情報スペクトル理論に基づく大規模言語モデルの生成挙動解析

太刀岡 勇気
デンソーアイティーラボラトリ
tachioka.yuki@core.d-itlab.co.jp

概要

本稿では，大規模言語モデル (LLM) を非定常な一般情報源として捉え，情報スペクトル理論に基づく生成挙動解析を提案する．生成系列ごとに系列レベル符号化レートを定義し，その分布を経験的情報スペクトルとして評価することで，正誤差が平均的指標では捉えにくく，高 surprisal 側の tail に顕著に現れることを示す．さらに生成系列を固定した教師強制に基づく情報利得スペクトルを導入し，追加情報 (択一選択肢・RAG・知識グラフ) の効果が系列ごとに不均一であり，有効に作用する系列を識別できることを確認した．また温度や条件付与が生成分布の尾部と性能へ与える影響を，分布指標と性能指標の両面から統一的に整理する枠組みを与える．

1 はじめに

大規模言語モデル (LLM) はプロンプトや文脈に応じて生成分布が大きく変化し，推論時条件の変動を伴う非定常な情報源として振る舞う．このような状況では，対数尤度やクロスエントロピーといった平均的指標のみでは，生成挙動のばらつきや失敗系列を十分に捉えられないことがある．実際，正誤や追加情報 (選択肢提示，RAG，知識グラフ (KG) など) の効果は平均的には小さく見えても，系列ごとの挙動，特に高 surprisal 側の tail では大きく異なり得る．

そこで本稿では，LLM を一般情報源として捉え，系列レベルの自己情報量密度 (符号化レート) の分布に基づいて生成挙動を診断する枠組みを提案する．具体的には，有限サンプルから情報スペクトルを推定する経験的情報スペクトルを導入し，平均ではなく tail に現れる差異を可視化する．さらに，生成系列を固定した教師強制により追加情報あり/なしの尤度差を測る情報利得スペクトル (IG スペクトル) を導入し，追加情報が「どの系列に」有効であったかを系列レベルで識別する．

本稿の貢献は以下の通りである：(1) 系列レベル符号化レートに基づく経験的情報スペクトルと分位点統計量を定義する．(2) 追加情報効果を系列対応を保った介入応答として定量化する IG スペクトルを提案する．(3) 複数モデル・複数タスクにおいて，正誤差が平均ではなく tail に現れること，追加情報がスペクトル形状を体系的に変化させ，効果が系列ごとに不均一であること，および指示チューニングにより性能を規定する要因が平均から分布幅/tail へシフトし得ることを示す．

関連研究　LLM の性能評価では，負対数尤度やクロスエントロピー，perplexity といった平均的指標が広く用いられてきた [1]．一方で，LLM の生成過程ではプロンプト依存性や履歴依存性が強く，平均指標のみでは系列間のばらつきを捉えにくいことが指摘されている [2, 3, 4]．またトークン単位の surprisal を用いて生成文の典型性や逸脱を解析する研究 [5] や，温度・top-p などのサンプリングが生成分布の tail を活性化しうることを示す研究 [6] も報告されている．推論タスクでは自己整合性やサンプリング戦略が性能に影響することが知られ [7, 8]，追加情報（選択肢提示，検索結果，外部知識）を用いた条件付与により生成挙動を制御する枠組みも多く提案されている [9, 10]．さらに，LLM を圧縮器として捉え，符号化率 (負対数尤度) に基づき学習やアライメントによる分布変化を解析する試みもある [11]．しかしこれらの多くは平均的なエントロピー低下や尤度改善に基づく議論が中心であり，追加情報がどの系列に効いたか，あるいは正誤差が分布のどの領域に現れるかを分布構造として同定する枠組みは十分に整理されていない．

2 情報スペクトルによる LLM の解析

一般情報源に対し，系列長 n のシンボル列 X^n の自己情報量密度 (系列レベル符号化レート)

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

の分布を情報スペクトルと呼ぶ [12, 13].

2.1 情報スペクトルの確率的上・下極限

情報スペクトル理論では，この分布の tail を特徴づける量として，ϵ を許容誤差確率のパラメータとしたときの確率的上極限 $H_\epsilon^{\sup} := \inf\{\alpha \mid \limsup_{n\to\infty} \Pr[Z_n > \alpha] \le \epsilon\}$ および確率的下極限 $H_\epsilon^{\inf} := \sup\{\beta \mid \liminf_{n\to\infty} \Pr[Z_n < \beta] \le \epsilon\}$ を用いて情報源の性質を特徴づける．直感的には，$H_\epsilon^{\sup}$ は高 surprisal 側 (tail) の符号化レート，$H_\epsilon^{\inf}$ は典型的系列の符号化レートに対応し，その幅 $H_w := H_\epsilon^{\sup} - H_\epsilon^{\inf}$ は情報源の非定常性・混合性に由来する系列間ばらつきを表す．定常無記憶情報源では Z_n がエントロピレートに収束し，$H_\epsilon^{\sup} = H_\epsilon^{\inf}$ となるが，LLM のような非定常情報源では一般に幅が残る．

2.2 経験的情報スペクトル

本稿では有限長系列のみを扱うため，生成系列 $t^{(s)} = (t_1^{(s)}, \ldots, t_{T_s}^{(s)})$ とプロンプト $y^{(s)}$ に対し，系列レベル符号化レート (平均 surprisal) を

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

で定義し，この集合 $\mathfrak{Z} = \{Z^{(s)}\}_{s=1}^{S}$ を経験的情報スペクトルと呼ぶ．

また $H_\epsilon^{\sup}, H_\epsilon^{\inf}$ の操作的近似として，$\mathfrak{Z}$ の分位点を用いる．具体的には，ϵ に対応するパラメータ θ を導入し[1)]，$(1-\theta)$ 分位点を $\hat{H}_\theta^{\sup}$，θ 分位点を $\hat{H}_\theta^{\inf}$ とし，$\hat{H}_w := \hat{H}_\theta^{\sup} - \hat{H}_\theta^{\inf}$ を経験的な幅と定義する．具体的には，$\hat{H}_\theta^{\sup}$ および $\hat{H}_\theta^{\inf}$ は経験分布の分位点として

$$\hat{H}_\theta^{\sup} := \inf\Big\{\alpha \;\Big|\; \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}(Z^{(s)} > \alpha) \le \theta\Big\},$$

$$\hat{H}_\theta^{\inf} := \sup\Big\{\beta \;\Big|\; \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}(Z^{(s)} < \beta) \le \theta\Big\}$$

で定義される．$\mathbf{1}$ は指示関数．

2.3 情報利得スペクトル (IG スペクトル)

追加情報 $\Delta y^{(s)}$ を含むプロンプト $y'^{(s)} = y^{(s)} + \Delta y^{(s)}$ で生成した系列 $t'^{(s)}$ に対し，教師強制により追加情報あり/なしの符号化レート差

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

を計算する．この分布 $\Delta\mathfrak{Z} = \{\Delta Z^{(s)}\}_{s=1}^{S}$ を IG スペクトルと呼び，$\Delta Z^{(s)} < 0$ は追加情報が当該系列の尤度を改善したことを意味する．IG スペクトルにより，追加情報が一様に効くのではなく「効く系列/効かない系列」が混在することを分布として可視化できる．なお IG スペクトルは系列対応を保った教師強制に基づく操作的指標であり，生成分布そのものの変形ではなく，同一系列に対する尤度改善を測る．

1) 本稿では $\theta = 0.01$ を用い，bootstrap により分位点推定の安定性を確認した (詳細は付録).

3 実験

LLM を一般情報源として捉えたとき，タスク構造および追加情報の形式が情報スペクトルと IG スペクトルに与える影響を検証する．QA タスクとして常識 QA(自由回答/択一)，知識 QA(RAG なし/あり)，HopQA(KG なし/あり) を用い，モデルとして Swallow-8B(base/instruct) および Qwen3-8B を比較する．各設定で生成系列の符号化レート分布を推定し，正誤差が平均ではなく tail に現れるか，および追加情報効果が系列ごとに不均一に現れるかを評価する．実験詳細 (問題数，温度設定，分位点推定，プロンプト形式など) は付録に示す．

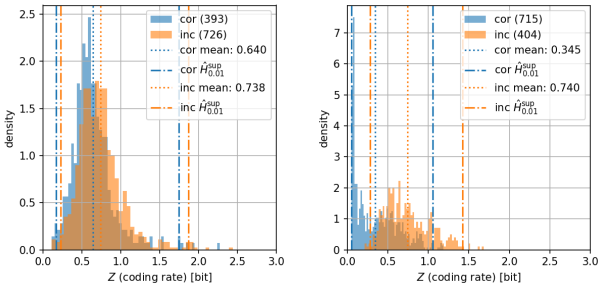
4 結果

4.1 正誤差は平均ではなく tail に現れる

図 1 は常識 QA における情報スペクトルを正解/不正解で比較したものである．正解系列は平均的に低い符号化レートを示すが，分布は大きく重なり，平均値のみでは正誤を十分に分離できない．一方，高 surprisal 側の tail に着目すると差が顕著であり，不正解系列が tail 領域を支配するのに対し，正解系列は低 surprisal 領域に集中する．この結果は，誤りが「平均の悪化」としてではなく「非典型系列 (tail) の出現」として現れることを示唆する．したがって，平均値に加えて tail を特徴づける分位点統計量や分布幅により，失敗系列の出現領域を診断できる．

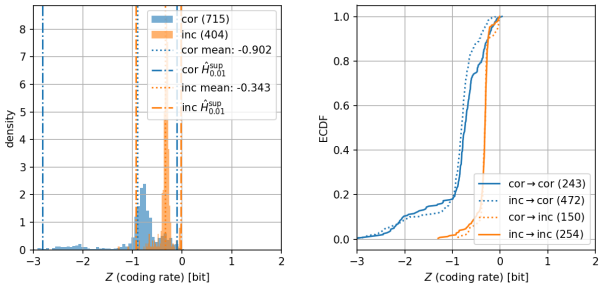
4.2 追加情報はスペクトル形状を体系的に変える

追加情報 (択一選択肢，RAG，KG) を導入すると，多くの設定で情報スペクトルが収縮し，tail が抑制される傾向が観測された．しかしその効果は系列ごとに一様ではない．図 2(a) は選択肢提示による IG スペクトルを示す．IG スペクトルは，各系列に対して追加情報が符号化レートをどの程度変化させたかを表す分布であり，$\Delta Z^{(s)} < 0$ は追加情報により尤度が改善したことを意味する．観測される分布は単峰ではなく混合的であり，追加情報が大きく効く



(a) 自由回答　　(b) 択一

図 1: 常識 QA における情報スペクトル (Swallow-8B)．選択肢付与による分布形状および tail の変化を示す．図中 cor: 正解，inc: 不正解を示す．カッコ内は該当のサンプル数である．



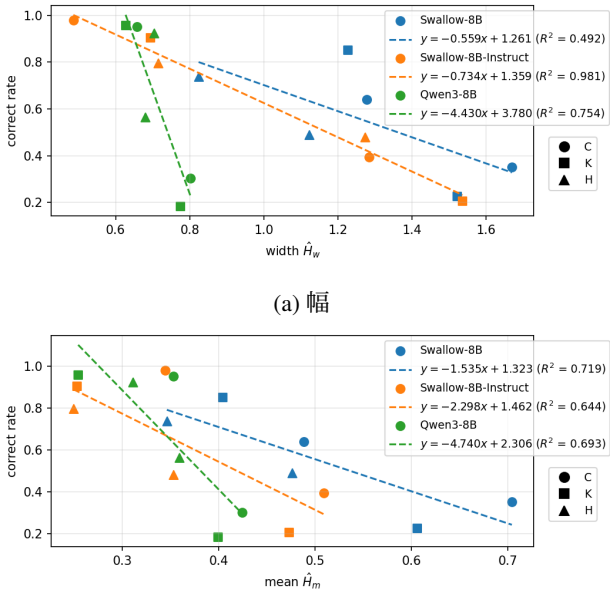
(a) IG スペクトル　　(b) ECDF

図 2: 常識 QA における IG スペクトル (Swallow-8B)．追加情報 (選択肢) 付与による符号化レート差の分布を，正誤遷移で比較する．

系列とほとんど効かない系列が同時に存在することが分かる．このように，IG スペクトルにより追加情報効果の不均一性を系列レベルで可視化できる．図 2(b) の ECDF では cor → cor / inc → cor / cor → inc / inc → inc の 4 群で比較し，修正群 (inc → cor) が最も負方向にシフトすることが分かる．

4.3 スペクトル統計量は性能指標と相関する

図 3 は，各モデルにおける情報スペクトル統計量 (平均/幅) と QA 正解率の関係を示す．その結果，平均および幅が大きい条件ほど正解率が低下する傾向が確認された．さらに，温度や追加情報の有無を制御した回帰分析 (表 1) により，分布指標の有効性 (増分説明力) はモデルに依存することが示唆された．Swallow-8B(base) では平均の寄与が相対的に大きい一方，Swallow-8B-Instruct および Qwen3-8B では幅がより強く性能を説明する傾向が見られた．この結果は，指示チューニングにより性能を規定する手掛かりが平均から分布幅/tail 構造へシフトし得ることを示唆する．



(a) 幅

(b) 平均

図 3: スペクトル幅/平均と QA(C: 常識，K: 知識，H:HopQA) 正解率の関係．各モデルにつき，ベース条件 3 点 (追加情報なし) と追加情報付与条件 3 点の計 6 点を比較した．

表 1: 温度 τ と追加情報の有無 info を制御した回帰における，平均 (H_m)・幅 (H_w) の増分説明力 (leave-one-out 推定)．

Metric	Swallow-8B		Qwen3-8B
	base	instruct	
$R^2(\tau, \text{info})$	0.736	0.855	0.851
$R^2(\tau, \text{info}, H_m)$	0.827	0.853	0.867
$R^2(\tau, \text{info}, H_w)$	0.793	0.917	0.883
$\Delta R^2(H_m)$	0.345	-0.014	0.107
$\Delta R^2(H_w)$	0.216	0.428	0.215

5 まとめ

本稿では，LLM を一般情報源として捉え，系列レベル符号化レートの分布 (経験的情報スペクトル) に基づく診断手法を提案した．実験により，誤りは平均的指標ではなく高 surprisal 側の tail に強く現れることを示した．また，追加情報の効果は系列ごとに不均一であり，IG スペクトルにより「どの系列で効いたか」を分布として可視化できることを示した．さらに，指示チューニングにより性能を規定する要因が平均から分布幅/tail へシフトし得ることを示唆した．以上より，LLM の生成挙動を平均ではなく分布構造として診断することの有用性が示された．今後は多様なタスクへの拡張や，副情報がある条件下でスペクトルを扱う方法 [14] との接続を検討する．

参考文献

[1] Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling?, 2025.

[2] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. **ACM Comput. Surv.**, Vol. 55, No. 12, March 2023.

[3] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. **ACM Trans. Inf. Syst.**, Vol. 43, No. 2, January 2025.

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

[5] Stefania Degaetano-Ortlieb and Elke Teich. Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns. In Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Feldman, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, **Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature**, pp. 68–77, Vancouver, Canada, August 2017. Association for Computational Linguistics.

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

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

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

[9] Sohee Yang, Jonghyeon Kim, Joel Jang, Seonghyeon Ye, Hyunji Lee, and Minjoon Seo. Improving probability-based prompt selection through unified evaluation and analysis. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 664–680, 05 2024.

[10] Hongfu Liu and Ye Wang. Towards informative few-shot prompt with maximum information gain for in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 15825–15838, Singapore, December 2023. Association for Computational Linguistics.

[11] Jiaming Ji, Kaile Wang, Tianyi Alex Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Josef Dai, Yunhuai Liu, and Yaodong Yang. Language models resist alignment: Evidence from data compression. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 23411–23432, Vienna, Austria, July 2025. Association for Computational Linguistics.

[12] T.S. Han and S. Verdu. Approximation theory of output statistics. **IEEE Transactions on Information Theory**, Vol. 39, No. 3, pp. 752–772, 1993.

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

[14] Shigeaki Kuzuoka and Shun Watanabe. An information-spectrum approach to weak variable-length source coding with side-information. **IEEE Transactions on Information Theory**, Vol. 61, No. 6, pp. 3559–3573, 2015.

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

[16] ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. JaQuAD: Japanese Question Answering Dataset for Machine Reading Comprehension, 2022.

[17] Ai Ishii, Naoya Inoue, Hisami Suzuki, and Satoshi Sekine. JEMHopQA: Dataset for Japanese explainable multi-hop question answering. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9515–9525, Torino, Italia, May 2024. ELRA and ICCL.

A　情報スペクトルの模式図

図 4 に情報スペクトルの模式図を示す．横軸は系列レベル自己情報量密度 Z_n の実現値 (符号化レート) であり，非定常情報源では Z_n の分布が幅を持つ．本文で用いる上側分位点 $\hat{H}_\theta^{\sup}$ は高 surprisal 側 (tail) の符号化レートを，下側分位点 $\hat{H}_\theta^{\inf}$ は典型的系列の符号化レートを特徴づける．幅 $\hat{H}_w = \hat{H}_\theta^{\sup} - \hat{H}_\theta^{\inf}$ は系列間ばらつきの指標として解釈できる．

B　スペクトル統計量の推定

B.1　分位点パラメータ θ の決定

本文では経験分位点 $\hat{H}_\theta^{\sup}, \hat{H}_\theta^{\inf}$ および幅 $\hat{H}_w$ を安定に推定するため，tail 確率 $\theta = 0.01$ を用いた．θ の選択は以下の bootstrap により決定した：(1) 元サンプル集合 $\mathfrak{X}$ から復元抽出により B 回サンプル集合 $\mathfrak{X}_b$ を生成，(2) 各 $\mathfrak{X}_b$ について $\hat{H}_\theta^{\sup}, \hat{H}_\theta^{\inf}, \hat{H}_w$ を推定，(3) 推定値の分散が十分小さくなる θ を選択した．図 5 に，θ に対する推定値の変動を示す．

B.2　サンプルサイズ S と推定の安定性

経験分位点は有限サンプルから推定するため，サンプル数 S に依存して分散が変化する．図 6 に，S に対する $\hat{H}_w$ および $\hat{H}_\theta^{\sup}$ の推定安定性を示す．$S \gtrsim 300$ で推定が安定することを確認した．

C　実験設定の詳細

生成 2 タスクと QA 3 タスク設定した．

- 常識[2] (1119 問)[15]: 自由回答[3] および択一条件．
- 知識[4] (3939 問)[16]: RAG なし/あり[5] 条件．
- Hop[6] (1059 問)[17]: KG なし/あり[7]．

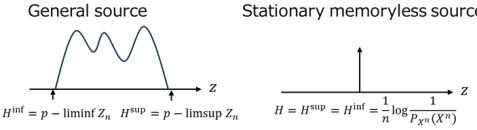


図 4: 情報スペクトルの模式図．横軸は自己情報量密度 Z_n の実現値 (符号化レート) を表す．

2)　https://huggingface.co/datasets/sbintuitions/JCommonsenseQA の (検証セット)
3)　もともと選択問題なものを選択肢を外して評価した．
4)　https://huggingface.co/datasets/SkelterLabsInc/JaQuAD の (検証セット)
5)　回答を導くうえで参考となる RAG 情報が提供されている．
6)　https://huggingface.co/datasets/sbintuitions/JEMHopQA の (学習セット)．検証セットの問題数が少なかったため，学習セットを使った．
7)　回答を導くうえで必要な KG が提供されている．

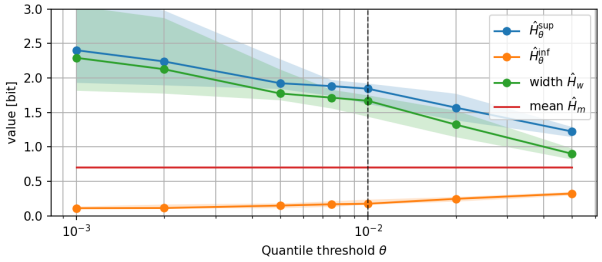


図 5: 分位点パラメータ θ に対する推定値の安定性 (bootstrap, $B = 5,000$).

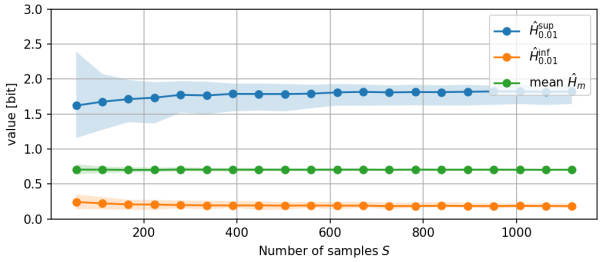


図 6: サンプルサイズ S に対する分位点統計量の推定安定性 (bootstrap, $B = 200$).

表 2: 使用したプロンプトテンプレート．

タスク	プロンプト
QA 共通 (追加情報なし)	あなたは質問応答システムです．次の質問に，日本語で簡潔に 1 文で答えてください． 質問: *question* 回答:
常識 QA (選択肢)	あなたは質問応答システムです．次の質問に，適切と思われる回答を選択肢のなかから一つ選んでください． 質問: *question* 選択肢: *choices* 回答:
知識 QA (RAG)	あなたは質問応答システムです．次の質問に，以下に示すコンテキストを参考にして，日本語で簡潔に 1 文で答えてください． 質問: *question* コンテキスト: *context* 回答:
Hop QA (KG)	あなたは質問応答システムです．次の質問に，日本語で簡潔に 1 文で答えてください． 質問: *question* ただし *deriv* 回答:
news	あなたは新聞記者です．架空のイベントに基づき，300 字程度で記事を書いてください．
poem	あなたは詩人です．恋愛をテーマにした 300 字程度の詩を書いてください．

モデルは Swallow-8B(base/instruct) および Qwen3-8B を用い，温度 $\tau \in \{0.4, 0.6, 0.8, 1.0, 1.2\}$ を比較した[8]．最大トークン数は QA:128，生成:1024 とした[9]．

8)　推論タスクでよく推奨される $\tau = 0.6$ を中心に報告した．
9)　回答に対して十分長くすることで意図せぬ打ち切りの影響を受けにくくした．