-heat$_{t=0.1}$	0.2703	0.2116	0.3309	0.3197
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2700	0.2115	0.3313	0.3190
Temperature = 0.5				
Discrete semantic entropy	0.2838	0.2217	0.3388	0.3325
Discrete U_{eig}	0.2829	0.2160	0.3392	0.3311
Discrete U_{deg}	0.2821	0.2141	0.3364	0.3298
KLE-heat$_{t=0.1}$	0.2825	0.2151	0.3365	0.3300
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2819	0.2156	0.3363	0.3298
Temperature = 1				
Discrete semantic entropy	0.2672	0.2204	0.3365	0.3018
Discrete U_{eig}	0.2648	0.2187	0.3387	0.3024
Discrete U_{deg}	0.2645	0.2147	0.3370	0.3015
KLE-heat$_{t=0.1}$	0.2649	0.2146	0.3370	0.3015
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2644	0.2146	0.3371	0.3015
Temperature = 1.5				
Discrete semantic entropy	0.2687	0.2058	0.3061	0.2774
Discrete U_{eig}	0.2814	0.2241	0.3174	0.2933
Discrete U_{deg}	0.2786	0.2222	0.3127	0.2867
KLE-heat$_{t=0.1}$	0.2777	0.2222	0.3119	0.2851
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2788	0.2226	0.3130	0.2873
Temperature = 2				
Discrete semantic entropy	0.0329	0.0054	0.0942	0.1246
Discrete U_{eig}	0.0478	0.0056	0.0930	0.1234
Discrete U_{deg}	0.0462	0.0054	0.0939	0.1243
KLE-heat$_{t=0.1}$	0.0427	0.0054	0.0936	0.1241
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0445	0.0055	0.0941	0.1243
Temperature = 2.5				
Discrete semantic entropy	0.0529	0.0258	0.0756	0.1297
Discrete U_{eig}	0.0536	0.0272	0.0746	0.1319
Discrete U_{deg}	0.0531	0.0255	0.0747	0.1306
KLE-heat$_{t=0.1}$	0.0530	0.0259	0.0749	0.1302
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0532	0.0254	0.0746	0.1299
Temperature = 3				
Discrete semantic entropy	0.0544	0.0217	0.1128	0.2182
Discrete U_{eig}	0.0533	0.0218	0.1132	0.2170
Discrete U_{deg}	0.0539	0.0205	0.1141	0.2168
KLE-heat$_{t=0.1}$	0.0536	0.0215	0.1144	0.2159
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0536	0.0211	0.1138	0.2155

Table 41: SUQ experimental result with LLAMA 3.1-70B INSTRUCT with different generation temperature. Panel: Translation.

Model / method	cs	de	fi	gu	kk	lt	ru	zh
Temperature = 0.1								
Discrete semantic entropy	0.5850	0.7162	0.5311	0.6329	0.4899	0.5717	0.6419	0.5482
Discrete U_{eig}	0.5864	0.7173	0.5313	0.6309	0.4907	0.5752	0.6416	0.5498
Discrete U_{deg}	0.5864	0.7173	0.5313	0.6308	0.4905	0.5748	0.6416	0.5487
KLE-heat$_{t=0.1}$	0.5864	0.7173	0.5313	0.6308	0.4903	0.5743	0.6416	0.5510
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5864	0.7173	0.5313	0.6308	0.4905	0.5749	0.6416	0.5495
Temperature = 0.5								
Discrete semantic entropy	0.6084	0.7319	0.5414	0.6510	0.4991	0.5870	0.6423	0.5686
Discrete U_{eig}	0.5999	0.7352	0.5466	0.6494	0.4951	0.5845	0.6464	0.5613
Discrete U_{deg}	0.6094	0.7352	0.5488	0.6483	0.4945	0.5869	0.6463	0.5664
KLE-heat$_{t=0.1}$	0.6095	0.7352	0.5486	0.6500	0.4942	0.5845	0.6462	0.5624
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.6077	0.7352	0.5484	0.6467	0.4924	0.5844	0.6462	0.5620
Temperature = 1								
Discrete semantic entropy	0.5884	0.7257	0.5329	0.6241	0.4770	0.5644	0.6349	0.5361
Discrete U_{eig}	0.5908	0.7285	0.5347	0.6274	0.4833	0.5770	0.6381	0.5406
Discrete U_{deg}	0.5941	0.7272	0.5288	0.6280	0.4823	0.5685	0.6351	0.5410
KLE-heat$_{t=0.1}$	0.5900	0.7281	0.5283	0.6292	0.4821	0.5716	0.6363	0.5418
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5878	0.7275	0.5281	0.6261	0.4818	0.5719	0.6358	0.5395
Temperature = 1.5								
Discrete semantic entropy	0.4857	0.6333	0.3563	0.2225	0.2110	0.3215	0.5318	0.2623
Discrete U_{eig}	0.4973	0.6499	0.3749	0.2844	0.2474	0.3585	0.5412	0.3198
Discrete U_{deg}	0.4971	0.6490	0.3731	0.2634	0.2408	0.3503	0.5401	0.3089
KLE-heat$_{t=0.1}$	0.4976	0.6458	0.3744	0.2562	0.2401	0.3491	0.5439	0.3064
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.4973	0.6472	0.3739	0.2653	0.2417	0.3507	0.5417	0.3093
Temperature = 2								
Discrete semantic entropy	0.0167	0.0382	0.0103	0.0423	0.0174	0.0076	0.0166	0.0078
Discrete U_{eig}	0.0165	0.0400	0.0109	0.0416	0.0153	0.0075	0.0178	0.0090
Discrete U_{deg}	0.0174	0.0413	0.0106	0.0424	0.0158	0.0076	0.0175	0.0086
KLE-heat$_{t=0.1}$	0.0174	0.0418	0.0104	0.0426	0.0159	0.0076	0.0174	0.0084
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0175	0.0414	0.0105	0.0425	0.0161	0.0077	0.0176	0.0085
Temperature = 2.5								
Discrete semantic entropy	0.0409	0.0818	0.0436	0.0603	0.0561	0.0903	0.0300	0.0113
Discrete U_{eig}	0.0417	0.0828	0.0420	0.0604	0.0533	0.0879	0.0290	0.0110
Discrete U_{deg}	0.0423	0.0820	0.0426	0.0607	0.0570	0.0864	0.0279	0.0109
KLE-heat$_{t=0.1}$	0.0423	0.0824	0.0427	0.0611	0.0567	0.0865	0.0284	0.0110
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0423	0.0818	0.0425	0.0612	0.0567	0.0872	0.0284	0.0108
Temperature = 3								
Discrete semantic entropy	0.0764	0.0924	0.0766	0.0783	0.0958	0.0536	0.0866	0.0367
Discrete U_{eig}	0.0777	0.0946	0.0804	0.0797	0.1038	0.0539	0.0862	0.0369
Discrete U_{deg}	0.0762	0.0961	0.0763	0.0798	0.1041	0.0552	0.0873	0.0373
KLE-heat$_{t=0.1}$	0.0777	0.0971	0.0768	0.0798	0.1032	0.0544	0.0864	0.0372
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.0772	0.0968	0.0776	0.0802	0.1049	0.0536	0.0864	0.0372

H.8 THINKING-ENABLED AND THINKING-DISABLED GENERATION

Discussion: Appendix C.6.

Table 42: Thinking comparison with QWEN3 8B. We compute the semantic uncertainty exhibited in the QWEN3 8B outputs under both thinking-enabled and thinking-disabled settings. Panel: QA and math.

Model / method	TriviaQA	SQUAD	BioASQ	CoQA	NQ	SVAMP	GSM8k
Enable Thinking							
Discrete semantic entropy	0.7493/0.8311	0.7266/0.8982	0.5790/0.6236	0.6616/0.7294	0.6834/0.4075	0.2988/0.8225	0.5108/0.6035
Discrete semantic density	0.7937/0.8576	0.7269/0.8922	0.6265/0.6519	0.6654/0.7358	0.6941/0.4240	0.2914/0.8288	0.4937/0.5680
KLE-heat$_{t=0.1}$	0.8021/0.8667	0.7339/0.8943	0.6457/0.6754	0.6747/0.7344	0.7081/0.4290	0.2916/0.8290	0.4937/0.5801
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8113/0.8684	0.7361/0.8930	0.6549/0.6705	0.6763/0.7352	0.7147/0.4342	0.2932/0.8290	0.4942/0.5814
U_{eig}	0.8381/0.8856	0.7646/0.9177	0.6665/0.6920	0.6882/0.7492	0.7175/0.4446	0.4774/0.8918	0.4222/0.5478
U_{deg}	0.8367/0.8859	0.7635/0.9168	0.6603/0.6879	0.6873/0.7450	0.7200/0.4446	0.4781/0.8923	0.4263/0.5520
LUQ-pair	0.7985/0.8725	0.7458/0.9131	0.6328/0.6660	0.6730/0.7355	0.6990/0.4304	0.7027/0.9410	0.5498/0.6760
SNNE	0.8227/0.8823	0.7623/0.9197	0.6572/0.6849	0.6865/0.7456	0.7027/0.4378	0.4818/0.8932	0.4279/0.5536
SNNE (only_num)	0.8237/0.8844	0.7627/0.9203	0.6572/0.6850	0.6866/0.7444	0.7035/0.4371	0.4819/0.8932	0.4277/0.5533
SNNE (only_denom)	0.8352/0.8866	0.7635/0.9217	0.6582/0.6855	0.6872/0.7473	0.7213/0.4445	0.4818/0.8931	0.4274/0.5530
Disable Thinking							
Discrete semantic entropy	0.8118/0.7746	0.6650/0.9001	0.7193/0.6844	0.6349/0.6975	0.7129/0.3381	0.6627/0.9511	0.4941/0.6101
Discrete semantic density	0.8214/0.7889	0.6732/0.9038	0.6991/0.6672	0.6392/0.6994	0.7131/0.3352	0.7145/0.9579	0.4944/0.6158
KLE-heat$_{t=0.1}$	0.8225/0.7842	0.6695/0.9018	0.6885/0.6581	0.6407/0.7043	0.7167/0.3403	0.7227/0.9539	0.4943/0.6154
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8234/0.7852	0.6698/0.9003	0.6899/0.6585	0.6417/0.7051	0.7155/0.3377	0.7254/0.9629	0.4941/0.6153
U_{eig}	0.8205/0.7827	0.6981/0.8931	0.6977/0.6602	0.6526/0.7101	0.7100/0.3393	0.8025/0.9615	0.7139/0.7918
U_{deg}	0.8223/0.7842	0.6976/0.8931	0.6960/0.6482	0.6520/0.7098	0.7177/0.3450	0.8037/0.9573	0.7369/0.8028
C_{ecc}	0.5604/0.6221	0.6543/0.8852	0.7392/0.6661	0.6135/0.6894	0.5816/0.2680	0.7231/0.9507	0.6500/0.7553
LUQ-pair	0.7776/0.7521	0.6504/0.8947	0.7384/0.6916	0.6269/0.6967	0.7034/0.3240	0.7518/0.9520	0.8258/0.8467
SNNE	0.8146/0.7774	0.6980/0.8931	0.7027/0.6524	0.6526/0.7063	0.7059/0.3391	0.8061/0.9576	0.7434/0.8055
SNNE (only_num)	0.8152/0.7805	0.6979/0.8931	0.7016/0.6516	0.6522/0.7062	0.7064/0.3395	0.8058/0.9559	0.7421/0.8050
SNNE (only_denom)	0.8220/0.7824	0.6976/0.8931	0.7010/0.6609	0.6522/0.7061	0.7193/0.3407	0.8062/0.9576	0.7407/0.8044

Table 42: Thinking comparison with QWEN3 8B. We compute the semantic uncertainty exhibited in the QWEN3 8B outputs under both thinking-enabled and thinking-disabled settings. Panel: Summarization.

Model / method	XSUM	AESLC	CNN/DM	Multi-News
Enable Thinking				
Discrete semantic entropy	0.2661	0.2004	0.3640	0.2911
Discrete semantic density	0.2668	0.2136	0.3632	0.2910
KLE-heat$_{t=0.1}$	0.2800	0.2265	0.3763	0.3063
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2799	0.2264	0.3762	0.3061
U_{eig}	0.2687	0.2245	0.3710	0.2933
U_{deg}	0.2680	0.2234	0.3707	0.2927
LUQ-pair	0.2681	0.2201	0.3700	0.2926
SNNE	0.2678	0.2231	0.3706	0.2927
SNNE (only_num)	0.2679	0.2231	0.3706	0.2927
SNNE (only_denom)	0.2679	0.2231	0.3707	0.2927
Disable Thinking				
Discrete semantic entropy	0.2566	0.2378	0.3312	0.2968
Discrete semantic density	0.2530	0.1875	0.3156	0.2913
KLE-heat$_{t=0.1}$	0.2540	0.2375	0.3276	0.2913
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2537	0.2403	0.3272	0.2909
U_{eig}	0.2549	0.2379	0.3370	0.3014
U_{deg}	0.2553	0.2371	0.3361	0.3007
C_{ecc}	0.2529	0.2280	0.3210	0.2962
LUQ-pair	0.2575	0.2403	0.3372	0.3025
SNNE	0.2556	0.2389	0.3365	0.3010
SNNE (only_num)	0.2555	0.2390	0.3365	0.3010
SNNE (only_denom)	0.2556	0.2384	0.3364	0.3009

Table 42: Thinking comparison with QWEN3 8B. We compute the semantic uncertainty exhibited in the QWEN3 8B outputs under both thinking-enabled and thinking-disabled settings. Panel: Translation.

Model / method	cs	de	fi	gu	kk	lt	ru	zh
Enable Thinking								
Discrete semantic entropy	0.5410	0.6760	0.4762	0.4857	0.4011	0.5047	0.5938	0.5236
Discrete semantic density	0.5398	0.6819	0.4813	0.5224	0.4107	0.5140	0.5973	0.5202
KLE-heat$_{t=0.1}$	0.5803	0.7119	0.5168	0.5974	0.4427	0.5444	0.6208	0.5431
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5779	0.7122	0.5163	0.5968	0.4425	0.5450	0.6208	0.5426
U_{eig}	0.5773	0.7126	0.5158	0.5876	0.4407	0.5414	0.6195	0.5416
U_{deg}	0.5767	0.7123	0.5158	0.5877	0.4391	0.5425	0.6199	0.5406
LUQ-pair	0.5770	0.7107	0.5112	0.6013	0.4459	0.5440	0.6172	0.5389
SNNE	0.5760	0.7126	0.5146	0.5893	0.4407	0.5422	0.6200	0.5410
SNNE (only_num)	0.5760	0.7126	0.5150	0.5884	0.4398	0.5434	0.6205	0.5409
SNNE (only_denom)	0.5760	0.7126	0.5149	0.5878	0.4391	0.5424	0.6192	0.5413
Disable Thinking								
Discrete semantic entropy	0.5559	0.6890	0.5052	0.6093	0.4321	0.5547	0.6075	0.5380
Discrete semantic density	0.5310	0.6328	0.4557	0.5254	0.3857	0.5032	0.5765	0.5033
KLE-heat$_{t=0.1}$	0.5593	0.6827	0.4985	0.6121	0.4378	0.5467	0.6100	0.5368
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.5584	0.6821	0.5069	0.6110	0.4391	0.5405	0.6092	0.5377
U_{eig}	0.5850	0.7114	0.5206	0.6325	0.4569	0.5619	0.6216	0.5518
U_{deg}	0.5821	0.7100	0.5184	0.6326	0.4563	0.5631	0.6237	0.5514
C_{ecc}	0.5824	0.7064	0.5132	0.6244	0.4557	0.5580	0.6229	0.5492
LUQ-pair	0.5819	0.7009	0.5181	0.6238	0.4553	0.5605	0.6184	0.5504
SNNE	0.5832	0.7100	0.5196	0.6321	0.4561	0.5630	0.6214	0.5515
SNNE (only_num)	0.5832	0.7100	0.5196	0.6322	0.4561	0.5626	0.6214	0.5514
SNNE (only_denom)	0.5833	0.7100	0.5202	0.6308	0.4573	0.5626	0.6216	0.5514

H.9 Complete Multilingual Translation Results

Discussion: Appendix D.

Table 43: Per-direction multilingual translation performance for the seven Llama-3.1, Qwen3, and Phi-3 models. We report ROUGE-L; Chinese and Japanese targets use character tokenization and all other targets use whitespace tokenization. The final column in each panel is sample-weighted.

OPUS-100: non-English→non-English

Model	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Overall
Llama 3.1-8B Instruct	32.19	37.51	31.39	40.34	26.03	32.04	33.25
Llama 3.1-70B Instruct	37.25	45.38	45.43	45.26	36.05	45.80	42.53
Llama 3.1-405B Instruct FP8	39.84	47.50	50.23	46.19	38.02	50.45	45.37
Qwen 3-1.7B	27.59	33.72	42.32	35.68	32.72	44.75	36.13
Qwen 3-8B	35.88	42.25	49.35	42.12	36.98	50.34	42.82
Qwen 3-32B	38.17	44.71	51.22	43.67	38.59	51.64	44.67
Phi-3-medium-128k-instruct (14B)	29.16	34.99	36.48	41.41	34.10	44.57	36.79

WMT24++: English→multilingual

Model	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Overall
Llama 3.1-8B Instruct	10.33	44.20	55.58	51.84	27.27	19.63	33.12	43.75	35.71
Llama 3.1-70B Instruct	19.63	51.92	62.68	58.92	37.06	45.44	40.07	55.48	46.40
Llama 3.1-405B Instruct FP8	23.90	54.06	64.46	61.35	37.40	49.34	42.49	59.82	49.10
Qwen 3-1.7B	12.94	41.04	51.11	47.05	16.86	37.84	31.06	53.13	36.38
Qwen 3-8B	18.74	49.72	59.90	55.94	29.76	46.16	37.29	59.63	44.64
Qwen 3-32B	19.23	50.71	60.08	57.12	32.32	48.16	38.50	61.48	45.95
Phi-3-medium-128k-instruct (14B)	12.91	46.43	57.96	54.59	21.55	42.42	35.75	47.43	39.88

Table 44: Per-direction multilingual PRR for LLAMA 3.1-8B INSTRUCT. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.3499	0.4017	0.3516	0.4087	0.2918	0.3769	0.3634
Discrete semantic density	0.3707	0.4098	0.3616	0.4323	0.3012	0.3815	0.3762
Discrete U_{eig}	0.3673	0.4040	0.3548	0.4198	0.3024	0.3750	0.3706
Discrete U_{deg}	0.3639	0.4029	0.3584	0.4177	0.3061	0.3771	0.3710
Discrete C_{ecc}	0.3695	0.3993	0.3460	0.4304	0.3063	0.3744	0.3710
KLE-heat$_{t=0.1}$	0.3627	0.4003	0.3586	0.4220	0.3024	0.3786	0.3708
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.3647	0.4023	0.3580	0.4192	0.3037	0.3796	0.3712
Discrete LUQ-pair	0.3275	0.3804	0.3387	0.4059	0.2702	0.3325	0.3425
Discrete SNNE	0.3485	0.3972	0.3524	0.4058	0.2931	0.3616	0.3598
Discrete SNNE (only-num)	0.3947	0.4378	0.3695	0.4545	0.3104	0.3867	0.3922
Discrete SNNE (only-denom)	0.4010	0.4501	0.3600	0.4646	0.3082	0.3637	0.3913
Semantic density	0.3734	0.4154	0.3556	0.4339	0.3005	0.3769	0.3760
U_{eig}	0.3763	0.4197	0.3582	0.4383	0.3033	0.3789	0.3791
U_{deg}	0.3766	0.4173	0.3575	0.4304	0.3017	0.3784	0.3770
U_{ecc}	0.2719	0.3456	0.2849	0.3886	0.2198	0.2785	0.2982
LUQ	0.3747	0.4005	0.3573	0.4326	0.3066	0.3809	0.3754
SNNE	0.3754	0.4175	0.3576	0.4310	0.3004	0.3781	0.3767
SNNE (only-num)	0.3761	0.4168	0.3568	0.4331	0.3002	0.3779	0.3768
SNNE (only-denom)	0.3794	0.4179	0.3575	0.4352	0.3013	0.3785	0.3783

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.1122	0.4591	0.5474	0.5179	0.2974	0.2436	0.3332	0.4526	0.3704
Discrete semantic density	0.1163	0.4598	0.5613	0.5227	0.3029	0.2537	0.3452	0.4561	0.3772
Discrete U_{eig}	0.1174	0.4616	0.5605	0.5136	0.2988	0.2528	0.3428	0.4588	0.3758
Discrete U_{deg}	0.1167	0.4604	0.5600	0.5122	0.3017	0.2544	0.3421	0.4549	0.3753
Discrete C_{ecc}	0.1158	0.4640	0.5655	0.5089	0.3023	0.2374	0.3418	0.4650	0.3751
KLE-heat$_{t=0.1}$	0.1174	0.4676	0.5587	0.5099	0.2994	0.2561	0.3387	0.4555	0.3754
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1171	0.4648	0.5587	0.5169	0.3005	0.2548	0.3392	0.4565	0.3760
Discrete LUQ-pair	0.0976	0.4392	0.5351	0.5200	0.2782	0.1982	0.3192	0.4378	0.3532
Discrete SNNE	0.1086	0.4558	0.5480	0.5146	0.2868	0.2504	0.3212	0.4331	0.3648
Discrete SNNE (only-num)	0.1203	0.4944	0.5921	0.5469	0.3130	0.2432	0.3680	0.4594	0.3922
Discrete SNNE (only-denom)	0.1572	0.5288	0.6340	0.5861	0.3323	0.1681	0.4065	0.4667	0.4100
Semantic density	0.1221	0.4666	0.5642	0.5133	0.2989	0.2646	0.3512	0.4546	0.3795
U_{eig}	0.1223	0.4707	0.5622	0.5157	0.2984	0.2669	0.3534	0.4563	0.3807
U_{deg}	0.1238	0.4672	0.5633	0.5127	0.2983	0.2696	0.3508	0.4548	0.3801
U_{ecc}	0.1046	0.4422	0.5681	0.5318	0.2586	0.1301	0.3434	0.4439	0.3528
LUQ	0.1232	0.4579	0.5619	0.5229	0.2958	0.2658	0.3462	0.4579	0.3790
SNNE	0.1228	0.4679	0.5636	0.5137	0.2975	0.2697	0.3516	0.4549	0.3802
SNNE (only-num)	0.1231	0.4668	0.5633	0.5128	0.2975	0.2697	0.3508	0.4546	0.3798
SNNE (only-denom)	0.1243	0.4664	0.5628	0.5141	0.2984	0.2698	0.3502	0.4546	0.3801

Table 45: Per-direction multilingual PRR for LLAMA 3.1-70B INSTRUCT. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.4074	0.4781	0.4759	0.4578	0.3705	0.4742	0.4440
Discrete semantic density	0.4285	0.4874	0.4803	0.4734	0.3816	0.4849	0.4560
Discrete U_{eig}	0.4285	0.4933	0.4836	0.4664	0.3773	0.4814	0.4551
Discrete U_{deg}	0.4195	0.4864	0.4753	0.4646	0.3772	0.4855	0.4514
Discrete C_{ecc}	0.4306	0.4887	0.4826	0.4601	0.3766	0.4759	0.4524
KLE-heat$_{t=0.1}$	0.4214	0.4856	0.4779	0.4647	0.3764	0.4825	0.4514
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.4202	0.4859	0.4830	0.4647	0.3732	0.4819	0.4515
Discrete LUQ-pair	0.3810	0.4528	0.4605	0.4431	0.3647	0.4556	0.4263
Discrete SNNE	0.3997	0.4697	0.4671	0.4549	0.3740	0.4617	0.4378
Discrete SNNE (only-num)	0.4595	0.5167	0.4956	0.4991	0.3826	0.4941	0.4746
Discrete SNNE (only-denom)	0.4650	0.5272	0.4970	0.5080	0.3851	0.4877	0.4784
Semantic density	0.4388	0.4886	0.4828	0.4802	0.3698	0.4809	0.4569
U_{eig}	0.4396	0.4926	0.4852	0.4761	0.3745	0.4828	0.4585
U_{deg}	0.4379	0.4884	0.4822	0.4771	0.3695	0.4800	0.4559
U_{ecc}	0.3114	0.4242	0.4448	0.4417	0.3520	0.4445	0.4031
LUQ	0.4328	0.4898	0.4781	0.4780	0.3698	0.4708	0.4532
SNNE	0.4378	0.4908	0.4812	0.4736	0.3705	0.4802	0.4557
SNNE (only-num)	0.4372	0.4873	0.4814	0.4788	0.3713	0.4795	0.4559
SNNE (only-denom)	0.4380	0.4896	0.4819	0.4788	0.3700	0.4801	0.4564

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.2148	0.5332	0.6285	0.5932	0.4007	0.4514	0.4049	0.5698	0.4746
Discrete semantic density	0.2182	0.5515	0.6346	0.6061	0.4064	0.4550	0.4022	0.5772	0.4814
Discrete U_{eig}	0.2174	0.5473	0.6357	0.5931	0.4056	0.4545	0.3993	0.5740	0.4783
Discrete U_{deg}	0.2161	0.5448	0.6396	0.6058	0.4105	0.4542	0.3965	0.5721	0.4800
Discrete C_{ecc}	0.2184	0.5447	0.6419	0.5963	0.4045	0.4471	0.4103	0.5672	0.4788
KLE-heat$_{t=0.1}$	0.2180	0.5430	0.6370	0.6045	0.4098	0.4555	0.3958	0.5705	0.4793
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2175	0.5424	0.6372	0.6007	0.4035	0.4515	0.3966	0.5707	0.4775
Discrete LUQ-pair	0.1960	0.5234	0.6068	0.5844	0.3722	0.4398	0.3811	0.5448	0.4561
Discrete SNNE	0.2035	0.5325	0.6206	0.5977	0.3994	0.4494	0.3923	0.5558	0.4689
Discrete SNNE (only-num)	0.2355	0.5750	0.6705	0.6408	0.4244	0.4576	0.4313	0.5857	0.5026
Discrete SNNE (only-denom)	0.2627	0.5962	0.6978	0.6639	0.4366	0.4646	0.4797	0.5912	0.5241
Semantic density	0.2211	0.5494	0.6423	0.6124	0.4030	0.4582	0.4125	0.5655	0.4831
U_{eig}	0.2221	0.5503	0.6429	0.6140	0.4037	0.4556	0.4131	0.5659	0.4834
U_{deg}	0.2215	0.5472	0.6412	0.6128	0.4008	0.4577	0.4115	0.5640	0.4821
U_{ecc}	0.1831	0.5138	0.6344	0.5926	0.3577	0.4685	0.4214	0.5565	0.4660
LUQ	0.2186	0.5469	0.6384	0.6069	0.3995	0.4583	0.4059	0.5619	0.4796
SNNE	0.2206	0.5483	0.6410	0.6113	0.4005	0.4575	0.4117	0.5636	0.4818
SNNE (only-num)	0.2207	0.5478	0.6431	0.6114	0.4021	0.4582	0.4123	0.5637	0.4824
SNNE (only-denom)	0.2216	0.5489	0.6424	0.6125	0.4015	0.4583	0.4125	0.5636	0.4826

Table 46: Per-direction multilingual PRR for LLAMA 3.1-405B INSTRUCT FP8. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.4321	0.4862	0.5231	0.4679	0.3902	0.5052	0.4675
Discrete semantic density	0.4396	0.5063	0.5247	0.4811	0.3963	0.5248	0.4788
Discrete U_{eig}	0.4455	0.4948	0.5199	0.4759	0.3951	0.5237	0.4758
Discrete U_{deg}	0.4444	0.4937	0.5231	0.4853	0.3907	0.5249	0.4770
Discrete C_{ecc}	0.4372	0.4971	0.5195	0.4792	0.3961	0.5222	0.4752
KLE-heat$_{t=0.1}$	0.4442	0.4966	0.5230	0.4749	0.3911	0.5185	0.4747
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.4448	0.4945	0.5215	0.4734	0.3932	0.5238	0.4752
Discrete LUQ-pair	0.4053	0.4850	0.5052	0.4580	0.3867	0.5014	0.4569
Discrete SNNE	0.4189	0.4817	0.5069	0.4578	0.3902	0.5096	0.4608
Discrete SNNE (only-num)	0.4678	0.5287	0.5346	0.5141	0.4034	0.5236	0.4954
Discrete SNNE (only-denom)	0.4770	0.5388	0.5346	0.5203	0.4058	0.5244	0.5001
Semantic density	0.4567	0.4997	0.5234	0.5017	0.3860	0.5119	0.4799
U_{eig}	0.4584	0.5014	0.5251	0.5029	0.3899	0.5132	0.4818
U_{deg}	0.4538	0.5009	0.5235	0.4998	0.3888	0.5116	0.4797
U_{ecc}	0.3389	0.4559	0.4934	0.4391	0.3773	0.5015	0.4344
LUQ	0.4639	0.5007	0.5201	0.4931	0.3925	0.5076	0.4796
SNNE	0.4562	0.4960	0.5217	0.5028	0.3869	0.5092	0.4788
SNNE (only-num)	0.4536	0.4999	0.5236	0.4993	0.3878	0.5105	0.4791
SNNE (only-denom)	0.4561	0.5011	0.5232	0.4924	0.3862	0.5102	0.4782

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.2674	0.5403	0.6360	0.6241	0.3976	0.4842	0.4168	0.5900	0.4946
Discrete semantic density	0.2719	0.5558	0.6442	0.6289	0.3982	0.4957	0.4386	0.5930	0.5033
Discrete U_{eig}	0.2713	0.5433	0.6436	0.6213	0.3965	0.4863	0.4224	0.5909	0.4969
Discrete U_{deg}	0.2731	0.5461	0.6443	0.6321	0.3934	0.4993	0.4179	0.5942	0.5001
Discrete C_{ecc}	0.2700	0.5622	0.6528	0.6329	0.3886	0.5001	0.4201	0.6076	0.5043
KLE-heat$_{t=0.1}$	0.2725	0.5556	0.6347	0.6179	0.3916	0.4905	0.4202	0.5984	0.4977
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2707	0.5554	0.6463	0.6261	0.3947	0.4919	0.4235	0.6002	0.5011
Discrete LUQ-pair	0.2356	0.5324	0.6275	0.6057	0.3900	0.4819	0.4016	0.5869	0.4827
Discrete SNNE	0.2611	0.5449	0.6268	0.6137	0.3868	0.4849	0.4173	0.5850	0.4901
Discrete SNNE (only-num)	0.2769	0.5814	0.6765	0.6540	0.4220	0.4948	0.4524	0.6085	0.5208
Discrete SNNE (only-denom)	0.2820	0.6027	0.7074	0.6746	0.4354	0.4977	0.4879	0.6223	0.5387
Semantic density	0.2674	0.5638	0.6479	0.6329	0.3971	0.4945	0.4308	0.5900	0.5031
U_{eig}	0.2703	0.5662	0.6445	0.6355	0.3988	0.4934	0.4297	0.5909	0.5037
U_{deg}	0.2676	0.5628	0.6471	0.6332	0.3977	0.4936	0.4302	0.5904	0.5028
U_{ecc}	0.2181	0.5407	0.6581	0.6174	0.3650	0.5145	0.4448	0.6181	0.4971
LUQ	0.2675	0.5620	0.6376	0.6271	0.3939	0.4896	0.4376	0.5960	0.5014
SNNE	0.2676	0.5629	0.6440	0.6326	0.3973	0.4918	0.4300	0.5901	0.5020
SNNE (only-num)	0.2677	0.5645	0.6444	0.6347	0.3962	0.4925	0.4295	0.5892	0.5023
SNNE (only-denom)	0.2676	0.5614	0.6463	0.6304	0.3971	0.4940	0.4267	0.5887	0.5015

Table 47: Per-direction multilingual PRR for QWEN 3-1.7B. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.3224	0.3633	0.4399	0.3765	0.3339	0.4606	0.3828
Discrete semantic density	0.3293	0.3765	0.4449	0.3957	0.3396	0.4731	0.3932
Discrete U_{eig}	0.3278	0.3648	0.4445	0.3873	0.3401	0.4634	0.3880
Discrete U_{deg}	0.3213	0.3647	0.4462	0.3850	0.3415	0.4639	0.3871
Discrete C_{ecc}	0.3300	0.3651	0.4417	0.3853	0.3373	0.4652	0.3874
KLE-heat$_{t=0.1}$	0.3257	0.3658	0.4461	0.3848	0.3398	0.4692	0.3886
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.3246	0.3654	0.4436	0.3857	0.3378	0.4636	0.3868
Discrete LUQ-pair	0.2866	0.3398	0.4249	0.3581	0.3276	0.4519	0.3648
Discrete SNNE	0.3213	0.3580	0.4361	0.3693	0.3326	0.4506	0.3780
Discrete SNNE (only-num)	0.3438	0.3912	0.4556	0.4123	0.3434	0.4750	0.4035
Discrete SNNE (only-denom)	0.3486	0.4095	0.4604	0.4234	0.3460	0.4756	0.4106
Semantic density	0.3338	0.3869	0.4513	0.3990	0.3402	0.4701	0.3969
U_{eig}	0.3360	0.3878	0.4519	0.3989	0.3431	0.4712	0.3981
U_{deg}	0.3319	0.3896	0.4497	0.3981	0.3392	0.4695	0.3963
U_{ecc}	0.2256	0.3146	0.4149	0.3329	0.3300	0.4378	0.3426
LUQ	0.3336	0.3857	0.4489	0.3983	0.3329	0.4713	0.3951
SNNE	0.3308	0.3888	0.4529	0.3991	0.3434	0.4699	0.3975
SNNE (only-num)	0.3376	0.3825	0.4489	0.3978	0.3406	0.4686	0.3960
SNNE (only-denom)	0.3340	0.3864	0.4500	0.4002	0.3410	0.4685	0.3967

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.1496	0.4448	0.5164	0.4811	0.2153	0.3756	0.3299	0.5228	0.3794
Discrete semantic density	0.1479	0.4379	0.5257	0.4918	0.2224	0.3785	0.3375	0.5373	0.3849
Discrete U_{eig}	0.1441	0.4506	0.5160	0.4885	0.2187	0.3802	0.3252	0.5317	0.3819
Discrete U_{deg}	0.1484	0.4495	0.5152	0.4888	0.2201	0.3776	0.3278	0.5322	0.3825
Discrete C_{ecc}	0.1495	0.4432	0.5236	0.4919	0.2104	0.3750	0.3363	0.5329	0.3829
KLE-heat$_{t=0.1}$	0.1467	0.4434	0.5139	0.4924	0.2215	0.3777	0.3245	0.5291	0.3812
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1474	0.4407	0.5221	0.4942	0.2193	0.3775	0.3326	0.5290	0.3829
Discrete LUQ-pair	0.1313	0.4177	0.5082	0.4704	0.1802	0.3697	0.3062	0.5214	0.3631
Discrete SNNE	0.1390	0.4259	0.5137	0.4829	0.2091	0.3886	0.3084	0.5237	0.3739
Discrete SNNE (only-num)	0.1567	0.4616	0.5392	0.5103	0.2274	0.3843	0.3492	0.5456	0.3968
Discrete SNNE (only-denom)	0.1799	0.4938	0.5779	0.5287	0.2506	0.3858	0.3818	0.5533	0.4190
Semantic density	0.1490	0.4521	0.5201	0.4938	0.2199	0.3794	0.3354	0.5322	0.3852
U_{eig}	0.1473	0.4505	0.5236	0.4950	0.2223	0.3815	0.3371	0.5320	0.3862
U_{deg}	0.1493	0.4536	0.5210	0.4940	0.2230	0.3807	0.3363	0.5303	0.3860
U_{ecc}	0.1259	0.3985	0.5150	0.4643	0.1421	0.3864	0.3212	0.5464	0.3625
LUQ	0.1504	0.4514	0.5204	0.4899	0.2184	0.3801	0.3342	0.5285	0.3842
SNNE	0.1486	0.4544	0.5225	0.4935	0.2218	0.3821	0.3340	0.5295	0.3858
SNNE (only-num)	0.1492	0.4531	0.5242	0.4907	0.2217	0.3804	0.3362	0.5291	0.3856
SNNE (only-denom)	0.1504	0.4555	0.5203	0.4908	0.2241	0.3801	0.3324	0.5300	0.3855

Table 48: Per-direction multilingual PRR for QWEN 3-8B. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.3963	0.4471	0.4977	0.4451	0.3817	0.5031	0.4452
Discrete semantic density	0.3960	0.4530	0.5010	0.4388	0.3694	0.5078	0.4443
Discrete U_{eig}	0.3952	0.4614	0.4989	0.4355	0.3824	0.5053	0.4464
Discrete U_{deg}	0.3967	0.4542	0.5004	0.4377	0.3790	0.5047	0.4455
Discrete C_{ecc}	0.3904	0.4476	0.4970	0.4317	0.3821	0.5053	0.4424
KLE-heat$_{t=0.1}$	0.3912	0.4538	0.4993	0.4375	0.3776	0.4999	0.4432
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.3852	0.4564	0.4994	0.4358	0.3790	0.5041	0.4433
Discrete LUQ-pair	0.3633	0.4271	0.4967	0.4114	0.3768	0.5119	0.4312
Discrete SNNE	0.3845	0.4364	0.4909	0.4208	0.3758	0.5065	0.4358
Discrete SNNE (only-num)	0.4304	0.4772	0.5153	0.4675	0.3826	0.5358	0.4681
Discrete SNNE (only-denom)	0.4436	0.4831	0.5190	0.4737	0.3888	0.5403	0.4747
Semantic density	0.4177	0.4662	0.5062	0.4534	0.3777	0.5236	0.4575
U_{eig}	0.4246	0.4657	0.5051	0.4562	0.3812	0.5246	0.4596
U_{deg}	0.4156	0.4631	0.5093	0.4534	0.3783	0.5239	0.4572
U_{ecc}	0.3035	0.3971	0.4908	0.4014	0.3763	0.5033	0.4121
LUQ	0.4328	0.4554	0.5043	0.4417	0.3801	0.5198	0.4557
SNNE	0.4175	0.4688	0.5074	0.4550	0.3833	0.5240	0.4593
SNNE (only-num)	0.4201	0.4646	0.5130	0.4564	0.3787	0.5238	0.4594
SNNE (only-denom)	0.4192	0.4647	0.5097	0.4565	0.3836	0.5219	0.4593

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.2071	0.5131	0.6155	0.5701	0.3383	0.4546	0.3722	0.5951	0.4583
Discrete semantic density	0.2183	0.5146	0.6087	0.5787	0.3410	0.4682	0.3868	0.6076	0.4655
Discrete U_{eig}	0.2146	0.5119	0.6038	0.5744	0.3454	0.4635	0.3760	0.5983	0.4610
Discrete U_{deg}	0.2142	0.5142	0.6045	0.5752	0.3402	0.4604	0.3754	0.5951	0.4599
Discrete C_{ecc}	0.2102	0.5214	0.6159	0.5842	0.3435	0.4679	0.3852	0.5954	0.4655
KLE-heat$_{t=0.1}$	0.2132	0.5186	0.6045	0.5767	0.3435	0.4630	0.3762	0.6031	0.4624
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2137	0.5134	0.6088	0.5745	0.3429	0.4623	0.3899	0.5943	0.4625
Discrete LUQ-pair	0.1877	0.5155	0.5871	0.5458	0.3135	0.4434	0.3561	0.5925	0.4427
Discrete SNNE	0.2050	0.5133	0.5983	0.5703	0.3291	0.4582	0.3781	0.5898	0.4553
Discrete SNNE (only-num)	0.2323	0.5482	0.6249	0.6073	0.3571	0.4641	0.4134	0.6091	0.4820
Discrete SNNE (only-denom)	0.2484	0.5719	0.6546	0.6260	0.3732	0.4662	0.4456	0.6155	0.5002
Semantic density	0.2161	0.5273	0.6100	0.5913	0.3408	0.4618	0.3955	0.6000	0.4679
U_{eig}	0.2156	0.5296	0.6069	0.5886	0.3411	0.4632	0.3961	0.6010	0.4678
U_{deg}	0.2149	0.5294	0.6049	0.5942	0.3414	0.4616	0.3969	0.5970	0.4675
U_{ecc}	0.1765	0.4915	0.6058	0.5553	0.2758	0.4802	0.3828	0.6110	0.4474
LUQ	0.2171	0.5242	0.6139	0.5785	0.3430	0.4596	0.3890	0.6030	0.4660
SNNE	0.2131	0.5294	0.6093	0.5923	0.3403	0.4617	0.3957	0.5985	0.4675
SNNE (only-num)	0.2128	0.5252	0.6067	0.5874	0.3417	0.4627	0.3952	0.5983	0.4662
SNNE (only-denom)	0.2128	0.5294	0.6086	0.5947	0.3416	0.4616	0.3922	0.5974	0.4673

Table 49: Per-direction multilingual PRR for QWEN 3-32B. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.4223	0.4749	0.5210	0.4456	0.3913	0.5285	0.4639
Discrete semantic density	0.4311	0.4823	0.5270	0.4607	0.4050	0.5346	0.4734
Discrete U_{eig}	0.4238	0.4830	0.5274	0.4537	0.3980	0.5311	0.4695
Discrete U_{deg}	0.4266	0.4842	0.5287	0.4534	0.3986	0.5336	0.4709
Discrete C_{ecc}	0.4199	0.4879	0.5270	0.4586	0.3968	0.5269	0.4695
KLE-heat$_{t=0.1}$	0.4244	0.4887	0.5266	0.4520	0.3996	0.5305	0.4703
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.4265	0.4854	0.5261	0.4536	0.3999	0.5271	0.4698
Discrete LUQ-pair	0.3928	0.4601	0.5142	0.4271	0.3930	0.5109	0.4497
Discrete SNNE	0.4117	0.4526	0.5182	0.4374	0.3949	0.5159	0.4551
Discrete SNNE (only-num)	0.4564	0.5128	0.5381	0.4921	0.4071	0.5508	0.4929
Discrete SNNE (only-denom)	0.4642	0.5221	0.5439	0.4959	0.4035	0.5479	0.4963
Semantic density	0.4464	0.4957	0.5254	0.4714	0.4026	0.5448	0.4810
U_{eig}	0.4490	0.5006	0.5255	0.4802	0.4054	0.5442	0.4842
U_{deg}	0.4457	0.4996	0.5289	0.4732	0.4037	0.5434	0.4824
U_{ecc}	0.3211	0.4140	0.5093	0.4130	0.3898	0.5122	0.4266
LUQ	0.4534	0.4872	0.5142	0.4626	0.3946	0.5354	0.4746
SNNE	0.4452	0.4976	0.5247	0.4708	0.4035	0.5405	0.4804
SNNE (only-num)	0.4448	0.5002	0.5266	0.4768	0.4041	0.5421	0.4824
SNNE (only-denom)	0.4441	0.4963	0.5275	0.4779	0.4047	0.5415	0.4820

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.2102	0.5299	0.5982	0.5884	0.3716	0.4838	0.3904	0.6221	0.4743
Discrete semantic density	0.2128	0.5413	0.6068	0.5983	0.3803	0.4946	0.4042	0.6243	0.4828
Discrete U_{eig}	0.2140	0.5368	0.6115	0.5893	0.3818	0.4910	0.4054	0.6229	0.4816
Discrete U_{deg}	0.2142	0.5314	0.6129	0.5895	0.3752	0.4921	0.4010	0.6237	0.4800
Discrete C_{ecc}	0.2152	0.5413	0.6120	0.5999	0.3717	0.4920	0.4003	0.6257	0.4823
KLE-heat$_{t=0.1}$	0.2128	0.5311	0.6062	0.5935	0.3798	0.4924	0.3995	0.6243	0.4800
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2147	0.5462	0.6115	0.5903	0.3736	0.4960	0.4020	0.6238	0.4823
Discrete LUQ-pair	0.1764	0.5215	0.5836	0.5654	0.3423	0.4861	0.3742	0.6149	0.4581
Discrete SNNE	0.1969	0.5211	0.5944	0.5785	0.3580	0.4835	0.3923	0.6210	0.4682
Discrete SNNE (only-num)	0.2254	0.5688	0.6398	0.6309	0.3911	0.5015	0.4301	0.6363	0.5030
Discrete SNNE (only-denom)	0.2715	0.5877	0.6669	0.6480	0.4109	0.4995	0.4656	0.6406	0.5238
Semantic density	0.2158	0.5387	0.6133	0.6051	0.3737	0.4961	0.4083	0.6189	0.4837
U_{eig}	0.2169	0.5412	0.6156	0.6083	0.3750	0.4970	0.4060	0.6169	0.4846
U_{deg}	0.2159	0.5377	0.6137	0.6067	0.3726	0.4951	0.4081	0.6155	0.4832
U_{ecc}	0.1809	0.4989	0.6041	0.5648	0.2954	0.4825	0.3967	0.6236	0.4559
LUQ	0.2129	0.5421	0.6089	0.6023	0.3720	0.4886	0.3924	0.6164	0.4794
SNNE	0.2136	0.5397	0.6101	0.6072	0.3717	0.4954	0.4085	0.6188	0.4831
SNNE (only-num)	0.2142	0.5365	0.6077	0.6068	0.3710	0.4950	0.4061	0.6192	0.4821
SNNE (only-denom)	0.2160	0.5403	0.6155	0.6072	0.3719	0.4949	0.4067	0.6186	0.4839

Table 50: Per-direction multilingual PRR for PHI-3-MEDIUM-128K-INSTRUCT (14B). The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.3420	0.3811	0.3872	0.4344	0.3589	0.4589	0.3937
Discrete semantic density	0.3526	0.3823	0.3906	0.4430	0.3671	0.4689	0.4007
Discrete U_{eig}	0.3479	0.3834	0.3923	0.4334	0.3664	0.4679	0.3986
Discrete U_{deg}	0.3532	0.3838	0.3901	0.4342	0.3619	0.4686	0.3986
Discrete C_{ecc}	0.3537	0.3851	0.3962	0.4458	0.3648	0.4600	0.4009
KLE-heat$_{t=0.1}$	0.3517	0.3826	0.3932	0.4425	0.3605	0.4669	0.3996
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.3502	0.3810	0.3906	0.4388	0.3613	0.4669	0.3981
Discrete LUQ-pair	0.3042	0.3514	0.3681	0.4077	0.3427	0.4491	0.3705
Discrete SNNE	0.3310	0.3654	0.3779	0.4163	0.3617	0.4584	0.3851
Discrete SNNE (only-num)	0.3779	0.3959	0.3971	0.4651	0.3659	0.4795	0.4136
Discrete SNNE (only-denom)	0.3910	0.4168	0.3979	0.4694	0.3653	0.4793	0.4199
Semantic density	0.3607	0.3810	0.3926	0.4438	0.3587	0.4666	0.4006
U_{eig}	0.3640	0.3855	0.3938	0.4467	0.3578	0.4694	0.4029
U_{deg}	0.3594	0.3824	0.3919	0.4432	0.3567	0.4666	0.4000
U_{ecc}	0.2391	0.3343	0.3503	0.3852	0.3249	0.4331	0.3445
LUQ	0.3513	0.3807	0.3891	0.4442	0.3634	0.4631	0.3986
SNNE	0.3593	0.3800	0.3909	0.4420	0.3556	0.4646	0.3987
SNNE (only-num)	0.3597	0.3823	0.3913	0.4418	0.3549	0.4641	0.3990
SNNE (only-denom)	0.3616	0.3824	0.3924	0.4409	0.3549	0.4674	0.3999

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.1416	0.4907	0.5828	0.5449	0.2413	0.4337	0.3679	0.4816	0.4106
Discrete semantic density	0.1388	0.4957	0.5901	0.5555	0.2540	0.4398	0.3799	0.4881	0.4177
Discrete U_{eig}	0.1323	0.4810	0.5899	0.5562	0.2504	0.4354	0.3799	0.4808	0.4132
Discrete U_{deg}	0.1357	0.4911	0.5857	0.5522	0.2508	0.4404	0.3743	0.4843	0.4143
Discrete C_{ecc}	0.1351	0.5001	0.5880	0.5647	0.2411	0.4449	0.3844	0.4895	0.4185
KLE-heat$_{t=0.1}$	0.1362	0.4836	0.5846	0.5577	0.2497	0.4392	0.3737	0.4826	0.4134
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1362	0.4840	0.5845	0.5565	0.2492	0.4352	0.3771	0.4833	0.4132
Discrete LUQ-pair	0.1235	0.4603	0.5686	0.5400	0.2249	0.4106	0.3495	0.4656	0.3929
Discrete SNNE	0.1262	0.4791	0.5791	0.5474	0.2306	0.4313	0.3566	0.4705	0.4026
Discrete SNNE (only-num)	0.1423	0.5164	0.6262	0.5920	0.2565	0.4381	0.3982	0.4911	0.4326
Discrete SNNE (only-denom)	0.1775	0.5466	0.6583	0.6215	0.2916	0.4255	0.4333	0.5052	0.4574
Semantic density	0.1381	0.5010	0.5938	0.5719	0.2527	0.4380	0.3833	0.4817	0.4201
U_{eig}	0.1352	0.5027	0.5934	0.5718	0.2527	0.4369	0.3854	0.4826	0.4201
U_{deg}	0.1389	0.5021	0.5919	0.5693	0.2527	0.4364	0.3819	0.4808	0.4192
U_{ecc}	0.1332	0.4591	0.5787	0.5435	0.1999	0.4289	0.3613	0.4862	0.3988
LUQ	0.1436	0.4989	0.5916	0.5696	0.2504	0.4376	0.3786	0.4849	0.4194
SNNE	0.1383	0.4993	0.5939	0.5693	0.2526	0.4365	0.3809	0.4808	0.4190
SNNE (only-num)	0.1388	0.5005	0.5920	0.5682	0.2521	0.4374	0.3810	0.4808	0.4188
SNNE (only-denom)	0.1397	0.4999	0.5925	0.5710	0.2523	0.4360	0.3826	0.4808	0.4193

Table 51: Per-direction multilingual translation performance for the five Mistral and Phi-4 models. We report ROUGE-L; Chinese and Japanese targets use character tokenization and all other targets use whitespace tokenization. The final column in each panel is sample-weighted.

OPUS-100: non-English→non-English

Model	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Overall
MISTRAL-7B-INSTRUCT-V0.3	21.57	29.10	20.60	36.91	23.62	25.00	26.13
MISTRAL-NEMO-INSTRUCT-2407 12B	34.42	40.96	39.92	42.49	33.76	42.74	39.05
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	38.68	44.92	47.85	44.90	37.30	48.52	43.69
PHI-4-MINI-INSTRUCT 3.8B	27.36	33.16	36.81	38.21	31.34	40.94	34.64
PHI-4 14B	35.97	43.42	46.74	43.77	36.53	48.25	42.45

WMT24++: English→multilingual

Model	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Overall
MISTRAL-7B-INSTRUCT-V0.3	7.77	39.04	48.72	46.29	13.44	24.99	29.95	35.14	30.67
MISTRAL-NEMO-INSTRUCT-2407 12B	7.35	47.99	58.61	56.81	22.14	36.66	38.11	48.44	39.51
MISTRAL-SMALL-3.2-24B-INSTRUCT-2506	13.46	53.53	63.32	61.12	32.58	47.13	41.64	57.74	46.31
PHI-4-MINI-INSTRUCT 3.8B	11.65	42.26	52.73	49.76	21.42	38.63	32.51	44.76	36.71
PHI-4 14B	18.80	50.49	60.70	57.56	33.15	48.60	37.82	55.57	45.34

Table 52: Per-direction multilingual PRR for MISTRAL-7B-INSTRUCT-V0.3. PRR is evaluated against per-example ROUGE-L. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.2631	0.3239	0.2336	0.3957	0.2787	0.2790	0.2957
Discrete semantic density	0.2738	0.3233	0.2313	0.4052	0.2851	0.2850	0.3006
Discrete U_{eig}	0.2703	0.3158	0.2271	0.4053	0.2876	0.2832	0.2982
Discrete U_{deg}	0.2757	0.3206	0.2302	0.4015	0.2846	0.2841	0.2995
Discrete C_{ecc}	0.2730	0.3217	0.2246	0.3986	0.2815	0.2839	0.2972
KLE-heat$_{t=0.1}$	0.2680	0.3203	0.2322	0.4061	0.2850	0.2843	0.2993
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2704	0.3194	0.2303	0.4032	0.2888	0.2841	0.2994
Discrete LUQ-pair	0.2213	0.2926	0.2125	0.3682	0.2433	0.2559	0.2656
Discrete SNNE	0.2539	0.3165	0.2302	0.3877	0.2712	0.2793	0.2898
Discrete SNNE (only-num)	0.2903	0.3367	0.2342	0.4219	0.2865	0.2851	0.3091
Discrete SNNE (only-denom)	0.3042	0.3627	0.2218	0.4304	0.2648	0.2548	0.3064
Semantic density	0.2770	0.3228	0.2318	0.4023	0.2832	0.2849	0.3003
U_{eig}	0.2781	0.3242	0.2252	0.4081	0.2833	0.2822	0.3002
U_{deg}	0.2765	0.3234	0.2285	0.4037	0.2831	0.2831	0.2997
U_{ecc}	0.1595	0.2678	0.2062	0.3369	0.2007	0.2306	0.2336
LUQ	0.2778	0.3257	0.2339	0.4019	0.2853	0.2820	0.3011
SNNE	0.2770	0.3232	0.2285	0.4035	0.2828	0.2827	0.2996
SNNE (only-num)	0.2777	0.3224	0.2286	0.4052	0.2821	0.2830	0.2999
SNNE (only-denom)	0.2785	0.3248	0.2297	0.4054	0.2821	0.2835	0.3006

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.1219	0.4109	0.5205	0.4819	0.1381	0.2592	0.3109	0.3834	0.3283
Discrete semantic density	0.1203	0.4211	0.5138	0.4864	0.1390	0.2604	0.3180	0.3854	0.3305
Discrete U_{eig}	0.1164	0.4143	0.5129	0.4757	0.1353	0.2612	0.3154	0.3811	0.3265
Discrete U_{deg}	0.1198	0.4188	0.5100	0.4744	0.1376	0.2611	0.3191	0.3830	0.3280
Discrete C_{ecc}	0.1083	0.4189	0.5016	0.4822	0.1416	0.2616	0.3217	0.3842	0.3275
KLE-heat$_{t=0.1}$	0.1214	0.4187	0.5131	0.4756	0.1379	0.2615	0.3173	0.3839	0.3287
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1191	0.4169	0.5168	0.4816	0.1374	0.2612	0.3189	0.3821	0.3293
Discrete LUQ-pair	0.1226	0.3976	0.4651	0.4418	0.1256	0.2489	0.2876	0.3551	0.3055
Discrete SNNE	0.1146	0.4019	0.5105	0.4644	0.1301	0.2665	0.2979	0.3745	0.3201
Discrete SNNE (only-num)	0.1215	0.4341	0.5363	0.5059	0.1453	0.2671	0.3320	0.3819	0.3405
Discrete SNNE (only-denom)	0.1687	0.4686	0.5749	0.5417	0.1757	0.2573	0.3673	0.3661	0.3651
Semantic density	0.1197	0.4174	0.5131	0.4841	0.1360	0.2656	0.3102	0.3767	0.3278
U_{eig}	0.1171	0.4204	0.5081	0.4824	0.1350	0.2636	0.3100	0.3755	0.3265
U_{deg}	0.1204	0.4175	0.5121	0.4829	0.1350	0.2657	0.3090	0.3752	0.3272
U_{ecc}	0.1206	0.3828	0.4732	0.4589	0.1428	0.2507	0.3023	0.3464	0.3097
LUQ	0.1238	0.4193	0.5069	0.4795	0.1357	0.2690	0.3082	0.3792	0.3277
SNNE	0.1207	0.4181	0.5098	0.4823	0.1355	0.2659	0.3083	0.3741	0.3268
SNNE (only-num)	0.1207	0.4185	0.5105	0.4819	0.1354	0.2660	0.3096	0.3740	0.3271
SNNE (only-denom)	0.1212	0.4170	0.5114	0.4828	0.1354	0.2665	0.3078	0.3754	0.3272

Table 53: Per-direction multilingual PRR for MISTRAL-NEMO-INSTRUCT-2407 12B. PRR is evaluated against per-example ROUGE-L. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.3891	0.4323	0.4253	0.4311	0.3526	0.4349	0.4109
Discrete semantic density	0.4032	0.4323	0.4248	0.4580	0.3569	0.4424	0.4196
Discrete U_{eig}	0.3953	0.4269	0.4200	0.4387	0.3591	0.4373	0.4129
Discrete U_{deg}	0.3987	0.4321	0.4287	0.4385	0.3613	0.4384	0.4163
Discrete C_{ecc}	0.3980	0.4362	0.4195	0.4459	0.3616	0.4399	0.4169
KLE-heat$_{t=0.1}$	0.3975	0.4292	0.4253	0.4385	0.3552	0.4362	0.4136
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.3980	0.4335	0.4237	0.4385	0.3586	0.4414	0.4156
Discrete LUQ-pair	0.3557	0.4124	0.4033	0.4179	0.3429	0.4315	0.3939
Discrete SNNE	0.3713	0.4170	0.4205	0.4200	0.3498	0.4362	0.4024
Discrete SNNE (only-num)	0.4327	0.4573	0.4235	0.4760	0.3756	0.4413	0.4344
Discrete SNNE (only-denom)	0.4341	0.4657	0.4175	0.4807	0.3758	0.4396	0.4356
Semantic density	0.4131	0.4364	0.4199	0.4475	0.3564	0.4407	0.4190
U_{eig}	0.4128	0.4378	0.4203	0.4547	0.3590	0.4408	0.4209
U_{deg}	0.4118	0.4398	0.4197	0.4511	0.3572	0.4396	0.4199
U_{ecc}	0.2794	0.3922	0.3810	0.4034	0.3191	0.4192	0.3657
LUQ	0.4128	0.4252	0.4253	0.4499	0.3613	0.4415	0.4193
SNNE	0.4110	0.4387	0.4184	0.4524	0.3579	0.4393	0.4196
SNNE (only-num)	0.4177	0.4364	0.4190	0.4520	0.3575	0.4397	0.4204
SNNE (only-denom)	0.4116	0.4384	0.4195	0.4534	0.3567	0.4410	0.4201

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.1048	0.4919	0.5878	0.5853	0.2545	0.3761	0.3902	0.4979	0.4111
Discrete semantic density	0.1040	0.4984	0.5901	0.5904	0.2575	0.3873	0.3857	0.5086	0.4152
Discrete U_{eig}	0.0995	0.4917	0.5840	0.5850	0.2552	0.3821	0.3863	0.4991	0.4104
Discrete U_{deg}	0.1030	0.4878	0.5843	0.5794	0.2563	0.3814	0.3880	0.5055	0.4107
Discrete C_{ecc}	0.0974	0.4928	0.5938	0.5905	0.2499	0.3807	0.3931	0.5132	0.4139
KLE-heat$_{t=0.1}$	0.1048	0.4895	0.5950	0.5852	0.2556	0.3816	0.3881	0.5024	0.4128
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1030	0.4865	0.5956	0.5803	0.2564	0.3799	0.3872	0.5012	0.4113
Discrete LUQ-pair	0.0800	0.4807	0.6042	0.5546	0.2316	0.3713	0.3665	0.4778	0.3958
Discrete SNNE	0.0967	0.4870	0.5778	0.5661	0.2422	0.3718	0.3784	0.4797	0.4000
Discrete SNNE (only-num)	0.1056	0.5290	0.6226	0.6077	0.2684	0.3862	0.4052	0.5053	0.4287
Discrete SNNE (only-denom)	0.1388	0.5670	0.6547	0.6364	0.3012	0.3724	0.4399	0.5079	0.4523
Semantic density	0.1075	0.4940	0.5844	0.5862	0.2577	0.3814	0.3861	0.5079	0.4131
U_{eig}	0.1065	0.4963	0.5855	0.5859	0.2582	0.3832	0.3868	0.5082	0.4138
U_{deg}	0.1115	0.4939	0.5852	0.5834	0.2585	0.3816	0.3862	0.5072	0.4134
U_{ecc}	0.0710	0.4869	0.5998	0.5774	0.2108	0.3780	0.3972	0.4888	0.4012
LUQ	0.1154	0.4873	0.5764	0.5838	0.2550	0.3883	0.3870	0.5092	0.4128
SNNE	0.1116	0.4933	0.5845	0.5846	0.2579	0.3817	0.3870	0.5074	0.4135
SNNE (only-num)	0.1120	0.4940	0.5843	0.5828	0.2580	0.3812	0.3868	0.5069	0.4133
SNNE (only-denom)	0.1126	0.4940	0.5839	0.5848	0.2586	0.3814	0.3865	0.5079	0.4137

Table 54: Per-direction multilingual PRR for MISTRAL-SMALL-3.2-24B-INSTRUCT-2506. PRR is evaluated against per-example ROUGE-L. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.4247	0.4589	0.4828	0.4552	0.3811	0.4910	0.4490
Discrete semantic density	0.4342	0.4710	0.4879	0.4675	0.3919	0.4882	0.4568
Discrete U_{eig}	0.4284	0.4692	0.4839	0.4576	0.3868	0.4928	0.4531
Discrete U_{deg}	0.4173	0.4652	0.4897	0.4634	0.3925	0.4912	0.4532
Discrete C_{ecc}	0.4302	0.4654	0.4880	0.4567	0.3945	0.4949	0.4550
KLE-heat$_{t=0.1}$	0.4264	0.4655	0.4876	0.4634	0.3899	0.4923	0.4542
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.4262	0.4693	0.4836	0.4637	0.3889	0.4920	0.4539
Discrete LUQ-pair	0.3985	0.4510	0.4845	0.4390	0.3785	0.4869	0.4397
Discrete SNNE	0.4134	0.4446	0.4796	0.4393	0.3865	0.4827	0.4410
Discrete SNNE (only-num)	0.4617	0.4969	0.4906	0.4908	0.3924	0.5040	0.4727
Discrete SNNE (only-denom)	0.4617	0.5094	0.4982	0.4976	0.3894	0.5009	0.4762
Semantic density	0.4522	0.4688	0.4842	0.4785	0.3806	0.4891	0.4589
U_{eig}	0.4556	0.4701	0.4870	0.4776	0.3829	0.4910	0.4607
U_{deg}	0.4553	0.4680	0.4857	0.4786	0.3815	0.4893	0.4597
U_{ecc}	0.3312	0.4411	0.4750	0.4292	0.3705	0.4881	0.4225
LUQ	0.4510	0.4721	0.4839	0.4683	0.3870	0.4902	0.4587
SNNE	0.4465	0.4699	0.4834	0.4777	0.3817	0.4899	0.4582
SNNE (only-num)	0.4514	0.4726	0.4833	0.4783	0.3807	0.4887	0.4592
SNNE (only-denom)	0.4481	0.4694	0.4860	0.4791	0.3828	0.4886	0.4590

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.1599	0.5434	0.6343	0.6311	0.3561	0.4668	0.4125	0.5844	0.4736
Discrete semantic density	0.1614	0.5460	0.6403	0.6324	0.3656	0.4720	0.4181	0.5844	0.4775
Discrete U_{eig}	0.1558	0.5425	0.6353	0.6250	0.3671	0.4708	0.4144	0.5871	0.4747
Discrete U_{deg}	0.1598	0.5409	0.6366	0.6258	0.3643	0.4748	0.4177	0.5817	0.4752
Discrete C_{ecc}	0.1597	0.5528	0.6453	0.6150	0.3645	0.4732	0.4255	0.5781	0.4767
KLE-heat$_{t=0.1}$	0.1578	0.5384	0.6377	0.6326	0.3642	0.4707	0.4129	0.5836	0.4747
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1563	0.5447	0.6393	0.6226	0.3679	0.4659	0.4183	0.5857	0.4751
Discrete LUQ-pair	0.1308	0.5307	0.6195	0.6095	0.3231	0.4586	0.4043	0.5800	0.4571
Discrete SNNE	0.1448	0.5315	0.6295	0.6224	0.3443	0.4674	0.4103	0.5652	0.4644
Discrete SNNE (only-num)	0.1694	0.5671	0.6686	0.6546	0.3808	0.4847	0.4380	0.5970	0.4950
Discrete SNNE (only-denom)	0.2129	0.5922	0.6994	0.6673	0.4134	0.4822	0.4682	0.6059	0.5177
Semantic density	0.1604	0.5411	0.6461	0.6356	0.3628	0.4718	0.4177	0.5788	0.4768
U_{eig}	0.1590	0.5452	0.6445	0.6390	0.3624	0.4737	0.4194	0.5782	0.4777
U_{deg}	0.1614	0.5385	0.6445	0.6364	0.3613	0.4714	0.4196	0.5774	0.4763
U_{ecc}	0.1320	0.5402	0.6433	0.6076	0.3215	0.4884	0.4301	0.5929	0.4695
LUQ	0.1603	0.5565	0.6464	0.6348	0.3546	0.4696	0.4213	0.5699	0.4767
SNNE	0.1593	0.5397	0.6459	0.6360	0.3594	0.4719	0.4194	0.5760	0.4760
SNNE (only-num)	0.1594	0.5400	0.6452	0.6363	0.3595	0.4709	0.4176	0.5767	0.4757
SNNE (only-denom)	0.1617	0.5396	0.6422	0.6367	0.3615	0.4719	0.4185	0.5760	0.4760

Table 55: Per-direction multilingual PRR for PHI-4-MINI-INSTRUCT 3.8B. PRR is evaluated against per-example ROUGE-L. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.3286	0.3632	0.3916	0.4093	0.3337	0.4295	0.3760
Discrete semantic density	0.3359	0.3663	0.3970	0.4175	0.3343	0.4356	0.3811
Discrete U_{eig}	0.3342	0.3619	0.3977	0.4156	0.3373	0.4344	0.3802
Discrete U_{deg}	0.3332	0.3599	0.3947	0.4206	0.3389	0.4315	0.3798
Discrete C_{ecc}	0.3381	0.3672	0.3951	0.4131	0.3430	0.4306	0.3812
KLE-heat$_{t=0.1}$	0.3316	0.3644	0.3967	0.4210	0.3357	0.4337	0.3805
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.3347	0.3644	0.3970	0.4200	0.3372	0.4298	0.3805
Discrete LUQ-pair	0.2838	0.3328	0.3747	0.3777	0.3169	0.4177	0.3506
Discrete SNNE	0.3194	0.3540	0.3891	0.3794	0.3298	0.4179	0.3649
Discrete SNNE (only-num)	0.3576	0.3894	0.4090	0.4389	0.3372	0.4462	0.3964
Discrete SNNE (only-denom)	0.3743	0.4178	0.4055	0.4556	0.3367	0.4454	0.4059
Semantic density	0.3419	0.3682	0.4003	0.4173	0.3250	0.4365	0.3815
U_{eig}	0.3443	0.3672	0.3995	0.4211	0.3276	0.4386	0.3831
U_{deg}	0.3435	0.3673	0.3992	0.4177	0.3260	0.4364	0.3817
U_{ecc}	0.2142	0.3086	0.3495	0.3546	0.2968	0.3935	0.3195
LUQ	0.3450	0.3700	0.3987	0.4179	0.3284	0.4351	0.3825
SNNE	0.3409	0.3672	0.4003	0.4145	0.3248	0.4348	0.3804
SNNE (only-num)	0.3376	0.3682	0.4004	0.4190	0.3251	0.4352	0.3809
SNNE (only-denom)	0.3431	0.3689	0.4010	0.4184	0.3251	0.4368	0.3822

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.1209	0.4513	0.5367	0.5159	0.2259	0.3884	0.3310	0.4570	0.3784
Discrete semantic density	0.1225	0.4697	0.5426	0.5236	0.2378	0.3953	0.3517	0.4578	0.3876
Discrete U_{eig}	0.1209	0.4649	0.5398	0.5232	0.2368	0.3973	0.3456	0.4505	0.3849
Discrete U_{deg}	0.1201	0.4581	0.5418	0.5199	0.2361	0.3957	0.3429	0.4518	0.3833
Discrete C_{ecc}	0.1210	0.4631	0.5410	0.5219	0.2346	0.3959	0.3475	0.4605	0.3857
KLE-heat$_{t=0.1}$	0.1220	0.4618	0.5414	0.5211	0.2362	0.3978	0.3403	0.4562	0.3846
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.1215	0.4608	0.5370	0.5190	0.2356	0.3966	0.3432	0.4525	0.3833
Discrete LUQ-pair	0.1142	0.4351	0.5339	0.5026	0.2115	0.3700	0.3242	0.4299	0.3652
Discrete SNNE	0.1177	0.4485	0.5366	0.5056	0.2188	0.3920	0.3122	0.4437	0.3719
Discrete SNNE (only-num)	0.1309	0.4890	0.5764	0.5488	0.2443	0.4062	0.3692	0.4582	0.4029
Discrete SNNE (only-denom)	0.1677	0.5158	0.6166	0.5858	0.2768	0.4033	0.4162	0.4767	0.4324
Semantic density	0.1249	0.4626	0.5487	0.5285	0.2371	0.3981	0.3521	0.4586	0.3888
U_{eig}	0.1234	0.4645	0.5506	0.5297	0.2387	0.4007	0.3540	0.4567	0.3898
U_{deg}	0.1261	0.4630	0.5480	0.5280	0.2367	0.3984	0.3513	0.4578	0.3887
U_{ecc}	0.1189	0.4065	0.5225	0.4891	0.2099	0.3891	0.3348	0.4615	0.3666
LUQ	0.1293	0.4623	0.5407	0.5219	0.2322	0.3970	0.3417	0.4515	0.3846
SNNE	0.1251	0.4616	0.5482	0.5267	0.2365	0.3978	0.3508	0.4577	0.3881
SNNE (only-num)	0.1251	0.4618	0.5492	0.5268	0.2364	0.3979	0.3509	0.4569	0.3881
SNNE (only-denom)	0.1268	0.4630	0.5471	0.5276	0.2363	0.3980	0.3512	0.4586	0.3886

Table 56: Per-direction multilingual PRR for PHI-4 14B. PRR is evaluated against per-example ROUGE-L. The 19 uncertainty measures and notation follow Table 20. Higher is better; Macro is the unweighted mean over directions within each corpus.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.3989	0.4512	0.4804	0.4512	0.3685	0.4819	0.4387
Discrete semantic density	0.4061	0.4610	0.4888	0.4561	0.3909	0.4985	0.4502
Discrete U_{eig}	0.3998	0.4547	0.4883	0.4518	0.3834	0.5001	0.4464
Discrete U_{deg}	0.4049	0.4587	0.4840	0.4523	0.3825	0.4913	0.4456
Discrete C_{ecc}	0.3987	0.4601	0.4871	0.4475	0.3851	0.4978	0.4460
KLE-heat$_{t=0.1}$	0.4064	0.4568	0.4899	0.4512	0.3826	0.4976	0.4474
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.4094	0.4559	0.4896	0.4460	0.3837	0.4928	0.4462
Discrete LUQ-pair	0.3683	0.4477	0.4732	0.4277	0.3710	0.4876	0.4293
Discrete SNNE	0.3774	0.4513	0.4699	0.4399	0.3748	0.4816	0.4325
Discrete SNNE (only-num)	0.4492	0.4902	0.4966	0.4924	0.3885	0.5143	0.4719
Discrete SNNE (only-denom)	0.4549	0.5033	0.4975	0.5023	0.3913	0.5168	0.4777
Semantic density	0.4229	0.4743	0.4929	0.4713	0.3800	0.5000	0.4569
U_{eig}	0.4258	0.4735	0.4937	0.4709	0.3812	0.4994	0.4574
U_{deg}	0.4238	0.4730	0.4915	0.4762	0.3764	0.4992	0.4567
U_{ecc}	0.3015	0.4083	0.4544	0.4233	0.3604	0.4737	0.4036
LUQ	0.4291	0.4640	0.4860	0.4700	0.3837	0.4966	0.4549
SNNE	0.4234	0.4728	0.4885	0.4657	0.3776	0.5000	0.4547
SNNE (only-num)	0.4249	0.4728	0.4891	0.4681	0.3747	0.4987	0.4547
SNNE (only-denom)	0.4249	0.4754	0.4893	0.4720	0.3761	0.5004	0.4563

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.2090	0.5160	0.6190	0.5834	0.3718	0.4838	0.3841	0.5568	0.4655
Discrete semantic density	0.2133	0.5240	0.6248	0.5823	0.3691	0.4811	0.3960	0.5671	0.4697
Discrete U_{eig}	0.2081	0.5262	0.6254	0.5843	0.3691	0.4808	0.3923	0.5664	0.4691
Discrete U_{deg}	0.2106	0.5150	0.6125	0.5879	0.3641	0.4778	0.3922	0.5619	0.4653
Discrete C_{ecc}	0.2070	0.5308	0.6213	0.5907	0.3711	0.4800	0.3921	0.5663	0.4699
KLE-heat$_{t=0.1}$	0.2109	0.5204	0.6099	0.5852	0.3716	0.4770	0.3912	0.5653	0.4664
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.2089	0.5234	0.6199	0.5897	0.3713	0.4741	0.3908	0.5600	0.4673
Discrete LUQ-pair	0.1969	0.5104	0.5814	0.5652	0.3531	0.4747	0.3664	0.5391	0.4484
Discrete SNNE	0.1965	0.5183	0.6099	0.5795	0.3691	0.4844	0.3675	0.5485	0.4592
Discrete SNNE (only-num)	0.2234	0.5635	0.6598	0.6230	0.3805	0.4901	0.4180	0.5751	0.4917
Discrete SNNE (only-denom)	0.2600	0.5918	0.6900	0.6434	0.3931	0.4955	0.4527	0.5813	0.5135
Semantic density	0.2127	0.5311	0.6286	0.5930	0.3643	0.4815	0.3987	0.5611	0.4714
U_{eig}	0.2127	0.5308	0.6293	0.5939	0.3655	0.4818	0.4012	0.5624	0.4722
U_{deg}	0.2126	0.5271	0.6267	0.5931	0.3634	0.4810	0.3980	0.5609	0.4703
U_{ecc}	0.1830	0.5009	0.6128	0.5762	0.3111	0.5020	0.3827	0.5673	0.4545
LUQ	0.2110	0.5247	0.6239	0.5818	0.3623	0.4816	0.3984	0.5578	0.4677
SNNE	0.2130	0.5280	0.6258	0.5910	0.3613	0.4806	0.3962	0.5601	0.4695
SNNE (only-num)	0.2127	0.5281	0.6263	0.5908	0.3630	0.4811	0.3976	0.5597	0.4699
SNNE (only-denom)	0.2138	0.5287	0.6258	0.5920	0.3638	0.4817	0.3950	0.5611	0.4702

H.10 Task-Specific Quality Scores and Recomputed PRRs

Discussion: Appendix F.

Table 57: Task-specific generation quality on the same four summarization datasets and eight WMT19 directions as Table 16. We report SEAHORSE Average and reference-based COMET, both multiplied by 100 for readability. Higher is better; bold marks the best available result in each column, and "–" denotes a missing generation artifact.

Model	Summarization (SEAHORSE Average ×100)				Translation (COMET ×100)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
Llama 3.1-8B Instruct	68.16	29.01	66.93	52.13	83.04	86.75	83.74	82.74	80.90	78.39	84.55	82.12
Llama 3.1-70B Instruct	71.51	35.88	68.47	53.13	84.83	88.06	85.85	87.14	84.86	83.30	85.48	82.77
Llama 3.1-405B Instruct FP8	71.36	26.65	70.79	53.62	85.72	**88.73**	87.07	88.76	85.85	85.53	86.24	84.05
Qwen 3-1.7B	73.35	54.47	72.51	54.99	79.30	84.59	76.10	81.28	75.15	73.88	82.64	81.95
Qwen 3-8B	73.67	**57.40**	73.14	55.71	84.05	87.54	84.33	87.31	83.45	83.31	85.21	83.79
Qwen 3-32B	73.30	54.69	71.25	54.80	85.09	88.16	85.81	88.41	84.76	84.45	85.72	84.15
Phi 3-14B	70.57	49.82	71.24	47.08	28.17	27.83	30.22	35.06	53.07	59.33	81.80	78.61
Phi 4-mini-Instruct 3.8B	65.17	49.15	62.93	44.76	64.55	73.21	73.57	48.29	71.76	56.10	80.60	78.36
Phi 4-14B	69.91	37.32	70.13	53.92	84.00	87.77	85.26	84.36	82.21	81.74	85.40	83.25
Mistral-7B-Instruct-v0.2	71.88	52.88	69.44	53.97	81.71	85.44	76.35	58.64	65.63	66.82	83.54	80.34
Mistral-Nemo-Instruct-2407 12B	67.28	49.41	**74.57**	**57.17**	82.27	86.87	82.13	79.37	82.40	78.71	84.95	82.42
Mistral-Small-3.2-24B	70.07	51.48	72.37	56.07	84.63	88.22	85.44	87.24	85.22	82.38	85.79	83.63
GPT-4o-mini	**74.89**	53.92	71.09	55.08	**85.78**	88.73	**87.26**	**89.01**	**86.67**	**85.87**	**86.37**	**84.28**
Qwen 3-8B (thinking)	74.11	53.28	70.04	53.81	83.02	86.84	83.24	82.24	81.34	82.33	84.63	83.32
DeepSeek-R1-8B (`deepseek-r1:8b`)	58.45	30.85	59.33	45.50	65.66	65.85	62.78	59.45	60.81	58.89	65.16	67.44
DeepSeek-R1-8B (`deepseek-r1-8b`)	48.55	46.51	31.45	29.57	28.80	32.10	30.96	32.36	31.34	30.01	28.88	29.06

Table 58: Reference-based COMET (×100) for six models on the 14 multilingual directions. Macro is the unweighted mean of the displayed direction scores within each corpus; full 19-measure COMET-based UQ results are reported separately for Llama 3.1-8B.

OPUS-100: non-English→non-English

Model	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Llama 3.1-8B Instruct	68.95	72.28	64.33	71.96	60.66	63.85	67.01
Llama 3.1-70B Instruct	72.46	76.72	78.08	74.70	72.44	79.96	75.73
Llama 3.1-405B Instruct FP8	73.54	77.59	80.45	75.18	73.64	81.82	77.04
Qwen 3-1.7B	63.94	69.04	77.25	68.17	70.22	79.77	71.40
Qwen 3-8B	71.49	75.57	80.87	73.64	73.34	82.28	76.20
Qwen 3-32B	73.10	77.21	81.69	74.70	74.01	82.76	77.25

WMT24++: English→multilingual

Model	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Llama 3.1-8B Instruct	44.84	75.44	78.79	75.89	54.77	49.80	71.44	72.05	65.38
Llama 3.1-70B Instruct	66.75	81.08	82.33	80.11	69.80	83.68	80.08	82.54	78.30
Llama 3.1-405B Instruct FP8	70.22	82.09	83.17	81.28	71.28	85.69	82.08	83.93	79.97
Qwen 3-1.7B	59.46	70.87	76.06	72.98	44.60	78.72	70.74	80.56	69.25
Qwen 3-8B	71.09	80.10	81.83	80.25	63.23	85.28	79.90	84.34	78.25
Qwen 3-32B	65.91	81.06	82.38	81.06	65.67	86.50	80.40	84.76	78.47

Table 59: PRR of the 19 SUQ measures against task-specific quality metrics on the same columns as Table 16. Summarization uses per-example SEAHORSE Average and translation uses per-example COMET. Each dataset/direction cell is the unweighted macro average over available generation models; the two Macro columns average all 74 summarization tasks and all 144 WMT19 translation tasks, respectively. Higher is better; bold marks the best SUQ measure in each column.

Uncertainty Method	Summarization (SEAHORSE Average PRR)					Translation (COMET PRR)								
	XSUM	AESLC	CNN/DM	Multi-News	Macro	cs	de	fi	gu	kk	lt	ru	zh	Macro
Discrete semantic entropy	0.6939	0.4874	0.6560	0.5467	0.5973	0.6821	0.7099	0.6839	0.6651	0.6598	0.6346	0.7089	0.7105	0.6819
Discrete semantic density	0.7063	0.4937	0.6597	0.5496	0.6037	0.6879	0.7179	0.6874	0.6703	0.6649	0.6385	0.7166	0.7144	0.6872
Discrete U_{eig}	0.7068	0.4951	0.6619	0.5546	0.6060	0.6853	0.7136	0.6859	0.6699	0.6630	0.6379	0.7120	0.7135	0.6851
Discrete U_{deg}	0.7102	0.5012	0.6719	0.5561	0.6113	0.6846	0.7137	0.6860	0.6702	0.6666	0.6368	0.7114	0.7136	0.6854
Discrete C_{ecc}	**0.7161**	**0.5016**	0.6809	0.5585	0.6158	0.6919	0.7220	0.6935	0.6808	0.6672	0.6416	0.7187	0.7125	0.6910
KLE-heat$_{t=0.1}$	0.7067	0.5009	0.6822	0.5905	0.6209	0.6968	0.7292	0.6915	0.6845	0.6684	0.6411	0.7258	0.7211	0.6948
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7068	0.5011	**0.6822**	0.5905	0.6209	0.6969	0.7291	0.6915	0.6844	0.6685	0.6411	0.7259	0.7211	0.6948
Discrete LUQ-pair	0.6984	0.4818	0.6577	0.5458	0.5973	0.6780	0.7056	0.6786	0.6569	0.6550	0.6236	0.7052	0.6985	0.6752
Discrete SNNE	0.6977	0.4880	0.6586	0.5477	0.5994	0.6767	0.7037	0.6792	0.6610	0.6553	0.6286	0.7034	0.6976	0.6757
Discrete SNNE (only-num)	0.7009	0.4922	0.6634	0.5614	0.6056	0.6952	0.7251	0.6938	0.6787	0.6713	0.6453	0.7222	0.7173	0.6936
Discrete SNNE (only-denom)	0.7048	0.4961	0.6814	**0.6109**	**0.6236**	0.6959	0.7242	**0.6947**	**0.6864**	**0.6751**	**0.6485**	0.7216	0.7169	0.6954
Semantic density	0.6907	0.4853	0.6586	0.5510	0.5976	0.6874	0.7234	0.6786	0.6707	0.6375	0.6256	0.7213	0.7076	0.6815
U_{eig}	0.7100	0.4985	0.6739	0.5784	0.6162	0.6960	0.7290	0.6923	0.6792	0.6698	0.6439	0.7254	0.7196	0.6944
U_{deg}	0.7087	0.4976	0.6728	0.5771	0.6151	0.6961	0.7290	0.6925	0.6792	0.6703	0.6447	0.7251	0.7195	0.6946
U_{ecc}	0.6791	0.4683	0.6624	0.5114	0.5822	0.6519	0.6851	0.6476	0.6455	0.6128	0.5774	0.6876	0.6696	0.6472
LUQ	0.7029	0.4948	0.6699	0.5749	0.6116	**0.6984**	**0.7305**	0.6936	0.6818	0.6722	0.6478	**0.7287**	**0.7225**	**0.6969**
SNNE	0.7081	0.4976	0.6725	0.5771	0.6148	0.6963	0.7292	0.6922	0.6791	0.6705	0.6442	0.7252	0.7199	0.6946
SNNE (only-num)	0.7081	0.4977	0.6725	0.5771	0.6149	0.6962	0.7292	0.6922	0.6790	0.6703	0.6443	0.7251	0.7195	0.6945
SNNE (only-denom)	0.7081	0.4979	0.6725	0.5769	0.6149	0.6961	0.7293	0.6926	0.6796	0.6706	0.6452	0.7253	0.7198	0.6948

Table 60: Per-direction PRR against COMET for LLAMA 3.1-8B INSTRUCT on the 14 multilingual translation directions. The 19 SUQ measures match Table 59; Macro is the unweighted mean over directions within each corpus. Higher is better; bold marks the best measure in each column.

OPUS-100: non-English→non-English							
Uncertainty Method	ar-de	ar-fr	ar-zh	de-fr	de-zh	fr-zh	Macro
Discrete semantic entropy	0.6368	0.6361	0.6498	0.6701	0.6020	0.6702	0.6442
Discrete semantic density	0.6494	0.6443	0.6630	0.6844	0.6129	0.6797	0.6556
Discrete U_{eig}	0.6499	0.6420	0.6607	0.6752	0.6124	0.6777	0.6530
Discrete U_{deg}	0.6474	0.6428	0.6625	0.6752	0.6136	0.6782	0.6533
Discrete C_{ecc}	0.6512	0.6394	0.6463	0.6808	0.6137	0.6689	0.6500
KLE-heat$_{t=0.1}$	0.6483	0.6472	0.6584	0.6858	0.6069	0.6803	0.6545
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.6508	0.6476	0.6601	0.6827	0.6077	0.6812	0.6550
Discrete LUQ-pair	0.5991	0.6130	0.6279	0.6515	0.5595	0.6172	0.6114
Discrete SNNE	0.6305	0.6276	0.6495	0.6638	0.6016	0.6612	0.6390
Discrete SNNE (only-num)	**0.6597**	0.6608	0.6605	0.6967	**0.6154**	0.6751	**0.6613**
Discrete SNNE (only-denom)	0.6584	**0.6670**	0.6372	**0.6974**	0.6056	0.6276	0.6488
Semantic density	0.6525	0.6488	0.6610	0.6866	0.6072	0.6795	0.6559
U_{eig}	0.6558	0.6513	0.6637	0.6905	0.6113	0.6822	0.6591
U_{deg}	0.6543	0.6509	0.6633	0.6848	0.6087	**0.6824**	0.6574
U_{ecc}	0.5206	0.5511	0.5414	0.6053	0.4824	0.5180	0.5365
LUQ	0.6517	0.6393	0.6630	0.6882	0.6120	0.6808	0.6558
SNNE	0.6543	0.6506	**0.6638**	0.6855	0.6073	0.6814	0.6572
SNNE (only-num)	0.6555	0.6506	0.6631	0.6875	0.6073	0.6811	0.6575
SNNE (only-denom)	0.6552	0.6508	0.6632	0.6891	0.6083	0.6822	0.6581

WMT24++: English→multilingual									
Uncertainty Method	en-ar_EG	en-de_DE	en-es_MX	en-fr_FR	en-hi_IN	en-ja_JP	en-ru_RU	en-zh_CN	Macro
Discrete semantic entropy	0.3971	0.6826	0.6897	0.6483	0.5012	0.5576	0.6495	0.7243	0.6063
Discrete semantic density	0.3957	0.6934	0.7072	0.6536	0.5011	0.5778	0.6613	0.7266	**0.6146**
Discrete U_{eig}	0.3891	0.6876	0.7016	0.6536	0.5007	0.5730	0.6620	**0.7306**	0.6123
Discrete U_{deg}	0.3934	0.6822	0.7022	0.6509	0.5000	0.5758	0.6599	0.7274	0.6115
Discrete C_{ecc}	0.3839	0.6896	0.7073	0.6562	0.4968	0.5437	0.6653	0.7275	0.6088
KLE-heat$_{t=0.1}$	0.4068	0.6863	0.6968	0.6448	0.4900	0.5879	0.6566	0.7243	0.6117
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.4058	0.6862	0.6971	0.6454	0.4894	0.5888	0.6563	0.7249	0.6117
Discrete LUQ-pair	0.3669	0.6397	0.6535	0.6204	0.4474	0.4684	0.6199	0.6986	0.5643
Discrete SNNE	0.3783	0.6713	0.6743	0.6322	0.4682	0.5575	0.6332	0.7069	0.5902
Discrete SNNE (only-num)	0.3928	0.7016	0.7089	0.6643	0.5016	0.5525	0.6682	0.7263	0.6145
Discrete SNNE (only-denom)	**0.4150**	**0.7130**	**0.7262**	**0.6864**	**0.5178**	0.4177	**0.6872**	0.7124	0.6094
Semantic density	0.3979	0.6919	0.7022	0.6447	0.4957	0.5848	0.6618	0.7256	0.6131
U_{eig}	0.3955	0.6944	0.7041	0.6482	0.4965	0.5863	0.6638	0.7276	0.6145
U_{deg}	0.3999	0.6911	0.7032	0.6444	0.4957	0.5909	0.6614	0.7261	0.6141
U_{ecc}	0.3400	0.6068	0.6385	0.6008	0.3933	0.3410	0.6006	0.6763	0.5246
LUQ	0.4006	0.6885	0.6984	0.6454	0.4943	0.5844	0.6520	0.7264	0.6113
SNNE	0.3991	0.6903	0.7021	0.6438	0.4939	0.5903	0.6595	0.7254	0.6130
SNNE (only-num)	0.3994	0.6902	0.7012	0.6425	0.4944	0.5906	0.6597	0.7252	0.6129
SNNE (only-denom)	0.4015	0.6913	0.7029	0.6455	0.4959	**0.5913**	0.6610	0.7259	0.6144

Table 61: Model-specific SUQ results using task-specific quality metrics (Part I). Summarization reports PRR against per-example SEAHORSE Average, and WMT19 translation reports PRR against per-example COMET. Higher is better; "–" denotes an unavailable generation artifact. The method and column order follows Table 20.

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
LLAMA 3.1-8B INSTRUCT												
Discrete semantic entropy	0.7510	0.3384	0.7232	0.5851	0.7851	0.8094	0.7926	0.7807	0.7335	0.7741	0.7570	0.7785
Discrete semantic density	0.7676	0.3476	0.7260	0.5836	0.7949	0.8234	0.7960	0.7863	0.7398	0.7747	0.7674	0.7831
Discrete U_{eig}	0.7683	0.3504	0.7273	0.5922	0.7892	0.8157	0.7961	0.7834	0.7411	0.7731	0.7630	0.7832
Discrete U_{deg}	0.7658	0.3483	0.7248	0.5884	0.7876	0.8157	0.7961	0.7850	0.7393	0.7730	0.7608	0.7835
Discrete C_{ecc}	0.7539	0.3444	0.7339	0.5650	0.7876	0.8140	0.7927	0.7806	0.7338	0.7715	0.7607	0.7815
KLE-heat$_{t=0.1}$	0.7582	0.3527	0.7445	0.5947	0.8016	0.8348	0.8008	0.7899	0.7426	0.7752	0.7777	0.7896
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7583	0.3531	0.7445	0.5948	0.8017	0.8348	0.8003	0.7898	0.7424	0.7749	0.7785	0.7892
Discrete LUQ-pair	0.7521	0.3387	0.7212	0.5764	0.7686	0.7899	0.7744	0.7503	0.7057	0.7479	0.7452	0.7618
Discrete SNNE	0.7547	0.3468	0.7242	0.5857	0.7758	0.8000	0.7810	0.7729	0.7243	0.7658	0.7494	0.7626
Discrete SNNE (only-num)	0.7580	0.3421	0.7288	0.5978	0.8056	0.8306	0.8052	0.7914	0.7478	0.7825	0.7752	0.7880
Discrete SNNE (only-denom)	0.7572	0.3235	0.7515	0.6352	0.8069	0.8296	0.8078	0.7945	0.7501	0.7836	0.7746	0.7897
Semantic density	0.7360	0.3309	0.7236	0.5752	0.7958	0.8289	0.7869	0.7684	0.7105	0.7544	0.7757	0.7738
U_{eig}	0.7712	0.3514	0.7376	0.6095	0.8028	0.8358	0.8005	0.7900	0.7428	0.7778	0.7795	0.7881
U_{deg}	0.7693	0.3494	0.7366	0.6083	0.8022	0.8356	0.8010	0.7910	0.7440	0.7790	0.7791	0.7884
U_{ecc}	0.6725	0.3176	0.7094	0.5318	0.7027	0.7289	0.6946	0.6480	0.6344	0.6420	0.6907	0.7128
LUQ	0.7585	0.3425	0.7330	0.6066	0.8009	0.8296	0.8014	0.7924	0.7425	0.7786	0.7802	0.7895
SNNE	0.7685	0.3501	0.7361	0.6084	0.8024	0.8353	0.8014	0.7901	0.7428	0.7778	0.7794	0.7883
SNNE (only-num)	0.7685	0.3499	0.7361	0.6084	0.8019	0.8354	0.8012	0.7904	0.7429	0.7778	0.7789	0.7884
SNNE (only-denom)	0.7684	0.3488	0.7362	0.6083	0.8025	0.8354	0.8013	0.7918	0.7438	0.7795	0.7794	0.7889
LLAMA 3.1-70B INSTRUCT												
Discrete semantic entropy	0.7812	0.4386	0.7187	0.5672	0.7946	0.8255	0.8025	0.8544	0.7555	0.7793	0.7845	0.7688
Discrete semantic density	0.7905	0.4375	0.7199	0.5684	0.8062	0.8356	0.8100	0.8586	0.7606	0.7877	0.7987	0.7731
Discrete U_{eig}	0.7883	0.4364	0.7237	0.5704	0.7999	0.8292	0.8101	0.8547	0.7613	0.7869	0.7915	0.7744
Discrete U_{deg}	0.7871	0.4345	0.7202	0.5675	0.8008	0.8285	0.8078	0.8549	0.7573	0.7832	0.7904	0.7730
Discrete C_{ecc}	0.7875	0.4243	0.7411	0.5760	0.7994	0.8290	0.8062	0.8514	0.7556	0.7829	0.7904	0.7695
KLE-heat$_{t=0.1}$	0.7841	0.4533	0.7602	0.6135	0.8096	0.8490	0.8157	0.8608	0.7639	0.7932	0.8045	0.7828
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7841	0.4532	0.7603	0.6137	0.8101	0.8489	0.8164	0.8602	0.7636	0.7931	0.8047	0.7827
Discrete LUQ-pair	0.7800	0.4234	0.7196	0.5626	0.7787	0.8074	0.7921	0.8370	0.7318	0.7615	0.7730	0.7507
Discrete SNNE	0.7781	0.4387	0.7230	0.5656	0.7862	0.8132	0.7964	0.8441	0.7409	0.7718	0.7778	0.7467
Discrete SNNE (only-num)	0.7847	0.4400	0.7274	0.5751	0.8154	0.8441	0.8157	0.8617	0.7701	0.7971	0.7966	0.7729
Discrete SNNE (only-denom)	0.8022	0.4255	0.7533	0.6230	0.8144	0.8428	0.8151	0.8630	0.7704	0.7983	0.7952	0.7712
Semantic density	0.7748	0.4457	0.7263	0.5701	0.8051	0.8416	0.7975	0.8402	0.7365	0.7758	0.7968	0.7572
U_{eig}	0.7884	0.4442	0.7415	0.6036	0.8114	0.8496	0.8156	0.8601	0.7657	0.7926	0.8053	0.7815
U_{deg}	0.7871	0.4447	0.7394	0.6022	0.8106	0.8493	0.8158	0.8606	0.7659	0.7922	0.8046	0.7818
U_{ecc}	0.7374	0.3624	0.7144	0.5236	0.7379	0.7574	0.7278	0.7583	0.6668	0.6808	0.7299	0.6984
LUQ	0.7808	0.4512	0.7349	0.5979	0.8101	0.8467	0.8138	0.8615	0.7648	0.7945	0.8039	0.7784
SNNE	0.7863	0.4466	0.7386	0.6019	0.8109	0.8492	0.8154	0.8598	0.7657	0.7921	0.8049	0.7816
SNNE (only-num)	0.7864	0.4462	0.7387	0.6019	0.8108	0.8492	0.8150	0.8598	0.7658	0.7921	0.8045	0.7817
SNNE (only-denom)	0.7865	0.4456	0.7388	0.6019	0.8102	0.8493	0.8156	0.8610	0.7658	0.7923	0.8049	0.7820

Table 61: Model-specific SUQ results using task-specific quality metrics (Part I). Summarization reports PRR against per-example SEAHORSE Average, and WMT19 translation reports PRR against per-example COMET. Higher is better; "–" denotes an unavailable generation artifact. The method and column order follows Table 20. (Continued.)

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
LLAMA 3.1-405B INSTRUCT FP8												
Discrete semantic entropy	0.7693	0.2984	0.7635	0.5880	0.7946	0.8496	0.7961	0.8528	0.7378	0.7844	0.7739	0.7808
Discrete semantic density	0.7748	0.3019	0.7632	0.5867	0.8021	0.8609	0.7977	0.8590	0.7418	0.8038	0.7876	0.7867
Discrete U_{eig}	0.7758	0.3031	0.7677	0.5898	0.7992	0.8532	0.7955	0.8561	0.7447	0.7976	0.7836	0.7850
Discrete U_{deg}	0.7753	0.3027	0.7643	0.5851	0.7935	0.8530	0.7925	0.8502	0.7431	0.7973	0.7858	0.7856
Discrete C_{ecc}	0.7703	0.3013	0.7605	0.5810	0.7922	0.8525	0.7934	0.8516	0.7378	0.7955	0.7831	0.7840
KLE-heat$_{t=0.1}$	0.7713	0.3061	0.7856	0.6361	0.8068	0.8685	0.8049	0.8638	0.7495	0.8057	0.8041	0.7936
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7714	0.3063	0.7856	0.6363	0.8068	0.8685	0.8049	0.8638	0.7494	0.8057	0.8032	0.7936
Discrete LUQ-pair	0.7694	0.2984	0.7603	0.5816	0.7782	0.8399	0.7863	0.8341	0.7266	0.7782	0.7715	0.7599
Discrete SNNE	0.7713	0.2957	0.7638	0.5824	0.7845	0.8435	0.7907	0.8406	0.7278	0.7815	0.7735	0.7583
Discrete SNNE (only-num)	0.7765	0.3004	0.7700	0.5924	0.8117	0.8661	0.8074	0.8681	0.7503	0.8079	0.7968	0.7836
Discrete SNNE (only-denom)	0.7877	0.2989	0.7916	0.6360	0.8109	0.8647	0.8056	0.8666	0.7516	0.8071	0.7954	0.7818
Semantic density	0.7666	0.2891	0.7575	0.5853	0.8086	0.8654	0.7952	0.8575	0.7248	0.8040	0.7982	0.7783
U_{eig}	0.7756	0.3036	0.7826	0.6314	0.8143	0.8725	0.8042	0.8665	0.7471	0.8071	0.8041	0.7929
U_{deg}	0.7747	0.3027	0.7810	0.6299	0.8132	0.8724	0.8054	0.8665	0.7469	0.8072	0.8033	0.7932
U_{ecc}	0.7469	0.2643	0.7405	0.5330	0.7439	0.8002	0.7443	0.7686	0.6747	0.7337	0.7309	0.7274
LUQ	0.7700	0.3021	0.7769	0.6265	0.8117	0.8681	0.8049	0.8659	0.7456	0.8087	0.7991	0.7914
SNNE	0.7741	0.3028	0.7804	0.6296	0.8139	0.8724	0.8052	0.8663	0.7473	0.8066	0.8028	0.7934
SNNE (only-num)	0.7741	0.3027	0.7805	0.6296	0.8139	0.8724	0.8050	0.8661	0.7469	0.8065	0.8028	0.7927
SNNE (only-denom)	0.7742	0.3023	0.7806	0.6296	0.8135	0.8724	0.8043	0.8666	0.7478	0.8066	0.8036	0.7934
QWEN 3-1.7B												
Discrete semantic entropy	0.8079	0.6367	0.7675	0.5778	0.7291	0.7710	0.6961	0.7946	0.7024	0.6825	0.8081	0.7454
Discrete semantic density	0.8158	0.6401	0.7704	0.5758	0.7330	0.7825	0.6935	0.7921	0.7049	0.6786	0.8172	0.7491
Discrete U_{eig}	0.8150	0.6446	0.7688	0.5746	0.7334	0.7761	0.6948	0.7934	0.7033	0.6776	0.8106	0.7511
Discrete U_{deg}	0.8142	0.6412	0.7662	0.5706	0.7356	0.7757	0.6942	0.7941	0.7046	0.6792	0.8104	0.7503
Discrete C_{ecc}	0.8093	0.6310	0.7688	0.5931	0.7257	0.7707	0.6875	0.7892	0.7018	0.6746	0.8107	0.7504
KLE-heat$_{t=0.1}$	0.8040	0.6455	0.7874	0.6425	0.7382	0.7934	0.6903	0.7965	0.7076	0.6783	0.8207	0.7546
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8040	0.6447	0.7874	0.6423	0.7393	0.7929	0.6893	0.7957	0.7076	0.6779	0.8212	0.7546
Discrete LUQ-pair	0.8060	0.6240	0.7598	0.5658	0.7126	0.7538	0.6653	0.7695	0.6904	0.6654	0.7967	0.7326
Discrete SNNE	0.8077	0.6366	0.7644	0.5695	0.7241	0.7633	0.6847	0.7859	0.6955	0.6786	0.8025	0.7302
Discrete SNNE (only-num)	0.8103	0.6426	0.7710	0.5870	0.7407	0.7896	0.6962	0.7968	0.7083	0.6805	0.8191	0.7531
Discrete SNNE (only-denom)	0.8117	0.6346	0.7914	0.6623	0.7389	0.7888	0.6993	0.7954	0.7138	0.6785	0.8175	0.7543
Semantic density	0.7949	0.6249	0.7569	0.5872	0.7246	0.7827	0.6695	0.7826	0.6889	0.6560	0.8160	0.7438
U_{eig}	0.8154	0.6454	0.7849	0.6155	0.7410	0.7912	0.6946	0.7975	0.7084	0.6804	0.8216	0.7530
U_{deg}	0.8148	0.6443	0.7839	0.6140	0.7421	0.7918	0.6960	0.7975	0.7087	0.6809	0.8215	0.7539
U_{ecc}	0.7524	0.5575	0.7323	0.5421	0.6432	0.6926	0.5930	0.7006	0.6314	0.5851	0.7560	0.6807
LUQ	0.8107	0.6337	0.7791	0.6084	0.7374	0.7862	0.6891	0.7943	0.7089	0.6804	0.8182	0.7551
SNNE	0.8145	0.6438	0.7835	0.6135	0.7410	0.7917	0.6958	0.7971	0.7084	0.6809	0.8214	0.7539
SNNE (only-num)	0.8145	0.6447	0.7836	0.6136	0.7414	0.7917	0.6950	0.7962	0.7082	0.6792	0.8214	0.7539
SNNE (only-denom)	0.8145	0.6442	0.7836	0.6136	0.7406	0.7918	0.6952	0.7980	0.7086	0.6817	0.8212	0.7540

Table 61: Model-specific SUQ results using task-specific quality metrics (Part I). Summarization reports PRR against per-example SEAHORSE Average, and WMT19 translation reports PRR against per-example COMET. Higher is better; "–" denotes an unavailable generation artifact. The method and column order follows Table 20. (Continued.)

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
QWEN 3-8B												
Discrete semantic entropy	0.8099	0.6856	0.7798	0.6079	0.7703	0.8666	0.7776	0.8382	0.7248	0.7369	0.7611	0.7753
Discrete semantic density	0.8147	0.6864	0.7802	0.6032	0.7785	0.8705	0.7726	0.8417	0.7326	0.7366	0.7658	0.7797
Discrete U_{eig}	0.8155	0.6888	0.7800	0.5981	0.7746	0.8659	0.7748	0.8397	0.7246	0.7387	0.7631	0.7755
Discrete U_{deg}	0.8145	0.6894	0.7770	0.5932	0.7720	0.8658	0.7757	0.8398	0.7298	0.7329	0.7633	0.7773
Discrete C_{ecc}	0.8044	0.6741	0.7780	0.6075	0.7675	0.8665	0.7694	0.8340	0.7219	0.7354	0.7641	0.7708
KLE-heat$_{t=0.1}$	0.8029	0.6897	0.7915	0.6435	0.7858	0.8799	0.7776	0.8482	0.7336	0.7481	0.7779	0.7850
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8029	0.6904	0.7914	0.6433	0.7857	0.8798	0.7772	0.8486	0.7332	0.7474	0.7782	0.7840
Discrete LUQ-pair	0.8088	0.6645	0.7709	0.5867	0.7529	0.8570	0.7605	0.8224	0.7163	0.7180	0.7489	0.7552
Discrete SNNE	0.8106	0.6870	0.7749	0.5909	0.7629	0.8591	0.7710	0.8256	0.7159	0.7274	0.7526	0.7533
Discrete SNNE (only-num)	0.8124	0.6877	0.7812	0.6097	0.7859	0.8772	0.7796	0.8456	0.7375	0.7483	0.7754	0.7811
Discrete SNNE (only-denom)	0.8153	0.6836	0.7983	0.6715	0.7855	0.8779	0.7784	0.8446	0.7375	0.7469	0.7750	0.7815
Semantic density	0.8016	0.6766	0.7728	0.6020	0.7804	0.8785	0.7653	0.8425	0.7208	0.7404	0.7763	0.7790
U_{eig}	0.8155	0.6912	0.7902	0.6288	0.7874	0.8793	0.7802	0.8491	0.7342	0.7467	0.7794	0.7856
U_{deg}	0.8147	0.6914	0.7889	0.6271	0.7866	0.8808	0.7796	0.8468	0.7340	0.7469	0.7794	0.7849
U_{ecc}	0.7565	0.6079	0.7498	0.5476	0.7114	0.8239	0.6983	0.7692	0.6626	0.6594	0.7082	0.7224
LUQ	0.8111	0.6849	0.7871	0.6304	0.7818	0.8730	0.7749	0.8426	0.7335	0.7472	0.7746	0.7828
SNNE	0.8145	0.6914	0.7888	0.6277	0.7875	0.8802	0.7796	0.8487	0.7344	0.7467	0.7770	0.7854
SNNE (only-num)	0.8145	0.6905	0.7889	0.6277	0.7875	0.8802	0.7796	0.8479	0.7344	0.7483	0.7770	0.7845
SNNE (only-denom)	0.8144	0.6911	0.7889	0.6276	0.7875	0.8808	0.7802	0.8490	0.7347	0.7472	0.7782	0.7863

Table 62: Model-specific SUQ results using task-specific quality metrics (Part II). Summarization reports PRR against per-example SEAHORSE Average, and WMT19 translation reports PRR against per-example COMET. Higher is better; "–" denotes an unavailable generation artifact. The method and column order follows Table 21.

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
QWEN 3-32B												
Discrete semantic entropy	0.8007	0.6512	0.7616	0.5873	0.7948	0.8196	0.8123	0.8105	0.7615	0.7743	0.7586	0.7889
Discrete semantic density	0.8101	0.6555	0.7625	0.5858	0.8023	0.8302	0.8152	0.8170	0.7641	0.7781	0.7693	0.7987
Discrete U_{eig}	0.8111	0.6556	0.7634	0.5920	0.7971	0.8241	0.8115	0.8123	0.7637	0.7786	0.7570	0.7926
Discrete U_{deg}	0.8103	0.6530	0.7604	0.5873	0.7999	0.8240	0.8156	0.8171	0.7634	0.7704	0.7578	0.7942
Discrete C_{ecc}	0.8042	0.6454	0.7650	0.5871	0.7969	0.8199	0.8133	0.8093	0.7605	0.7798	0.7519	0.7914
KLE-heat$_{t=0.1}$	0.7964	0.6565	0.7853	0.6323	0.8121	0.8443	0.8233	0.8232	0.7697	0.7845	0.7768	0.8054
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7965	0.6563	0.7853	0.6322	0.8112	0.8448	0.8235	0.8235	0.7697	0.7842	0.7778	0.8051
Discrete LUQ-pair	0.8049	0.6305	0.7526	0.5754	0.7810	0.8053	0.8037	0.7846	0.7397	0.7545	0.7425	0.7714
Discrete SNNE	0.8051	0.6481	0.7613	0.5821	0.7857	0.8096	0.8068	0.7986	0.7491	0.7601	0.7501	0.7698
Discrete SNNE (only-num)	0.8065	0.6570	0.7684	0.6013	0.8126	0.8423	0.8268	0.8254	0.7749	0.7821	0.7761	0.8009
Discrete SNNE (only-denom)	0.8077	0.6618	0.7933	0.6691	0.8123	0.8420	0.8270	0.8248	0.7732	0.7843	0.7765	0.8003
Semantic density	0.7884	0.6480	0.7578	0.5871	0.8057	0.8397	0.8140	0.8137	0.7428	0.7772	0.7761	0.7954
U_{eig}	0.8105	0.6575	0.7773	0.6210	0.8110	0.8452	0.8236	0.8239	0.7711	0.7834	0.7775	0.8040
U_{deg}	0.8095	0.6577	0.7758	0.6189	0.8116	0.8448	0.8236	0.8243	0.7708	0.7836	0.7775	0.8038
U_{ecc}	0.7451	0.5712	0.7352	0.5463	0.7349	0.7545	0.7393	0.7169	0.6749	0.6939	0.6995	0.7360
LUQ	0.8034	0.6515	0.7734	0.6157	0.8079	0.8412	0.8220	0.8218	0.7671	0.7840	0.7769	0.8020
SNNE	0.8091	0.6576	0.7755	0.6185	0.8117	0.8452	0.8231	0.8236	0.7706	0.7824	0.7773	0.8044
SNNE (only-num)	0.8091	0.6581	0.7755	0.6186	0.8114	0.8447	0.8236	0.8238	0.7706	0.7828	0.7776	0.8041
SNNE (only-denom)	0.8090	0.6578	0.7756	0.6186	0.8120	0.8447	0.8239	0.8239	0.7710	0.7835	0.7773	0.8037
MISTRAL-7B-INSTRUCT-V0.2												
Discrete semantic entropy	0.7766	0.5956	0.7302	0.5844	0.7657	0.7969	0.7319	0.5198	0.5361	0.5932	0.7620	0.7689
Discrete semantic density	0.7947	0.6122	0.7330	0.5848	0.7719	0.8043	0.7353	0.5191	0.5396	0.5938	0.7686	0.7680
Discrete U_{eig}	0.7936	0.6179	0.7372	0.5933	0.7676	0.8023	0.7360	0.5166	0.5415	0.5912	0.7621	0.7672
Discrete U_{deg}	0.7904	0.6115	0.7341	0.5906	0.7690	0.8016	0.7333	0.5197	0.5428	0.5910	0.7606	0.7639
Discrete C_{ecc}	0.7956	0.6166	0.7422	0.5725	0.7683	0.8007	0.7341	0.5137	0.5265	0.5811	0.7588	0.7676
KLE-heat$_{t=0.1}$	0.8002	0.6337	0.7657	0.6123	0.7792	0.8168	0.7313	0.5114	0.5199	0.5775	0.7800	0.7755
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8003	0.6337	0.7657	0.6123	0.7788	0.8165	0.7312	0.5115	0.5199	0.5779	0.7805	0.7759
Discrete LUQ-pair	0.7813	0.5853	0.7305	0.5809	0.7419	0.7715	0.6984	0.5027	0.5277	0.5679	0.7407	0.7392
Discrete SNNE	0.7815	0.6022	0.7340	0.5857	0.7565	0.7828	0.7179	0.5143	0.5365	0.5845	0.7514	0.7487
Discrete SNNE (only-num)	0.7858	0.6091	0.7395	0.5999	0.7818	0.8150	0.7419	0.5239	0.5459	0.6032	0.7773	0.7740
Discrete SNNE (only-denom)	0.7974	0.6149	0.7671	0.6635	0.7824	0.8150	0.7466	0.5295	0.5558	0.6108	0.7764	0.7786
Semantic density	0.7760	0.5894	0.7347	0.5802	0.7650	0.8095	0.7074	0.5000	0.5091	0.5484	0.7721	0.7555
U_{eig}	0.7983	0.6241	0.7541	0.6233	0.7798	0.8177	0.7363	0.5189	0.5438	0.5933	0.7801	0.7721
U_{deg}	0.7965	0.6223	0.7525	0.6223	0.7814	0.8181	0.7373	0.5225	0.5456	0.5966	0.7806	0.7741
U_{ecc}	0.7280	0.5422	0.7160	0.5382	0.6737	0.6970	0.6134	0.4499	0.4492	0.4353	0.6814	0.6835
LUQ	0.7854	0.6070	0.7462	0.6152	0.7770	0.8152	0.7366	0.5228	0.5456	0.5993	0.7772	0.7742
SNNE	0.7952	0.6203	0.7515	0.6219	0.7807	0.8179	0.7367	0.5200	0.5468	0.5946	0.7808	0.7737
SNNE (only-num)	0.7953	0.6206	0.7516	0.6219	0.7804	0.8180	0.7362	0.5202	0.5468	0.5945	0.7806	0.7732
SNNE (only-denom)	0.7955	0.6213	0.7518	0.6220	0.7811	0.8181	0.7375	0.5237	0.5466	0.5978	0.7804	0.7746

Table 62: Model-specific SUQ results using task-specific quality metrics (Part II). Summarization reports PRR against per-example SEAHORSE Average, and WMT19 translation reports PRR against per-example COMET. Higher is better; "–" denotes an unavailable generation artifact. The method and column order follows Table 21. (Continued.)

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
MISTRAL-NEMO-INSTRUCT-2407 12B												
Discrete semantic entropy	0.7438	0.5820	0.7919	0.6131	0.7983	0.8331	0.7975	0.7690	0.7979	0.7802	0.7653	0.8031
Discrete semantic density	0.7628	0.5858	0.7972	0.6139	0.8060	0.8422	0.8005	0.7693	0.7993	0.7814	0.7819	0.8076
Discrete U_{eig}	0.7631	0.5866	0.8027	0.6377	0.8012	0.8387	0.7961	0.7671	0.7988	0.7793	0.7730	0.8084
Discrete U_{deg}	0.7610	0.5857	0.8002	0.6312	0.8006	0.8377	0.7995	0.7692	0.7988	0.7794	0.7720	0.8065
Discrete C_{ecc}	0.7461	0.5703	0.7888	0.5941	0.7987	0.8347	0.7946	0.7559	0.7918	0.7717	0.7706	0.8068
KLE-heat$_{t=0.1}$	0.7473	0.5887	0.8065	0.6409	0.8098	0.8520	0.8016	0.7637	0.7985	0.7759	0.7859	0.8158
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7473	0.5885	0.8066	0.6409	0.8098	0.8511	0.8018	0.7635	0.7986	0.7758	0.7857	0.8156
Discrete LUQ-pair	0.7454	0.5644	0.7964	0.6210	0.7786	0.8149	0.7696	0.7459	0.7782	0.7567	0.7548	0.7835
Discrete SNNE	0.7476	0.5776	0.7963	0.6226	0.7850	0.8208	0.7897	0.7580	0.7848	0.7679	0.7575	0.7875
Discrete SNNE (only-num)	0.7529	0.5853	0.8006	0.6368	0.8112	0.8496	0.8097	0.7731	0.8052	0.7859	0.7837	0.8150
Discrete SNNE (only-denom)	0.7558	0.5989	0.8127	0.6782	0.8121	0.8493	0.8114	0.7790	0.8060	0.7887	0.7830	0.8171
Semantic density	0.7239	0.5682	0.7872	0.6104	0.8005	0.8481	0.7825	0.7315	0.7744	0.7543	0.7838	0.8014
U_{eig}	0.7645	0.5905	0.8146	0.6687	0.8106	0.8529	0.8042	0.7695	0.8015	0.7828	0.7873	0.8137
U_{deg}	0.7633	0.5910	0.8136	0.6666	0.8103	0.8523	0.8051	0.7726	0.8020	0.7834	0.7872	0.8139
U_{ecc}	0.6560	0.5111	0.7668	0.5546	0.7060	0.7614	0.6716	0.6338	0.7142	0.6553	0.7048	0.7416
LUQ	0.7531	0.5889	0.8075	0.6532	0.8089	0.8516	0.8045	0.7737	0.8003	0.7847	0.7886	0.8145
SNNE	0.7626	0.5919	0.8131	0.6660	0.8102	0.8520	0.8044	0.7713	0.8019	0.7831	0.7865	0.8138
SNNE (only-num)	0.7627	0.5917	0.8132	0.6660	0.8102	0.8528	0.8044	0.7714	0.8020	0.7830	0.7869	0.8140
SNNE (only-denom)	0.7626	0.5913	0.8132	0.6661	0.8104	0.8528	0.8054	0.7734	0.8024	0.7838	0.7870	0.8142
MISTRAL-SMALL-3.2-24B												
Discrete semantic entropy	0.7645	0.6100	0.7582	0.5819	0.7923	0.8437	0.8011	0.8329	0.7602	0.7672	0.7495	0.7776
Discrete semantic density	0.7742	0.6142	0.7591	0.5805	0.7964	0.8551	0.8098	0.8380	0.7667	0.7745	0.7597	0.7862
Discrete U_{eig}	0.7761	0.6152	0.7626	0.5907	0.7922	0.8468	0.8033	0.8328	0.7564	0.7752	0.7529	0.7834
Discrete U_{deg}	0.7749	0.6102	0.7588	0.5833	0.7916	0.8467	0.8019	0.8349	0.7600	0.7748	0.7529	0.7838
Discrete C_{ecc}	0.7677	0.6143	0.7754	0.5973	0.7872	0.8458	0.8013	0.8319	0.7607	0.7695	0.7501	0.7810
KLE-heat$_{t=0.1}$	0.7657	0.6363	0.7967	0.6636	0.8067	0.8661	0.8147	0.8462	0.7760	0.7817	0.7736	0.7955
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7657	0.6362	0.7966	0.6635	0.8066	0.8665	0.8150	0.8462	0.7765	0.7818	0.7734	0.7953
Discrete LUQ-pair	0.7659	0.5898	0.7563	0.5773	0.7698	0.8324	0.7890	0.8077	0.7525	0.7438	0.7407	0.7578
Discrete SNNE	0.7668	0.6035	0.7590	0.5872	0.7778	0.8364	0.7911	0.8204	0.7578	0.7548	0.7445	0.7561
Discrete SNNE (only-num)	0.7714	0.6169	0.7660	0.6064	0.8052	0.8626	0.8159	0.8460	0.7785	0.7807	0.7681	0.7902
Discrete SNNE (only-denom)	0.7798	0.6341	0.7948	0.6745	0.8048	0.8620	0.8165	0.8462	0.7773	0.7809	0.7668	0.7901
Semantic density	0.7533	0.6069	0.7638	0.5917	0.8010	0.8626	0.8051	0.8393	0.7617	0.7689	0.7712	0.7860
U_{eig}	0.7763	0.6213	0.7780	0.6346	0.8066	0.8670	0.8152	0.8447	0.7758	0.7828	0.7743	0.7956
U_{deg}	0.7751	0.6202	0.7761	0.6320	0.8063	0.8665	0.8150	0.8456	0.7755	0.7829	0.7728	0.7956
U_{ecc}	0.7073	0.5379	0.7459	0.5335	0.7240	0.7904	0.7144	0.7372	0.6958	0.6663	0.6974	0.7210
LUQ	0.7678	0.6210	0.7728	0.6262	0.8067	0.8635	0.8131	0.8463	0.7722	0.7810	0.7761	0.7948
SNNE	0.7743	0.6207	0.7756	0.6314	0.8078	0.8669	0.8150	0.8450	0.7760	0.7832	0.7735	0.7958
SNNE (only-num)	0.7744	0.6206	0.7757	0.6314	0.8075	0.8669	0.8150	0.8451	0.7760	0.7835	0.7735	0.7956
SNNE (only-denom)	0.7743	0.6203	0.7758	0.6316	0.8075	0.8670	0.8151	0.8461	0.7761	0.7833	0.7736	0.7954

Table 62: Model-specific SUQ results using task-specific quality metrics (Part II). Summarization reports PRR against per-example SEAHORSE Average, and WMT19 translation reports PRR against per-example COMET. Higher is better; "–" denotes an unavailable generation artifact. The method and column order follows Table 21. (Continued.)

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
GPT-4O-MINI												
Discrete semantic entropy	0.8073	0.6405	0.7529	0.5838	0.7828	0.8125	0.8068	0.8438	0.7479	0.7362	0.7452	0.7692
Discrete semantic density	0.8110	0.6413	0.7532	0.5820	0.7931	0.8244	0.8100	0.8460	0.7489	0.7375	0.7525	0.7849
Discrete U_{eig}	0.8119	0.6408	0.7546	0.5874	0.7833	0.8163	0.8076	0.8449	0.7489	0.7381	0.7475	0.7777
Discrete U_{deg}	0.8112	0.6393	0.7524	0.5831	0.7830	0.8169	0.8079	0.8448	0.7489	0.7381	0.7475	0.7780
Discrete C_{ecc}	0.8107	0.6364	0.7650	0.6024	0.7833	0.8160	0.8071	0.8423	0.7471	0.7385	0.7462	0.7707
KLE-heat$_{t=0.1}$	0.8086	0.6511	0.7812	0.6417	0.8036	0.8425	0.8168	0.8570	0.7596	0.7539	0.7690	0.7944
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.8086	0.6520	0.7812	0.6416	0.8036	0.8417	0.8168	0.8570	0.7585	0.7539	0.7705	0.7944
Discrete LUQ-pair	0.8103	0.6165	0.7515	0.5757	0.7749	0.8048	0.8023	0.8279	0.7472	0.7244	0.7392	0.7620
Discrete SNNE	0.8106	0.6335	0.7535	0.5849	0.7766	0.8073	0.8025	0.8310	0.7480	0.7268	0.7394	0.7627
Discrete SNNE (only-num)	0.8122	0.6445	0.7603	0.6031	0.8024	0.8374	0.8175	0.8540	0.7607	0.7514	0.7658	0.7899
Discrete SNNE (only-denom)	0.8183	0.6517	0.7872	0.6756	0.8012	0.8382	0.8191	0.8525	0.7619	0.7523	0.7653	0.7909
Semantic density	0.7987	0.6349	0.7543	0.5949	0.7998	0.8409	0.8130	0.8509	0.7523	0.7558	0.7697	0.7858
U_{eig}	0.8116	0.6455	0.7641	0.6064	0.8029	0.8419	0.8175	0.8578	0.7600	0.7546	0.7692	0.7937
U_{deg}	0.8110	0.6456	0.7629	0.6051	0.8029	0.8420	0.8175	0.8565	0.7576	0.7544	0.7706	0.7931
U_{ecc}	0.7936	0.5630	0.7486	0.5697	0.7504	0.7599	0.7678	0.7660	0.7133	0.6864	0.6972	0.7291
LUQ	0.8079	0.6424	0.7611	0.6064	0.8011	0.8323	0.8142	0.8542	0.7585	0.7568	0.7676	0.7877
SNNE	0.8106	0.6444	0.7626	0.6052	0.8029	0.8413	0.8176	0.8575	0.7605	0.7541	0.7713	0.7933
SNNE (only-num)	0.8106	0.6445	0.7627	0.6052	0.8029	0.8428	0.8176	0.8575	0.7582	0.7557	0.7713	0.7931
SNNE (only-denom)	0.8106	0.6458	0.7628	0.6051	0.8029	0.8434	0.8176	0.8575	0.7584	0.7557	0.7713	0.7931

Table 63: Model-specific SUQ results using task-specific quality metrics (Part III). Summarization reports PRR against per-example SEAHORSE Average, and WMT19 translation reports PRR against per-example COMET. Higher is better; "–" denotes an unavailable generation artifact. The method and column order follows Table 22.

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
PHI 4-14B												
Discrete semantic entropy	0.7544	0.4167	0.7396	0.5457	0.8028	0.8122	0.8036	0.8143	0.7753	0.7613	0.7564	0.7865
Discrete semantic density	0.7713	0.4154	0.7411	0.5404	0.8121	0.8236	0.8104	0.8181	0.7751	0.7614	0.7743	0.7921
Discrete U_{eig}	0.7701	0.4185	0.7432	0.5477	0.8064	0.8145	0.8021	0.8172	0.7757	0.7575	0.7647	0.7921
Discrete U_{deg}	0.7663	0.4161	0.7407	0.5440	0.8044	0.8163	0.8016	0.8168	0.7767	0.7598	0.7623	0.7884
Discrete C_{ecc}	0.7763	0.4229	0.7606	0.5759	0.8022	0.8135	0.8008	0.8115	0.7730	0.7547	0.7632	0.7907
KLE-heat$_{t=0.1}$	0.7858	0.4208	0.7787	0.6140	0.8163	0.8334	0.8135	0.8234	0.7760	0.7635	0.7823	0.7997
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7858	0.4208	0.7786	0.6140	0.8165	0.8333	0.8133	0.8231	0.7764	0.7635	0.7821	0.7992
Discrete LUQ-pair	0.7560	0.4082	0.7408	0.5427	0.7873	0.7893	0.7836	0.7900	0.7504	0.7265	0.7364	0.7627
Discrete SNNE	0.7598	0.4152	0.7434	0.5507	0.7911	0.7985	0.7894	0.8093	0.7643	0.7464	0.7511	0.7657
Discrete SNNE (only-num)	0.7669	0.4166	0.7510	0.5686	0.8191	0.8336	0.8174	0.8274	0.7824	0.7695	0.7821	0.8001
Discrete SNNE (only-denom)	0.7891	0.4149	0.7880	0.6488	0.8197	0.8334	0.8179	0.8302	0.7833	0.7700	0.7819	0.8024
Semantic density	0.7476	0.4196	0.7494	0.5628	0.8106	0.8326	0.8035	0.8041	0.7480	0.7481	0.7810	0.7868
U_{eig}	0.7807	0.4208	0.7555	0.5743	0.8161	0.8344	0.8146	0.8226	0.7780	0.7651	0.7827	0.7967
U_{deg}	0.7781	0.4200	0.7548	0.5735	0.8165	0.8336	0.8145	0.8238	0.7790	0.7649	0.7826	0.7979
U_{ecc}	0.6940	0.3909	0.7427	0.5301	0.7237	0.7326	0.7171	0.6974	0.6817	0.6432	0.6888	0.7267
LUQ	0.7685	0.4186	0.7533	0.5725	0.8160	0.8305	0.8112	0.8227	0.7786	0.7684	0.7823	0.7995
SNNE	0.7771	0.4200	0.7545	0.5735	0.8163	0.8344	0.8139	0.8231	0.7779	0.7648	0.7820	0.7979
SNNE (only-num)	0.7771	0.4199	0.7545	0.5734	0.8162	0.8342	0.8141	0.8232	0.7783	0.7645	0.7826	0.7974
SNNE (only-denom)	0.7772	0.4199	0.7545	0.5733	0.8164	0.8336	0.8147	0.8243	0.7788	0.7658	0.7827	0.7983
PHI 4-MINI-INSTRUCT 3.8B												
Discrete semantic entropy	0.7031	0.5539	0.6537	0.4578	0.6163	0.7208	0.7215	0.3257	0.6938	0.4988	0.7565	0.7210
Discrete semantic density	0.7520	0.5805	0.6739	0.4918	0.6233	0.7195	0.7181	0.3390	0.6940	0.5170	0.7594	0.7201
Discrete U_{eig}	0.7478	0.5800	0.6783	0.4984	0.6200	0.7170	0.7148	0.3494	0.6978	0.5163	0.7570	0.7189
Discrete U_{deg}	0.7440	0.5752	0.6751	0.4972	0.6189	0.7175	0.7204	0.3386	0.7020	0.5152	0.7569	0.7196
Discrete C_{ecc}	0.7538	0.5815	0.6751	0.4964	0.6063	0.7029	0.7071	0.3747	0.6729	0.4977	0.7478	0.7165
KLE-heat$_{t=0.1}$	0.7591	0.5880	0.6922	0.5463	0.6409	0.7285	0.7141	0.4366	0.7094	0.5180	0.7672	0.7277
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7592	0.5880	0.6923	0.5462	0.6406	0.7276	0.7139	0.4370	0.7095	0.5178	0.7671	0.7275
Discrete LUQ-pair	0.7169	0.5476	0.6619	0.4725	0.5747	0.6684	0.7027	0.3377	0.6832	0.4824	0.7289	0.6979
Discrete SNNE	0.7181	0.5522	0.6637	0.4793	0.6155	0.7129	0.7180	0.3272	0.6972	0.5071	0.7423	0.7120
Discrete SNNE (only-num)	0.7254	0.5638	0.6711	0.5011	0.6338	0.7291	0.7306	0.3382	0.7123	0.5237	0.7656	0.7297
Discrete SNNE (only-denom)	0.7446	0.6001	0.7027	0.5746	0.6470	0.7299	0.7380	0.4347	0.7327	0.5446	0.7694	0.7375
Semantic density	0.7028	0.5529	0.6469	0.5046	0.6039	0.7072	0.6833	0.3890	0.6584	0.4581	0.7561	0.6963
U_{eig}	0.7566	0.5929	0.6953	0.5155	0.6312	0.7240	0.7185	0.3639	0.7075	0.5234	0.7644	0.7232
U_{deg}	0.7546	0.5921	0.6945	0.5158	0.6370	0.7288	0.7241	0.3581	0.7118	0.5274	0.7660	0.7252
U_{ecc}	0.6213	0.4987	0.6057	0.3758	0.4126	0.5376	0.5592	0.3938	0.5604	0.3491	0.6417	0.6107
LUQ	0.7354	0.5773	0.6835	0.5165	0.6423	0.7303	0.7273	0.3568	0.7171	0.5315	0.7672	0.7288
SNNE	0.7531	0.5911	0.6939	0.5164	0.6337	0.7273	0.7195	0.3581	0.7106	0.5247	0.7659	0.7245
SNNE (only-num)	0.7532	0.5912	0.6940	0.5163	0.6343	0.7279	0.7200	0.3579	0.7107	0.5250	0.7658	0.7246
SNNE (only-denom)	0.7536	0.5916	0.6942	0.5161	0.6390	0.7304	0.7255	0.3571	0.7132	0.5290	0.7668	0.7261

Table 63: Model-specific SUQ results using task-specific quality metrics (Part III). Summarization reports PRR against per-example SEAHORSE Average, and WMT19 translation reports PRR against per-example COMET. Higher is better; "–" denotes an unavailable generation artifact. The method and column order follows Table 22. (Continued.)

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
PHI 3-14B												
Discrete semantic entropy	0.7631	0.5540	0.7513	0.5487	0.3643	0.3120	0.3800	0.4269	0.5045	0.5183	0.7923	0.7476
Discrete semantic density	0.7977	0.5941	0.7591	0.5820	0.3740	0.3170	0.3823	0.4285	0.5424	0.5181	0.8022	0.7520
Discrete U_{eig}	0.7974	0.5914	0.7586	0.5897	0.3759	0.3202	0.3852	0.4289	0.5427	0.5144	0.8009	0.7511
Discrete U_{deg}	0.7934	0.5830	0.7561	0.5895	0.3751	0.3199	0.3863	0.4289	0.5450	0.5206	0.7981	0.7512
Discrete C_{ecc}	0.7984	0.5939	0.7647	0.5708	0.3693	0.3180	0.3835	0.4293	0.5035	0.4993	0.7976	0.7484
KLE-heat$_{t=0.1}$	0.7985	0.6128	0.7754	0.6134	0.3730	0.3231	0.3826	0.4317	0.5339	0.4947	0.8072	0.7611
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7986	0.6126	0.7754	0.6135	0.3728	0.3228	0.3825	0.4315	0.5338	0.4947	0.8074	0.7610
Discrete LUQ-pair	0.7802	0.5490	0.7505	0.5704	0.3712	0.3165	0.3832	0.4242	0.5083	0.4974	0.7714	0.7256
Discrete SNNE	0.7778	0.5633	0.7518	0.5664	0.3716	0.3155	0.3841	0.4268	0.4984	0.5110	0.7840	0.7370
Discrete SNNE (only-num)	0.7821	0.5719	0.7562	0.5854	0.3691	0.3137	0.3851	0.4257	0.5318	0.5266	0.8083	0.7601
Discrete SNNE (only-denom)	0.7903	0.6044	0.7847	0.6341	0.3703	0.3175	0.3810	0.4291	0.5605	0.5492	0.8086	0.7666
Semantic density	0.7617	0.5628	0.7534	0.5343	0.3676	0.3136	0.3814	0.4302	0.3089	0.4716	0.8006	0.7352
U_{eig}	0.8041	0.5985	0.7743	0.6146	0.3762	0.3203	0.3857	0.4291	0.5450	0.5186	0.8078	0.7568
U_{deg}	0.8019	0.5949	0.7730	0.6124	0.3758	0.3196	0.3857	0.4289	0.5482	0.5247	0.8075	0.7577
U_{ecc}	0.6933	0.5260	0.7317	0.3735	0.3685	0.3208	0.3815	0.4260	0.2548	0.4057	0.7078	0.6573
LUQ	0.7869	0.5737	0.7664	0.6060	0.3754	0.3188	0.3845	0.4279	0.5488	0.5296	0.8062	0.7601
SNNE	0.8006	0.5929	0.7723	0.6137	0.3775	0.3206	0.3849	0.4300	0.5485	0.5220	0.8070	0.7574
SNNE (only-num)	0.8007	0.5929	0.7723	0.6137	0.3775	0.3205	0.3849	0.4300	0.5487	0.5224	0.8074	0.7574
SNNE (only-denom)	0.8008	0.5933	0.7724	0.6115	0.3758	0.3195	0.3855	0.4287	0.5489	0.5267	0.8078	0.7582

Table 64: Consistency-based UQ under task-specific quality metrics, matched to the five populated model blocks and 12 task columns of Table 24. We report PRR against SEAHORSE Average for summarization and COMET for translation. The final two columns are unweighted mean changes relative to the ROUGE-L PRRs in Table 24 over the four summarization and eight translation columns, respectively. These cross-target differences are descriptive, not improvements in uncertainty ranking.

Uncertainty Method / PRR	Summarization (SEAHORSE Average)				Translation (COMET)								Mean Δ	
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh	Summ.	Trans.
LLAMA 3.1-8B INSTRUCT														
Exact Match	0.7354	0.3353	0.7236	0.5749	0.7909	0.8218	0.7767	0.7552	0.7003	0.7385	0.7668	0.7652	+0.3400	+0.2226
BERT similarity	0.7612	0.3198	0.7548	0.6191	0.8079	0.8269	0.8128	0.7952	0.7602	0.7861	0.7694	0.7894	+0.3431	+0.2220
Cosine Similarity	0.7593	0.3401	0.7496	0.6244	0.8059	0.8291	0.8094	0.7953	0.7568	0.7871	0.7764	0.7910	+0.3432	+0.2238
LLAMA 3.1-70B INSTRUCT														
Exact Match	0.7732	0.4359	0.7263	0.5701	0.8041	0.8364	0.7926	0.8323	0.7263	0.7627	0.7936	0.7477	+0.3524	+0.1956
BERT similarity	0.8073	0.3718	0.7407	0.5739	0.8154	0.8394	0.8218	0.8565	0.7735	0.7914	0.7910	0.7712	+0.3427	+0.1797
Cosine Similarity	0.7988	0.4093	0.7479	0.6054	0.8158	0.8446	0.8226	0.8625	0.7741	0.7990	0.7991	0.7758	+0.3461	+0.1839
LLAMA 3.1-405B INSTRUCT FP8														
Exact Match	0.7671	0.2842	0.7573	0.5853	0.8050	0.8620	0.7927	0.8475	0.7195	0.8016	0.7932	0.7736	+0.2961	+0.1743
BERT similarity	0.7931	0.2997	0.7902	0.6018	0.8139	0.8645	0.8125	0.8601	0.7548	0.8058	0.7933	0.7823	+0.2983	+0.1593
Cosine Similarity	0.7863	0.3042	0.7954	0.6271	0.8150	0.8678	0.8113	0.8652	0.7530	0.8080	0.7988	0.7862	+0.3003	+0.1641
QWEN 3-1.7B														
Exact Match	0.7931	0.6201	0.7579	0.5863	0.7207	0.7748	0.6651	0.7770	0.6845	0.6563	0.8111	0.7381	+0.4036	+0.2525
BERT similarity	0.8165	0.6318	0.7811	0.5886	0.7443	0.7905	0.7160	0.8027	0.7293	0.6874	0.8194	0.7579	+0.4088	+0.2530
Cosine Similarity	0.8145	0.6418	0.7824	0.6247	0.7454	0.7905	0.7084	0.7998	0.7251	0.6856	0.8207	0.7558	+0.4153	+0.2529
QWEN 3-8B														
Exact Match	0.7977	0.6677	0.7698	0.5968	0.7782	0.8746	0.7649	0.8394	0.7155	0.7383	0.7741	0.7715	+0.4345	+0.2065
BERT similarity	0.8170	0.6817	0.7934	0.6203	0.7894	0.8769	0.7884	0.8459	0.7442	0.7458	0.7734	0.7823	+0.4409	+0.2046
Cosine Similarity	0.8174	0.6883	0.7928	0.6457	0.7889	0.8788	0.7844	0.8467	0.7406	0.7464	0.7739	0.7811	+0.4467	+0.2054

Table 65: Task-specific PRR for LLAMA 3.1-405B INSTRUCT FP8. The table combines the 19 SUQ measures with Exact Match, BERT similarity, and Cosine Similarity on the same four summarization datasets and eight WMT19 directions. Summarization uses SEAHORSE Average and translation uses COMET as per-example quality; consistency confidence is converted to uncertainty as $1 - \text{confidence}$. Higher is better, and bold marks the best of the 22 methods in each column.

Uncertainty Method	Summarization (SEAHORSE Average PRR)				WMT19 Translation (COMET PRR)							
	XSUM	AESLC	CNN/DM	Multi-News	cs	de	fi	gu	kk	lt	ru	zh
Discrete semantic entropy	0.7693	0.2984	0.7635	0.5880	0.7946	0.8496	0.7961	0.8528	0.7378	0.7844	0.7739	0.7808
Discrete semantic density	0.7748	0.3019	0.7632	0.5867	0.8021	0.8609	0.7977	0.8590	0.7418	0.8038	0.7876	0.7867
Discrete U_{eig}	0.7758	0.3031	0.7677	0.5898	0.7992	0.8532	0.7955	0.8561	0.7447	0.7976	0.7836	0.7850
Discrete U_{deg}	0.7753	0.3027	0.7643	0.5851	0.7935	0.8530	0.7925	0.8502	0.7431	0.7973	0.7858	0.7856
Discrete C_{ecc}	0.7703	0.3013	0.7605	0.5810	0.7922	0.8525	0.7934	0.8516	0.7378	0.7955	0.7831	0.7840
KLE-heat$_{t=0.1}$	0.7713	0.3061	0.7856	0.6361	0.8068	0.8685	0.8049	0.8638	0.7495	0.8057	0.8041	**0.7936**
KLE-Matérn$_{\kappa=1,\,\nu=1}$	0.7714	**0.3063**	0.7856	**0.6363**	0.8068	0.8685	0.8049	0.8638	0.7494	0.8057	0.8032	0.7936
Discrete LUQ-pair	0.7694	0.2984	0.7603	0.5816	0.7782	0.8399	0.7863	0.8341	0.7266	0.7782	0.7715	0.7599
Discrete SNNE	0.7713	0.2957	0.7638	0.5824	0.7845	0.8435	0.7907	0.8406	0.7278	0.7815	0.7735	0.7583
Discrete SNNE (only-num)	0.7765	0.3004	0.7700	0.5924	0.8117	0.8661	0.8074	**0.8681**	0.7503	0.8079	0.7968	0.7836
Discrete SNNE (only-denom)	0.7877	0.2989	0.7916	0.6360	0.8109	0.8647	0.8056	0.8666	0.7516	0.8071	0.7954	0.7818
Semantic density	0.7666	0.2891	0.7575	0.5853	0.8086	0.8654	0.7952	0.8575	0.7248	0.8040	0.7982	0.7783
U_{eig}	0.7756	0.3036	0.7826	0.6314	0.8143	**0.8725**	0.8042	0.8665	0.7471	0.8071	**0.8041**	0.7929
U_{deg}	0.7747	0.3027	0.7810	0.6299	0.8132	0.8724	0.8054	0.8665	0.7469	0.8072	0.8033	0.7932
U_{ecc}	0.7469	0.2643	0.7405	0.5330	0.7439	0.8002	0.7443	0.7686	0.6747	0.7337	0.7309	0.7274
LUQ	0.7700	0.3021	0.7769	0.6265	0.8117	0.8681	0.8049	0.8659	0.7456	**0.8087**	0.7991	0.7914
SNNE	0.7741	0.3028	0.7804	0.6296	0.8139	0.8724	0.8052	0.8663	0.7473	0.8066	0.8028	0.7934
SNNE (only-num)	0.7741	0.3027	0.7805	0.6296	0.8139	0.8724	0.8050	0.8661	0.7469	0.8065	0.8028	0.7927
SNNE (only-denom)	0.7742	0.3023	0.7806	0.6296	0.8135	0.8724	0.8043	0.8666	0.7478	0.8066	0.8036	0.7934
Exact Match	0.7671	0.2842	0.7573	0.5853	0.8050	0.8620	0.7927	0.8475	0.7195	0.8016	0.7932	0.7736
BERT similarity	**0.7931**	0.2997	0.7902	0.6018	0.8139	0.8645	**0.8125**	0.8601	**0.7548**	0.8058	0.7933	0.7823
Cosine Similarity	0.7863	0.3042	**0.7954**	0.6271	**0.8150**	0.8678	0.8113	0.8652	0.7530	0.8080	0.7988	0.7862