Diagnosing LLMs via Information Spectrum Analysis: Tail Behavior and the Effects of Side Information

Yuuki Tachioka
Denso IT Laboratory / 4-3-1 Shimbashi, Minato-ku, Tokyo, Japan
tachioka.yuki@core.d-itlab.co.jp

Abstract

Large language models (LLMs) exhibit non-stationary generation: their output distributions shift with prompts, retrieved documents, and decoding conditions. Under such variability, average likelihood metrics can obscure heterogeneous behaviors across samples, especially in high-surprisal tails where failures often occur. We propose an information-spectrum-based diagnostic framework that treats LLMs as general sources without assuming stationarity, ergodicity, or the asymptotic equipartition property. We define sequence-level self-information density (coding rate; mean surprisal) and construct an empirical information spectrum from finite samples, enabling operational estimates of spectrum quantiles and width. We further introduce an information gain spectrum, a teacher-forced likelihood-based measure that evaluates the same generated sequence with and without side information. Across multiple Japanese LLMs and QA settings, we observe that correctness differences are often more visible in the high-surprisal tail than in the mean coding rate, and that side information can reshape tail behavior in heterogeneous ways across sequences. We also observe that instruction tuning changes the spectrum structure, making tail statistics and spectrum width more predictive of correctness than the mean coding rate. Overall, our analysis illustrates how spectrum-based diagnostics complement average-based metrics for understanding conditional generation.

1 Introduction

Large language models (LLMs) exhibit substantial shifts in their generation distributions depending on prompts and contextual inputs, and thus behave as *non-stationary* information sources that involve mixtures of training distributions and variations in inference-time conditions. Under such variability, average metrics such as log-likelihood or cross-entropy may fail to capture the diversity of generation behaviors and failure cases (e.g., hallucinations).

Information theory has traditionally developed around asymptotic analyses based on typical sets, focusing on stationary and ergodic sources for which the asymptotic equipartition property (AEP) holds (Shannon, 1948). However, such assumptions often do not apply to non-stationary sources such as LLMs. To remove these restrictions, Han and Verdú proposed information spectrum (IS) theory, a framework for analyzing general sources without assuming stationarity or ergodicity. By treating self-information density as a random variable, they showed that achievable performance and fundamental limits can be characterized through its distributional structure (the *spectrum*) (Han and Verdu, 1993; Han, 2003).

This paper views LLMs as general information sources and leverages the IS framework to diagnose their generation behaviors. This enables a practical diagnostic methodology that reveals tail dynamics and heterogeneous effects of side information that are difficult to observe through mean-based metrics alone. Specifically, we define a sequence-level self-information density (coding rate) for each generated sequence and construct an empirical IS from a finite set of generated samples. Furthermore, we introduce an *information gain spectrum* (IG spectrum), which evaluates spectrum differences with and without additional context such as retrieval-augmented generation (RAG) (Lewis et al., 2020) through teacher forcing on the same generated sequence, thereby identifying *which sequences* benefit from side information at the sequence level.

Our main contributions are as follows. 1) To treat LLMs as general information sources, we define a sequence-level self-information density (sequence-level coding rate) and provide an operational framework for estimating IS from fi-

Findings of the Association for Computational Linguistics: ACL 2026, pages 12231–12253

nite samples (empirical spectra and their quantile-based statistics). 2) Based on teacher forcing with fixed generated sequences, we propose the IG spectrum, which quantifies how effectively side information improves coding rates on a per-sequence basis. 3) Through experiments across multiple models and tasks, we demonstrate that the proposed framework consistently visualizes distinctions that are difficult to capture with average metrics. In particular, we show that (i) correctness differences emerge primarily in the high-surprisal tail rather than in the mean, (ii) side information such as answer choices, RAG, and knowledge graphs (KGs) systematically reshapes the spectrum, with highly heterogeneous benefits across sequences, and (iii) instruction tuning can shift the key determinants of performance from mean self-information to distributional width and tail structure. Overall, this paper provides an information-theoretic framework for analyzing LLMs as general sources and suggests a connection to extensions of IS theory that treat additional context as side information (Kuzuoka and Watanabe, 2015).

2 Related Work

Information theory under AEP and general sources Shannon's information theory has revealed fundamental properties of information sources and communication channels through quantities such as entropy and mutual information (Shannon, 1948). However, many classical results rely on assumptions of stationarity and ergodicity, and may not apply to non-stationary sources, including mixed sources (Han and Verdu, 1993; Verdu and Han, 1994; Ahlswede et al., 2006). To remove these restrictions, Han and Verdú proposed IS theory as a framework for analyzing general sources without stationarity or ergodicity assumptions, showing that achievable performance and fundamental limits can be characterized based on the distribution of self-information density (Han and Verdu, 1993; Han, 2003). While IS theory has been extensively developed for limit analyses such as source coding, its use as an empirical diagnostic tool for modern conditional generative models remains largely unexplored.

Uncertainty and calibration in LLMs A growing body of work analyzes LLM behavior through uncertainty and calibration. Confidence-based analyses examine the relationship between model likelihood and correctness (Kadavath et al., 2022; Farquhar et al., 2024), while survey works study hallucination phenomena and reliability issues in generation (Ji et al., 2023; Huang et al., 2025). Other approaches evaluate prompt or context effects using likelihood differences or conditional probability comparisons (Yang et al., 2024). These methods typically rely on point estimates (e.g., average log-likelihood, entropy reduction, or scalar uncertainty scores). In contrast, our work emphasizes the distributional structure of sequence-level self-information density and its tail behavior, aiming to capture inter-sample variability and heterogeneous conditioning effects.

LLMs as compressors and information-based analyses Another line of work views LLMs as compressors and studies alignment or fine-tuning effects through compression rates derived from negative log-likelihood (Ji et al., 2025). Similarly, perplexity and cross-entropy remain standard evaluation metrics (Fang et al., 2025). However, average compression-based metrics may obscure heterogeneous tail behavior across sequences. Our approach complements these analyses by explicitly examining spectrum width and tail quantiles rather than relying solely on mean quantities.

Positioning of this paper This paper focuses on the empirical distribution of sequence-level self-information density (the empirical IS) and investigates where correctness differences concentrate within that distribution. Furthermore, by introducing an IG spectrum based on teacher forcing, we provide a controlled, per-sequence likelihood-based diagnostic for analyzing heterogeneous effects of side information while preserving per-sequence correspondence.

3 Information-Spectrum (IS) Theory

3.1 Probabilistic limsup and liminf

For a general source, let $P_{X^n}(X^n)$ denote the probability assigned by the source to a symbol sequence X^n with length n. The self-information density is defined as

$$Z_n := -\frac{1}{n} \log P_{X^n}(X^n).$$

For stationary ergodic sources, Z_n concentrates around the entropy rate of the AEP, and a single mean quantity is often sufficient. In contrast, for non-AEP sources (e.g., non-stationary or mixture

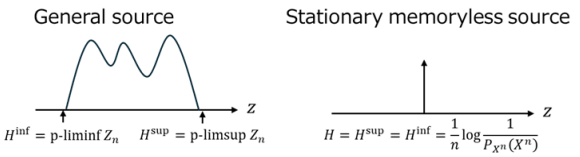


Figure 1: Schematic illustration of an information spectrum (IS). The horizontal axis is the per-symbol self-information density Z_n (coding rate). The bulk corresponds to typical sequences, while the right tail corresponds to rare high-surprisal sequences that often dominate failure cases (Appendix for intuition). For AEP sources, Z_n concentrates at the entropy rate and the spectrum degenerates to a point.

sources), Z_n may not concentrate and tail events can persist with non-vanishing probability, motivating distributional (spectrum) analysis.

The distribution of Z_n is called the IS (Han and Verdu, 1993; Han, 2003). The key properties of general sources can be characterized by the probabilistic limsup $H_\epsilon^{\sup}$ (weak upper entropy rate), the probabilistic liminf $H_\epsilon^{\inf}$ (weak lower entropy rate) and their width $H_\epsilon^w = H_\epsilon^{\sup} - H_\epsilon^{\inf}$:

$$\begin{aligned} H_\epsilon^{\sup} &= \mathrm{p}_\epsilon\text{-}\limsup_{n\to\infty} Z_n \\ &:= \inf\left\{\alpha \;\middle|\; \limsup_{n\to\infty} \Pr[Z_n > \alpha] \le \epsilon\right\}, \\ H_\epsilon^{\inf} &= \mathrm{p}_\epsilon\text{-}\liminf_{n\to\infty} Z_n \\ &:= \sup\left\{\beta \;\middle|\; \limsup_{n\to\infty} \Pr[Z_n < \beta] \le \epsilon\right\}. \end{aligned} \tag{1}$$

Here, ϵ relates to variable-length source coding that permits an error probability up to ϵ. Intuitively, $H_\epsilon^{\sup}$ reflects the coding rate of the high-surprisal (rare) tail, while $H_\epsilon^{\inf}$ reflects that of typical sequences. Importantly, two sources can share the same mean coding rate while exhibiting drastically different tail behavior; we provide toy examples and intuition for p_ϵ-$\limsup$ / $\liminf$ in Appendix A.2. Figure 1 shows that general sources yield a non-zero-width IS, whereas AEP sources concentrate at the entropy rate.

3.2 Finite-length approximation: empirical IS and statistics

To compute the quantities in Eq. (1) explicitly, one must consider the limit where the length of a single sequence goes to infinity ($n \to \infty$). However, in analyzing LLMs, we can generate only finite-length sequences. Therefore, for a finite collection of generated samples, we introduce an operational definition that corresponds to the probabilistic upper/lower limits of the IS, based on empirical quantiles, denoted as p_θ-$\sup$ / $\inf$ (where θ is a tail probability). For generated samples, we compute the self-information density $Z^{(s)}$ by Eq. (2) and empirically approximate the IS.

Let the output token sequence for sample $s = 1, \ldots, S$ with length T_s be $t^{(s)} = (t_1^{(s)}, \ldots, t_{T_s}^{(s)})$. For sample s, the conditional self-information density (i.e., coding rate / mean surprisal) conditioned on the token history and the prompt $y^{(s)}$, $Z^{(s)} = Z(t^{(s)}; y^{(s)})$, is normalized by T_s since sequence lengths differ across samples:

$$Z(t^{(s)}; y^{(s)}) := -\frac{1}{T_s}\sum_{k=1}^{T_s} \log P(t_k^{(s)} \mid t_{<k}^{(s)}, y^{(s)}). \tag{2}$$

Here, $y = (y^{(1)}, \ldots, y^{(S)})$ denotes the set of prompts that include task instructions. The collection $\mathcal{Z} = \{Z^{(s)}\}_{s=1}^S$ constitutes the *empirical* IS for finite samples, and we define its operational upper and lower limits as follows:

$$\begin{aligned} \hat{H}_\theta^{\sup} &= \mathrm{p}_\theta\text{-}\sup \mathcal{Z} \\ &:= \inf\left\{\alpha \;\middle|\; \frac{\sum_{s=1}^S \mathbf{1}(Z^{(s)} > \alpha)}{S} \le \theta\right\}, \\ \hat{H}_\theta^{\inf} &= \mathrm{p}_\theta\text{-}\inf \mathcal{Z} \\ &:= \sup\left\{\beta \;\middle|\; \frac{\sum_{s=1}^S \mathbf{1}(Z^{(s)} < \beta)}{S} \le \theta\right\}, \end{aligned} \tag{3}$$

where $\mathbf{1}$ is the indicator function and θ denotes the tail probability. These values correspond to the $(1-\theta)$-quantile and the θ-quantile of the empirical distribution, respectively. Thus, a smaller θ places greater emphasis on the high-surprisal tail. The width is defined as $\hat{H}_\theta^w = \hat{H}_\theta^{\sup} - \hat{H}_\theta^{\inf}$.

3.3 Information gain (IG) spectrum (teacher-forced likelihood gain)

To examine how side information such as RAG affects sequence-level information density (Liu and Wang, 2023), we evaluate the difference in Z for each sample s. First, we generate a token sequence $t'^{(s)}$ using a prompt that includes additional information $\Delta y^{(s)}$: $y'^{(s)} = y^{(s)} + \Delta y^{(s)}$, and compute $Z(t'^{(s)}; y'^{(s)})$. Next, to measure the effect of the added information, we fix $t'^{(s)}$ and, under teacher forcing, evaluate $Z(t'^{(s)}; y^{(s)})$ conditioned on the original prompt $y^{(s)}$ and the past tokens $t'^{(s)}_{<k}$. We

then compute

$$\Delta Z^{(s)} := Z(t'^{(s)}; y'^{(s)}) - Z(t'^{(s)}; y^{(s)}).$$

We call the distribution $\Delta\mathcal{Z} = \{\Delta Z^{(s)}\}_{s=1}^{S}$ the *IG spectrum*. The IG spectrum measures how much side information changes the sequence-level negative log-likelihood (coding rate) of a fixed generated sequence, capturing heterogeneity across samples. Since Z is the negative logarithmic likelihood of the sequence, $\Delta Z^{(s)} < 0$ indicates that side information increases the likelihood of the *same* sequence under teacher forcing.

Note (IG spectrum as an operational definition) The IG spectrum is an operational (finite-sample) measure that compares the likelihood of the *same* generated sequence with and without side information under teacher forcing. It should therefore be interpreted as a controlled likelihood-based diagnostic rather than a causal estimate of how side information reshapes the full generation distribution. In particular, it does not directly quantify how the probability mass shifts across different sequences when side information is introduced. We adopt this formulation to preserve per-sequence correspondence across conditions and to avoid ambiguities arising from independently sampled generations. This allows us to diagnose which generated sequences are strongly supported by the added information and which are not, while maintaining stable and computationally efficient estimation.

Finally, note that the IS in this paper empirically treats the *distribution of sequence-level negative log-likelihood*. Its core value lies not in the mean but in tail-aware distributional diagnostics. IS theory provides a language for organizing and comparing distributional structures of conditional, non-stationary generation behaviors such as those exhibited by LLMs, without assuming stationarity.

4 Experiments

4.1 Setup

To investigate how IS changes depending on the task structure and the form of side information when LLMs are regarded as general sources, we design experiments that combine two task families while systematically varying the type of conditioning[1]. For QA, we judge correctness by checking whether the generated answer contains the gold string and compute accuracy accordingly. For generation tasks, we evaluate diversity using Distinct-2 (the number of unique 2-grams divided by the total number of 2-grams in the generated sequences) (Li et al., 2016).

QA tasks For QA, we select three tasks that differ in how side information is provided:

1. Commonsense QA[2] [C-QA] (1,119 questions) (Kurihara et al., 2022). We compare free-form generation[3] and a multiple-choice setting. This contrast isolates the effect of constraining the output space via explicit answer options (free-form vs. choice-constrained).

2. Knowledge QA[4] [K-QA] (3,939 questions) (So et al., 2022). We compare conditions without and with RAG[5]. This contrast evaluates the effect of the RAG context on reasoning.

3. Hop QA[6] [H-QA] (1,059 questions) (Ishii et al., 2024). We compare conditions without and with a KG[7]. This contrast evaluates the effect of structured knowledge for multi-hop reasoning.

This design allows us to analyze, within a unified framework, how different forms of side information—output-space constraints, retrieved documents, and structured knowledge—affect both the IS and the IG spectrum. Although QA answers are typically only a few words long, we set the maximum generation length to 128 tokens to mitigate spectrum distortions caused by premature truncation.

Generation tasks For generation, we use two tasks: *news* generation, which tends to be more templated with relatively fixed structure, and *poem*

[1] https://github.com/DensoITLab/information_spectrum_LLM

[2] https://huggingface.co/datasets/sbintuitions/JCommonsenseQA (validation split)

[3] The original dataset is multiple-choice; we evaluate it in a free-form setting by removing the choices.

[4] https://huggingface.co/datasets/SkelterLabsInc/JaQuAD (validation split)

[5] The dataset provides retrieval context that can support answering.

[6] https://huggingface.co/datasets/sbintuitions/JEMHopQA (train split); The validation split is small, so we use the training split, but since the models are not trained on it, this does not constitute data leakage.

[7] The dataset provides KG information required for answering.

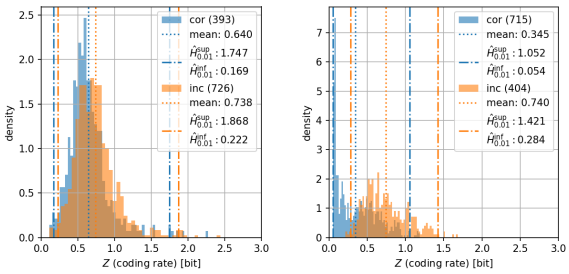
generation, which is more creative with a broader output space. We compare whether spectrum statistics behave differently depending on whether the output is structured/templated or not. For each prompt topic, we generate 1,000 sequences for both news and poem. Although the target length is 300 Japanese characters, we set the maximum generation length to 1,024 tokens to avoid spectrum artifacts due to truncation.

Model comparison and spectrum computation We take `Llama-3.1-Swallow-8B-v0.5` (Fujii et al., 2024), a widely used Japanese LLM, as the primary reference model. We compare the base model with its instruction-tuned variant (`Llama-3.1-Swallow-8B-Instruct-v0.5`) to examine how instruction tuning affects IS structure. In addition, we include `Qwen3-8B` (Team, 2025), which differs in training data and alignment policies, to verify whether the proposed diagnostics and observed trends are generalized between model families.

We vary the temperature τ from 0.4 to 1.2, using $\tau = 0.6$ as the default, since it is commonly recommended for reasoning tasks. For the quantile parameter θ, we determine $\theta = 0.01$ using bootstrap analysis; details are provided in Appendix C.2. IS computation uses only the output token sequence $t^{(s)}$. All prompts are provided as plain text, and we do not use chat templates or explicit role indicators such as system/user/assistant. This unifies the input format across models and eliminates confounding effects due to template design when comparing the impact of side information. We assess finite-sample stability via bootstrap and find that estimates stabilize around $S \gtrsim 300$ (Appendix Fig. 9).

4.2 Error signals appear in the tail rather than in the mean; temperature controls the tail.

Figure 2(a) compares the IS of correct vs. incorrect outputs for C-QA (free-form generation). The two distributions overlap substantially, so the mean alone cannot cleanly separate the correctness. In contrast, the difference becomes pronounced in the high-surprisal tail: incorrect sequences dominate the high-surprisal region, whereas correct sequences concentrate in the low-surprisal region. This suggests that errors manifest not primarily as a "worsening of the mean," but as the emergence of atypical sequences in the tail. Consequently,



(a) Free-form (b) Multiple-choice

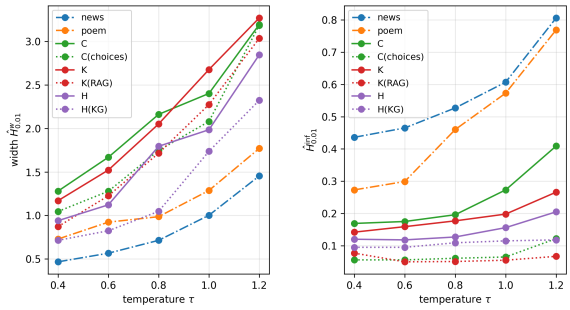
Figure 2: IS for commonsense QA (C-QA) with Swallow-8B under two answer formats: (a) free-form generation and (b) multiple-choice (answer choices provided). Histograms show the empirical distribution of sequence-level coding rates Z (mean surprisal) over generated answers; "cor" and "inc" denote correct and incorrect samples, respectively (counts shown in parentheses). Vertical lines mark the mean and tail-focused quantiles ($\hat{H}^{\mathrm{inf}}_{0.01}$ and $\hat{H}^{\mathrm{sup}}_{0.01}$; legend), which summarize typical behavior and the high-surprisal tail. Providing answer choices changes the distribution shape and is associated with a reduced upper-tail mass compared to free-form generation, making correctness differences more interpretable in the tail region.

we analyze both the quantile statistic that characterizes the tail ($\hat{H}^{\mathrm{sup}}_{\theta}$) and the IS width ($\hat{H}^{w}_{\theta}$), in addition to the mean, to diagnose where errors occur and how unstable the generation is.

Figure 2(b) shows the IS for the same C-QA task when answer choices are provided (multiple-choice condition). Because the output space is strongly constrained by the choices, the overall distribution contracts and the expansion of the IS width (tail thickening) is suppressed. Such spectrum-shape changes induced by side information are also observed in other tasks (RAG and KG; Figures 4 and 11).

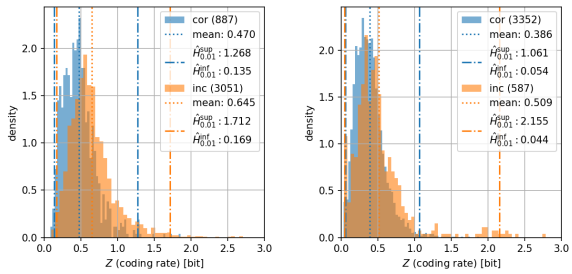
Temperature and the IS Figure 3 shows the relationship between the decoding temperature τ and the IS statistics ($\hat{H}^{w}_{\theta}$ / $\hat{H}^{\mathrm{inf}}_{\theta}$). We vary the temperature τ on five levels from 0.4 to 1.2. In summary, increasing temperature monotonically expands the IS width (tail thickness): it directly leads to increased errors (performance degradation) in QA, while forming a trade-off with improved diversity in generation tasks.

As shown in Figure 3(a), $\hat{H}^{w}_{\theta}$ increases monotonically with temperature on all tasks, indicating that the IS widens as τ increases. This occurs because a higher temperature activates the tail of the output distribution and increases sample-wise variability in surprisal. Moreover, intro-



(a) Width $\hat{H}_\theta^w$ (b) $\hat{H}_\theta^{\inf}$

Figure 3: Temperature dependence of the IS width and lower quantile across multiple QA tasks with Swallow-8B (C: Commonsense, K: Knowledge, H: Hop).



(a) Without RAG (b) With RAG

Figure 4: IS for knowledge QA (K-QA) with Swallow-8B, comparing (a) without RAG and (b) with RAG. RAG primarily affects the high-surprisal tail: incorrect samples concentrate more strongly in the upper tail, and providing retrieved context is associated with a reduced tail expansion and a shift toward lower coding rates for a subset of samples.

ducing side information such as answer choices, RAG, or KGs reduces the width even at the same temperature, confirming that side information can constrain the generation distribution. In contrast, generation tasks such as news and poem exhibit smaller widths than QA tasks, and in particular, news generation has a narrow width and is highly templatic.

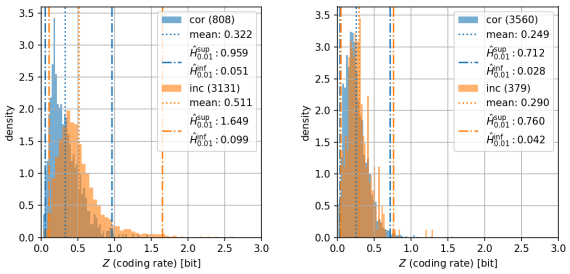
Figure 3(b) shows the behavior of $\hat{H}_\theta^{\inf}$, which represents the low-surprisal (typical) side. Changes in $\hat{H}_\theta^{\inf}$ are relatively moderate compared to those in width, suggesting that temperature mainly affects not the center of the typical set but the increase in atypical sequences (i.e., expansion of the upper tail). In addition, $\hat{H}_\theta^{\inf}$ exhibits strong task dependence and, in generation tasks, it tends to decrease dramatically as the temperature decreases. This indicates that under low-temperature decoding, the output distribution concentrates strongly on the typical set, making templatic sequences more likely to be generated.

Although high-temperature sampling activates the tail and increases diversity (Holtzman et al., 2020), uncontrolled increases in temperature can degrade performance on reasoning tasks (Minh et al., 2025; Wang et al., 2023). Our results are consistent with these and suggest that temperature is a primary factor that balances "concentration on the typical set" against "tail inflation."

4.3 Side information systematically reshapes the spectrum, and the IG spectrum separates sequences that benefit from side information

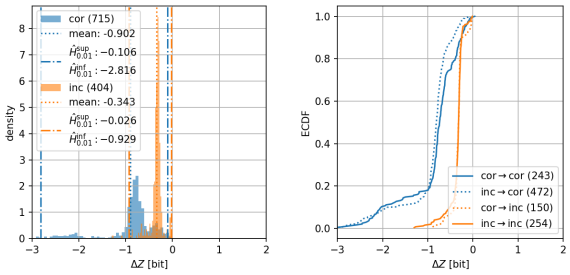
In this section, we analyze how side information (choices, RAG, and KG) affects the IS and IG spectrum. Figure 4 shows the IS for K-QA with and without RAG. Figures 2(b) and 4 show that the introduction of side information tends to shrink the distribution in many settings, suppressing the IS width (i.e., inflation of the tail). For correct sequences, adding RAG suppresses the high-surprisal tail and lowers $\hat{H}_\theta^{\sup}$, whereas for incorrect sequences, the tail can expand instead, suggesting that when retrieved evidence is not properly utilized, the output may shift toward more atypical sequences. The tendency that side information such as answer choices and RAG reduces IS width and suppresses tail inflation is also reproduced consistently across model families (Appendix Figures 14 and 16). These results indicate that side information appears not only as a reduction in the mean but also as a systematic change in distributional shape (especially in the tail). In the following, we use the IG spectrum to analyze *which sequences* benefit from side information, treating the effect as an intervention response while preserving sample-wise correspondence.

Next, Figure 6 shows the IG spectrum induced by the answer choices (side information). The IG spectrum is a distribution of intervention responses that describes how much side information changes the coding rate for each sequence. Here, $\Delta Z^{(s)} < 0$ indicates that side information improves likelihood (i.e., reduces the coding rate). In Figure 6(a), the correct sequences are generally more negatively shifted; however, this reflects not that "being correct" itself causes the coding rate to decrease, but rather that sequences with larger coding rate reductions due to side information are more likely to transition to being correct. Therefore, the IG spectrum provides a framework



(a) without RAG (b) with RAG

Figure 5: IS for K-QA with Swallow-8B-Instruct. Compared to the base model (Fig. 4), the instruct-tuned model exhibits a different spectrum structure, where correctness differences are more strongly reflected in tail statistics and width rather than the mean alone.

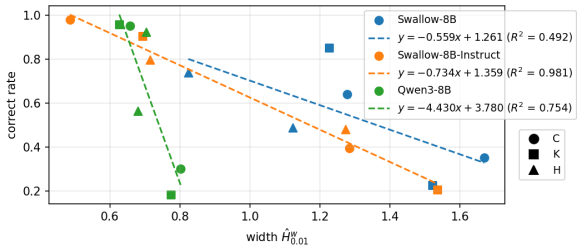


(a) IG spectrum (b) ECDF

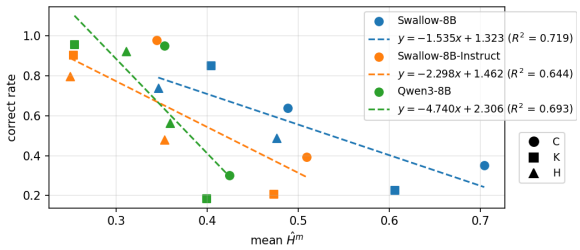
Figure 6: IG spectrum for C-QA with Swallow-8B. The distribution of coding-rate differences induced by side information (answer choices) is compared across correctness transitions.

for distinguishing, as a distribution, sequences for which side information is effective.

The empirical cumulative distribution function (ECDF) in Figure 6(b) partitions the sequences into four groups, which are confirm (cor→cor), fix (inc→cor), break (cor→inc), and none (inc→inc), according to the correctness transitions from without to with side information, and visualizes *which groups* shift toward the negative side in ΔZ. In the multiple-choice setting, *fix* exhibits the largest negative shift, indicating that answer choices selectively capture sequences for which side information contributes to error correction. In contrast, *none* concentrates near zero, suggesting that likelihood improvements are unlikely even with added choices. *confirm* also shifts slightly negative, consistent with the interpretation that hypotheses formed under free-form answering are corroborated by the candidate set. The separation strength depends on task structure: in H-QA, missing information can be supplemented and *fix* is more clearly separated, while in K-QA the separation is weaker because even *none* can exhibit likelihood improvements (Appendix Figure 12). We report identification performance of the IG-based indicators for other tasks in the Appendix G.2. Overall, the IG spectrum enables visualization of side-information effects as sequence-level intervention responses.



(a) Width $\hat{H}^w_\theta$



(b) Mean

Figure 7: Relationship between IS width/mean and QA accuracy. For each model, we compare six settings: three baseline conditions (without side information) and three conditions with side information.

4.4 IS statistics correlate with performance, and instruction tuning shifts the informative region

In this section, we investigate how IS statistics relate to task-level performance (accuracy). First, we show via correlation analysis that when conditions such as the presence/absence of side information and temperature change, performance varies accordingly, and IS width as well as upper/lower quantiles also change systematically. Next, after controlling for generation conditions (temperature τ and presence/absence of side information), we evaluate how much the mean ($\hat{H}^m$) and width ($\hat{H}^w_\theta$) can additionally explain the accuracy (incremental explanatory power).

Correlation between spectrum statistics and accuracy Figure 7 shows, for each model, the relationship between IS statistics and accuracy under six settings: three "without side information" conditions plus three "with side information" con-

ditions[8]. We observe that the accuracy tends to decrease as the IS width/mean increases. This suggests that as variability in generation behavior (atypical sequences / tail mass) grows, stable generation for QA becomes more difficult. Furthermore, we verified robustness of the correlations via leave-one-out (LOO) regression: the slopes for $\hat{H}_\theta^{\text{sup}}$ and width are consistently negative across all models, confirming that the observed correlations do not rely on a single outlier (Appendix Tables 5 and 6).

Incremental explanatory power under controlled conditions (mean vs. width) Next, after controlling for temperature τ and the presence/absence of side information (info), we evaluated how much distributional indicators (mean $\hat{H}^m = \frac{1}{S}\sum_s Z^{(s)}$ and width $\hat{H}_\theta^w$) can additionally explain accuracy y. Specifically, we compared the following linear regression models: (i) condition-controlled baseline (τ, info), (ii) baseline + $\hat{H}^m$, (iii) baseline + $\hat{H}_\theta^w$, (iv) baseline + $\{\hat{H}^m, \hat{H}_\theta^w\}$, and evaluated generalization performance using LOO cross-validated R^2.

To compare incremental explanatory power under controlled conditions, we define the normalized incremental coefficient of determination as

$$\Delta R^2(z) = \frac{R^2(\tau, \text{info}, z) - R^2(\tau, \text{info})}{1 - R^2(\tau, \text{info})}$$

where $z \in \{\hat{H}^m, \hat{H}_\theta^w\}$. As shown in Table 1, the condition-controlled baseline (τ, info) already achieves high explanatory power, indicating that temperature and side information are the main determinants of accuracy. However, effective distribution indicators depend on the model. For Swallow-8B (base), adding the mean $\hat{H}^m$ yields the largest improvement, and the width contribution is relatively small. In contrast, for Swallow-8B-Instruct, the mean contributes little and width becomes dominant. A similar trend is observed in the LOO evaluation, and when comparing prediction errors, the width is significantly smaller than the mean[9]. Qwen3-8B also shows a tendency that width performs better than mean.

Interpretation: instruction tuning shifts cues from "mean" to "distributional shape" Overall, these results suggest that for base models close to pretraining, mean self-information can be a primary driver of performance, whereas for instruction-tuned models, performance differences become harder to capture using the mean alone, and information based on distribution width and tail structure becomes more important. This is consistent with an IS perspective in which instruction tuning may rearrange the geometry of the typical and tail regions of the generation distribution, shifting error-detection cues from the mean toward shape/tail characteristics.

Table 1: For the baseline model with temperature τ and presence/absence of side information info as explanatory variables, we report the coefficient of determination R^2 and the normalized incremental coefficient of determination ΔR^2 when additionally including the spectrum mean ($= \hat{H}^m$) or width ($= \hat{H}_\theta^w$). Leave-one-out estimates are also shown.

Metric	Swallow-8B base	Swallow-8B instruct	Qwen3-8B
in sample R^2			
$R^2(\tau, \text{info})$	0.787	0.883	0.879
$R^2(\tau, \text{info}, \hat{H}^m)$	0.869	0.890	0.900
$R^2(\tau, \text{info}, \hat{H}_\theta^w)$	0.840	0.937	0.914
$R^2(\tau, \text{info}, \hat{H}^m, \hat{H}_\theta^w)$	0.870	0.938	0.915
$\Delta R^2(\hat{H}^m)$	**0.388**	0.056	0.167
$\Delta R^2(\hat{H}_\theta^w)$	0.251	**0.460**	**0.290**
leave-one-out R^2			
$R^2(\tau, \text{info})$	0.736	0.855	0.851
$R^2(\tau, \text{info}, \hat{H}^m)$	0.827	0.853	0.867
$R^2(\tau, \text{info}, \hat{H}_\theta^w)$	0.793	0.917	0.883
$R^2(\tau, \text{info}, \hat{H}^m, \hat{H}_\theta^w)$	0.820	0.913	0.881
$\Delta R^2(\hat{H}^m)$	**0.345**	-0.0138	0.107
$\Delta R^2(\hat{H}_\theta^w)$	0.216	**0.428**	**0.215**

4.5 Discussion

Our analysis suggests that performance-relevant differences are often more visible in the high-surprisal region (the tail) than in the mean coding rate alone. In particular, variability between sequences, quantified by the width of the spectrum $\hat{H}_\theta^w = \hat{H}_\theta^{\text{sup}} - \hat{H}_\theta^{\text{inf}}$, reflects heterogeneity in generation behavior and can be interpreted as a risk-sensitive characteristic of the model. Instruction tuning appears to shift the relative importance from the mean coding rate toward distributional properties such as width and tail structure, suggesting that improvements may stem not only from raising overall likelihood but also from suppressing high-surprisal (failure-prone) sequences.

The IG spectrum further visualizes the heterogeneous, sequence-dependent effects of side information. Beyond descriptive analysis, this perspective enables practical diagnostics. For example,

[8] Conditions with side information appear as points with higher accuracy than those without it.

[9] paired t-test $p = 5.9 \times 10^{-5}$, Wilcoxon $p = 1.1 \times 10^{-4}$

consistent left-shifts in coding-rate distributions or strongly negative IG values can serve as operational signals to select between alternative generations (e.g., with vs. without RAG) even without gold labels. In addition, comparing IG distributions across models or prompting strategies can reveal which configurations are more responsive to side information, supporting prompt design and model comparison. Overall, spectrum-based analysis provides a complementary lens to average-based metrics to understand and control conditional generation. Such risk-aware diagnostics may also inform decoding control (e.g., temperature selection) by explicitly monitoring tail expansion.

5 Conclusion

In this paper, we treat LLMs as general information sources and proposed a diagnostic method based on the distribution of sequence-level coding rates (the information spectrum). By constructing empirical IS from a finite set of generated samples, we demonstrated through experiments across multiple models and tasks that errors and low-quality generations concentrate in the high-surprisal region rather than being captured by average metrics. We further showed that the effects of side information (multiple-choice options, RAG, and KGs) are heterogeneous across sequences and that the IG spectrum enables visualization of the distribution of these effects. In addition, we found that instruction tuning can amplify the tendency for distribution width to dominate performance more than the mean, revealing that different models may exhibit distinct characteristics in suppressing failure sequences. These results highlight the usefulness of comparing and diagnosing LLM generation behavior as a distributional structure rather than relying solely on averages. Future work includes extending the analysis to a wider range of tasks and languages and exploring connections to analyses that directly capture deformations of the generation distribution.

Limitation

In this paper, we analyze the behavior of LLM generation as a distributional structure in self-information density based on information spectrum theory. However, several limitations remain.

Reliance on summary statistics of the distribution Because we summarized the characteristics of the information spectrum using quantile-based statistics ($\hat{H}_\theta^{\sup\,/\,\inf}$ and the width), we did not fully exploit the information contained in the distributional shape itself (e.g., skewness, multimodality, or tail asymmetry). Such shape information may enable for a more detailed characterization of how error sequences or hallucinations arise. Therefore, an important direction for future work is to introduce distributional distances (e.g., Wasserstein distance, KL approximations) and shape-related statistics, and to perform spectrum comparisons in a more direct manner.

The IG spectrum as a teacher-forcing approximation The IG spectrum used in this paper is an operational metric that measures *the likelihood improvement of the same fixed sequence* under side information via teacher forcing. It does not directly evaluate how conditioning transforms the generation distribution itself (i.e., the selection probabilities of sequences). In this work, we adopted this approximation to prioritize intervention visualization with preserved sequence alignment and stable estimation. Future work should extend the framework to metrics that more directly capture deformation of the generation distribution (e.g., distances between conditional distributions or intervention-induced reweighting), and it would be valuable to connect our empirical IG spectrum to the IS framework with side information (Kuzuoka and Watanabe, 2015), which provides an operational characterization of the coding limits under additional conditioning. Such a connection may enable for a more principled treatment of how auxiliary information reshapes the entire generation distribution.

Limitations of automatic evaluation In our experiments, the correctness was mechanically judged by checking whether the reference answer string was contained in the generated output. However, QA tasks may admit alternative correct answers or formulations, and therefore human evaluation or semantic equivalence modeling is preferable for strict correctness assessment (Ariyama et al., 2024).

In our main QA setting, the target answers are typically short and canonical (e.g., named entities, dates, factual phrases), and thus substantial paraphrasing is relatively uncommon. We also apply basic normalization (e.g., whitespace/punctuation normalization and canonical form matching) be-

fore string matching to reduce superficial mismatches.

Although automatic evaluation may still introduce misclassifications (false positives / false negatives), we applied the same rule across all models and all conditions. Hence, we expect our interpretation, which primarily focuses on relative comparisons (e.g., trends induced by temperature or side information), to be reasonably robust.

Finally, the goal of this paper is to provide an operational framework for treating LLMs as general information sources and to demonstrate the diagnostic potential of the distributional structure. The above limitations are therefore consistent with this objective.

Ethics and Broader Impacts

Our work provides diagnostic tools for analyzing the behavior of LLMs as information sources via empirical information spectra and teacher-forced likelihood gain spectra. Although we do not introduce new generative models, data collection, or deployment mechanisms, our analysis could influence how practitioners evaluate or trust LLM outputs. A potential risk is that our metrics may be misinterpreted as providing guaranties of factual correctness or safety; in particular, low tail surprisal does not necessarily imply factual reliability, and the teacher-forced IG spectrum does not capture distributional shifts in generation behavior. The misuse for cross-model comparisons without accounting for tokenization differences or dataset dependence could also lead to misleading conclusions. Our results highlight that side information (e.g., retrieval-augmented context or KG constraints) can both reduce and sometimes amplify high-surprisal tails, which suggests that poorly curated or biased external information sources could exacerbate misinformation or bias. We encourage users of these diagnostic methods to treat them as complementary signals, to validate conclusions with additional evaluation protocols, and to carefully audit any external information sources used in downstream systems.

Acknowledgments

We used generative AI tools for limited assistance with English proofreading and clarity improvements. All technical content, experimental design, analysis, and claims were developed and verified by the authors.

References

Rudolf Ahlswede, Lars Bäumer, Ning Cai, Harout Aydinian, Vladimir Blinovsky, Christian Deppe, and Haik Mashurian, editors. 2006. *General Theory of Information Transfer and Combinatorics*, 1 edition. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg.

Tomoki Ariyama, Jun Suzuki, Masatoshi Suzuki, Ryota Tanaka, Reina Akama, and Kyosuke Nishida. 2024. Achievements and challenges in Japanese question answering: Insights from quiz competition results. *Natural Language Processing*, 31(1):47–78.

Stefania Degaetano-Ortlieb and Elke Teich. 2017. Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 68–77, Vancouver, Canada. Association for Computational Linguistics.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. 2025. What is wrong with perplexity for long-context language modeling? *Preprint*, arXiv:2410.23771.

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

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities. In *First Conference on Language Modeling*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Te Sun Han. 2003. *Information-Spectrum Methods in Information Theory*, 1 edition. Stochastic Modelling and Applied Probability. Springer Berlin, Heidelberg. Original Japanese edition published by Baifukan, Tokyo, 1998.

T.S. Han and S. Verdu. 1993. Approximation theory of output statistics. *IEEE Transactions on Information Theory*, 39(3):752–772.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Ai Ishii, Naoya Inoue, Hisami Suzuki, and Satoshi Sekine. 2024. JEMHopQA: Dataset for Japanese explainable multi-hop question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9515–9525, Torino, Italia. ELRA and ICCL.

Jiaming Ji, Kaile Wang, Tianyi Alex Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. 2025. Language models resist alignment: Evidence from data compression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23411–23432, Vienna, Austria. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Shigeaki Kuzuoka and Shun Watanabe. 2015. An information-spectrum approach to weak variable-length source coding with side-information. *IEEE Transactions on Information Theory*, 61(6):3559–3573.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Hongfu Liu and Ye Wang. 2023. Towards informative few-shot prompt with maximum information gain for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15825–15838, Singapore. Association for Computational Linguistics.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*.

CE Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656.

ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. 2022. JaQuAD: Japanese Question Answering Dataset for Machine Reading Comprehension. *Preprint*, arXiv:2202.01764.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

S. Verdu and Te Sun Han. 1994. A general formula for channel capacity. *IEEE Transactions on Information Theory*, 40(4):1147–1157.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Sohee Yang, Jonghyeon Kim, Joel Jang, Seonghyeon Ye, Hyunji Lee, and Minjoon Seo. 2024. Improving probability-based prompt selection through unified evaluation and analysis. *Transactions of the Association for Computational Linguistics*, 12:664–680.

A Coding-Theoretic Interpretation of Information-Spectrum Quantities

Summary: The IS quantities $H_\epsilon^{\sup\,/\,\inf}$ characterize the achievable rates of a ϵ-variable-length source coding, and our $\hat{H}_\theta^{\sup\,/\,\inf}$ can be interpreted as a finite-sample operational approximation of them. In addition, we provide simple toy examples to build intuition for the information spectrum and the probabilistic $\limsup/\liminf$.

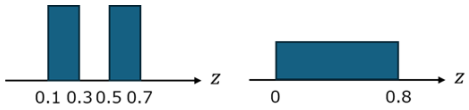


Figure 8: Toy examples where the mean coding rate is identical but the spectrum structure differs. Left: a bimodal mixture with mean $\mathbb{E}[Z] = 0.4$ but no probability mass near the mean. Right: a unimodal source with the same mean but broader support, yielding heavier extremes. This illustrates why tail quantiles and spectrum width provide information beyond the mean.

A.1 Operational interpretation via variable-length coding

The probabilistic upper and lower limits used in this paper, $H_\epsilon^{\sup}$ and $H_\epsilon^{\inf}$, can be interpreted in connection with ϵ-variable-length source coding for general sources, i.e., coding that allows an error probability not greater than ϵ (Han and Verdu, 1993; Han, 2003). Intuitively, $H_\epsilon^{\sup}$ corresponds to the high-surprisal side (rare sequences), whereas $H_\epsilon^{\inf}$ corresponds to the low-surprisal side (typical sequences).

More concretely, at any rate larger than $H_\epsilon^{\sup}$, there exists a code whose error probability converges to zero, so that all sequences can be encoded correctly with high probability. In contrast, at rates between $H_\epsilon^{\inf}$ and $H_\epsilon^{\sup}$, a code can be constructed whose error probability converges to zero for typical sequences, but errors for the remaining sequences (on the tail side) are unavoidable. Furthermore, at any rate below $H_\epsilon^{\inf}$, even for high-probability sequences, including the typical ones, it is impossible to make the probability of error converge to zero. Consequently, our $\hat{H}_\theta^{\sup\,/\,\inf}$ can be interpreted as a finite-sample operational approximation that corresponds to $H_\epsilon^{\sup\,/\,\inf}$.

A.2 Toy Examples for Intuition: Mixtures and Persistent Tails

Although the coding-theoretic interpretation provides an operational meaning, the definitions of p_ϵ-lim inf / sup can still feel abstract; here, we provide simple toy examples to build intuition for the information spectrum and the probabilistic limsup/liminf.

A.2.1 Mean Can Be Misleading: A Mixture Example

Consider two sources on a sequence-level coding rate Z, as shown in Fig. 8. The first is a bimodal mixture:

$$Z \sim \tfrac{1}{2}\mathrm{Unif}[0.1, 0.3] + \tfrac{1}{2}\mathrm{Unif}[0.5, 0.7].$$

The mean coding rate is $\mathbb{E}[Z] = 0.4$, yet there is no probability mass near 0.4; typical samples concentrate around two separate regions.

Now compare this with another source having the same mean but broader support:

$$Z \sim \mathrm{Unif}[0, 0.8].$$

Both sources share the same mean coding rate, yet their spectrum structure differs substantially. The second source exhibits heavier extremes, while the first has no mass near the mean. This illustrates that the mean alone cannot characterize heterogeneity or tail risk; spectrum width and tail quantiles provide complementary information.

A.2.2 Intuition for p-liminf and p-limsup

Consider a sequence of random variables Z_n.

Persistent tail. Suppose that

$$Z_n = \begin{cases} 0.2 & \text{with probability } 0.99, \\ 0.8 & \text{with probability } 0.01, \end{cases}$$

for all n. A high-surprisal tail remains with a constant probability even as $n \to \infty$. For any $\epsilon < 0.01$, we have

$$\mathrm{p}_\epsilon\text{-}\liminf_{n\to\infty} Z_n = 0.2, \quad \mathrm{p}_\epsilon\text{-}\limsup_{n\to\infty} Z_n = 0.8.$$

The spectrum retains non-zero width because tail events persist. That is, even if we ignore an ϵ fraction of rare events, the high-surprisal mode at 0.8 persists with non-vanishing probability. As ϵ decreases, these quantities approach the endpoints of the distribution support.

Vanishing tail. Now suppose that

$$Z_n = \begin{cases} 0.2 & \text{with probability } 1 - \frac{1}{n}, \\ 0.8 & \text{with probability } \frac{1}{n}. \end{cases}$$

Here, the probability of the high-surprisal event vanishes as $n \to \infty$. Indeed, for any fixed $\epsilon > 0$, there exists $N(\epsilon)$ such that for all $n \geq N(\epsilon)$,

$$\Pr[Z_n > 0.2] = \frac{1}{n} \leq \epsilon.$$

Therefore, for any $\epsilon > 0$,

$$\mathrm{p}_\epsilon\text{-}\liminf_{n\to\infty} Z_n = \mathrm{p}_\epsilon\text{-}\limsup_{n\to\infty} Z_n = 0.2.$$

Although outliers exist at finite n, they disappear asymptotically and thus do not contribute to the width of the IS.

These examples show that probabilistic $\liminf$ and $\limsup$ capture whether tail events persist with non-vanishing probability, rather than merely whether extreme values occur at finite length.

B Experimental Setup and Reproducibility

Summary: In our experiments, we controlled conditions using plain-text prompts and fixed random seeds, and reproducibly estimated IS statistics solely from self-information computed via teacher forcing.

B.1 Implementation Details

All experiments were implemented in Python using the HuggingFace Transformers library for model inference and teacher-forced log-likelihood evaluation. We used the model checkpoints listed in 4.1 and all experiments were run with the latest snapshot available on 2025/12/20. For decoding, we varied only the temperature τ as specified in Section 4.1 and kept other decoding settings fixed. We evaluate QA correctness by judging whether a gold answer is included in the answer string or not. We normalize both predictions and gold answers by Unicode NFKC, removing whitespace and punctuation/brackets, converting Kanji numerals to Arabic numerals, and adding common variants (e.g., YES/NO $\leftrightarrow$ はい/いいえ; Japanese era years $\rightarrow$ Gregorian years) before checking inclusion. For bootstrap-based estimation, we used 200 resamples for sample-size bootstrapping and 5,000 resamples for θ estimation, as described in Appendix C.

B.2 Prompts

In this paper, we use plain-text prompts for all models, and do not employ chat templates or role specifications such as system/user/assistant. Table 2 lists the prompts used in our experiments.

B.3 Details of Sampling and Teacher Forcing

About the random seed To ensure that the generated samples remain aligned when comparing between models, tasks, and experimental conditions, we fixed the random seed of LLM for each sample based on its sample ID. This enables us to evaluate the IS and IG spectrum in a way that supports sample-wise comparisons.

Table 2: Prompt templates used in the experiments.

Task	Prompt (given in Japanese)
QA tasks (no additional information)	You are a question answering system. Please answer the following question in Japanese concisely in one sentence. Question: $question$ Answer:
Commonsense QA (multiple choice)	You are a question answering system. For the following question, choose one appropriate answer from the options. Question: $question$ Options: $choices$ Answer:
Knowledge QA (RAG)	You are a question answering system. For the following question, refer to the context shown below and answer concisely in Japanese in one sentence. Question: $question$ Context: $context$ Answer:
Hop QA (KG)	You are a question answering system. Please answer the following question in Japanese concisely in one sentence. Question: $question$ However, $deriv$ Answer:
news	You are a newspaper reporter. Based on a fictional event, write an article of approximately 300 Japanese characters.
poem	You are a poet. Write a poem of approximately 300 Japanese characters on the theme of romance.

Handling tokenizer differences Because Swallow and Qwen use different tokenizers, even with the same prompt, the resulting token sequence y is not exactly identical. However, in this paper, the prompt sequence is used only as additional information, and spectrum computation is based solely on the self-information of the output sequence. Therefore, tokenizer differences are treated as part of the model differences.

B.4 Summary of Accuracy and Distinct-2 by Model

Table 3 reports, for each model, the diversity metric (Distinct-2) for the generation tasks and the accuracy for the QA tasks (C: commonsense, K: knowledge, and H: hop). For generation tasks, Qwen3-8B shows lower Distinct-2, suggesting that its output tends to be relatively formulaic. Among the Swallow-family models, Distinct-2 decreases after instruction tuning (Swallow-8B-Instruct), implying that generation diversity may be suppressed in exchange for improved instruction-following behavior.

In contrast, in QA tasks, instruction tuning improves accuracy under conditions with additional information such as multiple-choice options and RAG/KG, indicating more stable utilization of the provided information. Qwen3-8B consistently achieves high accuracy in settings with additional information, suggesting that its reasoning with external knowledge may be stronger.

C Stability of Empirical IS Estimation

Summary: Empirical IS quantile estimates become stable with relatively few samples on the typical side, whereas tail-side statistics exhibit much larger uncertainty and require a sufficient number of sequences and an appropriate choice of the quantile parameter.

C.1 Dependence on the Sample Size

Figure 9 shows the estimated IS statistics ($\hat{H}_\theta^{\sup}$, $\hat{H}_\theta^{\inf}$, and the mean Z) as a function of the sample size S, together with bootstrap-based uncertainty (confidence intervals)[10]. The mean Z and $\hat{H}_\theta^{\inf}$ are already stable from relatively small S, and their confidence intervals shrink rapidly as S increases. In contrast, $\hat{H}_\theta^{\sup}$ corresponds to a tail-side quantile, so its estimation uncertainty is relatively large, and the confidence interval is wide, especially when S is small. However, as S increases, the confidence interval consistently narrows, indicating that $\hat{H}_\theta^{\sup}$ can also be stably estimated given a sufficient number of samples. These results suggest that estimating tail-based spectrum statistics is inherently more difficult than estimating typical-side statistics, but that reliable estimation can be obtained when the number of sequences is sufficiently large, roughly $S \geq 300$.

C.2 Confidence-Interval Analysis for the Quantile Parameter θ via Bootstrap

We used the bootstrap method to evaluate the uncertainty of the estimation. From the observed data Z, we generated $B = 5,000$ bootstrap resamples of size S by sampling with replacement and computed $\hat{H}_\theta^{\sup\,/\,\inf}$ and the width for each resample: $\hat{H}_\theta^{\sup} = Q_{1-\theta}(Z)$, $\hat{H}_\theta^{\inf} = Q_\theta(Z)$, and the width $\hat{H}_\theta^{w} = \hat{H}_\theta^{\sup} - \hat{H}_\theta^{\inf}$. From the resulting bootstrap distributions, we estimated the 95% confidence intervals using the percentile method.

[10]Confidence intervals were estimated via bootstrap ($B = 200$), and we report the 95% interval.

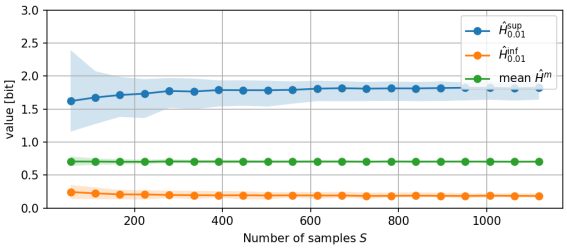


Figure 9: Estimated IS statistics and confidence intervals as a function of sample size S (C-QA, Swallow-8B). Tail-side statistics ($\hat{H}_\theta^{\sup}$) have high uncertainty for small S, but become increasingly stable as S grows. (Bootstrap $B = 200$)

To demonstrate that the results are not specific to a particular model or task, we show analyses for two models on different tasks. Figure 10 presents (a) C-QA with Swallow-8B and (b) H-QA with Qwen3-8B. In the small region θ ($\theta < 0.01$), the confidence intervals for $\hat{H}_\theta^{\sup}$ and the width become large and the estimates become unstable because they depend strongly on extreme values. In addition, for $\theta > 0.01$, the width becomes sharply smaller and exhibits strong dependence on θ. By contrast, around $\theta \sim 0.01$, the confidence intervals are relatively small and the estimation is more stable. Therefore, considering the trade-off between approximating the upper bound and estimation stability, we adopt $\theta = 0.01$ in this paper.

D Additional Experiments (Tasks and Models)

Summary: Differences due to side information and model choice are difficult to capture using only mean values, but can be systematically observed as changes in the full spectrum (distributional shape and tails).

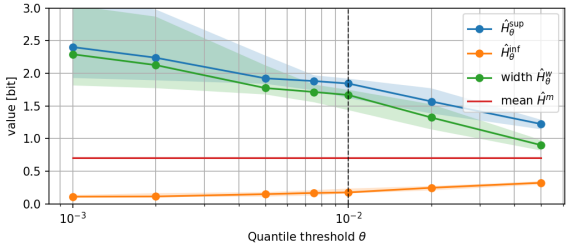
D.1 Differences across Tasks (QA)

Figure 11 shows the IS for H-QA with and without a KG. Unlike K-QA, in this setting we also observe a tendency for $\hat{H}_\theta^{\sup}$ to decrease even for incorrect sequences, indicating that side information can suppress the high-surprisal tail of incorrect sequences as well.

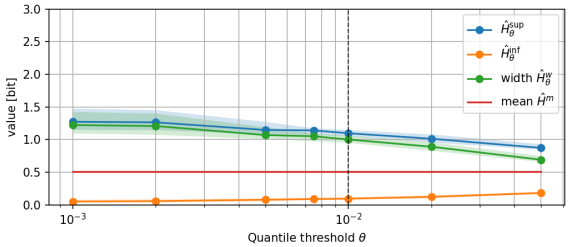
Figure 12(a) shows, for K-QA, that the IG spectrum shifts substantially to the negative side only for sequences that transition from incorrect to correct when RAG is provided. For sequences that remain incorrect, as well as those that transition from correct to incorrect, the coding-rate differ-

Table 3: Performance comparison across models ($\tau = 0.6$). Distinct-2 is used for generation tasks, and accuracy for QA tasks.

	Generation tasks		QA tasks					
	news	poem	Commonsense (C)		Knowledge (K)		Hop (H)	
	-	-	-	choices	-	RAG	-	KG
Swallow-8B	0.146	0.099	0.351	0.639	0.225	0.851	0.489	0.739
Swallow-8B-Instruct	0.131	0.069	0.394	0.979	0.205	0.904	0.480	0.797
Qwen3-8B	0.047	0.042	0.302	0.951	0.183	0.957	0.564	0.924

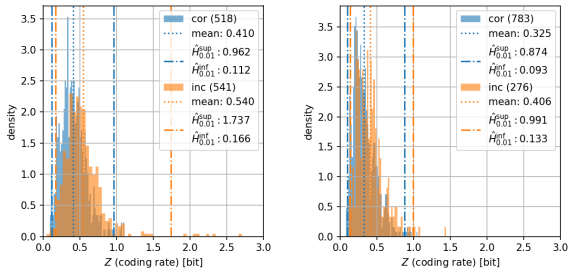


(a) C-QA (free-form), Swallow-8B, temperature $\tau = 0.6$)



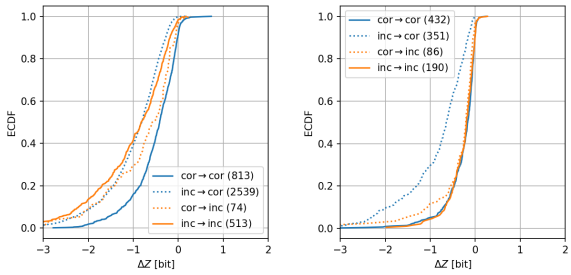
(b) H-QA (without KG), Qwen3-8B, temperature $\tau = 1.0$)

Figure 10: Bootstrap-estimated 95% confidence intervals for $\hat{H}_\theta^{\sup\,/\,\inf}$ and the width. For $\theta < 0.01$, uncertainty for the upper quantile and width increases sharply, whereas for $\theta > 0.01$ the width becomes rapidly smaller; thus we adopt $\theta = 0.01$. (Bootstrap $B = 5,000$)



(a) without KG (b) with KG

Figure 11: IS in H-QA (Swallow-8B). The plots illustrate how adding a KG changes the distributional shape and tail behavior.



(a) K-QA (with RAG) (b) H-QA (with KG)

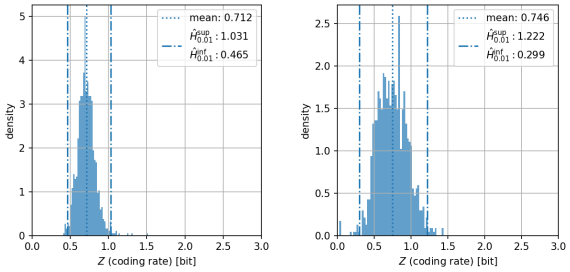
Figure 12: ECDFs of the IG spectrum for K-QA (with RAG) and H-QA (with KG) (Swallow-8B).

ences concentrate around zero, suggesting that RAG is not effectively used or may even act as a misleading signal. This indicates that RAG is not uniformly helpful across all sequences; rather, correctness transitions occur only when the side information actually improves the typicality of the sequence.

Figure 12(b) shows the ECDF of the IG spectrum for H-QA. Only the sequences that transition from incorrect to correct (*fix*) due to KG injection clearly shift to the negative side in the IG spectrum. In contrast, sequences that remain incorrect (*none*) or transition from correct to incorrect (*break*) have coding-rate differences concentrated near zero, implying that the KG has only a limited effect on the spectrum structure for those cases. This demonstrates that even in multi-hop reasoning, the effectiveness of side information varies substantially across sequences: only when conditioning works effectively does typicality improve and the output transition to a correct answer.

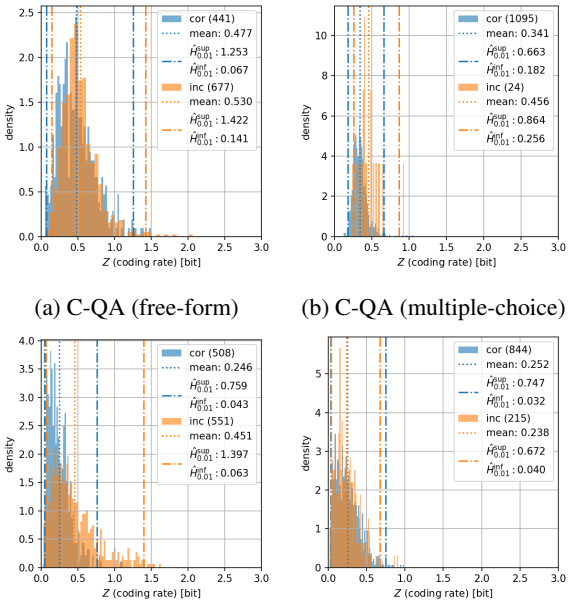
D.2 Differences across Tasks (Generation)

Figure 13 shows the IS for the generation tasks (news / poem). Although the differences in the mean coding rate (mean surprisal) are not large, the distributional shapes differ substantially between tasks, confirming the importance of evaluating the full distribution. For news, the low-surprisal side ($\hat{H}_\theta^{\inf}$) is relatively high and the outputs are strongly typical, while poem has a relatively higher high-surprisal side ($\hat{H}_\theta^{\sup}$), indicating a greater variation in generation. These trends are consistent with the discussion in the main text.



(a) news (b) poem

Figure 13: IS for generation tasks (Swallow-8B). The distributional shapes differ substantially between news and poem.



(a) C-QA (free-form) (b) C-QA (multiple-choice)

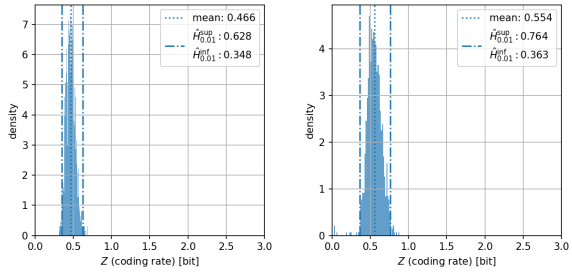
(c) H-QA (without KG) (d) H-QA (with KG)

Figure 14: IS for QA tasks (Swallow-8B-Instruct).

D.3 Model Comparison (Swallow-8B-Instruct / Qwen3-8B)

Figures 14 and 15 show the IS for Swallow-8B-Instruct. After instruction tuning, the IS width shrinks in many tasks, and this tendency is particularly pronounced when side information, such as choices, RAG, or a KG, is provided. A similar effect is also observed in generation tasks: the gap in $\hat{H}^{\sup}_{\theta}$ between news and poem becomes smaller, indicating that the variability of the output distribution is suppressed. These results suggest that instruction tuning stabilizes generation under side information and reduces the high-surprisal tail.

Next, Figures 16 and 17 show the IS for Qwen3-8B. Compared with the Swallow-family models, Qwen3-8B exhibits smaller IS widths, and this tendency becomes especially pronounced when side information (choices, RAG, KG) is provided.



(a) news (b) poem

Figure 15: IS for generation tasks (Swallow-8B-Instruct).

In addition, for C-QA, the distributions for correct and incorrect outputs are close to each other, suggesting that the model assigns relatively high likelihood even to incorrect outputs (i.e., it may be prone to overconfidence under errors).

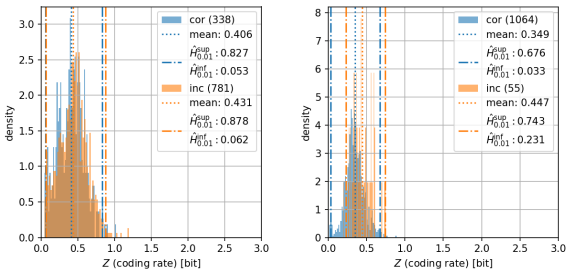
Summarizing the model differences together with the performance results in Table 3, Swallow-8B achieves higher diversity (Distinct-2) in generation tasks, but also shows relatively larger IS widths and larger high-surprisal quantiles ($\hat{H}^{\sup}_{\theta}$), indicating heavier tails in the generation distribution. In contrast, Qwen3-8B achieves high accuracy in QA tasks under side information, and its smaller IS width suggests a stronger concentration in the typical set. Swallow-8B-Instruct exhibits an intermediate behavior between the two: tail suppression (width reduction) due to instruction tuning corresponds to stabilized QA performance, while diversity decreases in generation tasks.

E Information Spectrum and Performance Metrics

Summary: In QA, IS statistics tend to be negatively correlated with performance, whereas in generation they tend to be positively correlated. This suggests that spectrum-based quantities may serve as task-agnostic diagnostic signals, with interpretation depending on task characteristics.

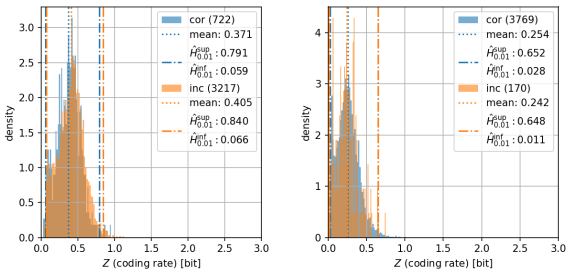
E.1 Correlation Analysis

Table 4 reports the Pearson/Spearman correlation coefficients (and their p-values) between IS statistics and performance metrics (QA: accuracy; generation: Distinct-2). For each combination of temperature setting (five levels) and task condition (QA: three tasks $\times$ with/without side information; generation: two tasks), we obtained a set of generated sequences and computed both spectrum statistics and performance metrics.



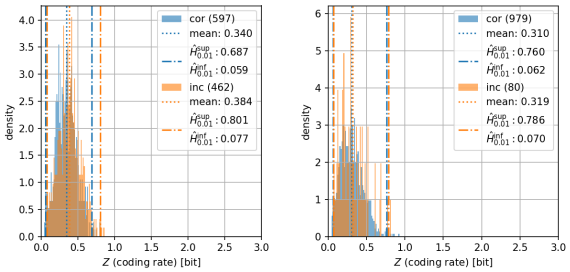
(a) C-QA (free-form)

(b) C-QA (multiple-choice)



(c) K-QA (without RAG)

(d) K-QA (with RAG)

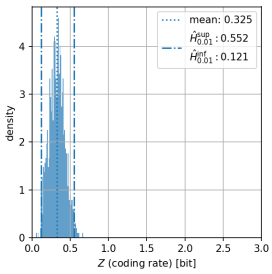


(e) H-QA (without KG)

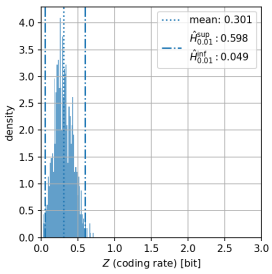
(f) H-QA (with KG)

Figure 16: IS for QA tasks (Qwen3-8B).

In QA tasks, many models show negative correlations between spectrum statistics and accuracy. In particular, for Swallow-8B (base), $\hat{H}^{\inf}_{\theta}$ exhibits a strong negative correlation. Since $\hat{H}^{\inf}_{\theta}$ corresponds to a low-surprisal quantile, a larger value indicates that the coding rate on the typical side increases (i.e., the typical set shifts toward a higher surprisal). Although the separation between correct and incorrect distributions is more pronounced in the high-surprisal tail, correlations with the performance metric may sometimes be more strongly tied to the typical side ($\hat{H}^{\inf}_{\theta}$). This implies that, in QA, higher values on the low-surprisal (typical) side tend to be associated with lower accuracy, suggesting that QA performance may be strongly constrained by behavior around the typical set. In contrast, for Swallow-8B-Instruct, the correlation with $\hat{H}^{\inf}_{\theta}$ disappears, suggesting that the spectrum region that governs performance can shift from the typical side to the tail side (e.g., $\hat{H}^{\sup}_{\theta}$ and the width) after the instruction tuning.



(a) news



(b) poem

Figure 17: IS for generation tasks (Qwen3-8B).

In generation tasks, both $\hat{H}^{\sup}_{\theta}$ and $\hat{H}^{\inf}_{\theta}$ show positive correlations with Distinct-2, and $\hat{H}^{\inf}_{\theta}$ in particular exhibits a strong positive correlation. That is, when the coding rate on the typical side is higher (i.e., the output distribution does not concentrate strongly onto the typical set), output diversity tends to increase. However, note that the generation setting contains only 10 data points, resulting in limited statistical power; therefore, we interpret these results primarily in terms of sign and magnitude.

Overall, these results indicate that spectrum statistics are linked to performance in different ways for QA and generation, and may be usable as diagnostic signals tailored to task properties. For example, in generation tasks, relatively higher temperatures can be advantageous for maintaining diversity, whereas in QA tasks, activating atypical sequences (the tail) directly leads to errors, suggesting that lower temperatures are preferable.

E.2 Stability Analysis of Leave-One-Out Regression

Table 5 reports, for a simple regression using the six points (three without side information and three with side information), the range of the estimated slope (min, max) and the range of the coefficient of determination R^2 on the training points when we re-estimate the regression while leaving out one point at a time. For all models, the slope for $\hat{H}^{\sup}_{0.01}$ and the width is always negative, confirming that the observed relationship does not depend on a single outlier. In contrast, for Swallow-8B-Instruct, $\hat{H}^{\inf}_{0.01}$ exhibits an unstable slope sign when the temperature τ is fixed at 0.6, and while the slope becomes consistent when combining all temperatures, the resulting R^2 is close to zero. Therefore, for instruction-tuned models, the typical-side statistic $\hat{H}^{\inf}_{0.01}$ is not a robust explanatory signal for performance.

Table 4: Correlations between IS statistics and performance metrics (QA ($N = 30$): accuracy; generation ($N = 10$): Distinct-2). Each data point corresponds to a combination of temperature setting (five levels) and task condition (QA: $3\times$ with/without side information; generation: two tasks).

Target	Statistic	Pearson		Spearman	
		r	p	ρ	p
Swallow-8B					
QA (Accuracy)	$\hat{H}^{w}_{0.01}$	-0.442	1.45e-02	-0.485	6.57e-03
	$\hat{H}^{\mathrm{sup}}_{0.01}$	-0.491	5.92e-03	-0.519	3.33e-03
	$\hat{H}^{\mathrm{inf}}_{0.01}$	-0.751	1.78e-06	-0.863	8.33e-10
Generation (Distinct-2)	$\hat{H}^{w}_{0.01}$	0.405	2.46e-01	0.455	1.87e-01
	$\hat{H}^{\mathrm{sup}}_{0.01}$	0.563	8.99e-02	0.673	3.30e-02
	$\hat{H}^{\mathrm{inf}}_{0.01}$	0.835	2.65e-03	0.915	2.04e-04
Swallow-8B-Instruct					
QA (Accuracy)	$\hat{H}^{w}_{0.01}$	-0.823	2.34e-08	-0.892	3.73e-11
	$\hat{H}^{\mathrm{sup}}_{0.01}$	-0.806	7.64e-08	-0.874	2.94e-10
	$\hat{H}^{\mathrm{inf}}_{0.01}$	0.227	2.28e-01	-0.105	5.80e-01
Generation (Distinct-2)	$\hat{H}^{w}_{0.01}$	0.058	8.72e-01	0.164	6.51e-01
	$\hat{H}^{\mathrm{sup}}_{0.01}$	0.248	4.89e-01	0.248	4.89e-01
	$\hat{H}^{\mathrm{inf}}_{0.01}$	0.635	4.83e-02	0.571	8.44e-02
Qwen3-8B					
QA (Accuracy)	$\hat{H}^{w}_{0.01}$	-0.361	4.98e-02	-0.436	1.60e-02
	$\hat{H}^{\mathrm{sup}}_{0.01}$	-0.361	4.98e-02	-0.416	2.22e-02
	$\hat{H}^{\mathrm{inf}}_{0.01}$	-0.244	1.94e-01	-0.429	1.80e-02
Generation (Distinct-2)	$\hat{H}^{w}_{0.01}$	0.739	1.47e-02	0.830	2.94e-03
	$\hat{H}^{\mathrm{sup}}_{0.01}$	0.845	2.07e-03	0.867	1.17e-03
	$\hat{H}^{\mathrm{inf}}_{0.01}$	0.950	2.56e-05	0.879	8.14e-04

Table 5: Robustness of the relationship between spectrum statistics and accuracy under leave-one-out (LOO) regression (ranges of slope and R^2) (temperature fixed at 0.6, $N = 6$).

Model	Statistic	Slope range	R^2 range
Swallow-8B	$\hat{H}^{w}_{0.01}$	[-0.722, -0.411]	[0.395, 0.699]
	$\hat{H}^{\mathrm{sup}}_{0.01}$	[-0.677, -0.424]	[0.507, 0.756]
	$\hat{H}^{\mathrm{inf}}_{0.01}$	[-4.899, -3.417]	[0.707, 0.884]
Swallow-8B Instruct	$\hat{H}^{w}_{0.01}$	[-0.759, -0.701]	[0.971, 0.989]
	$\hat{H}^{\mathrm{sup}}_{0.01}$	[-0.797, -0.718]	[0.965, 0.996]
	$\hat{H}^{\mathrm{inf}}_{0.01}$	[-8.729, 2.477]	[0.026, 0.693]
Qwen3-8B	$\hat{H}^{w}_{0.01}$	[-5.106, -3.682]	[0.693, 0.865]
	$\hat{H}^{\mathrm{sup}}_{0.01}$	[-4.273, -3.109]	[0.696, 0.925]
	$\hat{H}^{\mathrm{inf}}_{0.01}$	[-17.316, -10.561]	[0.343, 0.791]

Table 6: Robustness of the relationship between spectrum statistics and accuracy under LOO regression (ranges of slope and R^2) (all temperatures, $N = 30$).

Model	Statistic	Slope range	R^2 range
Swallow-8B	$\hat{H}^{w}_{0.01}$	[-0.151, -0.110]	[0.142, 0.260]
	$\hat{H}^{\mathrm{sup}}_{0.01}$	[-0.153, -0.119]	[0.185, 0.305]
	$\hat{H}^{\mathrm{inf}}_{0.01}$	[-2.891, -1.988]	[0.526, 0.676]
Swallow-8B Instruct	$\hat{H}^{w}_{0.01}$	[-0.422, -0.373]	[0.654, 0.718]
	$\hat{H}^{\mathrm{sup}}_{0.01}$	[-0.420, -0.368]	[0.624, 0.689]
	$\hat{H}^{\mathrm{inf}}_{0.01}$	[0.618, 1.114]	[0.014, 0.070]
Qwen3-8B	$\hat{H}^{w}_{0.01}$	[-0.551, -0.392]	[0.081, 0.178]
	$\hat{H}^{\mathrm{sup}}_{0.01}$	[-0.509, -0.359]	[0.081, 0.177]
	$\hat{H}^{\mathrm{inf}}_{0.01}$	[-4.418, -1.843]	[0.037, 0.159]

Table 6 shows the corresponding results aggregated over all temperatures. In particular, for Swallow-8B-Instruct, both the width and $\hat{H}^{\mathrm{sup}}_{0.01}$ maintain $R^2 \approx 0.65$ even at all temperatures, confirming a strong association with the performance metric.

F Effects of Temperature

Summary: Increasing the temperature activates the tail, enlarges the IS width, and provides a unified explanation of QA performance degradation and increased diversity in generation.

F.1 Temperature and the Information Spectrum (Model Differences)

In the following, we present IS at temperature $\tau = 1.0$ for each model. The spectrum of Swallow-8B is shown in Fig. 18(a)(b). Compared with the low-temperature case in Fig. 2, temperature affects not only the mean coding rate but also the distributional shape, in particular the IS width and the thickness of the tail. Under low temperature, correct sequences concentrate around the typical set, while incorrect sequences are more likely to appear in the high-surprisal tail. As a result, the separation between cor/inc by $\hat{H}^{\mathrm{sup}}_{\theta}$ is relatively clear. In contrast, under high temperature, the tail is activated for both correct and incorrect sequences, and the increased overlap of the distributions weakens the spectral separation between cor and inc. This trend is especially pronounced in the free-form setting. Although side information such as multiple-choice options partially suppresses tail expansion, it cannot fully counteract the increase of atypical sequences induced by high temperature. These observations suggest that the degradation of QA performance cannot be explained solely by changes in average uncertainty but may instead be driven by the activation of atypical sequences (the tail).

The spectrum of Swallow-8B-Instruct is shown in Fig. 18(c)(d). Instruction tuning does not eliminate tail activation itself as temperature increases, but it substantially reshapes the baseline spectrum at low temperature. At low temperature, the instruction-tuned model exhibits a smaller IS width, with both correct and incorrect sequences strongly concentrated near the typical set, leaving

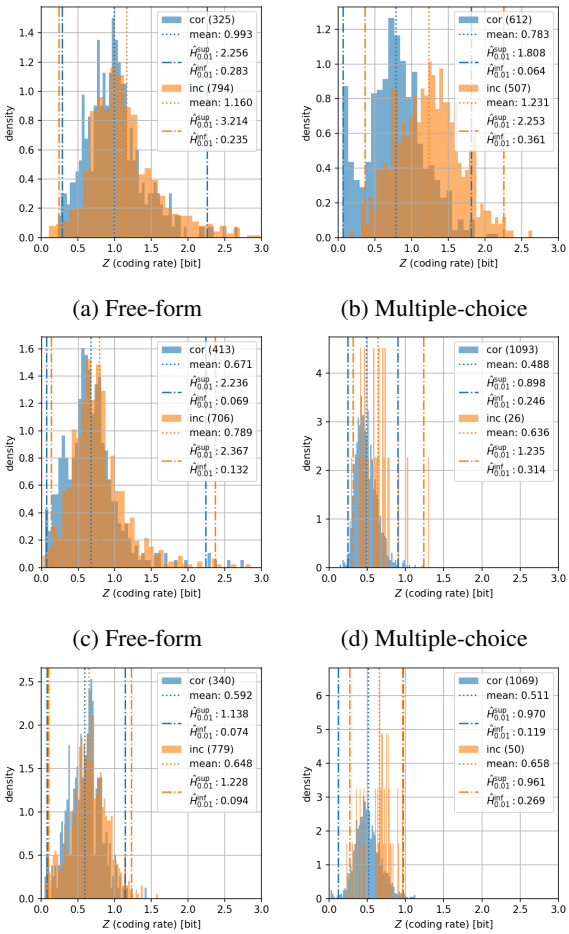
only limited tail mass. Consequently, the regime in which errors are dominated only by extreme surprisal events (tail) becomes weaker, and simple sequence-level diagnostics based on tail statistics (e.g., $\hat{H}_\theta^{\mathrm{sup}}$) may be less discriminative than for the base model. At high temperature, however, tail activation re-emerges even after instruction tuning, and the overlap between cor/inc again increases, although the degree of expansion tends to be smaller than in the base model. Overall, these results suggest that instruction tuning suppresses the tail under typical decoding conditions, while temperature continues to control the activation of atypical sequences.

The spectrum of Qwen3-8B is shown in Fig. 18(e)(f). For Qwen3, the IS width also increases with temperature, but the overlap between cor/inc distributions is relatively large, and incorrect sequences may still receive a comparatively high likelihood. As a result, tail-based separation is often less pronounced than in the Swallow family.

F.2 Tasks, Temperature, and Spectrum Statistics

Figure 19 shows, for multiple tasks, the relationship between temperature τ and spectrum statistics (width $\hat{H}_\theta^w$ and lower quantile $\hat{H}_\theta^{\mathrm{inf}}$) for each model. See also the Swallow-8B results in Fig. 3. For Swallow-8B-Instruct, as in Swallow-8B (base), $\hat{H}_\theta^w$ increases with temperature, but in the low-temperature regime the width is suppressed compared to the base model. That is, instruction tuning strengthens concentration around the typical set, particularly under low temperature, thus suppressing tail generation. However, at high temperature, H_ϵ^w increases sharply again even for the instruction-tuned model, and the basic property that higher temperature induces more atypical sequences remains. Hence, instruction tuning does not "remove" temperature dependence; rather, it compresses the spectrum under low temperature and reorganizes the distributional shape toward greater stability. Moreover, $\hat{H}_\theta^{\mathrm{inf}}$ shows a relatively small variation depending on the task, indicating that the dominant effect of temperature is expressed again in the width (i.e., tail activation).

For Qwen3-8B, we also observed an increase in H_ϵ^w with temperature, but the behavior exhibits distinct characteristics compared to the Swallow family. In particular, in QA tasks the width is already relatively large at low temperature, and the slope of increase with temperature is steep. This suggests that Qwen3-8B is highly sensitive to decoding temperature and that raising the temperature rapidly increases deviations from the typical set. In addition, the variation of $\hat{H}_\theta^{\mathrm{inf}}$ is greater than for the Swallow family, implying that even the typical-set side (low surprisal region) may fluctuate substantially with temperature. This aligns with the observation that correct/incorrect distributions tend to overlap (i.e., incorrect sequences can still receive relatively high likelihood), so tail separation does not necessarily become clear.



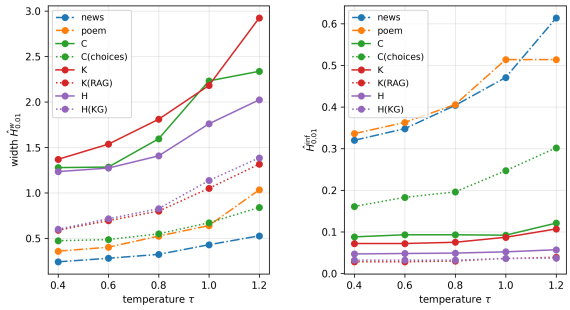
(a) Free-form (b) Multiple-choice

(c) Free-form (d) Multiple-choice

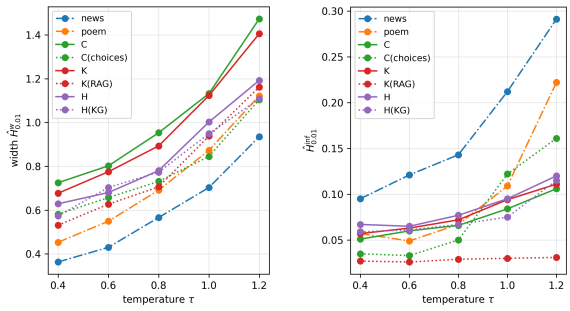
(e) Free-form (f) Multiple-choice

Figure 18: IS for the Commonsense QA task at $\tau = 1.0$ (Swallow-8B (a)(b) / Swallow-8B-Instruct (c)(d) / Qwen3-8B (e)(f)).

Figure 20 shows the relationship between temperature and $\hat{H}_\theta^{\mathrm{sup}}$. As the temperature increases, $\hat{H}_\theta^{\mathrm{sup}}$ grows monotonically for all tasks, reflecting the behavior of the width. This suggests that the increase in IS width is primarily driven by activa-



(a) Width $\hat{H}_\theta^w$ (b) $\hat{H}_\theta^{\inf}$



(c) Width $\hat{H}_\theta^w$ (d) $\hat{H}_\theta^{\inf}$

Figure 19: Relationships between temperature and IS width/lower bound across multiple tasks (Swallow-8B-Instruct (a)(b) / Qwen3-8B (c)(d)).

tion of the high-surprisal tail[11], i.e., an increase of $\hat{H}_\theta^{\sup}$.

These results show that temperature acts more strongly on the IS width (tail activation) than on the mean coding rate, and it becomes a major factor corresponding to QA performance degradation. Instruction tuning compresses and stabilizes the spectrum under low-temperature conditions, but tail activation remains dominant at high temperature and cannot be fully suppressed. Qwen3-8B exhibits high temperature sensitivity even on the typical-set side, and its distributional changes differ from those of the Swallow family.

Temperature and Accuracy Figure 21 (left axis) shows the relationship between temperature and QA accuracy. From this, the extent of performance degradation with respect to temperature differs between models. For Swallow-8B, the accuracy drop with increasing temperature is pronounced, and in particular under the no-side-information setting, performance decreases monotonically. In contrast, Swallow-8B-Instruct maintains a high accuracy across the entire temperature range, and degradation with increasing temperature is relatively mild. This suggests that instruction tuning makes the output distribution concentrate more strongly on the typical set, rendering it less sensitive to temperature-induced diffusion of the probability distribution (tail activation). For Qwen3-8B, a decrease in accuracy is also observed with increasing temperature, but the trend is weaker depending on the task and its temperature sensitivity can be lower than that of the Swallow family.

[11] $\hat{H}_\theta^{\sup}$ exhibits stronger temperature dependence than $\hat{H}_\theta^{\inf}$, implying that the temperature dependence of the width is dominated by the tail-side statistic.

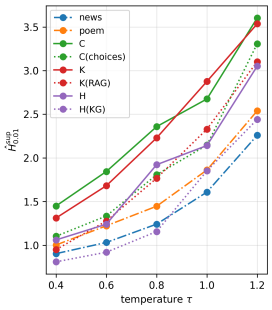
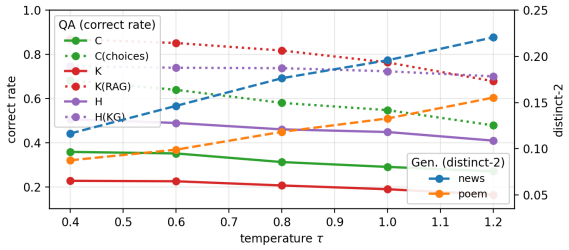


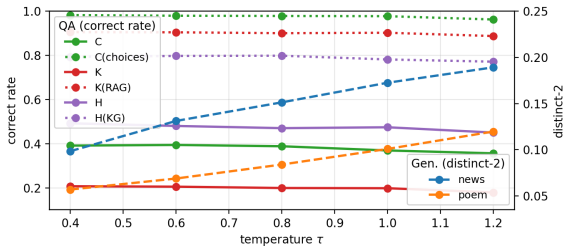
Figure 20: Relationship between temperature and $\hat{H}_\theta^{\sup}$ across multiple tasks (Swallow-8B).

Trade-off Between Width and Diversity in Generation Tasks Figure 21 (right axis) shows the relationship between temperature and Distinct-2 (the number of distinct 2-grams in generated sequences divided by the total number of 2-grams), which is used as a diversity measure in generation tasks(Li et al., 2016). As temperature increases, the IS width grows and Distinct-2 tends to increase. That is, under low-temperature decoding, the output distribution concentrates strongly on the typical set, reducing expressive diversity and making stereotyped text more likely to be generated. This occurs because temperature increases the probability of sequence generation outside the typical set, thereby expanding variability in lexical choice and expression. Therefore, in generation tasks, an increase in width yields a positive effect in the form of greater diversity, whereas in QA tasks, the same increase in width manifests itself as a rise in incorrect answers. Overall, the IS width provides a unified diagnostic quantity that explains temperature-induced performance changes, and it offers a clue for task-dependent optimization strategies such as "low temperature + suppressed width for QA" and "moderately high temperature + sufficient width for generation."

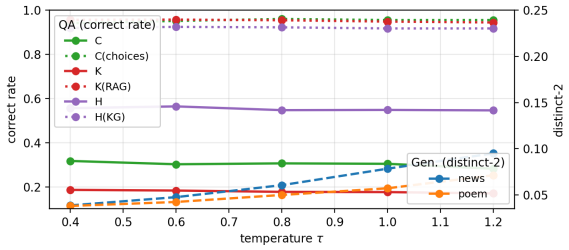
Interpretation: Linking Width to Performance Increasing the temperature τ affects the IS width



(a) Swallow-8B



(b) Swallow-8B-Instruct



(c) Qwen3-8B

Figure 21: Relationship between temperature τ and QA accuracy (left axis), and between τ and Distinct-2 in generation tasks (right axis). As temperature increases, accuracy tends to decrease, side information lifts performance, and diversity (Distinct-2) increases.

$\hat{H}_\theta^w$ (tail activation) more strongly than the mean coding rate, thus increasing the variability of the cross-sequence. In QA, an increase in the tail not only makes incorrect sequences more likely to appear, but also causes correct sequences to be dragged into the tail, increasing the overlap between correct and incorrect distributions and thus reducing accuracy. In contrast, in generation tasks, tail growth increases expressive fluctuation, which appears as higher diversity measured by Distinct-2. Hence, temperature does not simply "change the mean coding rate"; rather, it influences performance by controlling the IS width—the proportion of atypical sequences—and this structural change is reflected in different task metrics (accuracy vs. diversity) for QA and generation. Therefore, $\hat{H}_\theta^w$ serves as a task-agnostic diagnostic quantity that captures temperature-induced structural changes in the output distribution.

G Auxiliary Evaluation

Summary: Z is useful as a simple diagnostic signal for correctness based only on sequence likelihood, while ΔZ is a specialized metric for detecting intervention effects (error correction vs. misguidance).

G.1 Area under the curve (AUC) Tables (Correct vs. Incorrect Classification)

Although the main focus of this paper is the analysis of the IS-based distributional structure, we also evaluate the sequence-level diagnostic performance of self-information Z and the quantity of the IG-spectrum ΔZ using AUC. Table 7 compares the AUC of the coding-rate difference ΔZ derived from the IG spectrum with those of existing uncertainty metrics[12] for distinguishing correct vs. incorrect outputs[13].

Table 7 reports the same evaluation for Qwen3-8B. For Qwen3-8B, the overall correctness discrimination performance is lower than Swallow-8B, and some settings are close to the chance level (0.5). In some conditions, accuracy is extremely high ($> 95\%$), so the number of error samples is small and the AUC estimate can become unstable[14]. In addition, the model may assign relatively high likelihood even to incorrect sequences, which is consistent with the IS observation that the correct/incorrect spectral separation is weak.

These results confirm that Z and ΔZ achieve discrimination performance comparable to or better than existing metrics in many QA settings and can serve as effective sequence-level diagnostic signals. However, ΔZ measures "how much the intervention of side information improved the likelihood of the given sequence," and is not a metric designed to directly classify the static attribute of correctness. For example, even for correct sequences (cor→cor), side information may not change the likelihood, and conversely, even for in-

[12] Since entropy- and margin-based quantities form distributions within a sequence, multiple summary statistics were computed, such as mean/max/var for entropy and mean/min for margin.

[13] Note that entropy, margin, and top-K (with $K = 5$) mass require the vocabulary distribution (logits) at each token step, whereas Z and ΔZ in this paper can be computed solely from token-level negative log-likelihoods (teacher-forced log probabilities) on the generated sequence. That is, our metrics remain applicable even when only sequence scores are available, and thus their prerequisites differ from logits-dependent metrics.

[14] For space reasons, we report point estimates only, but confidence intervals can be estimated via bootstrap.

Table 7: Area under the curve (AUC) for correctness classification across QA settings (three models).

		C		K		H	
		-	choice	-	RAG	-	KG
Swallow-8B							
Z		0.617	0.843	0.690	**0.596**	0.669	**0.643**
ΔZ		-	0.892	-	0.524	-	0.611
entropy	(mean)	0.619	0.867	0.698	0.594	0.685	0.622
	(max)	0.600	0.752	0.741	0.511	0.652	0.532
	(var)	0.608	0.883	**0.762**	0.567	**0.711**	0.592
margin	(mean)	0.581	0.763	0.584	0.582	0.600	0.560
	(min)	0.510	0.700	0.569	0.528	0.527	0.548
top-1 prob (mean)		0.605	0.812	0.645	0.595	0.649	0.638
top-K mass (mean)		**0.623**	**0.896**	0.750	0.586	0.710	0.605
Swallow-8B-Instruct							
Z		0.566	0.778	0.716	**0.588**	0.718	0.479
ΔZ		-	0.679	-	0.461	-	**0.622**
entropy	(mean)	0.538	0.760	0.725	0.580	**0.738**	0.481
	(max)	**0.658**	0.733	0.781	0.546	0.682	0.470
	(var)	0.582	**0.798**	**0.783**	0.567	0.725	0.470
margin	(mean)	0.536	0.696	0.605	0.579	0.694	0.493
	(min)	0.616	0.525	0.626	0.533	0.575	0.496
top-1 prob (mean)		0.547	0.745	0.686	0.586	0.722	0.485
top-K mass (mean)		0.541	0.782	0.777	0.558	0.731	0.447
Qwen3-8B							
Z		0.536	0.715	0.557	0.468	0.584	0.501
ΔZ		-	0.653	-	0.485	-	0.552
entropy	(mean)	0.546	0.706	0.571	0.475	0.610	0.518
	(max)	**0.558**	0.406	0.560	0.449	0.570	0.464
	(var)	**0.558**	0.553	0.590	0.446	0.589	0.499
margin	(mean)	0.528	0.719	0.548	**0.524**	**0.622**	**0.562**
	(min)	0.508	0.523	0.508	0.483	0.527	0.483
top-1 prob (mean)		0.537	**0.720**	0.555	0.480	0.599	0.517
top-K mass (mean)		0.555	0.633	**0.593**	0.453	0.606	0.507

Table 8: Conditional AUC for predicting correctness transitions (three models).

		confirm/break			fix/none		
		C	K	H	C	K	H
Swallow-8B							
Z		0.835	0.583	0.639	0.847	**0.605**	0.644
ΔZ		0.846	0.409	0.438	**0.915**	0.512	**0.798**
entropy	(mean)	0.860	0.593	0.637	0.871	0.600	0.602
	(max)	0.768	0.509	0.580	0.745	0.505	0.472
	(var)	0.876	0.578	**0.660**	0.886	0.568	0.526
margin	(mean)	0.743	0.552	0.539	0.773	0.593	0.557
	(min)	0.682	0.504	0.514	0.710	0.530	0.561
top-1 prob (mean)		0.803	0.586	0.639	0.818	0.602	0.639
top-K mass (mean)		**0.887**	**0.594**	0.653	0.901	0.588	0.558
Swallow-8B-Instruct							
Z		0.774	**0.685**	0.518	0.777	0.581	0.421
ΔZ		0.723	0.436	0.493	0.678	0.489	**0.787**
entropy	(mean)	0.613	0.646	0.520	0.796	0.578	0.417
	(max)	**0.887**	0.557	0.438	0.675	0.547	0.467
	(var)	0.759	0.612	0.485	**0.801**	0.565	0.429
margin	(mean)	0.467	0.609	0.486	0.754	0.581	0.459
	(min)	0.546	0.545	0.497	0.511	0.532	0.483
top-1 prob (mean)		0.595	0.675	**0.531**	0.780	**0.582**	0.417
top-K mass (mean)		0.719	0.587	0.466	0.797	0.558	0.402
Qwen3-8B							
Z		**0.792**	0.429	0.442	0.689	0.470	0.503
ΔZ		0.580	**0.734**	**0.538**	0.676	0.490	**0.620**
entropy	(mean)	0.678	0.404	0.457	0.697	0.479	0.518
	(max)	0.402	0.445	0.400	0.402	0.443	0.490
	(var)	0.601	0.391	0.428	0.530	0.448	0.513
margin	(mean)	0.612	0.557	0.523	**0.725**	**0.525**	0.549
	(min)	0.603	0.575	0.455	0.509	0.479	0.483
top-1 prob (mean)		0.697	0.411	0.471	0.709	0.484	0.512
top-K mass (mean)		0.661	0.384	0.435	0.613	0.456	0.516

correct sequences (inc→inc), likelihood improvement ($\Delta Z < 0$) can occur. Thus, ΔZ is not necessarily optimal for the overall correct-vs. incorrect classification. In the next section G.2, we show that it is better suited for identifying sequences in which the intervention contributed to "error correction" or "confirmation." This is also consistent with the IS observation that correct/incorrect spectral separation can be weak.

G.2 Conditional AUC (Analysis by Correctness Transitions)

In this section, we categorize correctness transitions induced by adding side information into four types: confirm (cor→cor), break (cor→inc), fix (inc→cor), and none (inc→inc), and evaluate whether ΔZ can distinguish "error correction" (fix) and "misguidance" (break). Table 8 reports the conditional AUC results for Swallow-8B. As discussed above, ΔZ is not necessarily optimal for the discrimination of correctness (correct vs. incorrect), but it is useful for analyzing the effects of the intervention. Specifically, under the conditional evaluation restricted to sequences that are incorrect without side information (cond_incorrect), ΔZ can accurately distinguish whether the sequence transitions to correct under RAG (wrong→correct vs. wrong→wrong) (Figure 12(b)), and it achieves the highest AUC particularly in the Hop QA setting. However, in conditional evaluation restricted to sequences that are correct without side information, the discriminative performance of ΔZ decreases, suggesting that ΔZ is better suited for identifying error corrections than for judging correctness maintenance.

In K-QA, the conditional ECDF of ΔZ shows that, among sequences that are initially incorrect, the distributions for those that become correct with RAG (inc→cor) and those that remain incorrect (inc→inc) largely overlap (Figure 12(a)), indicating that ΔZ cannot selectively capture error correction in this task. Similar results for other models are shown in the same table.

G.3 Interpretation

The AUC results suggest the following:

1. Z achieves discrimination performance comparable to existing uncertainty metrics in many settings and is useful as a simple diagnostic signal obtainable solely from the likelihood of the sequence.

2. ΔZ is not necessarily suitable for the general classification of correctness, but it captures sequences whose likelihood is actually improved by side information and is particularly effective in identifying *fix* (wrong$\rightarrow$correct) transitions.

3. Therefore, ΔZ should not be interpreted as a "correctness classifier," but as a metric for intervention-effect analysis (e.g., diagnosing side-information design or RAG/KG behavior).