

周波数エントロピーによる位置埋込みの解明

岡 佑依^{1*} 花房 健太郎^{2*} 長谷川 拓¹ 西田 京介¹
¹NTT 株式会社 人間情報研究所 ²愛媛大学大学院理工学研究科
 yui.oka@ntt.com, hanafusa@ai.cs.ehime-u.ac.jp

概要

RoPE における周波数次元の役割は、高周波・低周波という粗い二分に基づく分析により、先行研究間で矛盾した解釈が与えられてきた。本研究は、各回転ペアを信号として周波数エントロピーで評価することで、周波数帯と周期性という二つの構造を分離して捉える。その結果、性能に本質的なのは周波数帯であり、周期性は冗長であることを示し、これらの特徴は高周波・低周波に依存しないことを示し、従来の矛盾した結果を解消する。

1 はじめに

Transformer に基づいた大規模言語モデルは様々な生成タスクにおいて優れた能力を発揮している [1, 2]。回転行列を用いた位置符号化である RoPE (Rotary Position Embedding) [3] は大規模言語モデルの標準的な位置表現であり、Llama, Qwen, Gemma モデル等で広く採用されている。しかし RoPE の「どの周波数次元が生成性能に寄与しているのか」について先行研究の主張は割れている。例えば、高周波が重要で低周波成分は不要 [4] という報告、逆に高周波は使われないという報告 [5]、低周波が長距離依存に寄与しやすいという報告 [6] が並立している。これら既存研究の多くは、高周波や低周波といった粗い二分法に依存しており、この混乱の根本原因は、次元単位で比較可能な定量指標が欠けていた点にある。また、Llama-4 [7] では、RoPE 層と NoPE 層を交互に併用する iRoPE を導入したが、その内部における周波数的挙動は十分に解明されていない。このように、RoPE 各次元の役割に関する知見は一貫せず、観察的な議論に留まっており、理論と実証をつなぐ統一的な解析枠組みが求められる。

そこで本稿では、RoPE を構成する各回転ペアを「トークン位置に沿って変化する 1 次元の信号」と

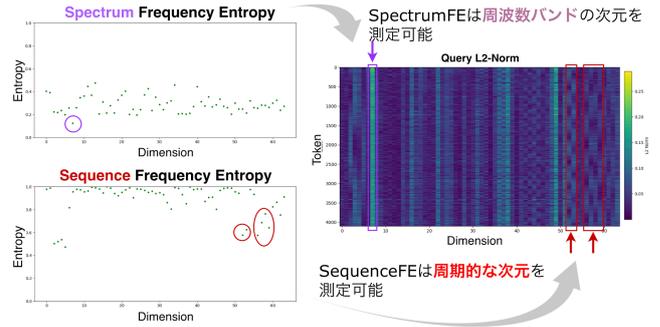


図1 RoPE の回転ペア j ごとの SpectrumFE と SequenceFE の分布図とクエリの l_2 ノルムのヒートマップ

みなし、その信号の周波数成分の偏りを、シャノンエントロピーに基づく周波数エントロピーとして数値化する。これにより、回転ペアごとに「どの周波数かどの程度テキスト生成性能に寄与しているのか」を定量的に評価可能であり、RoPE だけでなく、NoPE の評価にも適用可能である。FE は (i) 局所スペクトルの集中度 (周波数帯) を測る **SpectrumFE** と、(ii) トークン軸の周期性 (周期的な次元) を測る **SequenceFE** からなる。これにより、次元単位で周波数の比較が可能となる。Llama-4 モデルの解析から、周期的な次元は RoPE 層で顕在化し NoPE 層で消える一方、周波数帯は層・ヘッド・モデルに依存して出現することを示す。さらに推論時のみの重み付け介入 (重み付け RoPE) により、低 SpectrumFE となる周波数帯は性能に必須だが、高 SpectrumFE 外れ値となる周期的な次元は冗長であることを示す。さらに、周波数帯や周期性といった特徴は、低周波・高周波という区別に基づいて決まるものではないことを示す。これにより、従来の粗い二分的分析が生んできた矛盾を解消する。

2 背景

RoPE 多くの近年の大規模言語モデルで事実上の標準となっている RoPE [3] は自クエリ・キーに回転を適用することで位置情報を付与する。

1) *Equal contribution. 本研究は 2025 年 8 月 NTT 株式会社 人間情報研究所でのインターンシップの成果を含む。

$$A_{n,n} = (R_{n,\theta} q_n)^\top (R_{n,\theta} k_m) = q_n^\top R_{n-n,\theta} k_m \quad (1)$$

ここで $R_{n,\theta} \in \mathbb{R}^{d \times d}$ は周波数ごとに 2×2 回転を並べたブロック対角行列であり、各回転ペア j の角周波数は通常 $\theta_j = 10000^{-2j/d}$ として与えられる。

周波数エントロピー 周波数エントロピー H は、離散信号のパワースペクトルに対してシャノンエントロピーを適用し、周波数分布の「偏り」を1つの値で要約する指標である [8]。離散周波数 f_i のパワー $P(f_i)$ を正規化して $p_i = \frac{P(f_i)}{\sum_j P(f_j)}$ とすると、 $H = -\sum_i p_i \log_2 p_i$ で定義される。必要に応じて $\log_2 N$ で正規化すれば $0 \leq H \leq 1$ となる。 H が小さいほど特定周波数にエネルギーが集中し（規則的）、大きいほど広い周波数に分散する（雑音的）。信号処理・情報理論で広く用いられている [9, 10]。

3 FE による RoPE/NoPE の分析

本節では、RoPE の各回転ペアがトークン列に沿って示す振る舞いが、内容に依存した広帯域な変動なのか、それとも RoPE に起因する狭帯域・周期的な変動なのかを定量的に評価する。そこで、各回転ペアの特徴を定量的に捉えるため、2種類の指標 **SpectrumFE** と **SequenceFE** を導入する。**SpectrumFE** は「局所的にどの周波数成分が現れているか」を、**SequenceFE** は「トークン方向にどれだけ周期的に振動しているか」を評価する。これらの指標は次のように計算される。(1) トークン軸に沿った1次元の離散信号を作り、(2) その信号のパワースペクトルを計算し、(3) そのパワースペクトルのシャノンエントロピーを計算する。回転ペアごとに得られる FE の値はモデルに依存しないため、層・ヘッド・アーキテクチャ間で比較可能である。

3.1 RoPE を離散信号へ変換する

[4] に従い、Cauchy-Schwarz の不等式に基づいて RoPE における周波数成分の利用状況を定量化する。 j 番目の周波数成分がアテンションスコア $A_{n,m}$ に与える寄与は、次の Cauchy-Schwarz の不等式から、対応するクエリおよびキー成分の ℓ_2 ノルムによって上界付けられる。

$$|\langle q_n^{(j)}, k_m^{(j)} \rangle| \leq \|q_n^{(j)}\|_2 \|k_m^{(j)}\|_2, \quad j = 0, \dots, \frac{d}{2} - 1.$$

この上界に基づき、各回転ペア j について、トークン位置方向に沿ってこれらのノルムを計算する。

$$\mathbf{s}_j := [\|q_0^{(j)}\|_2, \|q_1^{(j)}\|_2, \dots, \|q_{L-1}^{(j)}\|_2]^\top \in \mathbb{R}^L, \quad (2)$$

ここで L は系列長である。この $\|q_n^{(j)}\|_2$ のトークン位置 n に沿った列は、RoPE の各周波数成分 j の $A_{n,m}$ への寄与の大きさを反映する系列として解釈でき、これを離散信号とみなして解析する。

3.2 パワースペクトルの計算

STFT によるスペクトル 局所的な周波数構造を評価するために、回転ペアごとの信号 \mathbf{s}_j に対し、短時間フーリエ変換 (STFT) を用いてスペクトルを計算する。STFT により得られるパワースペクトル $S_{k,t}$ を時間方向に平均化し、 S_k を計算する。

$$S_k = \frac{1}{T} \sum_{t=0}^{T-1} S_{k,t}, \quad S_{k,t} = \left| \sum_{n=0}^{F-1} \mathbf{s}_j[tH+n] w[n] e^{-i\frac{2\pi}{F}kn} \right|^2 \quad (3)$$

t ($t = 0, 1, \dots, T-1$) はフレーム番号、 $w[n]$ は解析窓、 k ($k = 0, 1, \dots, K-1$) は周波数ビンを表す。¹⁾

DFT によるスペクトル トークン軸方向の周期性を評価するために、信号 \mathbf{s}_j 全体に対して離散フーリエ変換 (DFT) を用いてスペクトルを計算する。

$$S_k = \left| \sum_{n=0}^{L-1} \mathbf{s}_j[n] e^{-i\frac{2\pi}{L}kn} \right|^2 \quad (4)$$

この時、DC 成分を除いた正の周波数成分 S_k ($1 \leq k \leq \lfloor L/2 \rfloor - 1$) のみを用いる。

3.3 シャノンエントロピーの計算

離散信号から計算された各パワースペクトル $\{S_k\}$ に対し、正規化された分布

$$p_k = \frac{S_k}{\sum_k S_k} \quad (5)$$

を計算する。その分布 p_k のシャノンエントロピー

$$H = -\sum_k p_k \log_2 p_k \quad (6)$$

を最大値 $H_{\max} = \log_2 K$ で正規化することで、スケール非依存な指標 $\tilde{H} = H/H_{\max}$ を得る。そして、各回転ペア $j \in \{0, \dots, \frac{d}{2} - 1\}$ について \tilde{H}_j を計算し、RoPE の各回転ペアごとのエントロピー分布を得る。そして、STFT によるスペクトルから得られるエントロピー分布を **Spectrum Frequency Entropy (SpectrumFE)** と呼び、DFT によるスペクトルから得られるエントロピー分布を **Sequence Frequency Entropy (SequenceFE)** と呼ぶ。このエントロピー

¹⁾ 本研究では、フレーム長 $F = 1024$ 、ホップ長 $H = 512$ 、系列長 $L = 4096$ と設定する。このとき、周波数ビン数は $K = \frac{F}{H} + 1 = 513$ 、フレーム数は $T = \lfloor \frac{L}{H} \rfloor + 1 = 7$ となる。

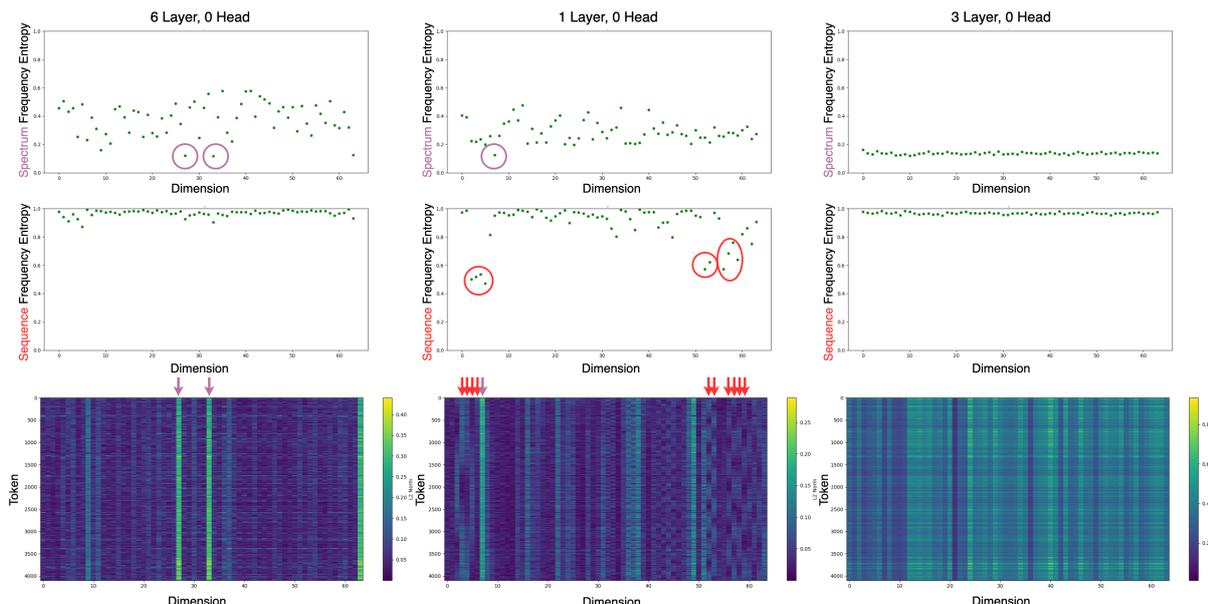


図2 Llama-4-Scout-17B-16E-Instruct モデルにおける各 FE の分布図とクエリの l_2 ノルムマップ. 層 6, 層 1, 層 3 の 0 番目のヘッドのみを示している. FE の分布図では, 横軸が RoPE における回転ペアのインデックス j , 縦軸がエントロピー値 \tilde{H}_j , クエリの l_2 ノルムマップでは, 縦軸が回転ペアのインデックス j , 横軸がトークン長 $n(0 \leq n \leq 4096)$ である.

分布とクエリ成分の l_2 ノルムのヒートマップと比較することで, 周波数成分の特徴を調査する.

3.4 分析

RoPE と NoPE のそれぞれの効果を同一モデル内で比較するため, Llama-4 モデルを使い分析を行った. 実験設定は付録 A.1 に記載している. 図 2 に Llama-4 モデルにおける SpectrumFE, SequenceFE, そして l_2 -ノルムのヒートマップを示す.

観察 1: SpectrumFE は周波数帯を捉える 図 1 左の層 6 において, SpectrumFE が最小となる次元は, クエリ l_2 ノルムが持続的に高い回転ペアの連続領域と一致している. これは周波数帯と呼ばれ, 先行研究でも観測されている [4]. SpectrumFE は各回転ペアに対して, STFT によって短い時間区間を切り出すことで「どの振動成分にエネルギーが集まっているか」を計測しているため, このような周波数帯が観測可能である.

観察 2: SequenceFE は周期的な次元を捉える 図 1 中央の層 1 において, SequenceFE が最小となる次元はクエリ l_2 ノルムが明確な周期振動を示す領域と一致している. SequenceFE は「クエリの大きさが, トークンに沿って同じリズムで揺れているか」を系列全体を DFT することで計測しているため, このような周期的な次元を観測可能である.

観察 3: NoPE 層は周期次元を消し, 周波数帯を多く持つ 図 2 の下段のクエリ l_2 ノルムのヒート

マップから, RoPE 層(層 1,6)と NoPE 層(層 3)では挙動が大きく異なることがわかる. SpectrumFE の観点では, NoPE 層は RoPE 層よりも多くの周波数帯を示す. 一方, SequenceFE の観点では, NoPE 層においては明確な周期振動を示す回転ペアは観測されない. 層 1 などで見られた周期的な次元を打ち消す可能性がある.

4 周波数帯と周期性の影響

SpectrumFE と SequenceFE の分析から観測された周波数帯と周期性はモデルにとって冗長な要素なのか, それとも不可欠な構成要素なのか?そこで, 各 FE の値に基づき, 特定の回転ペアの寄与を選択的に重み付けすることで RoPE に介入し, 周波数帯と周期的な次元の役割を検証する.

4.1 重み付け RoPE

全層全ヘッドのクエリとキーの FE をそれぞれ計算する. 次に, 層 l ・ヘッド h ・RoPE の回転ペア j の FE を $\tilde{H}^{(l,h,j)}$ とし, 閾値 τ により, 該当ペアの q, k を推論時だけ α で減衰させる. ここで, α は 0 から 1.0 まで 0.1 刻みで変化させる.

$$\alpha^{(l,h,j)} = \begin{cases} \alpha & (\tilde{H}^{(l,h,j)} < \tau) \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

$q_m^{(j)*} = \alpha^{l,h,j} R_{m,\theta}^{(l,h,j)} q_m^{(j)}$, $k_n^{(j)*} = \alpha^{l,h,j} R_{n,\theta}^{(l,h,j)} k_n^{(j)}$
学習は一切行わず, 推論時のみこの操作を行う.

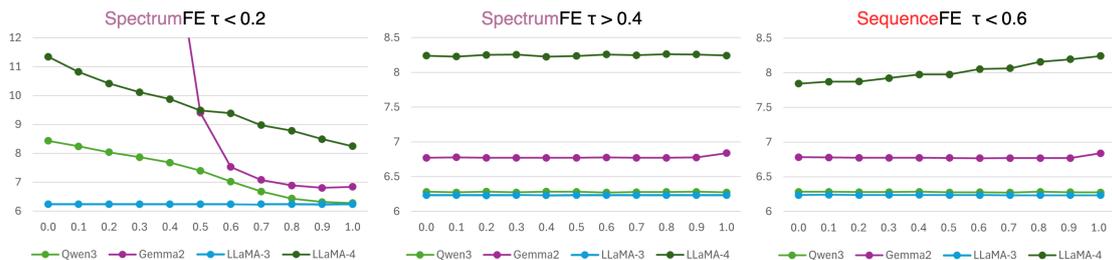


図3 重み α に対するパープレキシティの変化を示す。横軸は重み付け RoPE における重み α ，縦軸はパープレキシティである。Sequence FE については， $\tau > 0.4$ および $\tau < 0.2$ の2条件で，Spectrum FE については， $\tau < 0.6$ で検証した。

4.2 実験

実験設定 Llama-4-Scout-17B-16E[7], gemma-2-9B[11], Qwen3-8B[12], および Llama-3-8B[13] の4モデルを評価した。評価指標として，wikitext-103のテストセット [14, 15] のパープレキシティを計測した。閾値 τ は，FE に応じて異なる設定を採用した。SequenceFE は， $\tau > 0.4$ および $\tau < 0.2$ の二つの条件を評価した。SpectrumFE では，外れ値が低いエントロピー値にのみ現れることが観測されたため， $\tau < 0.6$ の条件を評価対象とした。

実験結果 重み α 毎のパープレキシティを図3に示す。SpectrumFE に基づき $\tau < 0.2$ の次元を抑制すると (図3左)， α の減少に伴いパープレキシティが増加する。よって， $\tau < 0.2$ の次元，すなわち周波数帯がモデル性能に重要であることが確認された。一方， $\tau > 0.4$ の次元を抑制すると (図3中央)，パープレキシティはほぼ変わらず，高い τ をもつ外れ値次元は冗長である可能性が示唆される。また，SequenceFE に基づき $\tau < 0.6$ の次元を抑制した場合 (図3右)，パープレキシティは若干低下し，特に Llama-4 ではその傾向が顕著であった。この結果は，周期的な次元がモデルにとって必須ではないことを示している。

4.3 下流タスクでの検証

前節の結果から，SpectrumFE における $\tau > 0.4$ の周波数帯と，SequenceFE における $\tau < 0.6$ の周期的次元は，モデルにとって冗長であると仮定する。これらの次元の影響を調べるため，複数の下流タスクで重み付け RoPE を評価した。重み付け RoPE では，SpectrumFE の $\tau > 0.4$ に該当する次元と，SequenceFE の $\tau < 0.6$ に該当する次元の双方に対し，重み $\alpha = 0.1$ を適用した。評価には，HellaSwag [16], TruthfulQA [17], および MMLU [18] を用いた。

表1 RoPE(-) と重み付け RoPE(+W) の比較

モデル	HellaSwag		TruthfulQA		MMLU	
	-	+W	-	+W	-	+W
Llama-4	66.67	66.67	97.32	97.99	60.05	60.81
Llama-3	60.16	60.16	84.85	84.85	34.21	34.21
Qwen3	58.94	58.92	95.31	95.31	57.89	57.89
Gemma-2	61.02	61.21	98.83	98.83	72.81	72.81

実験結果 表1に示すように，RoPE と重み付け RoPE の性能はほぼ同等であった。この結果から，SpectrumFE の外れ値次元および SequenceFE の周期的次元は冗長である可能性が高い。

5 先行研究の矛盾の解消

FE の分析から，性能を左右する周波数帯は高周波・低周波のいずれかに偏って現れるのではなく，層ごとに異なる次元に形成されることが分かった。したがって，「高周波か低周波か」という粗い区分で次元の有効性を議論すると，どの層・どの次元を見たかに依存して結論が揺れ，その結果，先行研究の矛盾が生じたと考えられる。また，層・ヘッド・次元が膨大な LLM では目視解析は現実的ではなく，FE のような次元単位の定量指標が有用である。

6 おわりに

本稿は，RoPE を周波数の使い方という観点で統一的に測る周波数エントロピー (FE) を提案し，(i) 周波数帯 (低 SpectrumFE) と (ii) 周期性 (低 SequenceFE) を分離して捉えられること，さらに推論時のみの介入で周波数帯は必須だが周期性は冗長になり得ることを示した。また，FE は RoPE だけでなく NoPE などの他の位置符号化手法にも使え，汎用的な評価指標になり得る。我々が提案した FE は単なる可視化ではなく，RoPE の冗長次元削減による KV キャッシュ圧縮への応用や，位置表現の探索指針になると考える。今後は，FE を解析に使う指標ではなく設計原理に組み込むことを検討する。

参考文献

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.
- [2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. **ArXiv**, Vol. abs/2302.13971, , 2023.
- [3] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.
- [4] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful?, 2024.
- [5] Ting-Rui Chiang and Dani Yogatama. The rotary position embedding may cause dimension inefficiency in attention heads for long-distance retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 13552–13562, Vienna, Austria, July 2025. Association for Computational Linguistics.
- [6] Xiangyu Hong, Che Jiang, Biqing Qi, Fandong Meng, Mo Yu, Bowen Zhou, and Jie Zhou. On the token distance modeling ability of higher RoPE attention dimension. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 5877–5888, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [7] Meta. Meta-llama-4-scout-17b-16e-instruct. <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct>, 2025. Hugging Face model repository.
- [8] Claude E. Shannon and Warren Weaver. **The Mathematical Theory of Communication**. University of Illinois Press, 1949.
- [9] H. Misra, S. Iqbal, H. Bourlard, and H. Hermansky. Spectral entropy based feature for robust asr. In **2004 IEEE International Conference on Acoustics, Speech, and Signal Processing**, Vol. 1, pp. I–193, 2004.
- [10] Victor Sucic, Nicoletta Saulig, and Boualem Boashash. Analysis of local time-frequency entropy features for non-stationary signal components time supports detection. **Digital Signal Processing**, Vol. 34, pp. 56–66, 2014.
- [11] Google. Gemma-2-9b-it: Model card. <https://huggingface.co/google/gemma-2-9b-it>, 2024. Accessed: 2026-01.
- [12] Qwen Team. Qwen3-8b: Model card. <https://huggingface.co/Qwen/Qwen3-8B>, 2025. Accessed: 2026-01.
- [13] Meta. Meta-llama-3-8b: Model card. <https://huggingface.co/meta-llama/Meta-Llama-3-8B>, 2024. Accessed: 2026-01.
- [14] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.
- [15] Salesforce. Wikitext dataset. <https://huggingface.co/datasets/Salesforce/wikitext>, 2024. Accessed: 2026-01.
- [16] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, 2019.
- [17] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. **Proceedings of the International Conference on Learning Representations (ICLR)**, 2021.
- [19] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In **International Conference on Learning Representations**, 2017.
- [20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **arXiv e-prints**, 2019.

A 付録

A.1 節 3.4 の実験設定

Llama-4-Scout-17B-16E-Instruct モデル [7] を用いて実験を行った。本モデルのヘッド次元は 128 であり、RoPE は 64 個の回転パターンをもつ。また、レイヤー数は 48、アテンションヘッド数は 40 である。評価には、WikiText-103 データセット [19] の wikitext-103-raw-v1 から複数のテキストを抽出し、それらを連結することで長さが正確に 4096 トークンとなる系列を構成した。各系列をモデルに入力し、推論時におけるアテンションのクエリベクトルを抽出し、これらのクエリに基づいて、それぞれの FE を算出した。

A.2 キーの分析

本節では節 3 と同様の分析をキーに対しても行う。図 4 にキーに対する各 FE の分布と l_2 ノルムのヒートマップを示す。RoPE が適用された層では、クエリと同様に、キーにおいても周波数帯および周期的次元の双方が観測される。FE はこれらの構造と整合しており、キーに対しても同様の傾向を示す。特に、キーにはクエリと比較して周期的次元がより多く含まれており（例えば第 5 層において顕著である）。一方、NoPE 層では、クエリの場合と同様に、周期的次元はまれである一方、周波数帯は頻繁に検出される。

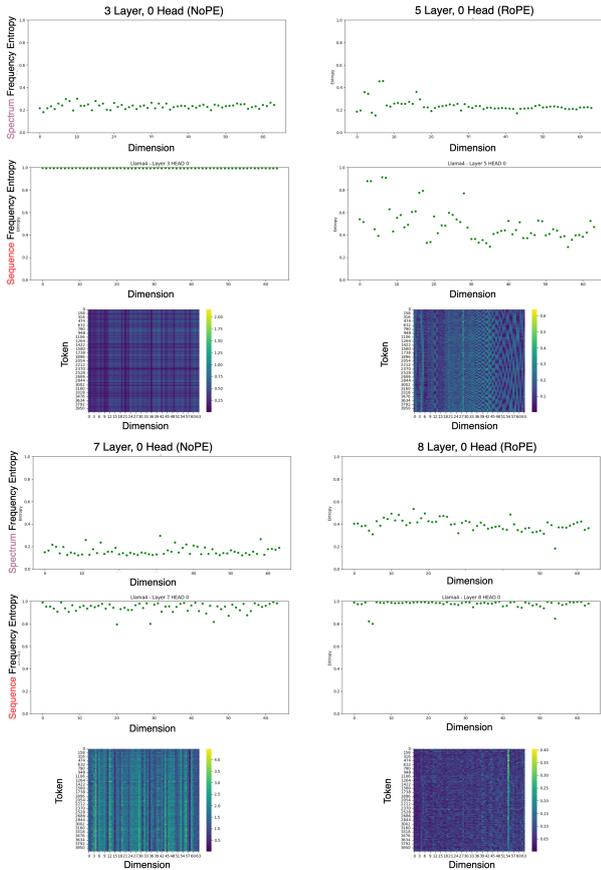


図 4 Llama-4-Scout-17B-16E-Instruct モデルにおける各 Frequency Entropy (FE) の分布図を示す。列は左から順に層 6, 層 1, 層 3 である。

A.3 他モデルにおける検証

Llama-4 モデル以外の FE についても節 3 と同様の検証を行った。ページの都合上、Gemma-2 モデルのみの例を示す。図 5 から、周波数帯は多くのヘッドにおいて観測され、一貫して明瞭な周波数帯を形成している。しかし、周波数帯が存在する場所は Llama-4 モデルと異なる。一方で、周期的次元は確認されず、SequenceFE は全体として高い値を維持することがわかった。

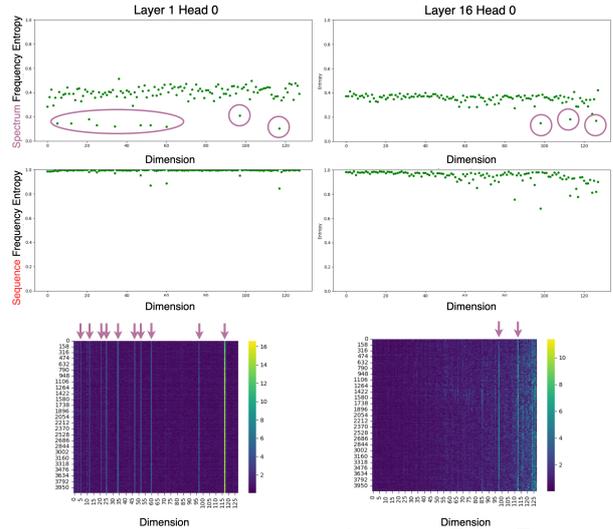


図 5 Gemma-2-9b-it モデルの FE 解析結果。

A.4 FE の頑健性

データセット、系列長を変えた場合、周波数帯や周期的な次元は変化するのだろうか？節 3 と異なる設定 (C4 データセット [20], 系列長 8192) で節 3 と同様の Llama-4 モデルの検証結果を図 6 に示す。結果から、データセット、系列長を変えた場合でも周波数帯や周期的な次元は観測され、その存在次元も変わらないことがわかった。NoPE 層についても同様であった。よって、FE が捉えている周波数帯・周期的次元は、データセット依存でも系列長依存でもなく、RoPE にや NoPE に内在する構造である。

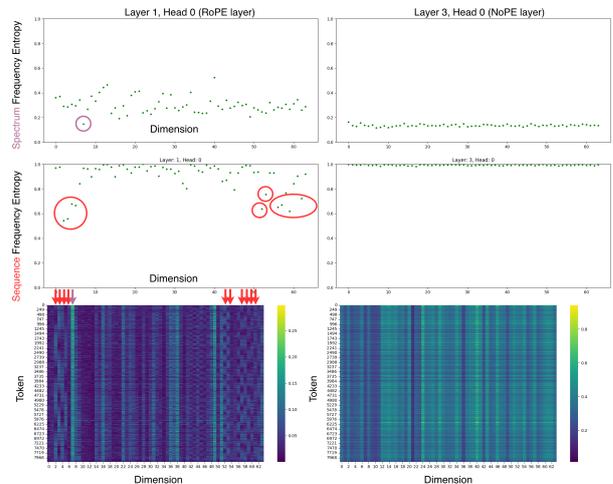


図 6 C4 データセット [20], 系列長 8192 の Llama-4 モデルの FE 解析結果。