ACCOUNTING FOR THE UNSEEN: FRAGILITY-AWARE RECALIBRATION OF SEMANTIC UNCERTAINTY FOR HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite their impressive ability to generate natural and coherent text, large language models (LLMs) can produce plausible but fabricated responses. The same prompt, when repeatedly applied, may yield different answers, undermining the confidence one may place in an individual generation. The common approach of detecting such confabulations by observing multiple generations to estimate the uncertainty over their semantic content must necessarily rely on a finite sample of all generations that the model may produce. Such estimates can be fragile when rare or singleton answers dominate observed semantic support. To address this, we propose Semantic Uncertainty Recalibration accounting for Fragility (SURF), a framework that corrects semantic uncertainty estimates using the *probability* of observed singletons as a proxy for unobserved semantic support. SURF incorporates this fragility proxy into an entropy-based optimization that recalibrates token-sequence probabilities according to their sensitivity, yielding uncertainty estimates that are more robust against finite observations. We also provide an estimate of the *number* of semantically distinct answers as an actionable signal for when additional observations or human intervention may be warranted. When evaluated across 112 different settings (involving four datasets, seven model variants, and different generation lengths and model quantization levels), SURF shows consistent improvement in AUROC and headroom gain (HRG), and maintains stable AURAC performance, which demonstrates its reliability even under resource constraints.

1 INTRODUCTION

Large language models (LLMs) are increasingly serving as the foundation for *agentic AI* systems, where models autonomously reason, plan, and interact with external tools to perform complex tasks. Beyond traditional text generation, LLMs now underpin applications such as automated web agents, coding assistants, and multi-step decision-making systems. Trained on large-scale corpora, they exhibit strong capabilities in natural language understanding and generation, fundamentally reshaping human interaction with intelligent systems across domains (Hosseini et al., 2023; Lai et al., 2023).

Despite these advances, the reliability of LLM outputs remains a critical concern. Errors and inconsistencies in generated responses can propagate downstream, particularly in interactive settings where outputs are assumed to be both fluent and factually grounded (Lei et al., 2023). This is particularly concerning in high-stakes safety-critical applications such as healthcare, defense, law, journalism, and autonomous systems, where incorrect outputs can lead to significant practical and ethical consequences. A central cause for failure in this context is *hallucinations,* wherein models generate responses that are syntactically plausible but factually incorrect, unverifiable, or fabricated.

Methods of Hallucination detection rely on finite model outputs that have been seen, which constitute only a finite sample from the space of all possible generations. This finite sample cannot capture every semantically distinct response the model might produce leaving plausible generations unobserved and missing. Moreover, even among the observed outputs, those that are rarely observed can substantially alter the uncertainty measures that guide hallucination detection schemes. We refer to this *finite sampling/observations* effect, viz., sensitivity to the composition of a finite sample/observation set, particularly to rare observations, as *fragility.*

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

Semantic Uncertainty Recalibration accounting for Fragility (SURF) is a framework that mitigates fragility by correcting for semantic support that is yet unseen. It uses the probability mass of observed singleton semantic clusters as a proxy for the mass of unobserved semantic support to recalibrate semantic uncertainty estimates. A high prevalence of singleton clusters hints at incomplete coverage and a fragile distribution which may substantially alter the uncertainty estimates as more responses are observed. This effect is particularly pronounced in heavy-tailed response distributions of LLMs, where considerable probability mass may be distributed across rare or previously unseen responses.

1.1 Background and Related Work

Existing approaches to detect hallucinations broadly fall into two categories. The first, which relies on supervised or auxiliary models trained to identify hallucinations (Azaria & Mitchell, 2023; Quevedo et al., 2024; Su et al., 2024), often require additional computational cost and does not generalize well. The second line of work, which estimates uncertainty directly from multiple stochastic generations of the same prompt (Lin et al., 2022; Chen et al., 2024; Manakul et al., 2023), replies on the premise that higher output variability signals a higher hallucination likelihood. Bayesian approaches have also been explored, but they remain difficult to calibrate for LLMs and often fail to reliably reflect factual correctness (Kang et al., 2025).

A primary measure of uncertainty in uncertainty-based methods is entropy. Token-level entropy offers a natural and interpretable measure (Kadavath et al., 2022), but it fails to account for semantic equivalence between different surface realizations (Xiao et al., 2020). *Semantic entropy (SE),* where outputs are grouped based on meaning rather than lexical similarity, attempts to address this limitation (Farquhar et al., 2024; Kossen et al., 2024). Smoother similarity measures and probabilistic weighting schemes build on this idea to improve robustness (e.g., Kernelized Likelihood Entropy (KLE) (Nikitin et al., 2024), Semantic Nearest Neighbour Entropy (SNNE) (Nguyen et al., 2025), and Semantic Density (SD) (Qiu & Miikkulainen, 2024)). Other approaches explore perturbation-based strategies (Gao et al., 2024; Cox et al., 2025), which probe sensitivity to input variations, and structure-based methods that use knowledge graphs (Jiang et al., 2024; Haskins & Adams, 2025). However, all these methods suffer from the finite sampling/observation effect and do not explicitly address fragility.

Most available approaches also focus on *global* uncertainty measures and overlook how token-sequence (TS) probabilities respond to *local* perturbations. Recent work has shown that sensitivity-aware measures can better capture such *local* instabilities and improve hallucination detection (Vipulanandan et al., 2026). But, these methods suffer from the same finite sampling/observation effect. By failing to account for fragility, entropy estimates can become unreliable when observations yield semantically distinct low-frequency answers. This behavior is closely related to the notion of *missing-mass,* the probability mass associated with unseen responses, where persistent emergence of novel answers indicates a fragile distribution and elevated hallucination risk (Lee & Boehme, 2025).

The problem of estimating unseen elements from limited sampling/observations has been extensively studied in the context of unseen-species and missing-mass estimation (Good, 1953; Orlitsky et al., 2016; Hao & Li, 2020). Classical estimators rely heavily on rare events and can exhibit high variance in small-sample regimes. More recent work shows that missing-mass can be largely inferred from observed frequency or occupancy statistics, with minimal dependence on the underlying distribution (Lee & Boehme, 2025). But these estimators are often fragile in that small changes in sample size can lead to large fluctuations in the estimates (Weltz et al., 2024; Schmitz & Rahmann, 2025).

1.2 Our Contributions

We therefore approach hallucination detection from a finite sample/observations perspective and seek uncertainty estimates that capture both observed variability and fragility. SURF operationalizes this perspective by treating repeated generations as samples from an underlying semantic response distribution. Instead of using token-level logits to speculate on the unobserved, SURF uses the probability mass of singleton semantic clusters as a signal of incomplete semantic support and hence fragility. It is this occupancy-based fragility signal that SURF combines with the local sensitivity of the observed TS probabilities to recalibrate semantic uncertainty.

Our work thus identifies and integrates two complementary sources of uncertainty: the local sensitivity of probability estimates of observed generations and the fragility arising from incomplete coverage

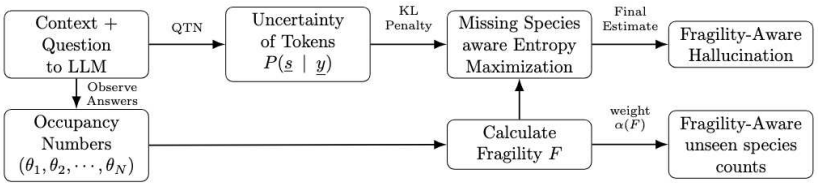


Figure 1: Overview of the SURF hallucination-detection pipeline. Generated sequences are clustered by bidirectional entailment; TS-level UQ characterizes observed TS probability sensitivity, while observed singleton mass provides a fragility proxy for unobserved semantic support. Both signals enter the entropy-regularized probability adjustment before semantic uncertainty is evaluated.

of the underlying semantic support. To the best of our knowledge, SURF is the first framework to connect fragility with semantic uncertainty quantification for hallucination detection in LLMs. The main contributions and key advantages of our work are the following:

(a) We introduce a novel occupancy-based *fragility index* that is linked to the probability mass of singleton semantic clusters. It quantifies the sensitivity of the empirical answer distribution to additional sampling, thus providing a principled proxy for detecting when sampled outputs are unreliable due to incomplete coverage.

(b) We develop closed-form formulations of unseen-mass and unseen-species count estimates which account for fragility under practical sampling regimes, and validate these estimates through extensive experiments on both synthetic and real-world datasets (App. F). The results demonstrate strong agreement between theoretical predictions and empirical behavior, highlighting the reliability of the proposed estimators.

(c) We introduce SURF (Semantic Uncertainty Recalibration accounting for Fragility), an interpretable fragility-aware framework for hallucination detection. It uses the newly introduced fragility index as a proxy for unobserved support, allowing semantic uncertainty to account for finite sampling without explicitly estimating unseen answers.

(d) We further show how the estimated number of unseen semantics (i.e., the 'unseen-species' count) can serve as a diagnostic, indicating when further sampling are likely to reveal additional semantically distinct answers.

Fig. 1 depicts overall pipeline of SURF. Extensive evaluations across datasets, model families, and quantization levels demonstrate consistent gains over the evaluated baselines, including in resource-constrained settings. An ablation study (Sec. 4.3) isolates the contributions of observed TS probability sensitivity and fragility.

2 SURF: Semantic Uncertainty Recalibration Accounting for Fragility

2.1 Preliminaries

Token-Sequence (TS) Probability. Let $\underline{s} = (s_1, \ldots, s_T)$ be a TS generated by an LLM in response to input $\underline{y}$. Under standard autoregressive factorization, the likelihood of $\underline{s}$ conditioned on $\underline{y}$ is

$$p_s \triangleq P(\underline{s} \mid \underline{y}) = \prod_{i=1}^{T} P(s_i \mid \underline{s}_{<i}, \underline{y}), \tag{1}$$

where $\underline{s}_{<i} = (s_1, \ldots, s_{i-1})$ denotes the token prefix. Let $\{p_s^{(1)}, \ldots, p_s^{(R)}\}$ be the TS probabilities associated with the generations all produced in response to the same input $\underline{y}$ applied R times.

The physics-inspired uncertainty quantification (UQ) framework in Principe (2010) calls for a non-parametric feature mapping of these sampled TS probabilities, specifically its kernel mean embedding (KME) in a kernel-specified reproducing kernel Hilbert space(RKHS) (Aronszajn, 1950; Schölkopf et al., 2015). This KME can then be interpreted as a data-dependent wave function. Vipulanandan et al. (2026) then utilizes the first-order corrections of an operator-theoretic perturbation analysis of

this KME to construct a local measure of sensitivity $\mathrm{UQ}(p_s^{(r)}),\ r = 1, \ldots, R$, associated with the probability of each TS (more details appear in App. B).

Semantic Cluster Probability. To account for semantic variability across generations, as in Farquhar et al. (2024) and Kossen et al. (2024), the set $\{p_s^{(1)}, \ldots, p_s^{(R)}\}$ of generated TSs is partitioned into semantically equivalent clusters as determined by bidirectional entailment (implemented via DeBERTa). This yields a set $\mathcal{Z} = \{z^{(1)}, \ldots, z^{(C)}\}$ of C semantic clusters; each cluster represents a set of semantically consistent generations.

Treating the cluster label as the random variable Z over the space of $\mathcal{Z}$, its probability distribution (conditioned on input $\underline{y}$) and the Shannon SE it induces are, respectively,

$$p_c^{(\ell)} \triangleq P(Z = z^{(\ell)} \mid \underline{y}) = \frac{\sum_{\underline{s} \in z^{(\ell)}} P(\underline{s} \mid \underline{y})}{\sum_{\ell=1}^{C} \sum_{\underline{s} \in z^{(\ell)}} P(\underline{s} \mid \underline{y})};\ \ \mathrm{SE}_S(Z \mid \underline{y}) = -\sum_{\ell=1}^{C} p(z^{(\ell)})\, \log p(z^{(\ell)}). \tag{2}$$

When token-level probabilities $P(s_i \mid \underline{s}_{<i}, \underline{y})$ are available, the cluster probabilities $p_c(z^{(\ell)})$ are computed via (1); otherwise, as Farquhar et al. (2024) do, they are approximated using normalized cluster counts as $p_c(z^{(\ell)}) \approx \mathbb{1}(z = z^{(\ell)})/C$, where $\mathbb{1}(\cdot)$ is the indicator function.

Sensitivity-Aware Recalibration. Instead of the Shannon SE-based method in Farquhar et al. (2024) and Kossen et al. (2024), Vipulanandan et al. (2026) rely on the Rényi SE

$$\mathrm{SE}_R(Z \mid \underline{y}) = -\log \sum_{\ell=1}^{C} \left(p(z^{(\ell)})\right)^2 = -\log\left(\mathbb{E}[p(^{(\ell)})]\right) \tag{3}$$

of the semantic cluster distribution. Here, $\mathbb{E}[\cdot]$ denotes expectation (over the space of semantic classes). They then employ the sensitivity measure $\mathrm{UQ}(p_s^{(r)})$ for recalibration of TS (and hence semantic cluster) probabilities: a larger $\mathrm{UQ}(p_s^{(r)})$ value indicates that the probability of the observed TS is more locally sensitive and should therefore receive a weaker fidelity constraint during recalibration. This method, which we will refer to as *sensitivity-aware Rényi SE method,* improves hallucination detection relative to prior methods and remains robust across short- and long-form responses as well as different quantization levels.

However, this sensitivity-aware hallucination detection baseline captures the probability sensitivity of only the observed TSs, but not the fragility of the sampled semantic support under further sampling. Indeed, neither Vipulanandan et al. (2026), nor Farquhar et al. (2024) or Kossen et al. (2024), account for unseen generations arising from finite sampling (i.e., finiteness of R), where missing responses may carry non-negligible probability mass and alter the entropy estimate. This limitation calls for a method that explicitly incorporates unseen mass, the focus of our work.

2.2 Fragility Under Finite Samples/Observations

Finite Observations. In generative models such as LLMs, finite observations provides only a partial view of the underlying response distribution, leaving a potentially non-trivial probability mass in unseen but plausible generations. Thus, entropy estimated from observations can underestimate uncertainty. In factual question-answer (QA) settings, prompts generate answers that are typically concentrated on a small set of semantically consistent clusters. Persistent emergence of novel rare semantic clusters signals the response distribution's instability and can be associated with hallucination-prone behavior.

It is important to distinguish this from observed TS probability sensitivity. The TS-level UQ score captures the local sensitivity of the probability assigned to each observed generation. Under limited or finite observations, the inclusion or exclusion of low-probability but semantically distinct answers can substantially alter cluster-level probabilities and entropy estimates, particularly in the heavy-tailed distributions typical of LLMs. Therefore, we posit that sensitivity measured solely on observed TS probabilities is insufficient.

Fragility. In contrast, *fragility* refers to the sensitivity of the *empirical semantic distribution* to finite observations. Even when observed TS probabilities are stable, the observed generations may omit or even over-represent semantic clusters, leading to instability in the induced semantic distribution. Thus,

TS-level UQ concerns local probability sensitivity associated with observed answers, whereas fragility concerns incomplete coverage of the response distribution under finite sampling and the instability of the empirical distribution that it gives rise to, particularly when the observed responses are dominated by low-frequency events. This characterization highlights a key limitation of probability-sensitivity measures defined solely over observed responses. Even when observed TS probabilities are locally stable, incomplete semantic coverage can make the induced distribution and its entropy sensitive to further observations.

What is a metric that captures this fragility? Let $\{p_\ell\}$ denote the (unknown) probability of the semantic cluster $c_\ell \in \mathcal{C}$. Given n i.i.d. samples, let $N_\ell = \sum_{i=1}^{n} \mathbb{1}(X_i = s_\ell)$ denote the observed counts and $\Phi_k = \sum_\ell \mathbb{1}(N_\ell = k)$ the occupancy numbers. Define the observed probability mass and the mass of k-tons as

$$P_{\text{obs}} = \sum_\ell p_\ell \, \mathbb{1}(N_\ell > 0); \qquad P_k = \sum_\ell p_\ell \, \mathbb{1}(N_\ell = k). \tag{4}$$

The missing mass $M_0 = 1 - P_{\text{obs}}$ represents the total probability of unseen semantic clusters (Lee & Boehme, 2025). We then define fragility as the proportion of probability mass concentrated in singletons relative to the observed mass:

$$F_{\text{mass}} = P_1 / P_{\text{obs}}, \tag{5}$$

where P_1 denotes the probability mass of singleton (observed exactly once) clusters. This measure directly characterizes the reliability of the observed support: higher F_{mass} values indicate that the observed distribution is dominated by rare events, implying limited information about the underlying distribution and increased sensitivity of semantic distribution to additional sampling. This means that small perturbations in sampling (e.g., the inclusion or exclusion of a few rare generations) can induce large changes in the empirical distribution and downstream entropy estimates.

SURF, the method being proposed, does not presume to recover exactly the missing mass M_0 of unseen clusters or their identities from the observed sample. Instead, SURF interprets high observed singleton mass as evidence of incomplete coverage and uses fragility as a proxy for unobserved support and the associated finite-coverage risk by augmenting TS-level sensitivity $\text{UQ}(p_s^{(r)})$ with F_{mass} in (5).

2.3 Objective Function

To account for sampling-induced fragility, SURF combines observed TS probability sensitivity with a fragility-aware correction to obtain more reliable estimates of semantic uncertainty under finite sampling. Towards this end, it recalibrates TS probabilities as

$$p_s^{(r)^\dagger} = \arg\max_{\widehat{p}_s^{(r)}} \Bigg\{ -\log\left(\widehat{p}_s^{(r)^2} + (1 - \widehat{p}_s^{(r)})^2\right) - \lambda \cdot \frac{1}{1 + \text{UQ}(p_s^{(r)})} \cdot D_{\text{KL}}(\widehat{p}_s^{(r)} \| p_s^{(r)}) - \gamma \cdot F_{\text{mass}} \cdot \widehat{p}_s^{(r)} \Bigg\}, \tag{6}$$

which constitutes an entropy-regularized optimization that incorporates two distinct signals: TS-level sensitivity $\text{UQ}(p_s^{(r)})$ which reflects local sensitivity of observed TS probabilities and fragility F_{mass} in (5) which captures instability of the empirical distribution due to incomplete sampling, particularly when observed mass is dominated by singleton events.

Interpretation. We now interpret and explain the role of each term in (6).

1. The quadratic Rényi entropy term discourages extreme probability assignments. The parameter $\lambda > 0$ controls fidelity to the observed TS probability by scaling the KL-divergence penalty $D_{\text{KL}}(\hat{p}_s^{(r)} \| p_s^{(r)})$. Larger λ heavily penalizes deviations from $p_s^{(r)}$, thus keeping the adjusted probability closer to its observed value.
2. $D_{\text{KL}}(\cdot \| \cdot)$ is the Kullback-Liebler (KL) divergence. It preserves fidelity to the empirical TS probabilities. The more the adjusted probability $\widehat{p}_s^{(r)}$ moves away from the observed probability $p_s^{(r)}$, the higher the penalty.

3. As the TS probability sensitivity measure $\mathrm{UQ}(p_s^{(r)})$ increases, the coefficient $1/(1+\mathrm{UQ}(p_s^{(r)}))$ decreases, weakening the KL fidelity penalty because the observed TS probability is less locally reliable. (The additive constant prevents a singular denominator to keep this modulation bounded.) Thus, this strategy allows locally sensitive samples to deviate more from their empirical values.
4. The higher the F_{mass} value (which implies that the observed support is singleton-dominated), the more the correction is boosted. $\gamma > 0$ controls the strength of this fragility-aware regularization. So, this strategy allows fragility, which is employed as as a proxy for unobserved support, to act as a global correction that penalizes overconfident assignments when singleton mass indicates incomplete coverage.

SURF does not claim to identify unseen answers or recover their exact probability mass. Rather, it combines observed TS probability sensitivity with a distinct finite-coverage signal. It does not instantiate unseen clusters; rather, it adjusts the concentration of the observed distribution in proportion to an empirical proxy for unobserved support. In particular, when fragility is high—indicating that the observed distribution is dominated by singleton mass—the optimization applies a stronger correction to overconfident observed probability assignments, yielding a semantic uncertainty score that is less dependent on an apparently complete finite sample.

The same occupancy statistics can also estimate the number and probability mass of semantic answers likely to remain unseen under finite sampling. As these estimators are not required to define the SURF semantic uncertainty score, their formulation, analysis, and dedicated experiments are relegated to App. A which also provides explicit expressions for missing-mass and unseen-count estimators.

3 Experiments

Datasets. Our evaluation spans multiple QA settings, including trivia-based knowledge (TriviaQA in Joshi et al. (2017)), general reading comprehension (SQuAD 1.1 in Stanford NLP Group (2018)), and open-domain question answering (NQ-Open in Lee et al. (2019)) derived from real-world Google search queries (Kwiatkowski et al., 2019). We further include mathematical reasoning tasks using SVAMP (Patel et al., 2021).

Models and Evaluation Settings. We evaluate seven open-weight model variants obtained via HuggingFace: LLaMA-3.2 1B (Grattafiori et al., 2024); LLaMA-2 7B and 13B, each in base and chat form (Touvron et al., 2023); and Mistral-7B in base and instruct form (Jiang, 2024). We draw 28 configurations from the seven model variants, precisions in $\{16, 8, 4\}$ bit, and brief or single-sentence prompting (not every possible combination was run). These 28 realized model/precision/prompt configurations across the four datasets give us exactly 112 evaluation settings. We intentionally include smaller-scale and quantized models because they enable extensive repeated sampling and provide a controlled setting for studying semantic uncertainty under limited model capacity and reduced precision. All evaluations[1] were conducted on a workstation equipped with 64 GB RAM and an NVIDIA A6000 GPU. The complete configuration and sampling protocol appear in App. D.1.

Sampling, Prompting, and Entailment. For every question, one sampled answer at temperature 0.1 is graded, while ten additional answers sampled at temperature 1.0 are semantically clustered. All calls use sampling (`do_sample=True`); the temperature 0.1 answer is not greedy. App. D provides the prompt templates and complete protocol. Semantic clustering uses `deberta-v2-xlarge-mnli` with the question prepended to each generation. Two answers are grouped only when the hard argmax MNLI class is entailed in both directions; no entailment-probability threshold is used. Assignment is greedy and sequential. This dedicated model provides a fixed and reproducible clustering rule across all experiments.

Benchmarks. For evaluation, we compare the proposed SURF approach against several established baselines: naive entropy (NE), Shannon SE (SE_S), and discrete Shannon SE (DSE_S) from Farquhar et al. (2024), and sensitivity-aware Rényi SE (SE_{RU}) from Vipulanandan et al. (2026), plus two strong supervised and prompting-based baselines, viz., embedding regression (ER) and verifier $p\,(\mathrm{True})$ (Kadavath et al., 2022). ER follows the P(IK) paradigm, but instead of fine-tuning the full model, we train a logistic regression classifier on final-layer representations to predict answer correctness, improving simplicity and reproducibility. While effective in-distribution, its performance degrades under distribution shift. The $p\,(\mathrm{True})$ method estimates confidence by prompting the model

[1] Code available at: https://anonymous.4open.science/r/Q-FASU-8F60/README.md

Cluster	# of TSs	$p_s^{(r)}$	$p_c^{(\ell)}$	$p_c^{(\ell)^*}$	$p_c^{(\ell)^\dagger}$	NE	SE_S	SE_{RU}	$\mathrm{SE}_{RU}^\dagger$
1. Russia	1	0.01814	0.01814	0.02223	0.02313	-0.03159	-0.03159	0.00049	0.00053
2. S. Arabia	5	0.17765	0.88824	0.85880	0.85480	-0.13331	-0.04572	0.73754	0.73068
3. Iran	1	0.02223	0.02223	0.02697	0.02807	-0.03674	-0.03674	0.00073	0.00079
4. Kuwait	1	0.02749	0.02749	0.03488	0.03558	-0.04291	-0.04291	0.00122	0.00127
5. Qatar	1	0.00672	0.00672	0.01214	0.01264	-0.01460	-0.01460	0.00015	0.00016
6. Iraq	1	0.03717	0.03717	0.04498	0.04578	-0.05315	-0.05315	0.00202	0.00210
Total	10	1.00000	1.00000	1.00000	1.00000	**0.84557**	**0.22471**	**0.12951**	**0.13340**

Table 1: An example calculation of TS probability and entropies for the query in Sec. 4.1. Values for only one representative TS (or LLM generation) per cluster is shown. *Columns:* $p_s^{(r)}$ are the normalized TS probabilities; $p_c^{(\ell)}$ are the normalized cluster probabilities; $p_c^{(\ell)^*}$ are the cluster probabilities after integrating sensitivity of observed TS probability; $p_c^{(\ell)^\dagger}$ are the cluster probabilities after applying the F_{mass} fragility correction. *Methods:* NE, SE_S, and SE_{RU} denote naive entropy, semantic entropy, and sensitivity-aware Rényi SE methods, respectively (Farquhar et al., 2024; Vipulanandan et al., 2026); $\mathrm{SE}_{RU}^\dagger$ denotes the newly proposed SURF-generated recalibration.

to judge whether its top answer is "True" or "False" based on multiple sampled candidates, optionally enhanced with few-shot examples (up to 20), subject to context length constraints. Unseen species comparison methods appear in App. F.

Evaluation Metrics. We use four primary metrics for purposes of evaluation:

(a) *AUROC (Area Under Receiver Operating Characteristic)* evaluates a classifier's ability to discriminate between correct and incorrect predictions. It can be interpreted as the probability that a randomly chosen correct instance is assigned a higher confidence score than a randomly chosen incorrect one. An AUROC of 1 indicates perfect separability (App. C.2 has detailed results).

(b) *Headroom Gain (HRG),* computed on AUROC, measures the relative improvement of one method over another by accounting for the remaining performance headroom of the baseline (Knights, 2026). It provides a normalized view of gains across different regimes (see App. C.3).

(c) *RAC (Rejection Accuracy Curve)* evaluates prediction accuracy as a function of confidence by retaining only a fraction of the most confident outputs. A well-calibrated uncertainty measure should yield increasing accuracy as low-confidence predictions are rejected (see App. C.4.1).

(d) *AURAC (Area Under Rejection Accuracy Curve)* summarizes RAC by measuring how accuracy evolves as uncertain predictions are removed. Higher values indicate more effective prioritization of correct predictions. Unlike AUROC, AURAC depends on the underlying predictive accuracy, offering a complementary perspective (see App. C.4).

4 Results

4.1 SURF: A Worked Example

This section presents the same worked example from Vipulanandan et al. (2026). The query used for this illustration is, *"Which oil producer is a close ally of the United States?"* repeated ten times. The corresponding TS probabilities are in Table 1.

The results demonstrate that incorporating the fragility correction into SE (i.e., $\mathrm{SE}_{RU}^\dagger$) yields a lower entropy estimate than both NE and the standard Shannon formulation SE_S. This reduction arises because semantically equivalent generations are clustered together, while observed TS probability sensitivity and singleton mass modulate the adjusted TS probabilities. Notably, $\mathrm{SE}_{RU}^\dagger$ is slightly higher than SE_{RU} in this example. This is the effect of a stronger correction when the observed support is singleton-dominated, not the direct assignment of probability to identified unseen clusters.

4.2 SURF-Based Confabulation Detection

This section presents results on the TriviaQA, NQ, SVAMP, and SQuAD across the seven model variants and 28 realized model/precision/prompt configurations described in Section 3. Fig. 2 sum-

Table 2: Component-level ablation across 112 paired evaluation settings. win-rate and HRG are computed relative to SE_R.

Method	Entropy	TS Prob. Sensitivity	Fragility	Mean AUROC ↑	Std ↓	Win-Rate ↑	Mean HRG ↑
SE_R	✓	–	–	0.7695	0.0572	–	0.0000
SE_{RU}	✓	✓	–	0.7732	**0.0556**	0.8795	0.0610
SE_R + fragility	✓	–	✓	0.7775	0.0562	0.8750	0.1477
SURF ($\mathrm{SE}^{\dagger}_{RU}$)	✓	✓	✓	**0.7784**	0.0558	**0.8929**	**0.1591**

marizes the evaluated hallucination-detection methods using two complementary metrics: pairwise win-rate and HRG computed over AUROC.

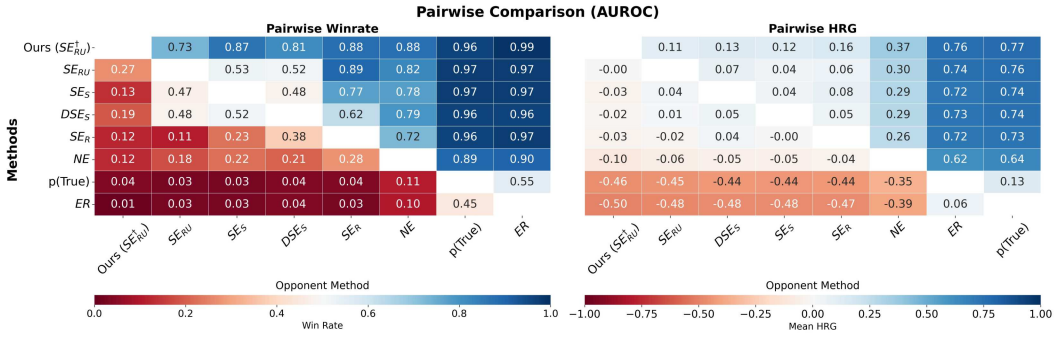


Figure 2: Pairwise win-rate (left) and HRG (right) matrices across SOTA hallucination detection methods on diverse datasets and LLM models. This summarizes **112 paired evaluation settings**. Each cell in the left heatmap shows the fraction of evaluation settings where the method on the row achieves higher AUROC than the method on the column (win-rate); the right heatmap shows the corresponding normalized performance gain (HRG). SURF generally outperforms baselines, even surpassing methods reliant on supervised learning. Additional details appear in App. C.

The results show that SURF generally outperforms SOTA baselines across all comparisons. The win-rate matrix provides a direct indicator of performance superiority, measuring how often one method achieves a higher AUROC than another across evaluation settings. Under this criterion, SURF achieves the highest win-rate against competing methods, including strong baselines such as SE_s, DSE_S, and SE_{RU}.

While win-rate measures how often a method outperforms another, HRG quantifies the magnitude of improvement relative to the available performance headroom. Again, SURF consistently achieves positive HRG values across pairwise comparisons, indicating both frequent and substantial gains. Additional results appear in App. C, together with hyper parameter selection in App. C.6.

4.3 Component Ablation

To isolate the contributions that TS probability sensitivity and fragility play in recalibration, Table 2 compares four configurations across all 112 paired evaluation settings. The base Rényi SE formulation (SE_R) contains entropy only; SE_{RU} incoporates TS probability sensitivity; the fragility-only variant adds fragility without TS probability sensitivity; and SURF ($\mathrm{SE}^{\dagger}_{RU}$) combines all three components. Results are summarized using mean AUROC, standard deviation, pairwise win-rate, and mean HRG.

Relative to SE_R, incorporating TS probability sensitivity alone increases mean AUROC from 0.7695 to 0.7732; incorporating fragility alone increases it to 0.7775. These single-component ablations show that both observed TS probability sensitivity and the singleton-mass proxy contribute information beyond entropy alone. Combining both components in SURF yields the highest mean AUROC (0.7784) and mean HRG (0.1591). All four configurations have comparable standard deviations, suggesting that the aggregate gains are not accompanied by increased variability across datasets, model families, and quantization levels.

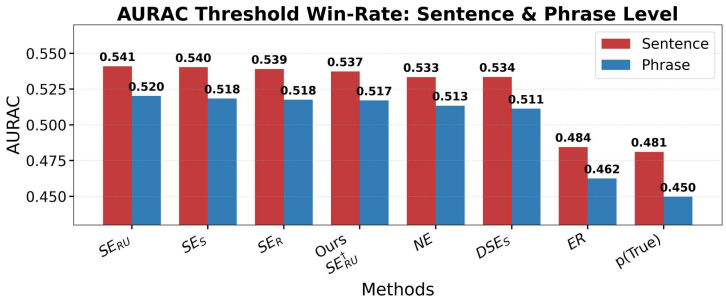


Figure 3: Overall AURAC comparison across hallucination-detection methods over **112 paired evaluation settings**, summarized across sentence- and phrase-level generations.

Paired Statistical Analysis. Across the 112 paired evaluation settings, SURF improves over Shannon SE (SE_S) by 0.0069 ± 0.0009 AUROC (99/112 positive differences; two-sided Wilcoxon $p = 1.5 \times 10^{-12}$) and over Rényi SE (SE_R) by 0.0089 ± 0.0010 (98/112; $p = 2.2 \times 10^{-13}$). The paired scatter plots and resampling diagnostics appear in App. C.1.

Fig. 3 summarizes selective-rejection performance across sentence- and phrase-level hallucination-detection experiments. SE-based methods consistently outperform probability-only baselines such as $p(\mathrm{True})$ and ER, highlighting the importance of semantic-level uncertainty estimation for reliable rejection behavior. SURF ($\mathrm{SE}^{\dagger}_{RU}$) remains highly competitive overall, although the results indicate a trade-off between maximizing AUROC and selective-rejection performance. Detailed sentence- and phrase-level AURAC breakdowns appear in App. C.4.

Theoretical properties of fragility appear in App. A; detailed AUROC and HRG evaluations, AURAC and RAC results, and insights into how quantization affects hallucinations are in App. C; implementation details and pseudocode are in App. E; and tests of reliability are in App. F.

5 Conclusion

We introduces a fragility-aware framework for hallucination detection in LLMs that extends Rényi SE to account for sampling-induced instability. Rather than claiming to recover unseen answers directly, SURF uses the probability mass of observed singleton semantic clusters as a proxy for unobserved support and combines it with observed TS probability sensitivity in an entropy-regularized correction.

Extensive evaluations across multiple datasets, model families, and quantization levels (16-bit, 8-bit, and 4-bit) show that the framework achieves strong and competitive performance across AUROC, AURAC, RAC, and HRG over 112 evaluation settings without supervised fine-tuning. The nested component analysis shows progressive AUROC gains from entropy-only SE_R to SE_{RU} with observed TS probability sensitivity and finally to SURF with the additional singleton-mass fragility correction.

Overall, the results highlight the value of observed singleton mass as an interpretable proxy for incomplete semantic coverage in hallucination detection.

Future Work. While the proposed framework shows strong performance, several future exploration on limitations remain. First, semantic clustering relies on an external entailment model, whose errors may propagate and affect both singleton occupancy and entropy estimates. Second, the method assumes access to TS probabilities, limiting the complete formulation to models that expose sequence likelihoods. Third, SURF uses singleton mass as a proxy for unobserved support and does not identify unseen semantic clusters or recover their exact probability mass. Finally, repeated stochastic sampling increases inference cost relative to single-pass uncertainty estimation methods.

Acknowledgments

To be inserted at the conclusion of the review stage.

REFERENCES

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, May 1950.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying. *arXiv preprint arXiv:2304.13734*, 2023.

Harald Baayen. *Word Frequency Distributions*. Springer, 2001.

George Casella and Roger Berger. *Statistical Inference*. Duxbury, 2002.

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

William G. Cochran. *Sampling Techniques*. Wiley, 3 edition, 1977.

Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23696–23703, 2025.

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

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. Spuq: Perturbation-based uncertainty quantification for large language models. *arXiv preprint arXiv:2403.02509*, 2024.

Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

Irving J Good and George H Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

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

Yi Hao and Ping Li. Optimal prediction of the number of unseen species with multiplicity. *Advances in Neural Information Processing Systems*, 33:8553–8564, 2020.

Reilly Haskins and Benjamin Adams. Kea explain: Explanations of hallucinations using graph kernel analysis. In *International Conference on Neurosymbolic Learning and Reasoning.*, volume 37, 2025.

Mohammad Hosseini, Catherine A Gao, David M Liebovitz, Alexandre M Carvalho, Faraz S Ahmad, Yuan Luo, Ngan MacDonald, Kristi L Holmes, and Abel Kho. An exploratory survey about using chatgpt in education, healthcare, and research. *Plos one*, 18(10):e0292216, 2023.

Fengqing Jiang. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington, 2024.

Mingjian Jiang, Yangjun Ruan, Prasanna Sattigeri, Salim Roukos, and Tatsunori B Hashimoto. Graph-based uncertainty metrics for long-form language model generations. *Advances in Neural Information Processing Systems*, 37:32980–33006, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 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.

Sungmin Kang, Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, and Salman Avestimehr. Uncertainty quantification for hallucination detection in large language models: Foundations, methodology, and future directions. *arXiv preprint arXiv:2510.12040*, 2025.

E. Knights. Cognitive alignment at no cost: Inducing human attention biases for interpretable vision transformers. *arXiv preprint arXiv:2604.20027*, 2026.

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.

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.

Zeqiang Lai, Xizhou Zhu, Jifeng Dai, Yu Qiao, and Wenhai Wang. Mini-dalle3: Interactive text to image by prompting large language models. *arXiv preprint arXiv:2310.07653*, 2023.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*, 2019.

Seongmin Lee and Marcel Boehme. How much is unseen depends chiefly on information about the seen. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uqWM9hBDAE.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting llm uncertainty quantification with pairwise semantic similarity. *arXiv preprint arXiv:2506.00245*, 2025.

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.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021.

J. C. Principe. Information theoretic learning: Renyi's entropy and kernel perspectives. In M. Jordan, R. Novak, and B. Schölkopf (eds.), *Information Science and Statistics*. Springer, New York, NY, 2010.

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.

Ernesto Quevedo, Jorge Yero Salazar, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach. In *World Congress in Computer Science, Computer Engineering & Applied Computing*, pp. 154–173. Springer, 2024.

Johanna Elena Schmitz and Sven Rahmann. A comprehensive review and evaluation of species richness estimation. *Briefings in Bioinformatics*, 26(2):bbaf158, 2025.

B. Schölkopf, K. Muandet, K. Fukumizu, S. Harmeling, and J. Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25:755–766, June 2015.

Stanford NLP Group. SQuAD: The stanford question answering dataset. `https://rajpurkar.github.io/SQuAD-explorer/`, 2018. [Accessed 28 Apr 2025].

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

P. Vipulanandan, K. Premaratne, and D. Sarkar. Semantic uncertainty quantification of hallucinations in llms: A quantum tensor network based method. In *The International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil, 2026.

Justin Weltz, Eric Laber, and Alexander Volfovsky. Hidden population estimation with indirect inference and auxiliary information. *Proceedings of machine learning research*, 244:3730, 2024.

Tim Z Xiao, Aidan N Gomez, and Yarin Gal. Wat zei je? detecting out-of-distribution translations with variational transformers. *arXiv preprint arXiv:2006.08344*, 2020.

A FRAGILITY: PROPERTIES

A.1 UNSEEN-SPECIES COUNT ESTIMATE

In hallucination-prone settings, repeated LLM sampling may continue to reveal semantically distinct unseen answers, indicating incomplete coverage of the underlying response distribution. The number of *significant unseen answers* is estimated using frequency-of-frequencies statistics derived from repeated LLM sampling. Let t denote the number of additional generations and n the size of the observed sample. We use the classical Good–Toulmin estimator (Good & Toulmin, 1956)

$$\widehat{U}_{\mathrm{GT}}(t) = -\sum_{k=1}^{n} \left(-\frac{t}{n}\right)^{k} \Phi_k, \tag{7}$$

where Φ_k is the number of semantic species observed exactly k times (see (4)). Thus, the estimator is an alternating weighted sum over the complete frequency-of-frequencies profile and predicts the number of new semantic species expected among the next t generations. In our experiments, $n = 10$ and we set $t = 10$ to estimate the number of unseen semantic species expected in an additional sample of the same size; hence, the extrapolation ratio is $t/n = 1$.

A large estimated unseen count indicates that repeated sampling is expected to continue producing semantically distinct answers, suggesting incomplete coverage and increased instability in the generated outputs.

This estimate is further refined through fragility-aware entropy maximization:

$$u_0^{\star} = \arg\max_{u \in (0,1)} \left\{ h(u) - \alpha(F_{\mathrm{count}})\, D_{\mathrm{KL}}(u \| U_{GT_o}(t)) \right\}, \tag{8}$$

where $h(u)$ is the binary entropy. The regularization strength is modulated by a fragility term:

$$\alpha(F_{\mathrm{count}}) = \alpha_0\, e^{\gamma\, F_{\mathrm{count}}}, \text{ where } F_{\mathrm{count}} = \frac{\Phi_1}{\Phi_{\mathrm{obs}}}. \tag{9}$$

Here, $\alpha_0 > 0$ and $\gamma < 0$ are hyperparameters, Φ_{obs} denotes the total number of observed species, and F_{count} captures the sensitivity of observed answer counts to repeated sampling. The final unseen-species estimate is given by $\widehat{U}_0 = u_0^{\star}\, U_{\max}$, which adapts classical estimators by accounting for fragility in the observed answer distribution.

A.2 MISSING-MASS ESTIMATE

Analogously, a fragility-aware unseen-mass formulation can be defined by replacing count-based occupancy statistics with probability-mass statistics:

$$M_0^{\star} = \arg\max_{M_0 \in (0,1)} \left\{ h(M_0) - \alpha(F_{\mathrm{mass}})\, D_{\mathrm{KL}}(M_0 \| \widehat{M}_{\mathrm{GT}}) \right\}. \tag{10}$$

where $\widehat{M}_{\mathrm{GT}} = \Phi_1/n$ is a base missing-mass estimator (Good, 1953). This formulation quantifies fragility directly in probability space and is particularly useful when rare answers carry non-negligible probability mass. Its detailed evaluation is presented in Sec. F.2.

A.3 EXPECTED VALUE

This section analyzes the expectation of the fragility measure via a second-order delta approximation (Cochran, 1977):

$$\mathbb{E}[\widehat{F}] \approx \frac{\mathbb{E}[P_1]}{\mathbb{E}[P_{\mathrm{obs}}]} - \frac{\mathrm{Cov}(P_1, P_{\mathrm{obs}})}{\mathbb{E}[P_{\mathrm{obs}}]^2} + \frac{\mathbb{E}[P_1]\,\mathrm{Var}\,(P_{\mathrm{obs}})}{\mathbb{E}[P_{\mathrm{obs}}]^3}. \tag{11}$$

This expression provides an approximation to $\mathbb{E}[\widehat{F}]$ once the moments of A and B are computed.

Recall that the empirical fragility estimator is defined as

$$\widehat{F} = \frac{P_1}{P_{\mathrm{obs}}}, \text{ where } P_1 = \sum_{\ell=1}^{C} p_\ell\, \mathbb{1}(N_\ell = 1) \text{ and } P_{\mathrm{obs}} = \sum_{\ell=1}^{C} p_\ell\, \mathbb{1}(N_\ell > 0),$$

where C denotes the number of semantic clusters. Since P_1 and P_{obs} are correlated, $\mathbb{E}[\widehat{F}] = \mathbb{E}[P_1/P_{\mathrm{obs}}]$ does not admit a simple closed form in general. Instead, we employ a second-order delta approximation around $(\mu_1, \mu_{\mathrm{obs}})$, where $\mu_1 = \mathbb{E}[P_1]$ and $\mu_{\mathrm{obs}} = \mathbb{E}[P_{\mathrm{obs}}]$.

A.3.1 MULTINOMIAL SAMPLING APPROXIMATION

Step 1: Expectations of Singleton and Observed Mass Consider a sampling process consisting of n independent draws from a discrete distribution over species with probabilities $\{p_\ell\}_{\ell=1}^{C}$. Recall that N_ℓ denotes the number of times species ℓ appears in the sample.

Under this multinomial sampling model, the marginal count for each species is binomial, i.e., $N_\ell \sim \text{Binomial}(n, p_\ell)$:

$$\Pr(N_\ell = 1) = n\, p_\ell (1-p_\ell)^{n-1}; \qquad \Pr(N_\ell > 0) = 1 - (1-p_\ell)^n.$$

Therefore,

$$\mu_1 \quad = \mathbb{E}[P_1] \quad = \sum_{\ell=1}^{C} p_\ell \, \Pr(N_\ell = 1) = n \sum_{\ell=1}^{C} p_\ell^2 (1-p_\ell)^{n-1}; \tag{12}$$

$$\mu_{\text{obs}} = \mathbb{E}[P_{\text{obs}}] = \sum_{\ell=1}^{C} p_\ell \Pr(N_\ell > 0) \; = 1 - \sum_{\ell=1}^{C} p_\ell (1-p_\ell)^n. \tag{13}$$

Step 2: Variance of Observed Mass Let $I_\ell = \mathbb{1}(N_\ell > 0)$ so that $P_{\text{obs}} = \sum_\ell p_\ell I_\ell$. Then $\mathbb{E}[I_\ell] = 1 - (1-p_\ell)^n$. For $\ell \neq m$,

$$\mathbb{E}[I_\ell I_m] = 1 - (1-p_\ell)^n - (1-p_m)^n + (1-p_\ell - p_m)^n.$$

Hence

$$\mathbb{E}[P_{\text{obs}}^2] = \sum_\ell p_\ell^2 (1 - (1-p_\ell)^n) + 2 \sum_{1 \le \ell < m \le C} p_\ell \, p_m \left(1 - (1-p_\ell)^n - (1-p_m)^n + (1-p_\ell - p_m)^n\right), \tag{14}$$

and therefore

$$\text{Var}(P_{\text{obs}}) = \mathbb{E}[P_{\text{obs}}^2] - \mu_{\text{obs}}^2. \tag{15}$$

Step 3: Covariance Between Singleton and Observed Masses Let $J_\ell = \mathbb{1}(N_\ell = 1)$ so that $P_1 = \sum_\ell p_\ell J_\ell$. Then

$$\mathbb{E}[P_1 P_{\text{obs}}] = \sum_{\ell, m} p_\ell \, p_m \Pr(N_\ell = 1, \; N_m > 0).$$

If $\ell = m$,

$$\Pr(N_\ell = 1, \; N_\ell > 0) = \Pr(N_\ell = 1).$$

If $\ell \neq m$,

$$\Pr(N_\ell = 1, \; N_m > 0) = n\, p_\ell \left((1-p_\ell)^{n-1} - (1-p_\ell - p_m)^{n-1}\right).$$

Thus,

$$\mathbb{E}[P_1 P_{\text{obs}}] = n \sum_\ell p_\ell^3 (1-p_\ell)^{n-1} + n \sum_{\substack{\ell, m \\ \ell \neq m}} p_\ell^2 p_m \left((1-p_\ell)^{n-1} - (1-p_\ell - p_m)^{n-1}\right),$$

and hence

$$\text{Cov}(P_1, P_{\text{obs}}) = \mathbb{E}[P_1 P_{\text{obs}}] - \mu_1 \mu_{\text{obs}}. \tag{16}$$

Step 4: Approximated Expected Value of Fragility Estimator Combine (13), (15), and (16) to get

$$\mathbb{E}[\widehat{F}] \approx \frac{\mu_1}{\mu_{\text{obs}}} - \frac{\text{Cov}(P_1, P_{\text{obs}})}{\mu_{\text{obs}}^2} + \frac{\mu_1 \, \text{Var}(P_{\text{obs}})}{\mu_{\text{obs}}^3} \tag{17}$$

$$= \frac{n \sum_\ell p_\ell^2 (1-p_\ell)^{n-1}}{1 - \sum_\ell p_\ell (1-p_\ell)^n} - \frac{\mathbb{E}[P_1 P_{\text{obs}}] - \mu_1 \mu_{\text{obs}}}{\left(1 - \sum_\ell p_\ell (1-p_\ell)^n\right)^2} + \frac{\mu_1 \left(\mathbb{E}[P_{\text{obs}}^2] - \mu_{\text{obs}}^2\right)}{\left(1 - \sum_\ell p_\ell (1-p_\ell)^n\right)^3}, \tag{18}$$

where

$$\begin{aligned}
\mu_1 &= n\sum_\ell p_\ell^2(1-p_\ell)^{n-1};\\
\mu_{\text{obs}} &= 1-\sum_\ell p_\ell(1-p_\ell)^n;\\
\mathbb{E}[P_{\text{obs}}^2] &= \sum_\ell p_\ell^2\left(1-(1-p_\ell)^n\right)\\
&\qquad + 2\sum_{1\le\ell<m\le C} p_\ell\, p_m\left(1-(1-p_\ell)^n-(1-p_m)^n+(1-p_\ell-p_m)^n\right);\\
\mathbb{E}[P_1 P_{\text{obs}}] &= n\sum_\ell p_\ell^3(1-p_\ell)^{n-1} + n\sum_{\substack{\ell,m\\ \ell\neq m}} p_\ell^2 p_m\left(1-p_\ell\right)^{n-1}-(1-p_\ell-p_m)^{n-1}\big).
\end{aligned}$$

A.3.2 Poisson Sampling Approximation

Instead of drawing exactly n samples, one could also assume the occupancy counts are independent,

$$N_\ell \overset{\text{indep}}{\sim} \text{Poisson}(\lambda_\ell) \text{ with } \lambda_\ell \triangleq np_\ell \text{ and } \Pr(N_\ell = k) = \frac{e^{-np_\ell}(np_\ell)^k}{k!},\ k\in\{0,\dots,n\}. \quad (19)$$

This approximation is accurate when C, the number of species, is large and each p_ℓ is small, and it eliminates cross-species dependence.

Step 1: Expectations Under Poissonization For $\text{Poisson}(\lambda)$, we have

$$\Pr(N=0)=e^{-\lambda};\quad \Pr(N=1)=\lambda e^{-\lambda};\quad \Pr(N>0)=1-e^{-\lambda}.$$

Hence

$$\begin{aligned}
\mu_1 = \mathbb{E}[P_1] &= \sum_\ell p_\ell \Pr(N_\ell=1) = \sum_\ell p_\ell\lambda_\ell e^{-\lambda_\ell} &&= n\sum_\ell p_\ell^2 e^{-np_\ell};\\
\mu_{\text{obs}} = \mathbb{E}[P_{\text{obs}}] &= \sum_\ell p_\ell \Pr(N_\ell>0) = \sum_\ell p_\ell(1-e^{-\lambda_\ell}) &&= 1-\sum_\ell p_\ell e^{-np_\ell};\\
\mathbb{E}[M_0] &= \sum_\ell p_\ell \Pr(N_\ell=0) &&= \sum_\ell p_\ell e^{-np_\ell}.
\end{aligned} \quad (20)$$

Step 2: Variance Under Poissonization Let $I_\ell = \mathbb{1}(N_\ell>0)$. Note that, under (19), $\{I_\ell\}_\ell$ are independent Bernoulli random variables, and therefore

$$\text{Var}(I_\ell) = \Pr(N_\ell>0)\ \Pr(N_\ell=0) = (1-e^{-\lambda_\ell})\,e^{-\lambda_\ell}.$$

Since $P_{\text{obs}} = \sum_\ell p_\ell I_\ell$,

$$\text{Var}\left(\sum_\ell X_\ell\right) = \sum_\ell \text{Var}(X_\ell) + 2\sum_{\ell<m}\text{Cov}(X_\ell, X_m).$$

Applying this identity with $X_\ell = p_\ell I_\ell$ yields

$$\text{Var}(P_{\text{obs}}) = \sum_\ell \text{Var}(p_\ell I_\ell) + 2\sum_{\ell<m}\text{Cov}(p_\ell I_\ell, p_m I_m).$$

The fact that $\{I_\ell\}_\ell$ are independent implies that the cross-covariances vanish. So,

$$\text{Var}(P_{\text{obs}}) = \sum_\ell p_\ell^2\,\text{Var}(I_\ell) = \sum_\ell p_\ell^2\,(1-e^{-np_\ell})\,e^{-np_\ell}. \quad (21)$$

Step 3: Covariance Under Poissonization Let $J_\ell = \mathbb{1}(N_\ell = 1)$. Under independence across species, all cross terms with $\ell \neq m$ vanish yielding

$$\mathrm{Cov}\,(P_1, P_{\mathrm{obs}}) = \sum_\ell \mathrm{Cov}\,(p_\ell J_\ell,\, p_\ell I_\ell) = \sum_\ell p_\ell^2\, \mathrm{Cov}\,(J_\ell, I_\ell).$$

But $J_\ell I_\ell = J_\ell$ (since $N_\ell = 1 \implies N_\ell > 0$). So,

$$\begin{aligned}\mathrm{Cov}\,(J_\ell, I_\ell) &= \mathbb{E}[J_\ell] - \mathbb{E}[J_\ell]\,\mathbb{E}[I_\ell] \\ &= \Pr\,(N_\ell = 1)\,(1 - \Pr\,(N_\ell > 0)) = (\lambda_\ell\, e^{-\lambda_\ell})\,(e^{-\lambda_\ell}) = \lambda_\ell e^{-2\lambda_\ell}.\end{aligned}$$

Therefore

$$\mathrm{Cov}\,(P_1, P_{\mathrm{obs}}) = \sum_\ell p_\ell^2\, \lambda_\ell e^{-2\lambda_\ell} = n \sum_\ell p_\ell^3 e^{-2np_\ell}. \tag{22}$$

Step 4: Approximated Expected Value of Fragility Estimator Substituting from (20)–(22) into (11) yields

$$\begin{aligned}\mathbb{E}[\widehat{F}] \approx \Bigg[&\frac{n\sum_\ell p_\ell^2 e^{-np_\ell}}{1 - \sum_\ell p_\ell e^{-np_\ell}} - \frac{n\sum_\ell p_\ell^3 e^{-2np_\ell}}{(1 - \sum_\ell p_\ell e^{-np_\ell})^2} \\ &+ \frac{\left(n\sum_\ell p_\ell^2 e^{-np_\ell}\right)\left(\sum_\ell p_\ell^2 (1 - e^{-np_\ell})\, e^{-np_\ell}\right)}{(1 - \sum_\ell p_\ell e^{-np_\ell})^3} \Bigg].\end{aligned} \tag{23}$$

A.3.3 Zipf (Power-Law) Species Probabilities under Poissonized Sampling

If the species probabilities are assumed to follow a Zipf's power-law, we have

$$p_\ell = \frac{\ell^{-\alpha}}{H_{S,\alpha}},\ \text{with } H_{S,\alpha} \triangleq \sum_\ell \ell^{-\alpha},\ \alpha > 0. \tag{24}$$

This captures heavy-tailed regimes often observed in language-model answer distributions, where a few clusters have large mass and many clusters have very small mass.

We get the corresponding approximated expectation of the fragility measure by first combining Zipf probabilities in (24) with Poissonization in (19) and then substituting in (23).

A.4 Convergence Results Regarding the Fragility Measure

First, note that the objectives in (10) and (8) share the same entropy–KL regularized structure. Therefore, the following analysis is presented for the generic objective function $J(\cdot)$, which applies to both formulations.

Lemma 1. *Let $P_1 \in (0,1)$ and $\alpha(F) > 0$. Then, the objective function*

$$J(M) = h(M) - \alpha(F_{\mathrm{mass}})\, D_{\mathrm{KL}}(M \| P_1),$$

where

$$\begin{aligned}h(M) &= -M\,\log(M) - (1-M)\,\log(1-M); \\ D_{\mathrm{KL}}(M\|P_1) &= M\,\log\frac{M}{P_1} + (1-M)\,\log\frac{1-M}{1-P_1},\end{aligned}$$

is strictly concave in $M \in (0,1)$ and hence admits a unique maximizer.

Proof. The first- and second-order derivatives of h w.r.t. M are

$$h'(M) = \log\frac{1-M}{M};\ \ h''(M) = -\frac{1}{M} - \frac{1}{1-M}.$$

The first- and second-order derivatives of the KL divergence w.r.t. M are

$$\frac{\partial}{\partial M} D_{\mathrm{KL}} = \log\frac{M\,(1-P_1)}{P_1(1-M)};\ \ \frac{\partial^2}{\partial M^2} D_{\mathrm{KL}} = \frac{1}{M} + \frac{1}{1-M}.$$

Therefore, the second derivative of J is

$$J''(M) = h''(M) - \alpha(F)\,\frac{\partial^2}{\partial M^2}\,D_{\mathrm{KL}} = -(1+\alpha(F))\left(\frac{1}{M}+\frac{1}{1-M}\right).$$

Since $\alpha(F) > 0$ and $1/M+1/(1-M) > 0,\ \forall\, M \in (0,1)$, it follows that $J''(M) < 0,\ \forall M \in (0,1)$. Hence J is strictly concave and admits a unique maximizer. □

This directly leads us to

Lemma 2. *Let $M_0^\star(\alpha)$ denote the unique maximizer of (8) for a given $\alpha > 0$. Then, whenever $P_1 \in (0,1)$,*

$$\lim_{\alpha\to\infty} M_0^\star(\alpha) = P_1.$$

Proof. Write the objective as

$$J_\alpha(M) = h(M) - \alpha\, D_{\mathrm{KL}}(M\|P_1).$$

But J_α is strictly concave on $(0,1)$ (see Lemma 1). So, the maximizer $M_0^\star(\alpha)$ is unique and satisfies the first-order condition $J'_\alpha(M_0^\star(\alpha)) = 0$. A direct calculation of the derivatives $h'(M)$ and $\partial D_{\mathrm{KL}}(M\|P_1)/\partial M$ yields that $M_0^\star(\alpha)$ satisfies

$$\log\frac{1-M_0^\star(\alpha)}{M_0^\star(\alpha)} = \alpha\,\log\frac{M_0^\star(\alpha)\,(1-P_1)}{(1-M_0^\star(\alpha))\,P_1}.$$

We can express this as

$$-\mathrm{logit}\,(M_0^\star(\alpha)) = \alpha\,(\mathrm{logit}\,(M_0^\star(\alpha)) - \mathrm{logit}\,(P_1)),$$

where $\mathrm{logit}(\centerdot)$ is the logit function, i.e., $\mathrm{logit}\,(x) = \log\,(x/(1-x))$. This gives us

$$\mathrm{logit}\,(M_0^\star(\alpha)) = \frac{\alpha}{\alpha+1}\,\mathrm{logit}\,(P_1).$$

So, as $\alpha\to\infty$, $\mathrm{logit}\,(M_0^\star(\alpha)) \to \mathrm{logit}\,(P_1)$ and hence $M_0^\star(\alpha)\to P_1$. □

B Full Construction of Rényi Semantic UQ

When invoking the physics-inspired framework for UQ in Principe (2010), the TS probabilities are represented via kernel mean embeddings (KMEs) in a reproducing kernel Hilbert space (RKHS) defined by a positive semi-definite kernel (Aronszajn, 1950; Schölkopf et al., 2015).

This formulation is extended to semantic uncertainty in Vipulanandan et al. (2026). Using a Gaussian kernel $\kappa_\sigma(p_s, x)$, the empirical KME is given by

$$\widehat{\psi}_{\underline{y}}(x) = \frac{1}{R}\sum_{r=0}^{R-1}\kappa_\sigma(p_s^{(r)}, x)\,p_s^{(r)}, \tag{25}$$

where $\{p_s^{(r)}\},\ r = 0,\dots R-1$, correspond to R independent generations. The Rényi SE to which this UQ construction is applied is defined in (3).

The KME $\widehat{\psi}_{\underline{y}}(x)$ further admits a physics-inspired interpretation as a data-dependent wave function, enabling an operator-theoretic perturbation analysis. A Hamiltonian $\underline{\widehat{H}}$ constructed such that $\{\widehat{\underline{\psi}}_{\underline{y}}\}$ is an eigenmode, followed by first-order perturbation theory, yields the local-sensitivity feature vectors

$$\underline{V}_m^{(1)}(x) = E_m^{(1)} + \frac{\sigma^2}{2}\,\frac{\nabla_m^2|\underline{\psi}_m^{(1)}(x)|}{|\underline{\psi}_m^{(1)}(x)|}, \tag{26}$$

which capture local sensitivity in the RKHS. Averaging these features over the M modes yields the observed TS probability sensitivity score $\mathrm{UQ}(p_s^{(r)})$.

The sensitivity score $\mathrm{UQ}(p_s^{(r)})$ underpins the sensitivity-aware Rényi SE method (SE_{RU}) introduced in Vipulanandan et al. (2026). By incorporating local sensitivity at the level of observed TS probabilities, SE_{RU} improves hallucination detection relative to prior methods and remains robust across short- and long-form responses as well as different quantization levels (Vipulanandan et al., 2026).

However, this formulation does not account for unseen generations arising from finite sampling (e.g., a small number of LLM outputs), where missing responses may carry non-negligible probability mass and alter the entropy estimate. This limitation motivates the need for a method that explicitly incorporates unseen mass. This is the focus of our work.

C Extended Evaluation of SURF

In this section, we provide an extended evaluation of SURF across a diverse array of settings.

C.1 Paired Statistical and Resampling Analysis

Fig. 4 reports paired comparisons of SURF with SE_S and SE_R using within-setting AUROC differences over the 112 evaluation settings; the component-level results are reported once in Table 2.

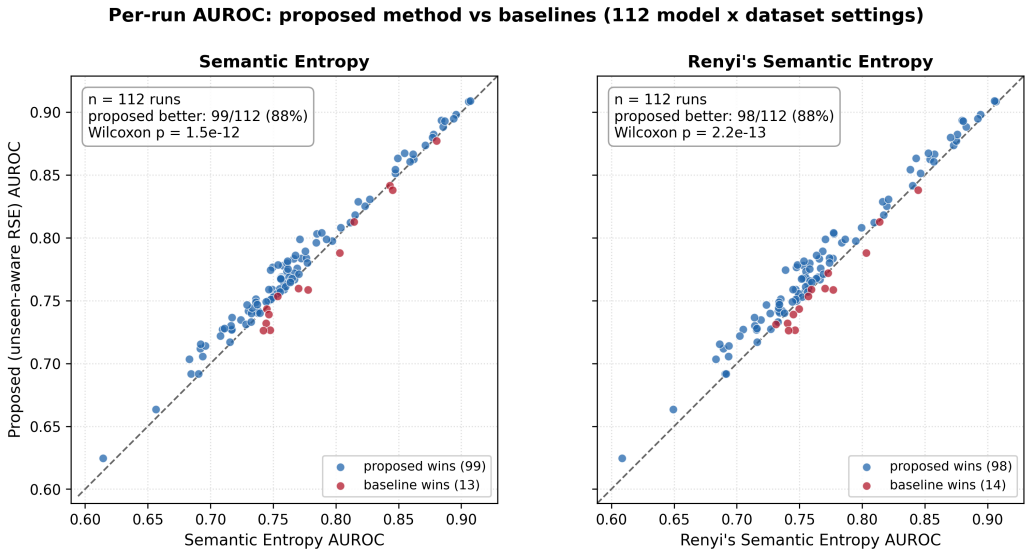


Figure 4: Paired AUROC comparison of SURF against Shannon SE (left) and Rényi SE (right) across 112 paired evaluation settings. Each point represents one paired setting; points above the diagonal favor SURF. The displayed p-values are from two-sided Wilcoxon signed-rank tests on the within-setting AUROC differences; "runs" in the plot denotes these paired evaluation settings.

We additionally analyze two distinct sources of variability. First, evaluation-example variability is assessed separately within every setting using a paired bootstrap ($B = 1000$ replicates), with the same resampled examples used for SURF and each baseline. Second, sensitivity to the finite set of sampled generations is assessed by deleting one of the ten generations for every query, recomputing the clustering-dependent uncertainty scores, and measuring the resulting variation in AUROC.

Fig. 5 reports the per-setting paired-bootstrap intervals for $\Delta\mathrm{AUROC}$. Although the point estimate favors SURF in 99/112 settings against SE_S and 98/112 against SE_R, most individual 95% intervals overlap zero: 7/112 are entirely positive against SE_S and 12/112 against SE_R, while one interval is entirely negative in each comparison. Thus, the evidence supports a small but systematic aggregate improvement rather than uniformly significant gains within every individual setting.

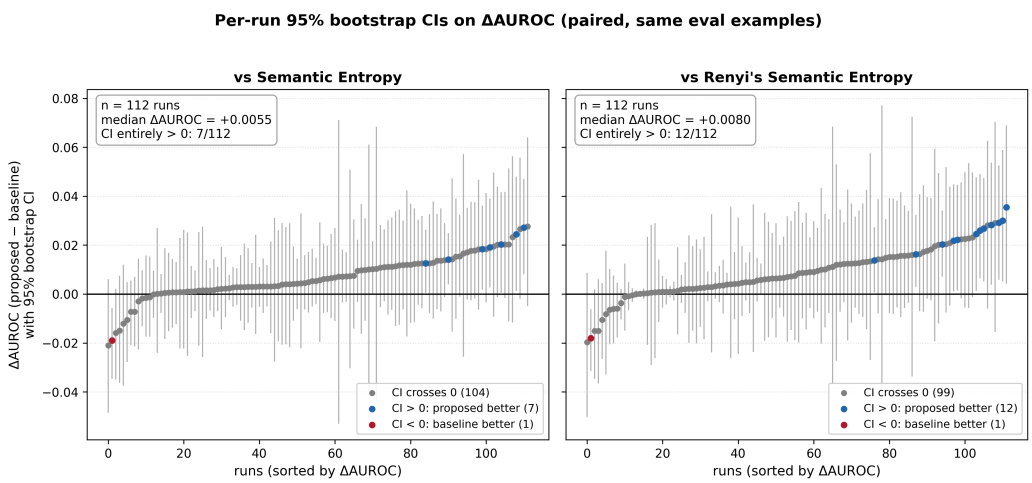


Figure 5: Per-setting paired-bootstrap 95% confidence intervals for $\Delta\text{AUROC} = \text{AUROC}(\text{SURF}) - \text{AUROC}(\text{baseline})$. Both methods are evaluated on the same resampled examples within each of the 112 settings. Settings are sorted by the observed AUROC difference; blue and red markers denote intervals entirely above and below zero, respectively.

Fig. 6 presents the leave-one-generation-out sensitivity analysis. This diagnostic isolates sensitivity to the finite generation set without requiring additional model inference. It should not be interpreted as a substitute for fully independent end-to-end reruns, because all replicates are constructed from the same original ten generations. Nevertheless, the predominance of positive point estimates and the reported deletion-based variability indicate that the aggregate comparison is not driven solely by a small number of sampled generations.

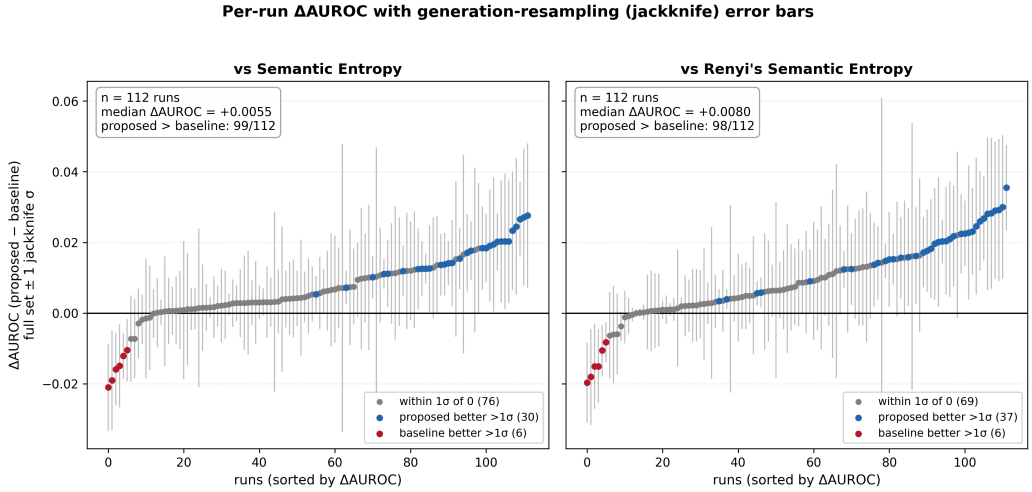


Figure 6: Leave-one-generation-out sensitivity analysis for the paired AUROC difference across 112 settings. Error bars show one deletion-based standard deviation after recomputing the clustering-dependent scores with one of the ten generations removed. This is a finite-sample sensitivity diagnostic, not an estimate obtained from independent model reruns.

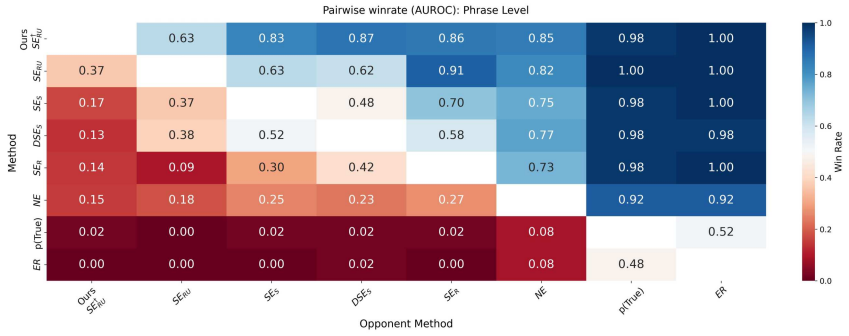
C.2 AUROC Performance Across Various Precision Models and Datasets

Figs. 7b and 7a summarize pairwise AUROC win-rates across all sentence- and phrase-level experiments, respectively. Each entry denotes the fraction of experimental settings in which the method

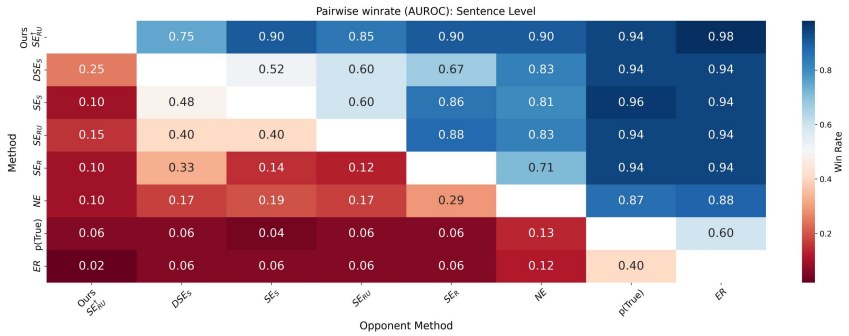
on the row achieves a higher AUROC than the method on the column. Values above 0.5 therefore indicate consistent superiority across datasets, model families, and quantization levels.

Overall, the proposed SURF ($\mathrm{SE}^{\dagger}_{RU}$) achieves the competitive and most consistent performance across both settings. At the sentence level, SURF outperforms competing semantic entropy methods in up to 90% of experiments and dominates probability-based baselines such as $p(\mathrm{True})$ and ER in 94% and 98% of cases, respectively. Similar trends are observed at the phrase level, where SURF maintains consistently high win-rates across nearly all baselines.

The results further show that semantic uncertainty methods substantially outperform probability-only approaches, highlighting the importance of semantic clustering for hallucination detection. Moreover, incorporating the singleton-mass fragility proxy yields systematic improvements over the evaluated semantic entropy formulations by identifying samples whose observed support is likely incomplete. This advantage becomes particularly pronounced at the phrase level, where repeated sampling induces greater semantic variability and distributional instability.



(a) Pairwise AUROC-based win-rate matrices across phrase length generation.



(b) Pairwise AUROC-based win-rate matrices across sentence length generation.

Figure 7: Evaluation of SOTA hallucination detection methods across different lengths of answer generation. This **summarizes 112 paired evaluation settings**. Each subfigure shows pairwise win-rate matrices under a different generation length. Our SURF ($\mathrm{SE}^{\dagger}_{RU}$)-based method generally outperforms the evaluated baselines at both lengths in AUROC.

C.2.1 AUROC Performance: Sentence Length

Figs. 8, 9, and 10 present AUROC comparisons across multiple hallucination detection methods under 16-bit, 8-bit, and 4-bit quantization settings, respectively, evaluated over sentence-length generations spanning multiple LLM architectures and benchmark datasets. In total, **60 experiments** were conducted across all quantization levels, providing a comprehensive evaluation of robustness under varying precision settings. Several important observations emerge from these experiments.

1. The proposed fragility-aware semantic uncertainty framework, SURF ($\mathrm{SE}^{\dagger}_{RU}$), consistently achieves competitive or superior AUROC performance across most datasets and quantization levels. In particular, the improvements become more pronounced under lower-precision regimes, suggesting that correcting in proportion to observed singleton mass provides additional robustness when model probabilities become noisier due to quantization.

2. Semantic uncertainty-based methods generally outperform approaches relying solely on TS probabilities ($p(\text{True})$) or embedding-regression baselines. This reinforces the importance of modeling semantic variability and distributional uncertainty, rather than relying only on raw confidence estimates from the model.
3. Performance degradation under aggressive quantization is notably smaller for SURF ($\text{SE}^{\dagger}_{RU}$) compared to several baselines such as NE and standard semantic entropy variants. This indicates that fragility-aware corrections help stabilize uncertainty estimation even when model precision is substantially reduced, making the framework suitable for resource-constrained deployment scenarios.
4. The gap between semantic methods and probability-based baselines becomes particularly visible on datasets such as TriviaQA, NQ, and SQuAD, where repeated sampling produces semantically diverse generations. This suggests that using observed singleton mass as a proxy for incomplete semantic coverage improves the ability to distinguish faithful responses from hallucinated ones.
5. Smaller models, particularly the LLaMA-3.2-1B variants, exhibit larger fluctuations across datasets and quantization levels relative to larger models, indicating that reduced representational capacity affects uncertainty calibration and semantic consistency. Nevertheless, SURF ($\text{SE}^{\dagger}_{RU}$) remains comparatively stable even in these low-capacity settings.
6. Across nearly all experimental configurations, SURF ($\text{SE}^{\dagger}_{RU}$) preserves strong AUROC performance without requiring supervised fine-tuning or auxiliary verification models, highlighting the practicality and scalability of the approach for real-world hallucination detection pipelines.

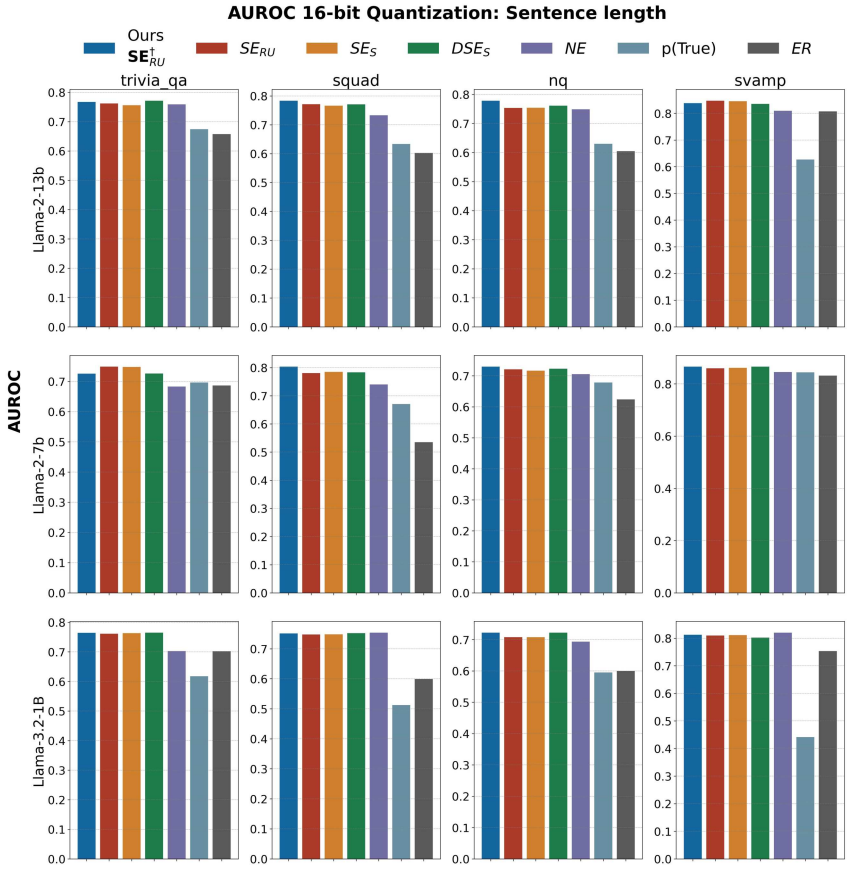


Figure 8: **Summarizing 12 evaluation settings**, AUROC scores for confabulation detection across three LLMs (LLaMA 2 13B, LLaMA 2 7B, LLaMA 3.2 1B) at 16-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for sentence-length output. The proposed $\text{SE}^{\dagger}_{RU}$ is competitive with the evaluated methods.

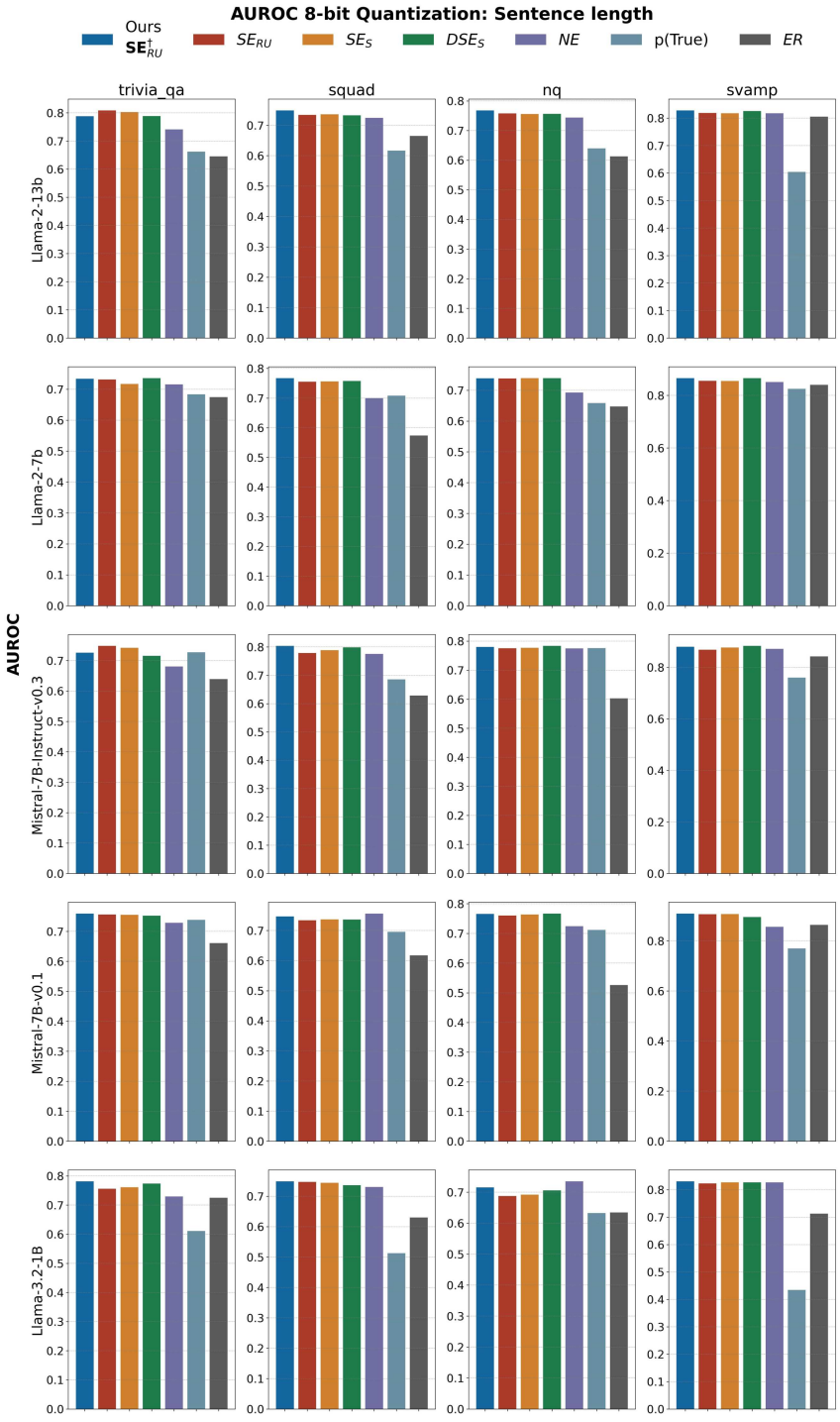


Figure 9: **Summarizing 20 evaluation settings**, AUROC scores for confabulation detection across five LLMs (LLaMA 2 13B, LLaMA 2 7B, LLaMA 3.2 1B, Mistral-7B, Mistral-7B-chat) at 8-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for sentence-length output. The proposed $\mathrm{SE}^{\dagger}_{RU}$ is competitive with the evaluated methods.

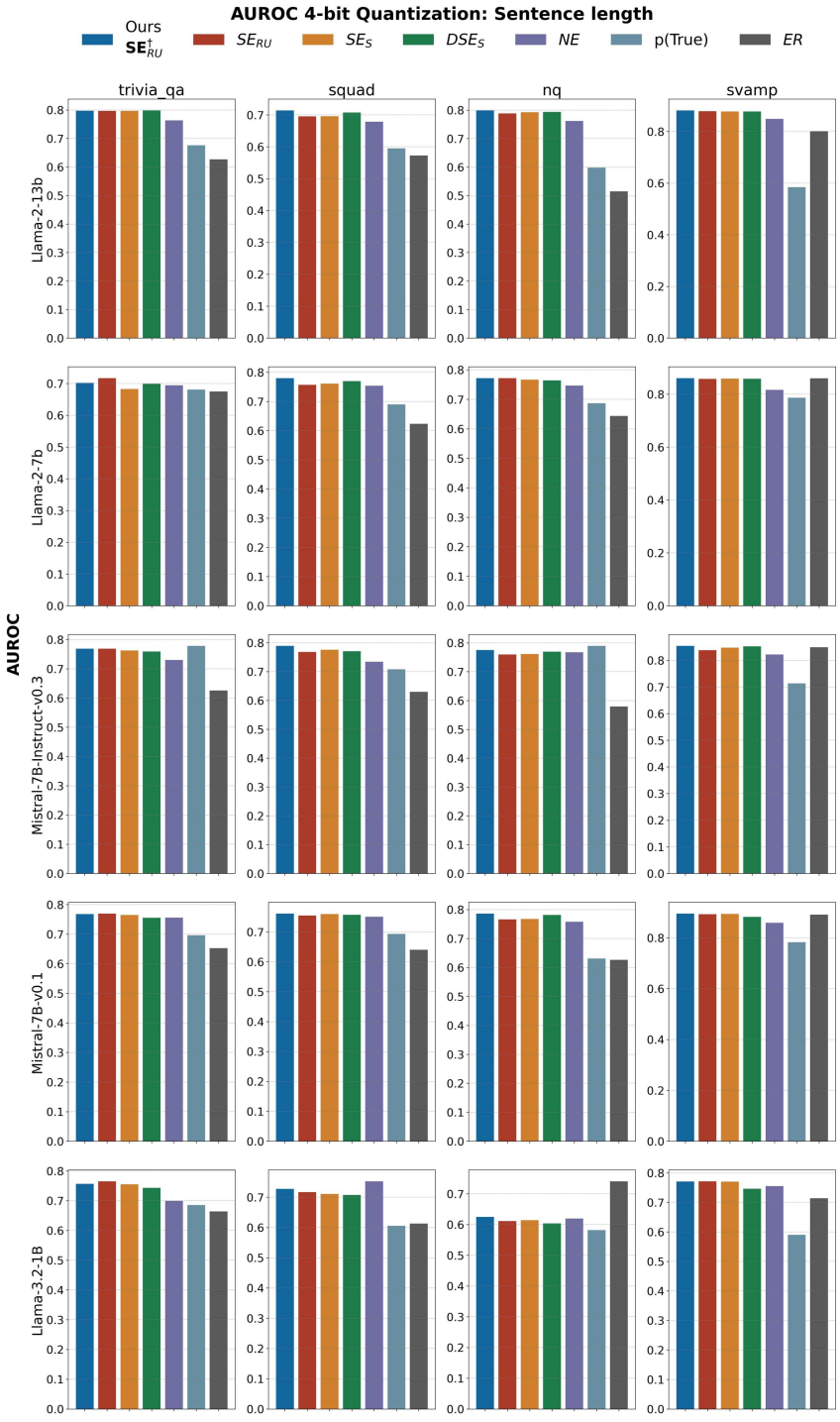


Figure 10: **Summarizing 20 evaluation settings**, AUROC scores for confabulation detection across five LLMs (LLaMA 2 13B, LLaMA 2 7B, LLaMA 3.2 1B, Mistral-7B, Mistral-7B-chat) at 4-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for sentence-length output. The proposed $\mathrm{SE}^{\dagger}_{RU}$ is competitive with the evaluated methods.

C.2.2 AUROC PERFORMANCE: PHRASE LENGTH ON INSTRUCT MODELS

Figs. 11, 12, and 13 present AUROC comparisons for phrase-length generations across instruction-tuned LLMs under 16-bit, 8-bit, and 4-bit quantization settings, respectively. In total, **24 experiments** were conducted spanning multiple datasets and quantization regimes. Several notable observations emerge from these evaluations.

1. SURF ($\text{SE}^{\dagger}_{RU}$) consistently achieves competitive or superior AUROC performance across nearly all datasets and quantization levels, demonstrating that fragility-aware semantic uncertainty estimation remains effective even for shorter phrase-level generations produced by instruction-tuned models.
2. Compared to probability-based baselines such as $p(\text{True})$ and embedding-regression methods, semantic uncertainty approaches provide substantially stronger hallucination detection performance. This highlights the importance of modeling semantic variability rather than relying solely on token-level confidence estimates.
3. The proposed fragility-aware correction remains stable under aggressive quantization. In particular, SURF ($\text{SE}^{\dagger}_{RU}$) exhibits smaller performance degradation from 16-bit to 4-bit precision relative to several baseline approaches, indicating improved robustness to quantization-induced noise in probability estimates.
4. Phrase-level generations exhibit slightly smaller performance gaps between methods compared to sentence-level outputs. This behavior is expected since shorter generations contain lower semantic diversity and reduced uncertainty. Nevertheless, the proposed framework consistently remains among the competitive-performing methods across nearly all experimental settings.
5. The benefits of fragility-aware uncertainty modeling become particularly visible on datasets such as TriviaQA, SQuAD, and SVAMP, where repeated sampling can still produce semantically distinct phrase completions despite the shorter output length.
6. Although lower-capacity models under heavy quantization exhibit increased instability, SURF ($\text{SE}^{\dagger}_{RU}$) maintains comparatively stable AUROC performance, highlighting its suitability for lightweight and resource-constrained deployment scenarios.

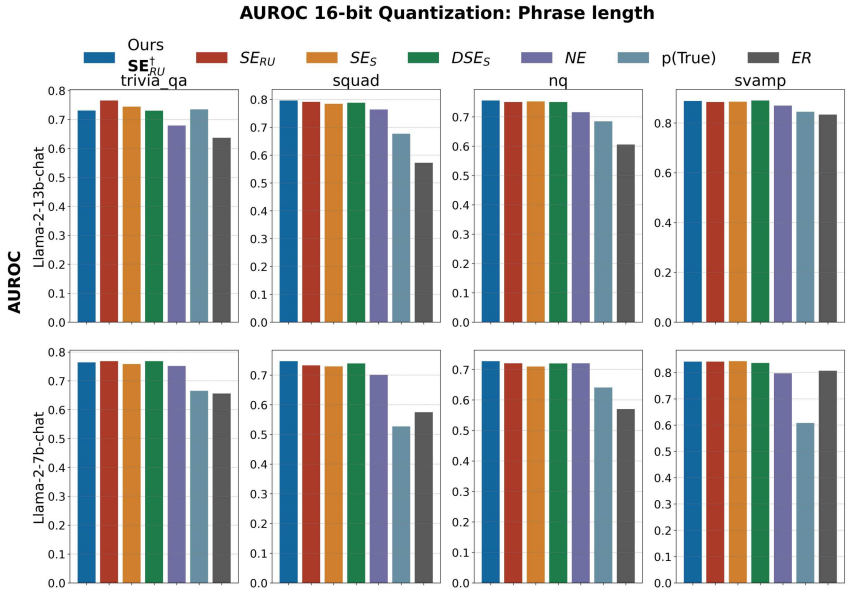


Figure 11: **Summarizing 8 evaluation settings**, AUROC scores for hallucination detection across instruction-tuned LLMs (LLaMA 2 13B-chat and LLaMA 2 7B-chat) at 16-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for phrase-length outputs. The proposed SURF ($\text{SE}^{\dagger}_{RU}$) achieves competitive performance across most settings.

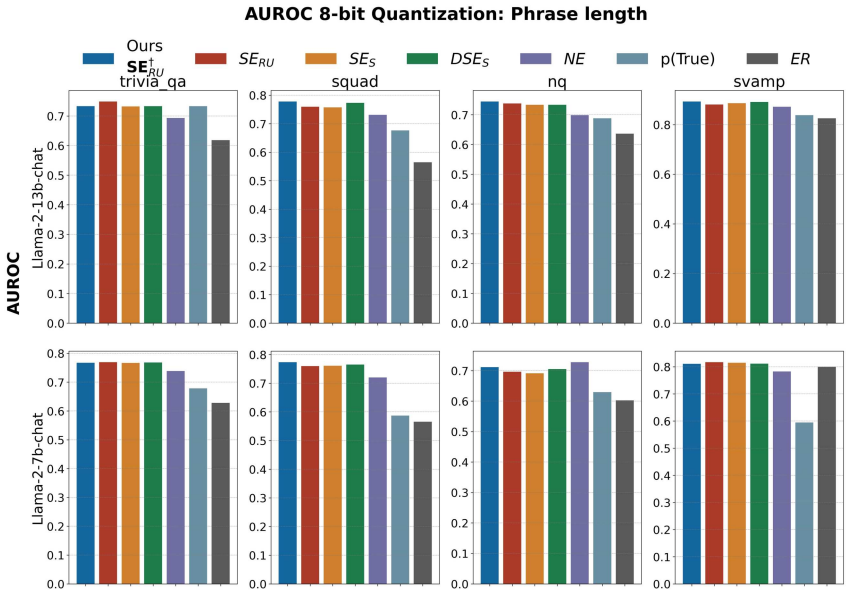


Figure 12: **Summarizing 8 evaluation settings**, AUROC scores for hallucination detection across instruction-tuned LLMs (LLaMA 2 13B-chat and LLaMA 2 7B-chat) at 8-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for phrase-length outputs. SURF ($\text{SE}^{\dagger}_{RU}$) remains competitive under reduced precision.

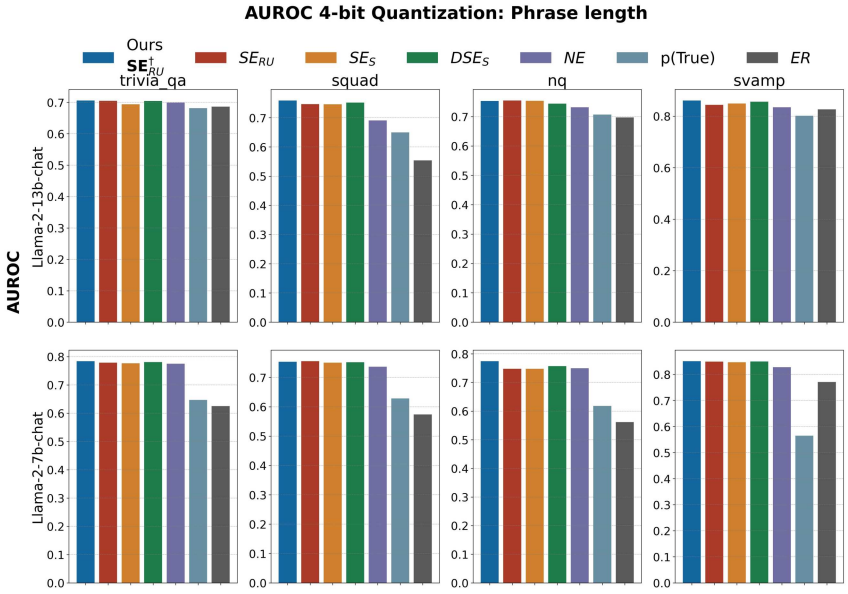


Figure 13: **Summarizing 8 evaluation settings**, AUROC scores for hallucination detection across instruction-tuned LLMs (LLaMA 2 13B-chat and LLaMA 2 7B-chat) at 4-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for phrase-length outputs. Under aggressive quantization, the proposed SURF ($\text{SE}^{\dagger}_{RU}$) remains competitive with the evaluated semantic uncertainty methods.

C.2.3 AUROC PERFORMANCE: PHRASE LENGTH ON NON INSTRUCT MODELS

Figs. 14, 15, and 16 illustrate the AUROC performance at 16-bit, 8-bit, and 4-bit precision, respectively, for phrase-length outputs on non-instruction-tuned LLMs. In total, these figures summarize **36 evaluation settings** across three model families, four datasets, and three quantization levels. Several important observations emerge from these results:

1. Semantic uncertainty-based methods consistently outperform TS probability baselines such as $p(\text{True})$ and embedding regression (ER), reinforcing the importance of semantic-level uncertainty modeling for hallucination detection in non instruction-tuned models.
2. The proposed SURF ($\text{SE}^{\dagger}_{RU}$) achieves highly competitive AUROC performance across nearly all datasets and quantization levels, frequently matching or exceeding existing semantic entropy approaches. This indicates that incorporating fragility-aware corrections preserves robust uncertainty estimation even when generation quality is less aligned compared to instruction-tuned models.
3. Performance degradation under aggressive quantization is relatively limited for SURF, particularly when transitioning from 16-bit to 8-bit and 4-bit precision. In contrast, methods relying heavily on raw TS probabilities exhibit larger fluctuations, especially for smaller models such as LLaMA-3.2-1B.
4. The proposed framework demonstrates particularly stable behavior on TriviaQA and SVAMP across all model scales, suggesting that fragility-aware uncertainty estimation remains effective even when semantic answer distributions are highly diverse.
5. Consistent with sentence-level experiments, smaller models exhibit larger variability in AUROC across datasets, highlighting the increased sensitivity of compact LLMs to quantization and sampling instability. Nevertheless, SURF maintains competitive performance even in these low-capacity regimes.
6. On datasets such as NQ and SQuAD, certain semantic entropy baselines occasionally achieve marginally higher AUROC values. However, these improvements are often inconsistent across quantization levels, whereas SURF exhibits stronger stability and robustness across the full experimental setting.

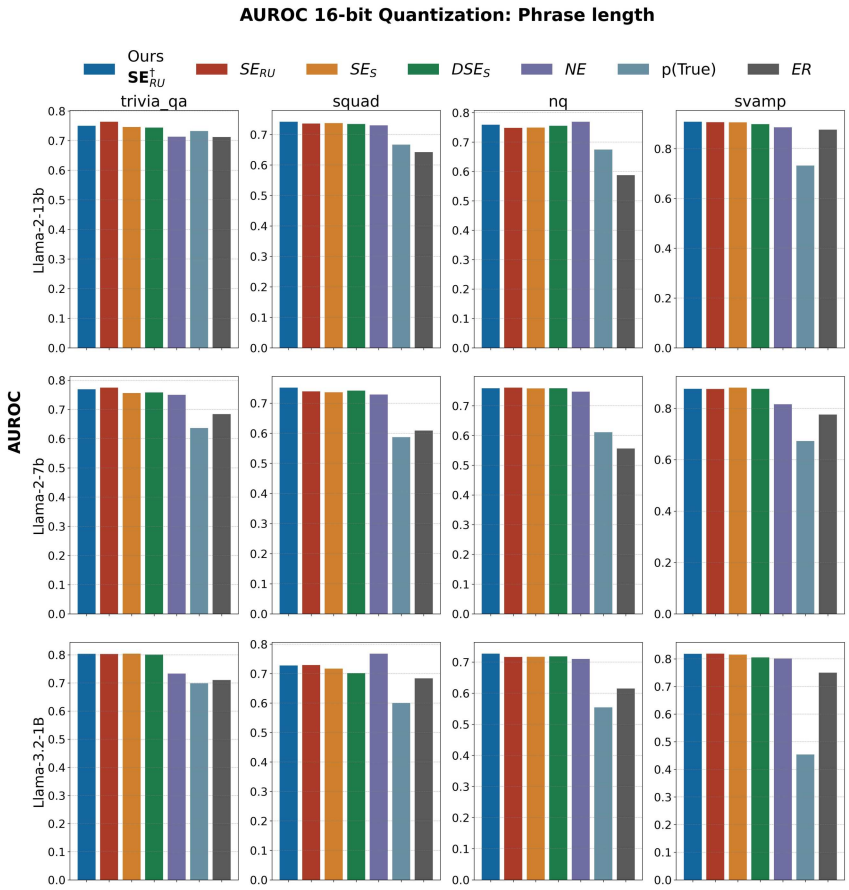


Figure 14: **Summarizing 12 evaluation settings**, AUROC scores for hallucination detection across non-instruction-tuned LLMs (LLaMA 2 13B, LLaMA 2 7B, LLaMA 3.2 1B) at 16-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for phrase-length outputs. The proposed SURF ($\mathrm{SE}^{\dagger}_{RU}$) achieves competitive performance relative to the evaluated semantic uncertainty methods.

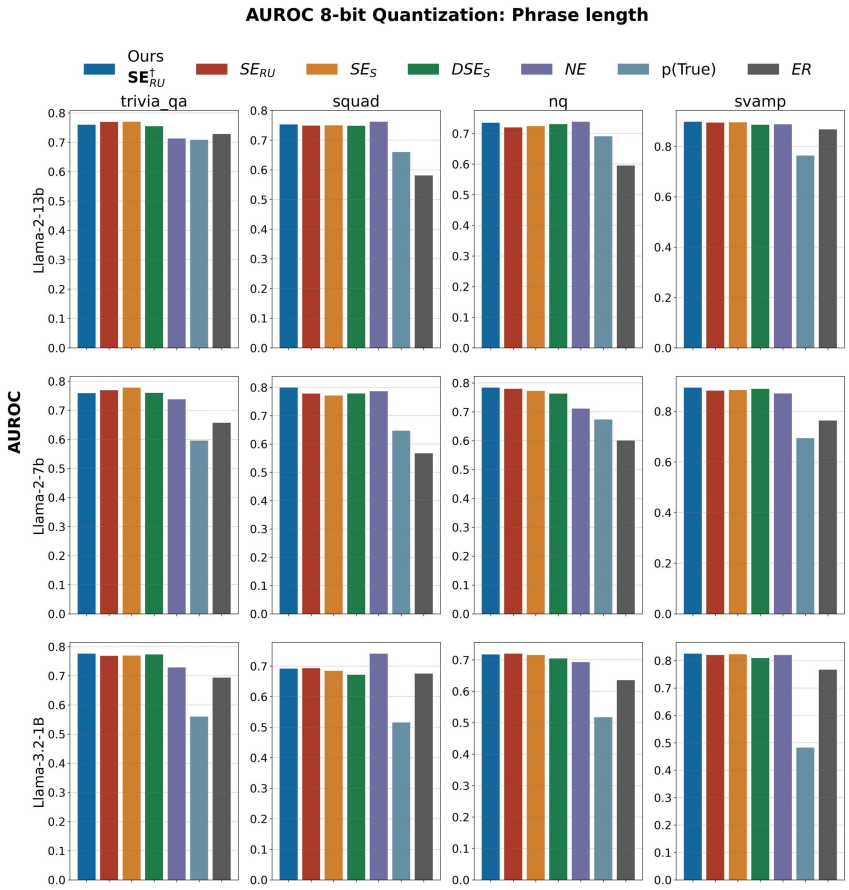


Figure 15: **Summarizing 12 evaluation settings**, AUROC scores for hallucination detection across non-instruction-tuned LLMs (LLaMA 2 13B, LLaMA 2 7B, LLaMA 3.2 1B) at 8-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for phrase-length outputs. The proposed SURF ($\mathrm{SE}^{\dagger}_{RU}$) remains competitive under moderate quantization.

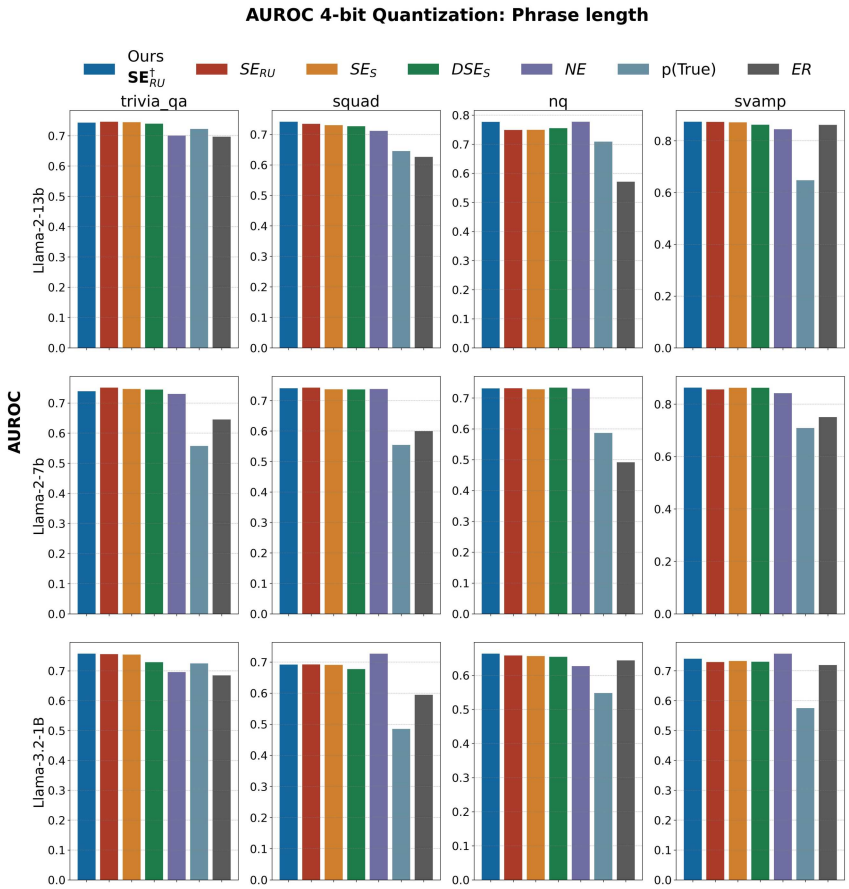


Figure 16: **Summarizing 12 evaluation settings**, AUROC scores for hallucination detection across non-instruction-tuned LLMs (LLaMA 2 13B, LLaMA 2 7B, LLaMA 3.2 1B) at 4-bit precision and four datasets (TriviaQA, SQuAD, NQ, SVAMP) for phrase-length outputs. Under aggressive quantization, the proposed SURF ($\mathrm{SE}^{\dagger}_{RU}$) maintains competitive AUROC performance.

C.3 HEADROOM GAIN (HRG) ACROSS VARIOUS LENGTH GENERATIONS

Complementing the AUROC performance in App. C.2 and to enable a fair comparison across models operating at different performance levels, we introduce *headroom gain (HRG)* as

$$\mathrm{HRG}(i,j) = \begin{cases} \tanh\left(\alpha \cdot \dfrac{A_i - A_j}{1 - A_j}\right), & \text{if } A_i \geq A_j; \\ \tanh\left(\alpha \cdot \dfrac{A_i - A_j}{A_j}\right), & \text{if } A_i < A_j. \end{cases} \tag{27}$$

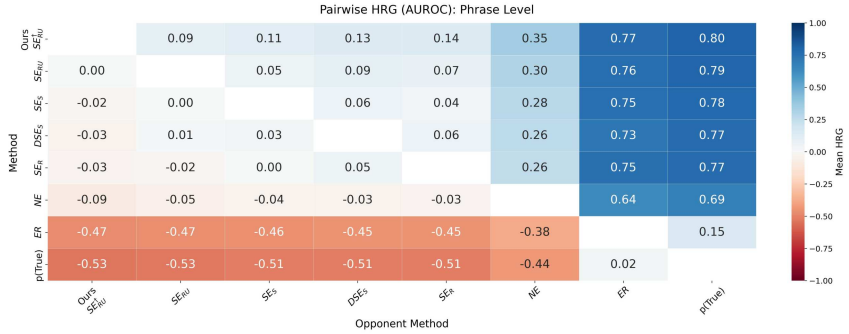
HRG as defined above is a directed metric that captures relative improvement with respect to the available performance headroom of a baseline. Unlike absolute differences in AUROC, which can be misleading near saturation, HRG normalizes gains by the remaining distance to the upper bound and losses by the distance to the lower bound. This yields a scale-aware measure that reflects how much of the achievable improvement (or degradation) is realized.

As a consequence, HRG is inherently asymmetric: the improvement of method i over j is not equal in magnitude to the reverse comparison, since the available headroom depends on the reference method. Importantly, this formulation ensures that HRG is naturally bounded. In its unscaled form, HRG lies in $[-1, 1]$, where 1 indicates that the full available headroom has been closed (i.e., reaching the optimal performance), and -1 corresponds to a complete degradation to the lower bound. The nonlinear $\tanh(\cdot)$ transformation further stabilizes the metric and preserves these bounds while mitigating the effect of extreme values. This bounded and asymmetric measure provides a more meaningful and robust comparison in settings such as AUROC, where equal absolute changes can correspond to vastly different relative effects depending on the operating regime.

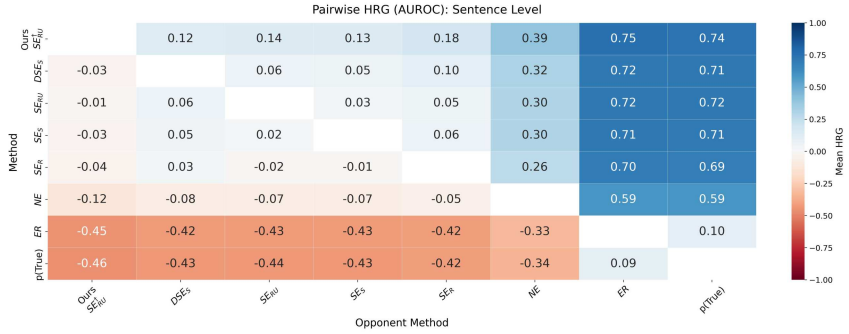
The parameter $\alpha > 0$ controls the sensitivity of the transformation, scaling the magnitude of the normalized gain before applying the $\tanh(\cdot)$ nonlinearity. In this work, we fix $\alpha = 3$, which provides a good balance between preserving meaningful differences and preventing saturation, ensuring that moderate improvements are sufficiently emphasized while extreme values remain bounded.

Figs. 17b and 17a summarize the pairwise AUROC HRG comparisons across **112 paired evaluation settings** for sentence- and phrase-level outputs, complementing the win-rate analysis in App. C.3 by quantifying the *magnitude* of within-setting improvement rather than only the frequency of positive differences. Positive HRG values indicate that the method on the row achieves higher relative AUROC gains than the method on the column, normalized by the available performance headroom.

Consistent with the win-rate results, the proposed SURF ($\mathrm{SE}^{\dagger}_{RU}$) achieves the competitive overall performance across both sentence- and phrase-level evaluations. In particular, SURF exhibits substantial positive HRG margins against probability-based baselines such as $p(\mathrm{True})$ and ER, while also maintaining consistent gains over competing semantic entropy methods. The advantage is especially pronounced in phrase-level settings, where fragility-aware corrections provide larger relative improvements under increased semantic variability and sampling instability. Overall, the HRG analysis confirms that the gains achieved by SURF are not only frequent but also statistically meaningful in magnitude across diverse datasets, model families, and quantization levels.



(a) Pairwirse HRG across phrase length generation.



(b) Pairwise HRG across sentence length generation.

Figure 17: Pairwise HRG comparison of hallucination detection methods at phrase (Fig. 17a) and sentence (Fig. 17b) levels using AUROC. Results are aggregated over 112 paired evaluation settings. Each matrix entry measures the within-setting HRG of one method over another. The proposed SURF ($\text{SE}^{\dagger}_{RU}$) generally achieves positive HRG against the evaluated baselines across both granularities.

C.4 AURAC AND RAC PERFORMANCE ACROSS VARIOUS PRECISION MODELS AND DATASETS

This section analyzes hallucination detection performance using AURAC, a complementary evaluation metric to AUROC in App. C.2, that measures how prediction accuracy evolves as increasingly uncertain samples are progressively rejected. Unlike AUROC, which evaluates ranking quality globally, AURAC directly reflects the practical usefulness of an uncertainty estimation method for selective prediction and decision-making under uncertainty. Higher AURAC values indicate that the method more effectively concentrates incorrect or hallucinated generations within the low-confidence region.

C.4.1 RAC PERFORMANCE: SENTENCE LENGTH

Figs. 18, 19, and 20 illustrate the RAC performance across different rejection thresholds for sentence-level outputs under 16-bit, 8-bit, and 4-bit quantization, respectively. RAC evaluates how effectively an uncertainty estimation method retains accurate predictions as increasingly uncertain generations are rejected. In total, these plots summarize **60 evaluation settings** spanning multiple datasets, model families, and quantization levels. Several important observations emerge from these results.

1. Semantic uncertainty methods consistently outperform probability-only baselines such as $p(\text{True})$ and ER across most datasets and quantization levels, indicating the importance of semantic-level variability in identifying hallucinations.
2. The proposed fragility-aware framework, SURF ($\text{SE}^{\dagger}_{RU}$), achieves consistently strong RAC performance and retains high-confidence correct predictions more effectively across rejection thresholds.
3. SURF remains robust under aggressive quantization (8-bit and 4-bit), whereas several competing baselines exhibit larger performance degradation and less stable rejection behavior.
4. The benefits of fragility-aware uncertainty estimation are particularly visible for smaller models and more challenging datasets, where repeated sampling exposes higher semantic variability and incomplete answer coverage.

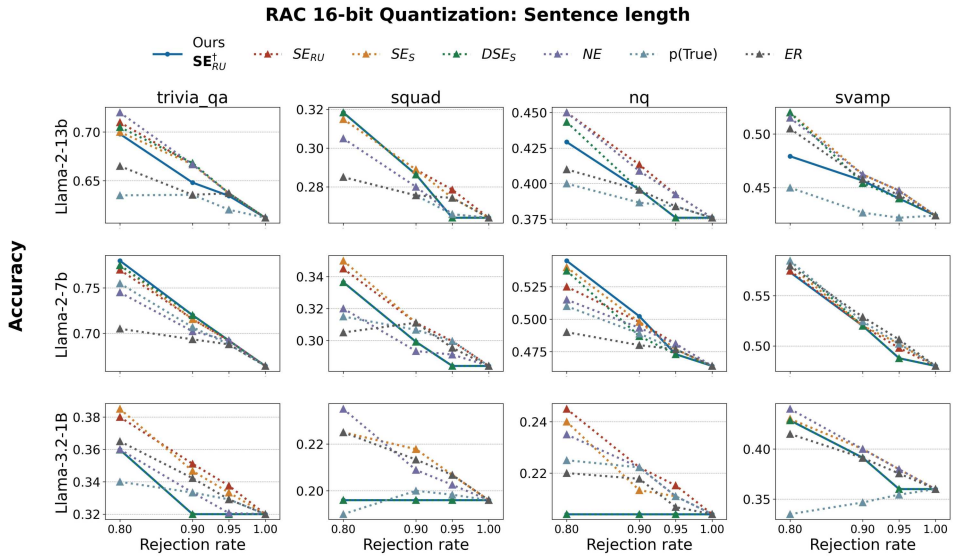


Figure 18: **Summarizing 16 evaluation settings**, RAC curves for hallucination detection across multiple LLMs and four datasets (TriviaQA, SQuAD, NQ, and SVAMP) under 16-bit quantization for sentence-level outputs. SURF ($\text{SE}^{\dagger}_{RU}$) maintains competitive retained accuracy across rejection thresholds relative to the evaluated semantic uncertainty baselines.

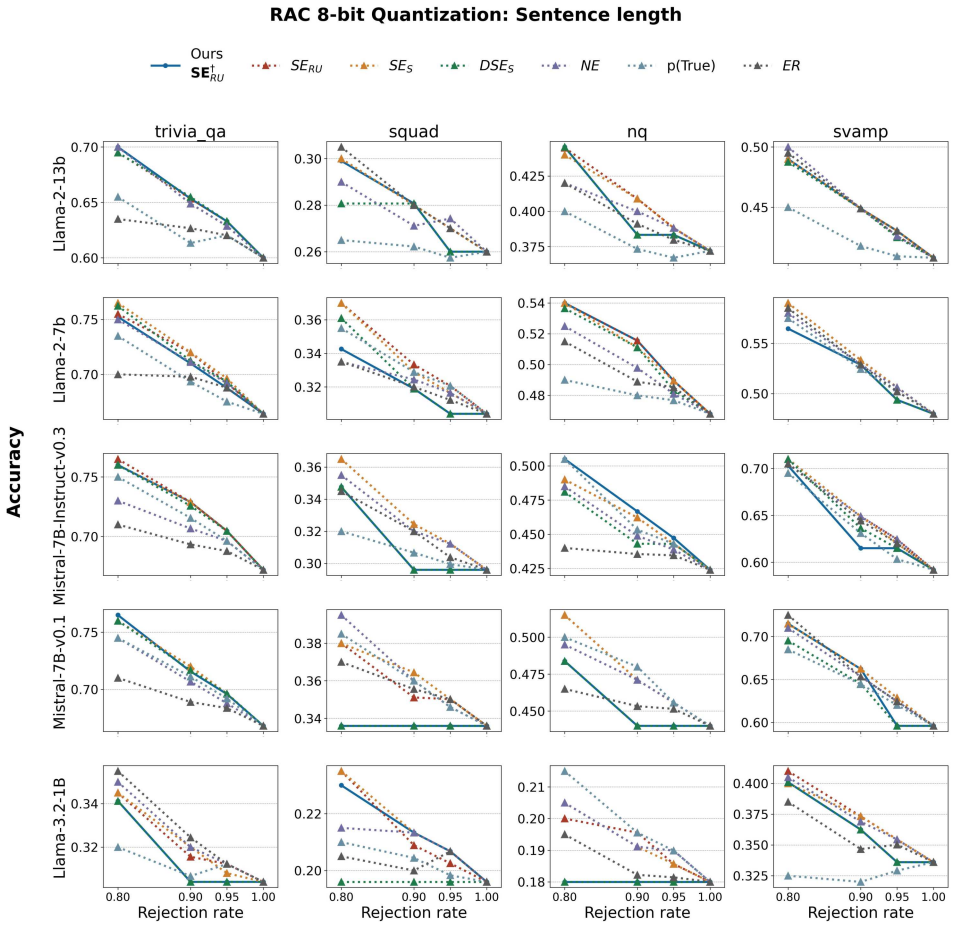


Figure 19: **Summarizing 20 evaluation settings**, RAC curves for hallucination detection across multiple LLMs and four datasets (TriviaQA, SQuAD, NQ, and SVAMP) under 8-bit quantization for sentence-level outputs. SURF ($\mathrm{SE}^{\dagger}_{RU}$) maintains competitive retained accuracy under reduced precision.

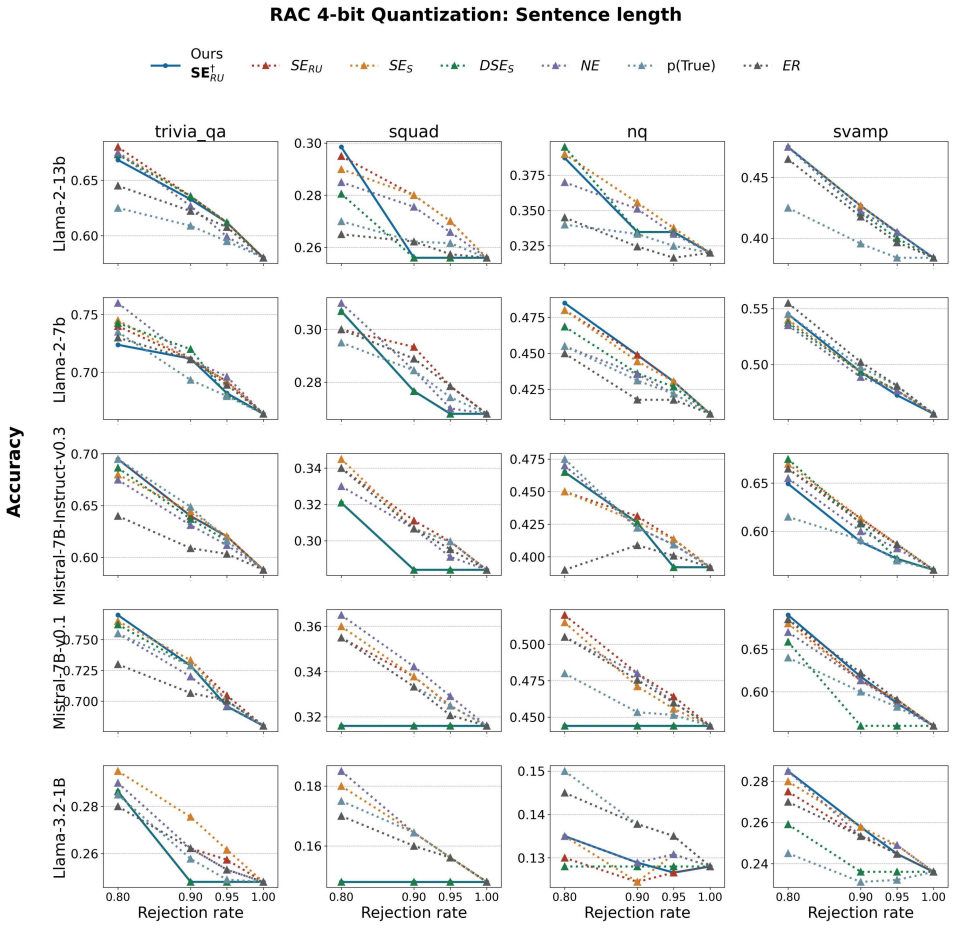


Figure 20: **Summarizing 20 evaluation settings**, RAC curves for hallucination detection across multiple LLMs and four datasets (TriviaQA, SQuAD, NQ, and SVAMP) under 4-bit quantization for sentence-level outputs. Under aggressive quantization, SURF ($\mathrm{SE}^{\dagger}_{RU}$) preserves competitive RAC performance relative to the evaluated semantic entropy methods.

C.4.2 RAC Performance: Phrase Length on Instruct Models

Figs. 21, 22, and 23 illustrate the RAC performance across rejection thresholds for phrase-level outputs under 16-bit, 8-bit, and 4-bit quantization, respectively. These plots summarize **24 evaluation settings** across multiple instruct-tuned LLMs and datasets.

1. Semantic uncertainty-based approaches consistently outperform probability-only baselines such as $p(\text{True})$ and ER, highlighting the importance of semantic-level variability for reliable hallucination detection at the phrase level.
2. The proposed fragility-aware framework, SURF ($\text{SE}^{\dagger}_{RU}$), maintains competitive RAC performance across nearly all rejection thresholds and quantization levels, often preserving higher retained accuracy in low- and moderate-rejection regimes.
3. Phrase-level settings exhibit larger variability across methods compared to sentence-level outputs, reflecting the increased semantic diversity introduced by shorter generations. In these settings, fragility-aware corrections help stabilize uncertainty estimation under repeated sampling.
4. Even under aggressive quantization, SURF demonstrates robust rejection behavior relative to existing semantic entropy baselines, indicating that the singleton-mass fragility correction improves uncertainty calibration in compressed deployment settings.

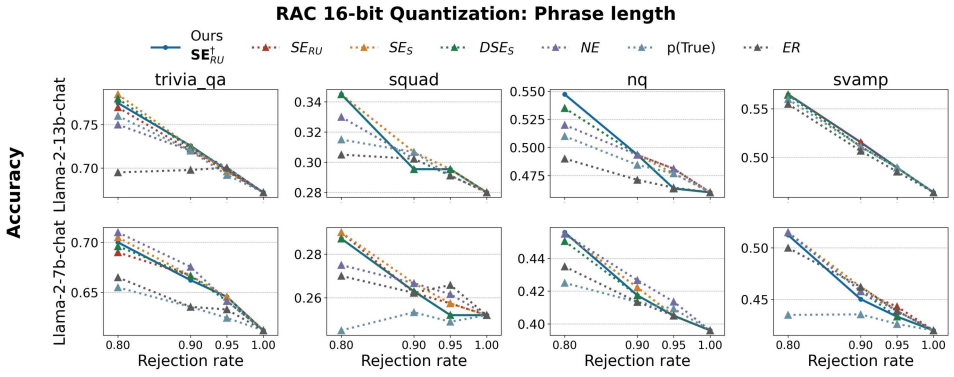


Figure 21: **Summarizing 8 evaluation settings**, RAC curves for hallucination detection across instruct-tuned LLMs and four datasets (TriviaQA, SQuAD, NQ, and SVAMP) under 16-bit quantization for phrase-level outputs. SURF ($\text{SE}^{\dagger}_{RU}$) maintains competitive retained accuracy across rejection thresholds.

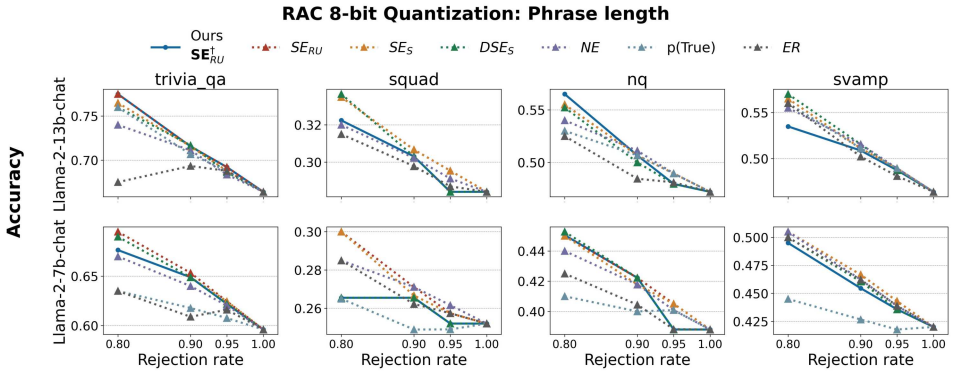


Figure 22: **Summarizing 8 evaluation settings**, RAC curves for hallucination detection across instruct-tuned LLMs and four datasets under 8-bit quantization for phrase-level outputs. The proposed SURF ($\text{SE}^{\dagger}_{RU}$) remains competitive with the evaluated semantic uncertainty baselines under reduced precision.

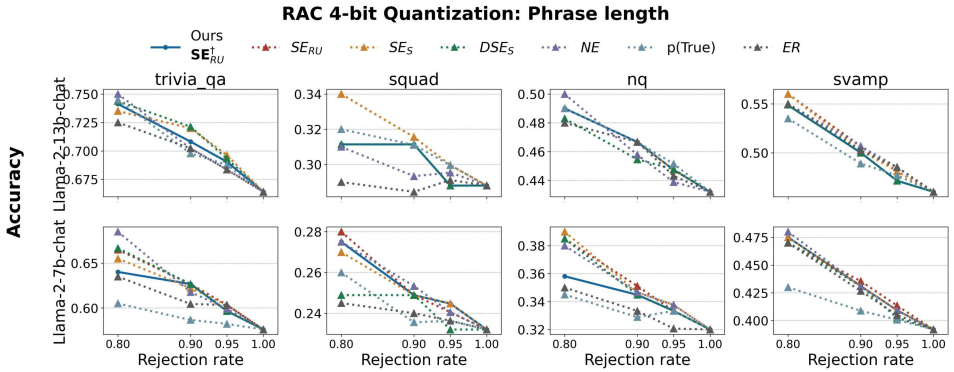


Figure 23: **Summarizing 8 evaluation settings**, RAC curves for hallucination detection across instruct-tuned LLMs and four datasets under 4-bit quantization for phrase-level outputs. Under aggressive quantization, SURF ($\mathrm{SE}^{\dagger}_{RU}$) remains competitive across rejection thresholds.

C.4.3 RAC Performance: Phrase Length on Non Instruct Models

Figs. 24, 25, and 26 illustrate the RAC performance across rejection thresholds for phrase-level outputs on non-instruction-tuned LLMs under 16-bit, 8-bit, and 4-bit quantization, respectively. These plots summarize **36 evaluation settings** spanning multiple model families and datasets.

1. Semantic uncertainty methods consistently outperform probability-only baselines such as $p(\mathrm{True})$ and ER, particularly at lower rejection thresholds where reliable ranking of uncertain generations is most critical.
2. The proposed fragility-aware framework, SURF ($\mathrm{SE}^{\dagger}_{RU}$), maintains competitive RAC behavior across most datasets and quantization levels, often achieving smoother retained-accuracy degradation as increasingly uncertain outputs are rejected.
3. Compared to instruction-tuned models, non-instruction-tuned LLMs exhibit larger variability across rejection thresholds, reflecting increased instability and semantic dispersion in generated phrase-level outputs. In these settings, fragility-aware corrections provide improved robustness under repeated sampling.
4. Even under aggressive quantization, SURF remains competitive with existing semantic entropy approaches, demonstrating that the singleton-mass fragility correction improves uncertainty calibration in compressed and lower-capacity deployment settings.

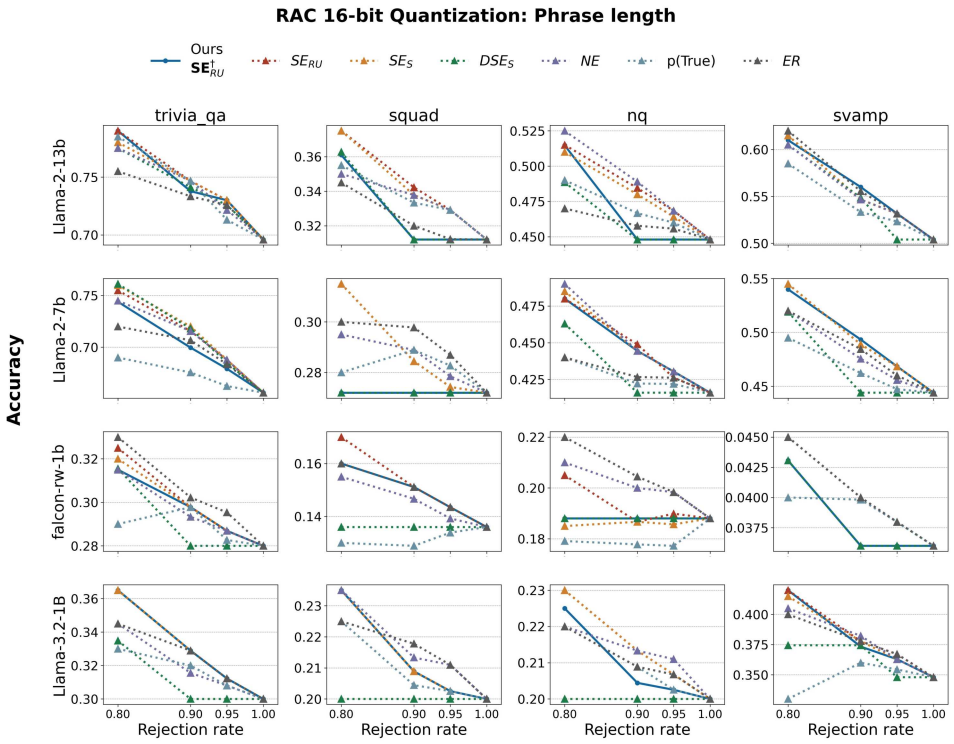


Figure 24: **Summarizing 12 evaluation settings**, RAC curves for hallucination detection across non-instruction-tuned LLMs and four datasets (TriviaQA, SQuAD, NQ, and SVAMP) under 16-bit quantization for phrase-level outputs. SURF ($\mathrm{SE}^{\dagger}_{RU}$) maintains competitive retained accuracy across rejection thresholds.

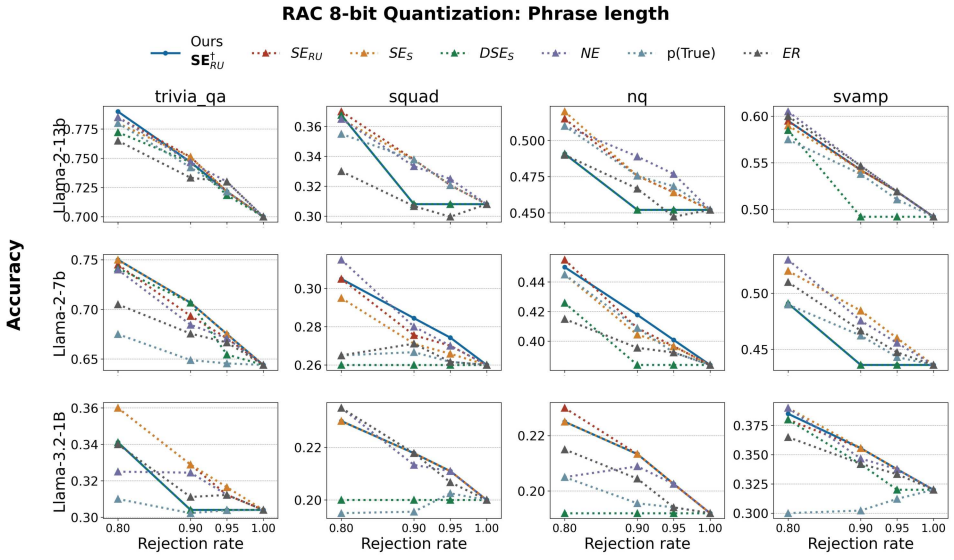


Figure 25: **Summarizing 12 evaluation settings**, RAC curves for hallucination detection across non-instruction-tuned LLMs and four datasets under 8-bit quantization for phrase-level outputs. The proposed SURF ($\mathrm{SE}^{\dagger}_{RU}$) remains competitive with the evaluated semantic uncertainty baselines under reduced precision.

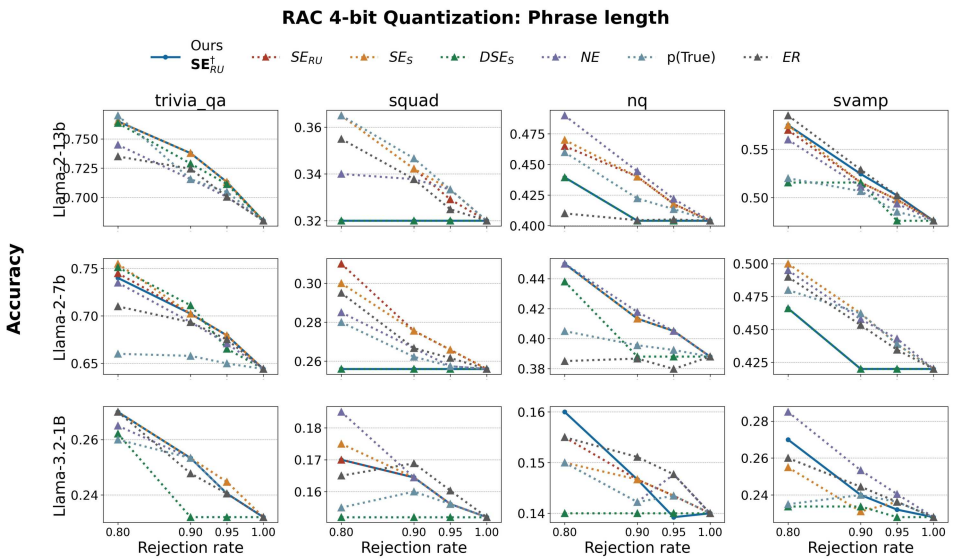


Figure 26: **Summarizing 12 evaluation settings**, RAC curves for hallucination detection across non-instruction-tuned LLMs and four datasets under 4-bit quantization for phrase-level outputs. Under aggressive compression, SURF ($\mathrm{SE}^{\dagger}_{RU}$) remains competitive across rejection thresholds.

C.5 Quantization Effects on Hallucinations

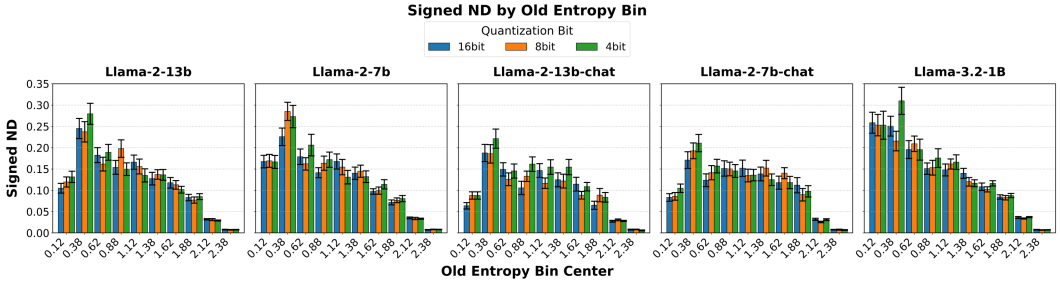


Figure 27: Signed normalized difference (ND) in entropy across old entropy bins for LLaMA-2 variants under different quantization settings (16-bit, 8-bit, 4-bit). Bars represent the mean signed ND with error bars indicating standard error. Notably, all models exhibit the largest variability in the 0.25–0.50 entropy range, suggesting this is a high-risk region where models frequently oscillate between multiple confident yet semantically divergent answers.

Fig. 27 shows that the effect of fragility-aware adjustments is non-uniform across entropy regions. The largest positive shifts in signed ND consistently occur in the mid-entropy range (=0.25–0.75), indicating that this is the regime where models exhibit the most semantic variability and are most sensitive to correction. In contrast, low and high-entropy regions remain stable, suggesting that highly confident predictions are largely unaffected. Importantly, this pattern is consistent across models and quantization levels, with even 4-bit settings preserving the same trend, demonstrating robustness under resource constraints. Overall, the results highlight that fragility adjustments are most impactful in high-risk intermediate regions, reinforcing the need for context-dependent rather than global thresholding strategies.

Figs. 28 and 29 illustrate the quantization trajectories of SURF ($\mathrm{SE}^{\dagger}_{RU}$) across different model families and datasets, showing the joint behavior of AUROC and AURAC-threshold performance under 16-bit, 8-bit, and 4-bit precision. Overall, quantization introduces a gradual degradation in both AUROC and retained selective accuracy, with the effect becoming more pronounced at 4-bit precision. Nevertheless, the trajectories remain relatively smooth, indicating that SURF preserves

stable uncertainty ranking behavior even under aggressive compression. Larger and instruction-tuned models generally maintain stronger robustness across quantization levels, while smaller non-instruction-tuned models exhibit larger shifts in both AUROC and AURAC. Importantly, the transition from 16-bit to 8-bit precision results in only marginal performance variation across most datasets, demonstrating that SURF can perform reliable hallucination detection under reduced precision settings without significant compromise in uncertainty estimation quality. Across most datasets, the movement from 16-bit to lower precision follows a consistent trade-off direction rather than abrupt collapse, suggesting that the SURF approach formulation remains resilient under deployment-constrained environments.

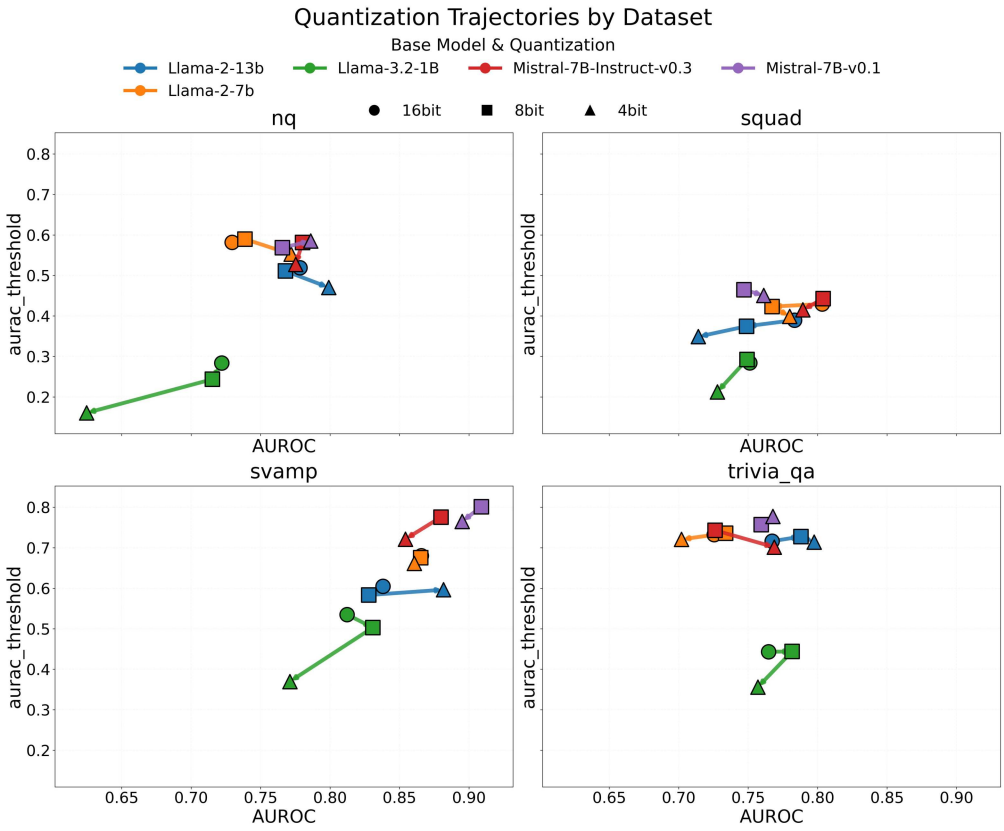


Figure 28: Quantization trajectories of our method at the **sentence level**, comparing AUROC (discrimination) and AURAC (rejection quality). Each trajectory corresponds to a base model, with markers indicating quantization levels (16-bit, 8-bit, 4-bit) and arrows showing the direction of quantization. We observe that quantization does not uniformly degrade performance; in several cases, lower-precision models achieve improved AUROC, indicating enhanced hallucination detection capability.

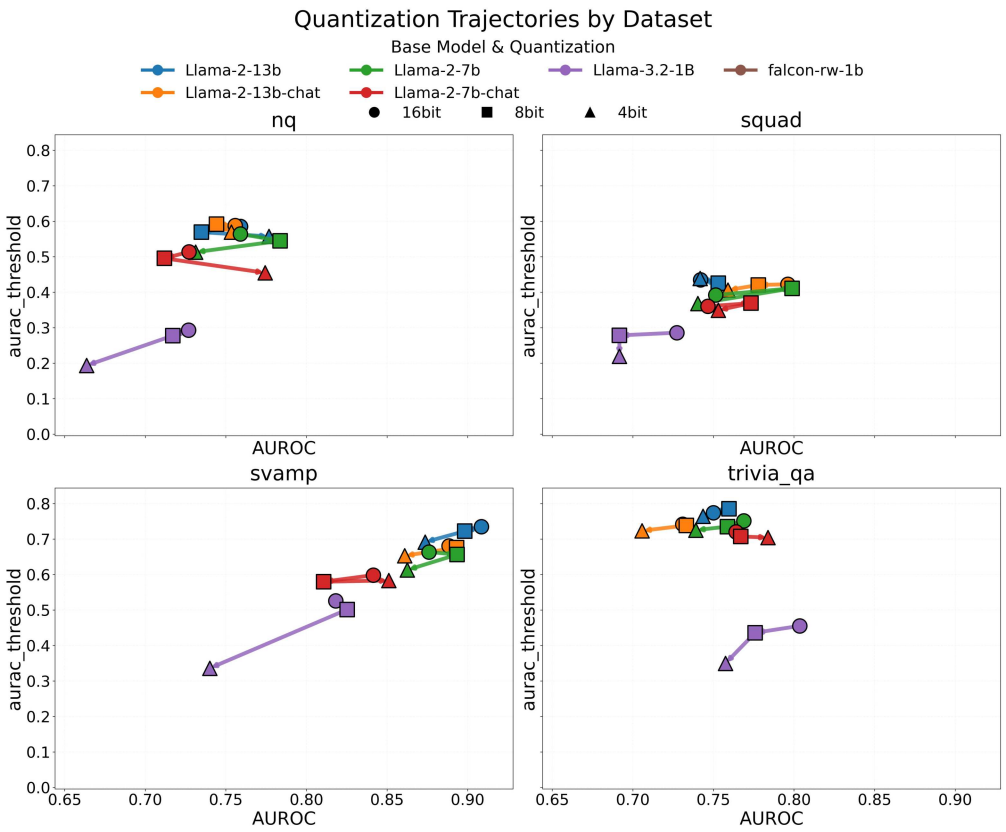


Figure 29: Quantization trajectories of our method at the **phrase level**, comparing AUROC and AURAC across datasets. Similar trends are observed as in the sentence-level setting: quantization can preserve or even improve discrimination performance (AUROC), though effects vary across datasets. This highlights that quantization acts as a data- and model-dependent transformation rather than a uniformly lossy compression step.

C.6 Hyperparameter Selection

Figs. 31 and 30 show the sensitivity of SURF ($\mathrm{SE}^{\dagger}_{RU}$) to the hyperparameters λ and γ in (6), using mean AUROC-HRG averaged across datasets. Positive values (red) indicate configurations where the fragility-aware adjustment improves over the reference method, whereas negative values (blue) indicate degradation.

For SURF, $\lambda \in \{5, 15, 25, 50, 75, 100\}$ and γ takes six logarithmically spaced values spanning 10^{-2} to 10^{4}, producing 36 candidate (λ, γ) combinations. The canonical configuration is $(\lambda, \gamma) = (25, 10)$.

The Rényi order is fixed at $\alpha = 2$. For the KME, the kernel bandwidth is set to five times the mean Silverman bandwidth of the standardized log-likelihood signal, with 0.01 used as a fallback when the bandwidth estimate is degenerate. The observed TS probability sensitivity construction uses $M = 8$ modes evaluated on a 128-point KME grid over $[-3, 3]$.

Overall, the heatmaps show that the best performance is obtained in red regions, where the entropy–fidelity trade-off and fragility correction are well balanced. Very blue regions correspond to overly weak or overly aggressive settings, where either fragility is insufficiently accounted for or the adjustment distorts the original TS probabilities. Across both sentence- and phrase-level settings, the presence of broad red regions for several models indicates that SURF can yield positive AUROC-HRG gains over a range of (λ, γ) choices, rather than depending on a single finely tuned configuration.

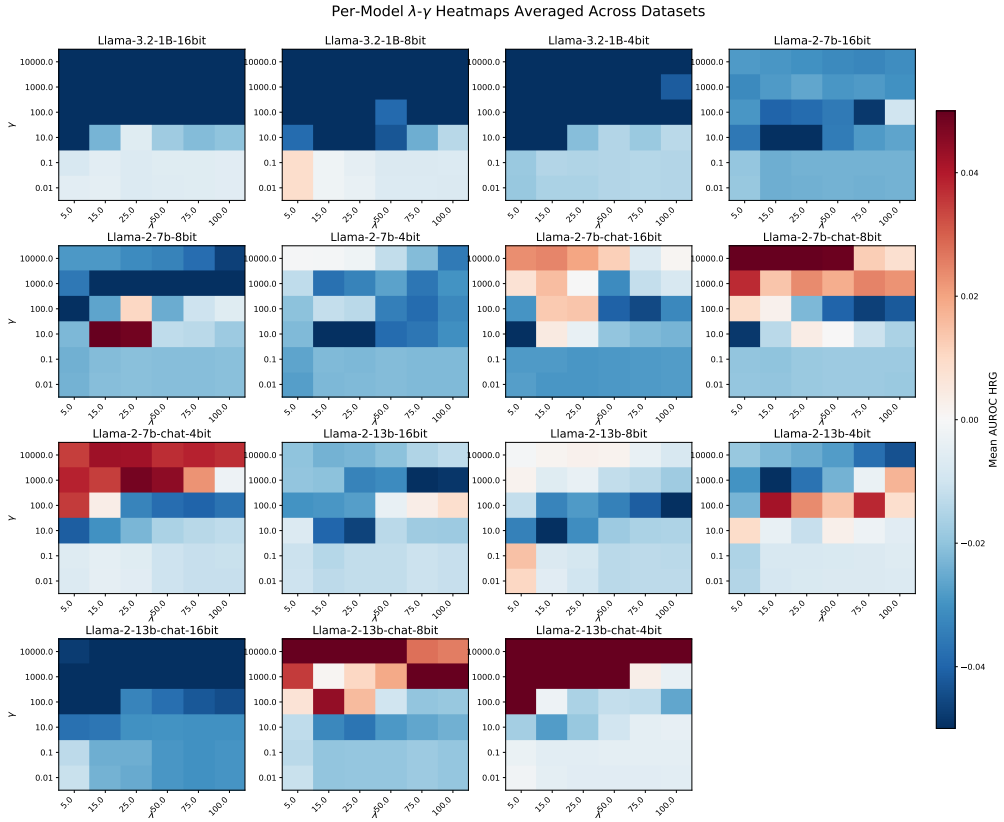


Figure 30: Effect of λ & γ hyperparameter selection on AUROC performance.

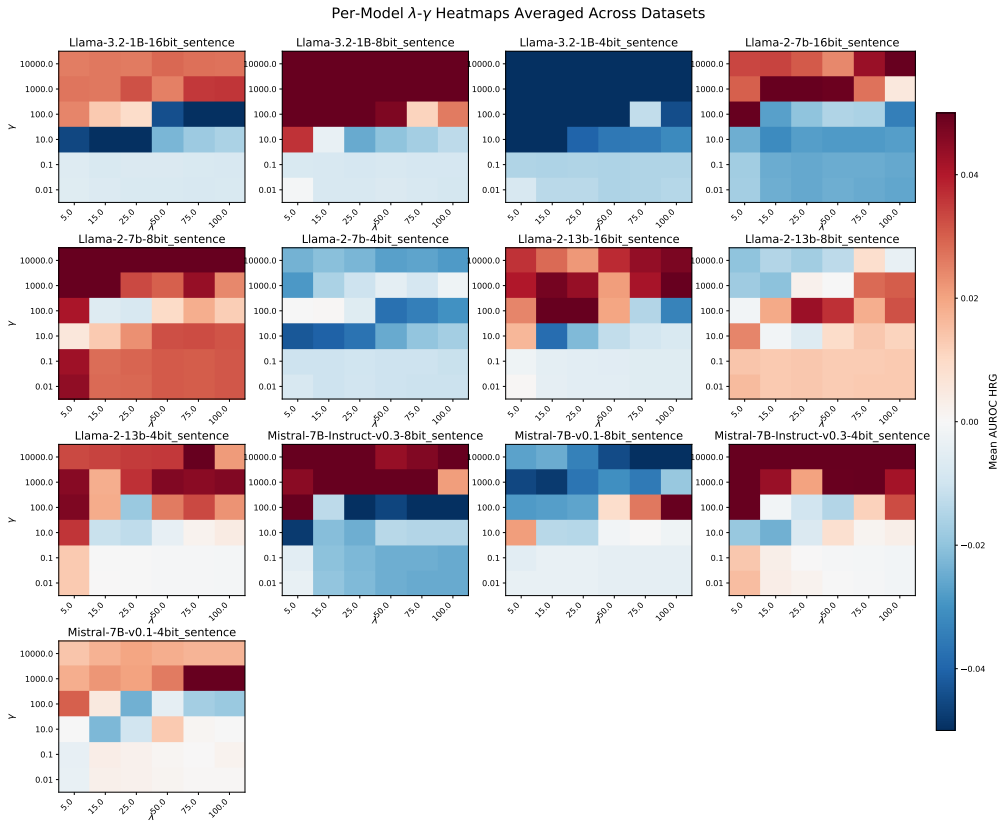


Figure 31: Effect of λ & γ hyperparameter selection on AUROC performance.

D Prompt Templates

D.1 Detailed Evaluation and Sampling Protocol

The 112 hallucination-detection settings consist of 28 realized model configurations crossed with the four datasets TriviaQA, NQ, SVAMP, and SQuAD. Each model configuration combines one of seven evaluated model variants with a precision in $\{16, 8, 4\}$ bit and either brief or single-sentence prompting. Because some model–precision–prompt combinations were not run, 28 of the possible combinations are present. In the implementation, runs carrying the `_sentence` suffix use `brief_prompt='chat'`.

For each question, generation serves two distinct roles. One answer is sampled at temperature $T = 0.1$ and graded against the reference answer; this is the answer whose correctness supplies the hallucination-detection label. Ten additional answers are sampled at $T = 1.0$ and used to construct the semantic clusters, occupancy statistics, and uncertainty score. Sampling is enabled for every generation call (`do_sample=True`), including the $T = 0.1$ graded answer; it is therefore a low-temperature sample rather than a greedy decode.

D.2 Generation Prompts and Semantic Clustering

Phrase Length Model Prompting and Answer Selection. The prompt template used for all datasets to generate LLM responses is

> *Answer the following question as briefly as possible:* {question}
> *Answer:*

Sentence length Model Prompting and Answer Selection. The prompt template used for all datasets to generate LLM responses is

Answer the following question in a single brief but complete sentence: {question}
Answer:

Entailment Estimation/Semantic Clustering. Semantic clustering uses `deberta-v2-xlarge-mnli`. The question is prepended to each generation before pairwise classification. The MNLI class is selected by hard argmax; there is no entailment-probability threshold. Two answers are placed in the same semantic cluster only if both directional comparisons are classified as entailment. Generations are processed sequentially and assigned greedily; when no existing cluster satisfies the strict bidirectional-entailment rule, a new cluster is created. The paired input is represented by the following template:

We are evaluating answers to the question: {question}

Here are two possible answers:
Possible Answer 1: {text1}
Possible Answer 2: {text2}

Does Possible Answer 1 semantically entail Possible Answer 2?
Respond with: entailment, contradiction, or neutral.

E Algorithmic Structures and Computational Cost

E.1 Computational Overhead

The computational overhead introduced by the observed TS probability sensitivity module is modest and remains well within real-time inference constraints. The method uses $M = 8$ perturbation modes evaluated on the 128-point KME grid described in App. C.6. Constructing the Hamiltonian associated with the sampled KME incurs a *one-time* pre-processing cost of approximately 45–60 s on CPU, though this can be substantially reduced through GPU-accelerated linear algebra routines. At inference time, the per-query overhead is minimal: evaluating the perturbation-based sensitivity features adds merely 6–10 ms on an NVIDIA A6000 GPU, which is negligible compared to the intrinsic latency of LLM decoding. Overall, this perturbation step introduces only a lightweight computational cost while enabling an observed TS probability sensitivity signal complementary to fragility.

E.2 Algorithmic Structure and Complexity

In this subsection, we present pseudocode illustrating key components of our approach, complementing the pipeline diagram provided in Fig. 1.

This includes pseudocode for (i) Hamiltonian construction from the sampled KME, (ii) extraction of first-order perturbation features, and (iii) the entropy-maximization procedure from Sec. 2.3.

Algorithm 1 Construction of Empirical KME $\widehat{\psi}(x)$

1: **Input:** Samples $\{x_i\}_{i=1}^{N}$ from random variable X; kernel bandwidth σ; sampling grid $\{x_j\}_{j=0}^{M-1}$
2: **Output:** Sampled KME vector $\widehat{\boldsymbol{\psi}} = \{\widehat{\psi}_j\}_{j=0}^{M-1}$
3: **for** $j = 0$ to $M - 1$ **do**
4: Compute Parzen PDF estimate:

$$\widehat{p}(x_j) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(x_j; x_i)$$

where $\kappa_\sigma(x_j; x_i)$ is the Gaussian kernel
5: Set KME value $\widehat{\psi}_j \leftarrow \widehat{p}(x_j)$
6: **end for**
7: **return** $\widehat{\psi}$

Algorithm 2 QTN Hamiltonian Construction via Quantum Correlation Matrix

1: **Input:** Sampled KME vector $\widehat{\boldsymbol{\psi}}$ of length $M = 2^L$; operator basis $\widehat{\mathcal{H}} = \{\widehat{\boldsymbol{H}}_k\}_{k=0}^{T-1}$
2: **Output:** Local Hamiltonian $\widehat{\boldsymbol{H}}$
3: Construct the QCM matrix using the inner products, where the (k, ℓ) entry is given by :

$$\mathbf{M}_{\widehat{\boldsymbol{\psi}}}(k, \ell) = \langle \widehat{\boldsymbol{H}}_k \widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{H}}_\ell \, \widehat{\boldsymbol{\psi}} \rangle$$

4: Compute null space of QCM:

$$\mathbf{w} = \{w_k\}_{k=0}^{T-1} \in \mathrm{Null}(\mathbf{M}_{\widehat{\boldsymbol{\psi}}})$$

5: Form Hamiltonian as linear combination:

$$\widehat{\boldsymbol{H}} = \sum_{k=0}^{T-1} w_k \widehat{\boldsymbol{H}}_k$$

6: **return** $\widehat{\boldsymbol{H}}$

Algorithm 3 First-Order Perturbation Corrections and TS UQ

1: **Input:** Base Hamiltonian

$$\widehat{\boldsymbol{H}} = \sum_{k=0}^{T-1} w_k \, \widehat{\boldsymbol{H}}_k,$$

where $\{w_k\}$ are the learned weights multiplying the Hermitian operator basis; perturbation vector $\Delta\mathbf{w}$ specifying perturbation of these weights.
2: **Output:** First-order corrected eigen-modes $\{\psi_m^{(1)}\}$ and the resulting UQ feature vector.
3: Construct the perturbed Hamiltonian:

$$\Delta \boldsymbol{H} = \sum_{k=0}^{T-1} \Delta w_k \, \widehat{\boldsymbol{H}}_k.$$

This corresponds to perturbing the coefficients (weights) associated with the Pauli-like operators forming the QTN Hamiltonian.
4: Compute eigen-decomposition of the unperturbed Hamiltonian:

$$\widehat{\boldsymbol{H}} \, \psi_m = E_m \, \psi_m, \qquad m = 0, \ldots, M-1.$$

5: **for** $m = 0$ to $M - 1$ **do**
6: Compute first-order energy correction:

$$E_m^{(1)} = \langle \psi_m, \, \Delta \boldsymbol{H} \, \psi_m \rangle.$$

7: Compute first-order eigenvector correction:

$$\psi_m^{(1)} = \sum_{n \neq m} \frac{\langle \psi_n, \, \Delta \boldsymbol{H} \, \psi_m \rangle}{E_m - E_n} \, \psi_n.$$

8: **end for**
9: We utilize (26) and $\{\psi_m^{(1)}\}$ to compute UQ feature vectors in the KME amplitude domain, yielding the local UQ scores $\mathrm{UQ}(p_s^{(r)})$.
10: **return** $\{\psi_m^{(1)}\}$ & $\mathrm{UQ}(p_s^{(r)})$.

Algorithm 4 Calibrated Entropy-Maximization Update for TS Probabilities

Input:
- Original TS probability $p_s^{(r)}$
- Uncertainty score $\mathrm{UQ}(p_s^{(r)})$
- Regularization parameter λ

Output: Adjusted probability $p_s^{(r)^*}$

Define objective:

$$\mathcal{L}(\widehat{p}) = -\log\big(\widehat{p}^2 + (1-\widehat{p})^2\big) - \frac{\lambda}{\mathrm{UQ}(p_s^{(r)})} \cdot D_{\mathrm{KL}}\left(\widehat{p} \,\|\, p_s^{(r)}\right)$$

where $D_{\mathrm{KL}}(\widehat{p}\|p) = \widehat{p}\log\frac{\widehat{p}}{p} + (1-\widehat{p})\log\frac{1-\widehat{p}}{1-p}$.

Initialize: $\widehat{p}^{(0)} \leftarrow p_s^{(r)}$

for $t = 1$ to $T_{\max}$ **do**
  Compute gradient $\nabla\mathcal{L}(\widehat{p}^{(t-1)})$
  Update via projected gradient ascent
  Stop early if $|\widehat{p}^{(t)} - \widehat{p}^{(t-1)}| < \delta$
end for

Set $p_s^{(r)^*} \leftarrow \widehat{p}^{(T)}$

return $p_s^{(r)^*}$

F Extended Evaluation of Fragility

In this section, we provide an extended evaluation of fragility across a diverse array of settings.

F.1 Upper Confidence Risk (UCR).

To summarize estimator reliability across fragility regimes, the *UCR* metric is used (Casella & Berger, 2002). Let the fragility range be partitioned into bins indexed by i. For each bin, let μ_i and σ_i denote the empirical mean and standard deviation of the prediction error across trials. Let w_i denote the fraction of samples falling into bin i. Then, the UCR is defined as

$$\mathrm{UCR} = \sum_i w_i \,(\mu_i + z\,\sigma_i), \tag{28}$$

where $z = 1.96$ corresponds to a 95% confidence bound. The UCR penalizes both systematic bias (via μ_i) and instability across trials (via σ_i). Lower values therefore indicate estimators that are both accurate and reliable.

In addition to scalar summaries, performance is analyzed as a function of the fragility measure. Errors are binned according to the observed fragility level, and mean error curves are plotted across bins. This analysis provides fine-grained insight into the regimes where fragility-aware corrections offer the largest improvement over classical estimators, particularly in high-fragility regions corresponding to singleton-dominated samples. This allows both global comparisons of estimator reliability and detailed inspection of performance across different data regimes.

Fig. 32 summarizes the unseen-species UCR win-rate analysis across synthetic distributions. The fragility-aware estimator consistently outperforms classical unseen-species estimators over the aggregated evaluation settings.

F.2 Fragility-Aware Missing-Mass

Alongside the unseen-species count formulation in App. A, one can define the unseen probability mass from a sampled pool.

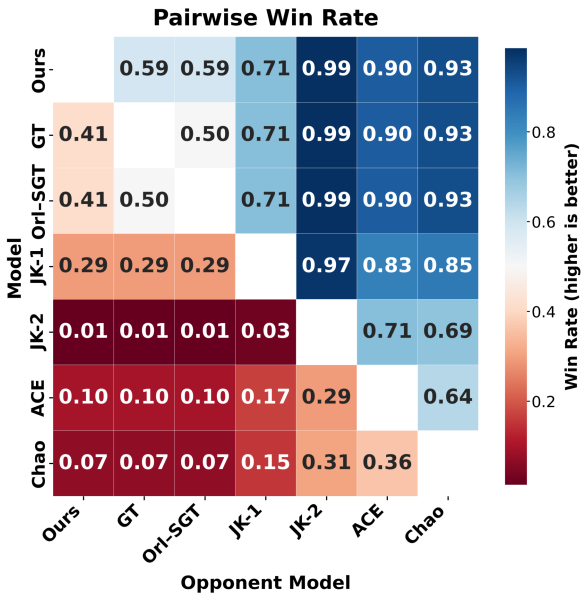


Figure 32: Pairwise UCR win-rate comparison across synthetic distributions and **144 experiments** for unseen-species estimation.

Consider the observed mass and the mass of k-tons:

$$P_{\text{obs}} = \sum_{\ell} p_\ell \, \mathbb{1}(N_\ell > 0); \;\; P_k = \sum_{\ell} p_\ell \, \mathbb{1}(N_\ell = k). \tag{29}$$

Expanding on the details mentioned in Sec. 2.2, the *fragility measure* we intend to utilize and the *regularization strength* it maps to are, respectively,

$$F_{\text{mass}} = P_1 / P_{\text{obs}}; \;\; \alpha\,(F_{\text{mass}}) = \alpha_0 \, e^{\gamma \, F_{\text{mass}}}, \tag{30}$$

where $\alpha_0 > 0$ and $\gamma < 0$ are hyperparameters. Higher F_{mass} values indicates that the observed region of the distribution is dominated by rare species and therefore provides limited information about the unseen region.

As for the (usually unknown) population frequencies p_ℓ, this study employs a proxy: for synthetic distributions, they are known; for real datasets, the empirical histogram is regularized using a Zipf model with exponent close to one, reflecting the well-established heavy-tailed structure of natural language corpora (Baayen, 2001).

The proposed fragility-aware missing-mass estimate ((10)) is given by

$$M_0^\star = \underset{M_0 \in (0,1)}{\arg\max} \left\{ h(M_0) - \alpha(F_{\text{mass}}) \, D_{\text{KL}}(M_0 \| A) \right\}.$$

Here, A is a base missing-mass estimator (taken to be the GT estimator $\widehat{M}_{\text{GT}} = \Phi_1/n$) that is subsequently adjusted according to the fragility of the observed sample, $h(M_0) = -M_0 \log(M_0) - (1 - M_0) \log(1 - M_0)$ is the Bernoulli entropy, and

$$D_{\text{KL}}(M_0 \| A) = M_0 \log(M_0/A) + (1 - M_0) \log((1 - M_0)/(1 - A)) \tag{31}$$

is the Kullback–Leibler divergence (KL) between Bernoulli parameters M_0 and A.

The entropy term in the fragility-aware formulation in (10) encourages uncertainty when the sample is fragile, while the KL term preserves proximity to the base estimate. In essence, balancing between the trust in the base estimator and maximizing entropy. It can be interpreted as a maximum-entropy correction of the base estimator, where the magnitude of the correction is controlled by a data-driven singleton-based fragility measure.

F.3 Synthetic Distributions

The synthetic datasets that we employed—uniform, Zipf, Dirichlet, and head-tail mixtures—span smooth, moderately skewed, and highly fragmented probability landscapes, enabling systematic analysis of estimator behavior under varying levels of rarity and tail heaviness.

Performance is summarized using pairwise win-rate comparisons based on the UCR, where each cell indicates the probability that the row model outperforms the column model. This aggregates both bias and variability of error across sampling regimes. win-rates therefore quantify how often one estimator yields lower risk than another across repeated trials.

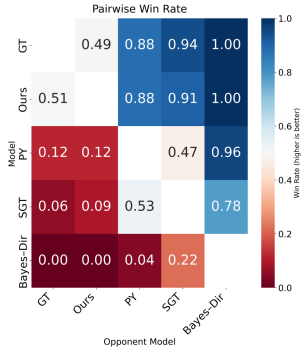


Figure 33: Pairwise win-rate comparison for fragility-aware missing-mass estimation across all synthetic distributions and 144 experiments. The proposed method generally outperforms competing baselines. This complements the unseen-species win-rate analysis shown in Fig. 32.

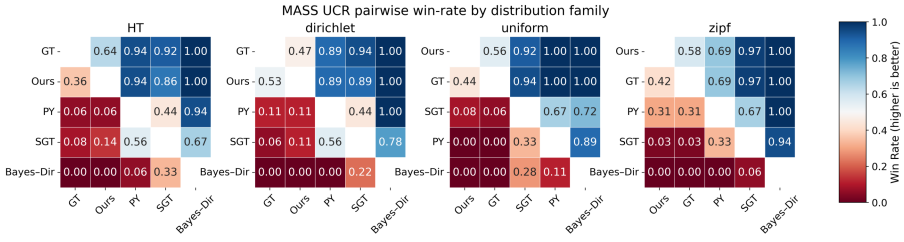


Figure 34: Pairwise win-rates (72 experiments) for missing-mass estimation across synthetic distribution families against SOTA methods. In fragmented regimes such as Zipf or Dirichlet, the fragility-aware estimator consistently maintains strong win-rates, reflecting its ability to down weight unstable singleton-driven signals. These results indicate that the proposed adjustment generally outperforms SOTA baselines.

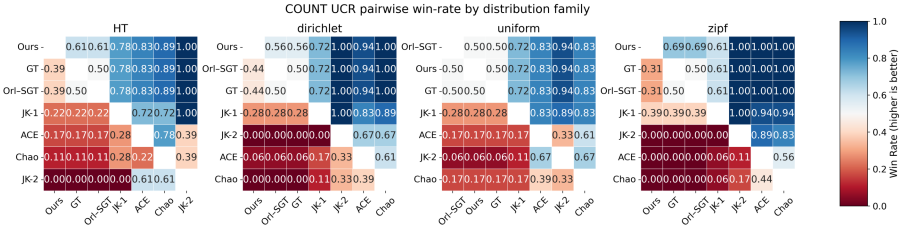


Figure 35: Pairwise win-rates (72 experiments) for unseen-species prediction across synthetic distribution families against SOTA methods. The fragility-aware estimator achieves consistently strong win-rates across SOTA methods, indicating that incorporating a stability signal mitigates the variance amplification associated with rare-event extrapolation.

Fig 33 shows that, in missing-mass estimation, the proposed fragility-aware method matches or exceeds SOTA estimators in most regimes, indicating that the entropy-based adjustment preserves

accuracy while improving stability. A finer-grained view in Fig 34 reveals that this behavior is consistent across distribution families, with one notable exception in HT distribution. The structured mass split makes singleton counts relatively stable, allowing GT estimator to outperform; further analysis of this behavior is presented in App F.3.4.

Fig 32 illustrates that our method remains competitive with discovery-oriented estimators such as GT and SGT, while avoiding the high variance observed for jackknife and coverage-based estimators. These results suggest that incorporating fragility yields a more uniformly reliable estimator across heterogeneous sampling conditions. As shown in Fig 35, the fragility-aware method performs consistently against SOTA estimators, across all distributions.

A more detailed analysis across individual parameter settings for each distribution, including per-sample-size breakdowns, appears in App. F.3, where the same qualitative trends remain. The proposed fragility-aware adjustment explicitly accounts for this sensitivity, leading to improved reliability in difficult regimes while preserving accuracy in easier ones. This suggests that fragility captures an intrinsic property rather than a distribution-specific heuristic, making it a useful signal for adaptive estimator design.

F.3.1 Uniform Distribution

To study estimator behavior under controlled distributional structure, experiments were conducted on synthetic uniform distributions with varying support size S. Here, S denotes the total number of species in the population while the total probability mass is fixed. Increasing S produces increasingly sparse observations for a fixed sample size N. This setting isolates the effect of distributional fragmentation: larger S disperses probability mass across many low-frequency events, increasing sensitivity to sampling variability and rare-event noise. Table 3, Table 4 report the UCR across sample sizes and support sizes for both missing-mass estimation and unseen-species prediction, allowing the examination on how estimator reliability evolves under increasing sparsity in terms of both probability mass and discovery counts. Figures 36 and 37 provide a complementary perspective by visualizing the mean estimation error and its variability fragility regions, highlighting how estimator stability evolves with increasing sparsity for both missing mass and unseen-species prediction.

Parameter S	N	**Ours**	**GT**	**SGT**	**BD**	**PY**
200	25	**0.168**	**0.168**	0.184	0.483	0.942
200	50	**0.168**	**0.168**	0.172	0.421	0.854
200	100	**0.123**	0.124	0.230	0.287	0.676
200	200	**0.107**	0.108	0.315	0.119	0.434
200	500	**0.047**	**0.047**	0.203	0.136	0.120
200	1000	**0.013**	**0.013**	0.023	0.104	0.017
500	25	0.112	**0.111**	**0.111**	0.512	0.985
500	50	0.113	**0.112**	0.160	0.489	0.956
500	100	**0.088**	0.089	0.160	0.413	0.858
500	200	**0.086**	0.088	0.275	0.299	0.707
500	500	**0.057**	**0.057**	0.457	0.079	0.391
500	1000	**0.037**	**0.037**	0.373	0.108	0.160
1000	25	**0.085**	**0.085**	**0.085**	0.510	0.986
1000	50	0.094	**0.093**	0.105	0.505	0.984
1000	100	**0.081**	**0.081**	0.105	0.472	0.940
1000	200	**0.072**	**0.072**	0.289	0.400	0.849
1000	500	**0.066**	0.067	0.468	0.233	0.618
1000	1000	**0.044**	**0.044**	0.595	0.063	0.370

Table 3: **Uniformly distributed population:** UCR (lower is better) for missing-mass estimation across support sizes S and sample sizes N. Larger S induces greater sparsity and rare-event dominance. The fragility-aware estimator remains competitive with classical methods while improving stability under increasing fragmentation.

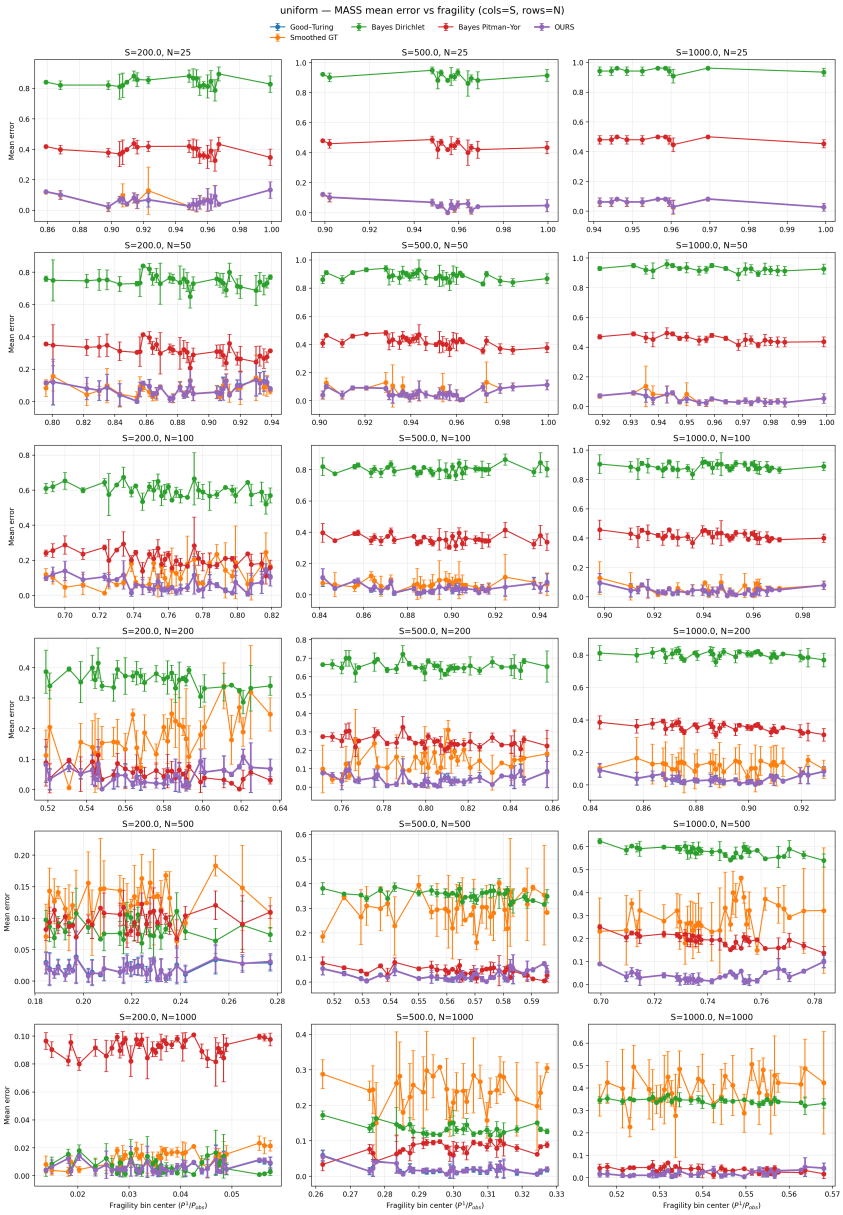


Figure 36: **summarizing 18 experimental** scenarios, under uniform distributions, mean estimation error across fragility regimes is shown for varying population sizes S and sample sizes N. Each subplot corresponds to a specific (S, N) pair, with error bars indicating variability across repeated trials. The fragility-aware estimator remaining competitive against classical methods in low-sample regimes as sample size grows.

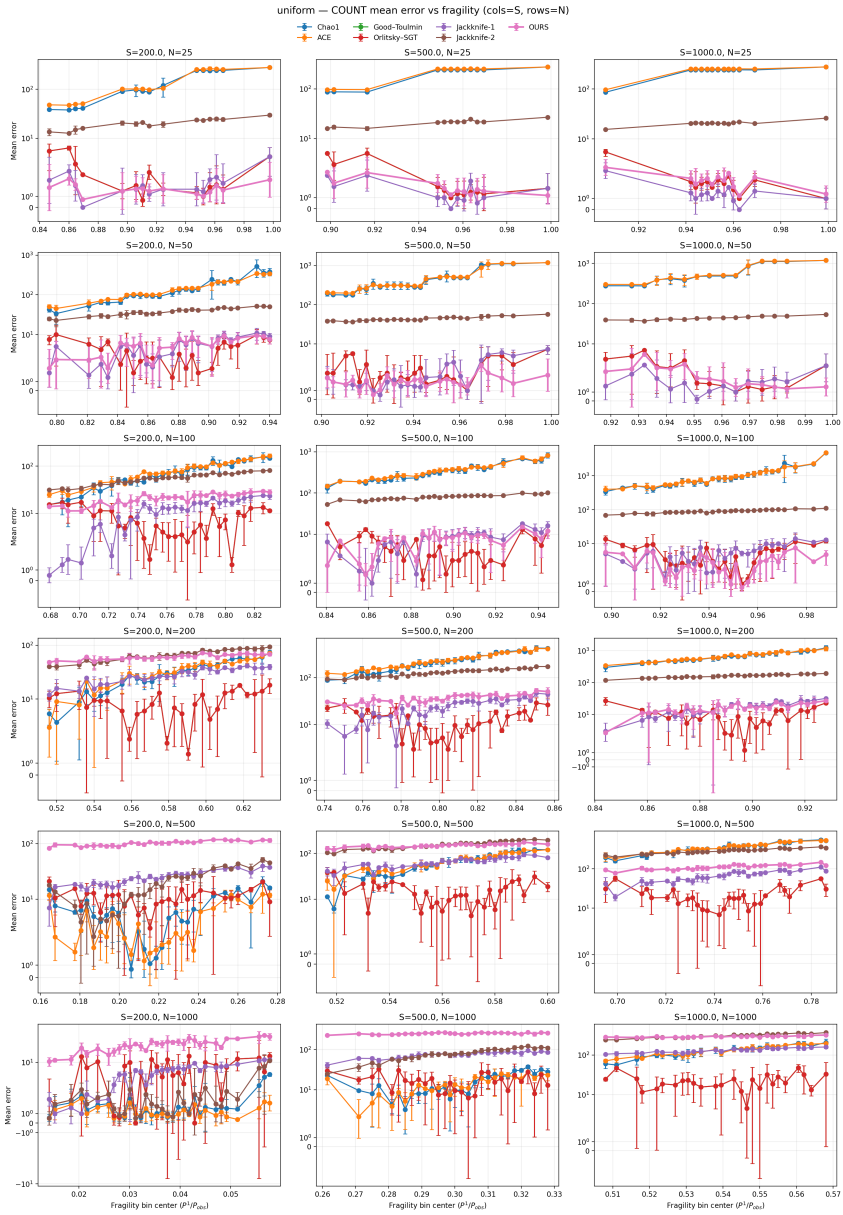


Figure 37: **summarizing 18 experimental** scenarios, under uniform distributions, mean unseen-species estimation error is shown across fragility regimes for varying population sizes S and sample sizes N. Each subplot corresponds to a specific (S, N) pair, with error bars (on a symmetric logarithmic (symlog) scale) reflecting variability across trials. The fragility-aware estimator demonstrates stable performance across regimes, often outperforming classical estimators in data-scarce settings while remaining competitive at larger sample sizes.

Parameter S	N	**Ours**	**G-Toulmin**	**Orlitsky-SGT**	**Jackknife-1**	**Jackknife-2**	**Chao1**	**ACE**
200	25	**5.435**	5.442	5.442	5.577	26.737	211.597	215.907
200	50	**10.239**	10.258	10.258	10.902	46.118	255.169	185.015
200	100	**13.573**	13.604	13.604	19.381	65.113	101.960	96.771
200	200	**17.968**	18.026	18.026	37.116	73.424	39.992	43.215
200	500	21.293	21.146	21.146	30.291	25.780	9.795	**8.569**
200	1000	18.130	17.951	17.951	8.211	5.299	3.135	**2.562**
500	25	4.161	4.148	4.148	**3.720**	26.877	256.536	261.082
500	50	6.800	6.791	6.791	**6.264**	50.206	629.432	614.722
500	100	**10.258**	10.337	10.337	15.043	89.050	437.921	442.582
500	200	**21.731**	21.794	21.794	34.262	142.537	281.562	278.157
500	500	28.299	**28.219**	**28.219**	81.819	171.223	86.473	87.871
500	1000	38.736	38.630	38.630	87.401	95.639	24.040	**23.419**
1000	25	3.079	3.070	3.070	**2.767**	26.070	264.657	267.909
1000	50	5.863	5.862	5.862	**5.238**	51.535	921.149	939.512
1000	100	**11.051**	11.067	11.067	11.881	99.093	1,336.655	1,270.199
1000	200	**17.648**	17.677	17.677	26.509	174.058	792.281	791.569
1000	500	**36.959**	37.051	37.051	82.191	304.287	408.656	397.535
1000	1000	46.669	**46.466**	**46.466**	153.834	336.670	169.980	161.654

Table 4: **Uniformly distributed population:** UCR (lower is better) for unseen-species estimation across support sizes S and sample sizes N. Larger S induces greater sparsity and rare-event dominance. The fragility-aware estimator remains competitive with classical methods while improving stability under increasing fragmentation.

F.3.2 DIRICHLET DISTRIBUTION

To study estimator behavior under controlled random heterogeneity in discrete distributions, experiments were conducted on synthetic symmetric Dirichlet–multinomial populations. Specifically, a probability vector is sampled as $p \sim \mathrm{Dirichlet}(\alpha \mathbf{1})$. Smaller α produces higher variability in the realized probability vectors, leading to heterogeneous outcomes with uneven mass allocation across species, while larger α yields draws concentrated near the uniform distribution. Tables 5 and 6 report UCR across sample sizes and α for both missing-mass estimation and unseen-species prediction, enabling comparison of estimator stability under increasing distributional dispersion. Figures 38 and 39 complement these summaries by showing mean estimation error and its variability across fragility regimes for each (α, N) scenario.

Parameter α	N	**Ours**	**GT**	**SGT**	**BD**	**PY**
0.3	25	**0.209**	**0.209**	0.267	0.496	0.937
0.3	50	**0.186**	0.187	0.206	0.392	0.790
0.3	100	**0.144**	0.144	0.182	0.275	0.623
0.3	200	**0.106**	**0.106**	0.191	0.120	0.407
0.3	500	**0.053**	**0.053**	0.166	0.067	0.191
0.3	1000	**0.026**	0.027	0.051	0.069	0.094
1.0	25	0.155	**0.154**	**0.154**	0.510	0.973
1.0	50	**0.171**	0.172	0.213	0.458	0.898
1.0	100	**0.133**	**0.133**	0.236	0.356	0.766
1.0	200	**0.095**	0.096	0.281	0.222	0.575
1.0	500	0.052	**0.051**	0.308	0.050	0.282
1.0	1000	**0.030**	**0.030**	0.143	0.083	0.130
3.0	25	0.131	**0.130**	**0.130**	0.522	0.990
3.0	50	**0.132**	**0.132**	0.154	0.476	0.936
3.0	100	**0.106**	**0.106**	0.170	0.395	0.827
3.0	200	**0.108**	0.109	0.284	0.279	0.661
3.0	500	**0.060**	**0.060**	0.411	0.058	0.347
3.0	1000	0.032	**0.031**	0.270	0.095	0.140

Table 5: **Symmetric Dirichlet multinomial distributed population:** UCR (lower is better) for missing-mass estimation across α and sample sizes N. Smaller α produces higher variability in the realized probability vectors, leading to heterogeneous outcomes with uneven mass allocation across species. The fragility-aware estimator remains competitive with classical methods while improving stability under increasing fragmentation.

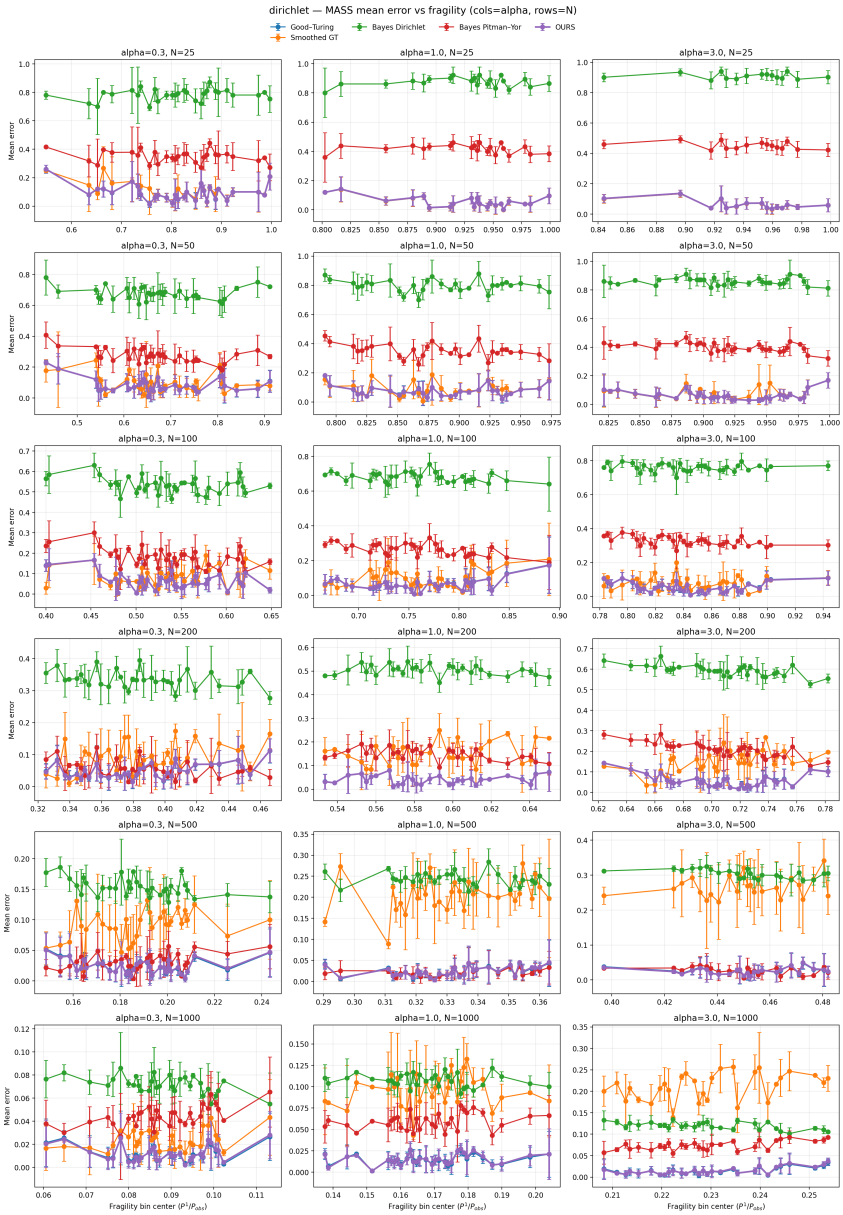


Figure 38: **summarizing 18 experimental** scenarios under Dirichlet-generated distributions, mean missing-mass estimation error is shown across fragility regimes for varying concentration parameters α and sample sizes N. Each subplot corresponds to a specific (α, N) pair, with error bars indicating variability across repeated trials. Across regimes, the fragility-aware estimator adapts to distributional uncertainty, maintaining competitive accuracy for smooth distributions and yielding improved stability in highly fragmented settings.

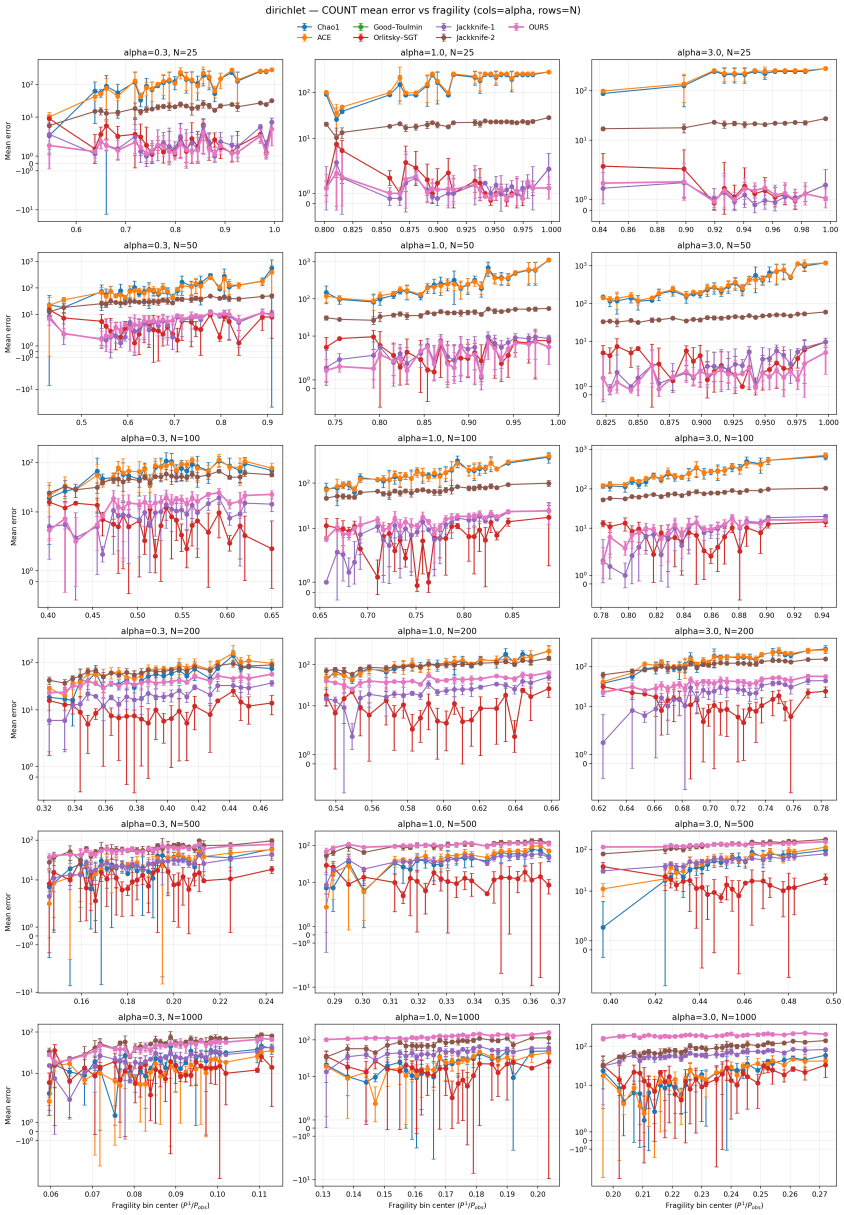


Figure 39: **summarizing 18 experimental** scenarios under Dirichlet-generated distributions, mean unseen-species estimation error is shown across fragility regimes for varying concentration parameters α and sample sizes N. Each subplot corresponds to a specific (α, N) pair, with error bars reflecting variability across trials (on a symmetric logarithmic (symlog) scale). As α decreases the classical estimators exhibit increasing variance, while the fragility-aware method maintains stable behavior and often achieves lower error in sparse regimes, remaining competitive as data size grows.

Parameter α	N	**Ours**	**G-Toulmin**	**Orlitsky-SGT**	**Jackknife-1**	**Jackknife-2**	**Chao1**	**ACE**
0.3	25	7.084	7.098	7.098	**6.417**	27.885	254.570	256.952
0.3	50	11.213	11.247	11.247	**11.049**	45.665	256.863	202.047
0.3	100	**14.704**	14.746	14.746	20.595	66.187	128.606	125.656
0.3	200	22.074	**22.070**	**22.070**	32.825	86.133	89.184	105.025
0.3	500	27.493	**27.377**	**27.377**	38.688	86.424	55.049	48.655
0.3	1000	31.260	**30.953**	**30.953**	37.395	77.312	43.453	26.231
1.0	25	4.848	**4.842**	**4.842**	4.679	27.248	256.567	267.411
1.0	50	**10.411**	10.435	10.435	10.830	52.669	511.866	459.114
1.0	100	14.048	**14.099**	**14.099**	18.832	82.163	250.517	239.518
1.0	200	**19.850**	19.923	19.923	34.966	114.104	153.271	149.221
1.0	500	**29.893**	29.912	29.912	58.912	131.877	78.466	76.954
1.0	1000	38.402	**38.341**	**38.341**	62.644	108.658	40.975	41.694
3.0	25	4.531	4.515	4.515	**3.887**	27.224	279.444	287.859
3.0	50	8.285	8.279	8.279	**7.711**	50.725	643.188	635.944
3.0	100	**14.504**	14.542	14.542	16.758	87.382	394.088	380.217
3.0	200	**25.693**	25.755	25.755	37.366	132.705	208.646	210.272
3.0	500	**31.231**	31.259	31.259	69.785	154.659	84.542	85.253
3.0	1000	**35.319**	34.974	34.974	75.799	115.731	36.853	36.242

Table 6: **Symmetric Dirichlet multinomial distributed population:** UCR (lower is better) for unseen species count across α and sample size N. Smaller α produces higher variability in the realized probability vectors, leading to heterogeneous outcomes with uneven count allocation across species. The fragility-aware estimator remains competitive with classical methods while improving stability under increasing fragmentation.

F.3.3 Zipf Distribution

To study estimator behavior under structured heavy-tailed distributions, experiments were conducted on synthetic Zipf populations. Specifically, probabilities are generated according to a power-law decay $p_i \propto i^{-s}$ over S species. The exponent s controls the degree of tail heaviness: smaller s produces flatter distributions with relatively balanced mass across ranks, while larger s concentrates probability on the highest-ranked species and yields stronger head–tail imbalance. Tables 7 and 8 report UCR across sample sizes and exponents s for both missing-mass estimation and unseen-species prediction, enabling comparison of estimator stability as distributions transition from near-uniform to strongly heavy-tailed regimes. Figures 40 and 41 complement these summaries by showing mean estimation error and its variability across fragility regimes for each (s,N) scenario.

Parameter s	N	Ours	GT	SGT	BD	PY
0.5	25	**0.149**	**0.149**	**0.149**	0.510	0.978
0.5	50	0.133	**0.132**	0.201	0.463	0.914
0.5	100	**0.123**	**0.123**	0.242	0.389	0.803
0.5	200	**0.099**	0.100	0.194	0.259	0.628
0.5	500	0.065	**0.064**	0.280	0.081	0.356
0.5	1000	**0.033**	**0.033**	0.255	0.076	0.168
1.0	25	**0.271**	0.272	0.395	0.406	0.758
1.0	50	**0.199**	0.200	0.375	0.321	0.644
1.0	100	**0.121**	0.122	0.259	0.221	0.504
1.0	200	**0.094**	0.095	0.182	0.175	0.402
1.0	500	0.050	**0.049**	0.068	0.069	0.240
1.0	1000	0.032	**0.031**	0.080	0.028	0.153
1.5	25	0.251	0.252	0.292	**0.235**	0.434
1.5	50	0.147	0.147	0.181	**0.138**	0.313
1.5	100	0.109	0.109	0.139	**0.100**	0.234
1.5	200	**0.062**	**0.062**	0.086	**0.062**	0.161
1.5	500	**0.034**	**0.034**	0.049	0.035	0.105
1.5	1000	0.018	0.019	0.026	**0.017**	0.068

Table 7: **Zipf distributed population:** UCR (lower is better) for missing-mass estimation across s and sample sizes N. Larger s induces stronger head–tail imbalance, concentrating probability mass on a few dominant species while leaving many low-probability events. The fragility-aware estimator remains competitive with classical methods while improving stability under increasing fragmentation.

Parameter s	N	Ours	G-Toulmin	Orlitsky-SGT	Jackknife-1	Jackknife-2	Chao1	ACE
0.5	25	5.323	5.320	5.320	**4.610**	28.306	290.706	301.441
0.5	50	**8.670**	8.696	8.696	9.251	51.845	743.289	649.588
0.5	100	**15.188**	15.237	15.237	15.957	86.381	403.385	379.102
0.5	200	**22.974**	23.039	23.039	35.691	134.974	262.261	328.278
0.5	500	**33.862**	33.942	33.942	69.954	166.451	110.984	156.909
0.5	1000	32.609	**32.527**	**32.527**	79.435	123.943	37.764	58.402
1.0	25	8.317	8.339	8.339	**6.951**	23.043	153.805	121.992
1.0	50	10.992	11.021	11.021	**10.683**	38.012	253.840	186.293
1.0	100	**13.307**	13.341	13.341	14.669	57.893	200.923	203.862
1.0	200	**21.430**	21.536	21.536	22.722	86.562	194.927	229.652
1.0	500	**30.608**	30.682	30.682	43.343	127.809	180.810	198.100
1.0	1000	30.321	**30.116**	**30.116**	56.093	141.454	113.372	148.172
1.5	25	7.363	7.373	7.373	**5.893**	14.103	40.336	37.820
1.5	50	7.830	7.830	7.830	**7.328**	19.801	64.457	47.897
1.5	100	12.239	12.238	12.238	**11.495**	29.997	115.519	73.717
1.5	200	13.533	13.534	13.534	**12.802**	39.690	114.960	86.982
1.5	500	18.445	**18.437**	**18.437**	19.987	59.780	116.750	99.287
1.5	1000	**20.793**	20.795	20.795	21.934	69.922	111.718	96.191

Table 8: **Zipf distributed populations:** UCR (lower is better) for unseen-species counts across s and sample sizes N. Larger s induces stronger head–tail imbalance, concentrating probability mass on a few dominant species while leaving many low-probability events. The fragility-aware estimator remains competitive with classical methods while improving stability under increasing fragmentation.

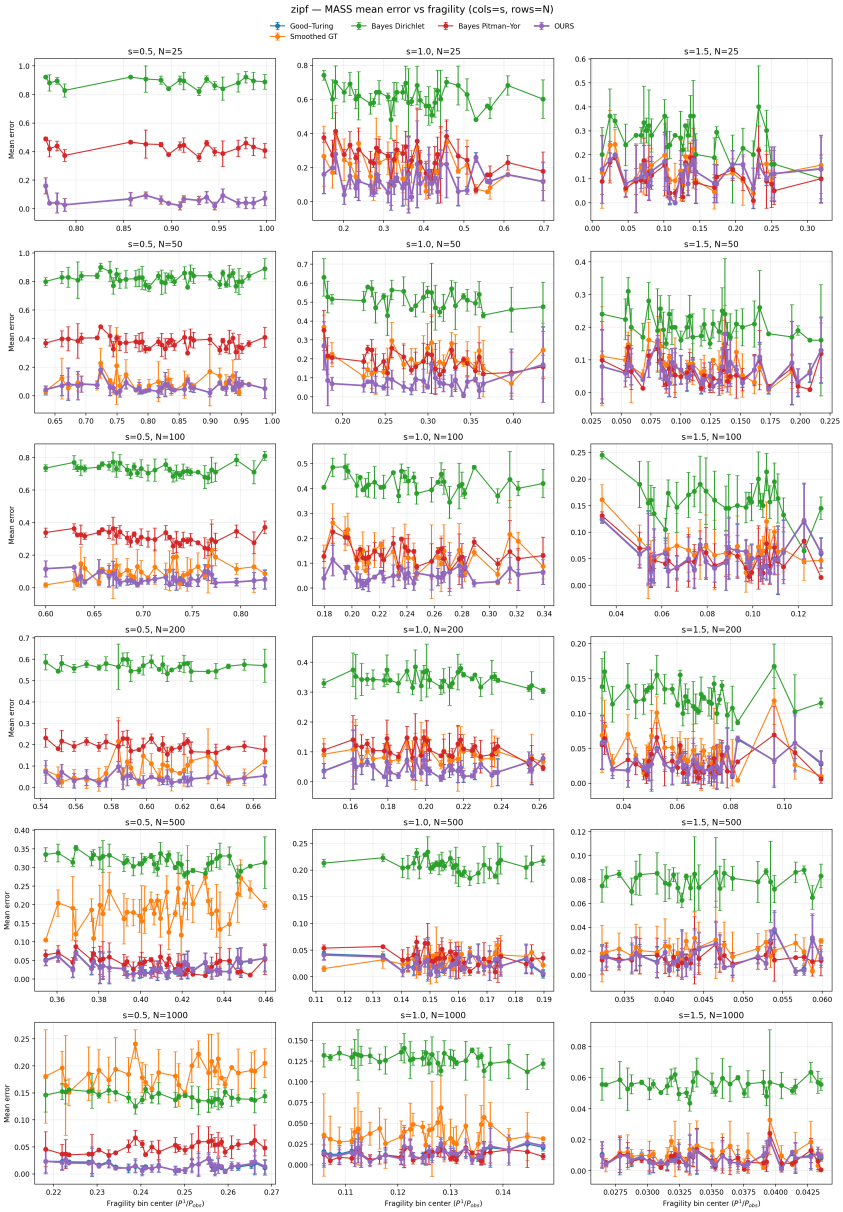


Figure 40: **summarizing 18 experimental** scenarios under Zipf distributions, mean missing-mass estimation error is shown across fragility regimes for varying tail exponents s and sample sizes N. Each subplot corresponds to a specific (s, N) pair, with error bars indicating variability across repeated trials. Smaller exponents ($s = 0.5$) produce flatter distributions with many moderately frequent species, whereas larger exponents ($s = 1.5$) yield strong head–tail imbalance with extremely rare types. Across these regimes, the fragility-aware estimator adapts to the underlying tail structure, maintaining competitive accuracy in smoother settings and improving stability when rare-event uncertainty dominates.

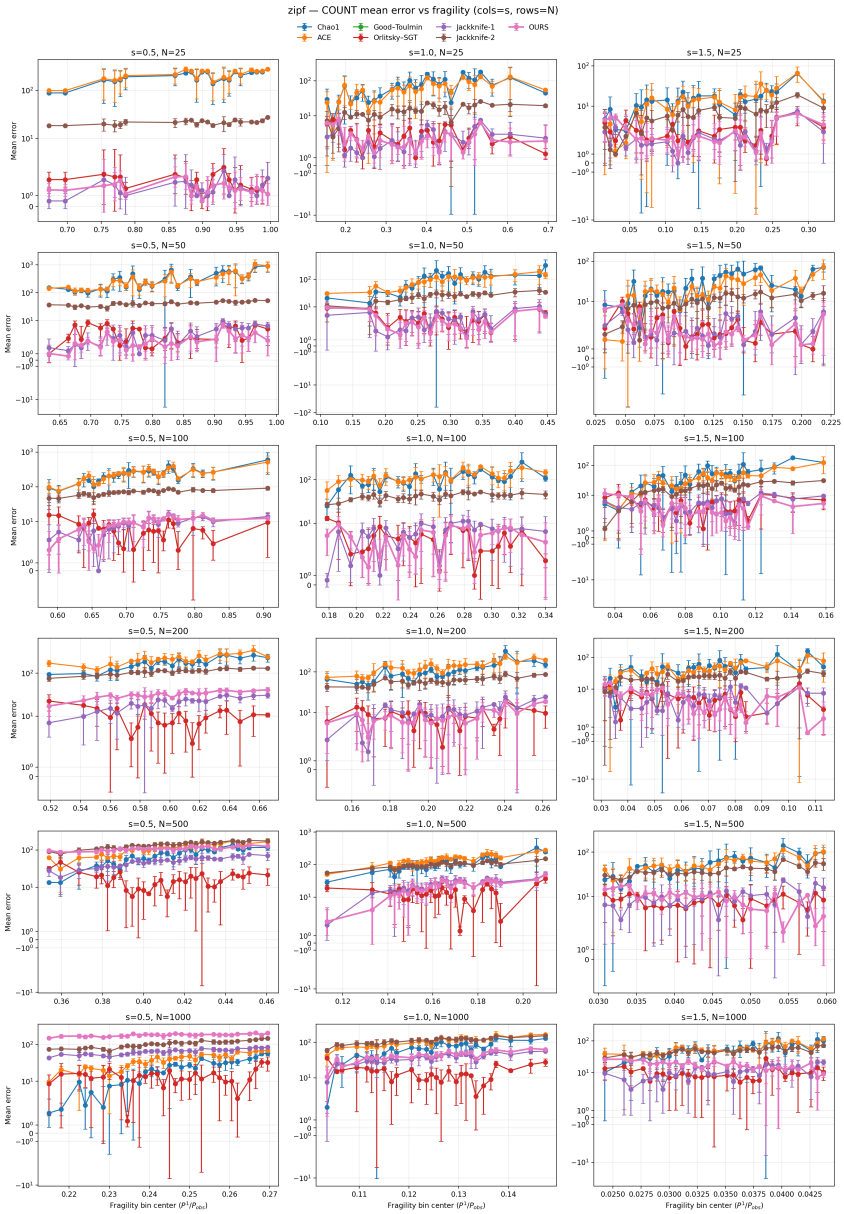


Figure 41: **summarizing 18 experimental** scenarios under Zipfian distributions, mean unseen-species estimation error is shown across fragility regimes for varying tail exponents s and sample sizes N. Each subplot corresponds to a specific (s, N) pair, with error bars reflecting variability across trials (on a symmetric logarithmic (symlog) scale). As the exponent increases and probability mass concentrates in a few frequent types, classical estimators exhibit larger variance due to the expansion of the rare-event tail. The fragility-aware estimator responds to this shift by stabilizing the predicted growth of unseen-species, often reducing variability in sparse sampling regimes while remaining competitive as sample size increases.

F.3.4 HEAD-TAIL DISTRIBUTION

To study estimator behavior under controlled head–tail mixtures, experiments were conducted on synthetic two-component distributions in which a fraction p of the total probability mass is assigned to a "head" subset of species, while the remaining mass $1 - p$ is spread across a "tail" subset, each containing $N/2$ species. The parameter $p \in (0, 1)$ therefore directly controls the strength of head dominance: smaller values yield relatively diffuse distributions, whereas larger values concentrate most probability on a few dominant species while leaving many low-probability tail events. Tables 9 and 10 report UCR across sample sizes and head-mass fractions for both missing-mass estimation and unseen-species prediction, enabling comparison of estimator stability as distributions transition from moderately imbalanced to strongly head-dominated regimes. Figures 42 and 43 complement these summaries by showing mean estimation error and its variability across fragility regimes for each (p, N) scenario.

Parameter p	N	**Ours**	**GT**	**SGT**	**BD**	**PY**
0.7	25	0.145	**0.144**	**0.144**	0.520	0.992
0.7	50	0.121	**0.120**	0.146	0.473	0.938
0.7	100	0.119	**0.118**	0.165	0.425	0.862
0.7	200	**0.105**	**0.105**	0.274	0.288	0.683
0.7	500	**0.062**	**0.062**	0.415	0.064	0.360
0.7	1000	0.033	**0.032**	0.311	0.098	0.149
0.85	25	0.154	**0.153**	**0.153**	0.512	0.977
0.85	50	**0.142**	**0.142**	0.202	0.466	0.921
0.85	100	**0.117**	0.119	0.228	0.378	0.804
0.85	200	**0.099**	**0.099**	0.295	0.239	0.611
0.85	500	**0.066**	**0.066**	0.361	0.059	0.302
0.85	1000	**0.036**	**0.036**	0.210	0.095	0.132
0.95	25	**0.164**	**0.164**	0.164	0.521	0.984
0.95	50	**0.152**	0.153	0.173	0.449	0.897
0.95	100	**0.119**	**0.119**	0.217	0.340	0.754
0.95	200	**0.093**	**0.093**	0.325	0.188	0.535
0.95	500	0.062	**0.061**	0.333	0.112	0.225
0.95	1000	0.022	**0.021**	0.112	0.102	0.075

Table 9: **Head-Tail distributed population:** UCR (lower is better) for missing-mass estimation across p and sample sizes N. Increasing head dominance concentrates probability mass on a small subset of species while distributing the remainder across many low-probability events, creating strong imbalance between frequent and rare categories. The fragility-aware estimator remains competitive with classical methods and demonstrates improved stability as the distribution becomes increasingly skewed toward dominant species.

Parameter p	N	Ours	G-Toulmin	Orlitsky-SGT	Jackknife-1	Jackknife-2	Chao1	ACE
0.7	25	4.604	4.588	4.588	**3.971**	27.137	261.977	266.892
0.7	50	**8.298**	8.314	8.314	8.341	51.437	690.024	663.327
0.7	100	**13.844**	13.858	13.858	14.949	86.418	382.896	369.169
0.7	200	**22.644**	22.706	22.706	37.053	138.328	242.087	236.338
0.7	500	**34.394**	34.480	34.480	71.294	156.876	81.398	76.494
0.7	1000	35.912	**35.691**	**35.691**	77.097	108.740	28.993	26.892
0.85	25	4.983	4.972	4.972	**4.421**	26.601	237.005	245.768
0.85	50	**8.146**	8.167	8.167	8.180	48.750	548.348	495.813
0.85	100	**13.080**	13.146	13.146	19.690	85.157	296.299	288.181
0.85	200	**20.433**	20.486	20.486	36.485	120.098	159.315	152.210
0.85	500	30.961	**30.825**	**30.825**	66.300	128.929	58.073	51.022
0.85	1000	**34.432**	34.392	34.392	58.233	99.140	37.576	24.561
0.95	25	5.101	5.102	5.102	**4.610**	25.874	211.294	221.125
0.95	50	**9.445**	9.458	9.458	9.559	48.646	442.268	370.640
0.95	100	**13.409**	13.446	13.446	18.842	77.086	213.865	195.890
0.95	200	**19.658**	19.697	19.697	35.642	98.348	90.427	85.213
0.95	500	26.138	**26.040**	**26.040**	57.837	88.021	31.430	26.408
0.95	1000	33.419	33.261	33.261	**29.513**	49.277	**20.247**	23.287

Table 10: **Head-Tail distributed population:** UCR (lower is better) for unseen-species counts across p and sample sizes N. Increasing head dominance concentrates probability mass on a small subset of species while distributing the remainder across many low-probability events, creating strong imbalance between frequent and rare categories. The fragility-aware estimator remains competitive with classical methods and demonstrates improved stability as the distribution becomes increasingly skewed toward dominant species.

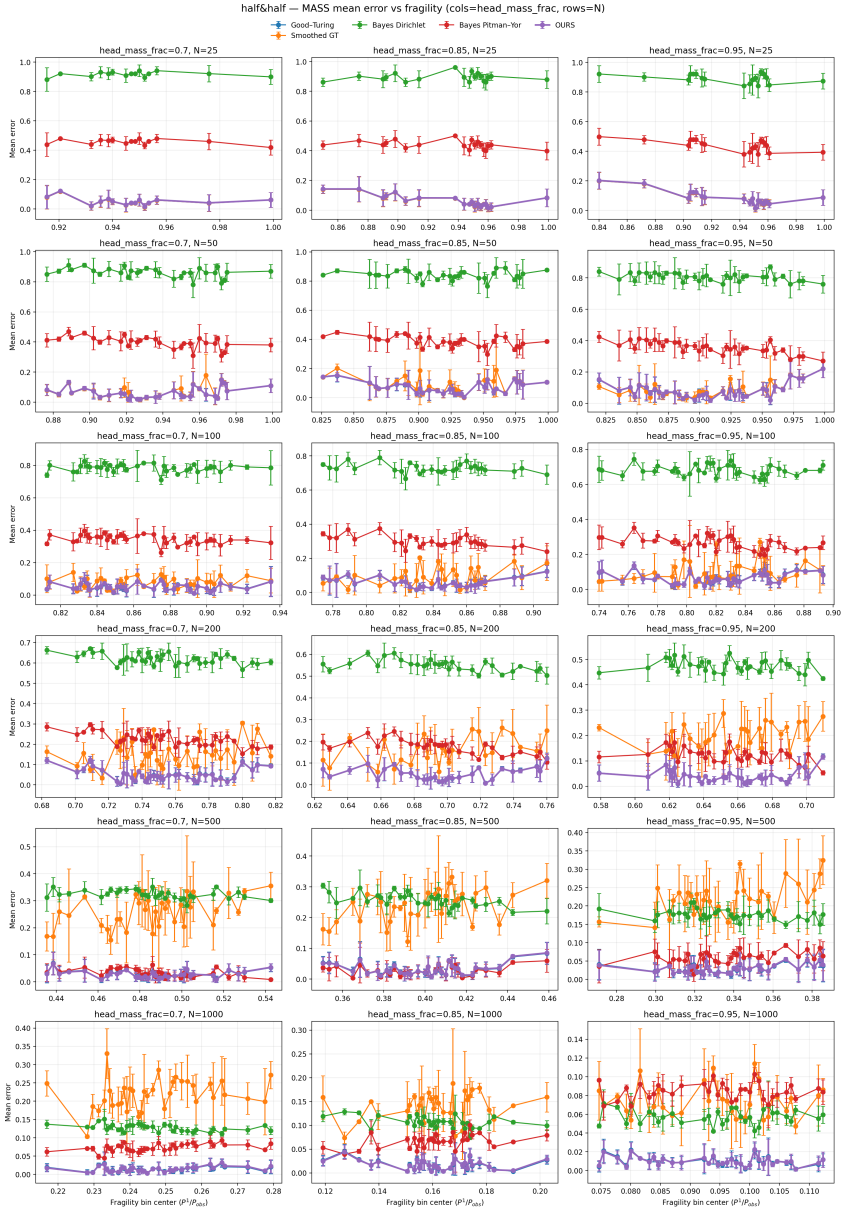


Figure 42: **summarizing 18 experimental** scenarios under head–tail mixture distributions, mean missing-mass estimation error is shown across fragility regimes for varying p fractions and sample sizes N. Each subplot corresponds to a specific p, N) pair, with error bars indicating variability across repeated trials. Increasing head dominance produces stronger imbalance between frequent and rare species, amplifying estimation difficulty in low-sample regimes. Across these regimes, the proposed fragility-aware estimator maintains stable behavior and consistently reduces error relative to classical methods, particularly when the distribution becomes highly skewed.

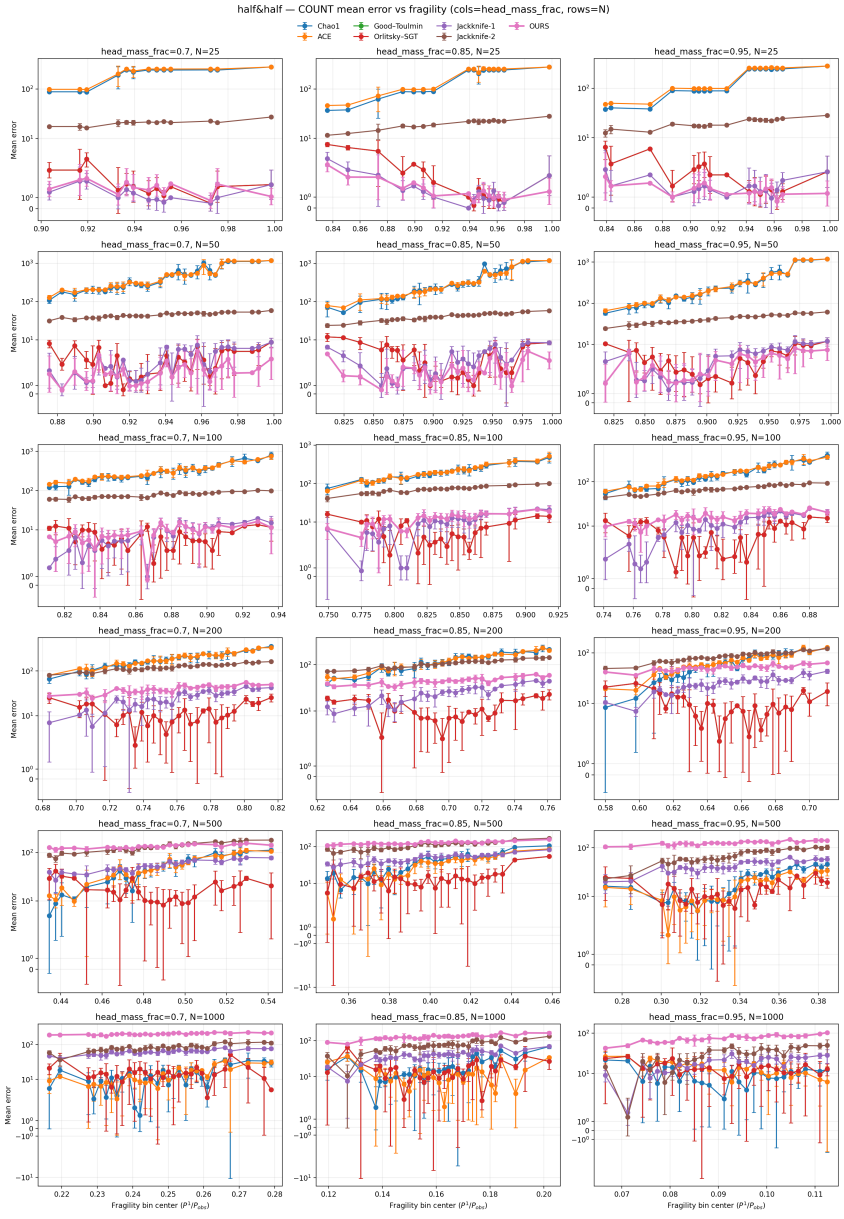


Figure 43: **summarizing 18 experimental** scenarios under head–tail mixture distributions, mean unseen-species estimation error is shown across fragility regimes for varying head-mass fractions and sample sizes N. Each subplot corresponds to a specific $(\text{head mass}, N)$ pair, with error bars reflecting variability across trials (on a symmetric logarithmic (symlog) scale). As head dominance increases, classical estimators exhibit larger variance and bias due to reduced tail support, whereas the fragility-aware estimator adapts to the changing singleton structure, yielding improved stability in sparse regimes while remaining competitive at larger sample sizes.

F.4 Real World Dataset Additional Results

To study how estimator behavior changes under different levels of extrapolation, experiments are conducted across a range of forecast horizons r_{future} on both the Shakespeare–Gutenberg and Google corpora. The parameter r_{future} specifies the relative size of the future sample, i.e the number of possible additional observations is $N_{\text{new}} = r_{\text{future}}\, N$, which determines the amount of additional data used to define the evaluation target and fragility-adjusted ground truth, rather than being directly incorporated into the estimators themselves. In this study, forecast horizons between 3% and 10% of the observed sample are considered, representing moderate extrapolation regimes where meaningful discovery may still occur while the observed sample remains informative. Varying this quantity directly changes both the probability that the next observation belongs to an unseen species M_0 in Eq 10 and the plausible upper bound on the number of undiscovered species U_{max} in Eq 8. Evaluating estimators across this spectrum allows the analysis of how fragility-aware adjustments influence both missing-mass and unseen-species estimates as the amount of future data.

F.4.1 Shakespeare Corpus: Missing Mass Estimation

Fig 44 illustrates how estimator comparisons vary with the forecast horizon r_{future} in Shakespeare Corpus missing mass estimation. As the horizon increases, the implied discovery process becomes more sensitive to rare events, making evaluation increasingly dependent on the stability of the observed statistics. Across these regimes, the fragility-aware estimator maintains consistently strong performance against variance-sensitive classical methods, suggesting that explicitly accounting for sampling instability leads to predictions that remain reliable as the extrapolation task becomes more demanding. In contrast, estimators that rely heavily on raw singleton counts exhibit more variable behavior across horizons, reflecting their sensitivity to small perturbations in the observed sample.

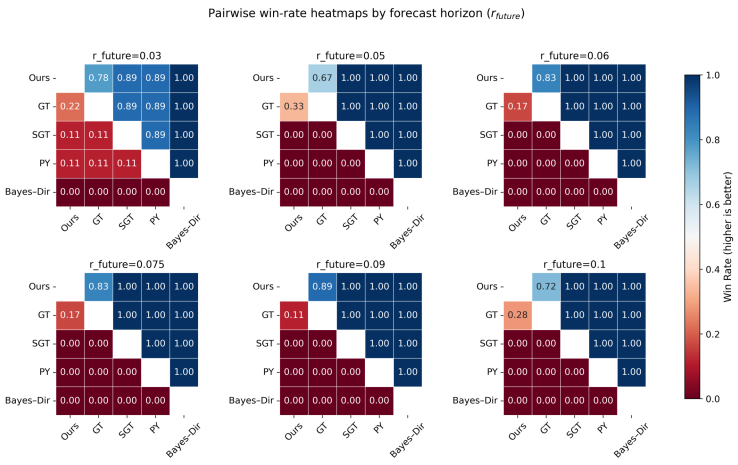
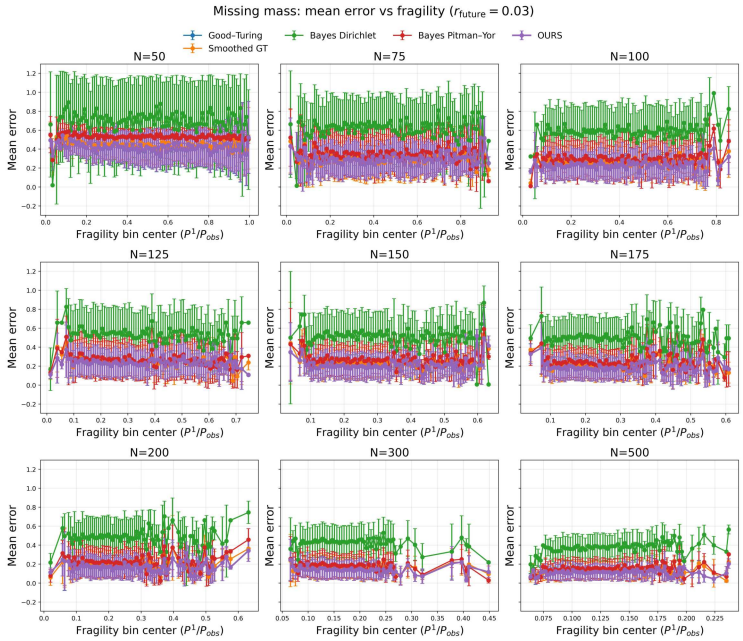


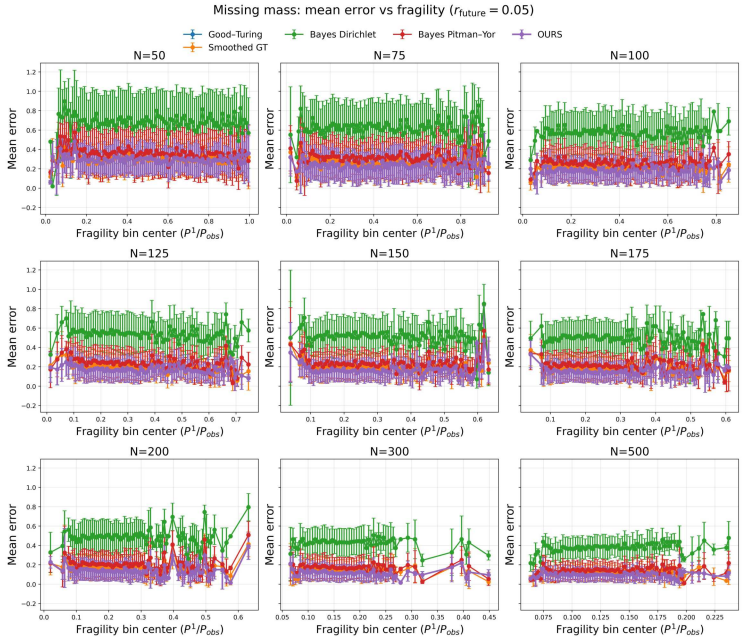
Figure 44: Pairwise win-rate heatmaps for missing-mass estimation on the Shakespeare–Gutenberg corpus across forecast horizons $r_{\text{future}} \in [0.03, 0.10]$. Each entry shows the fraction of experiments where the row estimator achieves lower UCR than the column estimator. The fragility-aware estimator remains competitive with classical smoothing methods across horizons and generally outperforms SOTA estimators.

Fig 45, Fig 46, and Fig 47 illustrate the performance of the proposed missing mass estimator against SOTA estimators across forecast horizons $r_{\text{future}} \in [0.03, 0.10]$. Each grid corresponds to a specific N value, with error bars reflecting variability across repeated trials. As N increases, all methods exhibit reduced mean error and narrower variability bands, reflecting improved reliability of empirical frequency statistics. However, classical estimators remain sensitive to fragility, with noticeable bias in regions dominated by singleton mass. In contrast, the fragility-aware estimator maintains comparatively lower error across bins, showing weaker dependence on fragility even in sparse regimes. This pattern holds across all forecast horizons, indicating that the proposed adjustment primarily improves robustness to sampling instability rather than exploiting any specific choice of

r_{future}. Overall, the results suggest that incorporating fragility information yields more consistent missing-mass estimates across both sample sizes and extrapolation regimes.

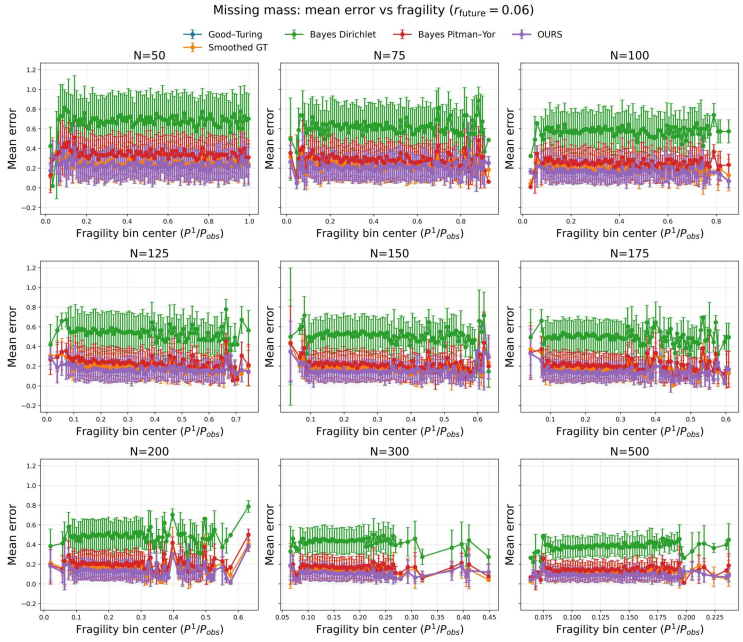


(a) Missing-mass error vs. fragility ($r_{\text{future}} = 0.03$)

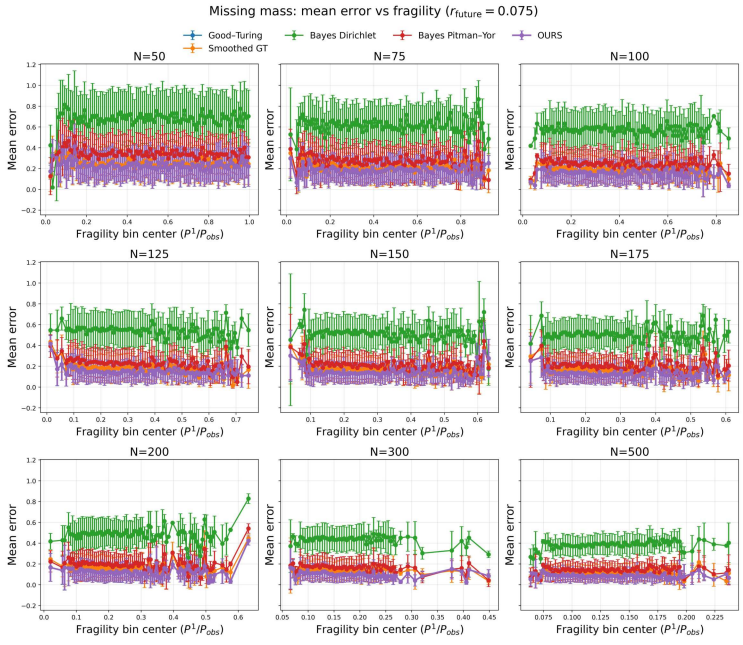


(b) Missing-mass error vs. fragility ($r_{\text{future}} = 0.04$)

Figure 45: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error is shown across fragility regimes for varying sample sizes N. The fragility-aware estimator consistently achieves lower error compared to SOTA methods.

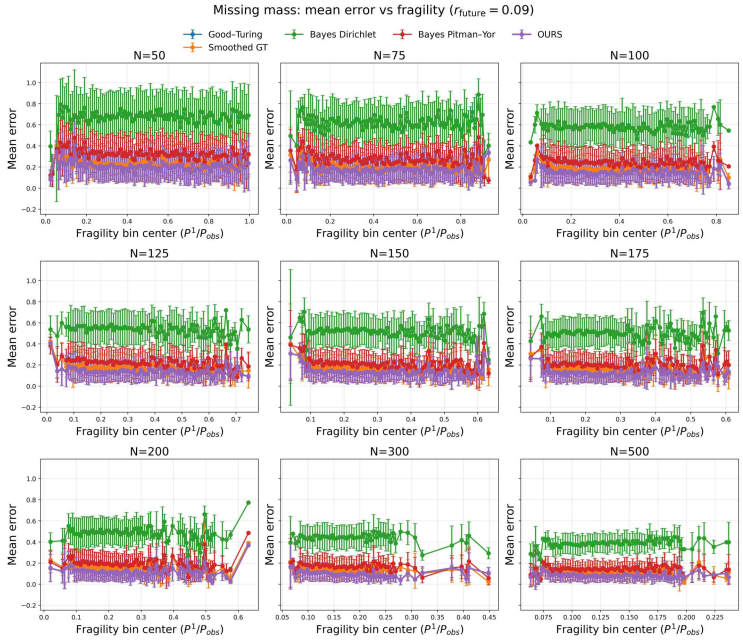


(a) Missing-mass error vs. fragility ($r_{\text{future}} = 0.06$)

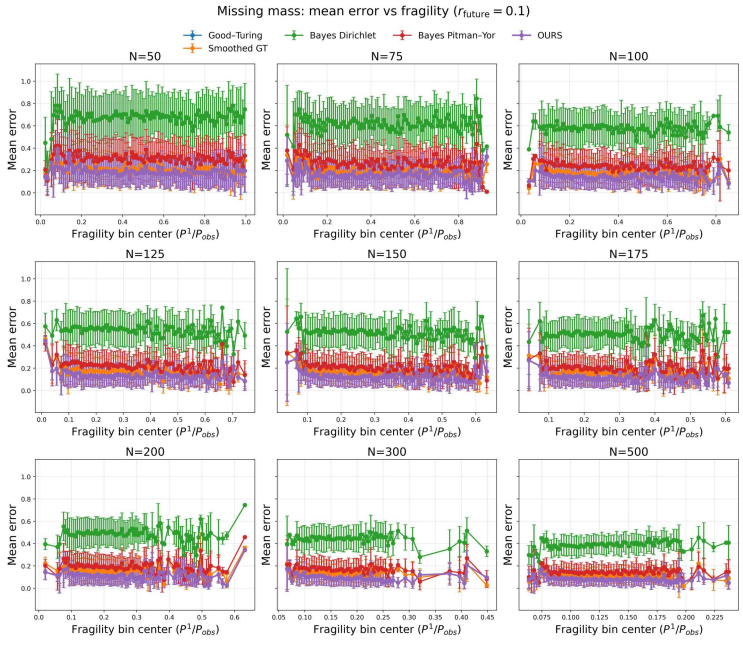


(b) Missing-mass error vs. fragility ($r_{\text{future}} = 0.075$)

Figure 46: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error is shown across fragility regimes for varying sample sizes N. The fragility-aware estimator consistently achieves lower error compared to SOTA methods.



(a) Missing-mass error vs. fragility $r_{\text{future}} = 0.09$)



(b) Missing-mass error vs. fragility ($r_{\text{future}} = 0.1$)

Figure 47: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error is shown across fragility regimes for varying sample sizes N. The fragility-aware estimator consistently achieves lower error compared to SOTA methods.

F.4.2 SHAKESPEARE CORPUS: MISSING COUNT ESTIMATION

Fig 48 illustrates how estimator comparisons vary with the forecast horizon r_{future} in Shakespeare Corpus unseen-species count estimation. As the horizon increases, the implied discovery process becomes more sensitive to rare events, making evaluation increasingly dependent on the stability of the observed statistics. Across these regimes, the fragility-aware estimator maintains consistently strong performance against variance-sensitive classical methods, suggesting that explicitly accounting for sampling instability leads to predictions that remain reliable as the extrapolation task becomes more demanding, and indicating robust extrapolation under long-tailed language distributions.

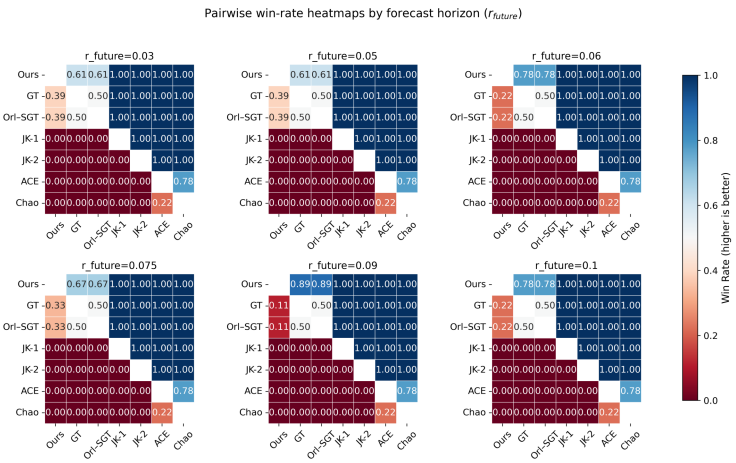
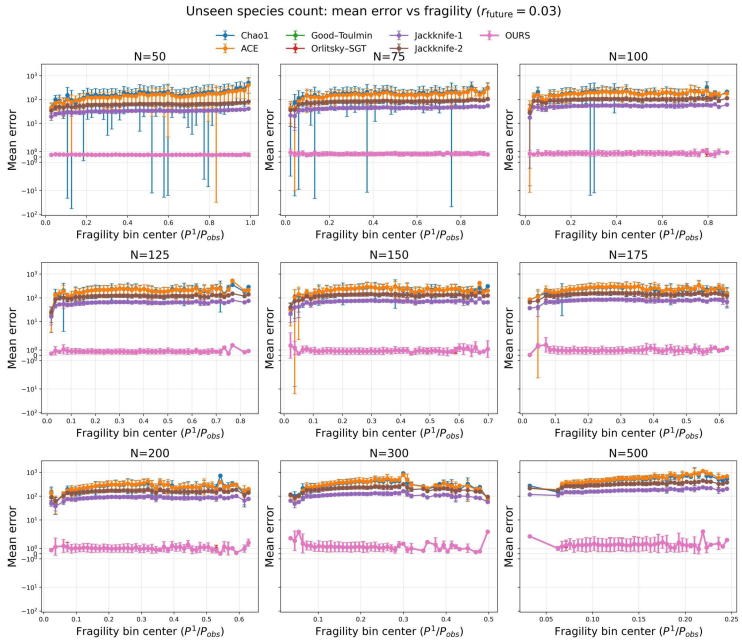


Figure 48: Pairwise win-rate heatmaps for unseen-species estimation on the Shakespeare–Gutenberg corpus across forecast horizons $r_{\text{future}} \in [0.03, 0.10]$. Each entry shows the fraction of experiments where the row estimator achieves lower UCR than the column estimator. The fragility-aware estimator remains competitive with classical smoothing methods across horizons and generally outperforms SOTA estimators.

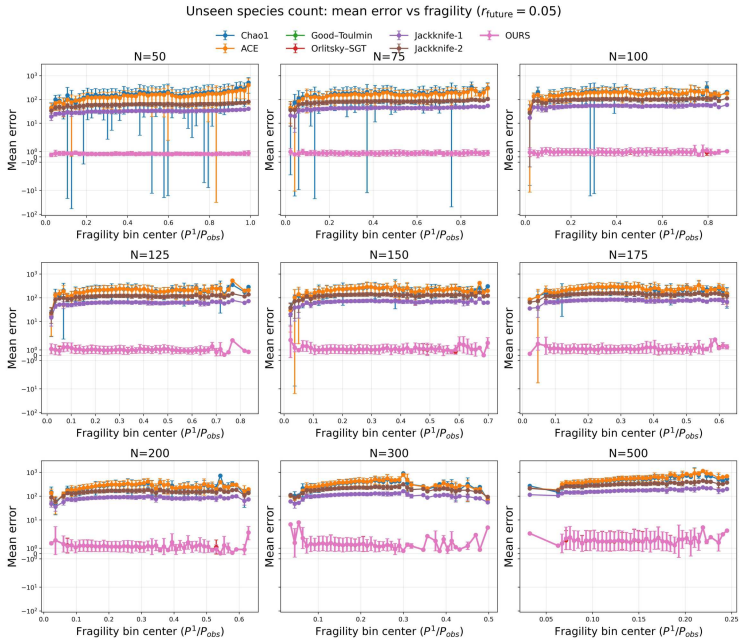
Figures 49–51 present unseen-species count estimation across forecast horizons $r_{\text{future}} \in [0.03, 0.10]$. In contrast to missing-mass results, the unseen-count task is substantially more unstable: classical extrapolation-based estimators exhibit large absolute errors, reflecting the difficulty of predicting discrete discovery counts in heavy-tailed language data.

Across sample sizes, increasing N reduces variance for all methods, but systematic differences persist. Smoothing-based estimators such as Good–Toulmin and Orlitsky–SGT remain comparatively stable, while resampling and coverage-based estimators often produce very large errors. The fragility-aware estimator shows a distinct behavior: although its mean error is sometimes slightly higher in low fragility regions, its variability across fragility bins is substantially reduced, indicating improved robustness to sampling instability.

This pattern holds consistently across forecast horizons, suggesting that fragility adjustment primarily regularizes sensitivity to rare-event fluctuations rather than optimizing raw discovery counts. Overall, the results highlight a key distinction between mass and count prediction: while fragility correction improves stability in both cases, unseen-species counts remain intrinsically harder to extrapolate, particularly in long-tail linguistic distributions.

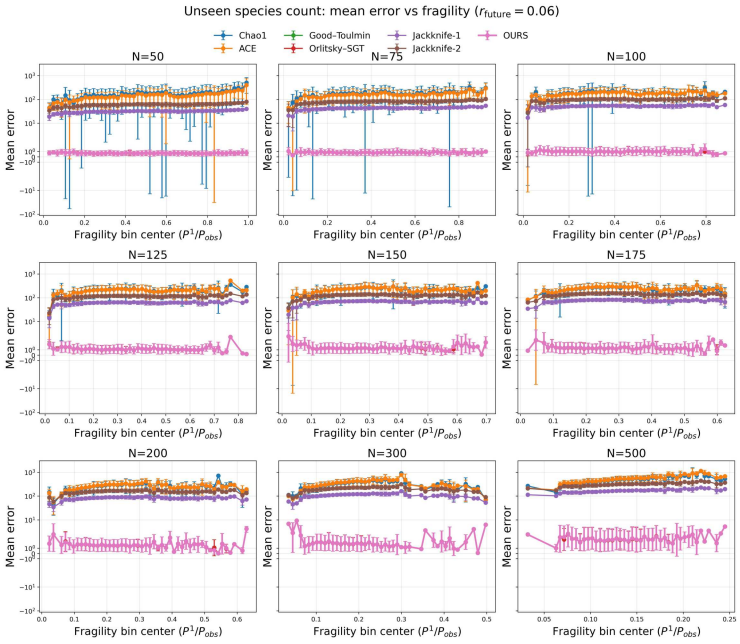


(a) Unseen-species error vs. fragility ($r_{\text{future}} = 0.03$)

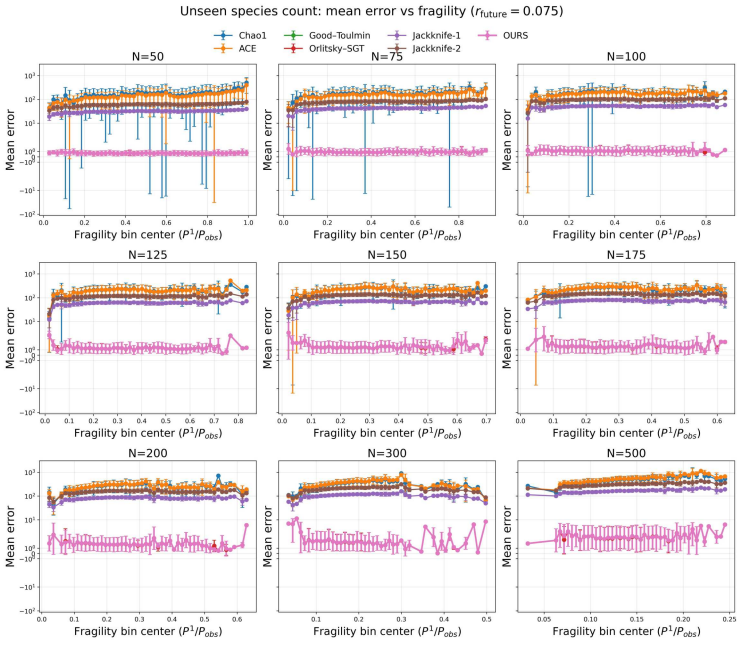


(b) Unseen-species error vs. fragility ($r_{\text{future}} = 0.05$)

Figure 49: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error (symlog scale) is shown across fragility regimes for varying sample sizes N. The fragility-aware estimator adapts to this regime shift, consistently achieving lower error in sparse settings while remaining competitive as sampling density improves.

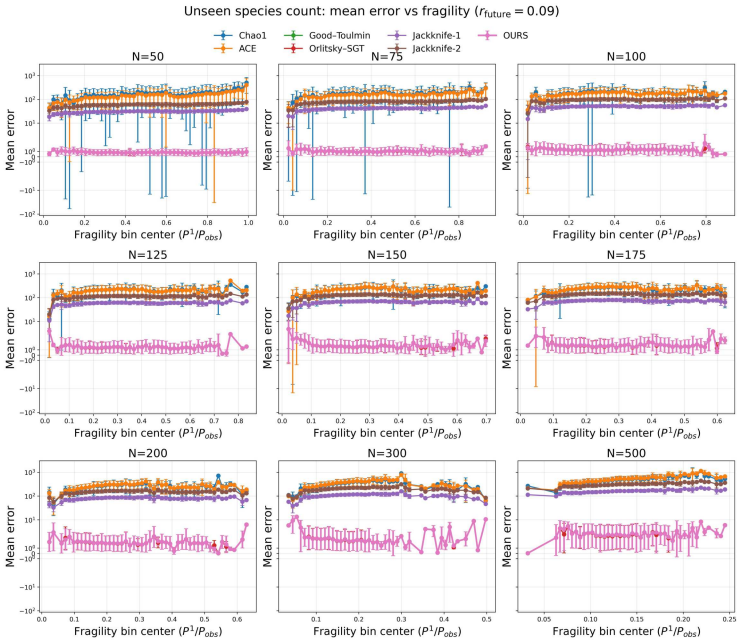


(a) Unseen-species error vs. fragility ($r_{\text{future}} = 0.06$)

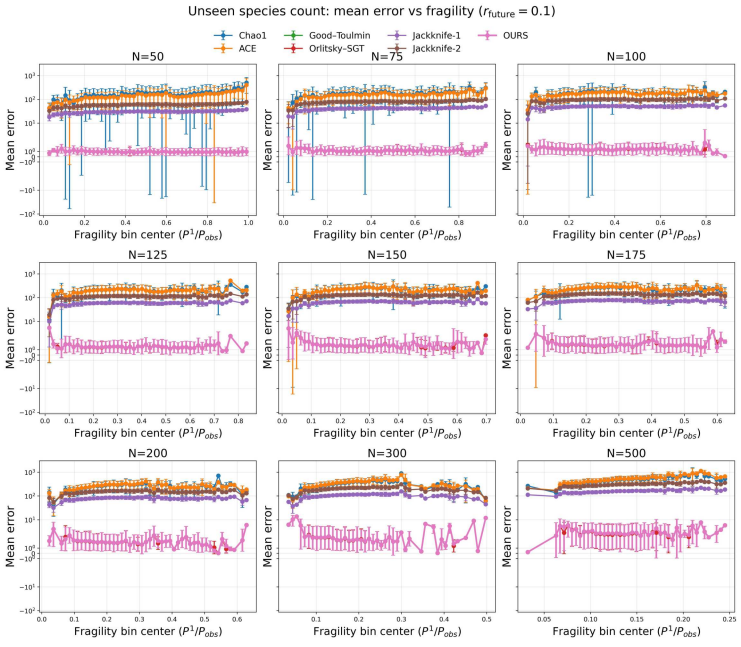


(b) Unseen-species error vs. fragility ($r_{\text{future}} = 0.075$)

Figure 50: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error (symlog scale) is shown across fragility regimes for varying sample sizes N. The fragility-aware estimator adapts to this regime shift, consistently achieving lower error in sparse settings while remaining competitive as sampling density improves.



(a) Unseen-species error vs. fragility ($r_{\text{future}} = 0.09$)



(b) Unseen-species error vs. fragility ($r_{\text{future}} = 0.1$)

Figure 51: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error (symlog scale) is shown across fragility regimes for varying sample sizes N. The fragility-aware estimator adapts to this regime shift, consistently achieving lower error in sparse settings while remaining competitive as sampling density improves.

F.4.3 GOOGLE N-GRAM CORPUS: MISSING MASS ESTIMATION

Fig 52 shows how pairwise estimator performance varies with forecast horizon r_{future} on the Google N-gram corpus for missing-mass estimation. As the horizon increases, the implied extrapolation emphasizes increasingly rare lexical events, making reliable estimation dependent on robustness to sampling noise rather than purely on observed frequencies. Across all horizons, the fragility-aware estimator consistently achieves the highest win-rates against SOTA methods, indicating stable comparative performance even as the extrapolation difficulty grows. In contrast, frequency-driven estimators exhibit greater sensitivity to horizon changes, reflecting their reliance on unstable singleton statistics in heavy-tailed language data. Overall, the results suggest that incorporating fragility information improves comparative robustness of missing-mass predictions in large-scale natural language distributions.

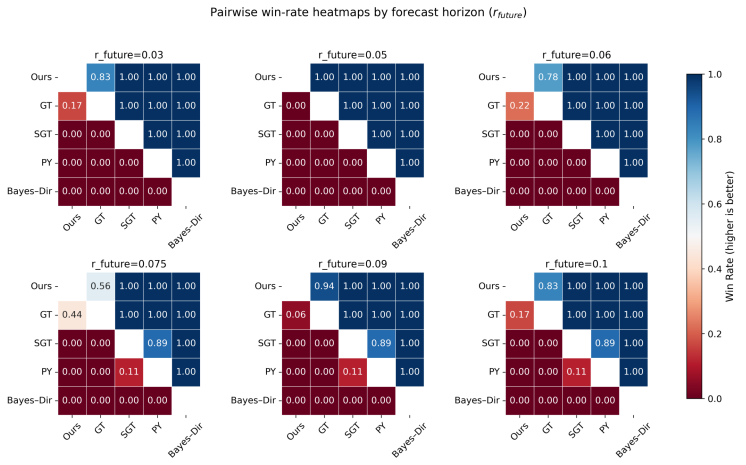
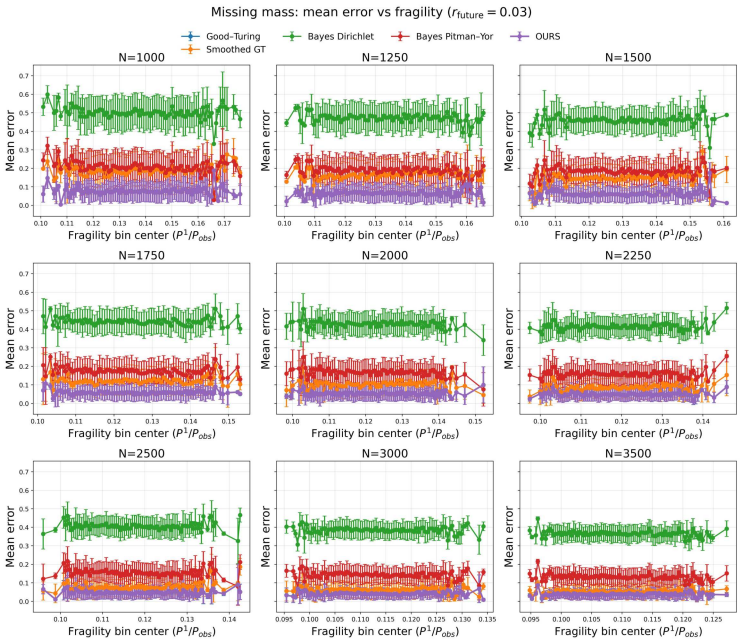


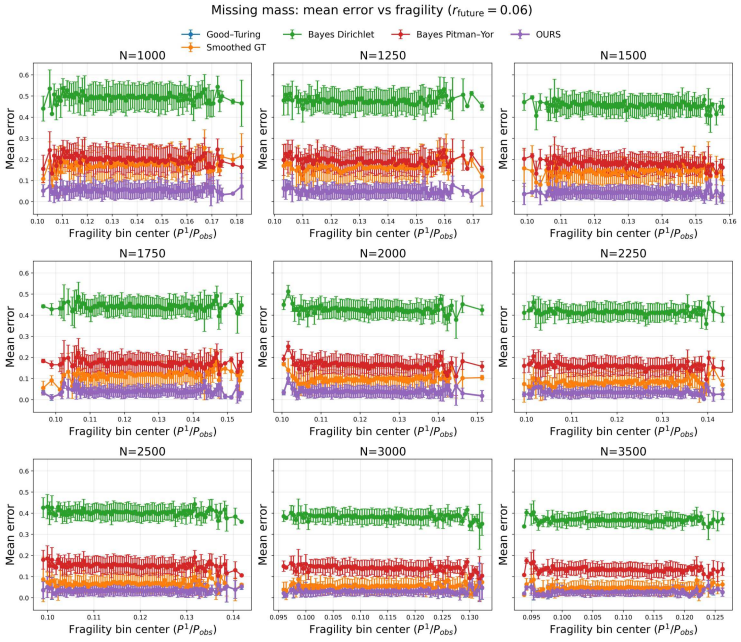
Figure 52: Pairwise win-rate heatmaps for missing-mass estimation on the Google n-gram corpus across forecast horizons $r_{\text{future}} \in [0.03, 0.10]$. Each entry shows the fraction of experiments where the row estimator achieves lower UCR than the column estimator. The fragility-aware estimator remains competitive with classical smoothing methods across horizons and generally outperforms SOTA estimators.

Fig 53, Fig 54 and Fig 55 present paired views of missing-mass error versus fragility on the Google-Ngram corpus across forecast horizons. Each subplot corresponds to a specific N value, with error bars reflecting variability across repeated trials. Each pair contrasts adjacent horizons while holding the evaluation protocol fixed, allowing the effect of extrapolation scale to be visually isolated. Across all pairs, similar to Shakespeare Corpus in Sec F.4.1, a consistent pattern emerges: Bayesian estimators exhibit comparatively large and nearly flat error curves, indicating limited responsiveness to fragility structure, while classical frequency-based methods achieve lower bias but still display moderate sensitivity in higher-fragility bins. In contrast, the proposed fragility-aware estimator maintains the lowest and most stable error profile throughout the fragility spectrum, with visibly tighter trajectories and weaker dependence on bin position.

As sample size increases within each paired grid, variance bands narrow and all estimators improve, reflecting the stabilizing effect of larger corpora. However, the relative ordering of methods remains unchanged across the paired figures, suggesting that the gains of the proposed approach arise from structural robustness rather than sample-specific effects. Notably, the similarity of trends across horizon pairs indicates that the estimator's advantage is not tied to a particular extrapolation level; instead, incorporating fragility information appears to systematically improve reliability of missing-mass estimation in heavy-tailed language data.

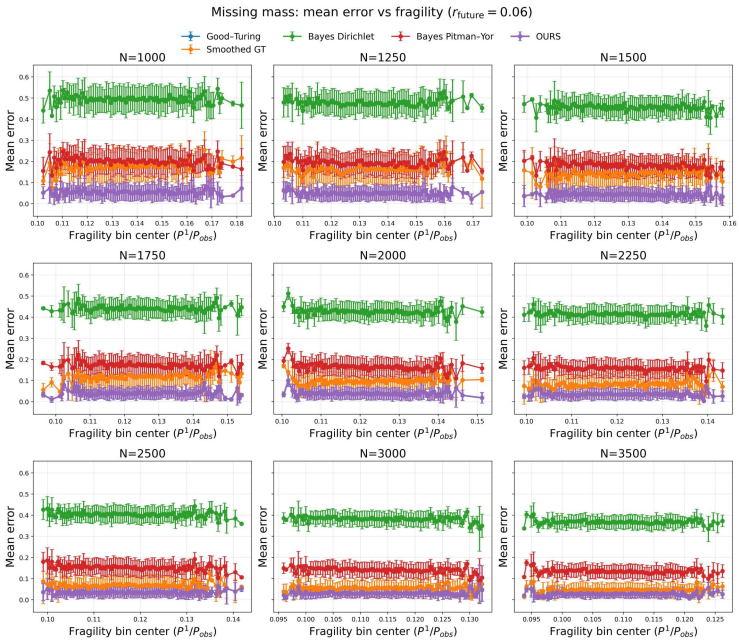


(a) Missing-mass error vs. fragility ($r_{\text{future}} = 0.03$)

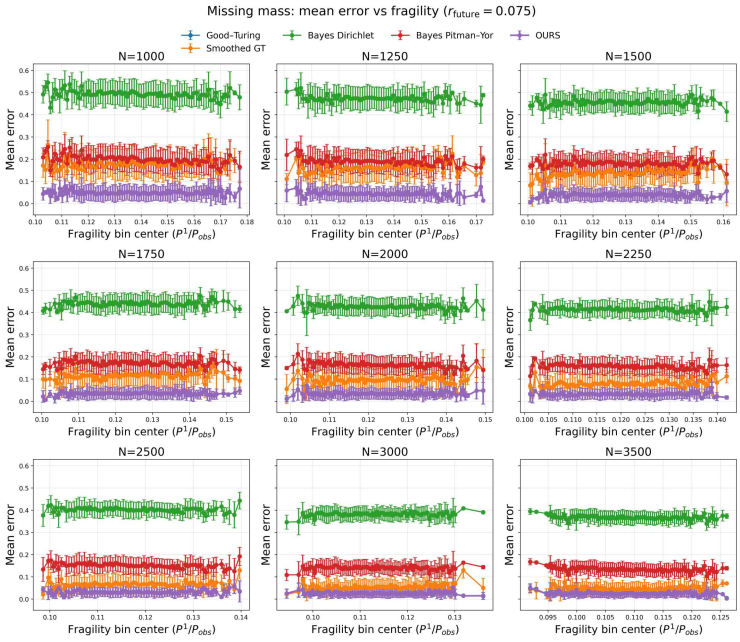


(b) Missing-mass error vs. fragility ($r_{\text{future}} = 0.05$)

Figure 53: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error is shown across fragility regimes for varying sample sizes N. With larger N, fragility falls and classical estimators stabilize, yet bias remains in singleton-heavy regions. The fragility-aware method maintains low error in sparse regimes and stays competitive as data densifies.

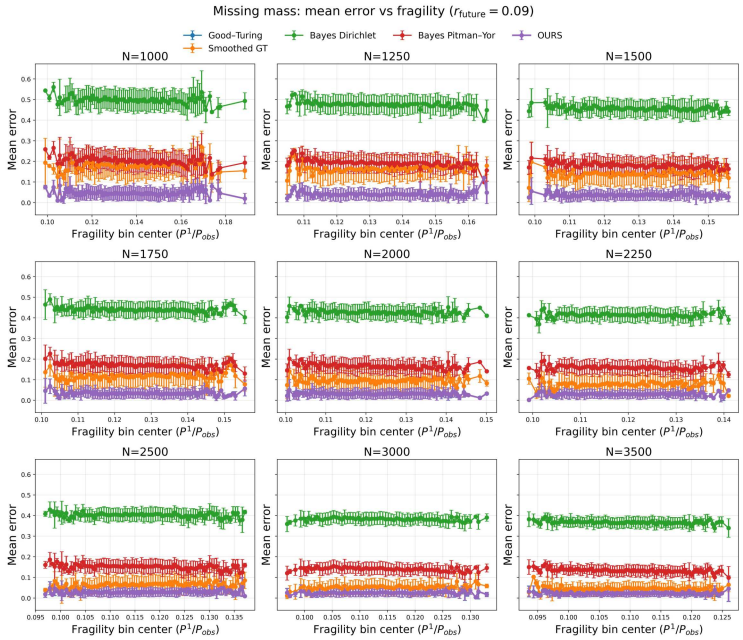


(a) Missing-mass error vs. fragility ($r_{\text{future}} = 0.06$)

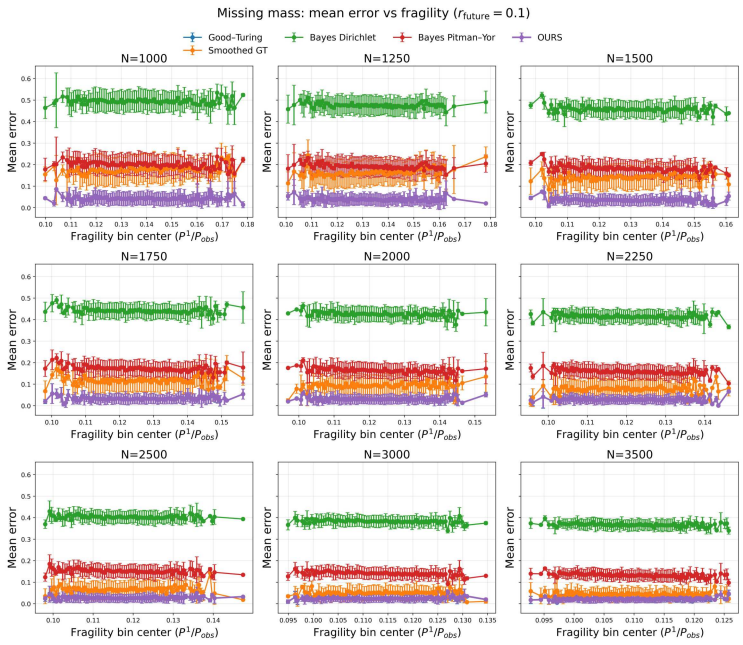


(b) Missing-mass error vs. fragility ($r_{\text{future}} = 0.075$)

Figure 54: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error is shown across fragility regimes for varying sample sizes N. With larger N, fragility falls and classical estimators stabilize, yet bias remains in singleton-heavy regions. The fragility-aware method maintains low error in sparse regimes and stays competitive as data densifies.



(a) Missing-mass error vs. fragility ($r_{\text{future}} = 0.09$)



(b) Missing-mass error vs. fragility ($r_{\text{future}} = 0.1$)

Figure 55: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean missing-mass estimation error is shown across fragility regimes for varying sample sizes N. With larger N, fragility falls and classical estimators stabilize, yet bias remains in singleton-heavy regions. The fragility-aware method maintains low error in sparse regimes and stays competitive as data densifies.

F.4.4 GOOGLE N-GRAM CORPUS: MISSING COUNT ESTIMATION

Fig 56 shows pairwise estimator performance across forecast horizons r_{future} for unseen-species count prediction on the Google N-gram corpus. As the horizon increases, the extrapolation task places greater emphasis on the discovery of previously unseen lexical types, making accurate prediction sensitive to sampling variability in the tail. Across all horizons, the fragility-aware estimator consistently attains the highest win-rates against competing methods, indicating stable comparative performance as the prediction task becomes more demanding. In contrast, classical count-based estimators display more variability across horizons, reflecting their dependence on unstable rare-event statistics in heavy-tailed language distributions. These results suggest that explicitly incorporating fragility improves robustness of unseen-species count prediction in large-scale natural language corpora.

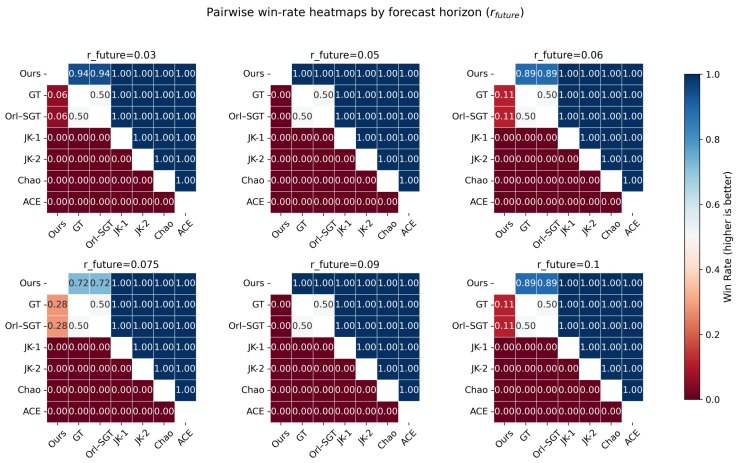
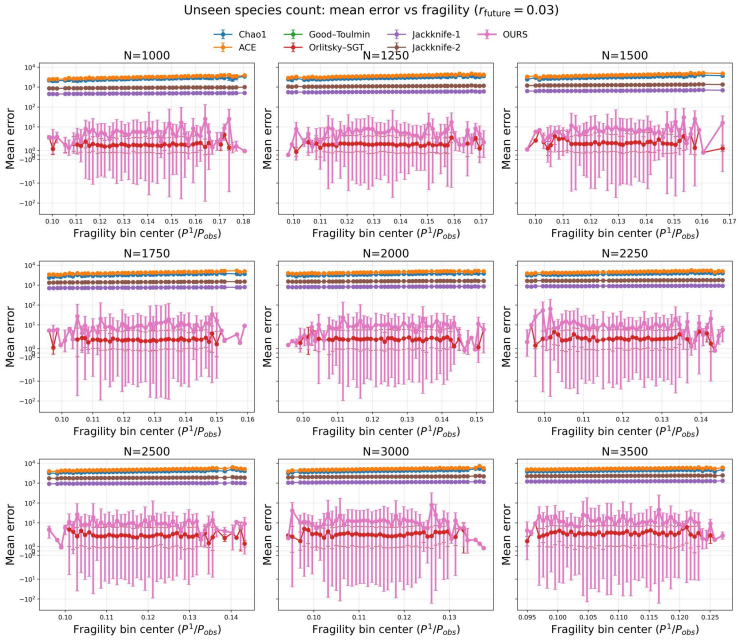
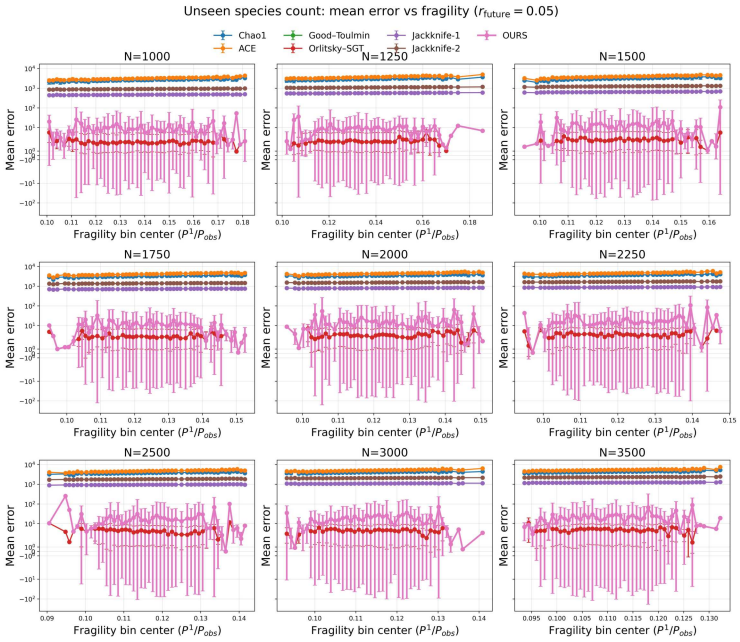


Figure 56: Pairwise win-rate heatmaps for unseen-count estimation on the Google n-gram corpus across forecast horizons $r_{\text{future}} \in [0.03, 0.10]$. Each entry shows the fraction of experiments where the row estimator achieves lower UCR than the column estimator. The fragility-aware estimator remains competitive with classical smoothing methods across horizons and generally outperforms SOTA estimators.

Across the Google N-gram experiments, unseen-count estimation exhibits a markedly different stability profile compared to missing-mass prediction discussion in Sec F.4.3. In low-sample regimes, classical estimators produce extremely large errors with minimal sensitivity to fragility, reflecting their reliance on asymptotic approximations that break down in heavy-tailed vocabularies. As sampling increases, these methods gradually compress in scale but retain substantial bias, indicating that improved frequency estimates alone do not fully resolve tail uncertainty. In contrast, the fragility-aware estimator displays a distinct behavior: although its variance is higher at very small N, its error grows more smoothly across fragility bins and stabilizes earlier as data accumulates. This suggests that explicitly modeling sampling instability provides a structural advantage in count extrapolation, where the dominant challenge is not merely estimating singleton mass but controlling the propagation of uncertainty into future vocabulary growth. Overall, the Google N-gram results reinforce that fragility awareness primarily improves robustness of unseen-event prediction rather than simply reducing variance, yielding estimates that remain better calibrated across both sparse and moderately dense sampling regimes.

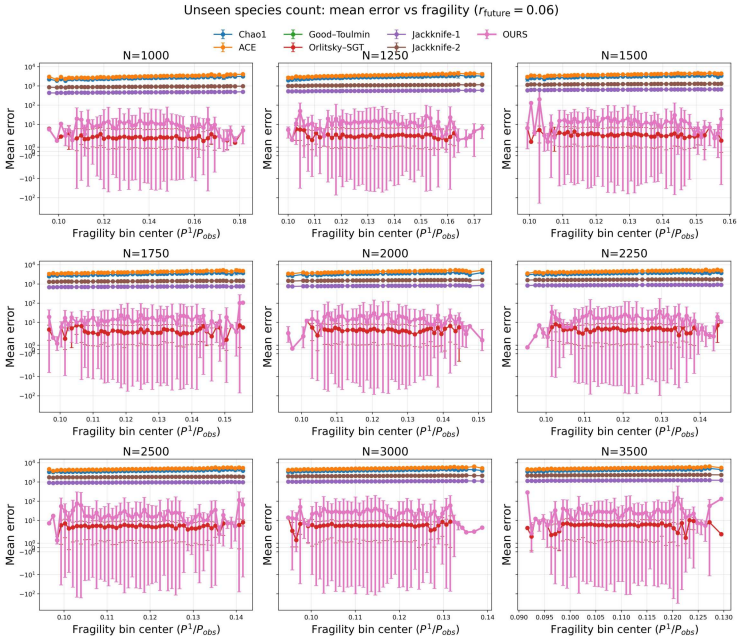


(a) Unseen-species error vs. fragility ($r_{\text{future}} = 0.03$)

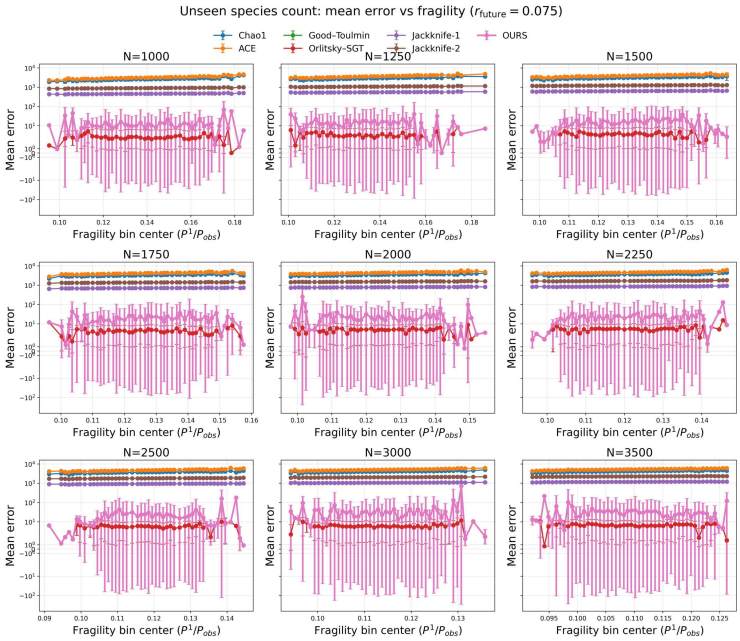


(b) Unseen-species error vs. fragility ($r_{\text{future}} = 0.05$)

Figure 57: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean unseen species estimation error (symlog scale) is shown across fragility regimes for varying sample sizes N. As N grows, fragility remains consistent, even though classical estimators stabilize they remain biased in singleton regimes; the fragility-aware estimator stays more accurate in sparse settings while remaining competitive overall.

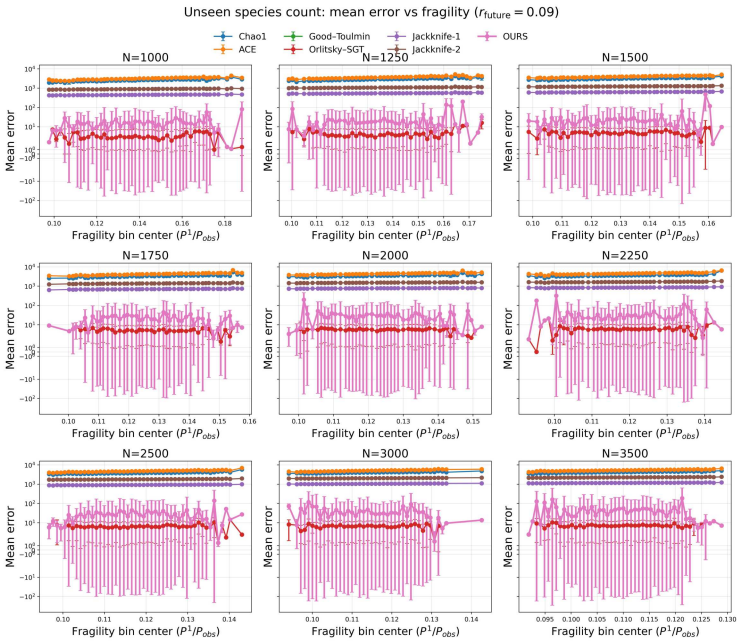


(a) Unseen-species error vs. fragility ($r_{\text{future}} = 0.06$)

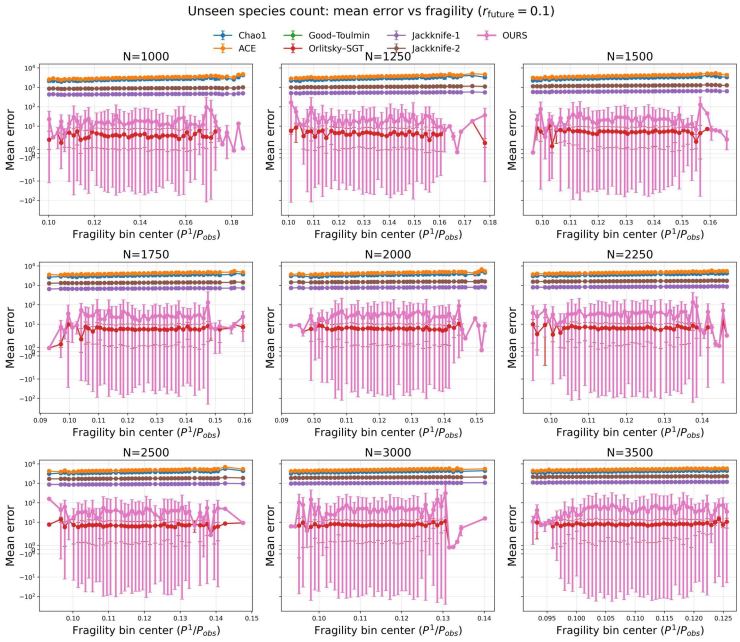


(b) Unseen-species error vs. fragility ($r_{\text{future}} = 0.075$)

Figure 58: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean unseen species estimation error (symlog scale) is shown across fragility regimes for varying sample sizes N. As N grows, fragility remains consistent, even though classical estimators stabilize they remain biased in singleton regimes; the fragility-aware estimator stays more accurate in sparse settings while remaining competitive overall.



(a) Unseen-species error vs. fragility ($r_{\text{future}} = 0.09$)



(b) Unseen-species error vs. fragility ($r_{\text{future}} = 0.1$)

Figure 59: **summarizing 18 experimental** scenarios (two forecast horizons shown) under increasing sample sizes, mean unseen species estimation error (symlog scale) is shown across fragility regimes for varying sample sizes N. As N grows, fragility remains consistent, even though classical estimators stabilize they remain biased in singleton regimes; the fragility-aware estimator stays more accurate in sparse settings while remaining competitive overall.

G NOTATION

Table 11 summarizes the principal notation used throughout the paper.

Table 11: Principal notation and abbreviations used in the proposed framework.

Notation	Definition
Generated sequences and semantic clusters	
$\underline{y}$	Input prompt or question supplied to the language model.
$\underline{s} = (s_1, \ldots, s_T)$	Generated TS, where s_i is its ith token and T is the sequence length.
$R,\ r$	Number of stochastic generations and the corresponding generation index, respectively.
$p_s^{(r)}$	Autoregressive probability of the rth generated TS conditioned on $\underline{y}$.
$\widehat{p}_s^{(r)}, p_s^{(r)\dagger}$	Optimization variable and resulting uncertainty- and fragility-adjusted TS probability.
$\mathcal{Z} = \{c^{(1)}, \ldots, c^{(C)}\}$	Set of semantic clusters, with C denoting the number of observed clusters.
$c^{(\ell)}$	The ℓ-th semantic cluster, whose members are mutually semantically consistent generations.
$p_c^{(\ell)}$	Probability assigned to semantic cluster $c^{(\ell)}$.
SE_S	Shannon SE over the semantic-cluster distribution.
SE_R	Base quadratic Rényi SE; in the ablation study, it contains neither TS-level UQ nor fragility correction.
Occupancy statistics, unseen mass, and fragility	
n	Number of i.i.d. sampled generations used to form the occupancy statistics; $n = 10$ in the experiments.
t	Number of additional generations for which new semantic species are predicted; $t = 10$ in the experiments, giving extrapolation ratio $t/n = 1$.
p_ℓ	Unknown population probability of semantic cluster or species ℓ.
N_ℓ	Number of times cluster or species ℓ appears among the n samples.
Φ_k	Number of clusters or species observed exactly k times; in particular, Φ_1 and Φ_2 are the singleton and doubleton counts.
Φ_{obs}	Total number of distinct observed clusters or species.
P_{obs}	Total population probability mass of clusters observed at least once.
P_k	Population probability mass of clusters observed exactly k times; P_1 is the singleton mass.
$M_0 = 1 - P_{\mathrm{obs}}$	Missing mass, i.e., the total probability of semantic clusters not observed in the finite sample.
$F_{\mathrm{mass}} = P_1/P_{\mathrm{obs}}$	Mass-based fragility: the fraction of observed probability mass assigned to singleton clusters.
$F_{\mathrm{count}} = \Phi_1/\Phi_{\mathrm{obs}}$	Count-based fragility: the fraction of observed clusters that are singletons.
$\widehat{M}_{\mathrm{GT}} = \Phi_1/n$	Good–Turing base estimator of the missing mass.
$\widehat{U}_{\mathrm{GT}}(t)$	Good–Toulmin estimate of the number of new semantic species expected among t additional generations.
$U_{GT_o}(t), U_{\max}$	Clipped, normalized Good–Toulmin reference used in the entropy–KL objective and its upper bound, with $U_{\max} = t = 10$.
ε	Small positive clipping constant that prevents singular Bernoulli KL terms.
$u_0^\star, \widehat{U}_0$	Optimized normalized unseen-count estimate and final unseen-species estimate, where $\widehat{U}_0 = u_0^\star U_{\max}$.
$M_0^\star$	Fragility-adjusted missing-mass estimate obtained by entropy–KL optimization.
Uncertainty adjustment and method variants	
$\mathrm{UQ}(p_s^{(r)})$	Observed TS probability sensitivity associated with generation r (retaining the TS-level UQ notation of the prior formulation).
$D_{\mathrm{KL}}(\cdot\|\|\cdot)$	Kullback–Leibler (KL) divergence used to preserve fidelity to a base probability or estimator.
$h(\cdot)$	Binary entropy function used in the unseen-count and missing-mass corrections.
λ	Weight of the UQ-modulated KL fidelity term in the SURF TS adjustment.
γ	Weight of the mass-fragility penalty in the SURF TS adjustment.

Continued on the next page

Table 11 – continued from the previous page

Notation	Definition
SE_{RU}	Rényi SE with observed TS probability sensitivity and without the fragility correction.
SE_R + fragility	Fragility-only ablation: Rényi SE with fragility correction and without TS-level UQ.
$\mathrm{SE}^{\dagger}_{RU}$ (SURF)	Full fragility-aware semantic uncertainty method combining TS-level UQ and fragility correction.
Evaluation quantities	
ΔAUROC	Paired AUROC difference between SURF and a comparison method.
HRG	Headroom gain: AUROC improvement normalized by the performance headroom of the reference method.
RAC, AURAC	Rejection accuracy curve and the area under that curve, respectively.