Frontier Language Models Struggle to Copy: Text Can Be Better Viewed in 2D

Anonymous Authors[1]

Abstract

While large language models (LLMs) can solve advanced reasoning problems in seconds, we show that even frontier models fail to perform a much simpler operation: exactly copying an input string that lies well within their context windows. We attribute this failure to positional encodings in Transformer architectures, whose inductive bias favors copying through a shortcut based on matching local contexts rather than carefully locating the corresponding input positions. To address this issue, we introduce 2D-RoPE, which organizes text into a 2D grid rather than a 1D sequence and assigns each token a row ID and a column ID. Under this view, copying becomes simply retrieving input tokens from the same column, which makes the task easy to learn. In synthetic copy experiments, shallow Transformers with 2D-RoPE achieve perfect copying at input lengths hundreds of times longer than those seen during training, whereas standard positional encodings fall far behind. We further pretrain 2D-RoPE language models on DCLM at scales up to 1.4B parameters and show that 2D-RoPE substantially improves performance on copy and NIAH tasks. Overall, our results suggest that viewing text in 2D can benefit language modeling, and we hope this encourages future work to further explore the potential of 2D positional encodings.

1. Introduction

While large language models (LLMs) have achieved tremendous success on a wide range of complex reasoning problems, we show that they still struggle with a seemingly simple task: given a string as input, reproduce the exact same string as output, possibly with minor adaptations to match a required output format. We refer to this as the *copy* task and specifically consider the following two variants:

- **Binary Copy.** We give the model a simple string consisting of two types of tokens (such as 0 and 1) and the model is instructed to output the exact same sequence of tokens.
- **Python List Conversion.** We generate a comma-separated list of data points from synthetic physical experiments, feed it to the LLM, and ask the LLM to convert the data into a Python list that can be used to generate plotting code for the data.

Although solving these copy tasks may not require the same level of intelligence as an LLM, it is arguably a fundamental capability that any intelligent agent should possess. For example, an agent may need to copy parameters from a configuration file for function calling, or organize unstructured user input into a specific format for downstream processing. After all, if current LLMs are already intelligent enough to achieve gold-medal performance in Olympiad-level mathematics and competitive programming (OpenAI, 2025; Luong & Lockhart, 2025; Lin & Cheng, 2025), why shouldn't we expect them to perform this very basic copying task perfectly?

However, as shown in Figure 1, all the frontier LLMs we evaluate, including GPT-5.5, Gemini 3.1 Pro, and DeepSeek V4 Pro, fail on a substantial fraction of input strings that are well within their context lengths. Their copying accuracy drops further on longer inputs, often falling well below 50%.

Understanding the Failure of Copying. Our experimental design is motivated by the following conjecture: LLMs may not copy a string by directly using the absolute index i to retrieve the i-th character from the input. Instead, copying may be implemented in a way closer to the induction head mechanism (Olsson et al., 2022; Chen et al., 2024b): the model searches for a previous occurrence of a similar local context and then predicts the token that appears right after this occurrence.

Under this view, repeated substrings create ambiguous local matches and can therefore interfere with exact copying. In-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

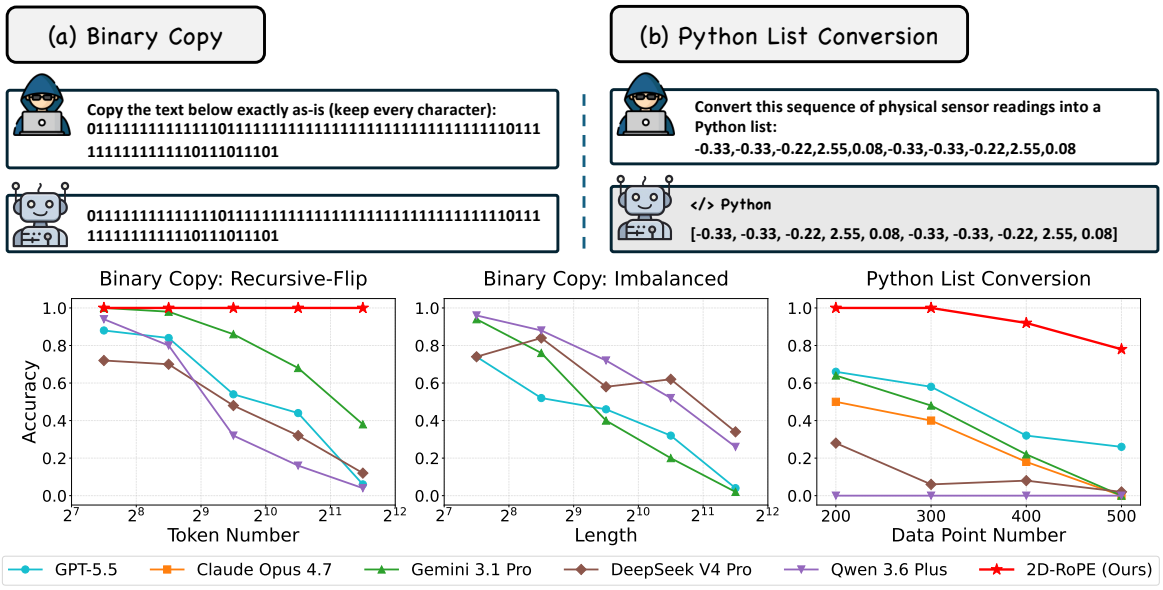


Figure 1. **Top-row:** Examples of Binary Copy and Python list conversion prompts. **Bottom-row:** Frontier language models struggle to copy strings perfectly, while our 2D-RoPE models achieve significantly better performance. See Section 3.1 for details on the benchmark construction.

deed, in our experiments, the tested strings are not arbitrary but are deliberately chosen to contain repeated substrings. For example, in the Python List Conversion task, part of the data is generated from periodic sine or triangular waves. Our further experiments in Figure 2 show that copying accuracy is negatively correlated with the degree of repetition in the input strings.

In Section 3, we develop expressivity theory and mechanistic interpretability results to support this conjecture, and show that the copying failure is closely tied to the positional encoding used in Transformers. Although RoPE has appealing properties such as shift-invariance, it does not provide a sufficiently strong inductive bias for the model to align the source and target tokens by their relative positions. As a result, the learned attention patterns are not cleanly aligned with the corresponding source positions, but are instead easily attracted by locally similar substrings.

Our Method: 2D Positional Encoding. This analysis points to a natural design goal: a positional encoding with a stronger inductive bias towards copying from the corresponding position in the input. To this end, a key observation is that the copy task has an implicit 2D structure. When copying an input string $x_1, \ldots, x_n$ into an output string $y_1, \ldots, y_n$, if we only view the whole context as a 1D sequence, for example $x_1, \ldots, x_n, <\backslash n>, y_1, \ldots, y_n$, then producing an output token y_k requires the model to locate the corresponding input token x_k in this 1D sequence through careful arithmetic over positions. To do so, the model may first have to identify the boundary between the input and output, since the relative positions of x_k and y_k depend on the input length n. It is thus no wonder that models may fall back to an easier shortcut such as matching similar substrings. Instead, if we arrange the input and output strings into two rows and use the position k within each string as the column index, then producing y_k reduces to simply retrieving a token from the same column.

Motivated by this key observation, we introduce 2D-RoPE. The construction is inspired by prior uses of 2D positional encodings in vision models that view images as a 2D grid (Jeevan & Sethi, 2022; Heo et al., 2024; Chu et al., 2024). Given any text, 2D-RoPE uses the line break token $<\backslash n>$ (also called the newline token) as the row separator and views the text as a 2D grid, where each token is assigned a 2D position ID pair $(\text{row ID}, \text{column ID})$. RoPE attention is then applied to this 2D grid to encode relative positions along both axes. In this view, copying a string can be implemented by a single attention layer that retrieves tokens from the same column, which is arguably a simpler structure for the model to learn.

Empirically, 2D-RoPE leads to strong length generalization on the copy task. Here, length generalization refers to the ability to generalize from short training sequences to longer test sequences. In synthetic copy experiments, we train 2D-RoPE models on the binary copy task and show that one-layer models achieve perfect copying with **up to** $1000\times$ **length generalization**, while 12-layer models maintain the perfect performance up to $100\times$ length generalization. Theoretically, we also prove that one-layer Transformer with 2D-RoPE can provably learn copying with length generalization.

Beyond this synthetic setting, many natural documents also contain explicit 2D structures induced by line breaks, such as lists, tables, and code, which today's LLMs do not directly encode in their positional encodings. We therefore pretrain 2D-RoPE models ranging from 0.3B to 1.3B parameters on DCLM (Li et al., 2024). After finetuning, these models exhibit length generalization on copy tasks and significantly improve the long-context retrieval performance on the Needle In A Haystack (NIAH) task (Hsieh et al., 2024) across lengths from 512 to 4K.

Finally, since the 2D structure in natural documents may not always be explicitly marked by line breaks, we introduce Auto-2D-RoPE, which learns a data-dependent transformation to automatically assign 2D coordinates to each token. Our experiments show that Auto-2D-RoPE maintains length generalization on binary copy even without line breaks, whereas the original 2D-RoPE does not.

Overall, our results suggest that viewing text in 2D can benefit language modeling, and we hope this encourages future work to further explore the potential of 2D positional encodings.

2. Related Works

Comparison with Jelassi et al. (2024). Jelassi et al. (2024) is most closely related to our work, which studies synthetic copy tasks and gives a two-layer Transformer construction based on context matching. Our focus is different in several aspects. First, we show that frontier LLMs fail to copy user-provided strings exactly, even within their context lengths. Second, their construction and analysis focus on the uniform data setting and do not imply perfect copying for arbitrary strings with repeated substrings or provide the length-generalization guarantee considered in our work. Finally, their construction is context-based, whereas our results suggest that repeated substrings make context matching unreliable, and further, we propose 2D-RoPE to solve copy. See Appendix B for additional related works.

3. Language Models Fail to Copy

In this section, we first introduce our copy benchmarks in detail, and then present our analysis to diagnose the failure of LLMs on the copy tasks.

3.1. Our Copy Benchmarks

For a testbed of LLMs' copy capabilities, we construct the following copy tasks. See example prompts in Figure 1 and full prompt templates used in our experiments in Appendix M.

Binary Copy Task. In binary copy, we fix a vocabulary $\mathcal{V}$ of size $|\mathcal{V}| = 2$, which means that the input string s only consists of two types of tokens. Given an input string s, the model is required to output an exact copy of s. For example, the input string can be $s =$ "011010101010100", with the vocabulary fixed to $\mathcal{V} = \{0, 1\}$. We generate such input strings from three different distributions:

- **Uniform Generation:** Each token is sampled independently and uniformly from $\mathcal{V}$.
- **Imbalanced Generation:** We fix a probability set $p \in \{0.05, 0.15, 0.3, 0.5, 0.7, 0.85, 0.95\}$. For each string, we first sample p uniformly from this set, and then each input token is sampled as the first token in $\mathcal{V}$ with probability p and the second token with probability $1 - p$.
- **Recursive-Flip Generation:** We initialize s with a single token sampled uniformly from $\mathcal{V}$. Then we iteratively update $s \leftarrow s + c + s$ for several rounds, where c is a randomly sampled token from $\mathcal{V}$ in each round. After K rounds, we obtain a string s with length $2^{K+1} - 1$, which contains a complicated recursive structure as well as many random flips. When we need to generate a string with an arbitrary length n, we simply truncate s to its first n tokens.

Python List Conversion Task. We generate a comma-separated list of data points from synthetic physical experiments and ask the LLM to convert the data into a Python list. This mimics the scenario where an agent needs to copy user inputs into a specific format for downstream processing with Python code, such as visualizing the data. Our evaluation includes 9 synthetic experiments in total, each data-point list is generated randomly from one of the 9 experiments. These experiments can be further categorized into three types: smooth periodic oscillations, piecewise or reset-driven oscillations, and nonlinear or pulse-like signal responses. See Appendix C.2 for details.

3.2. Evaluation of Frontier LLMs

Figure 1 shows that frontier LLMs fails to perform perfect copying on both binary copy and Python list conversion tasks. For binary copy, we test the models on Recursive-Flip and Imbalanced data. For each length interval $[2^k, 2^{k+1} - 1)$ with $k \in \{7, 8, 9, 10, 11\}$, we generate 50 samples with lengths sampled uniformly from that interval, and report the average copying accuracy within each interval. For Recursive-Flip, the vocabulary is either {"0", "1"} or {"000", "111"}, depending on whether the model's tokenizer tends to merge consecutive 01 characters into a single token. For Imbalanced data, we fix the same vocabulary for all models: {" a", " b"}. See details of the evaluation in Appendix C. For the Python List Conversion task, we evaluate four list lengths, 200, 300, 400, 500, with 50 samples for each length. See details of the evaluation in Appendix C.1.

For Recursive-Flip, the vocabulary is either {"0", "1"} or {"000", "111"}, depending on whether the model's tok-

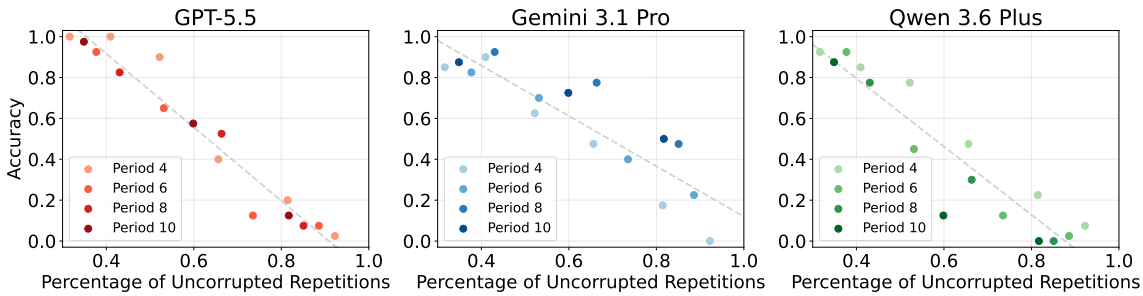


Figure 2. Repeat-structure copy test for frontier LLMs. Accuracy decreases as repetitions are preserved more strongly, suggesting that repetitions make exact copying harder; see Appendix C.3 for details.

enizer tends to merge consecutive 01 characters into a single token. For Imbalanced data, we fix the same vocabulary for all models: {“ a”, “ b”}.

There are two notable features in the results. First, as the input length increases, the accuracy of all models decreases. This suggests that the models may only have learned an ad-hoc copy rule that works for some specific input length range, rather than a rule that can be generalized to all lengths. Throughout the paper, we use the term *length generalization* to refer to the ability of a model to perform well on inputs with lengths beyond the training range.

Second, frontier LLMs perform especially poorly on strings with many repeated substrings. To see this, we sample a short binary base string, repeat it multiple times to generate a fixed-length long string, and independently flip each token with probability $1 - p$. As we increase p from 0 to 1, each repeated substring becomes less likely to be corrupted, and thus the input contains more repeated substrings. As shown in Figure 2, the copying accuracy consistently decreases as the number of repeated substrings increases in this way. Details of this experiments can be found in Appendix C.3

3.3. Why Do LLMs Fail to Copy?

Motivated by the above observations, we now state our conjecture on why frontier LLMs fail to copy. There are two natural algorithms for solving the copy task:

- *Position-based algorithm*: to output the k-th output token y_k, derive some representation of the index k and use it to retrieve the corresponding input token x_k.
- *Context-based algorithm*: first identify the local context around the current output token y_k, such as $y_{k-3}y_{k-2}y_{k-1}$, and then find a similar local context in the input, and then copy the token associated with that matched context.

The position-based algorithm sounds the most natural, but implementing it in a Transformer requires careful arithmetic over positions, since the relative position between x_k and y_k depends on the input length n, which may vary across input strings. Without handling this dependence properly, the model may learn to copy only for some input lengths but not others.

The context-based algorithm is more flexible and can be implemented with a few attention layers, but it is NOT a correct algorithm in general: if there are many repeated substrings in the input, the context-based algorithm may not be able to locate the correct input position.

It is clear from the imperfect copying accuracy that LLMs do not exactly implement either of the two algorithms above, but what do they do instead? We conjecture that LLMs instead implement a mixture of the two algorithms. In other words, LLMs exhibit an inductive bias towards a mixture of position-based and context-based copying algorithms. This conjecture is supproted by the two failure modes: copying accuracy decreases as either the input length or the degree of repetition increases.

3.4. Diagnosis via Expressivity Theory and Synthetic Experiments

To further support our conjecture and understand why the inductive bias does not favor the position-based algorithm exactly, we analyze the problem via expressivity theory and synthetic experiments.

3.4.1. Evidence from Expressivity Theory

First of all, we ask: can Transformers express the position-based copying algorithm? The answer is mixed. Given the apparent simplicity of the copy task, one might expect a single Transformer layer to be expressive enough to solve it. This expectation is natural: single attention layers are closely related to associative recall (Ramsauer et al., 2021; Bricken & Pehlevan, 2021; Smart et al., 2025), and positional encodings can in principle support retrieval tokens at a fixed relative offset (Vaswani et al., 2017; Weiss et al., 2021). Perhaps surprisingly, this is not the case for binary copy. We show that even for input strings with length at most 5, one-layer Transformers cannot exactly perform copying.

Theorem 3.1. *No one-layer Transformer with shift-*

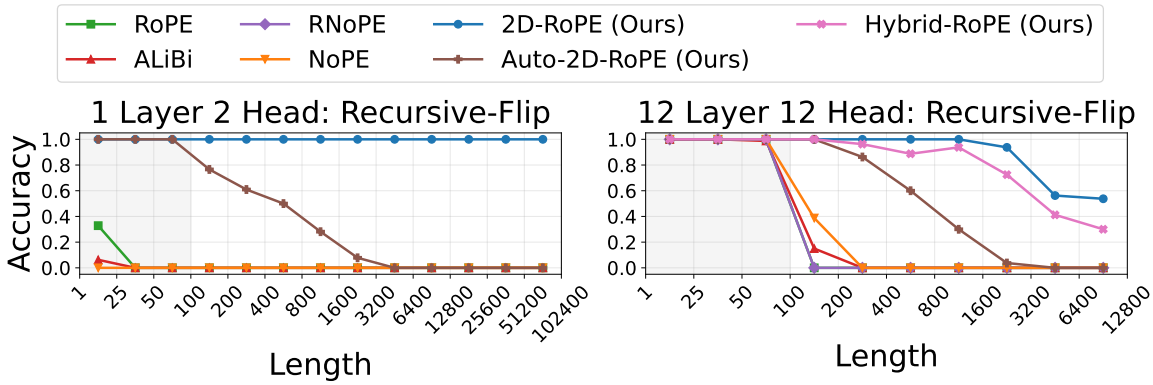


Figure 3. Length generalization on the Recursive-Flip distribution. All models are trained on Imbalanced strings with lengths from 1 to 100, so Recursive-Flip is an out-of-distribution. Full results on all test distributions are shown in Appendix D, Figure 6.

invariant positional encoding can perform the binary copy task for all input strings with length $n \leq 5$.

Note that shift-invariant positional encodings cover a wide range of positional encodings, including RoPE, NoPE, ALiBi, etc. This follows from a simple proof by contradiction. Intuitively, one-layer transformer fails to copy since copy requires retrieval at a length-dependent offset rather than a fixed offset. Thus, the failure of one-layer models reflects a lack of capability for processing length-dependent tasks. The formal statement of this theorem and its proof are in Appendix J.

Beyond one layer, we show that two-layer Transformers with RoPE can express binary copy.

Theorem 3.2. *There exists a constant embedding dimension* $d \geq 1$ *such that for any norm constraint* $\rho > 10^4$*, there exist RoPE frequencies* $\boldsymbol{\beta}$ *and a parameter vector* $\boldsymbol{\theta}$ *with* $\|\boldsymbol{\theta}\|_2 \leq \rho$ *such that the Transformer with embedding dimension* d*, RoPE frequencies* $\boldsymbol{\beta}$*, parameters* $\boldsymbol{\theta}$ *can perfectly perform copying for all lengths* $1 \leq L \leq O(\rho^2 / \log \rho)$.

The key intuition here is that with multiple layers, the model can construct intermediate positional features and use them to align each output position with its corresponding input position. The proof of this theorem is in Section K.

Putting these two theorems together, we can conclude that: (1) the inductive bias induced by standard positional encodings, including RoPE, are not strong enough to support perfect copying for one-layer Transformers; (2) Transformers with multiple layers can express the position-based copying algorithm, suggesting that training, rather than expressivity alone, biases Transformers away from position-based algorithms.

3.4.2. Synthetic Experiments

Setup. We further analyze what Transformers learn through synthetic experiments by training on the binary copy task. Here we consider two settings: Transformers with 1 layer and 2 attention heads, and Transformers with 12 layers and 12 attention heads. For each setting, we consider 4 position encodings: RoPE, ALiBi (Press et al., 2021), NoPE, and RNoPE (Yang et al., 2025b).

Uniform Data. We first train these models on uniform data with lengths from 1 to 100. While they perform perfectly on uniform data, they perform poorly on other data distributions, such as Imbalanced (see Figure 7). This suggests that the models may have learned a context-based algorithm.

Imbalanced Data. Instead, when we train these models on imbalanced data for sufficient iterations, they can perform almost perfectly on imbalanced data as well as the other data distributions we construct. However, all the position encodings mentioned above fail to generalize to longer inputs. See Appendix D for more details of experimental setups.

Attention Patterns. We further show that these models use a mixture of position-based and context-based algorithms when copying. To see this, we analyze the attention patterns of a trained RoPE model on binary copy under the imbalanced distribution. We consider a small model with 2 heads in the first layer and 1 head in the second layer, trained on sequences with lengths from 1 to 100. Figure 4 shows the attention maps of all layers and heads on representative copy examples.

We first inspect the last layer, since exact copying requires each output position to align with its corresponding input position. However, even for a failed sample with length 100, which is within the training range, the learned attention is not a clean position-based alignment. As shown in the zoomed attention map, the final-layer head attends not only to the correct diagonal positions, but also periodically assigns large attention scores to nearby wrong tokens. This

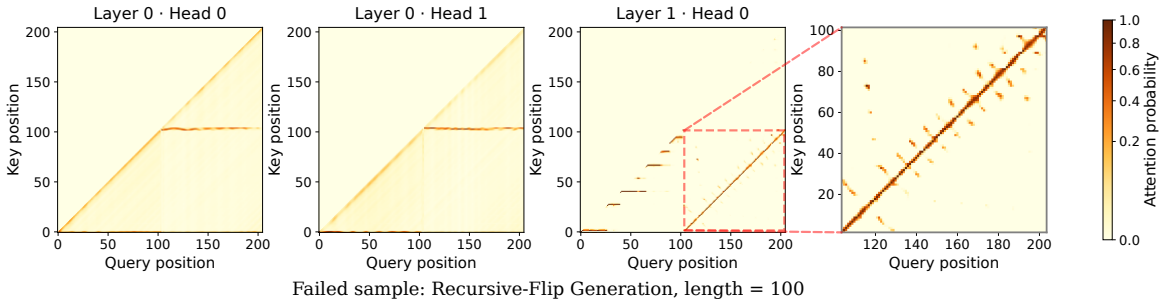


Figure 4. Attention map of a RoPE model on a failed sample with sequence length within training length. In Layer 0, attention heads exhibit both **horizontal patterns** (attending to special tokens before and after the input string) and **diagonal patterns** (semantic/induction-style behavior). In Layer 1, RoPE would attend to wrong tokens near themselves periodically.

suggests that the final copy decision is affected by ambiguous local patterns.

We then inspect Layer 0 to understand where this imperfect alignment comes from. We observe two types of patterns: **horizontal stripes** on special boundary tokens such as [<\n>, <I:>, <O:>], which behave like locator heads and provide position-related signals; and **diagonal patterns**, which resemble local-context or induction-style behavior (Olsson et al., 2022; Singh et al., 2024; Jelassi et al., 2024). Together, these attention maps suggest that the trained RoPE model does not implement a purely position-based copy rule. Instead, it combines locator-style positional signals with context-based matching, and this mixture can be confused by repeated local structures. We provide further discussion beyond the training length in Appendix E.

4. 2D Positional Information Encoding with 2D-RoPE

Given that current LLMs exhibit inductive bias towards a mixture of position-based and context-based algorithms when copying and can fail, a natural question is whether we can design a positional encoding that explicitly favors the position-based algorithm. In this section, we propose a novel architecture based on 2D-RoPE (Section 4.1). We first show the superior performance of 2D-RoPE in copying (Section 4.2). For readers who are interested in theory, we also present a theoretical analysis of 2D-RoPE (Section L) both in expressivity and global minima.

4.1. 2D-RoPE Architecture

2D Position ID Generation. The key idea of 2D-RoPE is to assign each token a pair of position IDs instead of a single position ID. We use the line break token <\n> as a row separator. Let the 2D position of token i be (r_i, c_i), where r_i is the row index and c_i is the column index within that row. We set the first token position to $(r_1, c_1) = (0, 0)$. For each subsequent token $i > 1$,

$$(r_i, c_i) = \begin{cases} (r_{i-1},\ c_{i-1} + 1), & i-1 \text{ is not line break}, \\ (r_{i-1} + 1,\ 0), & i-1 \text{ is a line break}. \end{cases}$$

Thus, tokens on the same line share the same row index, and their column index records the offset within the line. In a copy task, this representation makes the source token x_k and the target token y_k share the same column index, so copying can be reduced to retrival from the same column. This assumes that the input and output are separated by a line break. For the case where the separator is not a line break, we propose Auto 2D-RoPE to automatically identify the separator (see Appendix H).

From 1D RoPE to 2D-RoPE. Standard RoPE (Su et al., 2024) applies a positional rotation to the query and key vectors according to the 1D token position. In 2D-RoPE, we apply the same idea separately to the row and column coordinates. Concretely, we split the head dimension into two equal-sized parts: one encodes relative row positions, and the other encodes relative column positions. This gives the model direct access to both row differences and column differences. For copy, the column coordinate is especially useful, because source and target tokens that should be aligned have the same column ID, although their 1D distance depends on the input length.

We now give the formal definition. Let d be the head dimension, divisible by 4. For an input $\boldsymbol{X} \in \mathbb{R}^{T \times d}$, a 2D-RoPE attention layer is defined as

$$\mathsf{Attn}(\boldsymbol{X}) = \mathcal{S}\left(\mathrm{MASK}\left(R(\boldsymbol{XQ})\big(R(\boldsymbol{XK})\big)^\top\right)\right)\boldsymbol{XV},$$

where $\mathcal{S}$ is the row-wise softmax, MASK is the causal attention mask, and $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ are trainable projection matrices.

It remains to define the rotation function R. For any matrix $\boldsymbol{Z} \in \mathbb{R}^{T \times d}$, the i-th row $\boldsymbol{Z}_i$ is mapped to $\boldsymbol{Z}_i R_{r_i, c_i}^\top$, where $\boldsymbol{R}_{r_i, c_i} = \mathrm{diag}\big(\boldsymbol{R}_1(r_i), \ldots, \boldsymbol{R}_{d/4}(r_i), \boldsymbol{R}_1(c_i), \ldots, \boldsymbol{R}_{d/4}(c_i)\big)$.

Each block $\boldsymbol{R}_j(t) = \begin{bmatrix} \cos(t\beta_j) & -\sin(t\beta_j) \\ \sin(t\beta_j) & \cos(t\beta_j) \end{bmatrix}$ is the usual RoPE rotation matrix, where β_j is the frequency parameter. In our synthetic experiments in Section 3.4, we use $\beta_j = \theta^{-4j/d}$ where $\theta = 100$. In our LLM experiments in Section 5, we set θ to 1000. For the original RoPE architecture, see Appendix I. We also discuss differences from prior vision 2D-RoPE variants in Appendix F.

4.2. Empirical Results of 2D-RoPE

Results of Binary Copy Test. In Figure 1, the 2D-RoPE model maintains near-perfect accuracy on Recursive-Flip strings, suggesting that explicitly aligning source and target positions provides a more reliable inductive bias for copying than relying on local-context matching alone. The model in the figure has 1.4B parameters and is finetuned on the imbalanced binary copy task.

Results of Python List Conversion Test. In the right subfigure in Figure 1, our fine-tuned 2D-RoPE model in Section 5 is trained only on lists with 50 to 150 data points and periods between 2 and 8. Although the training data are easy, the 2D-RoPE model maintains substantially higher accuracy than all tested LLMs. This result is consistent with the binary copy experiments: when the input contains repeated local patterns, as in periodic or oscillatory sequences, exact copying is difficult for standard position encoding while 2D-RoPE provides an explicit positional alignment between the source data points and their copied output positions.

Results of Synthetic Experiments on Binary Copy. Figure 3 shows that 2D positional structure provides a strong inductive bias for binary copy. 2D-RoPE-based methods exhibit much stronger length generalization than standard positional encoding. In the 12-layer setting, positional encodings with 2D structure substantially improve over standard positional encodings. The advantage is even clearer in the 1-layer setting: the 2D-RoPE model maintains perfect accuracy up to more than $1000\times$ its training context length. This suggests that the row-column alignment introduced by 2D-RoPE directly supports the position-based copy rule needed for systematic extrapolation.

5. Experiments on LLM Pretraining

To validate the effectiveness of 2D-RoPE in empirical language modeling, we first pretrain Transformer language models from scratch and compare 2D-RoPE against RoPE. Then, we finetune our different models on the synthetic binary copy dataset and the Python List Conversion datasets.

Model Configuration. We experiment with the Qwen3 (Yang et al., 2025a) architecture, which is one of the most widely used open-source LLMs. To ensure fair comparison, we keep everything unchanged and only replace RoPE with 2D-RoPE. We experiment with different models sizes (350M, 730M, and 1.4B parameters). See Appendix G.1 for more details.

Pretraining Configuration. The models are pretrained with the DCLM corpus (Li et al., 2024) with 2K context length using the typical optimizer, LR scheduler, and hyperparameters (see Appendix G.2 for more details). We use a data-to-parameter ratio of 20 (i.e., Chinchilla law (Hoffmann et al., 2022)). Furthermore, we experiment with an overtrained setting that is common in state-of-the-art LLMs, where a 730M model is pretrained on 100B token.

Finetuning Configuration. Each model is finetuned on the imbalanced binary copy task (described in Section 3), using 200K examples and 2K context length (1K maximum input/output tokens). For each model, we swept learning rate (LR) values of $\{3 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}\}$ and arrive at 5×10^{-5} because it has best overall performance. See Appendix G.5 for more details.

Hybrid RoPE. In order to show the effectiveness of 2D-RoPE, we construct a Hybrid-RoPE (H-RoPE) model between RoPE and 2D-RoPE, in which RoPE and 2D-RoPE are used alternately across Transformer blocks. H-RoPE serves as an investigation on whether partial use of 2D positional structure is sufficient to improve length generalization in copying.

5.1. LLM Evaluation Details.

We evaluate the models on the following benchmarks. See Appendix G.4 for more details.

Common-Sense Reasoning (CSR). A set of eight zero-shot question-answering tasks that measure a model's general language modeling abilities.

Needle-in-a-Haystack (NIAH). A set of tasks where the objective is to retrieve a specific "needle" from a document. We use the NIAH (single) tasks from RULER (Hsieh et al., 2024), but with newline characters around the needle.

Binary Copy. We evaluate on the Imbalanced and Recursive-Flip binary copy tasks as described in Section 3.1. The first task is an in-domain copy task (since the models are finetuned on it), and the second task is an out-of-domain task that tests the transferability of our method on copying.

5.2. LLM Results

As shown in Table 1, LLMs with 2D-RoPE have superior length generalization in the binary copy tasks, in both in-domain (Imbalanced) and out-of-domain (Recursive-Flip) settings. This result shows that 2D-RoPE can exhibit its advantage in copying in practical large-scale LLM training.

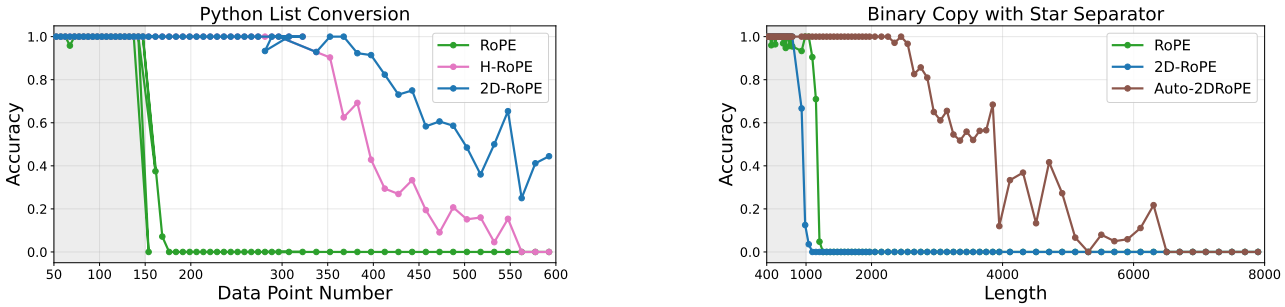


Figure 5. Fine-tuning results on copy-related tasks using 730M models pretrained on 100B DCLM tokens. The gray region indicates the training context length range during finetuning.

Model	Param	PT Data	CSR	NIAH 512	NIAH 1K	NIAH 2K	NIAH 4K	Copy (Imbalanced) 1K	Copy (Imbalanced) 2K	Copy (Imbalanced) 4K	Copy (Imbalanced) 8K	Copy (Recursive-Flip) 1K	Copy (Recursive-Flip) 2K	Copy (Recursive-Flip) 4K	Copy (Recursive-Flip) 8K
RoPE	350M	7B	**41.4**	**93.7**	92.1	87.0	1.9	97.2	14.8	0.0	0.0	14.4	7.8	0.0	0.0
H-RoPE (ours)	350M	7B	41.2	93.1	**98.8**	**94.4**	57.9	**99.6**	57.8	0.0	0.0	**96.9**	**92.2**	35.1	0.0
2D-RoPE (ours)	350M	7B	40.4	91.9	97.9	93.9	**92.8**	97.4	**69.2**	**33.4**	**2.4**	**96.9**	89.3	**56.4**	**36.4**
RoPE	730M	15B	44.0	**100.0**	**98.4**	**96.7**	26.4	97.0	11.0	0.0	0.0	15.5	7.8	1.1	0.0
H-RoPE (ours)	730M	15B	**45.5**	95.9	97.9	90.0	4.5	95.4	11.4	0.0	0.0	14.4	5.8	0.0	0.0
2D-RoPE (ours)	730M	15B	45.2	93.0	98.2	92.4	**91.3**	**99.4**	**74.6**	**12.8**	0.0	**92.8**	**76.7**	**45.7**	**10.9**
RoPE	1.4B	28B	45.9	84.5	87.2	82.3	10.0	99.6	26.8	0.0	0.0	21.6	8.7	0.0	0.0
H-RoPE (ours)	1.4B	28B	**47.2**	93.4	**98.5**	**94.5**	62.3	**100.0**	73.6	3.6	0.0	**100.0**	**98.1**	55.3	5.5
2D-RoPE (ours)	1.4B	28B	46.8	**98.3**	96.4	92.6	**86.0**	**100.0**	**100.0**	**92.0**	**61.8**	**100.0**	**98.1**	**92.6**	**87.3**
RoPE	730M	100B	46.0	76.9	80.9	70.5	0.4	99.6	29.2	0.0	0.0	76.9	80.9	70.5	0.4
H-RoPE (ours)	730M	100B	**46.2**	83.7	89.9	**89.3**	38.3	99.2	52.1	0.0	0.0	83.7	89.9	**89.3**	38.3
2D-RoPE (ours)	730M	100B	**46.2**	**87.3**	**93.1**	88.6	**86.5**	**100.0**	**96.3**	**72.8**	**15.8**	**87.3**	**93.1**	88.6	**86.5**

Table 1. Large-scale pretraining results of 2D-RoPE, H-RoPE, and RoPE. Each model is pretrained on DCLM, and then finetuned on the copying data with imbalanced distribution.

Models with 2D-RoPE (2D-RoPE and H-RoPE) also show better NIAH performance when the context length exceeds the maximum training context length, implying that copying abilities is beneficial to real-world in-context retrieval tasks. Finally, models with 2D-RoPE outperform RoPE models in CSR except for the smallest-scale model, implying that learning copying does not result in degraded general language modeling capabilities at scale.

5.3. Ablation Experiments

Python List Conversion Task. To demonstrate the practical importance of the ability to copy, we finetune the 730M model (overtrained) on the Python List Conversion task (described in Section 3.1). After finetuning, 2D-RoPE achieves perfect generalization to sequences up to 3 times longer than those seen during training. In comparison, H-RoPE also generalizes to about three times the training length, but underperforms 2D-RoPE. Meanwhile, RoPE shows limited length generalization.

Auto 2D-RoPE: Beyond the Line-Break Separator. 2D-RoPE constructs 2D positional coordinates using the line-break token <\n>, so it loses its main advantage when the input contains no explicit line breaks. To reduce this reliance, we introduce Auto 2D-RoPE, which learns to construct 2D position IDs directly from the input sequence. Auto 2D-RoPE updates the 2D position ID of each token through a cumulative product of learnable transformation matrices generated from hidden states. This allows the model to adaptively decide how position IDs should evolve along the sequence.

Auto 2D-RoPE Results. Empirically, when we replace <\n> with another character $*$, standard 2D-RoPE can no longer identify row boundaries and largely loses length generalization, while Auto 2D-RoPE still learns useful 2D positional structure and generalizes to longer sequences, as shown in Figure 5. We provide the full architectural details in Appendix H.

References

Anil, C., Wu, Y., Andreassen, A., Lewkowycz, A., Misra, V., Ramasesh, V., Slone, A., Gur-Ari, G., Dyer, E., and Neyshabur, B. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022.

Ataee Tarzanagh, D., Li, Y., Zhang, X., and Oymak, S. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, 36: 48314–48362, 2023.

Awasthi, P. and Gupta, A. Improving length-generalization in transformers via task hinting. *arXiv preprint arXiv:2310.00726*, 2023.

Bhattamishra, S., Ahuja, K., and Goyal, N. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7096–7116, 2020.

Bisk, Y., Zellers, R., Bras, R. L., Gao, J., and Choi, Y. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Bricken, T. and Pehlevan, C. Attention approximates sparse distributed memory. *Advances in Neural Information Processing Systems*, 34:15301–15315, 2021.

Chang, Y. and Bisk, Y. Language models need inductive biases to count inductively. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chen, L., Peng, B., and Wu, H. Theoretical limitations of multi-layer transformer. *arXiv preprint arXiv:2412.02975*, 2024a.

Chen, S., Sheen, H., Wang, T., and Yang, Z. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 66479–66567. Curran Associates, Inc., 2024b. doi: 10.52202/079017-2127.

Chen, W., Lin, Y., Zhou, Z., Huang, H., Jia, Y., Cao, Z., and Wen, J.-R. Icleval: evaluating in-context learning ability of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10398–10422, 2025.

Chiang, D. and Cholak, P. Overcoming a theoretical limitation of self-attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7654–7664, 2022.

Chu, X., Su, J., Zhang, B., and Shen, C. Visionllama: A unified llama backbone for vision tasks. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019. URL https://arxiv.org/abs/1905.10044.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Fan, C., Schmidt, M., and Thrampoulidis, C. Implicit bias of spectral descent and muon on multiclass separable data. In *High-dimensional Learning Dynamics 2025*, 2025.

Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., and Wang, L. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798, 2023.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Heo, B., Park, S., Han, D., and Yun, S. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pp. 289–305. Springer, 2024.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Cantor, A., et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30040–30053, 2022.

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., Zhang, Y., and Ginsburg, B. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Hu, S., Tu, Y., Han, X., Cui, G., He, C., Zhao, W., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhang, X., Thai, Z. L., Wang, C., Yao, Y., Zhao, C., Zhou, J., Cai, J., Zhai, Z., Ding, N., Jia, C., Zeng, G., dahai li, Liu, Z., and Sun, M. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3X2L2TFr0f.

Huang, X., Yang, A., Bhattamishra, S., Sarrof, Y., Krebs, A., Zhou, H., Nakkiran, P., and Hahn, M. A formal framework for understanding length generalization in transformers. In *The Thirteenth International Conference on Learning Representations*, 2024.

Jeevan, P. and Sethi, A. Resource-efficient hybrid x-formers for vision. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2982–2990, 2022.

Jelassi, S., d'Ascoli, S., Domingo-Enrich, C., Wu, Y., Li, Y., and Charton, F. Length generalization in arithmetic transformers. *arXiv preprint arXiv:2306.15400*, 2023.

Jelassi, S., Brandfonbrener, D., Kakade, S. M., and Malach, E. Repeat after me: Transformers are better than state space models at copying. In *International Conference on Machine Learning*, pp. 21502–21521. PMLR, 2024.

Kazemnejad, A., Padhi, I., Natesan Ramamurthy, K., Das, P., and Reddy, S. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Lin, H. M. and Cheng, H.-T. Gemini achieves gold-medal level at the International Collegiate Programming Contest World Finals. 2025.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Luong, T. and Lockhart, E. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. 2025.

Merrill, W. and Sabharwal, A. The expressive power of transformers with chain of thought. In *The Twelfth International Conference on Learning Representations*, 2024.

Mohamadi, M. A., Li, Z., Wu, L., and Sutherland, D. J. Why do you grok? a theoretical analysis on grokking modular addition. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 35934–35967, 2024.

Morwani, D., Edelman, B. L., Oncescu, C.-A., Zhao, R., and Kakade, S. M. Feature emergence via margin maximization: case studies in algebraic tasks. In *The Twelfth International Conference on Learning Representations*, 2024.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

OpenAI. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*, 2025.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. The lambada dataset: Word prediction requiring a broad discourse context, 2016. URL https://arxiv.org/abs/1606.06031.

Pérez, J., Barceló, P., and Marinkovic, J. Attention is turing-complete. *Journal of Machine Learning Research*, 22 (75):1–35, 2021.

Press, O., Smith, N., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2021.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Adler, T., Kreil, D., Kopp, M. K., et al. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Sanford, C., Hsu, D. J., and Telgarsky, M. Representational strengths and limitations of transformers. *Advances in Neural Information Processing Systems*, 36:36677–36707, 2023.

Singh, A. K., Moskovitz, T., Hill, F., Chan, S. C., and Saxe, A. M. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *Forty-first International Conference on Machine Learning*, 2024.

Smart, M., Bietti, A., and Sengupta, A. M. In-context denoising with one-layer transformers: Connections between attention and associative memory retrieval. *Proceedings of Machine Learning Research*, 267:55950–55971, 2025.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Ji, T., Wu, Y., Yan, H., Gui, T., Zhang, Q., Huang, X.-J., and Wang, X. Length generalization of causal transformers without position encoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14024–14040, 2024a.

Wang, W., Zhang, S., Ren, Y., Duan, Y., Li, T., Liu, S., Hu, M., Chen, Z., Zhang, K., Lu, L., et al. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565, 2024b.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems*, 32, 2019.

Weiss, G., Goldberg, Y., and Yahav, E. Thinking like transformers. In *International Conference on Machine Learning*, pp. 11080–11090. PMLR, 2021.

Wen, K., Dang, X., and Lyu, K. Rnns are not transformers (yet): The key bottleneck on in-context retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xie, S. and Li, Z. Implicit bias of adamw: ℓ_∞-norm constrained optimization. In *International Conference on Machine Learning*, pp. 54488–54510. PMLR, 2024.

Yang, A. and Chiang, D. Counting like transformers: Compiling temporal counting logic into softmax transformers. In *First Conference on Language Modeling*, 2024.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Yang, B., Venkitesh, B., Gnaneshwar, D., Lin, H., Cairuz, D., Blunsom, P., and Locatelli, A. Rope to nope and back again: A new hybrid attention strategy. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Yang, S., Shen, Y., Wen, K., Tan, S., Mishra, M., Ren, L., Panda, R., and Kim, Y. PaTH attention: Position encoding via accumulating householder transformations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=ZBlHEeSvKd.

Yao, S., Peng, B., Papadimitriou, C., and Narasimhan, K. Self-attention networks can process bounded hierarchical languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3770–3785, 2021.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Zhang, C., Zou, D., and Cao, Y. The implicit bias of adam on separable data. *Advances in Neural Information Processing Systems*, 37:23988–24021, 2024.

Zhou, H., Bradley, A., Littwin, E., Razin, N., Saremi, O., Susskind, J. M., Bengio, S., and Nakkiran, P. What algorithms can transformers learn? a study in length generalization. In *The Twelfth International Conference on Learning Representations*, 2023.

Zhou, Y., Alon, U., Chen, X., Wang, X., Agarwal, R., and Zhou, D. Transformers can achieve length generalization but not robustly. *arXiv preprint arXiv:2402.09371*, 2024.

Contents

A. Conclusion

In this paper, we show that even the most advanced LLMs still struggle with exact copy tasks together with a hard copy benchmark which makes the LLMs fail, even when the inputs are well within their context lengths. Our experiments and analysis suggest that repeated local structures make copying difficult, and that reliable copying requires a stronger position-based alignment between source and target tokens. Motivated by this observation, we introduce 2D-RoPE, which views text as a 2D grid and reduces copying to same-column retrieval. This simple positional bias leads to strong length generalization on synthetic copy tasks and improves long-context retrieval in pretrained LLMs. We further introduce Auto 2D-RoPE to reduce the reliance on explicit line breaks. For future works, it remains important to better understand why multi-layer RoPE Transformers fail to learn length-invariant copying, how to scale Auto-2D-RoPE to practical LLM training, and whether our global-minima analysis can be extended to broader classes of rotation-based positional encodings.

B. Additional Related Works

Positional Encoding. The attention mechanism relies on position encoding methods to model the positional information of tokens (Vaswani et al., 2017). Most state-of-the-art LLMs employ RoPE (Su et al., 2024), which have demonstrated strong language modeling performance and length generalization. Kazemnejad et al. (2023) shows that attention with just a causal mask can reconstruct absolute and relative position information. More recently, Yang et al. (2026) proposes a position encoding scheme based on accumulating Householder transformations and showed that it can solve NC^1-complete problems under AC^0-reductions. Critically, no previous works have explored the influence of position encoding on the copying abilities of language models.

Length Generalization. Many early studies showed that transformers sometimes succeed and sometimes fail at length-generalization, correlated with multiple factors such as positional encoding, data format, and other training hyperparameters (Bhattamishra et al., 2020; Anil et al., 2022; Kazemnejad et al., 2023; Awasthi & Gupta, 2023; Jelassi et al., 2023; Wang et al., 2024a; Zhou et al., 2023; 2024; Chang & Bisk, 2025; Jelassi et al., 2024). Specifically on the copy task, Transformers trained to copy short strings often do not generalize well to longer strings when the input string includes repeated substrings (Zhou et al., 2024; Morwani et al., 2024). Theoretically, Huang et al. (2024) introduced a formal framework, proving that Transformers can solve a class of problems expressible by the C-RASP formalism (Yang & Chiang, 2024). Unfortunately, the repeated copy is provably difficult in the sense of C-RASP expressiveness.

The Representation Power of Transformer. There is a long line of work studying the representation power of Transformer (Pérez et al., 2021; Yao et al., 2021; Chiang & Cholak, 2022; Sanford et al., 2023). Merrill & Sabharwal (2024); Feng et al. (2023) analyze the representation power of Transformer with Chain-of-Thought. Wen et al. (2025) compares the expressive ability of Transformer with Recurrent Neural Networks. Later, Chen et al. (2024a) showed the limitation of multi-layer Transformers' expressive power.

Copy Capability of Transformers. Copying from context has been studied in several related settings. Some works use synthetic copy tasks to study Transformer length generalization (Kazemnejad et al., 2023; Huang et al., 2024; Jelassi et al., 2024), while other benchmarks include copying as part of broader LLM evaluations (Chen et al., 2025; Wang et al., 2024b; Hsieh et al., 2024). However, these works do not isolate exact copying of user-provided strings as a standalone failure mode of frontier LLMs, nor do they provide a positional-encoding-based remedy. In contrast, we directly evaluate modern LLMs on exact copy tasks, connect the failure to positional alignment, and propose 2D-RoPE to better align source and target tokens.

C. Test Details of the LLMs in Binary Copy and Python List Conversion

We give the details of our Copy benchmarks. All the prompt templates can be found in Appendix M

Choice of $\mathcal{V}$. Our preliminary design is to set $\mathcal{V} = \{0, 1\}$ and use the above generation methods to test LLMs' copy abilities. However, we found some of the tested LLMs' tokenizers would merge three consecutive 01 characters into a single token. To make our comparison fair, for these models, we set $\mathcal{V} = \{000, 111\}$, and for others, we set $\mathcal{V} = \{0, 1\}$; thus the actual number of tokens in the tested input is aligned. In our fine-tuning experiments in Section 5, we use the Llama2 tokenizer (Touvron et al., 2023) for which no such merging happens, so we set $\mathcal{V} = \{0, 1\}$.

C.1. Binary Copy Evaluation

For the binary copy evaluation in Figure 1, we test frontier LLMs on two challenging string distributions: Recursive-Flip and Imbalanced Generation. In both settings, the model is asked to reproduce the input binary string exactly. The Recursive-Flip distribution creates strings with many repeated substrings due to its recursive construction, while the Imbalanced distribution makes one symbol appear much more frequently than the other. These two distributions create different forms of local ambiguity. In Recursive-Flip strings, the same local substring may appear at many positions due to the recursive structure. In Imbalanced strings, short contexts become less informative because many positions contain the dominant symbol. As the string length increases, the number of possible false local matches grows, and the tested LLMs show a clear drop in exact-copy accuracy. This supports the view that frontier LLMs often rely on local-context matching rather than a purely position-based copy algorithm.

C.2. Python List Conversion Evaluation

We provide more details for the Python List Conversion task used in Figure 1. Each example is generated from a synthetic one-dimensional physical signal. We first sample one scenario uniformly from 9 physical signal families:

- *simple harmonic motion*
- *triangle wave motion*
- *sawtooth scan motion*
- *rectified AC signal*
- *clipped sensor oscillation*
- *relaxation oscillator*
- *Fourier periodic waveform*
- *pulse train*
- *pendulum-like oscillation*

These scenarios cover three broad types of structured sequences: smooth periodic oscillations, piecewise or reset-driven oscillations, and nonlinear or pulse-like signal responses. For each example, we sample a period p from a specified range, in our fine-tuning experiment in Section 5.1, $p \in [2, 8]$, and in Figure 1, $p \in [2, 10]$. Then we generate one cycle of length p. The final sequence of length n is then obtained by repeating this cycle through indexing, i.e.,

$$y_k = \text{cycle}_{k \bmod p}, \qquad k = 0, \ldots, n-1.$$

This ensures that the periodic structure is exact at the value level, rather than relying on recomputing trigonometric functions at shifted phases. We then round each value to a fixed number of decimal places, using two decimals in our experiments, and represent the input as a comma-separated sequence of numeric values. The target output is the same sequence wrapped as a Python list, with the form

$$[y_1, y_2, \ldots, y_n]$$

Thus, the task requires the model to preserve every numeric token while changing only the output format.

The signal families are generated as follows. For smooth periodic signals, we use simple sinusoidal motion, Fourier periodic waveforms with 2 to 4 harmonics, and pendulum-like oscillations with a small third-harmonic component. For piecewise or reset-driven signals, we use triangle waves, sawtooth waves, and relaxation oscillators. For nonlinear or pulse-like signals, we use rectified AC signals, clipped sinusoidal sensor readings, and periodic Gaussian pulse trains. Each sample also includes randomized physical parameters such as amplitude, phase, offset, clipping level, relaxation time, pulse width, and pulse center when applicable. Finally, the user instruction is sampled from a set of prompt templates, all asking the model to convert or copy the given physical sensor sequence into a Python list. This keeps the task objective fixed while avoiding dependence on a single prompt wording.

Table 2. Copying accuracy of repeat flip 0 1.

Model	128-255	256-511	512-1023	1024-2047	2048-4095
GPT-5.5	0.88	0.84	0.54	0.44	0.06
Gemini 3.1 Pro	1.0	0.98	0.86	0.68	0.38
DeepSeek V4 Pro	0.72	0.70	0.48	0.32	0.12
Qwen 3.6 Plus	0.94	0.80	0.32	0.16	0.04

Table 3. Copying accuracy of 95% imbalanced a b.

Model	128-255	256-511	512-1023	1024-2047	2048-4095
GPT-5.5	0.74	0.52	0.46	0.32	0.04
Gemini 3.1 Pro	0.94	0.76	0.4	0.2	0.02
DeepSeek V4 Pro	0.74	0.84	0.58	0.62	0.34
Qwen 3.6 Plus	0.96	0.88	0.72	0.52	0.26

C.3. Repeat-structure copy test for frontier LLMs

In our test, we generate string by repeat a base string for multiple times until reaches length 1000. Each token in the base string is generated by uniformly selecting " a" and " b". The length of base string is among $\{4, 6, 8, 10\}$. Then, we choose a flipping ratio $p \in \{0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. Each token of the whole string is randomly flipped with probability p. For each length and flipping ratio, we sample 40 strings and test the models' copying accuracy. The x-axis means the ratio of base strings such that none of the tokens are flipped. The larger the ratio is, the structure of this string contains more repetition.

D. Details of Synthetic Experiments on Binary Copy

We conduct training-from-scratch experiments on the synthetic binary copy dataset. For each binary string $s \in \{0, 1\}^*$, the input sequence takes the form

$$\mathcal{S} := (s, \texttt{<\textbackslash n>}, \texttt{<O:>}, s),$$

where the model is trained to predict the next token autoregressively, including the copied string after <O:> and the final <EOS> token at the end of the sequence.

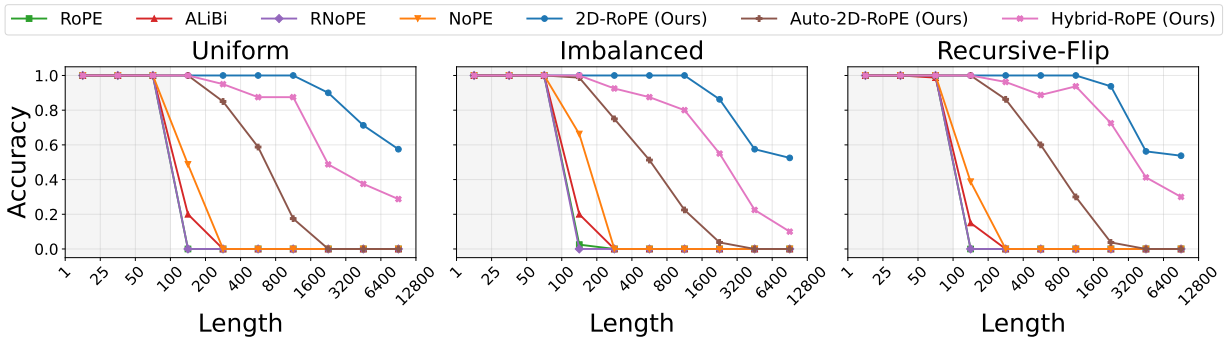
Data Distributions. As introduced in Section 3.1, we consider three data distributions: *uniform*, *imbalanced*, and *recursive-flip*. All training is conducted on imbalanced data with string lengths from 1 to 100. For the imbalanced distribution, the probability of sampling the first binary token is chosen from

$$\{0.05, 0.15, 0.3, 0.5, 0.7, 0.85, 0.95\},$$

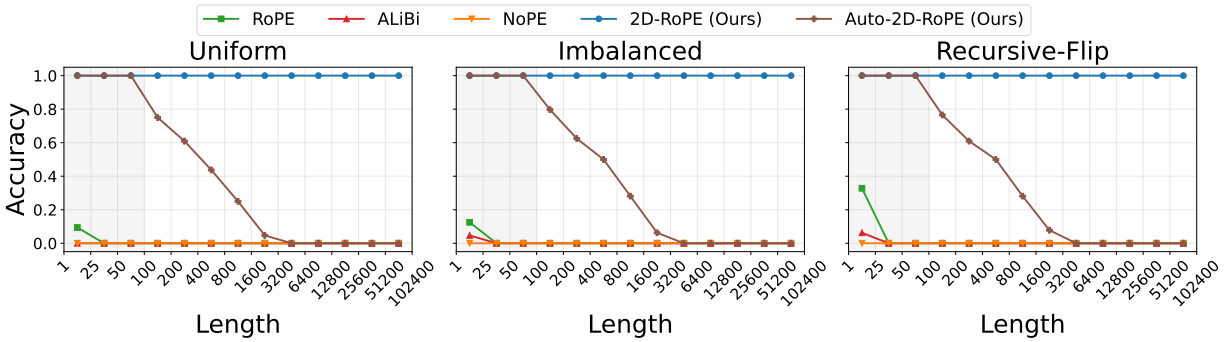
and tokens are then sampled independently according to this probability. At test time, we evaluate on all three distributions. Thus, uniform and recursive-flip test strings are out-of-distribution relative to the training distribution.

Model Architectures. We train Transformers from scratch under two model sizes. The shallow model has 1 layer and 2 attention heads, where each head has dimension 512. The deeper model has 12 layers and 12 attention heads, where each head has dimension 128. Thus, the embedding dimensions are 1024 and 1536, respectively. We set dropout to 0 and use bias-free linear layers. For each architecture and each positional encoding, we train 8 models with different random seeds and report the average accuracy. We average over seeds because training on the copy task shows noticeable variance across runs, especially near the boundary between interpolation and extrapolation.

Training Configuration. The two model sizes use different optimization settings. For the 12-layer, 12-head models, we train for 30000 iterations with learning rate decayed from 10^{-4} to 10^{-6}. For the 1-layer, 2-head models, we train for 60000 iterations with learning rate decayed from 5×10^{-4} to 5×10^{-5}. We use cosine learning rate decay. For both settings, we use weight decay 0.01, $\beta_2 = 0.95$, and a warmup length of 100 iterations. The training batch size is 64 with gradient accumulation over 4 steps. The maximum context length is set to 21000 for 1-layer model and 21000 for 12-layer model, which is large enough for all training and evaluation sequences considered in these experiments. We evaluate every 500 iterations, using 50 batches for training-set evaluation and 250 batches for test-set evaluation.



(a) Length generalization results for different positional encodings with 12 layers and 12 heads.



(b) Length generalization results for different positional encodings with 1 layer and 2 heads.

Figure 6. Length generalization for training-from-scratch experiments.

Evaluation. During training, strings have length at most 100. At evaluation time, we test models on longer strings to measure length generalization. We evaluate lengths beyond the training range up to the longest test length used in the corresponding figure, with test lengths binned by interval size 50. Exact-copy accuracy is computed by checking whether the generated copied string matches the ground-truth string exactly. This is a strict metric: a single wrong token makes the whole example incorrect.

Results. Figure 6 shows the full length-generalization results for all test distributions. The main text reports the Recursive-Flip setting because it is both out-of-distribution and especially challenging due to repeated substrings. Across the full results, standard positional encodings fail to extrapolate reliably beyond the training length range, while 2D-RoPE-based methods show substantially stronger length generalization.

Training Configuration. As discussed before, we train transformers from scratch on the binary copy task under two model sizes. The first is a shallow model with 1 layer and 2 attention heads, where each head has dimension 512. The second is a deeper model with 12 layers and 12 attention heads, where each head has dimension 128. For each architecture and each positional encoding, we train 8 models with different random seeds and report the average accuracy. We average over seeds because training on the copy task shows noticeable variance across runs, especially near the boundary between interpolation and extrapolation.

E. Further Discussion of Multi-Layer RoPE Learns Length-Dependent Pattern

We provide a more detailed discussion of the attention patterns in Figure 4 and Figure 8. The main text shows that RoPE does not learn a clean position-based copy map even on a failed example within the training length range. Here, we further examine what happens beyond the training length.

Within the Training Length. In Figure 4, the last-layer attention should ideally align each output position with its corresponding input position. However, the zoomed attention map is not a clean diagonal alignment. The head attends

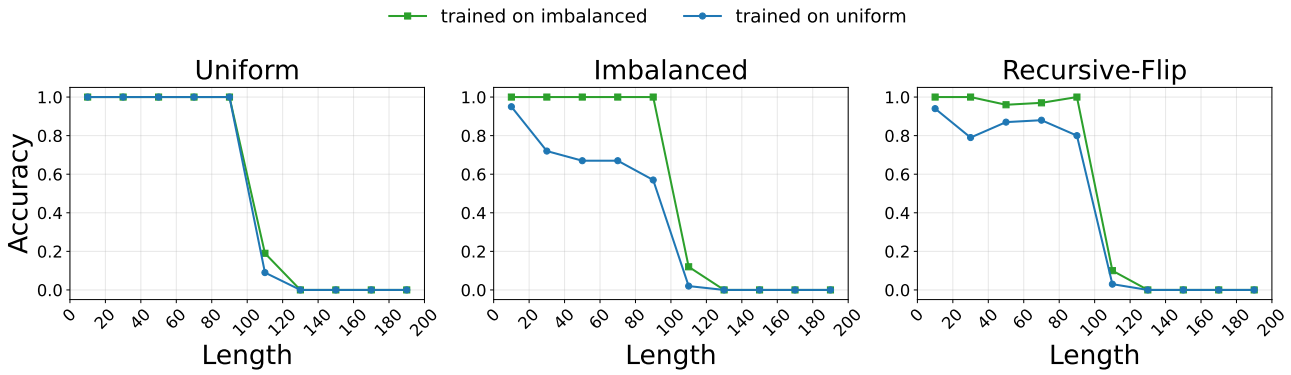


Figure 7. Length generalization for training-from-scratch experiments in 2 distributions.

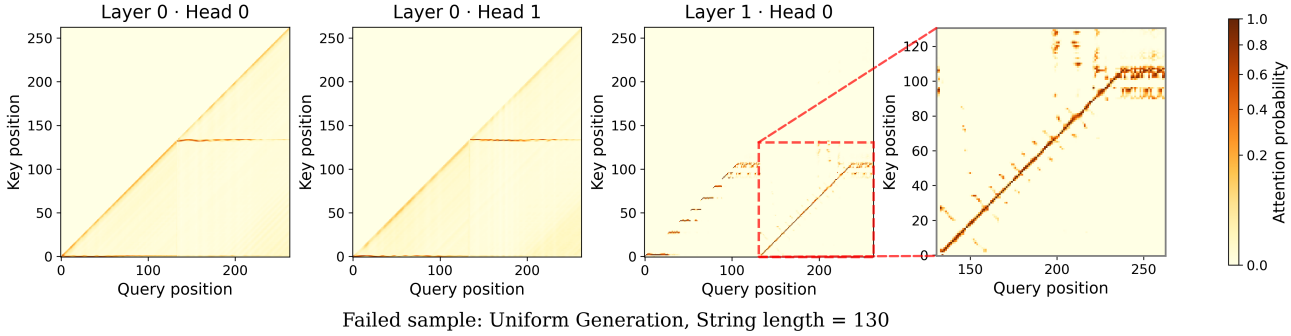


Figure 8. Attention map of a RoPE-based model on a failed sample with sequence length beyond the training length. In Layer 0, attention heads exhibit both **horizontal patterns** (attending to special tokens in between the input string and the output, and the tokens before the string as a locator) and **diagonal patterns** (semantic/induction-style behavior). In Layer 1, one can see that RoPE would attend to wrong tokens near themselves periodically.

not only to the correct source positions, but also periodically assigns large attention scores to nearby wrong tokens around the diagonal. This suggests that the final copy decision is affected by local-context matching: when repeated or similar substrings appear, locally similar positions can compete with the correct source position. Thus, even within the training length range, RoPE can fail because the learned copy mechanism is not purely position-based.

Beyond the Training Length. Figure 8 further shows why this learned mechanism does not length-generalize. In Layer 0, the two heads contain horizontal locator patterns that attend to special tokens around the boundary between the input string and the output string, such as the delimiters in [<\n>, <I:>, <O:>]. These patterns provide a position-related signal and can help the model infer source-target alignment within the seen length range. However, when the sequence length goes beyond the maximum training length, the locator pattern no longer remains stable: after the copied position exceeds the training range, the horizontal attention pattern becomes discontinuous and eventually breaks. This indicates that the model's position-based component is not a true length-invariant copy algorithm, but a locator-style shortcut tied to the training lengths.

At the same time, Layer 0 also contains diagonal patterns, which resemble local-context or induction-style behavior (Olsson et al., 2022; Singh et al., 2024; Jelassi et al., 2024). These patterns can attend to nearby or locally similar tokens, but they cannot uniquely identify the correct source position when repeated substrings are present. Therefore, the trained RoPE model combines two imperfect mechanisms: a locator-style positional shortcut that does not generalize in length, and a context-based matching mechanism that becomes ambiguous on repeated structures. This explains why the last-layer attention becomes noisy and why the model fails to implement the ideal position-based copy rule suggested by Theorem 3.2.

F. Implementation Details of 2D-RoPE

F.1. Rotational Frequency Hyperparameter.

In RoPE, there is a hyperparameter θ that determines the rotational frequency. For 2D-RoPE, the two dimensions can have separate rotational frequencies, θ_x, θ_y. However, for simplicity, we keep $\theta_x = \theta_y$. Moreover, we use the default hyperparameter from the original RoPE paper, which is $\theta = 10,000$, and we set $\theta_x = \theta_y = 1,000$. Our preliminary experiments show that setting $\theta_x = \theta_y = 100$ or $\theta_x = \theta_y = 10,000$ does not yield significantly different results.

F.2. Implementation Details and Comparison with Vision 2D-RoPE

RoPE-style 2D positional encodings have been explored in computer vision, where each image patch has a natural spatial coordinate. For example, prior vision models apply RoPE to patch tokens using their height and width coordinates, so that attention can encode relative spatial displacement on the image grid (Jeevan & Sethi, 2022; Heo et al., 2024; Chu et al., 2024). In this setting, the 2D coordinates are given by the image layout itself.

Our implementation differs in both motivation and coordinate construction. In language modeling, the input is normally treated as a 1D token sequence, and there is no fixed 2D grid analogous to image patches. We instead use line break tokens to induce a 2D structure in text. Each token receives a row ID, determined by the number of preceding line breaks, and a column ID, determined by its offset within the current line. Thus, the 2D coordinates are not externally given but are constructed from the textual structure. Further, we introduce an alternative termed Auto 2D-RoPE for 2D-RoPE to learn the 2D structure in context adaptively in Appendix H.

In our implementation, 2D-RoPE follows an axial decomposition: part of the head dimension is rotated according to the row ID, and the remaining part is rotated according to the column ID. We keep the standard causal attention mask for language modeling. This is also different from most vision applications, which typically use bidirectional attention over image patches.

G. LLM Experiment Details

Unless otherwise specified, the model training in the Section 5 follows the following settings.

Table 4. The hyperparameters of the three model of different sizes involved in Section 5.

Hyperparameter	**350M**	**730M**	**1.4B**
Vocab size	32000	32000	32000
Layers	24	24	28
Hidden size	1024	1536	2048
MLP intermediate dim	3072	4096	5120
MLP activation function	SwiGLU	SwiGLU	SwiGLU
Attention heads	8	12	16
KV heads	8	12	16
Attention head size	128	128	128
RoPE θ	10,000	10,000	10,000
2D-RoPE θ_x, θ_y	1,000	1,000	1,000
Training Configuration			
Pretraining batch size	256	256	256
Finetuning batch size	64	64	64
Max length	2K	2K	2K
Pretraining max LR	5e-4	5e-4	4e-4
Finetuning max LR	5e-5	5e-5	5e-5

G.1. LLM Model Configuration Details

The hyperparameters for the models involved in LLM experiments (Section 5) are described in Table 4. These hyperparameter values are chosen to be similar to popular open-source LLMs such as Qwen3 and Llama3 (Grattafiori et al., 2024). Each model uses the Llama3 tokenizer.

G.2. LLM Pretraining Configuration Details

Here, we describe more details for the LLM pretraining (Section 5) for reproducibility.

Learning Rate Scheduler. Each model was trained using the WSD LR scheduler (Hu et al., 2024), with different LR depending on the model size. The 350M, 730M, and 1.4B models use 5e-4, 5e-4, and 2e-4 max LR, respectively. Each of the LRs warms up (from 0) for 100 steps, and decays for the last 10% steps to $1/10$ of their max LR.

Data Batching. The batch size is always 32, and the maximum sequence length during pretraining is always 2048, so the number of tokens per batch is 512K. Each pretraining document is concatenated with an `<end_of_sequence>` delimiter until they exceed maximum sequence length, and then truncated down to 2048. The part that was truncated is discarded. Each token can attend to all other tokens in other documents that was concatenated into the same sequence.

Optimizer. We use the same optimizer for both pretraining and finetuning, which is AdamW (Loshchilov & Hutter, 2019) with beta values of (0.9, 0.95), a weight decay of 0.1, and gradient clipping at 1.0.

Hardware. All LLM experiments were conducted on machines equipped with NVIDIA A800-80GB GPUs, using PyTorch and HuggingFace Transformer libraries. We used BF16 precision during both training and evaluation.

G.3. LLM Finetuning Configuration Details

Learning Rate Scheduler. We use the Cosine LR scheduler for finetuning, since the cost of rerunning finetuning is much lower than pretraining. The LR warms up for 100 steps then decays down to 0.

Data Batching. For finetuning, each sequence might have different sequence length, and we use attention masking to ensure that padding tokens are not attended to. Since the training dataset is synthesized, it does not exceed 2048 tokens and there is no truncation. We always use batch size of 64 during finetuning.

Optimizer and Hardware. The optimizer for finetuning is the same as the one used for pretraining, and all experiments were conducted on the same hardware, based on the same precision and implementation libraries.

G.4. LLM Evaluation Details

Common-Sense Reasoning To evaluate the common-sense reasoning abilities of LLMs, we use the average accuracy in the following set of tasks: ARC-easy, ARC-challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2019), MMLU (Hendrycks et al., 2021), LAMBADA (Paperno et al., 2016), PIQA (Bisk et al., 2019), BoolQ (Clark et al., 2019). We use LM-Evaluation-Harness (Gao et al., 2024) for evaluation.

Needle-in-a-Haystack. Our needle-in-a-haystack task is the same as the NIAH-Single tasks from RULER, but with one change: we add newline characters before and after the needle (which is a sentence). This change enables 2D-RoPE to better copy the needle from the context.

G.5. LR Ablation in LLM Experiments

In the LLM experiments (Section 5), we swept different max LR in finetuning and chose the best LR that balanced the performance across different tasks. In this section, we report the results for each value of LR ($\{1\times10^{-4}, 5\times10^{-5}, 3\times10^{-5}\}$), which is shown in Table 5. Overall, the results show that using a larger LR leads to better copying abilities, but the CSR and NIAH performance degrades due to catastrophic forgetting. Interestingly, 2D-RoPE is generally more robust to the choice of LR, leading to less catastrophic forgetting in the case of a fairly large LR.

G.6. Pretraining Loss

In this section, we report the training and evaluation loss during pretraining (in Section 5) for completeness and reproducibility. We evaluate the pretraining checkpoints on FineWeb-Edu, another widely used pretraining corpus. The results are shown in Figure 9 and Table 6. One can see that the three models have roughly the same training and evaluation loss, which indicates that the performance of 2D-RoPE on next token prediction is comparable to RoPE.

Table 5. LR ablation results in the LLM experiments. Each model was trained with 3 different LR values, and the main content reports the LR that achieves best overall performance in the three kinds of tasks (CSR, NIAH, and Copy), which is selected to be 5e-5 and is highlighted in orange.

		PT			NIAH				Copy (Imbalanced)				Copy (Recursive-Flip)			
Model	Size	Data	LR	CSR	512	1K	2K	4K	1K	2K	4K	8K	1K	2K	4K	8K
Different Model Scales with Chinchilla law																
RoPE	350M	7B	1e-4	40.9	74.5	74.5	60.1	0.1	100.0	16.8	0.0	0.0	26.8	11.7	1.1	1.8
RoPE	350M	7B	5e-5	41.4	93.7	92.1	87.0	1.9	97.2	14.8	0.0	0.0	19.6	8.7	0.0	0.0
RoPE	350M	7B	3e-5	41.6	93.9	94.1	89.5	4.3	93.2	9.6	0.0	0.0	21.6	8.7	0.0	0.0
H-RoPE	350M	7B	1e-4	41.2	92.0	98.1	93.8	66.3	100.0	98.2	18.8	0.2	100.0	99.0	72.3	27.3
H-RoPE	350M	7B	5e-5	41.2	93.1	98.8	94.4	57.9	99.6	57.8	0.0	0.0	99.0	91.3	39.4	1.8
H-RoPE	350M	7B	3e-5	41.3	92.7	98.1	93.7	48.8	84.8	33.4	0.0	0.0	100.0	98.1	55.3	5.5
2D-RoPE	350M	7B	1e-4	40.1	89.8	96.7	92.7	88.7	99.8	94.8	81.8	27.2	100.0	100.0	100.0	100.0
2D-RoPE	350M	7B	5e-5	40.4	91.9	97.9	93.9	92.8	97.4	69.2	33.4	2.4	97.9	92.2	75.5	43.6
2D-RoPE	350M	7B	3e-5	40.5	92.3	98.6	94.3	93.1	81.2	57.2	14.8	0.4	100.0	98.1	92.6	87.3
RoPE	730M	15B	1e-4	43.3	93.6	97.6	91.7	21.9	100.0	20.8	0.0	0.0	20.6	6.8	0.0	1.8
RoPE	730M	15B	5e-5	44.0	100.0	98.4	96.7	26.4	97.0	11.0	0.0	0.0	12.4	5.8	1.1	0.0
RoPE	730M	15B	3e-5	44.2	97.2	97.1	93.1	27.4	88.8	4.4	0.0	0.0	14.4	7.8	0.0	0.0
H-RoPE	730M	15B	1e-4	44.1	96.4	77.7	67.1	7.1	93.6	13.6	0.0	0.0	100.0	100.0	78.7	3.6
H-RoPE	730M	15B	5e-5	45.5	95.9	97.9	90.0	4.5	95.4	11.4	0.0	0.0	76.3	72.8	17.0	0.0
H-RoPE	730M	15B	3e-5	45.4	95.9	99.5	92.1	8.0	84.0	4.0	0.0	0.0	96.9	92.2	35.1	0.0
2D-RoPE	730M	15B	1e-4	44.3	91.7	95.9	87.0	80.1	100.0	100.0	89.6	24.2	100.0	98.1	88.3	90.9
2D-RoPE	730M	15B	5e-5	45.2	93.0	98.2	92.4	91.3	99.4	74.6	12.8	0.0	78.4	73.8	47.9	5.5
2D-RoPE	730M	15B	3e-5	45.1	99.8	99.3	97.3	96.2	92.8	41.4	3.6	0.0	96.9	89.3	56.4	36.4
RoPE	1.4B	28B	1e-4	36.8	0.0	0.0	0.0	0.0	100.0	31.0	0.0	0.0	20.6	8.7	0.0	1.8
RoPE	1.4B	28B	5e-5	45.9	84.5	87.2	82.3	10.0	99.6	26.8	0.0	0.0	15.5	4.9	0.0	0.0
RoPE	1.4B	28B	3e-5	46.0	92.3	96.9	96.3	11.2	99.8	23.2	0.0	0.0	15.5	7.8	1.1	0.0
H-RoPE	1.4B	28B	1e-4	46.8	8.5	6.3	5.2	2.3	100.0	92.4	38.8	5.2	17.5	5.8	1.1	0.0
H-RoPE	1.4B	28B	5e-5	47.2	93.4	98.5	94.5	62.3	100.0	73.6	3.6	0.0	12.4	1.9	0.0	0.0
H-RoPE	1.4B	28B	3e-5	47.2	90.9	97.5	93.8	61.5	99.8	65.8	2.8	0.0	14.4	5.8	0.0	0.0
2D-RoPE	1.4B	28B	1e-4	45.9	2.7	3.7	5.4	4.6	100.0	100.0	100.0	99.0	99.0	95.1	93.6	69.1
2D-RoPE	1.4B	28B	5e-5	46.8	98.3	96.4	92.6	86.0	100.0	100.0	92.0	61.8	66.0	54.4	26.6	1.8
2D-RoPE	1.4B	28B	3e-5	46.9	98.7	97.5	92.7	92.3	100.0	92.0	51.8	14.6	92.8	76.7	45.7	10.9
Overtrained Setting																
RoPE	730M	100B	3e-5	46.9	81.9	94.1	86.5	0.2	99.5	23.3	0.0	0.0	15.5	7.8	1.1	1.8
RoPE	730M	100B	5e-5	46.0	76.9	80.9	70.5	0.4	99.6	29.2	0.0	0.0	24.7	8.7	1.1	1.8
RoPE	730M	100B	1e-4	34.3	0.0	0.0	0.0	0.0	100.0	22.6	0.0	0.0	63.9	15.5	0.0	1.8
H-RoPE	730M	100B	3e-5	46.4	99.9	95.2	96.9	31.1	100.0	41.4	0.0	0.0	92.8	86.4	6.4	0.0
H-RoPE	730M	100B	5e-5	46.2	83.7	89.9	89.3	38.3	99.2	52.1	0.0	0.0	100.0	98.1	28.7	1.8
H-RoPE	730M	100B	1e-4	43.3	0.1	0.0	0.0	0.0	100.0	74.6	1.7	0.0	100.0	100.0	50.0	1.8
2D-RoPE	730M	100B	3e-5	46.3	86.0	95.4	91.6	89.3	100.0	95.5	37.6	0.0	99.0	98.1	84.0	34.5
2D-RoPE	730M	100B	5e-5	46.2	87.3	93.1	88.6	86.5	100.0	96.3	72.8	15.8	100.0	99.0	94.7	74.5
2D-RoPE	730M	100B	1e-4	43.4	0.0	0.1	0.1	0.0	100.0	99.8	94.1	40.7	100.0	100.0	100.0	81.8

H. Auto 2D-RoPE Discussion

We now describe the architecture of Auto 2D-RoPE in more detail. The goal is to keep the main advantage of 2D-RoPE, namely assigning each token a 2D position ID, while removing the hard reliance on a manually specified line-break token. In standard 2D-RoPE, the position ID of each token is updated by a fixed rule: the column index increases within a line, and the row index increases after a line break. In Auto 2D-RoPE, we instead let the model learn how the 2D position IDs should evolve along the sequence.

For each token i, let its 2D position ID be

$$P_i = \left(p_i^{(1)}, p_i^{(2)}\right).$$

Intuitively, $p_i^{(1)}$ plays the role of a column-like coordinate, while $p_i^{(2)}$ plays the role of a row-like coordinate. Unlike rule-based 2D-RoPE, these coordinates are not computed from line breaks. Instead, at each layer, we generate two update values from the hidden representation of the current token.

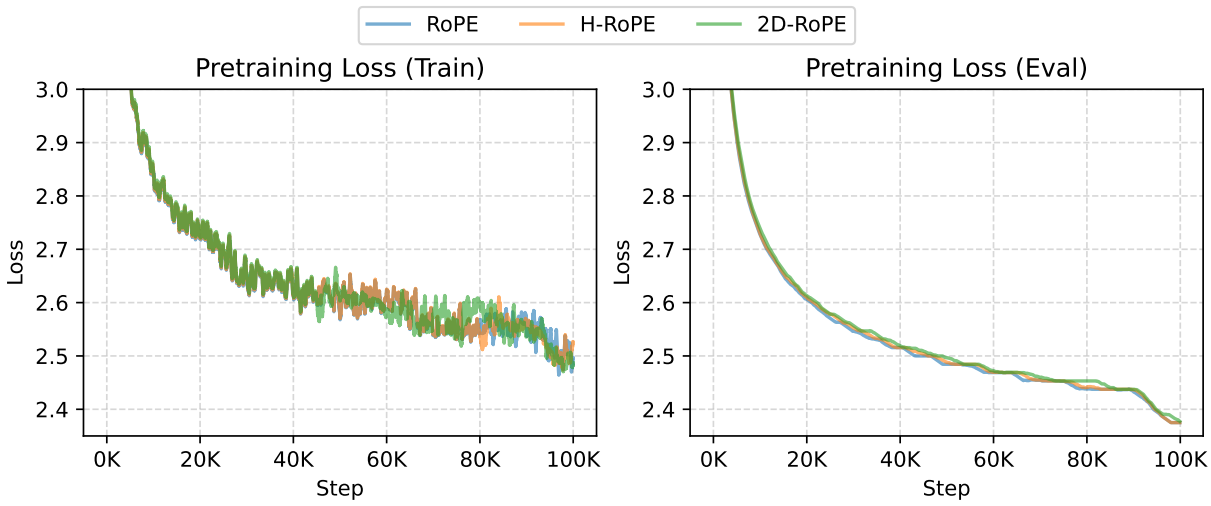


Figure 9. The training loss and evaluation loss during pretraining of the 2D-RoPE, H-RoPE, and RoPE models with 730M parameters (see Section 5). These models were trained on 100B tokens from DCLM, and evaluated on FineWeb-Edu.

Table 6. The training and evaluation loss during pretraining of the three overtrained models in Section 5. The training data is DCLM and the evaluation data is FineWeb-Edu.

Model	**Training Loss**	**Evaluation Loss**
RoPE	2.5080	2.3808
H-RoPE	2.5000	2.3808
2D-RoPE	2.4943	2.3876

Concretely, let $\boldsymbol{h}_i^{(\ell)}$ be the hidden representation of token i at layer ℓ. We apply a learnable linear projection to obtain

$$(a_i^{(\ell)}, b_i^{(\ell)}) = \boldsymbol{W}_{\text{pos}}^{(\ell)} \boldsymbol{h}_i^{(\ell)} + \boldsymbol{b}_{\text{pos}}^{(\ell)}.$$

We then pass these two scalars through a sigmoid function and a rescaling factor α:

$$u_i^{(\ell)} = \alpha \cdot S(a_i^{(\ell)}), \qquad v_i^{(\ell)} = \alpha \cdot S(b_i^{(\ell)}),$$

where $S(\cdot)$ denotes the sigmoid function. In our experiments, we set $\alpha = 2$ by default. This makes the update coefficients bounded and positive.

Given these updated values, the first coordinate is updated by an affine transformation

$$p_{i+1}^{(1)} = u_i^{(\ell)} p_i^{(1)} + v_i^{(\ell)}.$$

Equivalently,

$$\begin{bmatrix} p_{i+1}^{(1)} \\ 1 \end{bmatrix} = \begin{bmatrix} u_i^{(\ell)} & v_i^{(\ell)} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p_i^{(1)} \\ 1 \end{bmatrix}.$$

The second coordinate is updated by the complementary increment

$$p_{i+1}^{(2)} = p_i^{(2)} + \left(1 - u_i^{(\ell)}\right).$$

Putting the two coordinates together, we can write the update as

$$\begin{bmatrix} p_{i+1}^{(1)} \\ p_{i+1}^{(2)} \\ 1 \end{bmatrix} = \begin{bmatrix} u_i^{(\ell)} & 0 & v_i^{(\ell)} \\ 0 & 1 & 1 - u_i^{(\ell)} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} p_i^{(1)} \\ p_i^{(2)} \\ 1 \end{bmatrix}.$$

Thus, the position ID of token i is determined by the cumulative product of these learned affine transformations along the sequence.

This update rule can be viewed as a learnable relaxation of the line-break rule used in 2D-RoPE. For example, if $u_i^{(\ell)} \approx 1$ and $v_i^{(\ell)} \approx 1$, then

$$p_{i+1}^{(1)} \approx p_i^{(1)} + 1, \qquad p_{i+1}^{(2)} \approx p_i^{(2)},$$

which resembles moving to the next column within the same row. On the other hand, if $u_i^{(\ell)} \approx 0$ and $v_i^{(\ell)} \approx 0$, then

$$p_{i+1}^{(1)} \approx 0, \qquad p_{i+1}^{(2)} \approx p_i^{(2)} + 1,$$

which resembles resetting the column coordinate and moving to a new row. Therefore, Auto 2D-RoPE can in principle learn behavior similar to rule-based 2D-RoPE, but it is not restricted to using the newline token as the only row separator.

After the 2D position IDs are computed, we apply the same 2D-RoPE rotation as in Section 4. Specifically, for each layer ℓ, the query and key vectors are rotated according to the learned coordinates

$$\big(p_i^{(1)}, p_i^{(2)}\big),$$

where one part of the head dimension encodes relative differences in the first coordinate, and the other part encodes relative differences in the second coordinate. The attention layer is then computed in the usual causal way:

$$\mathsf{Attn}(\boldsymbol{H}^{(\ell)}) = \mathcal{S}\left(\mathrm{MASK}\left(R_{\mathrm{auto}}(\boldsymbol{H}^{(\ell)}\boldsymbol{Q})\big(R_{\mathrm{auto}}(\boldsymbol{H}^{(\ell)}\boldsymbol{K})\big)^\top\right)\right)\boldsymbol{H}^{(\ell)}\boldsymbol{V},$$

where R_{auto} denotes the 2D-RoPE rotation using the learned Auto 2D position IDs.

Importantly, the quantities $a_i^{(\ell)}$ and $b_i^{(\ell)}$ are not independent learnable parameters tied to absolute positions. They are generated by a shared linear map from the hidden representation of each token. Therefore, the number of parameters does not grow with the input length, and the coordinate-generation rule can be applied to longer sequences at test time. This is the main difference between Auto 2D-RoPE and simply learning a fixed position ID table.

Empirically, this learned coordinate update allows the model to recover useful 2D structure even when the input does not contain explicit newline separators. As discussed in the main text, when we replace the newline token <\n> with another character $*$, rule-based 2D-RoPE can no longer identify row boundaries and effectively loses its 2D structure. In contrast, Auto 2D-RoPE can still learn data-dependent position IDs and shows stronger length generalization on the copy task.

I. Preliminary for Theoretical Results

In this section, we formalize our general setting for our theory.

Theoretical Definition of Copying. First we define our copying task used in theories. We always have a 0/1 string s with length at least 1 and a copying position $1 \le i \le |s|$. The input of the model is a string generated by s, i, denoted by $\mathsf{Str}(s,i) := (s, <\backslash\mathrm{n}>, <\mathrm{O}:>, s_{:i})$. Here $s_{:i}$ means the length $(i-1)$ prefix of s. For this input, the last token is s_{i-1} (and when $i = 1$, the last token is <O:>). The vocabulary set is then defined as $\mathcal{V} = \{<\mathrm{O}:>, <\backslash\mathrm{n}>, 0, 1\}$. We model the predictor as a function $P_{\boldsymbol{\theta}} : \mathcal{V}^+ \to [0,1]^2$, parameterized by $\boldsymbol{\theta}$, which maps an input string $x \in \mathcal{V}^+$ to a distribution over the next token in $\{0,1\}$. Specifically, for any input x, the output $P_{\boldsymbol{\theta}}(x) = \big([P_{\boldsymbol{\theta}}(x)]_1, [P_{\boldsymbol{\theta}}(x)]_2\big)$ satisfies

$$[P_{\boldsymbol{\theta}}(x)]_1 + [P_{\boldsymbol{\theta}}(x)]_2 = 1, \quad \text{and} \quad 0 \le [P_{\boldsymbol{\theta}}(x)]_i \le 1 \text{ for } i \in \{1,2\}.$$

Thus, $P_{\boldsymbol{\theta}}(x)$ defines a valid probability distribution over the next token. For each pair of s, i, the corresponding loss is defined as

$$\mathcal{L}_{\boldsymbol{\theta}}(s,i) = -\log\left[P_{\boldsymbol{\theta}}(\mathsf{Str}(s,i))\right]_{s_i+1}.$$

Now given a distribution $\mathcal{D}$, the total loss is defined as

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(s,i)\sim\mathcal{D}}[\mathcal{L}_{\boldsymbol{\theta}}(s,i)].$$

Definition I.1. We say a model can successfully do copying in length L if for any $|s| \le L$ and $1 \le i \le |s|$, we have $[P_{\boldsymbol{\theta}}(\mathsf{Str}(s,i))]_{s_i+1} > \frac{1}{2}$.

Formalization of RoPE. Given a dimension d, RoPE attention is defined as a function $\mathsf{Attn} : \mathbb{R}^{T\times d} \to \mathbb{R}^{T\times d}$ for any L parameterized by $\boldsymbol{K}, \boldsymbol{Q}, \boldsymbol{V} \in \mathbb{R}^{d\times d}$. For any input $\boldsymbol{X}$, the RoPE attention is defined as

$$\mathsf{Attn}(\boldsymbol{X}) = \mathcal{S}\left(\mathrm{MASK}\left(R\left(\boldsymbol{X}\boldsymbol{Q}\right)\left(R\left(\boldsymbol{X}\boldsymbol{K}\right)\right)^\top\right)\right)\boldsymbol{X}\boldsymbol{V}^\top \in \mathbb{R}^{L\times d}.$$

Here, $R : \mathbb{R}^{T\times d} \to \mathbb{R}^{T\times d}$ is the RoPE function. Specifically, for an input $\boldsymbol{X} \in \mathbb{R}^{T\times d}$, the i-th row X_i for $1 \le i \le L$ is mapped to $X_i R_i^\top$, where

$$R_i = \begin{bmatrix} \cos(i\beta_1) & -\sin(i\beta_1) & & & \\ \sin(i\beta_1) & \cos(i\beta_1) & & & \\ & & \cos(i\beta_2) & -\sin(i\beta_2) & \\ & & \sin(i\beta_2) & \cos(i\beta_2) & \\ & & & & \ddots \end{bmatrix} \in \mathbb{R}^{d\times d}.$$

Here, β_j is the base frequency associated with the $(2j-1, 2j)$ coordinate pair. In our setting, these β_j's are arbitrary hyperparameters. Thus, R_i acts block-diagonally on X_i, rotating each two-dimensional subspace $(X_{i,2j-1}, X_{i,2j})$ by angle $i\beta_j$. The causal mask operator $\mathrm{MASK}(\cdot)$ enforces the autoregressive constraint: for $\boldsymbol{Z} \in \mathbb{R}^{T\times T}$,

$$\mathrm{MASK}(\boldsymbol{Z})_{ij} = \begin{cases} \boldsymbol{Z}_{ij}, & j \le i, \\ -\infty, & j > i, \end{cases}$$

so that position i cannot attend to any future position $j > i$. The operator $\mathcal{S}(\cdot)$ applies row-wise softmax normalization: for a matrix $\boldsymbol{A} \in \mathbb{R}^{T\times T}$,

$$\mathcal{S}(\boldsymbol{A})_{ij} = \frac{\exp(\boldsymbol{A}_{ij})}{\sum_{k=1}^{L}\exp(\boldsymbol{A}_{ik})}, \quad 1 \le i, j \le T.$$

J. Shift Invariant Positional Encoding Cannot Solve Copying

In this section, we prove that shift invariant positional encoding cannot do copying.

J.1. Shift Invariant PE Formalization

First, we mathematically define what is shift-invariant positional encoding.

The attention is also parameterized by $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^d$. For each $\boldsymbol{X} \in \mathbb{R}^{T\times d}$, the attention first computes the key vectors and query vectors

$$\boldsymbol{k}_i = \boldsymbol{K}^\top \boldsymbol{X}_{i,:}^\top, \boldsymbol{q}_i = \boldsymbol{K}^\top \boldsymbol{X}_{i,:}^\top.$$

Then, it computes a matrix $\boldsymbol{A} \in \mathbb{R}^{T\times T}$ such that

$$\boldsymbol{A}_{i,j} = \begin{cases} -\infty & i < j \\ f(i-j, \boldsymbol{q}_i, \boldsymbol{k}_j) & i \ge j \end{cases}$$

for some function $f : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Then, the output of the attention is

$$\mathsf{Attn}(\boldsymbol{X}) = \mathcal{S}(\boldsymbol{A})\boldsymbol{X}\boldsymbol{V}^\top.$$

When we treat this attention as a model for copying (as defined in Section I), we first have a token embedding $\boldsymbol{w}_\complement \in \mathbb{R}^d$ for each $\complement \in \mathcal{V}$. For any input s, i, we convert the string $\mathsf{Str}(s, i)$ into the input to the attention as $\boldsymbol{X}_{j,:} = \boldsymbol{w}_{\mathsf{Str}(s,i)_j}^\top$ (denote this matrix as $\boldsymbol{X}(s, i) \in \mathbb{R}^{(|s|+i+1)\times d}$, where $|s| + i + 1$ is the length of $\mathsf{Str}(s, i)$). We also add another matrix parameter $\boldsymbol{W}_O \in \mathbb{R}^{d\times 2}$, and the model is defined as a map from (s, i) to a 2-dimensional vector:

$$P_{\boldsymbol{\theta}}(s, i) = \mathrm{softmax}\left((\mathsf{Attn}(\boldsymbol{X}(s, i))\boldsymbol{W}_O)_{|s|+i+1,:}\right).$$

J.2. Failure of Shift Invariant PE for Copying

In this subsection, we prove Theorem 3.1.

First we give an equivalent form of this model successfully does copying. We consider the last token of $\mathsf{Str}(s,i)$ and denote it by $C(s,i)$. For each $\mathtt{C} \in \mathcal{V}$, notice that $\boldsymbol{w}_{\mathtt{C}}\boldsymbol{V}^\top\boldsymbol{W}_O$ is a 2-dimensional vector, we denote $\Delta_{\mathtt{C}}$ as the first coordinate of it minus the second coordinate of it. Moreover, we define $\boldsymbol{q}_{\mathtt{C}} = \boldsymbol{w}_{\mathtt{C}}\boldsymbol{Q}^\top, \boldsymbol{k}_{\mathtt{C}} = \boldsymbol{w}_{\mathtt{C}}\boldsymbol{K}^\top$. To avoid confusion, when we refer to $\boldsymbol{q}_{\mathtt{C}}$ and $\boldsymbol{k}_{\mathtt{C}}$ for $\mathtt{C} = 0, 1$, we write like $\boldsymbol{q}_{(0)}$ to indicate that the index is a token rather than a position. Therefore, $[P_{\boldsymbol{\theta}}(\mathsf{Str}(s,i))]_{s_i+1} > \frac{1}{2}$ is equivalent to that:

$$(-1)^{s_i}\sum_{j=1}^{|s|+i+1}\exp(\boldsymbol{A}_{|s|+i+1,j})\Delta_{\mathsf{Str}(s,i)_j} > 0, \tag{1}$$

where $\boldsymbol{A}$ is the attention matrix for input $\boldsymbol{X}(s,i)$.

Now, assume that the model can successfully solve copying with length at most 5. We prove it cannot be true by contradiction. We consider the above property for the following four pairs of (s,i):

- $s_1 = (00010), i_1 = 2$
- $s_2 = (01000), i_2 = 2$
- $s_3 = (100), i_3 = 3$
- $s_4 = (001), i_4 = 3$

Consider

$$u := \mathsf{Str}(s_1,i_1) = (0,0,0,1,0,\texttt{<\textbackslash n>},\texttt{<O:>},0),$$
$$v := \mathsf{Str}(s_2,i_2) = (0,1,0,0,0,\texttt{<\textbackslash n>},\texttt{<O:>},0).$$

We can rewrite Equation (1) for these two strings (notice that by definition of our shift-invariant model, $\boldsymbol{A}_{|s|+i+1,j} = f(|s|+i+1-j, \boldsymbol{q}_{|s|+i+1}, \boldsymbol{k}_j)$):

$$\sum_{\substack{1\le j\le 8\\ j\neq 2,4}}\exp(f(8-j,\boldsymbol{q}_{(0)},\boldsymbol{k}_{u_j}))\Delta_{u_j} + \exp(f(6,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(0)}))\Delta_{(0)} + \exp(f(4,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(1)}))\Delta_{(1)} > 0$$
$$\sum_{\substack{1\le j\le 8\\ j\neq 2,4}}\exp(f(8-j,\boldsymbol{q}_{(0)},\boldsymbol{k}_{v_j}))\Delta_{v_j} + \exp(f(6,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(1)}))\Delta_{(1)} + \exp(f(4,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(0)}))\Delta_{(0)} < 0.$$

By comparing the above two formula, and notice that $u_j = v_j$ for $j \neq 2, 4$, we have

$$\begin{aligned}&\exp(f(6,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(0)}))\Delta_{(0)} + \exp(f(4,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(1)}))\Delta_{(1)}\\ >&\exp(f(6,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(1)}))\Delta_{(0)} + \exp(f(4,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(0)}))\Delta_{(1)}.\end{aligned} \tag{2}$$

Now, we similarly compare Equation (1), and denote

$$u' := \mathsf{Str}(s_3,i_3) = (1,0,0,\texttt{<\textbackslash n>},\texttt{<O:>},1,0),$$
$$v' := \mathsf{Str}(s_4,i_4) = (0,0,1,\texttt{<\textbackslash n>},\texttt{<O:>},0,0).$$

We have

$$\sum_{\substack{1\le j\le 7\\ j\neq 1,3}}\exp(f(7-j,\boldsymbol{q}_{(0)},\boldsymbol{k}_{u'_j}))\Delta_{u'_j} + \exp(f(6,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(1)}))\Delta_{(1)} + \exp(f(4,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(0)}))\Delta_{(0)} > 0$$
$$\sum_{\substack{1\le j\le 7\\ j\neq 1,3}}\exp(f(7-j,\boldsymbol{q}_{(0)},\boldsymbol{k}_{v'_j}))\Delta_{v'_j} + \exp(f(6,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(0)}))\Delta_{(0)} + \exp(f(4,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(1)}))\Delta_{(1)} < 0.$$

Therefore,

$$\begin{aligned}&\exp(f(6,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(0)}))\Delta_{(0)}+\exp(f(4,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(1)}))\Delta_{(1)}\\<&\exp(f(6,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(1)}))\Delta_{(0)}+\exp(f(4,\boldsymbol{q}_{(0)},\boldsymbol{k}_{(0)}))\Delta_{(1)},\end{aligned}$$

which contradicts to Equation (2). Thus we proved Theorem 3.1.

□

K. 2-Layer RoPE can Represent Copy

In this section, we show the existence of a Transformer that can exactly perform the copy task for sequence lengths significantly larger than the norm constraint.

K.1. Prompt Template

Our proof in this section needs a new prompt template. We consider the copy task defined over the vocabulary

$$\mathcal{V}=\{\texttt{A1},\ \texttt{A2},\texttt{B1},\ \texttt{B2},0,1,\texttt{<\textbackslash n>},\texttt{<O:>}\}.$$

An input sequence has the form

$$(\texttt{A1},\ \texttt{A2},\ s,\ \texttt{<\textbackslash n>},\ \texttt{B1},\ \texttt{B2},\ \texttt{<O:>},\ s')$$

where $s\in\{0,1\}^*$ is a binary string and s' is a prefix of s. Here, $\texttt{A1},\texttt{A2},\texttt{B1},\texttt{B2}$ are special prompt tokens. The token $\texttt{<O:>}$ marks the beginning of the prefix s', while $\texttt{<\textbackslash n>}$ indicates the end of the original string s.

The objective is to predict the next token of s following a strict prefix s' of s. Therefore, the target output is the next bit in s after s'.

K.2. Architecture Setup

We consider the following simplified transformer architecture.

Let $d>50$ be an even number. The embeddings $\boldsymbol{w}_0,\boldsymbol{w}_1,\boldsymbol{w}_{\texttt{A1}},\boldsymbol{w}_{\texttt{A2}},\boldsymbol{w}_{\texttt{B1}},\boldsymbol{w}_{\texttt{B2}},\boldsymbol{w}_{\texttt{<O:>}},\boldsymbol{w}_{\texttt{<\textbackslash n>}}\in\mathbb{R}^d$ are unit vectors and orthogonal to each other. The input is converted to $\boldsymbol{X}^{(0)}\in\mathbb{R}^{L\times d}$, where L is the total input length and d is the embedding size. Then we have one RoPE attention layer

$$\boldsymbol{H}^{(1)}\ =\ \mathsf{Attn}_1(\boldsymbol{X}^{(0)})=\mathcal{S}\left(\mathrm{MASK}\left(R\left(\boldsymbol{X}^{(0)}\boldsymbol{Q}^{(0)}\right)\left(R\left(\boldsymbol{X}^{(0)}\boldsymbol{K}^{(0)}\right)\right)^\top\right)\right)\boldsymbol{X}^{(0)}(\boldsymbol{V}^{(0)})^\top\ \in\ \mathbb{R}^{L\times d}.$$

Here, $\boldsymbol{K}^{(0)},\boldsymbol{Q}^{(0)},\boldsymbol{V}^{(0)}\in\mathbb{R}^{d\times d}$, and $R:\mathbb{R}^{L\times d}\to\mathbb{R}^{L\times d}$ is the RoPE function. Specifically, for an input $\boldsymbol{X}\in\mathbb{R}^{L\times d}$, the i-th row X_i for $1\le i\le L$ is mapped to $X_iR_i^\top$, where

$$R_i\ =\ \begin{bmatrix}\cos(i\beta_1)&-\sin(i\beta_1)&&&\\\sin(i\beta_1)&\cos(i\beta_1)&&&\\&&\cos(i\beta_2)&-\sin(i\beta_2)&\\&&\sin(i\beta_2)&\cos(i\beta_2)&\\&&&&\ddots\end{bmatrix}\ \in\ \mathbb{R}^{d\times d}.$$

Here, β_j is the base frequency associated with the $(2j-1,2j)$ coordinate pair. In our setting, these β_j's are arbitrary hyperparameters. Thus, R_i acts block-diagonally on X_i, rotating each two-dimensional subspace $(X_{i,2j-1},X_{i,2j})$ by angle $i\beta_j$. The causal mask operator $\mathrm{MASK}(\cdot)$ enforces the autoregressive constraint: for $\boldsymbol{Z}\in\mathbb{R}^{L\times L}$,

$$\mathrm{MASK}(\boldsymbol{Z})_{ij}\ =\ \begin{cases}\boldsymbol{Z}_{ij},&j\le i,\\-\infty,&j>i,\end{cases}$$

so that position i cannot attend to any future position $j>i$. The operator $\mathcal{S}(\cdot)$ applies row-wise softmax normalization: for a matrix $\boldsymbol{A}\in\mathbb{R}^{L\times L}$,

$$\mathcal{S}(\boldsymbol{A})_{ij}\ =\ \frac{\exp(\boldsymbol{A}_{ij})}{\sum_{k=1}^{L}\exp(\boldsymbol{A}_{ik})},\quad 1\le i,j\le L.$$

Then, the attention layer output come through a normalization layer (here we don't consider MLP layer for simplicity), and then added to $\boldsymbol{X}^{(0)}$ by a residual link. Formally, we have

$$\bar{\boldsymbol{H}}^{(1)} = \mathrm{RMSNorm}(\boldsymbol{H}^{(1)})$$
$$\boldsymbol{X}^{(1)} = \boldsymbol{X}^{(0)} + \bar{\boldsymbol{H}}^{(1)}$$

where

$$\mathrm{RMSNorm}([\boldsymbol{h}_1, \ldots, \boldsymbol{h}_L]^\top) = \left[\frac{\lambda\, \boldsymbol{h}_1}{\|\boldsymbol{h}_1\|_2}, \ldots, \frac{\lambda\, \boldsymbol{h}_L}{\|\boldsymbol{h}_L\|_2}\right]^\top.$$

The second attention layer doesn't use positional encoding. Formally,

$$\boldsymbol{H}^{(2)} \;=\; \mathsf{Attn}_2\big(\boldsymbol{X}^{(1)}\big) = \mathcal{S}\left(\mathrm{MASK}\left((\boldsymbol{X}^{(1)}(\boldsymbol{K}^{(1)})^\top)\left((\boldsymbol{X}^{(1)}(\boldsymbol{Q}^{(1)})^\top)\right)^\top\right)\right)\boldsymbol{X}^{(1)}(\boldsymbol{V}^{(1)})^\top \;\in\; \mathbb{R}^{L\times d}.$$

Finally, the hidden representation $\boldsymbol{H}^{(2)} \in \mathbb{R}^{L\times d}$ is projected to the vocabulary space via a linear map $\boldsymbol{W}_O \in \mathbb{R}^{d\times|\mathcal{V}|}$:

$$\boldsymbol{Y} \;=\; \boldsymbol{H}^{(2)}\boldsymbol{W}_O \;\in\; \mathbb{R}^{L\times|\mathcal{V}|},$$

where $\boldsymbol{Y}$ contains the output logits.

Let $\mathcal{V}^+ = \mathcal{V}\cup\mathcal{V}^2\cup\mathcal{V}^3\cup\cdots$ denote the set of all nonempty finite sequences over $\mathcal{V}$. We denote the Transformer architecture as a function

$$F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})} : \mathcal{V}^+ \to \mathbb{R}^{L\times|\mathcal{V}|}, \quad F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x) = \boldsymbol{Y},$$

where $x \in \mathcal{V}^+$ is the input sequence, $\boldsymbol{\theta}$ are the learnable parameters, d is the embedding dimension, and $\boldsymbol{\beta}$ denotes the hyperparameters $\beta_i, 1 \le i \le d/2$.

The prediction distribution is obtained by applying the row-wise softmax:

$$P_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x) = \mathrm{softmax}\Big(F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x)\Big),$$

$$\Big[P_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x)\Big]_{i,j} = \frac{\exp\Big(F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x)_{i,j}\Big)}{\sum_{k=1}^{|\mathcal{V}|}\exp\Big(F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x)_{i,k}\Big)},$$

where $x \in \mathcal{V}^+$ is the input sequence, i indexes positions, and j indexes vocabulary symbols.

K.3. Existence of a Length-Generalizing Solution

In this section, we show the existence of a transformer architecture that can solve copy task for length larger than parameter norm. Therefore, we show that the failure of transformer to length generalize on copy task isn't an expressiveness issue.

Theorem K.1. *There exists a constant embedding dimension $d \ge 1$ such that for any norm constraint $\rho > 10000$, there exists RoPE freqencies $\boldsymbol{\beta}$ and a set of parameters $\boldsymbol{\theta} \in \{\|\boldsymbol{\theta}\|_2 \le \rho\}$ such that Transformer with embedding dimension d, RoPE freq $\boldsymbol{\beta}$, parameters $\boldsymbol{\theta}$ can perfectly perform copying for all lengths $1 \le L \le O(\rho^2/\log\rho)$:*

$$\forall i \in [n]: \quad \arg\max_{j\in\mathcal{V}}\Big[P_{\boldsymbol{\theta}}^{(d,\beta)}(\texttt{A1},\ \texttt{A2},\ s,\ \texttt{<\textbackslash n>},\ \texttt{B1},\ \texttt{B2},\ \texttt{<O:>},\ s)\Big]_{n+5+i,j} = s_i.$$

Proof. We fix $d = 100$. Let $\beta_1 = 0$ and β_i be $r_i\pi$, where r_i's are irrational numbers for $2 \le i \le d/2$. In this section, we fix $L_0 = \rho^2/(100\log\rho), \alpha = \rho/10$. We prove the existence of the parameters that can prefectly do copy for $1 \le L \le L_0$. Let $n = |s|$ be the length of the binary string to be copied.

As there are only 8 kinds of token, there exists $\boldsymbol{K}^{(0)}, \boldsymbol{Q}^{(0)}$ such that:

$$\boldsymbol{K}^{(0)}\boldsymbol{w}_{\texttt{0}} = \boldsymbol{K}^{(0)}\boldsymbol{w}_{\texttt{1}} = \boldsymbol{K}^{(0)}\boldsymbol{w}_{\texttt{<\textbackslash n>}} = \boldsymbol{K}^{(0)}\boldsymbol{w}_{\texttt{<O:>}} = \boldsymbol{0}, \boldsymbol{K}^{(0)}\boldsymbol{w}_{\texttt{A1}} = \alpha\,\mathbf{e}_1,$$
$$\boldsymbol{K}^{(0)}\boldsymbol{w}_{\texttt{A2}} = \alpha\mathbf{e}_1 + \mathbf{e}_4, \boldsymbol{K}^{(0)}\boldsymbol{w}_{\texttt{B1}} = 2\alpha\,\mathbf{e}_1, \boldsymbol{K}^{(0)}\boldsymbol{w}_{\texttt{B2}} = 2\alpha\,\mathbf{e}_1 + \mathbf{e}_4,$$
$$\boldsymbol{Q}^{(0)}\boldsymbol{w}_{\texttt{0}} = \boldsymbol{Q}^{(0)}\boldsymbol{w}_{\texttt{1}} = \boldsymbol{Q}^{(0)}\boldsymbol{w}_{\texttt{<\textbackslash n>}} = \boldsymbol{Q}^{(0)}\boldsymbol{w}_{\texttt{<O:>}} = \mathbf{e}_1 + \mathbf{e}_3,$$
$$\boldsymbol{Q}^{(0)}\boldsymbol{w}_{\texttt{A1}} = \boldsymbol{Q}^{(0)}\boldsymbol{w}_{\texttt{A2}} = \boldsymbol{Q}^{(0)}\boldsymbol{w}_{\texttt{B1}} = \boldsymbol{Q}^{(0)}\boldsymbol{w}_{\texttt{B2}} = \boldsymbol{0}.$$

Let $\boldsymbol{v}_{\texttt{a}}, \boldsymbol{v}_{\texttt{b}}, \boldsymbol{v}_{\texttt{c}}, \boldsymbol{v}_{\texttt{d}}$ be unit vectors that are orthogonal to $\boldsymbol{w}_0, \boldsymbol{w}_1, \boldsymbol{w}_{\texttt{a}}, \boldsymbol{w}_{\texttt{b}}, \boldsymbol{w}_{\texttt{c}}, \boldsymbol{w}_{\texttt{<\textbackslash n>}}, \boldsymbol{w}_{\texttt{<O:>}}$. Let

$$\boldsymbol{V}^{(0)}\boldsymbol{w}_{\texttt{C}} = \boldsymbol{v}_{\texttt{C}}, \texttt{C} = \texttt{A1,A2,B1,B2}; \boldsymbol{V}^{(0)}\boldsymbol{w}_{\texttt{0,1,<\textbackslash n>,<O:>}} = \boldsymbol{0}.$$

Therefore, we have

$$\|\boldsymbol{K}^{(0)}\|_2^2 + \|\boldsymbol{Q}^{(0)}\|_2^2 + \|\boldsymbol{V}^{(0)}\|_2^2 = 10\alpha^2 + 14$$

Let s_0' be <O:>. Let s_{n+1} be <\n>. In this construction, for $1 \le i \le n+1$, we have

$$\begin{aligned}
\boldsymbol{w}_{\texttt{A1}}^\top \left(\boldsymbol{K}^{(0)}\right)^\top R_{i+1}\boldsymbol{Q}^{(0)}\boldsymbol{w}_{s_i} &= \alpha \\
\boldsymbol{w}_{\texttt{A2}}^\top \left(\boldsymbol{K}^{(0)}\right)^\top R_{i}\boldsymbol{Q}^{(0)}\boldsymbol{w}_{s_i} &= \alpha + \sin(i\beta_2) \\
\boldsymbol{w}_{\texttt{B1}}^\top \left(\boldsymbol{K}^{(0)}\right)^\top R_{i+1}\boldsymbol{Q}^{(0)}\boldsymbol{w}_{s'_{i-1},\texttt{<O:>}} &= 2\alpha \\
\boldsymbol{w}_{\texttt{B2}}^\top \left(\boldsymbol{K}^{(0)}\right)^\top R_{i}\boldsymbol{Q}^{(0)}\boldsymbol{w}_{s'_{i-1}} &= 2\alpha + \sin(i\beta_2).
\end{aligned}$$

We define

$$\begin{aligned}
\boldsymbol{h}_i &= \frac{1}{\sqrt{1+\exp^2(\sin(i\beta_2))}}\boldsymbol{v}_{\texttt{A1}} + \frac{\exp(\sin(i\beta_2))}{\sqrt{1+\exp^2(\sin(i\beta_2))}}\boldsymbol{v}_{\texttt{A2}}, 1 \le i \le n+1 \\
\tilde{\boldsymbol{h}}_i &= \frac{1}{\sqrt{1+\exp^2(\sin(i\beta_2))}}\boldsymbol{v}_{\texttt{B1}} + \frac{\exp(\sin(i\beta_2))}{\sqrt{1+\exp^2(\sin(i\beta_2))}}\boldsymbol{v}_{\texttt{B2}}, 1 \le i \le n+1.
\end{aligned}$$

As $\boldsymbol{V}^{(0)}\boldsymbol{w}_{\texttt{0,1,<\textbackslash n>,<O:>}} = \boldsymbol{0}$, we have

$$\bar{\boldsymbol{H}}_{i+2}^{(1)} = \lambda \boldsymbol{h}_i, 1 \le i \le n+1.$$

Now we consider $\bar{\boldsymbol{H}}_{5+L+i}$. Notice that

$$\begin{aligned}
\frac{1}{\lambda}\bar{\boldsymbol{H}}_{5+L+i}^{(1)} =& \frac{\exp(-\alpha)}{\sqrt{1+\exp^2(\sin(i\beta_2)) + \exp^2(-\alpha) + \exp^2(\sin((i+3+L)\beta_2)-\alpha))}}\boldsymbol{v}_{\texttt{A1}} \\
&+ \frac{\exp(\sin((i+3+L)\beta_2)-\alpha)}{\sqrt{1+\exp^2(\sin(i\beta_2)) + \exp^2(-\alpha) + \exp^2(\sin((i+3+L)\beta_2)-\alpha))}}\boldsymbol{v}_{\texttt{A2}} \\
&+ \frac{1}{\sqrt{1+\exp^2(\sin(i\beta_2)) + \exp^2(-\alpha) + \exp^2(\sin((i+3+L)\beta_2)-\alpha))}}\boldsymbol{v}_{\texttt{B1}} \\
&+ \frac{\exp(\sin(i\beta_2))}{\sqrt{1+\exp^2(\sin(i\beta_2)) + \exp^2(-\alpha) + \exp^2(\sin((i+3+L)\beta_2)-\alpha))}}\boldsymbol{v}_{\texttt{B2}}.
\end{aligned}$$

Therefore,

$$\|\bar{\boldsymbol{H}}_{5+n+i}^{(1)} - \lambda\tilde{\boldsymbol{h}}_i\|_2^2 \le \lambda^2\exp(-2\alpha)(1+e^2+e^2(e^2+1)^2) \le 10000\lambda^2\exp(-2\alpha).$$

We need to prove the following lemma:

Lemma K.2. *Let $\boldsymbol{h}_0 = \frac{1}{\sqrt{2}}\boldsymbol{v}_{A1} + \frac{1}{\sqrt{2}}\boldsymbol{v}_{A2}, \boldsymbol{h}_{-1} = \boldsymbol{v}_{A1}$. There exists $\boldsymbol{\beta}$ such that for any $n < L_0$ and $-1 \le i < j \le n+1$, we have $\langle \boldsymbol{h}_i, \boldsymbol{h}_j\rangle \le 1 - \frac{1}{72L_0^2}$.*

Proof. Let $\beta_2 = \frac{\pi}{4L_0}$. For $0 \le i \le n+1$, we let $a_i = \exp(\sin(i\beta))$. Therefore, for any $0 \le i < j \le n+1$, as $n < L_0$,

$$\sin(j\beta) - \sin(i\beta) = \int_{j\beta}^{i\beta}\cos(x)dx \ge \frac{\beta}{\sqrt{2}}(j-i) \ge \frac{1}{2L_0}.$$

Therefore, for any $0 \le i < j \le n+1$,

$$a_j - a_i = \int_{\sin(i\beta)}^{\sin(j\beta)}\exp(x)dx \ge \frac{1}{e}(\sin(j\beta)-\sin(i\beta)) \ge \frac{1}{2eL_0} \ge \frac{1}{6L_0}.$$

This shows

$$|\arctan a_i - \arctan a_j| \geq \frac{1}{6L_0}.$$

As $\frac{1}{e} \leq a_i \leq e$, we have

$$\langle \boldsymbol{h}_i, \boldsymbol{h}_j \rangle = \cos(\arctan a_i - \arctan a_j) \leq 1 - \frac{1}{72L_0^2}.$$

Moreover, for each $0 \leq i \leq n+1$,

$$\langle \boldsymbol{h}_i, \boldsymbol{h}_{-1} \rangle = \frac{1}{\sqrt{1+a_i^2}} \leq 1 - \frac{1}{72L_0^2}.$$

Thus, the lemma follows. □

Let $\lambda = 10\sqrt{L_0 \log L_0}$. In the second layer, we let

$$\boldsymbol{K}^{(1)^\top} = 10\sqrt{L_0}\left(\boldsymbol{v}_{\mathtt{A}1}\boldsymbol{v}_{\mathtt{B}1}^\top + \boldsymbol{v}_{\mathtt{A}2}\boldsymbol{v}_{\mathtt{B}2}^\top\right), \boldsymbol{Q}^{(1)} = 10\sqrt{L_0}\left(\boldsymbol{v}_{\mathtt{B}1}\boldsymbol{v}_{\mathtt{B}1}^\top + \boldsymbol{v}_{\mathtt{B}2}\boldsymbol{v}_{\mathtt{B}2}^\top\right),$$

so

$$\boldsymbol{K}^{(1)^\top}\boldsymbol{Q}^{(1)} = 100L_0\left(\boldsymbol{v}_{\mathtt{A}1}\boldsymbol{v}_{\mathtt{B}1}^\top + \boldsymbol{v}_{\mathtt{A}2}\boldsymbol{v}_{\mathtt{B}2}^\top\right).$$

For any $1 \leq i \leq n+1, -1 \leq j \leq n+1, j \neq i$,

$$\langle \bar{\boldsymbol{H}}_{5+i+n}, \boldsymbol{h}_i - \boldsymbol{h}_j \rangle \geq \lambda \langle \boldsymbol{h}_i, \boldsymbol{h}_i - \boldsymbol{h}_j \rangle - 2 \cdot 100\lambda \exp(-\alpha) \geq \frac{\lambda}{100L_0^2}.$$

Therefore, for $1 \leq i \leq n+1$, for any $1 \leq j \leq L, j \neq i+2$, we have

$$\bar{\boldsymbol{H}}_j \boldsymbol{K}^{(1)^\top}\boldsymbol{Q}^{(1)}\bar{\boldsymbol{H}}_{5+i+h}^\top \leq \bar{\boldsymbol{H}}_{i+2}\boldsymbol{K}^{(1)^\top}\boldsymbol{Q}^{(1)}\bar{\boldsymbol{H}}_{5+i+h}^\top - 100\log L_0.$$

Therefore, in the second layer, the attention of s_i' to s_{i+1} is larger than $1/2$,$i = 0, 1, \cdots, L$. We let $\boldsymbol{v}_0, \boldsymbol{v}_1, \boldsymbol{v}_{<\backslash \mathtt{n}>}$ be unit vectors that are orthogonal to the $\boldsymbol{v}_{\mathtt{C}}, \boldsymbol{w}_{\mathtt{C}}$ defined before (here $\mathtt{C}$ denotes arbitrary characters). We let

$$\boldsymbol{V}^{(1)}\boldsymbol{w}_{\mathtt{C}} = \boldsymbol{v}_{\mathtt{C}}, \mathtt{C} = 0, 1, <\backslash \mathtt{n}>.$$

At last, we let $\boldsymbol{W}_O$ be

$$\boldsymbol{W}_O = \boldsymbol{v}_0\boldsymbol{e}_0^\top + \boldsymbol{v}_1\boldsymbol{e}_1^\top + \boldsymbol{v}_{<\backslash \mathtt{n}>}\boldsymbol{e}_{<\backslash \mathtt{n}>}^\top.$$

Here $\boldsymbol{e}_{\mathtt{C}}$ denotes the one-hot vector corresponding to token $\mathtt{C} \in \{0, 1, <\backslash \mathtt{n}>\}$. The prediction is always true.

For our construction, we have

$$\begin{aligned}\|\boldsymbol{\theta}\|^2 &= \|\boldsymbol{K}^{(0)}\|_2^2 + \|\boldsymbol{Q}^{(0)}\|_2^2 + \|\boldsymbol{V}^{(0)}\|_2^2 + \|\boldsymbol{K}^{(1)}\|_2^2 + \|\boldsymbol{Q}^{(1)}\|_2^2 + \|\boldsymbol{V}^{(1)}\|_2^2 + \|\boldsymbol{W}_O\|_2^2 + \lambda^2 \\ &\leq \frac{\rho^2}{10} + 100 + 400L_0 + 100L_0\log L_0 \leq \rho^2.\end{aligned}$$

Thus the construction satisfies the norm constraint, so the theorem follows. □

Remark K.3. One may argue that the above construction does not provide a solution that works for all sequence lengths. Indeed, such a solution cannot exist under bounded parameters: if all weights are uniformly bounded, then the attention assigned to any individual token is at most $O(1/L)$, which vanishes as the sequence length L grows. Consequently, the copy task cannot be perfectly solved for arbitrarily long sequences.

L. One-Layer 2D-RoPE Provably Length Generalize on Copy

L.1. Setup

In this subsection, we define our setup in detail.

For copying task, we still consider the input format $(s, \texttt{<\textbackslash n>}, \texttt{<O:>}, s_{:-1})$, where s is a binary string of length at most L, and $s_{:-1}$ denotes the string obtained by deleting the last symbol of s.

Let d be a positive integer divisible by 4. The embeddings $\boldsymbol{w}_0, \boldsymbol{w}_1, \boldsymbol{w}_{\texttt{<O:>}}, \boldsymbol{w}_{\texttt{<\textbackslash n>}} \in \mathbb{R}^d$ are one-hot embeddings of the four tokens. The input is converted to $\boldsymbol{X} \in \mathbb{R}^{T\times d}$, where T is the total input length and d is the embedding size. Then we have one 2D-RoPE attention layer

$$\boldsymbol{Y} = \mathsf{Attn}(\boldsymbol{X}) = \mathcal{S}\left(\mathrm{MASK}\left(R\left(\boldsymbol{X}\boldsymbol{Q}\right)\left(R\left(\boldsymbol{X}\boldsymbol{K}\right)\right)^\top\right)\right)\boldsymbol{X}\boldsymbol{V}^\top \in \mathbb{R}^{T\times 2}.$$

Here, $\boldsymbol{K}, \boldsymbol{Q} \in \mathbb{R}^{d\times d}$, $\boldsymbol{V} \in \mathbb{R}^{2\times d}$ and $R : \mathbb{R}^{L\times d} \to \mathbb{R}^{L\times d}$ is the 2D-RoPE function. Specifically, for an input $\boldsymbol{X} \in \mathbb{R}^{L\times d}$ and $1 \le i \le L$, let the i-th token of the input is at position (x_i, y_i) (which means that it is at x_ith column and y_ith row). Then the row X_i for $1 \le i \le L$ is mapped to $X_i\boldsymbol{R}_{x_i,y_i}^\top$, where

$$\boldsymbol{R}_{x_i,y_i} \;=\; \mathrm{diag}\big(\boldsymbol{R}_1(x_i), \ldots, \boldsymbol{R}_{d/4}(x_i),\; \boldsymbol{R}_{d/4+1}(y_i), \ldots, \boldsymbol{R}_{d/2}(y_i)\big) \;\in\; \mathbb{R}^{d\times d},$$

and for each $1 \le j \le d/2$, the block $\boldsymbol{R}_j(\cdot) \in \mathbb{R}^{2\times 2}$ is the RoPE rotation matrix

$$\boldsymbol{R}_j(t) \;:=\; \begin{bmatrix}\cos(t\beta_j) & -\sin(t\beta_j)\\ \sin(t\beta_j) & \cos(t\beta_j)\end{bmatrix}.$$

Here, β_j are the base frequencies associated with the $(2j-1, 2j)$ coordinate pair. In our setting, these frequencies are arbitrary hyperparameters. In other words, if we denote the attention matrix

$$\boldsymbol{A} := \mathrm{MASK}\left(R\left(\boldsymbol{X}\boldsymbol{Q}\right)\left(R\left(\boldsymbol{X}\boldsymbol{K}\right)\right)^\top\right),$$

then

$$\boldsymbol{A}_{i,j} = \begin{cases} -\infty & i < j \\ f(i-j, \boldsymbol{q}_i, \boldsymbol{k}_j) & i \ge j \end{cases}$$

for

$$f(i-j, \boldsymbol{q}_i, \boldsymbol{k}_j) = \boldsymbol{q}_i^\top R_{x_j-x_i, y_j-y_i}\boldsymbol{k}_j.$$

The causal mask operator $\mathrm{MASK}(\cdot)$ enforces the autoregressive constraint: for $\boldsymbol{Z} \in \mathbb{R}^{T\times T}$,

$$\mathrm{MASK}(\boldsymbol{Z})_{ij} \;=\; \begin{cases}\boldsymbol{Z}_{ij}, & j \le i,\\ -\infty, & j > i.\end{cases}$$

Let $\mathcal{V}^+ = \mathcal{V} \cup \mathcal{V}^2 \cup \mathcal{V}^3 \cup \cdots$ denote the set of all nonempty finite sequences over $\mathcal{V}$. We denote the Transformer architecture as a function

$$F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})} : \mathcal{V}^+ \to \mathbb{R}^{L\times|\mathcal{V}|}, \quad F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x) = \boldsymbol{Y},$$

where $x \in \mathcal{V}^+$ is the input sequence, $\boldsymbol{\theta}$ are the learnable parameters, d is the embedding dimension, and $\boldsymbol{\beta}$ denotes the hyperparameters $\beta_i, 1 \le i \le d/2$. Notice that each row of $\boldsymbol{Y}$ is a two-dimensional vector, where the first entry corresponds to the logit for 0 and the second corresponds to the logit for 1.

The prediction distribution is obtained by applying the row-wise softmax:

$$P_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x) = \mathrm{softmax}\Big(F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x)\Big),$$

$$\Big[P_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x)\Big]_{i,j} = \frac{\exp\Big(F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x)_{i,j}\Big)}{\sum_{k=1}^{|\mathcal{V}|}\exp\Big(F_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(x)_{i,k}\Big)},$$

where $x \in \mathcal{V}^+$ is the input sequence, i indexes positions, and j indexes vocabulary symbols.

For a parameter set $\boldsymbol{\theta}$, the loss is defined as the cross entropy loss for the last $|s|$ tokens:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{s\sim\mathcal{D}}\left[-\sum_{i=1}^{|s|}\log\left[P_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(s, \texttt{<\textbackslash n>}, \texttt{<O:>}, s_{:i})\right]_{|s|+1+i,s_i+1}\right]. \tag{3}$$

Here, $\mathcal{D}$ is the distribution of s. We make the following assumptions:

- For every s in the support of $\mathcal{D}$, we have $|s| \le L$.
- For every length $1 \le \ell \le L$, every string $s \in \{0,1\}^\ell$ has positive probability under $\mathcal{D}$. We denote the minimum such probability by p.

L.2. Expressive Power of One-Layer 2D-RoPE

In this section, we prove the following theorem:

Theorem L.1. *Assume that the 2D-RoPE dimension $d \ge 50$. For any $\rho > 1000$, let $\beta_i \sim \mathrm{Unif}[0, 2\pi]$ independently. Then with probability at least $\left(1 - \frac{1}{\rho}\right)$ there exists a 1-layer transformer $P_{\boldsymbol{\theta}}$ with 2D-RoPE such that:*

1. *$\|\boldsymbol{\theta}\|_\infty \le \rho$;*
2. *For any binary string with length $|s| \le \rho^{d/6}$, the transformer can correctly copy s.*

Since $\boldsymbol{w}_{\texttt{C}}$ are one-hot embeddings, there exists $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ such that

$$\boldsymbol{q}_{\texttt{C}} = \begin{bmatrix} M \\ M \\ \vdots \\ M \\ M \\ M \\ \vdots \\ M \end{bmatrix}, \boldsymbol{k}_{\texttt{C}} = \begin{bmatrix} M \\ M \\ \vdots \\ M \\ -M \\ -M \\ \vdots \\ -M \end{bmatrix}, \boldsymbol{v}_{(0)} = \begin{bmatrix} M \\ -M \end{bmatrix}, \boldsymbol{v}_{(1)} = \begin{bmatrix} -M \\ M \end{bmatrix}, \boldsymbol{v}_{\texttt{<\textbackslash n>}} = \boldsymbol{v}_{\texttt{<O:>}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

where there are $d/2$ M's and $d/2$ $(-M)$'s for $\boldsymbol{k}_{\texttt{C}}$.

We first need this technical lemma:

Lemma L.2. *Let $\theta_1, \ldots, \theta_m \sim \mathrm{Unif}[0, 2\pi]$ be independent. Then for any $\delta \in (0, 0.1)$, with probability at most $\delta^{m/2}$, we have*

$$\sum_{i=1}^{m}\cos(\theta_i) \ge m - \delta.$$

Proof. First consider one $\theta \sim \mathrm{Unif}[0, 2\pi]$. Notice that $\cos\theta = 1 - 2\sin^2\frac{\theta}{2}$. Therefore, if $\cos\theta \ge 1 - \delta$, then $\sin^2\frac{\theta}{2} \le \frac{\delta}{2}$. Since $\delta < 0.1$, we must have $0 \le \theta \le \sqrt{2\delta}$ or $2\pi - \sqrt{2\delta} \le \theta \le 2\pi$. Therefore, the probability of $\sin^2\frac{\theta}{2} \le \frac{\delta}{2}$ is at most $\sqrt{\delta}$.

Notice that $\sum_{i=1}^m \cos(\theta_i) \ge m - \delta$ indicates that $\cos(\theta_i) \ge 1 - \delta$ for each i. Therefore, the result follows by the independency of these θ_i's. □

Now we consider back our construction of $\boldsymbol{q}, \boldsymbol{k}$. For any s, i, we consider the attention matrix $\boldsymbol{A}$. By construction, we have

$$\boldsymbol{A}_{|s|+i+1,i} = \frac{d}{2}\cdot M^2 - 2M^2\sum_{j=d/4+1}^{d/2}\cos(\beta_j).$$

For any $1 \le j \le |s|+1$,

$$\boldsymbol{A}_{|s|+i+1,j} = 2M^2 \sum_{k=1}^{d/4} \cos((j-i)\beta_k) - 2M^2 \sum_{j=d/4+1}^{d/2} \cos(\beta_j).$$

For any $|s|+2 \le j \le |s|+i+1$,

$$\boldsymbol{A}_{|s|+i+1,j} = 2M^2 \sum_{k=1}^{d/4} \cos((j-|s|-2+i)\beta_k) - \frac{d}{2} \cdot M^2.$$

Notice that for any non-zero integer m, $\cos(m\beta_j)$ has the same distribution as $\cos(\theta)$ for $\theta \sim \mathrm{Unif}[0, 2\pi]$. Therefore, by Lemma L.2, with probability at least $1-(L+1)\cdot\delta^{d/4}$, we have:

- For each $1 \le m \le L$, $\sum_{k=1}^{d/4} \cos((j-i)\beta_k) \le d/4 - \delta$;
- $\sum_{j=d/4+1}^{d/2} \cos(\beta_j) \le d/4 - \delta$.

In this case, we always have

$$\boldsymbol{A}_{|s|+i+1,i} \ge \boldsymbol{A}_{|s|+i+1,j} + 2M^2\delta.$$

For each other position, with probability at most $\delta^{d/4}$, the $\boldsymbol{q}^\top R\boldsymbol{k} \le M^2(d/2-\delta)$. Therefore, we just need $2L \cdot \exp(-M^2\delta) \le 1$ and $L^2 \cdot \delta^{d/4} \le 1$.

Therefore, we let $\delta = \frac{1}{M^{1.5}}$ and $L = M^{d/6}$. When $M > 1000$ the result is true.

L.3. Global Minimum under ℓ_∞ Norm

First we discuss the regulariazation path we choose. The previous regularization path analysis mainly focused on ℓ_2 norm regularization (Wei et al., 2019; Ataee Tarzanagh et al., 2023). Here we consider the ℓ_∞ regularization. This setup is motivated by the implicit bias of Adam and AdamW (Xie & Li, 2024; Fan et al., 2025; Zhang et al., 2024) since the experiments in this paper are all conducted on AdamW. In detail, Xie & Li (2024) proved that AdamW can converge to the KKT points of ℓ_∞ constrained optimization, and Zhang et al. (2024); Fan et al. (2025) showed that Adam converges to max-margin solutions w.r.t. ℓ_∞ norm for linear models on separable data. Mohamadi et al. (2024) also introduced a similar ℓ_∞ regularization condition in their grokking analysis.

Now, we define our setting. Notice that the loss $\mathcal{L}$ is defined on a distribution $\mathcal{D}$ on some strings with length at most L. We make the following assumptions:

Assumption L.3. There exists a $p > 0$ such that for each string, $\Pr_{x\sim\mathcal{D}}[x = s] \ge p$.

Assumption L.4. The training length L and hidden dimension d satisfies $L > d^3$ and $d > \log^3 L$. Also, $L, d > 10000$.

We also need the following assumption that the tokens <\n> and <0:> have zero value vectors. This is a natural assumption because intuitively, <\n> and <0:> cannot give any guide to predicting the next binary token.

Assumption L.5. Let $\boldsymbol{w}_{<\backslash n>}$ and $\boldsymbol{w}_{<0:>}$ be the embeddings of corresponding tokens. Then the corresponding value vectors $\boldsymbol{V}^\top \boldsymbol{w}_{<\backslash n>} = \boldsymbol{V}^\top \boldsymbol{w}_{<0:>}$ are zero vectors.

Then we let $M \to \infty$, and obtain the corresponding solution $\boldsymbol{Q}^*, \boldsymbol{K}^*, \boldsymbol{V}^*$ in a limit sense. Eventually, we can see there exists a directional convergence for $\boldsymbol{Q}^*, \boldsymbol{K}^*, \boldsymbol{V}^*$, which leads to a good length generalization property in this binary copy problem.

In this section, we consider the global minimum under the ℓ_∞ constraint:

$$\|\boldsymbol{\theta}\|_\infty \le M.$$

Our main result is the following:

Theorem L.6. *Assume that each β_i is independently sampled from* $\mathrm{Unif}[0, 2\pi]$. *Under the above assumptions, with probability at least $1 - \frac{2}{d}$, there exists $M_0 > 0$ such that for all $M > M_0$, the transformer with parameter $\boldsymbol{\theta}_M^*$ can correctly copy all binary strings with length at most $L^{\sqrt{d}/2}$.*

For each $\mathtt{c} \in \{0, 1, \texttt{<O:>}, \texttt{<\n>}\}$, let $\boldsymbol{q}_\mathtt{c} = \boldsymbol{Q}^\top \boldsymbol{w}_\mathtt{c}, \boldsymbol{k}_\mathtt{c} = \boldsymbol{K}^\top \boldsymbol{w}_\mathtt{c}$. Since $\boldsymbol{w}_\mathtt{c}$ is one-hot embedding, the requirements on $\boldsymbol{Q}$ and $\boldsymbol{K}$ asks that:

$$\|\boldsymbol{q}_\mathtt{c}\|_\infty, \|\boldsymbol{k}_\mathtt{c}\|_\infty \le M.$$

Notice that $\boldsymbol{V}$ is a $2 \times d$ matrix. For the column corresponding to the one-hot embedding $\boldsymbol{w}_\mathtt{c}$, we denote it by $(V_{\mathtt{c},0}, , V_{\mathtt{c},1})^\top$. Therefore, the ℓ_∞ constraint requires $|V_{\mathtt{c},0}|, |V_{\mathtt{c},1}| \le M$.

Since other rows of $\boldsymbol{Q}, \boldsymbol{K}$ and other columns of $\boldsymbol{V}$ don't affect the loss, we can always assume that they are 0. Thus, $\|\boldsymbol{\theta}\|_\infty \le M$ is equivalent to:

$$\|\boldsymbol{q}_\mathtt{c}\|_\infty, \|\boldsymbol{k}_\mathtt{c}\|_\infty, |V_{\mathtt{c},0}|, |V_{\mathtt{c},1}| \le M. \tag{4}$$

Now we need to consider the different part of the loss term. For each string s and each index $i \le |s|$, we consider the attention distribution of the last token under the input

$$(s, \texttt{<\n>}, \texttt{<O:>}, s_{:i}),$$

where $s_{:i}$ means the length $(i-1)$ prefix of s. For this input, the last token is s_{i-1} (and when $i = 1$, the last token is $\texttt{<O:>}$). For each $\mathtt{c} \in \mathcal{V}$, let $A_\mathtt{c}(s, i)$ denote the total attention weight assigned to all tokens equal to $\mathtt{c}$. Then the logits for predicting s_i is

$$\boldsymbol{Y}_{|s|+2+i} = \left(\sum_{\mathtt{c} \in \mathcal{V}} A_\mathtt{c}(s,i) V_{\mathtt{c},0}, \sum_{\mathtt{c} \in \mathcal{V}} A_\mathtt{c}(s,i) V_{\mathtt{c},1} \right).$$

Therefore, the loss for predicting s_i is:

$$\begin{aligned}
\mathcal{L}(s,i) :&= -\log \left[P_{\boldsymbol{\theta}}^{(d,\boldsymbol{\beta})}(s, \texttt{<\n>}, \texttt{<O:>}, s_{:-1}) \right]_{|s|+1+i, s_i+1} \\
&= -\log \frac{\exp\left(\boldsymbol{Y}_{|s|+1+i, s_i+1}\right)}{\exp\left(\boldsymbol{Y}_{|s|+1+i,1}\right) + \exp\left(\boldsymbol{Y}_{|s|+1+i,2}\right)} \\
&= \log\left(1 + \exp\left(\sum_{\mathtt{c} \in \mathcal{V}} A_\mathtt{c}(s,i) \left(V_{\mathtt{c},1-s_i} - V_{\mathtt{c},s_i} \right) \right)\right).
\end{aligned}$$

Without loss of generality, we can always assume that $V_{\mathtt{c},0} + V_{\mathtt{c},1} = 0$. Since we have assumed that $V_{\texttt{<\n>},0} = V_{\texttt{<\n>},1} = V_{\texttt{<O:>},0} = V_{\texttt{<O:>},1} = 0$, we denote $V_0 := V_{0,0}$ and $V_1 := V_{1,1}$. Therefore,

$$\mathcal{L}(s,i) = \log\left(1 + \exp\left(-2A_{s_i}(s,i)V_{s_i} + 2A_{1-s_i}(s,i)V_{1-s_i}\right)\right).$$

By definition,

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{s \sim \mathcal{D}} \left[\sum_{i=1}^{|s|} \mathcal{L}(s,i) \right].$$

For such a pair (s, i) and $1 \le j \le 1 + |s| + i$, we use $S(s, i, j)$ to denote the jth token in sequence $(s, \texttt{<\n>}, \texttt{<O:>}, s_{:i})$, and we use $D(s, i, j)$ to denote the 2-D relative position between $S(s, i, j)$ and s_{i-1}.

We define a function to characterize the global minimum. For each s, i, we define

$$\begin{aligned}
g_{s,i}^+(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &:= \max_{S(s,i,j) = s_i} \left\{ \boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,j)} \boldsymbol{k}_{s_i} \right\}; \\
g_{s,i}^-(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &:= \max_{S(s,i,j) \ne s_i} \left\{ \boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,j)} \boldsymbol{k}_{S(s,i,j)} \right\}.
\end{aligned}$$

Notice that it is well defined, because the existence of such j is promised by $S(s, i, i) = s_i$ and $S(s, i, |s| + 1) \ne s_i$. Then, we define

$$f_L(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) := \min_{s, i, |s| \le L} \left\{ g_{s,i}^+(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) - g_{s,i}^-(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \right\}.$$

First we need to prove that the original problem can be approximated by optimizing the f_L function. We have this technical lemma:

Lemma L.7. *Consider the function $g(x) = \log(1 + \exp(-2M + x))$. Then for any $x > 0$,*

$$g(x) \geq \log(1 + \exp(-2M)) + \frac{\exp(-2M)}{1 + \exp(-2M)} \cdot x.$$

Moreover, when $0 \leq x \leq 0.1$, we have

$$g(x) \leq \log(1 + \exp(-2M)) + \frac{\exp(-2M)}{1 + \exp(-2M)} \cdot x + \frac{2\exp(-2M)}{1 + \exp(-2M)} \cdot x^2$$

Proof. Direct calculation shows that

$$\frac{dg(x)}{dx} = \frac{\exp(-2M + x)}{1 + \exp(-2M + x)},$$

which increases with x.

Therefore,

$$g(x) \geq \log(1 + \exp(-2M)) + \left.\frac{dg(x)}{dx}\right|_{x=0} \cdot x = \log(1 + \exp(-2M)) + \frac{\exp(-2M)}{1 + \exp(-2M)} \cdot x.$$

On the other hand, notice that when $x \leq 0.1$, $\exp(x) \leq 1 + x + x^2$. Therefore,

$$\begin{aligned}
g(x) &\leq \log(1 + \exp(-2M)) + x \cdot \frac{\exp(-2M + x)}{1 + \exp(-2M + x)} \\
&\leq \log(1 + \exp(-2M)) + x \cdot \frac{\exp(-2M)(1 + x + x^2)}{1 + \exp(-2M)(1 + x + x^2)} \\
&\leq \log(1 + \exp(-2M)) + x \cdot \frac{\exp(-2M)}{1 + \exp(-2M)}(1 + x + x^2) \\
&\leq \log(1 + \exp(-2M)) + \frac{\exp(-2M)}{1 + \exp(-2M)} \cdot x + \frac{2\exp(-2M)}{1 + \exp(-2M)} \cdot x^2.
\end{aligned}$$

□

The following lemma shows that if we have a big f_L, then the global minimum is small:

Lemma L.8. *For each tuple $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$, let $S := f_L(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$. Assume that $S > M$, then when M is large enough there exists $\boldsymbol{V}_0$ such that $\|\boldsymbol{V}_0\|_\infty \leq M$ and*

$$\mathcal{L} \leq \log(1 + \exp(-2M)) + \frac{\exp(-2M)}{1 + \exp(-2M)} \cdot 6ML\exp(-S) + \frac{72\exp(-2M)}{1 + \exp(-2M)} \cdot M^2L^2\exp(-2S)$$

when we replace $\boldsymbol{V}$ by $\boldsymbol{V}_0$.

Proof. We consider $V_0 = V_1 = M$.

To prove this lemma, we just need to prove the desired result for any $\mathcal{L}(s, i)$ such that $1 \leq i \leq |s| \leq L$. Without loss of generality, we assume that $s_i = 0$ (since token 0 and 1 are symmetric). Thus

$$\begin{aligned}
\mathcal{L}(s, i) &= \log\left(1 + \exp\left(-2A_{s_i}(s, i)V_{s_i} + 2A_{1-s_i}(s, i)V_{1-s_i}\right)\right) \\
&= \log\left(1 + \exp\left(-2A_0(s, i)V_0 + 2A_1(s, i)V_1\right)\right).
\end{aligned}$$

Notice that $g^+_{s,i}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ is the maximum attention score from s_{i-1} to some token 0, and $g^-_{s,i}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ is the maximum attention score to other tokens. Thus, since $f_L(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = S$, for each token other than 0, the corresponding attention is at

most $\exp(-S)$. Therefore, $A_1(s,i) \le 2L\exp(-S)$, and $A_0(s,i) \ge 1 - 2L\exp(-S)$. Therefore,

$$\begin{aligned}
&\mathcal{L}(s,i) \\
&= \log\left(1 + \exp\left(-2A_{s_i}(s,i)V_{s_i} + 2A_{1-s_i}(s,i)V_{1-s_i}\right)\right) \\
&\le \log(1 + \exp(-2M + 4L \cdot \exp(-S) \cdot M)) \\
&\le \log(1+\exp(-2M)) + \frac{\exp(-2M)}{1+\exp(-2M)} \cdot 4ML\exp(-S) + \frac{32\exp(-2M)}{1+\exp(-2M)} \cdot M^2L^2\exp(-2S).
\end{aligned}$$

□

We need the following technical lemma:

Lemma L.9. *Let $\theta_1, \dots, \theta_m \sim \mathrm{Unif}[0, 2\pi]$ be independent. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left|\sum_{i=1}^m \cos(\theta_i)\right| \le \sqrt{2m\log\frac{2}{\delta}}.$$

In particular, for any constant $c > 0$, with probability at least $1 - L^{-c}$,

$$\left|\sum_{i=1}^m \cos(\theta_i)\right| \le \sqrt{2m(c\log L + \log 2)}.$$

Proof. Let $X_i = \cos(\theta_i)$. Then $\mathbb{E}[X_i] = 0$ and $X_i \in [-1,1]$. By Hoeffding's lemma,

$$\mathbb{E}[\exp(\lambda X_i)] \le \exp\left(\frac{\lambda^2}{2}\right).$$

Therefore,

$$\Pr\left[\sum_{i=1}^m X_i \ge t\right] \le \exp\left(-\lambda t + \frac{m\lambda^2}{2}\right).$$

Optimizing over $\lambda > 0$ by taking $\lambda = t/m$, we get

$$\Pr\left[\sum_{i=1}^m X_i \ge t\right] \le \exp\left(-\frac{t^2}{2m}\right).$$

Applying the same argument to $-X_i$ gives

$$\Pr\left[\sum_{i=1}^m X_i \le -t\right] \le \exp\left(-\frac{t^2}{2m}\right).$$

Hence, by a union bound,

$$\Pr\left[\left|\sum_{i=1}^m \cos(\theta_i)\right| \ge t\right] \le 2\exp\left(-\frac{t^2}{2m}\right).$$

Taking

$$t = \sqrt{2m\log\frac{2}{\delta}}$$

proves the first claim.

Finally, setting $\delta = L^{-c}$ gives

$$\left|\sum_{i=1}^m \cos(\theta_i)\right| \le \sqrt{2m\log(2L^c)} = \sqrt{2m(c\log L + \log 2)}$$

with probability at least $1 - L^{-c}$. □

Lemma L.10. *Let $\theta_1, \dots, \theta_m \sim \mathrm{Unif}[0, 2\pi]$ be independent, and define*

$$h(\theta) = \max\{|\cos\theta|, |\sin\theta|\}.$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\sum_{i=1}^{m} h(\theta_i) \le \frac{2\sqrt{2}}{\pi}\, m + \sqrt{\frac{m}{2}\log\frac{1}{\delta}}.$$

Proof. Let

$$X_i := h(\theta_i) = \max\{|\cos\theta_i|, |\sin\theta_i|\}.$$

Then $X_1, \dots, X_m$ are independent and satisfy $0 \le X_i \le 1$.

We first compute the mean of X_i. By symmetry,

$$\mathbb{E}[X_i] = \frac{4}{2\pi}\int_0^{\pi/2} \max\{\cos\theta, \sin\theta\}\, d\theta.$$

On $[0, \pi/4]$, we have $\cos\theta \ge \sin\theta$, while on $[\pi/4, \pi/2]$, we have $\sin\theta \ge \cos\theta$. Hence

$$\mathbb{E}[X_i] = \frac{2}{\pi}\left(\int_0^{\pi/4} \cos\theta\, d\theta + \int_{\pi/4}^{\pi/2} \sin\theta\, d\theta\right).$$

Evaluating the integrals gives

$$\int_0^{\pi/4} \cos\theta\, d\theta = \sin(\pi/4) - \sin(0) = \frac{\sqrt{2}}{2},$$

and

$$\int_{\pi/4}^{\pi/2} \sin\theta\, d\theta = -\cos(\pi/2) + \cos(\pi/4) = \frac{\sqrt{2}}{2}.$$

Therefore,

$$\mathbb{E}[X_i] = \frac{2}{\pi}\cdot\sqrt{2} = \frac{2\sqrt{2}}{\pi}.$$

Now apply Hoeffding's inequality to the independent random variables $X_1, \dots, X_m \in [0, 1]$. For any $t > 0$,

$$\Pr\left(\sum_{i=1}^{m} X_i - \mathbb{E}\left[\sum_{i=1}^{m} X_i\right] \ge t\right) \le \exp\left(-\frac{2t^2}{m}\right).$$

Set

$$t = \sqrt{\frac{m}{2}\log\frac{1}{\delta}}.$$

Then

$$\Pr\left(\sum_{i=1}^{m} X_i \ge \frac{2\sqrt{2}}{\pi}\, m + \sqrt{\frac{m}{2}\log\frac{1}{\delta}}\right) \le \delta.$$

Equivalently, with probability at least $1 - \delta$,

$$\sum_{i=1}^{m} h(\theta_i) = \sum_{i=1}^{m} X_i \le \frac{2\sqrt{2}}{\pi}\, m + \sqrt{\frac{m}{2}\log\frac{1}{\delta}}.$$

This proves the lemma. □

Lemma L.11. *Let $\theta \sim \mathrm{Unif}[0, 2\pi]$, then for any $\delta < 0.1$, with probability at least,*

$$\left|\sum_{k=1}^{m} \cos(k\theta)\right|, \left|\sum_{k=1}^{m} \sin(k\theta)\right| \le \frac{3}{\delta}.$$

Proof. For each m, we have

$$\begin{aligned}\left|\sum_{k=1}^{m}\cos(k\theta)\right| &= \left|\frac{\sum_{k=1}^{m}\left(\sin\left(\frac{2k+1}{2}\theta\right)-\sin\left(\frac{2k-1}{2}\theta\right)\right)}{2\sin(\theta/2)}\right| \\ &= \left|\frac{\sin\left(\frac{2m+1}{2}\theta\right)-\sin\left(\frac{1}{2}\theta\right)}{2\sin(\theta/2)}\right| \\ &\le \frac{1}{|\sin(\theta/2)|}.\end{aligned}$$

Similarly,

$$\begin{aligned}\left|\sum_{k=1}^{m}\sin(k\theta)\right| &= \left|\frac{\sum_{k=1}^{m}\left(\cos\left(\frac{2k-1}{2}\theta\right)-\cos\left(\frac{2k+1}{2}\theta\right)\right)}{2\sin(\theta/2)}\right| \\ &= \left|\frac{\cos\left(\frac{1}{2}\theta\right)-\cos\left(\frac{2m+1}{2}\theta\right)}{2\sin(\theta/2)}\right| \\ &\le \frac{1}{|\sin(\theta/2)|}.\end{aligned}$$

As θ is uniformly sampled, with probability at least $1-\delta$, we have $|\sin(\theta/2)| \le \delta/3$. Therefore,

$$\left|\sum_{k=1}^{m}\cos(k\theta)\right|, \left|\sum_{k=1}^{m}\sin(k\theta)\right| \le \frac{3}{\delta},$$

and the result follows.

□

The following gives a lower bound of f_L.

Lemma L.12. *With probability at least $1-\frac{1}{L}$, there exists $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ such that $f_L(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \ge M^2d/2 - 2M^2\sqrt{d\log L}$.*

Proof. For each $\mathtt{C}$, we define

$$\boldsymbol{k}_{\mathtt{C}} = \begin{bmatrix} M \\ \vdots \\ M \\ M \\ \vdots \\ M \end{bmatrix}, \boldsymbol{q}_{\mathtt{C}} = \begin{bmatrix} M \\ \vdots \\ M \\ -M \\ \vdots \\ -M \end{bmatrix}$$

Notice that for any i,s, $D(s,i,i) = (0,1)$. Therefore, for any $j \ne 1+L+i$, there exists $-L \le t \le L$ such that $D(s,i,j) = (t,0)(t \ne 0)$ or $(t,1)$. If $D(s,i,j) = (t,0)$,

$$\begin{aligned}&\boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,i+1)}\boldsymbol{k}_{S(s,i,i+1)} - \boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,j)}\boldsymbol{k}_{S(s,i,j)} \\ =&M^2\cdot\frac{d}{2} - 2M^2\sum_{i=1}^{d/4}\cos(t\beta_i)\end{aligned}$$

If $D(s,i,j) = (t,1)$,

$$\begin{aligned}&\boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,i+1)}\boldsymbol{k}_{S(s,i,i+1)} - \boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,j)}\boldsymbol{k}_{S(s,i,j)} \\ =&M^2\cdot\frac{d}{2} - 2M^2\sum_{i=1}^{d/4}\cos(t\beta_i) - 2M^2\sum_{i=d/4+1}^{d/2}\cos(t\beta_i) + M^2\cdot\frac{d}{2}.\end{aligned}$$

Therefore, with high probability at least $1 - \frac{1}{L}$, we have

$$\boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,i+1)} \boldsymbol{k}_{S(s,i,i+1)} - \boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,j)} \boldsymbol{k}_{S(s,i,j)} \geq M^2 d/2 - 2M^2 \sqrt{d \log L}.$$

□

The following lemma shows that small loss requires a big $f_L(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$.

Lemma L.13. *For any $S > M$, if $f_L(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \leq S$, then we have*

$$\mathcal{L} \geq \log(1 + \exp(-2M)) + \frac{\exp(-2M)}{1 + \exp(-2M)} \cdot \frac{M \exp(-S)}{2L + 1} \cdot p.$$

Proof. First, if $V_0 \leq M - 1$, we consider the string s only containing 0, we know that the loss

$$\mathcal{L}(s, i) = \log(1 + \exp(-2A_0(s, i)V_0)) \geq \log(1 + \exp(-2M + 2)).$$

The result is similar if $V_1 \leq M - 1$.

In the following, we only consider the case such that $V_0, V_1 > M - 1$.

By definition, there exists a pair of (s, i) such that

$$g_{s,i}^+(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) - g_{s,i}^-(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \leq S.$$

Without loss of generality, we also assume that $s_i = 0$. The above inequality means that there exists a token $\texttt{C} \neq 0$ in s such that the attention score from s_i to it is at least the attention score to any other minus S. As there are at most $2L + 1$ tokens in the input, the attention to token $\texttt{C}$ is at least $\exp(-S)/(2L + 1)$. Therefore,

$$\begin{aligned}
\mathcal{L}(s, i) &= \log\left(1 + \exp\left(-2A_{s_i}(s, i)V_{s_i} + 2A_{1-s_i}(s, i)V_{1-s_i}\right)\right) \\
&\geq \log\left(1 + \exp\left(-2M + M \cdot \frac{\exp(-S)}{2L + 1}\right)\right) \\
&\geq \log(1 + \exp(-2M)) + \frac{\exp(-2M)}{1 + \exp(-2M)} \cdot M \cdot \frac{\exp(-S)}{2L + 1}.
\end{aligned}$$

The last inequality is by Lemma L.7. Notice that for any other (s', i'), we also have

$$\mathcal{L}(s', i') \geq \log(1 + \exp(-2M)).$$

Thus the result follows.

□

Lemma L.14. *Assume that $M > L^2$. Assume that $f_L(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \geq M^2 d/2 - 3M^2 \sqrt{d \log L}$. Then with probability at least $1 - \frac{2}{d}$, for any string with s, i with $|s| \leq L^{\sqrt{d}/2}$, $A_{s_i}(s, i) \geq 1 - \exp(-M^2)$.*

Proof. By the definition of f_L, we consider a random s, i with the length at most $L^{\sqrt{d}/2}$. Without loss of generality, we assume that $s_i = 0$. Now we consider the attention to each $s_j = 1$. We denote this token as s^*. We want to prove that it is small compared to the attention to s_i.

We consider the string with the same (s_{i-1}, s_i) pair, and other positions are all 1. For such string, $g_{s,i}^+(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ are fixed. We consider the

$$g_{s,i}^+(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) - \boldsymbol{q}_{s_{i-1}}^\top \boldsymbol{R}_{D(s,i,j)} \boldsymbol{k}_{S(s,i,j)} \tag{5}$$

for other token (1) in the same row with s^*. By Lemma L.11, with probability at least $1 - \frac{1}{d}$, for any $1 \leq j \leq d/4$,

$$\left|\sum_{j=1}^{L} \cos(j\theta)\right|, \left|\sum_{j=1}^{L} \sin(j\theta)\right| \leq d^2.$$

We consider the two contributions of the two parts in

$$g^{+}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) - g^{-}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$$

and denote that as $g^{+x}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) - g^{-x}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) + g^{+y}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) - g^{-y}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$. Now we sum up all possible j in Equation (5), and we know that with probability at least $1-\frac{1}{d}$,

$$g^{+x}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) + M^2\sqrt{d\log L} + g^{+y}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) - g^{-y}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \geq M^2 d/2 - 3M^2\sqrt{d\log L}.$$

Consider s^* back, notice that for $\|\boldsymbol{q}\|_\infty, \|\boldsymbol{k}\|_\infty \leq M$. For each RoPE 2*2 block, consider the corresponding θ as the base frequency times the relative position. We just use the maximum possible value given the θ to upper bound its absolute value. We can always assume that the relative position is nonzero since we only consider the generalization. Therefore, when $L' < L^{\sqrt{d}}$, by Lemma L.10, for each position, with probability at least $1 - L^{-\sqrt{d}}$ we have

$$g^{-x}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \leq M^2\left(\frac{\sqrt{2}}{\pi}d + 2d^{3/4}\sqrt{\log L}\right) \leq 0.47M^2 d.$$

Therefore, with probability at least $1 - L^{-1}$, for each length at most $L^{\sqrt{d}/2}$,

$$g^{+x}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) - g^{-x}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) + g^{+y}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) - g^{-y}_{s,i}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \geq 0.01M^2 d.$$

Since $M > L^2$, the result follows. □

Proof of Theorem L.6. By Lemma L.8 and Lemma L.12, there exists a parameter set such that

$$\mathcal{L} \leq \log(1+\exp(-2M)) + \frac{\exp(-2M)}{1+\exp(-2M)} \cdot 6ML\exp(-S) + \frac{72\exp(-2M)}{1+\exp(-2M)} \cdot M^2L^2\exp(-2S)$$

for $S = M^2d/2 - 2M^2\sqrt{d\log L}$. Combining with Lemma L.13, when M is large enough, the global minimum must satisfy

$$f_L(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \geq M^2d/2 - 3M^2\sqrt{d\log L}.$$

Also, it must satisfies $V_0, V_1 > M-1$ because otherwise we must have

$$\mathcal{L} \geq \log(1+\exp(-2M+2))$$

by considering a all 0 or all 1 string. Therefore, applying Lemma L.14 gives the desired result since M is large enough.

□

M. Prompt Templates of Copy Test for Frontier LLMs

Table 7. Prompt template for the **Recursive-Flip binary copy task** used for Gemini and Qwen. The input is a recursively generated binary string, and the Output is the exact same sequence.

System Prompt	User Prompt with Example Input	Output
You are taking a copying test. Your task is to copy the binary sequence exactly. Output only the copied binary sequence. Do not add any explanation, quotes, punctuation, or formatting. Do not add spaces or newlines inside the sequence.	Copy the following binary sequence exactly: `0101010101010101010101...`	`0101010101010101010101...`

Table 8. Prompt template for the **Recursive-Flip binary copy task** used for GPT and DeepSeek. We use a different input template because GPT and DeepSeek tokenizers tend to merge three consecutive digits into one token, as discussed in Appendix C.1. The target output is still the exact same binary sequence.

System Prompt	User Prompt with Example Input	Output
You are taking a copying test. Your task is to copy the binary sequence exactly. Output only the copied binary sequence. Do not add any explanation, quotes, punctuation, or formatting. Do not add spaces or newlines inside the sequence.	Copy the following binary sequence exactly: `000111000111000111000...`	`000111000111000111000...`

Table 9. Prompt template for the **binary copy** task under **Imbalanced Generation**. The example input is an imbalanced sequence over two symbols, where one symbol appears much more frequently than the other. The target output is the exact same sequence.

Prompt Template	Example Input	Output
Copy the following sequence exactly. Output only the copied sequence. Do not add any explanation, quotes, punctuation, or formatting.	a b a a a a a b a a a a a a b a a a a ...	a b a a a a a b a a a a a a b a a a a ...

Table 10. Prompt templates for the **Python List Conversion** task. Each prompt is followed by a comma-separated sequence of sensor values, and the target output is the same sequence formatted as a single-line Python list.

ID	Prompt Template	Example Input	Example Output
1	Copy the following sensor sequence exactly. Output only one single-line Python list, with no line breaks, no spaces, and no extra text:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
2	Rewrite the following sensor measurements as exactly one single-line Python list. Do not add any line breaks, spaces, explanation, or extra characters:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
3	Return the following numbers as one single-line Python list only. Keep the values and order exactly the same. No line breaks, no spaces, no extra text:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
4	Convert the following sensor readings into a Python list on a single line. Output only the list, with no spaces, no line breaks, and no other text:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
5	Repeat the following sequence exactly as a one-line Python list. Do not change any number. Do not insert line breaks, spaces, or commentary:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
6	Write the following measurements as exactly one Python list in a single line. Output only the list itself, with no spaces, no newlines, and no explanation:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
7	Copy the following numeric sequence exactly into a single-line Python list. Preserve every value and its order. No line breaks, no spaces, no extra text:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
8	Output exactly one single-line Python list containing the following numbers in the same order. Do not add spaces, newlines, or any surrounding text:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
9	Turn the following sensor values into a Python list written on one line only. Keep the sequence exactly unchanged. No spaces, no line breaks, no extra text:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`
10	Produce exactly the following sequence as a single-line Python list. Output only the list, with no spaces, no line breaks, and no additional content:<\n>	`1.55,1.71,...`	`[1.55,1.71,...,]`

Prompt Template	Example Input	Gold Output
Copy the following sequence exactly. Output only the copied sequence and nothing else. Do not add explanations, quotation marks, or newlines.	`Sequence: a b a b a b a b b b a ...`	`a b a b a b a b b b a ...`

Table 11. Prompt template for the **repeat-structure copy test**. We first sample a short base sequence over symbols `a` and `b`, repeat it to length 1000, and then corrupt tokens with a specified probability. The model is asked to output exactly the corrupted sequence.

N. Prompt Templates for Finetuning Data

Following Python list shows the prompts templates that was uniformly and randomly sampled from when generating finetuning data. The {s} indicates the portion that is to be replaced with the binary string to be copied.

```
[
    # With newline
    "Please echo the next sequence verbatim (no extra text). Here it is:\n{s}",
    "Copy the text below exactly as-is (keep every character):\n{s}",
    "Task: return the string shown below. Do not add or remove anything.\n{s}",
    "Print the following content exactly, unchanged. Output only the content:\n{s}",
    "Write the following string exactly as it is:\n{s}",
    "I want you to say the following string and nothing else:\n{s}",
    "Say the following string exactly as it is:\n{s}",
    "Here is a string:\n{s}\n\nPlease print the string exactly as it is.",
    "I will give you a string. You need to repeat it exactly as it is. Here it is:\n{s}",
    "{s}\n\nPlease print the above string exactly as it is.",
    "{s}\n\nCan you write out the above string and nothing else.",
    "Can you copy a string of characters? Here is the string I want you to copy:\n{s}",
    "Repeat this string and say nothing else:\n{s}",

    # Without newline
    "Repeat the following string after me (output only the string): {s}",
    "Echo this sequence exactly (no extra characters): {s}",
    "Return the exact same string and nothing else -> {s}",
    "Copy-paste the following payload exactly as written: {s}",
    "Output EXACTLY the string below, unchanged: {s}",
    "Just say \"{s}\" and nothing else.",
    "Say '{s}' and say nothing else.",
]
```

Following is the chat template used for finetuning:

```
### user:
{input}

### assistant:
{output}
```