位置符号化の基底拡大戦略は外挿性能を制限する

岡 佑依 [1,2] 斉藤 いつみ [2] 西田 京介 [1]
[1]NTT 株式会社 人間情報研究所 [2] 東北大学
yui.oka@ntt.com

概要

Rotary Position Embedding（RoPE）は LLM で広く用いられており，一般に基底 θ を大きい値に設定することで長文脈性能が向上すると考えられているが，コンテキスト拡張には依然として再学習が不可欠である．しかし，本研究では，基底 θ を事前学習時の最大系列長に設定し，推論時に大きくするだけで，微調整なしで外挿性能が向上し，コンテキスト拡張のための再学習が不要となることを示す．この設定の妥当性は，RoPE の周波数帯構造の実証的および理論的分析によって裏付けられる．

1 はじめに

Rotary Position Embedding（RoPE）[1] は，Transformer 系大規模言語モデル (LLM) [2, 3, 4, 5, 6] で広く用いられている位置符号化であり，基底周波数 θ に基づく回転行列によってトークン位置を表現する．この時，RoPE の基底 θ を従来の 10,000 や 500,000 以上へ拡大する戦略が一般的である [2, 4, 5]．この設計は，相対距離に伴う注意スコアの急激な減衰を抑制するという直感に基づいており，長系列に有効と考えられている．しかし，θ を拡大するだけでは事前学習時の最大系列長 (本稿では L_{train} と定義する) を超えた生成性能 (外挿性能) は十分に得られず，LLM が取り扱える系列長を拡張するために長系列データを用いて再学習を行う必要がある [4, 7]．一方，Barbero らは，相対距離に伴う注意スコアの急激な減衰は起こりえず，クエリおよびキーの低周波次元において，全トークンにわたって高い値が連続している**周波数帯**が存在することを指摘した [8]．また，この周波数帯より低周波の RoPE 次元を NoPE に置き換えても性能が劣化しないことを示し，これらの次元が位置情報を必要としない可能性を示した．

この背景から，θ の拡大は RoPE の表現力を高めるのか，それとも無効な次元を増やすだけなのか？という疑問が生じる．本研究ではこの疑問に答えるため，周波数帯に着目し，周波数帯が何が要因で，いつ，どこに，どのように存在するのかを実証的かつ理論的に明らかにした．その結果，RoPE は全次元を一様に利用しているのではなく，θ と事前学習時の最大系列長 L_{train} に依存して，限られた次元にのみ位置表現能力が集中することがわかった．さらに，θ を L_{train} と同じ値に設定し，RoPE 次元を最も効率的に活用することで，外挿性能が向上することを発見した．すなわち，従来，RoPE では外挿性能が低いためコンテキスト拡張のための再学習が不可欠と考えられてきたが，**RoPE の基底 θ の拡大戦略は外挿性能を制限しており，最大系列長 L_{train} と同じ値に設定すると外挿性能が備わることを示す．**

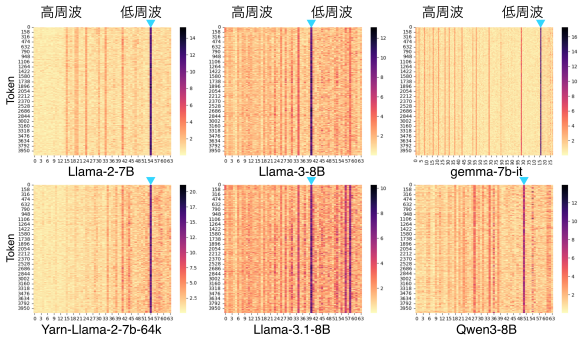

図 1 代表的な LLM における $L2$-ノルムのヒートマップ．縦軸は系列長，横軸は RoPE の各回転ペア ($i \in \{0, 1, \ldots, d/2\}$)．▼ が指す次元が周波数帯である．

2 背景

RoPE RoPE [1] は，多くの LLM で事実上の標準となっている位置表現であり，クエリ・キーに回転行列を適用することで位置情報を付与する．

$$\begin{bmatrix} \cos m\theta_j & -\sin m\theta_j \\ \sin m\theta_j & \cos m\theta_j \end{bmatrix} \begin{bmatrix} q^m_{2j-1} \\ q^m_{2j} \end{bmatrix} \tag{1}$$

$q^m \in \mathbb{R}^{1 \times d}$ は次元数が d の場合の m 番目のクエリであり，各回転ペア j の周波数は $\theta_j = \theta^{-2j/d}$ と計算され，基底 θ は 10,000（Gemma，Llama-2），500,000（Llama-3），1,000,000（Qwen-3）という値が使われている [4, 6, 5]．RoPE の外挿性能は低いため，LLM の

表 1 節 3 および節 4.3 における分析結果．YaRN と Llama3 は，微調整時に適用される位置補間手法である．Gemma モデルのヘッド次元数 d は 256 であり，それ以外のモデルは $d = 128$ である．

モデル	系列長 L_{train}		基底 θ	帯次元 (節 3)		p-RoPE のパープレキシティ (節 3)				予測帯次元 (節 4.3)	
	事前学習	微調整		i_{band}	$i_{band}/\frac{d}{2}$	r=1.0	r=0.9	r=0.75	r =0.50	$j^\star$	$c \times j^\star$
Gemma	8k	-	10000	116.68	0.91	2.52	2.70	81.66	> 100	107	117
Qwen3	40k	-	1000000	51.04	0.79	6.22	6.22	6.22	7.46	43	47
Llama-2	4k	-	10000	53.53	0.84	2.54	2.58	> 100	> 100	49	53
+YaRN	4k	64k	10000	51.93	0.81	2.81	5.08	> 100	> 100	-	-
Llama-3	8k	-	500000	43.43	0.68	2.29	2.29	2.29	84.50	38	41
+Llama3	8k	131k	500000	40.47	0.63	2.29	2.29	2.29	5.53	-	-

コンテキストを拡張する際 θ の値を変更する位置補間 [7, 9] が一般的に適用されるが，事前学習に加えて追加の微調整学習が必要である．

周波数帯 Barbero らは，RoPE 適用後のクエリ $q^m \in \mathbb{R}^{1\times d}$ の $L2$ ノルム $\|q^m\|_2$ において高い値が連続して現れる周波数帯が存在すること，低周波次元の RoPE を NoPE [10] に置き換えて事前学習しても性能が変化しないことを示した [8]．一方で，彼らの分析は短文や Gemma モデルに限定されており，位置補間や長文脈に対する検証は行われておらず，周波数帯が形成される機構についても未解明である．

3 代表的な LLM における周波数帯

他の LLM や位置補間を適用したモデルでも周波数帯は観測されるのか？先行研究 [8] で分析されていない LLM や位置補間を対象に分析を行う．

3.1 分析手法

L2 ノルムヒートマップ 周波数の利用状況を測るため，先行研究 [8] と同様に，主にクエリの $L2$ ノルム $\|q^m\|_2$ を計算し，ヒートマップとして可視化する．このとき，RoPE では各周波数が 2 次元の回転ペアとして表現されるため，ヘッド次元数が d の場合，独立な周波数成分は $d/2$ 個となる．

p-RoPE RoPE における周波数成分の寄与を分析するため，先行研究 [8] と同様に，低周波次元を無効化する p-RoPE [8] を用いてパープレキシティを測定した．p-RoPE は高周波側の上位 r 次元のみに RoPE を適用する．例えば，$r = 0$ の時 NoPE と一致し，$r = 1$ の時通常の RoPE と一致する．なお，本研究では推論時のみを評価し，学習は行わない．

周波数帯次元 i_{band} 次の手順で周波数帯が存在する次元を特定する．まず，各トークン位置 n におけるクエリ q^n について，RoPE の回転ペアに対応する $\frac{d}{2}$ 個の周波数次元の中から，2 ノルムが最大となる次元を次のように計算する．

$$\text{idx}_n = \underset{i\in\{0,1,\ldots,\frac{d}{2}-1\}}{\arg\max} \|k_i^n\|_2$$

次に，系列長 L 全体にわたって得られた $\{\text{idx}_n\}_{n=0}^{L-1}$ の中で，最も出現するインデックスを求める．

$$\hat{\text{idx}} = \underset{k^n\in\{k^0,k^1,\ldots,k^{L-1}\}}{\text{argmax}} (\text{count}(\text{idx}_n))$$

この操作を全ヘッドおよび全層に対して行い，得られた $\hat{i}$ の平均値を周波数帯次元 $i_{\text{band}} (0 \le i_{\text{band}} \le \frac{d}{2})$ として定義する．さらに，p-RoPE の r と対応づけるため，周波数帯次元 i_{band} を $\frac{d}{2}$ で割った相対的な周波数帯次元 $i_{band}/\frac{d}{2}$ についても報告する．

実験設定 複数のモデル（Gemma 8B，Llama-2 7B，Llama-3 8B，Qwen-3-8B）および複数の位置補間（YaRN[7]，Llama-3.1 における scaling[11]）を用いて分析を行なった．評価データセットには Wikitext-103[12] のテストセットを用い，すべてのモデルにおいて推論時の系列長は $L = 4096$ とした．

3.2 分析結果

図 1 に各モデルにおけるクエリの 2 ノルムを示す．第 1 層のヘッドからクエリを抽出した．その結果，すべてのモデルで周波数帯が確認され，周波数帯は一般的な現象であることが分かった．一方で，周波数帯が出現する次元位置はモデルごとに異なる．また，位置補間モデルにおいても，補間手法に依らず周波数帯が継承され不変であることを確認した．表 1 に p-RoPE と周波数帯次元 i_{band} の結果を示す．周波数帯より低周波側の RoPE 次元を NoPE に置き換えても ($r > i_{band}/\frac{d}{2}$)，性能低下は見られず，低周波数の RoPE が有効に利用されていないことが示された．しかし，周波数帯を含めた RoPE 次元を NoPE に置き換えると大幅に性能は低下する．

✓ **結論** 周波数帯は一般的に観測され，位置補間を伴う微調整後も継承され不変である．また，周波数帯は性能に寄与する重要な特徴であり，周波数帯より低周波の次元の RoPE は性能に寄与しない．

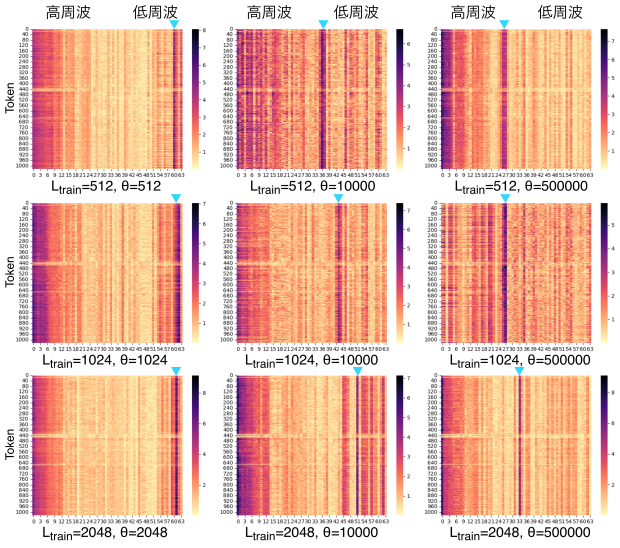

図 2　$\{L_{\text{train}}, 10000, 500000\} \times \{512, 1024, 2048\}$ の θ と L_{train} の組み合わせにおける $L2$-ノルムを示す．縦軸は系列長，横軸は RoPE の各回転ペア（$i \in \{0, 1, \ldots, d/2\}$）．

4　周波数帯の解明

次に，この重要な特徴である周波数帯がいつ，どこに，どのように形成されるのかを明らかにするため，$(L_{\text{train}}, \theta) \in \{512, 1024, 2048\} \times \{L_{\text{train}}, 10000, 500000\}$ の組み合わせにおける小規模な Transformer デコーダの事前学習を行ない，周波数帯を分析した．実験設定とその他の評価結果は付録 A に記載している．

4.1　実証分析：どこに存在するのか？

図 2 に，θ と L_{train} の組合せに対する 2 ノルムマップを示す．θ を固定すると（$L_{\text{train}} = 512 \to L_{\text{train}} = 2048$，図 2 の上から下），$L_{\text{train}}$ の増加に伴い周波数帯は高次元側へ移動し，周波数帯の位置が学習時最大系列長に依存することが分かる．一方，L_{train} を固定して θ を増加させると（図 2 の左から右），周波数帯は低次元側へ移動する．これから，周波数帯が存在する次元は基底 θ と L_{train} の両方に依存することが示唆される．また，$\theta = L_{\text{train}}$ とした場合，周波数帯はヘッド次元の最大インデックス付近に形成される．

4.2　実証分析：いつ出来るのか？

周波数帯が学習過程のどの段階で形成されるのかを明らかにするため，学習中の各エポックにおける周波数帯の挙動を分析した．図 3 は，$L_{\text{train}} = 512$，$\theta = 10{,}000$ としたモデルにおける，各エポックでのキーの 2 ノルムを示す．エポック 1 では周波数帯は未形成であり分布は散在しているが，初期収束が進むエポック 3 または 4 において明確な周波数帯が出

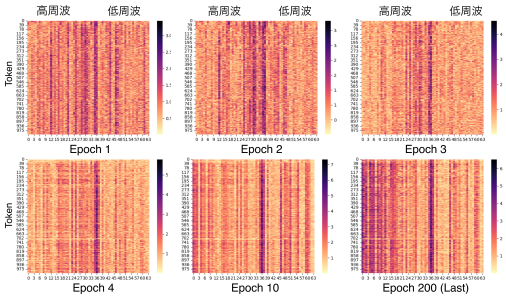

図 3　各エポックにおける $L2$-ノルムの推移．縦軸は系列長，横軸は RoPE の各回転ペア（$i \in \{0, 1, \ldots, d/2\}$）．

現する．その後，この周波数帯は学習の進行に伴っても崩れることなく，最終エポックまで維持される．すなわち，周波数帯は学習初期に形成され，学習全体を通じて安定に保たれる．

4.3　理論分析：どのように出来るのか？

これまでに観測した周波数帯の形成を理論的に説明するため，「学習時の最大系列長 L_{train} と基底 θ が与えられたとき，位置方向分散を最大化する RoPE 回転ペアが学習中に選択され，周波数帯が形成される」という仮説を立てる．この仮説を検証するため，L_{train} と θ のみを用いて，学習中にエネルギーが集中しやすい RoPE 回転ペアを予測し実際の周波数帯存在次元と比較する．

目的　議論を単純化するため，$\omega = \theta^{-2j/d}$ とし，RoPE 回転の 1 次元成分 $\cos(m\omega)$ に着目する．トークン位置 $m \in [0, L_{\text{train}}]$ を動かしたときの $\cos(m\omega)$ の値のばらつきを評価するため，$\cos(m\omega)$ の分散 $V(x)$ が最大となる周波数を求める．

$$V(x) \;:=\; \mathrm{Var}_{m \sim \mathrm{Unif}[0, L_{\text{train}}]}\big[\cos(m\omega)\big], \quad x := \omega L_{\text{train}},$$

Step 1　このときの期待値は次のように計算できる．

$$\mathbb{E}[\cos(m\omega)] = \frac{\sin x}{x}, \quad \mathbb{E}[\cos^2(m\omega)] = \frac{1}{2} + \frac{\sin(2x)}{4x}$$

これから分散は次のように計算できる．

$$V(x) = \frac{1}{2} + \frac{\sin(2x)}{4x} - \left(\frac{\sin x}{x}\right)^2$$

Step 2　分散 $V(x)$ を最大化する周波数を求めるため，$V(x)$ を微分し，停留条件 $V'(x) = 0$ を満たす点を導出する．$V'(x) = 0$ を解くと，$x^\star$ は $x^\star \approx 3.657210$ であり，このとき $V(x)$ は最大値となる．[1)]

Step 3　$x^\star$ に対応する角周波数 $j^\star$ が予測される周波数帯存在次元となる．$j^\star$ は $\theta^{-2j/d} = x^\star / L_{\text{train}}$ から求

1)　$V(x)$ は $x > 0$ において単峰性を持つため，得られる正の解のうち最小のものが $V(x)$ の大域的最大値を与える．

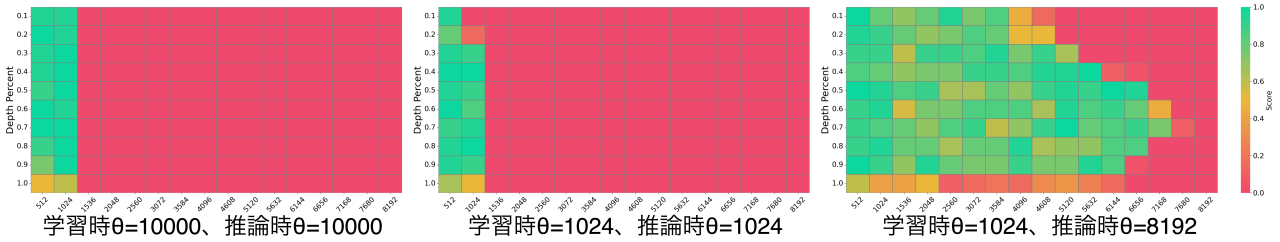

図 4 Needle-in-a-Haystack の結果. 下二段が提案手法である．推論時に θ を大きくすると外挿性能が向上する．

められ，次のように計算できる．

$$j^\star \approx \frac{d}{2}\log_\theta\left(\frac{L_{\text{train}}}{x^\star}\right), \quad x^\star \approx 3.657210$$

$j^\star$ は最も近い整数に丸め，対応する RoPE 次元は $(2j^\star, 2j^\star+1)$ である．この時，基底 θ，ヘッドの次元数 d，最大系列長 L_{train} はモデルによって異なる．

Step 4 3 節における各モデルにおける $j^\star$ を計算した結果を表 1 に示す．周波数帯存在次元 i_{band} は $j^\star$ と $i_{\text{band}} \approx c \times j^\star$（$c \approx 1.1$）の関係を示し，予測値 $c \times j^\star$ とほぼ一致する．モデル間のわずかな差異は，クエリ分布の違いに起因すると考えられる．以上より，周波数帯は L_{train} と θ で予測可能であり，我々の仮説は成り立つと考える．

✓ **結論** 周波数帯は，学習初期に位置方向分散を最大化する回転ペアが選択されることで形成され，その次元は L_{train} と θ によって決定される．$\theta = L_{\text{train}}$ の時，周波数帯は最低周波数付近に位置する．

5 $\theta = L_{\text{train}}$ は外挿を可能にする

3，4 節の結論から，$\theta = L_{\text{train}}$ は RoPE を全次元で効率的に活用できる設定であると考えられる．本節では，言語モデルにおけるその有効性を検証する．

実験設定 最大系列長 $L_{\text{train}} = 1024$ トークンで，1B 規模の Transformer デコーダを事前学習した．$\theta = 10000$ の通常 RoPE をベースラインとし，学習時 $\theta = 1024$ の提案手法，および推論時のみ $\theta = 8192$ へ拡張した場合を，再学習なしで比較する．その他実験設定は付録 A に記載している．Social IQA[13], PIQA[14], CommonsenseQA [15], HellaSwag[16] といった代表的なベンチマークで評価を行った．外挿性能の評価として，Needle-in-a-Haystack[17] による評価も行った．

実験結果 図 4 から，$\theta = 1024$ で学習し推論時も $\theta = 1024$ であるモデルは，ベースラインと比較すると内挿性能が若干悪いことがわかった．一方で，推論時に $\theta = 8192$ に変更すると，外挿性能が向上することがわかる．表 2 より，$\theta = 1024$ で学習したモデルは，ベースラインと同等の内挿性能を維持しつつ，推論時に θ を 8192 へ拡張しても下流タスクの性能が大きく下がることはなかった．実験結果から，学習時に θ を最大系列長 L_{train} に設定し，推論時にのみ θ を拡張することで，再学習なしに外挿性能を向上できることがわかった．

表 2 下流タスクの結果．提案手法は灰色の箇所である．

基底 θ		下流タスク			
学習時	推論時	SIQA	PIQA	CQA	HellaSwag
10000	10000	43.90	**70.78**	32.35	**45.00**
1024	1024	43.96	69.58	**33.66**	44.80
1024	8192	**44.16**	68.71	32.92	44.91

何故外挿可能なのか？ $\theta = L_{\text{train}}$ は，長さ制御可能な位置符号化 (LRPE) [18] と対応する設定である．LRPE では推論時に θ を変えることで、出力長を制御できる [19]．よって、本手法も LRPE と同様に，系列長に関する情報を θ を介して内部表現に取り込んでいる可能性があり，推論時の θ の変更が「有効長」の制御として働き，外挿性能の改善が生じた，と解釈できる．さらにこの解釈に立つと，RoPE の回転角が系列内でちょうど一周分の 2π に達する回転ペアが周波数帯として現れている可能性がある．

6 おわりに

本研究では周波数帯に着目し，RoPE における位置情報の有効利用を分析した．低周波成分は性能に寄与せず，周波数帯が有効・無効を分ける境界として機能し，その位置は基底 θ と学習時最大系列長 L_{train} によって決定されることを明らかにした．そして，RoPE の全周波成分が性能に寄与するよう $\theta = L_{\text{train}}$ とした場合，再学習なしに下流タスクの性能を下げることなく外挿性能が向上したことを明らかにした．以上から，従来の基底拡大戦略は RoPE の外挿性能を制限していたと結論づける．本研究の主要な貢献は，RoPE の基底調整のみで外挿が可能なことを実証および理論の両面から明らかにし，基底設定に関する新たな指針を与えた点にある．

参考文献

[1] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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.

[3] Hugo Touvron, et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[4] Aaron Grattafiori, et al. The llama 3 herd of models, 2024.

[5] An Yang, et al. Qwen3 technical report. **arXiv preprint arXiv:2505.09388**, 2025.

[6] Gemma Team. Gemma: Open models based on gemini research and technology, 2024.

[7] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In **The Twelfth International Conference on Learning Representations**, 2024.

[8] Federico Barbero, Alex Vitvitskyi, Christos Perivolaropoulos, Razvan Pascanu, and Petar Veličković. Round and round we go! what makes rotary positional encodings useful?, 2024.

[9] Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. Longrope: Extending llm context window beyond 2 million tokens, 2024.

[10] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In **Thirty-seventh Conference on Neural Information Processing Systems**, 2023.

[11] Meta. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/, 2024. Accessed: 2025-05-08.

[12] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In **International Conference on Learning Representations**, 2017.

[13] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[14] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In **AAAI Conference on Artificial Intelligence**, 2019.

[15] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[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] gkamradt. Llmtest needle in a haystack – pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023. Accessed: 2025-12-31.

[18] Sho Takase and Naoaki Okazaki. Positional encoding to control output sequence length. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 3999–4004, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[19] Yui Oka, Katsuki Chousa, Katsuhito Sudoh, and Satoshi Nakamura. Incorporating noisy length constraints into transformer with length-aware positional encodings. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, **Proceedings of the 28th International Conference on Computational Linguistics**, pp. 3580–3585, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[20] Alexei Baevski and Michael Auli. Adaptive input representations for neural language modeling. In **International Conference on Learning Representations**, 2019.

[21] Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In **International Conference on Learning Representations**, 2022.

[22] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of NAACL-HLT 2019: Demonstrations**, 2019.

[23] ofirpress. Attention with linear biases (alibi). https://github.com/ofirpress/attention_with_linear_biases, 2022. Accessed: 2025-12-31.

[24] 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, 2023.

[25] Yui Oka, Taku Hasegawa, Kyosuke Nishida, and Kuniko Saito. Wavelet-based positional representation for long context. In **The Thirteenth International Conference on Learning Representations**, 2025.

A　実験設定

節 4.1 の実験設定　Transformer ベースの言語モデル [20] を使って比較評価を行なった．データセットには，WikiText-103[12] を使った．WikiText-103 データセットは，1 億 300 万トークン以上の英語版 Wikipedia の記事から構成される．単語埋め込みの次元数 d_{model} は 1024，ヘッド数 n は 8，ヘッドの次元数 d は 128，レイヤー数は 16 である．外挿実験で用いたパラメタ設定は ALiBi の原論文 [21] と同じものを使った．学習エポック数は 205，バッチサイズは 9216 である．学習率は 1.0 とし，学習の過程で 16000step 毎に 1e-7 ずつ更新した．実装には文献 [21] が提供する fairseq[22] ベースのコード [23] を用い，ハイパーパラメタは全てにおいて文献 [21] と同じ設定とした．

節 5 の実験設定　RoPE と FlashAttention を用いた OLMo モデルベースの Transformer デコーダを学習した．隠れ次元数は 2048，層数 16，ヘッド数 16 とした．学習時の最大系列長 L_{train} は 1024 トークンとした．語彙サイズは 50,280 で，GPT-NeoX/OLMo Dolma v1.5 トークナイザを用いた．最適化には AdamW（$(\beta_1, \beta_2) = (0.9, 0.95)$）を用いた．学習率は最大 6×10^{-4} とし，10,000 ステップのウォームアップ後に cosine スケジュールで減衰させ，最終的に最大値の 0.1 倍まで低下させた．バッチサイズは 512 とした．事前学習には英語 C4 コーパス [24] を用い，1 エポック分の学習を行った．

B　$\theta = L_{train}$ の時の周波数帯分析

節 4.1 における $\theta = L_{train}$ の場合の追加分析として，節 3 と同じ分析を行なった．最大系列長 $L_{train} = 512$ の時の，帯存在次元，p-RoPE のパープレキシティを示す．$L_{train} = 512$ と固定の時，帯存在次元は θ の値が大きくなればなるほど，高周波次元に寄ることがわかった．さらに，p-RoPE の結果から，$\theta = L_{train}$ の時，周波数帯が最も低周波の次元にあるため，低周波次元を NoPE に置き換えると性能低下し (r=0.90 の時)，全ての次元の RoPE が性能に寄与していることがわかった．しかし，$\theta = 10000, 500000$ の時は，r=0.50 までは低周波を NoPE に置き換えても性能低下が見られなかった．これらの結果から，大規模な言語モデルだけでなく，小規模なモデルでも周波数帯は観測され，その周波数帯が θ の値に依存するという特徴は変わらないことがわかった．

表 3　節 3 と同じ分析を節 4.1 の $\theta = L_{train}$ のモデルに行なった分析結果．最大系列長 $L_{train} = 512$ の場合のみ示す．

基底 θ	帯存在次元		p-RoPE のパープレキシティ				
	i_{band}	$i_{band}/\frac{d}{2}$	r=1.0	r=0.90	r=0.75	r=0.50	r=0.25
512	60.5	0.94	19.58	20.18	24.28	35.11	98.26
10000	30.12	0.47	19.39	19.39	19.39	22.71	63.59
500000	17.00	0.26	19.35	19.35	19.35	19.35	34.46

C　コンテキスト拡張のための再学習を行なった場合の比較

実験設定　外挿性能の向上に有効とされる位置符号化手法として，Attention with Linear Biases（ALiBi）[21]，および Wavelet Positional Representation [25] との比較実験を行った．実験設定は節 4.1 と同一とし，学習時の最大系列長は $L_{\text{train}} = 512$ に統一した．外挿性能は，WikiText-103 データセット [12] のテストセットにおけるパープレキシティで評価した．また，従来の LLM におけるコンテキスト拡張と同様に，位置補間を用いた微調整学習も行った．このとき，最大系列長は $L_{\text{train}} = 1512$ とし，位置補間手法には Qwen[5] などで用いられている YaRN [7] を採用した．

実験結果　表 4 に実験結果を示す．外挿性能に着目すると，最も高い性能を示したのは Wavelet 位置符号化であった．RoPE 系手法の中では，提案手法である $\theta = L_{\text{train}}$ 設定が最も優れた外挿性能を示した．一方で，$\theta = L_{\text{train}}$ 設定は，通常の θ 設定と比較して内挿性能が低下することが確認され，θ の値に応じた内挿性能と外挿性能のトレードオフが存在することがわかる．さらに，位置補間手法 YaRN を用いたコンテキスト拡張を行った場合にも，このトレードオフは維持されることが確認された．コンテキスト拡張により，RoPE は学習時の系列長（1512）を超える長さにおいても一定の外挿性能を示すようになるが，その中でも $\theta = L_{\text{train}}$ 設定が最も高い外挿性能を達成している．

表 4　パープレキシティの結果を示す．YaRN はコンテキスト拡張時に適用される位置補間手法である．

	最大系列長 L_{train}		基底 θ		推論時の系列長 L			
	事前学習	微調整	学習	推論	512	1512	2512	3512
WaveletRPE	512	-	-	-	**19.20**	**17.99**	**18.00**	**18.21**
ALiBi	512	-	-	-	19.69	18.53	18.40	18.43
RoPE	512	-	10000	10000	19.39	43.63	84.45	>100
	512	-	500000	500000	19.35	40.39	77.90	>100
	512	-	1000000	1000000	19.35	37.94	74.26	>100
	512	-	512	512	19.58	21.19	24.20	27.42
	512	-	512	1512	20.02	19.09	21.40	24.00
	512	-	512	3512	21.28	20.27	20.37	23.00
RoPE+YaRN	512	1512	10000+YaRN	10000+YaRN	19.10	17.84	17.75	18.37
	512	1512	500000+YaRN	500000+YaRN	19.14	17.89	18.83	18.34
	512	1512	1000000+YaRN	1000000+YaRN	19.07	17.76	17.81	18.72
	512	1512	1512+YaRN	1512+YaRN	19.62	17.78	17.56	17.65
	512	1512	1512+YaRN	3512+YaRN	19.38	17.99	17.66	17.64