On the Power of Convolution Augmented Transformer

Mingchen Li [1] **Xuechen Zhang** [1] **Yixiao Huang** [2] **Samet Oymak** [1]

Abstract

The transformer architecture has catalyzed revolutionary advances in language modeling. However, recent architectural recipes, such as state-space models, have bridged the performance gap. Motivated by this, we examine the benefits of Convolution-Augmented Transformer (CAT) for recall, copying, and length generalization tasks. CAT incorporates convolutional filters in the K/Q/V embeddings of an attention layer. Through CAT, we show that the locality of the convolution synergizes with the global view of the attention. Unlike comparable architectures, such as Mamba or transformer, CAT can provably solve the associative recall (AR) and copying tasks using a single layer while also enjoying guaranteed length generalization. We also establish computational tradeoffs between convolution and attention by characterizing how convolution can mitigate the need for full attention by summarizing the context window and creating salient summary tokens to attend. Evaluations on real datasets corroborate our findings and demonstrate that CAT and its variations indeed enhance the language modeling performance.

1. Introduction

The attention mechanism, central to the transformer architecture (Vaswani et al., 2017), facilitates comprehensive token interactions across the context window in contemporary large language models. Nevertheless, devoid of positional encoding (PE), attention mechanism lacks inherent locality, rendering the self-attention layer permutation-equivariant with no bias towards proximal versus distant token interactions. The convolution operator, traditionally successful in vision applications, aggregates local features based on relative positions and recently has been adopted in language modeling (Dauphin et al., 2017), including innovative frameworks such as state-space models (Gu et al., 2021)

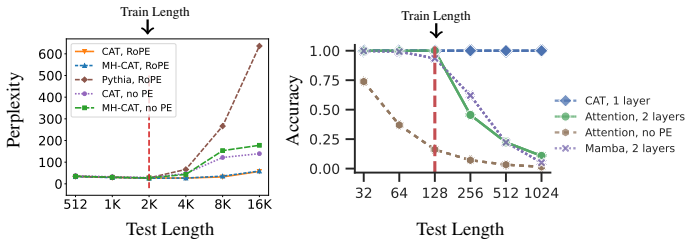

Figure 1. Evaluations on synthetic and real data. The models are trained on 2,048 and 128 context length (vertical dashed lines) and tested on varying context lengths respectively. **Left figure:** Evaluations on language modeling where we train CAT models by equipping Pythia with short convolutions (window size 21). Convolution allows the model to pretrain without positional encoding and further improves perplexity when combined with RoPE. Importantly, it also generalizes to longer context lengths more robustly with or without RoPE. For length generalization, we used YaRN (Peng et al., 2023) which incorporates position interpolation (Chen et al., 2023) (for RoPE only) and temperature scaling (see Sec. D.2). **Right figure:** We conduct synthetic experiments on the Associative Recall task and contrast 1-layer CAT with 2-layers of alternative architectures. The embedding dimension is 128. We find that CAT is the only model that solves AR with length generalization in line with our theory (also see Fig. 4).

and linear RNNs (Orvieto et al., 2023) designed for efficient long-range sequence modeling. These models, however, traditionally struggle with global context processing, a limitation that spurred the development of hybrid architectures that synergistically integrate both convolutional and attentional dynamics (De et al., 2024; Arora et al., 2024; Park et al., 2024; Arora et al., 2023).

In this work, we explore the synergy between attention and convolution which reveals new theoretical principles that inform hybrid architecture design. Specifically, we introduce an intuitive hybrid architecture called *Convolution-Augmented Transformer* (CAT)[1]. CAT incorporates convolutional filters to the K/Q/V embeddings of the attention layer as depicted on the left hand side of Figure 2. We explore the capabilities of the CAT layer through mechanistic tasks including associative recall (AR), selective copying (Gu & Dao, 2023; Jing et al., 2019), and length generalization. For

[1]University of Michigan [2]University of California, Berkeley. Correspondence to: Samet Oymak <oymak@umich.edu>.

Work presented at TF2M workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024.

[1]The transformer architecture consists of attention and MLP layers. For theoretical analysis and synthetic experiments, we will entirely focus on the *Convolution Augmented Attention* layer described in Fig. 2. For this reason, we will use the CAT acronym to refer to both Convolution-Augmented Transformer and Attention.

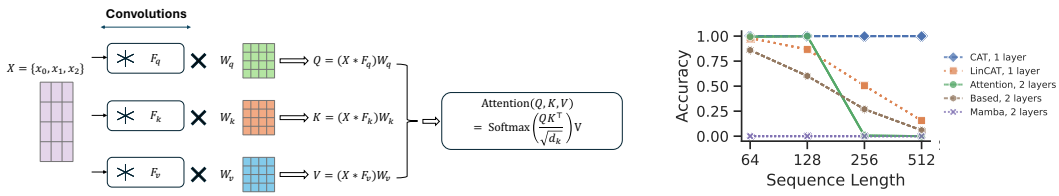

Figure 2. Left: Illustration of the Convolution-Augmentated Attention (CAT) block, where separate filters are applied to the K/Q/V embeddings, before self-attention (see Sec. 2.1 for details). Right: Performance of 1-layer CAT models trained on multi-query AR (MQAR, see Sec. C.3 for details) tasks with model embedding dimension 64 and varying sequence length. The LinCAT replaces the standard attention in CAT with linear attention. We observe that the CAT model outperforms the baseline models across all sequence lengths with only 1 layer compared to 2 layers baselines.

instance, AR is a fundamental task motivated from the associative memory in cognitive science (Ba et al., 2016). This task underpins critical applications such as bigram retrieval, where a specific sequence, such as 'Rings' following 'The Lord of the', must be correctly retrieved. Such tasks are known to be crucial for LLM functionality and mechanistic understanding (Olsson et al., 2022; Fu et al., 2022; Arora et al., 2024; Nichani et al., 2024; Poli et al., 2024).

We theoretically and empirically show that, within the CAT layer, attention and convolution exhibit strong synergy and complementarity to solve these mechanistic tasks while enjoying length generalization benefits. As a concrete example, the left side of Figure 1 displays the AR performance for various test-time sequence lengths. As the sequence length grows, we observe two distinct failure modes: Mamba's accuracy degrades due to its finite state dimension whereas attention-only models degrade due to the length extension bottlenecks of PE. In contrast, CAT maintains perfect accuracy and length generalization because attention and convolution patch these failure modes in a complementary fashion. Overall, we make the following contributions:

• We propose the convolution-augmented attention layer and prove that it can solve the N-gram AR (NAR) and Selective Copying tasks using a single layer (Theorems 1 and 4). Comparison to alternatives (Mamba, Based, attention, linear attention) reveals that CAT can uniquely solve NAR with length generalization.

• To explain this, we establish a length generalization result on the loss landscape (Theorem 2): Under mild assumptions, all CAT models that solve AR for a particular context length provably generalize to all other context lengths.

• We evaluate CAT on real data and demonstrate that even CAT noticeably aids language modeling: In line with theory, convolution enables the model to train stably without PE and achieve length generalization. (see Sec. D.2, Table 2).

Additionally, we present the detailed discussion on related works in Sec. A.

2. Problem Setup

2.1. Convolutional-Augmented Attention

Notation. Let us first introduce helpful notation. $\boldsymbol{I}_d$ is the identity matrix of size d. $\boldsymbol{D}_i$ denotes the causal delay filter that shifts a signal $\boldsymbol{x}$ i-timesteps forward i.e. $(\boldsymbol{x} * \boldsymbol{D}_i)_j = \boldsymbol{x}_{j-i}$. For an integer $n \geq 1$, we denote the set $\{0, \ldots, n-1\}$ by $[n]$. We use lower-case and upper-case bold letters (e.g., $\boldsymbol{m}$, $\boldsymbol{M}$) to represent vectors and matrices, respectively. m_i denotes the i-th entry of a vector $\boldsymbol{m}$.

Below, we introduce the Convolution-Augmented Attention layer, which incorporates learnable filters into the K/Q/V embeddings. Let $\boldsymbol{X} = [\boldsymbol{x}_0 \ \ldots \ \boldsymbol{x}_{L-1}]^\top \in \mathbb{R}^{L \times d}$ denote the input to the layer containing L tokens with embedding dimension d. Let $\boldsymbol{F} \in \mathbb{R}^W$ denote the convolutional filter with temporal length W. We examine two convolution types which handle multi-head attention in different ways:

1D per-head convolution: For each attention head, we have a distinct 1D filter $\boldsymbol{F} \in \mathbb{R}^W$. $\boldsymbol{F}$ is applied temporally to each of the d embedding dimensions. This results in $\boldsymbol{F} * \boldsymbol{X}$ where $(\boldsymbol{F} * \boldsymbol{X})_i = \sum_{j \in [W]} F_j \boldsymbol{x}_{i-j}$, with F_j being the j-th entry of $\boldsymbol{F}$.

Multi-head convolution: Suppose we have H sequences $\bar{\boldsymbol{X}} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_H] \in \mathbb{R}^{L \times d \times H}$ each corresponding to one of the H attention heads. We use a filter $\bar{\boldsymbol{F}} = [\boldsymbol{F}_1, \ldots, \boldsymbol{F}_H] \in \mathbb{R}^{W \times H \times H}$. Each $\boldsymbol{F}_i$ is convolved with $\bar{\boldsymbol{X}}$ to obtain the i-th head's output of size $L \times d$.

Definition 1 (Convolution-Augmented Attention (CAT))**.** *A CAT layer incorporates learnable convolutional filters to the key/query/value embeddings. For a single-head CAT, the key embeddings are given by* $\boldsymbol{K} = (\boldsymbol{X} * \boldsymbol{F}_k)\boldsymbol{W}_k$ *with weights* $\boldsymbol{F}_k, \boldsymbol{W}_k$ *(same for query and value embeddings).*

2.2. Mechanistic Tasks for Language Modeling

In this section, we will explore the efficacy of a *single CAT layer* in solving tasks such as AR and Selective Copying, inspired by recent works in sequence modeling literature (Gu & Dao, 2023; Arora et al., 2023). To this end, let us introduce and examine the N-gram AR task, which is a generalization of the AR task where the model needs to identify the copy of the last N tokens in the context window

and return the associated value.

Definition 2 (Associative Recall Problem)**.** *Consider a discrete input sequence $X = [x_0, x_1, \ldots, x_{L-1}]$, with tokens drawn from a vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$. The AR problem is defined as follows: Suppose that there is a unique index i ($0 \le i < L-1$) such that $x_i = x_{L-1}$. A model f successfully solves the AR problem if $f(X) = x_{i+1}$ for all inputs X. In this problem, x_i becomes the key, x_{i+1} is the associated value, and the last token x_{L-1} is the query.*

Definition 3 (N-gram AR Problem)**.** *Consider a discrete input sequence $X = [x_0, x_1, \ldots, x_{L-1}]$, with tokens drawn from a vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$. Let $X_{\{i,j\}} = [x_i, x_{i+1}, \ldots, x_j]$ denote the subsequence of X from index i to j. The N-gram associative recall (NAR) problem is formulated as follows: for $X_{\{L-N,L-1\}}$ (which are the last N tokens), there exists a unique index i ($0 \le i < L-N$) such that $X_{\{i,i+N-1\}} = X_{\{L-N,L-1\}}$. A model f solves NAR if $f(X) = x_{i+N}$ for all inputs X.*

Selective copying (SC) task is originally introduced by (Jing et al., 2019) and it is utilized by the recent Mamba (Gu & Dao, 2023) and Griffin (De et al., 2024) papers to assess their model's approximation capabilities. In SC, given an input sequence X containing noisy tokens, the model should denoise X and return the *signal tokens* within.

Definition 4 (Selective Copying)**.** *Consider a vocabulary $\mathcal{V}$ composed of a set of signal tokens $\mathcal{S}$, a set of noise tokens $\mathcal{N}$, and special token $\perp$ i.e. $\mathcal{V} = \mathcal{S} \cup \mathcal{N} \cup \{\perp\}$. Let X be a sequence whose tokens are drawn from $\mathcal{S} \cup \mathcal{N}$ and let $X_{\mathcal{S}}$ be the sub-sequence of X that includes all signal tokens in order. f solves* selective copying *over $\mathcal{S}$ if it autoregressively outputs $X_{\mathcal{S}}$ following the prompt $[X \perp]$ for all inputs X. f solves* unique selective copying *if it outputs all unique tokens of $X_{\mathcal{S}}$ in order for all X.*

Table 1 provides examples of the synthetic tasks we consider in this work. Specifically, we conduct AR and NAR experiment on their multi-queiry variants to evaluate the model's ability to recall multiple queries (detailed in Sec. C.3).

3. Provable Benefits of Convolution-Augmented Attention

Before diving into the theoretical results, we make a few clarifying remarks. We assume that all token embeddings have unit ℓ_2 norm. Secondly, a CAT layer maps each query to a vector-valued output $f(\boldsymbol{X}) \in \mathbb{R}^d$. To sample the discrete output token, we will simply return the nearest neighbor in the vocabulary of token embeddings. For associative recall problems, we will use a single head attention layer with weights $\boldsymbol{W}_q, \boldsymbol{W}_k$ are chosen as suitably scaled identity matrices. With this choice, attention essentially implements a *nearest neighbor retrieval*. It suffices for the theory thanks to the simple nature of the AR problem where we wish to identify the replica of a query within the context window. In general, we can easily contrive natural generalizations of AR and Selective Copy problems that necessitate a more sophisticated attention mechanism (see (Poli et al., 2024)). One such generalization is, given query q, we wish to retrieve a general key k (possibly $k \neq q$) and return the value associated with k.

N-gram AR. Our first result shows that a single CAT layer can solve the NAR problem under fairly general conditions.

Theorem 1 (Solving NAR)**.** *Let $\boldsymbol{F} \in \mathbb{R}^N$ be a causal 1-D convolutional filter of length N and $\texttt{norm}(\boldsymbol{X})$ normalize the rows of a matrix to unit ℓ_2 norm. Consider a single CAT layer $f(\boldsymbol{X}) = (\boldsymbol{X}_v \boldsymbol{W}_v)^\top \mathbb{S}(\boldsymbol{X}_k \boldsymbol{W}_k \boldsymbol{W}_q^\top \boldsymbol{q})$ where $\boldsymbol{q}$ is the final token of $\boldsymbol{X}_q$ and $\boldsymbol{X}_q = \texttt{norm}(\boldsymbol{X} * \boldsymbol{F}_q) \in \mathbb{R}^{L \times d}$ (same for $\boldsymbol{X}_k$,). Set $\boldsymbol{F}_q = \boldsymbol{F}$ and $\boldsymbol{W}_k = \boldsymbol{W}_q = \sqrt{c}\boldsymbol{I}_d$. Use either*

- ***Value delay:*** *$\boldsymbol{F}_k = \boldsymbol{F}_q$, $\boldsymbol{F}_v = \boldsymbol{D}_{-1}$ and $\boldsymbol{W}_v = 2\boldsymbol{I}_d$ or,*
- ***Key delay:*** *$\boldsymbol{F}_k = \boldsymbol{D}_1 * \boldsymbol{F}_q$, $\boldsymbol{F}_v = \boldsymbol{D}_0$ and $\boldsymbol{W}_v = \boldsymbol{I}_d$*

Let $\varepsilon > 0$ be the minimum ℓ_2 distance between two distinct tokens embeddings. For almost all choices of $\boldsymbol{F}$, there is a scalar $c_0 > 0$ depending on $\boldsymbol{F}$ such that, setting $c = c_0 \log(4L/\varepsilon)$, CAT layer solves the NAR problem of Def. 3 for all input sequences up to length L.

Corollary 1 (1-D CAT solves AR)**.** *Consider a CAT layer employing 1-D convolution on key embeddings with the delay filter $\boldsymbol{F}_k = \boldsymbol{D}_1 = [0\ 1\ 0\ \ldots\ 0]$ and $\boldsymbol{F}_q = \boldsymbol{F}_v = \boldsymbol{D}_0$. This model solves AR.*

Length generalization. The next theorem shows that global minima of CAT provably exhibit length generalization. That is, even if we train CAT for a fixed context length, it will work well for all other context lengths. This result is distinct from Theorem 1 because **it establishes length generalization for all CAT models** that approximately solve the AR problem for a context length, rather than constructing one such solution.

Theorem 2 (Length generalization)**.** *Let $\boldsymbol{F}_v \in \mathbb{R}_+^{2W+1}$ be a convolutional filter from time $t = -W$ to $t = W$ where $W \le L-1$. Consider a CAT layer of the form $f(\boldsymbol{X}) = \boldsymbol{X}_v^\top \mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_{L-1})$ where $\boldsymbol{X} \in \mathbb{R}^{L \times d}$, $\boldsymbol{X}_v = \boldsymbol{X} * \boldsymbol{F}_v \in \mathbb{R}^{L \times d}$ and $\boldsymbol{x}_{L-1}$ is the last token of $\boldsymbol{X}$ and $\boldsymbol{W} = \boldsymbol{W}_k \boldsymbol{W}_q^\top$. Suppose that token embeddings have unit norm. Consider any model $f = (\boldsymbol{W}, \boldsymbol{F}_v)$ that can solve the AR problem defined in Def. 2 up to ε-accuracy on all sequences of length $L \ge 3$. That is, for all $(\boldsymbol{X}, \boldsymbol{y})$ where query $\boldsymbol{x}_{L-1}$ repeats twice and $\boldsymbol{y}$ being the associated value token, we have $\|\boldsymbol{y} - f(\boldsymbol{X})\|_{\ell_2} \le \varepsilon$. Define the minimum embedding distance within vocabulary $\mathcal{V}$ as $\Delta = (1 - \max_{\boldsymbol{a} \neq \boldsymbol{b} \in \mathcal{V}} (\boldsymbol{a}^\top \boldsymbol{b})^2)^{1/2}$ and assume that $\Delta > 0$. There are absolute constants $R_0, R > 0$ such that, if $\varepsilon_0 := \varepsilon/\Delta \le R_0/L$, we have that*

- *The filter obeys $\|\boldsymbol{F} - \boldsymbol{D}_{-1}\|_{\ell_1} \le L\varepsilon_0$, which is in line with Theorem 1.*
- *Let $\boldsymbol{X}$ be an input sequence of length L' following Def. 2. Let $\boldsymbol{s}_\star(\boldsymbol{X}) \in \mathbb{R}^{L'}$ be the "golden attention map" with en-*

tries equal to $1/2$ *at the positions of the query* $\boldsymbol{x}_{L'-1}$ *and* 0 *otherwise. For all such* $\boldsymbol{X}$*, the attention map of* f *obeys* $\|\mathbb{S}(\boldsymbol{X}\boldsymbol{W}\boldsymbol{x}_{L'-1}) - \boldsymbol{s}_\star(\boldsymbol{X})\|_{\ell_1} \leq L'\varepsilon_0$.

• *For all* $\boldsymbol{X}$ *of length* L' *following Def. 2, we have that* $\|\boldsymbol{y} - f(\boldsymbol{X})\|_{\ell_2} \leq RL'\varepsilon_0$.

Selective Copy. Our next result shows that, 1-layer CAT model can solve the *unique selective copy* problem. That is, it can provably generate all signal tokens in the correct order as long as the input contains each distinct signal token at most once. Corroborating this, our experiments demonstrate that 1-layer CAT performs on par with or better than alternative architectural choices. The proof is deferred to Section E.4.

Theorem 3 (Selective Copy)**.** *Consider the setting of Def. 4. There is a 1-layer CAT using exponential-decay query-convolution (i.e.* $F_{q,i} = \rho^i$*) and* $d = |\mathcal{S}| + 3$ *dimensional token embeddings such that, it outputs all signal tokens in order for all inputs where signal tokens appear uniquely.*

Additionally, we show that long convolutions bring the benefit of *context summarization* and mitigate need for attention: We describe Landmark CAT (following Landmark Attention (Mohtashami & Jaggi, 2023)) which first attends on landmark tokens to locate the salient block within the subsequence and then applies full attention within that block. For the AR task, we establish fundamental theoretical tradeoffs between the embedding dimension (amount of memory), convolution/block length, and the sparsity of attention (recall capability), which shows that long convolutions can provably enable the success of sparse attention (Sec.B).

4. Experiments

In this section, we conduct synthetic experiments on N-gram AR and length generalization to evaluate the capability of CAT. We utilize convolution kernels with a width of $W = 3$ and explore model embedding sizes of $d = 32$, 64, and 128 across MQAR and MQNAR problems to assess the impact of model dimension on performance. In addition to the standard attention mechanism, we introduce a perturbation strategy by implementing linear attention on the convoluted Q, K, and V embeddings, referred to as LinCAT. We adhere strictly to the parameters set by (Arora et al., 2023). More detailed information on the training setup can be found in Section C. Additionally, Sec. 2 provides 370M-parameter models pretrain experiments on SlimPajama (Soboleva et al., 2023) with 15B tokens, which demonstrate the effectiveness of CAT in real-world language modeling tasks.

As illustrated in Fig. 3, the CAT model consistently outperforms all baseline models across a range of sequence lengths and model dimensions. Notably, both Mamba and Based models exhibit improved performance as the model dimension increases, particularly with shorter sequence lengths. This improvement is due to the memory-recall tradeoff (Arora et al., 2024) where models store and recall sequence information more as their dimensionalities expand. In contrast, thanks to the short convolution, the *single-layer* CAT model maintains 100% accuracy across all experimental settings, aligned with our theorem 1. Interestingly, aside from CAT, Mamba is the only model demonstrating the potential to effectively address the MQAR task within a single-layer network architecture. We will discuss this observation in further detail in Section D.

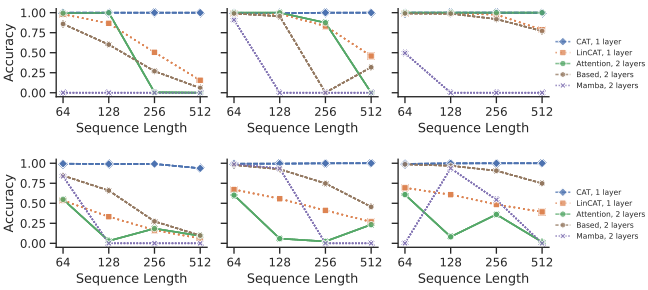

Figure 3. Evaluation of models on MQAR and MQNAR tasks with varying model dimensions and sequence lengths. Model dimensions are **32, 64, 128** for each column of the figures, from left to right. **Top:** Models trained on the MQAR setup. **Bottom:** Models trained on the MQNAR setup. Note that CAT models employ a single-layer architecture, whereas all other models utilize two layers. Refer to Section C for detailed setup descriptions.

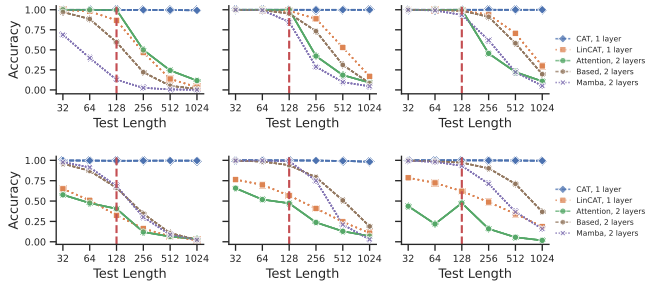

Figure 4. Evaluation of models on length generalization. Model dimensions are **32, 64, 128** for each column of the figures, from left to right. The models are trained with sequence length 128 (vertical red dashed lines) and tested on varying test length. **Top:** Models trained on the MQAR. **Bottom:** Models trained on the MQNAR. Note that CAT models establish length generalization aligned with Theorem 2 .

Evaluation of Length Generalization. In Fig. 4, we train models with 128 sequence length (the vertical red dashed line) and evaluate on varying sequence lengths from 32 to 1,024. Fig. 4 shows the results of length generalization, which is aligned with our Theorem 2: CAT models maintain 100% accuracy while all other models exhibit a sharp decline in performance as the sequence length increases. This decrease is due to the increased demand of recall which requires the model to store and retrieve more information as the sequence length grows. The CAT model, however, is able to maintain its performance by leveraging the convolutional filters to shift the context and retrieve the information.

5. Conclusion and Limitations

In this work, we have examined the synergy between the attention and convolution mechanisms by introducing Convolution-Augmented Attention where K/Q/V embeddings are equipped with convolution. We have shown that CAT enjoys strong theoretical guarantees when it comes to AR and copying tasks and also reveal insightful tradeoffs between attention and convolution. Importantly, real experiments confirm the benefit of CAT both in accuracy and in length generalization. Ultimately, we believe this work as well as the related recent literature (Gu & Dao, 2023; Arora et al., 2024; Poli et al., 2024) contributes to stronger design principles for the next generation of (hybrid) architectures.

Limitations and future work. This work has a few shortcomings. We have only focused on pretraining. However, Fig. 1 shows the potential of CAT in finetuning as a future direction. While K/Q convolution helps in theoretical constructions for N-gram AR, in real experiments, they don't provide noticeable performance benefits. We suspect that K/Q convolution might be *diluting* the attention scores and incorporating normalization or better parameterization can address this issue. An important parameterization to explore is replacing the short convolutions within CAT with SSMs. Finally, Section B introduced Landmark CAT as a sparse attention strategy. It would be interesting to evaluate this proposal on real language modeling tasks.

Acknowledgements

This work was supported in part by the National Science Foundation grants CCF-2046816, CCF-2403075, the Office of Naval Research award N000142412289, and gifts by Open Philanthropy and Google Research.

References

Arora, S., Eyuboglu, S., Timalsina, A., Johnson, I., Poli, M., Zou, J., Rudra, A., and Ré, C. Zoology: Measuring and improving recall in efficient language models. *arXiv:2312.04927*, 2023.

Arora, S., Eyuboglu, S., Zhang, M., Timalsina, A., Alberti, S., Zinsley, D., Zou, J., Rudra, A., and Ré, C. Simple linear attention language models balance the recall-throughput tradeoff. *arXiv preprint arXiv:2402.18668*, 2024.

Ba, J., Hinton, G. E., Mnih, V., Leibo, J. Z., and Ionescu, C. Using fast weights to attend to the recent past. *Advances in neural information processing systems*, 29, 2016.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Candes, E. J. The restricted isometry property and its implications for compressed sensing. *Comptes rendus. Mathematique*, 346(9-10):589–592, 2008.

Candes, E. J. and Tao, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

Chen, S., Wong, S., Chen, L., and Tian, Y. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Dao, T. and Gu, A. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

De, S., Smith, S. L., Fernando, A., Botev, A., Cristian-Muraru, G., Gu, A., Haroun, R., Berrada, L., Chen, Y., Srinivasan, S., et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A., and Ré, C. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Gu, A., Goel, K., Gupta, A., and Ré, C. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35: 35971–35983, 2022.

Gupta, A., Gu, A., and Berant, J. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022.

Hasani, R., Lechner, M., Wang, T.-H., Chahine, M., Amini, A., and Rus, D. Liquid structural state-space models. *arXiv preprint arXiv:2209.12951*, 2022.

Jelassi, S., Brandfonbrener, D., Kakade, S. M., and Malach, E. Repeat after me: Transformers are better than state space models at copying. *arXiv preprint arXiv:2402.01032*, 2024.

Jing, L., Gulcehre, C., Peurifoy, J., Shen, Y., Tegmark, M., Soljacic, M., and Bengio, Y. Gated orthogonal recurrent units: On learning to forget. *Neural computation*, 31(4): 765–783, 2019.

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, 2024.

Kosma, C., Nikolentzos, G., and Vazirgiannis, M. Time-parameterized convolutional neural networks for irregularly sampled time series. *arXiv preprint arXiv:2308.03210*, 2023.

Li, Y., Cai, T., Zhang, Y., Chen, D., and Dey, D. What makes convolutional models great on long sequence modeling? *arXiv preprint arXiv:2210.09298*, 2022.

Ma, X., Zhou, C., Kong, X., He, J., Gui, L., Neubig, G., May, J., and Zettlemoyer, L. Mega: moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*, 2022.

Ma, X., Yang, X., Xiong, W., Chen, B., Yu, L., Zhang, H., May, J., Zettlemoyer, L., Levy, O., and Zhou, C. Megalodon: Efficient llm pretraining and inference with unlimited context length. *arXiv preprint arXiv:2404.08801*, 2024.

Mohtashami, A. and Jaggi, M. Landmark attention: Random-access infinite context length for transformers. *NeurIPS*, 2023.

Nichani, E., Damian, A., and Lee, J. D. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 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.

Orvieto, A., Smith, S. L., Gu, A., Fernando, A., Gulcehre, C., Pascanu, R., and De, S. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, pp. 26670–26698. PMLR, 2023.

Park, J., Park, J., Xiong, Z., Lee, N., Cho, J., Oymak, S., Lee, K., and Papailiopoulos, D. Can mamba learn how to learn? a comparative study on in-context learning tasks. *arXiv preprint arXiv:2402.04248*, 2024.

Peng, B., Quesnelle, J., Fan, H., and Shippole, E. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Poli, M., Massaroli, S., Nguyen, E., Fu, D. Y., Dao, T., Baccus, S., Bengio, Y., Ermon, S., and Ré, C. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023.

Poli, M., Thomas, A. W., Nguyen, E., Ponnusamy, P., Deiseroth, B., Kersting, K., Suzuki, T., Hie, B., Ermon, S., Ré, C., et al. Mechanistic design and scaling of hybrid architectures. *arXiv preprint arXiv:2403.17844*, 2024.

Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

Ren, L., Liu, Y., Lu, Y., Shen, Y., Liang, C., and Chen, W. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *arXiv preprint arXiv:2406.07522*, 2024.

Slepian, D. The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.

Soboleva, D., Al-Khateeb, F., Myers, R., Steeves, J. R., Hestness, J., and Dey, N. Slimpajama: A 627b token cleaned and deduplicated version of redpajama, 2023.

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., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 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.

Yang, B., Bender, G., Le, Q. V., and Ngiam, J. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in neural information processing systems*, 32, 2019.

Yang, S., Wang, B., Shen, Y., Panda, R., and Kim, Y. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

A. Related Works

Convolution-like sequence models. Gated-convolutions (Dauphin et al., 2017) and state-space models, such as S4 (Gu et al., 2021), utilize long convolutions to reduce the computational demands associated with attention mechanisms. Performance enhancements have also been achieved through novel filter parametrization techniques (Gupta et al., 2022; Gu et al., 2022). Despite these innovations, challenges in Multi-query Associative Recall (MQAR) prompted the development of input-dependent convolution techniques. Notable developments in this area include, Liquid S4 (Hasani et al., 2022), Mamba (Gu & Dao, 2023; Dao & Gu, 2024) and (Yang et al., 2019; Kosma et al., 2023) where convolution filters are directly parametrized by inputs and include correlation terms between input tokens to enhance state mixing. (Li et al., 2022) empirically explores the reason underlying the success of convolutional models.

Expressivity, recall, length generalization. Recent works (Arora et al., 2024; Jelassi et al., 2024; Arora et al., 2023; Fu et al., 2022) explore the limitations of purely convolutional models, including Mamba, and demonstrate that, these models inherently lack the capability to solve recall problems unless they have large state dimensions (i.e. memory). (Jelassi et al., 2024) also provides a construction for 2-layer self-attention to solve AR with length generalization. Interestingly, this construction uses Hard Alibi, which is a variation of Alibi PE (Press et al., 2021) that utilize explicit linear biases in attention. Their Hard Alibi restricts the attention layer to focus on and aggregate only the recent N tokens. In this regard, this construction is related to our short convolution. On the other hand, while this work is constructive, we also prove that CAT has good loss landscape and all CAT solutions to AR provably length generalize. It has also been observed that PE can hurt length generalization and reasoning. In fact, (Kazemnejad et al., 2024) has found NoPE to be viable. On the other hand, in our real data evaluations, we have found pure NoPE to be highly brittle as it either fails to converge or optimization is unreasonably slow. Our AR experiments also corroborate that NoPE by itself is indeed not a viable strategy.

Hybrid architectures. There is a growing interest in integrating different language modeling primitives to obtain best-of-all-world designs. To this end, mechanistic tasks such as AR, copying, induction head, and in-context learning have been important to demystify the functionalities of language models (Olsson et al., 2022; Park et al., 2024) and have been utilized to guide architecture design (Arora et al., 2023; Poli et al., 2024). Gating mechanisms have been integrated within convolutional frameworks to enhance the model's selectivity. Models employing gating functions, have shown substantial improvements in AR tasks (Fu et al., 2022; Poli et al., 2023). Additionally, recent innovations on hybrid architecture, such as BaseConv (Arora et al., 2023; 2024), GLA (Yang et al., 2023), MambaFormer (Park et al., 2024), and (Ma et al., 2024; 2022; Ren et al., 2024) have provided more effective solutions to AR tasks. This comprehensive foundation of hybrid architectures informs our exploration into the convolution-attention synergy.

B. Benefits of Long Convolution for Enabling Sparse-Attention

So far we have discussed the benefits of short convolutions to equip transformer with local context to solve AR and its variations. During this discussion, we have used dense attention which has exact recall capabilities thanks to its ability to scan the full context window. In this section, we ask the following: Can convolution also help mitigate the need for dense attention? Intuitively, we should be able to tradeoff the accuracy of attention computation with computation. Here, we describe how long convolutions can enable this by effectively summarizing the context window so that we can identify where to attend in (extremely) long-context settings.

Specifically, we will prove that, long convolutions (such as SSMs) allow us to utilize sparse attention while retaining (high-probability) recall guarantees. These findings complement the recent research that establish the recall limitations of purely recurrent models (Arora et al., 2024; 2023). Our theory will also shed light on the mechanics of landmark attention (Mohtashami & Jaggi, 2023). While (Mohtashami & Jaggi, 2023) does not rely on convolution, we will describe how convolution can generate *landmark tokens* by summarizing/hashing the chunks of the context window, and attention can efficiently solve recall by attending only to these summary tokens.

Landmark Convolutional Attention (LCAT): Figure 5 describes the LCAT block that apply on input sequence $\boldsymbol{X}$. Let $\boldsymbol{F}_k \in \mathbb{R}^L$ be the convolutional filter on keys, B be the sampling rate, and $\bar{L} = \lceil L/B \rceil$. Setting $\boldsymbol{K} = (\boldsymbol{X} * \boldsymbol{F}_k)\boldsymbol{W}_k \in \mathbb{R}^{L\times d}$, we obtain $\boldsymbol{K}^{\text{ss}} \in \mathbb{R}^{\bar{L}\times d}$ by sampling $\boldsymbol{K}$ at every B tokens. Additionally, define $\boldsymbol{X}_i$ to be the ith block of $\boldsymbol{X}$ of size B spanning tokens $(i-1)B+1$ to iB. Let $\boldsymbol{V} = (\boldsymbol{F}_v * \boldsymbol{X})\boldsymbol{W}_v$ denote the value embeddings. For a query $\boldsymbol{q}_i$ for $i \in [L]$, the LCAT layer

outputs:

$$\text{(1) Hard Attention:} \quad b = \arg\max_{j \neq \lceil i/B \rceil} \boldsymbol{K}^{\text{ss}} \boldsymbol{q}_i \qquad \text{(LCAT)}$$

$$\text{(2) Local Attention:} \quad \boldsymbol{y} = \mathbb{S}(\boldsymbol{K}^{\text{loc}} \boldsymbol{q}_i)\boldsymbol{V}^l \quad \text{where} \quad \boldsymbol{K}^{\text{loc}} = \text{concat}(\boldsymbol{K}_{\lceil i/B \rceil}, \boldsymbol{K}_b).$$

Above, *hard attention* phase aims to retrieve the correct block associated to the query. This block is merged with the local block $\lceil i/B \rceil$ that contains the query itself similar to sliding window attention. We then apply *dense local attention* on the concatenated blocks $\boldsymbol{K}^{\text{loc}}$.

Computational complexity of LCAT: For a fixed query, (LCAT) requires $O(d(L/B + B))$ computations. This is in contrast to $O(dL)$ computations of vanilla attention. Choosing a suitable block size (e.g. $B = O(\sqrt{L})$), this model should save up to $\times \sqrt{L}$ in computational savings. Importantly, our theory will highlight the interplay between the embedding dimension d and the allowable acceleration by characterizing the exact performance of (LCAT) under a random context model.

Definition 5 (Random Context Model)**.** *The query token $\boldsymbol{x}_L$ occurs twice in the sequence and has unit ℓ_2 norm. All other tokens of $\boldsymbol{X}$ are IID and drawn with IID $\mathcal{N}(0, \sigma^2/d)$ entries.*

The following proposition shows that, (LCAT) will solve AR if and only if $\frac{d}{2B \log \bar{L}} \geq 1 + o(1)$.

Proposition 1. *Recall $\bar{L} = \lceil L/B \rceil$ is the number of blocks. Let $\boldsymbol{W}_v = 2\boldsymbol{I}_d$, $\boldsymbol{F}_v = \boldsymbol{D}_{-1}$, and $\boldsymbol{W}_k = \boldsymbol{W}_q = \sqrt{c} \cdot \boldsymbol{I}_d$ with $c \to \infty$. Set key convolution as $F_{k,i} = 1$ for $0 \leq i < B$ and zero otherwise.*
***(A)** If $d \geq 2\sigma^2 B(\sqrt{\log \bar{L}} + t)^2$, then (LCAT) solves AR for fixed $\boldsymbol{x}_L$ with probability at least $1 - 3e^{-t^2/4}$.*
***(B)** Conversely, for any $\varepsilon > 0$ there is $C_\varepsilon > 0$ as follows: If $\bar{L} \geq C_\varepsilon$ and $d \leq 2\sigma^2 B(\sqrt{(1-\varepsilon)\log \bar{L}} - t)^2$, then (LCAT) fails to solve AR with the same probability.*
***(C)** Finally, suppose we wish to solve AR **uniformly** for all queries $\boldsymbol{x}_L$ over a subspace S. This succeeds with the same probability whenever $d \geq 2\sigma^2 B(\sqrt{\log \bar{L}} + \sqrt{\dim(S)} + t)^2$.*

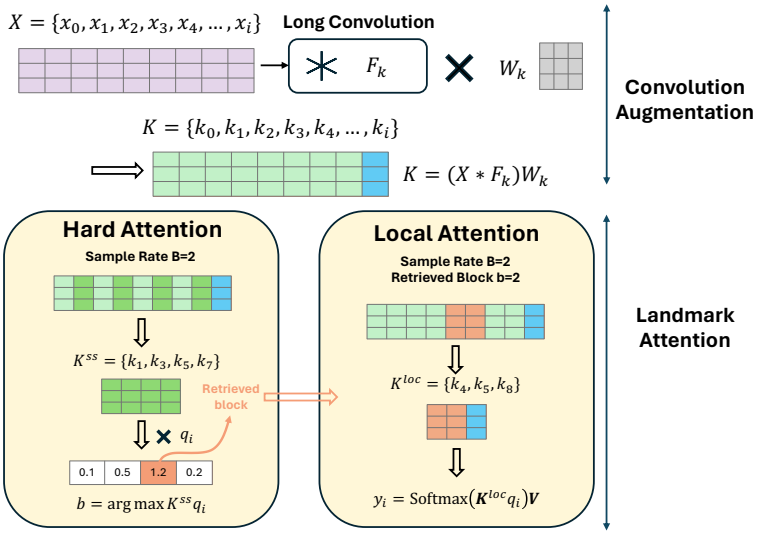

Figure 5. Illustration of the Landmark CAT. We first apply long convolution on the input sequence and subsample it to obtain landmark tokens representing individual blocks. Hard Attention computes the similarity between the query and landmarks to retrieve the most relevant block. Local Attention concatenates the retrieved block with the final block containing the query and computes the output token.

Figure 6 corroborates the predictive accuracy of Proposition 1: As the block size increases, the embedding dimension to maintain success of AR grows approximately linearly. One can expand on this proposition in two directions. Firstly, a fundamental bottleneck in (LCAT) is the requirement $d \gtrsim B \log \bar{L}$. This arises from a *memory-recall tradeoff* (Arora et al., 2024; Jelassi et al., 2024) as we are summarizing the information of block $\boldsymbol{X}_i$ of length B through its landmark token. However, once this requirement is satisfied, the model can identify the correct block in $O(\bar{L})$ cost. To avoid paying the additional $O(B)$ cost of local attention, we could apply the LCAT approach hierarchically within the selected block to reduce the compute cost to $d(\bar{L} + \log B)$ per token. The dominant term $d\bar{L}$ captures the recall capacity of the LCAT model: Consistent with our theorem and lower bounds of (Arora et al., 2024), for AR to succeed, we need

$$\text{recall\_capacity} = d\bar{L} \geq L = \text{required\_memory}$$

Secondly, Proposition (1) chooses a particular long convolution where landmarks become the mean of the input tokens within the block. In practice, we can use a state-space model (Gu et al., 2022) to parameterize convolution efficiently. A particular SSM choice of state dimension 1 is simply using exponential smoothing. This yields the following SSM variant of Proposition 1.

Proposition 2. *Consider the setting of Proposition 1 with the exponential smoothing filter $\boldsymbol{F}_i = \rho^i$ for $i \geq 0$. Set $\rho = e^{-1/B}$ so that $\rho^B = e^{-1}$. Suppose $d \geq 50B(\sqrt{\log \bar{L}} + t)^2$. Then, (LCAT) solves AR with probability at least $1 - 3e^{-t^2/4}$.*

Above, we fixed the decay rate ρ for exposition purposes. More generally, any ρ choice with an effective context size of $O(B)$ would result in similar guarantee.

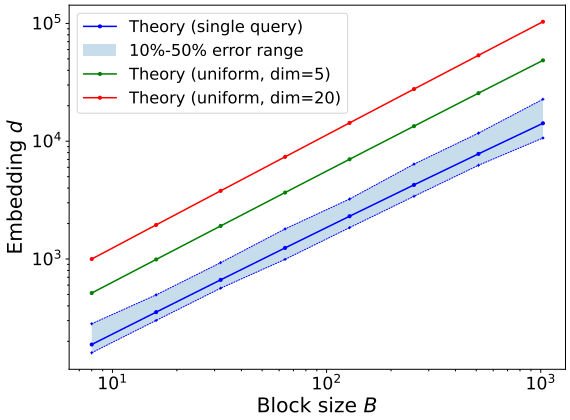

Figure 6. Behavior of the embedding dimension as a function of block size for context length $L = 2^{20} \approx 1$ million (noise level $\sigma^2 = 1$). Shaded region highlights te range of d that exhibits 10%-50% empirical success. Proposition 1 accurately captures the empirical behavior. For the success of uniform AR, we need larger d as the dimension of the query space S grows.

C. Detailed Experiment Setup

C.1. Associative Recall Experiments

We first introduce the training setup for the synthetic experiments. In our MQAR and MQNAR experiments, we create a dataset with a vocabulary size of 8,092 to ensure that the vocabulary replicates the scope of real language data. The dataset is constructed as described in Sec. C.3, with varying sequence lengths L of 64, 128, and 256, and 512. Specifically, we formulate the dataset in the form of key-value pairs accompanied by multiple queries where the keys are unique within each sequence. For each example, we initially select k keys from the vocabulary without replacement and subsequently draw the values from the remaining vocabulary. We then randomly shuffle the keys and associated values to form the input sequence. The number of queries is set to match k, ensuring each key in the sequence is queried. It should be noted that while the keys are unique within a single example, they may be repeated across different examples. For sequence lengths of $L = 64$, 128, 256, and 512, we set $k = 16$, 32, 64, and 128 respectively, indicating that the number of keys and queries scales with the sequence length, thus increasing the task complexity. We generate 100,000 training examples and 3,000 testing examples for each of the sequence lengths. For NAR experiment, we primarily focus on $N = 2$ to evaluate the performance. We construct the dataset similarly to the MQAR task with sequence lengths of 64, 128, and 256. Consequently, the number of keys and queries is reduced to $k = 10, 20, 40$ respectively, to accommodate the larger N. We generate 200,000 training examples and 3,000 testing examples for each sequence length.

For the training, we adhere strictly to the parameters set by (Arora et al., 2023), and their experimental setup and code, using learning rate sweep among $0.001, 0.01, 0.1$ and train the model for 64 epoches. The maximum accuracy achieved across these learning rate is reported.

We remark that for the length generalization experiments, we sweep the learning rate among $0.001, 0.003, 0.01, 0.03, 0.1$ and report the maximum accuracy over 5 runs to ensure the robustness and reproducibility of the results.

C.2. Language Modeling Experiments

For the language modeling experiments, we exactly follow the setup from (Yang et al., 2023). For the length generalization experiment, we train the model on sequences of length 2,048 and assess its zero-shot performance on the Wikitext dataset across varying test sequence lengths.

C.3. Multi-Query Synthetic Tasks

In this section, we introduce the multi-query versions of the AR and NAR tasks, denoted as MQAR and MQNAR, respectively. In the multi-query (MQ) scenario, a model receives multiple queries simultaneously and must generate corresponding outputs for each query. This approach was first introduced in (Arora et al., 2023), which demonstrated that while the Mamba model successfully addresses single-query AR tasks, it struggles with MQAR when operating with a limited model dimension. This highlights the increased complexity of multi-query tasks.

Definition 6 (Multi-Query Associative Recall (MQAR))**.** *Consider a discrete input sequence $X = [x_0, x_1, \ldots, x_{L-1}]$ with tokens drawn from a vocabulary $\mathcal{V}$. Let $X_{\{i,j\}} = [x_i, \ldots, x_j]$ denote a subsequence of X from index i to j. The* ***multi-query***

Table 1. Illustrative examples of synthetic tasks. In all AR-based tasks, keys and queries are highlighted in red and the values in green. For NAR tasks, parentheses denote N-gram queries; note that the parentheses are not part of the input. In SC tasks, signal tokens are in green and noise tokens in gray, and the model begins output when ⊥ appears in the sequence.

		Input	Query	Output
Single Query	AR	a 2 c 1	a	2
	NAR	(a b) 2 (b a) q (a a) 4	b a	q
	SC	a [n] [n] c [n] k	⊥	a c k
Multi Query	AR	a 2 c 1	c a	1 2
	NAR	(a b) 2 (b a) q (a a) 4	(b a) (a a)	q 4

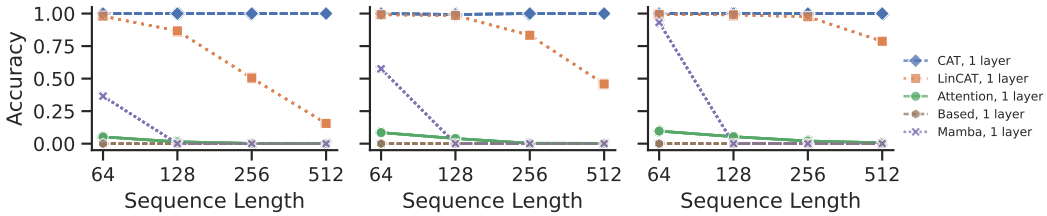

Figure 7. Performance of 1-layer models on MQAR tasks with varying model dimension and sequence length. Noted that all models are trained using 1-layer architecture.

N-gram associative recall (MQNAR) *problem is defined as follows: for every N-gram query* $Q_k = X_{k-N+1\ldots k}$, $N \leq k < L$, *determine if there exists a* $N \leq j < k$ *such that* $X_{\{j-N+1,j\}} = Q_k$. *If so, output the value* x_{j+1} *as the result, else output a special token to indicate no match is found. A model* f *solves MQNAR if it outputs the correct values for all N-gram queries and all inputs X. The standard* ***MQAR*** *problem (Arora et al., 2023) is a special instance of MQNAR by setting* $N = 1$.

D. Additional Experiments

D.1. Extended Synthetic Experiments

We conduct additional Experiments, Fig. 7 and 8 shows the result of 1-layer models on 1-gram and 2-gram MQNAR tasks with varying hidden sizes and sequence lengths. The model dimension is set to 32, 64, and 128 for each column of the figures, from left to right. All other models perform much worse compare to their 2-layer counterparts. Fig. 9 and 10 show the length generalization results of 1-layer models on 1-gram and 2-gram MQNAR tasks. The results are consistent with the 2-layer models.

D.2. Evaluations on Language Modeling

Based on the outcomes from the synthetic experiments, we further explore the efficacy of the CAT model in real-world NLP tasks by integrating a 1D CAT structure into the Pythia (Biderman et al., 2023) framework. We pretrain the modified

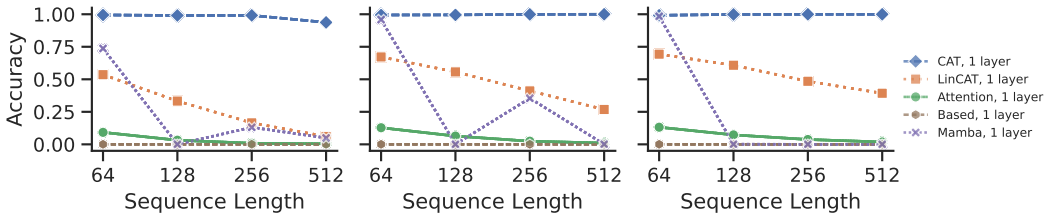

Figure 8. Performance of 1-layer CAT models on 2-gram MQNAR tasks with varying hidden sizes and sequence length. All models are trained using 1-layer architecture.

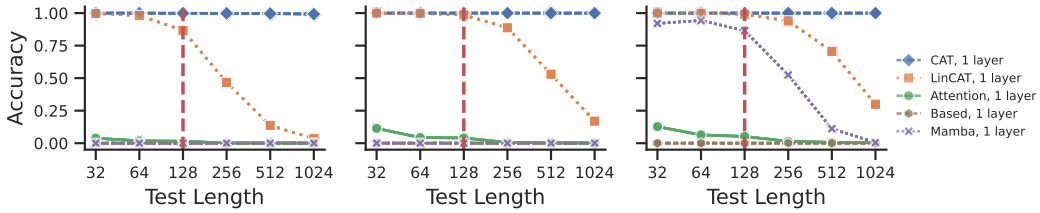

Figure 9. 1-gram Length generalization

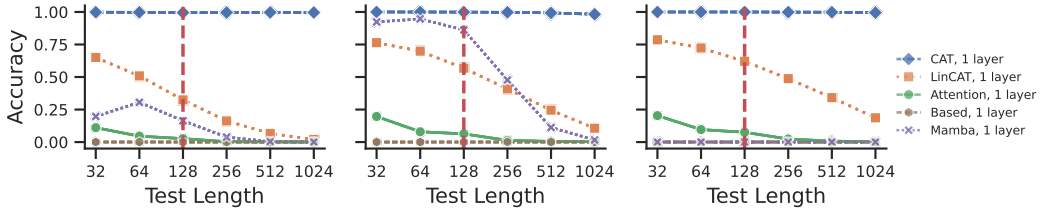

Figure 10. 2-gram Length generalization

370M-parameter model on the SlimPajama (Soboleva et al., 2023) dataset, involving 15 billion tokens. We then assess the model on a variety of downstream zero-shot tasks, including Wikitext, Lambada, Piqa, Hella, Winogrande, Arc-E, and Arc-C, a methodology commonly used in the field to evaluate generalization capabilities across diverse tasks (Biderman et al., 2023; Gu & Dao, 2023; Arora et al., 2023; 2024). The findings are compiled in Table 2.

In this series of experiments, the CAT model is trained in two variants: one incorporating rotary positional embedding (Su et al., 2024) (PE) and another without positional embedding (noPE). We observe that the CAT model with PE not only consistently outperforms the Pythia model but also achieves performance better than state-of-the-art models, including Mamba (Gu & Dao, 2023), TF++ (Touvron et al., 2023), and GLA (Yang et al., 2023). Notably, the CAT model secures a superior perplexity gain compared to the standard model while maintaining a similar level of parameters.

Regarding the noPE variant, training a Pythia model without positional encoding leads directly to divergence and extremely large losses during training, affirming the critical role of positional encoding in enabling standard transformer models to learn and converge. Intriguingly, despite the absence of positional encoding, the CAT model still performs competitively with the leading models. This suggests that the convolutional structure in the CAT model effectively captures positional information within the data. We conjecture that the short convolutions provide positional information for neighboring tokens, while the deep multi-layer network structure hierarchically aggregates this information to establish long-range positional information.

This observation aligns with our synthetic experiment results, where the CAT model demonstrated the capability to handle the AR task without positional encoding. These insights indicate that the convolutional structure could potentially replace positional encoding, which might benefit length extrapolation and generalization in the model. This offers a promising direction for further model design and optimization in the field of NLP.

•**Length Generalization** Figure 1 presents the results from a length generalization experiment with the CAT model, in which we trained the model on sequences of length 2,048 and assessed its zero-shot performance on the Wikitext dataset across varying test sequence lengths. As a baseline in our analysis, we implemented position interpolation (PI) (Chen et al., 2023) and YaRN (Peng et al., 2023) tempreture scaling on RoPE models, including CAT/MH-CAT RoPE, to facilitate length generalization. The results indicate that among the three RoPE models examined, the CAT model consistently demonstrates excellent performance across all test sequence lengths. In contrast, the Pythia model exhibits a sharp decline in performance as the sequence length increases. We suggest that is due to the additional positional embeddings introduced by PI that was absent during the training phase. Despite this, CAT models proficiently manage the relative positioning of tokens (especially overcome the new positional embeddings by leveraging convolution information), which significantly boosts

Table 2. Experiment results for model pretraining. * are results from (Yang et al., 2023), which uses a same dataset and training procedure as ours. We use the same hyperparameters as (Yang et al., 2023) for fair comparison. For perplexity, lower is better, and for accuracy, higher is better. The average accuracy in last column is calculated by averaging the accuracy across all tasks but excluding the perplexity tasks. The best and second best results are highlighted in boldface and underline, respectively.

Model	Wikitext ppl↓	Lambada_std ppl↓	Lambada_openai ppl↓	Lambada_std acc↑	Lambada_openai acc↑	Piqa acc↑	Hella acc_norm↑	Winogrande acc↑	Arc-E acc↑	Arc-C acc_norm↑	Avg Acc↑
Pythia	27.410	74.663	34.023	0.281	0.343	0.651	0.355	<u>0.529</u>	0.443	0.235	0.405
CAT, no PE	29.216	86.318	42.260	0.266	0.321	0.640	0.339	0.515	0.436	0.237	0.393
CAT, RoPE	<u>26.776</u>	65.423	38.557	0.288	0.341	<u>0.654</u>	<u>0.362</u>	0.507	0.461	0.239	0.407
MH-CAT, no PE	27.417	<u>58.959</u>	<u>32.822</u>	<u>0.296</u>	<u>0.355</u>	0.644	0.352	**0.531**	0.460	<u>0.240</u>	<u>0.411</u>
MH-CAT, RoPE	**25.858**	**47.593**	**28.273**	**0.330**	**0.377**	**0.662**	**0.376**	0.512	**0.466**	0.231	**0.422**
TF++ (Touvron et al., 2023)*	28.390	NA	42.690	NA	0.310	0.633	0.340	0.504	0.445	**0.242**	NA
Mamba (Gu & Dao, 2023)*	28.390	NA	39.660	NA	0.306	0.650	0.354	0.501	<u>0.463</u>	0.236	NA
GLA (Yang et al., 2023)*	28.650	NA	43.350	NA	0.303	0.648	0.345	0.514	0.451	0.227	NA

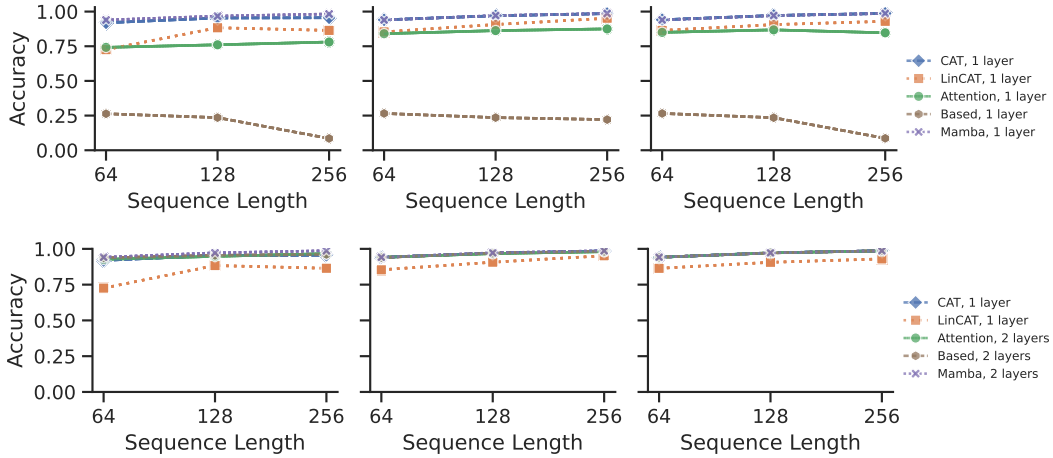

Figure 11. Evaluation of models on selective copying tasks with varying model dimensions and sequence lengths. Model dimensions are **32, 64, 128** for each column of the figures, from left to right. **Top:** Models trained on 1-layer architectures. **Bottom:** Models trained on 2-layer architectures. Note on 1-layer experiment, CAT and Mamba achieve nearly 100% and their curves are overlapped.

its ability for length generalization. Additionally, the CAT model without PE is superior to the Pythia model with RoPE, suggesting the effectiveness of the convolutional structure within the CAT model in capturing essential positional data in length extrapolation.

D.3. Model Evaluation on Selective Copying

Fig. 11 displays the selective copying results for 1-layer and 2-layer models. We train these models across a variety of model dimensions and sequence lengths. The models are required to copy 16 signal tokens from the input sequence and output them in the correct order. We observe that all 2-layer models perform well and show overlapping results, except for LinCAT. Among the 1-layer models, CAT and Mamba achieve nearly 100% accuracy, while the performance of other models is lower. These results are consistent with Theorem 4 and demonstrate that the 1-layer CAT model can solve the selective copying problem without repetitions.

E. Proofs on Associative Recall and Selective Copying

E.1. Proof of Theorem 1

Proof. Given an N-gram $\boldsymbol{Z} \in \mathbb{R}^{N\times d}$, let us define $\boldsymbol{s}(\boldsymbol{Z}) = \texttt{norm}(\sum_{i\in[N]} F_{N-i}\boldsymbol{z}_i)$ to be its signature. We will first show that for almost all $\boldsymbol{F}$, each N-gram admits a unique signature. To see this, let $\boldsymbol{A}, \boldsymbol{Z} \in \mathbb{R}^{N\times d}$ be two distinct N-grams. Let us write the difference between their signatures as a correlation coefficient. Set $\boldsymbol{s}'(\boldsymbol{Z}) = \sum_{i\in[N]} F_{N-i}\boldsymbol{z}_i$. Note that if $\boldsymbol{s}(\boldsymbol{A}) = \boldsymbol{s}(\boldsymbol{Z})$, we

would have the following function of $\boldsymbol{F}$ that arises from correlation coefficient as zero:

$$g_{\boldsymbol{A},\boldsymbol{Z}}(\boldsymbol{F}) = (\boldsymbol{s}'(\boldsymbol{Z})^\top \boldsymbol{s}'(\boldsymbol{A}))^2 - \|\boldsymbol{s}'(\boldsymbol{Z})\|_{\ell_2}^2 \|\boldsymbol{s}'(\boldsymbol{A})\|_{\ell_2}^2.$$

Now, observe that g is a fourth-order polynomial of the entries of $\boldsymbol{F} \in \mathbb{R}^N$ and we can expand $g(\boldsymbol{F})$ further as follows

$$g(\boldsymbol{F}) = (\sum_{i\in[N]} \sum_{j\in[N]} F_{N-i} F_{N-j} \boldsymbol{a}_i^\top \boldsymbol{z}_j)^2 - \|\sum_{i\in[N]} F_{N-i} \boldsymbol{a}_i\|_{\ell_2}^2 \|\sum_{i\in[N]} F_{N-i} \boldsymbol{z}_i\|_{\ell_2}^2. \tag{1}$$

Above, let c_i be the coefficient of the fourth moment term F_{N-i}^4. Note that

$$c_i = (\boldsymbol{a}_i^\top \boldsymbol{z}_i)^2 - \|\boldsymbol{a}_i\|_{\ell_2}^2 \|\boldsymbol{z}_i\|_{\ell_2}^2.$$

Since $\boldsymbol{A} \neq \boldsymbol{Z}$, there exists $i \in [N]$ such that $\boldsymbol{a}_i \neq \boldsymbol{z}_i$. This implies that $c_i \neq 0$ and $g(\boldsymbol{F})$ is a nonzero polynomial. As a result, $g(\boldsymbol{F}) \neq 0$ almost everywhere implying the same for $\boldsymbol{s}(\boldsymbol{Z}) \neq \boldsymbol{s}(\boldsymbol{A})$. Since there are finitely many N-grams, repeating the same argument for all N-gram pairs, we find that all N-gram signatures are unique for almost all $\boldsymbol{F}$.

Next, suppose we have an $\boldsymbol{F}$ resulting in unique signatures. We will prove the ability of CAT layer to solve the N-AR problem. Consider an arbitrary sequence $\boldsymbol{X}$ and denote the last N tokens by $\boldsymbol{Z}$. Let $\boldsymbol{X}_* = \texttt{norm}(\boldsymbol{X} * \boldsymbol{F})$ be the convolved sequence and let $\boldsymbol{q}$ be the final token of $\boldsymbol{X}_*$. By assumption, $\boldsymbol{q}$ repeats exactly twice in the sequence. Let α be the position of the $\boldsymbol{q}$ in the sequence. By definition, the target token $\boldsymbol{v} = \boldsymbol{x}_{\alpha+1}$. Let $\boldsymbol{\mathcal{I}}_i \in \mathbb{R}^L$ be the indicator function that has 1 at position i and 0 everywhere else. Since all N-grams are unique and their signatures have unit norm, we have that

$$\lim_{c\to\infty} \mathbb{S}(c\boldsymbol{X}_*\boldsymbol{q}) = \boldsymbol{s}_*(\boldsymbol{X}) := \frac{\boldsymbol{\mathcal{I}}_L + \boldsymbol{\mathcal{I}}_\alpha}{2}. \tag{2}$$

Above we use the standard fact that softmax will saturate at the top entry as the inverse-temperature goes to infinity. For the purposes of length generalization, we provide the precise temperature requirement. Let $\boldsymbol{a}, \boldsymbol{b}$ be two vectors in the normalized N-gram token set $\mathcal{S}_N$ (the set of tokens obtained after convolving with $\boldsymbol{F}$). Over all such $\boldsymbol{a}, \boldsymbol{b}$, define the minimum cosine distance to be

$$\Delta = 1 - \max_{\boldsymbol{a}\neq\boldsymbol{b}\in\mathcal{S}_N} \boldsymbol{a}^\top \boldsymbol{b}.$$

Given sequence $\boldsymbol{X}_*$, using the worst case likelihood ratios of $e^{-\Delta}$ between the two $\boldsymbol{q}$-tokens vs the remaining $L-2$ non-$\boldsymbol{q}$ N-grams tokens, for any $\boldsymbol{X}$, we have that

$$\|\text{map}(\boldsymbol{X}, c) - \boldsymbol{s}_*(\boldsymbol{X})\|_{\ell_1} = \|\mathbb{S}(c\boldsymbol{X}_*\boldsymbol{q}) - \boldsymbol{s}_*(\boldsymbol{X})\|_{\ell_1} \leq \frac{2(L-2)e^{-c\Delta}}{2 + (L-2)e^{-c\Delta}}. \tag{3}$$

To make the right hand side $\leq \varepsilon/2$ for all (admissible) sequences $\boldsymbol{X}$ of length at most L, we need $2(L-2)e^{-c\Delta} \leq \varepsilon$ which implies $c \geq \Delta^{-1} \log(\frac{2(L-2)}{\varepsilon})$.

Value delay. For value delay, we will use (3) as key and query embeddings use the same filter. Let $\boldsymbol{X}_v = 2 \cdot \boldsymbol{X} * \boldsymbol{D}_{-1}$. Using the fact that rows of $\boldsymbol{X}_v$ are unit norm, for $c \geq \Delta^{-1} \log(\frac{2(L-2)}{\varepsilon})$

$$\|\boldsymbol{X}_v^\top \text{map}(\boldsymbol{X}, c) - \boldsymbol{X}_v^\top \boldsymbol{s}_*(\boldsymbol{X})\|_{\ell_2} \leq 2\|\mathbb{S}(c\boldsymbol{X}_*\boldsymbol{q}) - \boldsymbol{s}_*(\boldsymbol{X})\|_{\ell_1} \leq \varepsilon.$$

Next, note that

$$\boldsymbol{X}_v^\top \boldsymbol{s}_*(\boldsymbol{X}) = \boldsymbol{X}_v^\top (\frac{\boldsymbol{\mathcal{I}}_L + \boldsymbol{\mathcal{I}}_\alpha}{2}) = \frac{\boldsymbol{v}_\alpha + \boldsymbol{v}_L}{2}.$$

Now observe that, thanks to -1 delay, $\boldsymbol{v}_\alpha = 2\boldsymbol{x}_{\alpha+1}$ and $\boldsymbol{v}_L = 2\boldsymbol{x}_{L+1} = 0$ resulting in $\lim_{c\to\infty} \boldsymbol{X}_v^\top \text{map}(\boldsymbol{X}, c) = \boldsymbol{x}_{\alpha+1}$. Combining the above results, we find that $\|\boldsymbol{V}^\top \text{map}(\boldsymbol{X}, c) - \boldsymbol{v}\|_{\ell_2} \leq \varepsilon$ for all $\boldsymbol{X}$.

Key delay. In this scenario, we are delaying $\boldsymbol{X}_*$ forward by one. Because of this, we have $\boldsymbol{X}_k = \boldsymbol{X}_* * \boldsymbol{D}_1$ and, within $\boldsymbol{X}_k$, $\boldsymbol{q}$ appears in positions $\alpha+1$ and $L+1$. Since the latter is out of bounds, repeating the argument (3) and defining $\text{map}(\boldsymbol{X}, c) := \mathbb{S}(c\boldsymbol{X}_k\boldsymbol{q})$, for any sequence $\boldsymbol{X}$, we find that

$$\|\mathbb{S}(c\boldsymbol{X}_k\boldsymbol{q}) - \boldsymbol{\mathcal{I}}_{\alpha+1}\|_{\ell_1} \leq \frac{2(L-1)e^{-c\Delta}}{1 + (L-1)e^{-c\Delta}}.$$

Similarly, the right hand side is upper bounded by ε, whenever $c \geq \Delta^{-1}\log(\frac{2(L-1)}{\varepsilon})$.

To conclude, using the fact that tokens are unit norm and the target value vector is $\boldsymbol{v} = \boldsymbol{x}_{\alpha+1}$, for any $\boldsymbol{X}$, we obtain

$$\|\boldsymbol{X}^\top \text{map}(\boldsymbol{X}, c) - \boldsymbol{v}\|_{\ell_2} \leq \|\mathbb{S}(c\boldsymbol{X}_*\boldsymbol{q}) - \mathcal{I}_{\alpha+1}\|_{\ell_1} \leq \varepsilon,$$

completing the proof that $\|\boldsymbol{X}^\top \text{map}(\boldsymbol{X}, c) - \boldsymbol{v}\|_{\ell_2}$ for all $\boldsymbol{X}$ of length at most L.

Concluding the proof of the theorem statement. So far, we have concluded that, for all input sequences $\boldsymbol{X}$, CAT layer output guarantees $\|f(\boldsymbol{X}) - \boldsymbol{v}\|_{\ell_2} < \varepsilon_0$ where $\boldsymbol{v}$ is the target value token and ε_0 is under our control by choosing $c = \Delta^{-1}\log(2L/\varepsilon_0)$. Since we assume the minimum distance between distinct token embeddings are ε, to accurately and uniquely decode the target $\boldsymbol{v}$, we choose $\varepsilon_0 = \varepsilon/2$ and apply nearest neighbor on $f(\boldsymbol{X})$ to conclude. □

E.2. Proof of Theorem 2

In this section, we will use the shorthand $\boldsymbol{F}$ to denote the value filter $\boldsymbol{F}_v$ for notational simplicity. Recall that $R_0 > 0$ is an absolute constant throughout the proof. Finally, the constant R used in Theorem 2's statement will be subsumed within the $O(\cdot)$ notation below.

Lemma 1. *Consider the same setting in Theorem 2. For any $f = (\boldsymbol{W}, \boldsymbol{F})$ that can solve the AR problem defined in Def. 2 up to ε-accuracy on all sequences of length $L \geq 3$, if $\varepsilon_0 := \varepsilon/\Delta \leq 1/8$, we have that*

$$\|\boldsymbol{F} - \boldsymbol{D}_{-1}\|_{\ell_1} \leq O(W\varepsilon_0(1 + L\varepsilon_0) + L\varepsilon_0) \leq O(L\varepsilon_0(1 + W\varepsilon_0)) \tag{4}$$

$$\|\boldsymbol{F}_{\geq 0}\|_{\ell_1} = \sum_{i=0}^{W} F_i \leq O(\varepsilon_0(1 + L\varepsilon_0)) \tag{5}$$

where we use $O(\cdot)$ notation to denote an upper bound up to a constant i.e. for some absolute $r > 0$, $O(x) \leq r \cdot x$. Moreover, let $\boldsymbol{q}$ be a token within vocabulary $\mathcal{V}$ and $\boldsymbol{v}$ be the top query not equal to $\boldsymbol{q}$ that maximizes the similarity $\boldsymbol{v}^\top \boldsymbol{W}\boldsymbol{q}$ i.e. $\boldsymbol{v} = \arg\max_{\boldsymbol{x}\in\mathcal{V}, \boldsymbol{x}\neq\boldsymbol{q}} \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{q}$, we have

$$o_q = s_v/s_q \leq \Gamma = \frac{\varepsilon_0}{1 - 4\varepsilon_0} = O(\varepsilon_0) \tag{6}$$

where s_q and s_v are the softmax values for $\boldsymbol{q}$ and $\boldsymbol{v}$.

Proof. Throughout, we assume $\varepsilon \leq \frac{\Delta}{8}$ and $\Delta > 0$ where Δ is the minimum embedding distance, i.e., $\Delta = (1 - \max_{\boldsymbol{a}\neq\boldsymbol{b}\in\mathcal{V}}(\boldsymbol{a}^\top\boldsymbol{b})^2)^{1/2}$. Before proceeding, we first note that, without losing generality, we can assume $L \geq W + 1$. The reason is that, if $L \leq W$, left or right end of the convolutional filter will never interact with features. Thus, we simply set them to zero, truncating the filter. Define sequence $\boldsymbol{X}^i \in \mathbb{R}^{L\times d}$ where $\boldsymbol{x}^i_{L-1} = \boldsymbol{q}$, $\boldsymbol{x}^i_i = \boldsymbol{q}$ and $\boldsymbol{x}^i_j = \boldsymbol{v}$ for all $j \neq i$. Let $\boldsymbol{Z}^i = \boldsymbol{F} * \boldsymbol{X}^i$. Let $\boldsymbol{s}^i = \mathbb{S}(\boldsymbol{X}^i\boldsymbol{W}\boldsymbol{q})$ and $s_q = \boldsymbol{s}^i_{L-1}$ and $s_v = (1 - 2s_q)/(L - 2)$. Here s_q and s_v are the softmax values for $\boldsymbol{q}$ and $\boldsymbol{v}$ respectively. Additionally, observe that

$$s_v/s_q = \exp((\boldsymbol{v} - \boldsymbol{q})^\top \boldsymbol{W}\boldsymbol{q}).$$

Finally, let $\mathcal{I} = [L] - \{L - 1, i\}$ and recalling value sequence $\boldsymbol{Z}^i$, note that

$$f(\boldsymbol{X}^i) = s_v \sum_{j\in\mathcal{I}} \boldsymbol{z}^i_j + s_q(\boldsymbol{z}^i_i + \boldsymbol{z}^i_L).$$

By assumption, we also have that

$$\|\boldsymbol{v} - f(\boldsymbol{X}^i)\|_{\ell_2} \leq \varepsilon \quad \text{for} \quad i < L - 2, \quad \|\boldsymbol{q} - f(\boldsymbol{X}^{L-2})\|_{\ell_2} \leq \varepsilon. \tag{7}$$

We will leverage these inequalities to prove the statement of the theorem. Let $\rho = \rho(\boldsymbol{q}, \boldsymbol{v}) = \boldsymbol{q}^\top\boldsymbol{v}$ be the correlation between $\boldsymbol{q}, \boldsymbol{v}$. Define $\boldsymbol{v}^\perp = \frac{\boldsymbol{q} - \rho\boldsymbol{v}}{\|\boldsymbol{q} - \rho\boldsymbol{v}\|_{\ell_2}}$. Observe that convolution output has the form $f(\boldsymbol{X}^i) = \alpha\boldsymbol{v} + \beta\boldsymbol{q}$ for some $\alpha = \alpha_i, \beta = \beta_i > 0$. For $i < L - 2$, we have that

$$\varepsilon \geq \|\boldsymbol{v} - f(\boldsymbol{X}^i)\|_{\ell_2} \geq |(\boldsymbol{v}^\perp)^\top(\boldsymbol{v} - f(\boldsymbol{X}^i))| \geq \beta|(\boldsymbol{v}^\perp)^\top\boldsymbol{q}| \geq \beta\sqrt{1 - \rho^2}.$$

Recalling that the minimum embedding distance is defined as $\Delta = \sqrt{1 - \max_{\boldsymbol{q},\boldsymbol{v}} \rho^2(\boldsymbol{q},\boldsymbol{v})} \le 1$ and setting $\varepsilon_0 = \varepsilon/\Delta$, this implies that

$$\beta_i \le \varepsilon_0 := \varepsilon/\Delta \quad \text{for} \quad i < L-2, \quad \alpha_{L-2} \le \varepsilon_0 := \varepsilon/\Delta. \tag{8}$$

Additionally, writing $\varepsilon \ge |\boldsymbol{v}^\top(\boldsymbol{v} - f(\boldsymbol{X}^i))| = |1 - \alpha_i - \beta_i \boldsymbol{v}^\top \boldsymbol{q}|$ for $i < L-2$ and using $|\boldsymbol{v}^\top \boldsymbol{q}| \le 1$, we can deduce

$$\alpha_i \ge 1 - (1 + 1/\Delta)\varepsilon \ge 1 - 2\varepsilon_0 \quad \text{for} \quad i < L-2, \quad \beta_{L-2} \ge 1 - (1 + 1/\Delta)\varepsilon \ge 1 - 2\varepsilon_0 \tag{9}$$

$$\alpha_i \le 1 + (1 + 1/\Delta)\varepsilon \le 1 + 2\varepsilon_0 \quad \text{for} \quad i < L-2, \quad \beta_{L-2} \le 1 + (1 + 1/\Delta)\varepsilon \le 1 + 2\varepsilon_0. \tag{10}$$

We note that when $L = W + 1$, the problem only has a subtle difference, which we discuss at the end.

Case 1: $L \ge W + 2$**.** For $i = 0$ and $i = L - 2$, the coefficients α_i, β_i can be written in terms of convolution as

$$\beta_0 = 2s_q F_0 + s_v \sum_{i \ne 0} F_i \tag{11}$$

$$\beta_{L-2} = s_q(2F_0 + F_{-1} + F_1) + s_v[2\sum_{i<0} F_i - (F_{-1} + F_{-W})]. \tag{12}$$

Let $\bar{F}_1 = F_{-1} + F_1$. Observing $2\sum_{i<0} F_i - (F_{-1} + F_{-W}) \le 2(\sum_{i\ne 0} F_i)$, we can write

$$1 - (1 + 1/\Delta)\varepsilon \le \beta_{L-2} \le s_q \bar{F}_1 + 2\beta_0 \le s_q \bar{F}_1 + 2\varepsilon/\Delta. \tag{13}$$

Combining these implies $s_q \bar{F}_1 \ge 1 - 4\varepsilon_0$. Also, we know the trivial bound $s_q \bar{F}_1 \le \beta_{L-2} \le 1 + 2\varepsilon_0$. Thus, we obtain

$$1 + 2\varepsilon_0 \ge s_q \bar{F}_1 \ge 1 - 4\varepsilon_0.$$

To proceed, we wish to prove that s_v is small. From (11), we have that $s_v \bar{F}_1 \le \varepsilon_0$. Consequently, we have that

$$\frac{s_v}{s_q} \le \Gamma = \frac{\varepsilon_0}{1 - 4\varepsilon_0}.$$

Using $2s_q + (L-2)s_v = 1$, we get

$$1 = 2s_q + (L-2)s_v \le (2 + (L-2)\Gamma)s_q \implies s_q \ge \frac{1}{2 + (L-2)\Gamma} = \frac{1 - 4\varepsilon_0}{2 + (L-10)\varepsilon_0}.$$

Since $s_q \le 1/2$ (due to query repeating twice), this also implies that

$$2\frac{(1 + L\varepsilon_0)(1 + 2\varepsilon_0)}{1 - 4\varepsilon_0} \ge 2\frac{(1 + 2\varepsilon_0)(1 + (L/2 - 5)\varepsilon_0)}{1 - 4\varepsilon_0} \ge \bar{F}_1 \ge 2(1 - 4\varepsilon_0).$$

Using above, in essence, so far we have established that $|\bar{F}_1 - 2| \le O(L\varepsilon_0)$ and $s_v/s_q \le O(\varepsilon_0)$. Both statements hold whenever $\varepsilon_0 \le 1/8$ (e.g. so that $1/(1 - 4\varepsilon_0) \le 1 + O(\varepsilon_0)$). The primary remaining item in the proof is establishing $|F_i| \le O(\varepsilon_0)$ for all $i \ne -1$.

To prove this, we utilize the following observations: First, by keeping track of the contributions of the last two $\boldsymbol{q}$ vectors on α_{L-2}, we observe that

$$s_q \sum_{i=1}^{W} F_i \le \alpha_{L-2} \le \varepsilon_0.$$

This implies $\sum_{i=1}^{W} F_i \le \varepsilon_0/s_q \le \varepsilon_0 \frac{2+(L-10)\varepsilon_0}{1-4\varepsilon_0} = O(\varepsilon_0(1 + L\varepsilon_0))$. We similarly find $F_0 \le \varepsilon_0/2s_q$ through (11). Finally, since $F_1 \le O(\varepsilon_0(1 + L\varepsilon_0)) \le O(L\varepsilon_0)$, we also find the critical bound

$$|F_{-1} - 2| \le O(L\varepsilon_0).$$

Finally, we wish to bound $\sum_{i\le -2} F_i$. To do so, we can bound the contribution of the first $\boldsymbol{q}$ vector on β_i as follows. For any $W \ge j \ge 2$, letting $i = L - 1 - j$, we have that

$$\varepsilon_0 \ge \beta_i \ge s_q F_{-j} \implies F_{-j} \le \varepsilon_0 \frac{2 + (L-10)\varepsilon_0}{1 - 4\varepsilon_0} = O(\varepsilon_0(1 + L\varepsilon_0)).$$

Aggregating these, we have found the advertised bounds:

$$\|\boldsymbol{F} - \boldsymbol{D}_{-1}\|_{\ell_1} \le O(W\varepsilon_0(1 + L\varepsilon_0) + L\varepsilon_0) \le O(L\varepsilon_0(1 + W\varepsilon_0)) \tag{14}$$

$$\|\boldsymbol{F}_{\ge 0}\|_{\ell_1} = \sum_{i=0}^{W} F_i \le O(\varepsilon_0(1 + L\varepsilon_0)) \tag{15}$$

$$o_q = s_v/s_q \le \Gamma = \frac{\varepsilon_0}{1 - 4\varepsilon_0} = O(\varepsilon_0) \tag{16}$$

where $\boldsymbol{v}$ is chosen to be the most similar token in terms of attention probabilities. Note that, the bound on left entries of $\boldsymbol{F}$ that retrieves the past values is tighter than the right entries.

Case 2: $L = W + 1$**.** In this scenario, the main difference is we have the following estimates rather than (11)

$$\beta_0 = s_q(2F_0 + F_W + F_{-W}) + s_v \sum_{i\neq 0, |i|<W} F_i \tag{17}$$

$$\beta_{L-2} = s_q(2F_0 + \bar{F}_1) + s_v[2\sum_{i<0} F_i - (F_{-1} + F_{-W})]. \tag{18}$$

So we can't immediately use the estimate provided right below (13) because of the missing $F_{-W}s_v$ term. On the other hand, considering $\boldsymbol{X}^1$ and contribution of the first $\boldsymbol{v}$ token on β_1, we have that $s_vF_{-W} \le \beta_1 \le \varepsilon_0$. As a result, we can instead use the fact that $\beta_0 + \beta_1 \le O(\varepsilon_0)$ and the fact that

$$2s_qF_0 + s_v(2\sum_{i<0} F_i - (F_{-1} + F_{-W})) \le 2(\beta_0 + \beta_1)$$

so that we have again established $|1 - s_q\bar{F}_1| \le O(\varepsilon_0)$ and can proceed similarly. □

Now that we have established the fine-grained control of the filter and attention map with Lemma 1, we can conclude with length generalization.

Proof of Theorem 1. Given a query $\boldsymbol{q}$ and a sequence of length L', let us define s_q similarly (i.e. attention probability that falls on the $\boldsymbol{q}$ token) and study the attention output. Let $\boldsymbol{q}$ appear at i for the first time, $\boldsymbol{v}$ be the token following $\boldsymbol{q}$, and $\mathcal{I} = [L'] - \{i, L' - 1\}$. Let $\boldsymbol{a} = \mathbb{S}(\boldsymbol{XWq}) \in \mathbb{R}^{L'}$ be softmax scores with $a_i = a_{L'-1} = s_q$. We write

$$f(\boldsymbol{X}) = \sum_{j\in\mathcal{I}} a_j\boldsymbol{z}_j + s_q(\boldsymbol{z}_i + \boldsymbol{z}_{L'-1}).$$

where $\boldsymbol{z}_j = \sum_{i=-W}^{W} F_i\boldsymbol{x}_{j-i}$ To proceed, let R be a universal constant and $\Xi = RL\varepsilon_0(1 + W\varepsilon_0)$ so that $\|\boldsymbol{F}\|_{\ell_1} \le 2 + \Xi$ from (14) in Lemma 1. Then we get $\|\boldsymbol{z}_j\|_{\ell_2} \le \|\boldsymbol{F}\|_{\ell_1} \le 2 + \Xi$ for all $j \in [L']$. Secondly, due to right-clipped convolution we have $\|\boldsymbol{z}_{L'-1}\|_{\ell_2} \le \|\sum_{i=0}^{W} F_i\|_{\ell_1} \le \Xi$ and thanks to value retrieval at i'th position, we get

$$\|\boldsymbol{z}_i - 2\boldsymbol{v}\|_{\ell_2} \le |F_{-1} - 2|\|\boldsymbol{v}\|_{\ell_2} + |\sum_{j\neq -1} F_j| \le \Xi \tag{19}$$

Next, observe that $a_j/s_q \le s_v/s_q \le \Gamma = \frac{\varepsilon_0}{1-4\varepsilon_0}$ for all $j \in \mathcal{I}$ and that $2s_q + \sum_{j\in\mathcal{I}} a_j = 1$, consequently, for some constant $R_0 > 0$,

$$\frac{1}{2} \ge s_q \ge \frac{1}{2 + (L' - 2)\Gamma} = \frac{1 - 4\varepsilon_0}{2 + (L' - 10)\varepsilon_0} \implies |2s_q - 1| \le R_0L'\varepsilon_0.$$

and

$$\sum_{j\in\mathcal{I}} a_j = 1 - 2s_q \le R_0L'\varepsilon_0.$$

Aggregating these, we find that

$$\|f(\boldsymbol{X}) - \boldsymbol{v}\|_{\ell_2} \overset{(a)}{\leq} \|\sum_{j\in\mathcal{I}} a_j \boldsymbol{z}_j\|_{\ell_2} + \|s_q(\boldsymbol{z}_i + \boldsymbol{z}_{L'-1}) - 2s_q\boldsymbol{v}\|_{\ell_2} + |2s_q - 1|\|\boldsymbol{v}\|_{\ell_2} \tag{20}$$

$$\overset{(b)}{\leq} |\sum_{j\in\mathcal{I}} a_j|(2+\Xi) + \|s_q(\boldsymbol{z}_i + \boldsymbol{z}_{L'-1}) - 2s_q\boldsymbol{v}\|_{\ell_2} + |2s_q - 1| \tag{21}$$

$$\leq R_0(2+\Xi)L'\varepsilon_0 + \Xi + R_0L'\varepsilon_0 \tag{22}$$

$$\leq 3R_0L'\varepsilon_0 + \Xi + R_0\Xi L'\varepsilon_0 \tag{23}$$

$$\leq 3\varepsilon_0(R_0L' + RL(1+W\varepsilon_0)(1+R_0L'\varepsilon_0)) \tag{24}$$

where (a) follows triangle inequality and (b) follows Cauchy-Schwarz inequality. Let c_0, c_1 be absolute constants to be determined. Assuming $W\varepsilon_0 \leq O(1)$ (i.e. bounded by constant), we have that

$$\|f(\boldsymbol{X}) - \boldsymbol{v}\|_{\ell_2} \leq c_0\varepsilon_0(L' + L + LL'\varepsilon_0)$$

where $c_0 \geq 3\max\{R_0, R(1+W\varepsilon_0), R_0R(1+W\varepsilon_0)\}$. Assuming the stronger bound $L\varepsilon_0 \leq O(1)$ and $c_1 \geq c_0(1 + L/L' + L\varepsilon_0)$, we have that

$$\|f(\boldsymbol{X}) - \boldsymbol{v}\|_{\ell_2} \leq c_1\varepsilon_0L'$$

This concludes the advertised proof. □

E.3. Proving Length Generalization for N-gram AR (Proposition 3)

In this section we use $\boldsymbol{F}_v = \boldsymbol{F}$ for the filter applied on value token and $\boldsymbol{F}_q = \boldsymbol{F}_k = \bar{\boldsymbol{F}}$ for filters on query and keys.

Assumption 1. *Recall that $\mathcal{V}$ is the vocabulary from which the token embeddings are drawn. We have the following two assumptions to make the output $f(\boldsymbol{X})$ more tractable:*

- ***a)*** *The filter weights are bounded and obey $\|\boldsymbol{F}\|_{\ell_1} \leq 1$. Besides, assuming that $\Delta = 1 - \max_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{V},\boldsymbol{a}\neq\boldsymbol{b}} \boldsymbol{b}^\top\boldsymbol{a} > 0$*
- ***b)*** *Any subset of $2N$ tokens within the vocabulary $\mathcal{V}$ is linearly independent.*

Note that Assumption 1.b is essentially a restricted isometry property condition on the embedding matrix induced by the vocabulary $\mathcal{V}$. Specifically, if embeddings are randomly chosen, as soon as the embedding dimension obeys $d \gtrsim O(N\log\frac{|\mathcal{V}|}{N})$, this assumption will hold with high probability (Candes, 2008; Candes & Tao, 2006). In the following analysis, we will leverage either one of the assumptions to establish the length generalization result.

Lemma 2. *Suppose Assumption 1.b holds. Let $\mathcal{B}$ be any subset of $2N$ tokens within $\mathcal{V}$ and $\mathcal{U} := \{\boldsymbol{u}_j | j \in [|\mathcal{U}|]\}$ be the orthonormal tokens obtained after applying the Gram-Schmidt process on $\mathcal{B}$ where $\boldsymbol{b}_j = \sum_{l=0}^{j}\beta_{j,l}\boldsymbol{u}_l$. Then we have $0 < \delta = \min_{j\in[|\mathcal{B}|]}|\beta_{j,j}| \leq 1$.*

Proof. First note that $\beta_{j,l} = \boldsymbol{b}_j^\top\boldsymbol{u}_l \leq 1$ for any $j, l \in [|\mathcal{B}|]$ and $\beta_{0,0} = 1$. Then $\delta = \min_{j\in[|\mathcal{B}|]}|\beta_{j,j}| \leq 1$. Moreover, we can prove $\delta > 0$ by contradiction. Assuming there exists $j \geq 1$ such that $\beta_{j,j} = 0$. This indicates that $\boldsymbol{b}_j$ can represented as a linear combination of the previously orthogonalized vectors $\{\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{j-1}\}$. In other words, $\boldsymbol{b}_j$ lies entirely in the span of these previous vectors. This contradicts the fact that tokens in $\mathcal{B}$ are linearly independent. As a result we have $\delta > 0$. □

Proposition 3. *Let $\bar{\boldsymbol{F}} \in \mathbb{R}^N$ be a 1-D causal convolutional filter and $\boldsymbol{F} \in \mathbb{R}_+^{2W+1}$ be a 1-D convolutional filter from time $t = -W$ to $t = W$ where $W \leq L - N$. Suppose that token embeddings have unit norm. Consider the same CAT Layer $f(\boldsymbol{X}) = (\boldsymbol{X}_v\boldsymbol{W}_v)^\top\mathbb{S}(\boldsymbol{X}_k\boldsymbol{W}_k\boldsymbol{W}_q^\top\boldsymbol{q})$ defined in Theorem 1 where $\boldsymbol{q}$ is the final token of $\boldsymbol{X}_q$ and $\boldsymbol{X}_q = \texttt{norm}(\boldsymbol{X} * \boldsymbol{F}_q) \in \mathbb{R}^{L\times d}$ (same for $\boldsymbol{X}_k, \boldsymbol{X}_v$). We set $\boldsymbol{F}_q = \boldsymbol{F}_k = \bar{\boldsymbol{F}}$, $\boldsymbol{W} = \boldsymbol{W}_k\boldsymbol{W}_q^\top$, and $\boldsymbol{W}_v = 2\boldsymbol{I}_d$ Consider any $f = (\boldsymbol{W}, \boldsymbol{F})$ that can solve the N-AR problem up to ε-accuracy on all sequences of length $L \geq O(N)$. That is, for all $(\boldsymbol{X}, \boldsymbol{y})$ where N-gram $\boldsymbol{Z}$ occurs within $\boldsymbol{X}$ exactly twice and $\boldsymbol{y}$ being the associated value token that follows the first occurrence of $\boldsymbol{Z}$, we have $\|\boldsymbol{y} - f(\boldsymbol{X})\|_{\ell_2} \leq \varepsilon$. Let $\mathcal{B}$ is any subset of $2N$ tokens within vocabulary $\mathcal{V}$ and $\beta_{j,j}$ be the corresponding projection coefficients defined in Lemma 2. Assume either Assumption 1.a or 1.b holds and define*

$$\varepsilon_0 = \begin{cases} \varepsilon/\Delta, & \Delta = 1 - \max_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{V},\boldsymbol{a}\neq\boldsymbol{b}} \boldsymbol{b}^\top\boldsymbol{a} \quad \textit{under Assumption 1.a} \\ \varepsilon\frac{e^{2N/\delta}}{\delta}, & \delta = \min_{j\in[|\mathcal{B}|]}|\beta_{j,j}| \quad \textit{under Assumption 1.b} \end{cases}$$

For almost all choices of $\bar{\boldsymbol{F}}$, there are absolute constants $R_0, R > 0$ such that, if $\varepsilon_0 \le R_0/L$, we have that

- $\|\boldsymbol{F} - \boldsymbol{D}_{-1}\|_{\ell_1} \le L\varepsilon_0$
- *Let $\boldsymbol{s}_\star \in \mathbb{R}^{L'}$ be a vector with entries equal to $1/2$ at the positions of query $\boldsymbol{q}$ in $\boldsymbol{X}_q$ and 0 otherwise. For all inputs $\boldsymbol{X}$ of arbitrary length L', attention map obeys $\|\mathbb{S}(\boldsymbol{X}_k\boldsymbol{W}\boldsymbol{q}) - \boldsymbol{s}_\star\|_{\ell_1} \le L'\varepsilon_0$.*
- *For all N-AR sequences $\boldsymbol{X}$ of arbitrary length L', we have that $\|\boldsymbol{y} - f(\boldsymbol{X})\|_{\ell_2} \le RL'\varepsilon_0$.*

Lemma 3. *Consider the same setting in Prop. 3, for any $f = (\boldsymbol{W}, \boldsymbol{F})$ that can solve the N-AR problem defined in Def. 3 up to ε-accuracy on all sequences of length $L \ge O(N)$. There are absolute constants $R_0 > 0$ such that, if $\varepsilon_0 \le R_0/N$, we have that*

$$\|\boldsymbol{F} - \boldsymbol{D}_{-1}\|_{\ell_1} \le O(N\varepsilon_0(1 + L\varepsilon_0) + N\varepsilon_0) \le O(L\varepsilon_0(1 + N\varepsilon_0)) \tag{25}$$

$$\|\boldsymbol{F}_{\ge 0}\|_{\ell_1} = \sum_{i=0}^{W} F_i \le O(N\varepsilon_0(1 + L\varepsilon_0)) \tag{26}$$

where we use $O(\cdot)$ notation to denote an upper bound up to a constant i.e. for some absolute $r > 0$, $O(x) \le r \cdot x$. Moreover, we consider N-gram $\boldsymbol{Z}_q \in \mathbb{R}^{N\times d}$ that ends with a token $\boldsymbol{q}'$, which can be any token from the vocabulary $\mathcal{B}_N$. Let $\boldsymbol{q}$ be the final token of $\texttt{norm}(\boldsymbol{Z}_q * \bar{\boldsymbol{F}})$ and $\boldsymbol{v}$ be the top query not equal to $\boldsymbol{q}$ that maximizes the similarity $\boldsymbol{v}^\top\boldsymbol{W}\boldsymbol{q}$. i.e. $\boldsymbol{v} = \arg\max_{\boldsymbol{x}\in\mathcal{B}_N, \boldsymbol{x}\neq\boldsymbol{q}} \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{q}$, we have

$$o_q = \frac{s_v}{s_q} \le \Gamma = \frac{O(\varepsilon_0)}{1 - O(N\varepsilon_0)} \le O(N\varepsilon_0) \tag{27}$$

where s_q and s_v are the softmax values for $\boldsymbol{q}$ and $\boldsymbol{v}$.

Proof. Following the proof of Thoerem 1, suppose that we have an $\bar{\boldsymbol{F}}$ that results in unique signatures. We argue that the length generalization fails when $W > L - N$, which is explained at the end. Throughout, we assume that $W = L - N$. When $W < L - N$, it is equivalent to the setting where $W = L - N$ and $F_j = 0$ for $W + 1 \le |j| \le L - N$. Denote the corresponding N-gram that results in $\boldsymbol{q}$ and $\boldsymbol{v}$ after convolving with $\bar{\boldsymbol{F}}$ be $\boldsymbol{Z}_q = [\boldsymbol{q}_0, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_{N-1}] \in \mathbb{R}^{N\times d}$ and $\boldsymbol{Z}_v = [\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{N-1}] \in \mathbb{R}^{N\times d}$ respectively, i.e., $\boldsymbol{q} = \texttt{norm}(\boldsymbol{Z}_q * \bar{\boldsymbol{F}})$ and $\boldsymbol{v} = \texttt{norm}(\boldsymbol{Z}_v * \bar{\boldsymbol{F}})$. $\boldsymbol{Z}_q$ and $\boldsymbol{Z}_v$ are unique due to the assumption on $\bar{\boldsymbol{F}}$. For brevity, let $\boldsymbol{q}' = \boldsymbol{q}_{N-1}, \boldsymbol{v}' = \boldsymbol{v}_{N-1}$ and $\boldsymbol{Z}_{v'}^k = [\boldsymbol{v}', \boldsymbol{v}', \ldots, \boldsymbol{v}'] \in \mathbb{R}^{k\times d}$, $\boldsymbol{Z}_q' = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{N-1}] \in \mathbb{R}^{(N-1)\times d}$, where $\boldsymbol{q}_0$ is removed from $\boldsymbol{Z}_q$.

$$\boldsymbol{X}^{i,k} = \begin{bmatrix} \boldsymbol{Z}_{v'}^{N-1+i} & \boldsymbol{Z}_v & \boldsymbol{Z}_q & \boldsymbol{Z}_{v'}^{N-1+k} & \boldsymbol{Z}_q' & \boldsymbol{q}_0 & \boldsymbol{Z}_{v'}^{N-1} & \boldsymbol{Z}_{v'}^{n_{i,k}} & \boldsymbol{Z}_v & \boldsymbol{Z}_q \end{bmatrix} \in \mathbb{R}^{L\times d} \tag{28}$$

$$\bar{\boldsymbol{X}}^{i,k} = \begin{bmatrix} \boldsymbol{Z}_{v'}^{N-1+i} & \boldsymbol{Z}_v & \boldsymbol{Z}_q & \boldsymbol{q}_0 & \boldsymbol{Z}_{v'}^{N-1+k} & \boldsymbol{Z}_q' & \boldsymbol{Z}_{v'}^{N-1} & \boldsymbol{Z}_{v'}^{n_{i,k}} & \boldsymbol{Z}_v & \boldsymbol{Z}_q \end{bmatrix} \in \mathbb{R}^{L\times d} \tag{29}$$

where $n_{i,k} = L - 8N + 3 - i - k \ge 0$, and this naturally introduces a lower bound for L, i.e., $L \ge O(N)$ and upper bounds for both i and k. Note that $\boldsymbol{X}^{i,k}$ and $\bar{\boldsymbol{X}}^{i,k}$ have different labels. By assumption, we have that

$$\|\boldsymbol{v}' - f(\boldsymbol{X}^{i,k})\|_{\ell_2} \le \varepsilon, \quad \|\boldsymbol{q}_0 - f(\bar{\boldsymbol{X}}^{i,k})\|_{\ell_2} \le \varepsilon. \tag{30}$$

Let $\boldsymbol{s}^{i,k} = \mathbb{S}(\boldsymbol{X}^{i,k}\boldsymbol{W}\boldsymbol{q})$, $\bar{\boldsymbol{s}}^{i,k} = \mathbb{S}(\bar{\boldsymbol{X}}^{i,k}\boldsymbol{W}\boldsymbol{q})$. Define the probability of selecting the j-th entry of $\boldsymbol{X}^{i,k}$ and $\bar{\boldsymbol{X}}^{i,k}$ as $s_j^{i,k}$ and $\bar{s}_j^{i,k}$ and selecting the token $\boldsymbol{q}$ and $\boldsymbol{v}$ as s_q, s_v. Here we omit i, k for s_q and s_v since it's invariant to the values of i, k. Additionally, observe that

$$s_v/s_q = \exp((\boldsymbol{v} - \boldsymbol{q})^\top\boldsymbol{W}\boldsymbol{q}) \text{ and } (L-2)s_v + 2s_q \ge 1$$

where the inequality comes from the fact that $\boldsymbol{v} = \arg\max_{\boldsymbol{x}\in\mathcal{B}_N, \boldsymbol{x}\neq\boldsymbol{q}} \boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{q}$. We will leverage these inequalities to prove the statement of the theorem. Define the vocabulary set $\mathcal{B} = \{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{N-1}, \boldsymbol{q}_0, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_{N-1}\}$ which includes all tokens in $\boldsymbol{X}^{i,k}$ and $\bar{\boldsymbol{X}}^{i,k}$. Note that the vocabulary $\mathcal{B}$ is a subset of tokens chosen from $\mathcal{V}$, i.e., $\mathcal{B} \subseteq \mathcal{V}$ and the vocabulary size $|\mathcal{B}|$ is at most $2N$, i.e., $|\mathcal{B}| := K \le 2N$. Observe that convolution output has the form $f(\boldsymbol{X}^{i,k}) = \sum_{j\in[|\mathcal{B}|]} m_j^{i,k}\boldsymbol{b}_j$ and $f(\bar{\boldsymbol{X}}^{i,k}) = \sum_{j\in[|\mathcal{B}|]} \bar{m}_j^{i,k}\boldsymbol{b}_j$ where $\{m_j^{i,k}, \bar{m}_j^{i,k}\}_{j\in[|\mathcal{B}|]}$ are non-negative coefficients due to the assumption that entries in $\bar{\boldsymbol{F}}$ and softmax probabilities $\boldsymbol{s}^{i,k}$ and $\bar{\boldsymbol{s}}^{i,k}$ are non-negative. In particular, we are interested in $m_q^{i,k}, \bar{m}_q^{i,k}$ and $m_v^{i,k}, \bar{m}_v^{i,k}$, which correspond to the coefficients of token $\boldsymbol{q}_0$ and $\boldsymbol{v}'$. To proceed, we leverage Assumption 1 to bound the coefficients:

When Assumption 1.a holds. By expanding the coefficients, we get

$$\sum_{j\in[|\mathcal{B}|]} m_j^{i,k} = \sum_{j\in[|\mathcal{B}|]}\sum_{t\in[L]} F_v t - j s_t^{i,k} \le \|\boldsymbol{F}\|_{\ell_1} \sum_{t\in[L]} s_t^{i,k} \le 1 \tag{31}$$

Combining this with the fact that

$$\varepsilon \ge \|\boldsymbol{v}' - f(\boldsymbol{X}^{i,k})\|_{\ell_2} \ge |\boldsymbol{v}'^\top(\boldsymbol{v}' - f(\boldsymbol{X}^{i,k}))| \ge |\sum_{j\in[|\mathcal{B}|],\boldsymbol{b}_j\neq\boldsymbol{v}'} \boldsymbol{v}'^\top \boldsymbol{b}_j m_j^{i,k} + m_v^{i,k} - 1| \tag{32}$$

, we have

$$\varepsilon \ge \sum_{j\in[|\mathcal{B}|],\boldsymbol{b}_j\neq\boldsymbol{v}'} (1 - \boldsymbol{v}'^\top \boldsymbol{b}_j) m_j^{i,k} \ge (1 - \boldsymbol{v}'^\top \boldsymbol{b}_j) m_j^{i,k} \quad \text{for any} \quad j \in \{j \mid j \in [|\mathcal{B}|], \boldsymbol{b}_j \neq \boldsymbol{v}'\} \tag{33}$$

$$\to m_j^{i,k} \le \varepsilon/\Delta := \varepsilon_0 \quad \text{for any} \quad j \in \{j \mid j \in [|\mathcal{B}|], \boldsymbol{b}_j \neq \boldsymbol{v}'\} \tag{34}$$

where $\Delta = 1 - \max_{\boldsymbol{a},\boldsymbol{b}\in\mathcal{B},\boldsymbol{a}\neq\boldsymbol{b}} \boldsymbol{b}^\top \boldsymbol{a} > 0$. In terms of $m_v^{i,k}$, we apply Triangle Inequality on (32) and (34):

$$|1 - m_v^{i,k}| \le |\sum_{j\in[|\mathcal{B}|],\boldsymbol{b}_j\neq\boldsymbol{v}'} \boldsymbol{v}'^\top \boldsymbol{b}_j m_j^{i,k} + m_v^{i,k} - 1| + |\sum_{j\in[|\mathcal{B}|],\boldsymbol{b}_j\neq\boldsymbol{v}'} \boldsymbol{v}'^\top \boldsymbol{b}_j m_j^{i,k}| \tag{35}$$

$$\le \varepsilon + 2N\varepsilon_0 \le (2N+1)\varepsilon_0 \tag{36}$$

Similarly for $\bar{\boldsymbol{X}}^{i,k}$ we have

$$\bar{m}_j^{i,k} \le \varepsilon_0 \quad \text{for any} \quad j \in \{j \mid j \in [|\mathcal{B}|], \boldsymbol{b}_j \neq \boldsymbol{q}_0\}, \quad |1 - \bar{m}_q^{i,k}| \le O(N\varepsilon_0) \tag{37}$$

When Assumption 1.b holds. Based on the linear independence property, we can apply the Gram–Schmidt process to transform the tokens in $\mathcal{B}$ to orthonormal tokens $\mathcal{U} = \{\boldsymbol{u}_j | j \in [|\mathcal{U}|]\}$ where $\boldsymbol{b}_j = \sum_{l=0}^{j} \beta_{j,l}\boldsymbol{u}_l$ where $\beta_{j,l} = \boldsymbol{b}_j^\top \boldsymbol{u}_l$. Since the order of tokens in $\mathcal{U}$ does not matter, we can set $\boldsymbol{u}_0 = \boldsymbol{v}'$. Then for any $j \ge 1$, $\boldsymbol{u}_j$ is orthogonal to $\boldsymbol{v}'$ and $\boldsymbol{b}_i$ for all $i < j$. Consider the case of $\boldsymbol{X}^{i,k}$ whose label is $\boldsymbol{v}'$, utilizing the orthogonality we get

$$\varepsilon \ge \|\boldsymbol{v}' - f(\boldsymbol{X}^{i,k})\|_{\ell_2} \ge |\boldsymbol{u}_j^\top(\boldsymbol{v}' - f(\boldsymbol{X}^{i,k}))| \ge |\sum_{l=j}^{|\mathcal{B}|-1} m_l^{i,k} \boldsymbol{u}_j^\top \boldsymbol{b}_l| \tag{38}$$

Using backward induction, we can then bound $m_j^{i,k}$ for $1 \le j \le K-1$. First consider $j = |\mathcal{B}| - 1 = K - 1$. Then we have:

$$\varepsilon \ge |m_{K-1}\boldsymbol{u}_{K-1}^\top \boldsymbol{b}_{K-1}| = |m_{K-1}\beta_{K-1,K-1}| \ge |m_{K-1}\delta| \tag{39}$$

where $\delta = \min_{j\in[|\mathcal{B}|]} |\beta_{j,j}| = \min_{j\in[|\mathcal{B}|]} |\boldsymbol{b}_j^\top \boldsymbol{u}_j|$. Following Lemma 2 we have $0 < \delta \le 1$. As a result we get $m_{K-1} \le \varepsilon/\delta$. Next we prove that if $j \ge 1$ and $m_l^{i,k} \le \varepsilon\frac{(1+1/\delta)^{K-l-1}}{\delta}$ for $j < l \le K-1$, $m_j \le \varepsilon\frac{(1+1/\delta)^{K-j-1}}{\delta}$. When $1 \le j \le K-2$, from equation (38) we can derive

$$\varepsilon \ge |\sum_{l=j}^{K-1} m_l^{i,k}\boldsymbol{u}_j^\top \boldsymbol{b}_l| = |\sum_{l=j+1}^{K-1} m_l^{i,k}\boldsymbol{u}_j^\top \boldsymbol{b}_l + m_j \boldsymbol{u}_j^\top \boldsymbol{b}_j| \tag{40}$$

For the first term we have

$$|\sum_{i=j+1}^{K-1} m_i \boldsymbol{u}_j^\top \boldsymbol{b}_i| \le \sum_{i=j+1}^{K-1} m_i \le \sum_{i=0}^{K-j-2} \varepsilon\frac{(1+1/\delta)^i}{\delta} = \varepsilon((1+1/\delta)^{K-j-1} - 1)$$

Using Triangle Inequality we get

$$|m_j \boldsymbol{u}_j^\top \boldsymbol{b}_j| \le \varepsilon(1+1/\delta)^{K-j-1} \to m_j \le \varepsilon\frac{(1+1/\delta)^{K-j-1}}{\delta} \le \varepsilon\frac{e^{\frac{K-j-1}{\delta}}}{\delta} \le \varepsilon\frac{e^{2N/\delta}}{\delta} \tag{41}$$

We can hereby bound $m_j^{i,k}$ for any $j \in \{j \mid \boldsymbol{b}_j \in \mathcal{B}, \boldsymbol{b}_j \neq \boldsymbol{v}'\}$. Let $\varepsilon_1 := \varepsilon \frac{e^{2N/\delta}}{\delta}$, we have

$$m_j^{i,k} \leq \varepsilon_1 \quad \text{for any} \quad j \in [|\mathcal{B}|], \boldsymbol{b}_j \neq \boldsymbol{v}' \tag{42}$$

Additionally, writing $\varepsilon \geq |\boldsymbol{v}'^\top (\boldsymbol{v}' - f(\boldsymbol{X}^{i,k}))| = |1 - m_v^{i,k} - \boldsymbol{v}'^\top \sum_{j \in [|\mathcal{B}|], \boldsymbol{b}_j \neq \boldsymbol{v}'} m_j^{i,k} \boldsymbol{b}_j|$ and using $|\boldsymbol{v}'^\top \boldsymbol{b}_j| \leq 1$ for any $\boldsymbol{b}_j \in \mathcal{B}$, we can deduce

$$|1 - m_v^{i,k}| \leq \varepsilon + \sum_{i=1}^{K-1} m_i \tag{43}$$

$$\leq \varepsilon + \sum_{i=0}^{K-2} \varepsilon \frac{(1 + 1/\delta)^i}{\delta} \tag{44}$$

$$\leq \varepsilon (1 + 1/\delta)^{K-1} \leq \varepsilon_1 \tag{45}$$

Similarly for $\bar{\boldsymbol{X}}^{i,k}$, we have

$$\bar{m}_j^{i,k} \leq \varepsilon_1 \quad \text{for any} \quad j \in [|\mathcal{B}|], \boldsymbol{b}_j \neq \boldsymbol{q}_0, \quad |1 - \bar{m}_q^{i,k}| \leq \varepsilon_1 \tag{46}$$

To summarize, using Assumption 1, we can have an upper bound on $\{m_j^{i,k}, \bar{m}_j^{i,k}\}_{j \in [|\mathcal{B}|]}$:

$$m_j^{i,k} \leq \varepsilon_0 \quad \text{for any} \quad j \in \{j \mid j \in [|\mathcal{B}|], \boldsymbol{b}_j \neq \boldsymbol{v}'\}, \quad |1 - m_v^{i,k}| \leq O(N\varepsilon_0) \tag{47}$$

$$\bar{m}_j^{i,k} \leq \varepsilon_0 \quad \text{for any} \quad j \in \{j \mid j \in [|\mathcal{B}|], \boldsymbol{b}_j \neq \boldsymbol{q}_0\}, \quad |1 - \bar{m}_q^{i,k}| \leq O(N\varepsilon_0) \tag{48}$$

where

$$\varepsilon_0 := \begin{cases} \varepsilon/\Delta, & \Delta = 1 - \max_{\boldsymbol{a},\boldsymbol{b} \in \mathcal{B}, \boldsymbol{a} \neq \boldsymbol{b}} \boldsymbol{b}^\top \boldsymbol{a} > 0 \quad \text{under Assumption 1.a} \\ \varepsilon \frac{e^{2N/\delta}}{\delta}, & \delta = \min_{j \in [|\mathcal{B}|]} |\beta_{j,j}| \quad \text{under Assumption 1.b} \end{cases} \tag{49}$$

We proceed by comparing $m_q^{i,k}$ and $\bar{m}_q^{i,k}$:

$$m_q^{i,k} = 2s_q(F_{-L+4N+i-2} + F_{-2N+1-k} + 2F_{N-1} + F_{L-5N-i-k} + F_{L-2N-i}) \tag{50}$$

$$+ 2 \sum_{j \in [L] - \{3N-2+i, L-1\}} s_j^{i,k} (F_{-L+N+j} + F_{-5N-i-k+j+3} + F_{-2N-i+j+1}) \tag{51}$$

$$\bar{m}_q^{i,k} = 2s_q(F_{-L+4N+i-2} + F_{-1} + 2F_{N-1} + F_{L-3N-i} + F_{L-2N-i}) \tag{52}$$

$$+ 2 \sum_{j \in [L] - \{3N-2+i, L-1\}} \bar{s}_j^{i,k} (F_{-L+N+j} + F_{-3N-i+j+1} + F_{-2N-i+j+1}) \tag{53}$$

Observing that

- $s_j^{i,k} = \bar{s}_j^{i,k}$ for $j \in [3N - 1 + i]$ and $6N - 3 + i + k \leq j \leq L - 1$
- $s_{j+1}^{i,k} = \bar{s}_j^{i,k}$ for $4N - 1 + i + k \leq j \leq 5N - 3 + i + k$
- $s_{j+(2N-1)}^{i,k} = \bar{s}_j^{i,k}$ for $5N - 2 + i + k_1 \leq j \leq 6N - 4 + i + k$
- $s_{j+2N-2+k_1+i_1-i_2}^{i_1,k_1} = \bar{s}_j^{i_2,k_2}$ for $3N - 1 + i_2 \leq j \leq 4N - 2 + i_2$

Utilizing these observations, we have

$$\sum_{j\in[3N-1+i]} 2\bar{s}_j^{i,k}(F_{-L+N+j}+F_{-3N-i+j+1}+F_{-2N-i+j+1}) \le m_q^{i,k}+m_q^{i+N,k}$$

$$\sum_{j\in[6N-3+i+k,L-1]} 2\bar{s}_j^{i,k}(F_{-L+N+j}+F_{-3N-i+j+1}+F_{-2N-i+j+1}) \le m_q^{i,k}+m_q^{i,k-N}$$

$$\sum_{4N-1+i+k\le j\le 5N-3+i+k} 2\bar{s}_j^{i,k}(F_{-L+N+j}+F_{-3N-i+j+1}+F_{-2N-i+j+1}) \le m_q^{i+1,k}+m_q^{i-N+1,k}+m_q^{i,k+1}$$

$$\sum_{5N-2+i+k\le j\le 6N-4+i+k} 2\bar{s}_j^{i,k}(F_{-L+N+j}+F_{-3N-i+j+1}+F_{-2N-i+j+1}) \le m_q^{i+(N-1),k}+m_q^{i+(2N-1),k}+m_q^{i,k+(2N-1)}$$

$$\begin{aligned}\sum_{3N-1+i\le j\le 4N-2+i} 2\bar{s}_j^{i,k}(F_{-L+N+j}+F_{-3N-i+j+1}+F_{-2N-i+j+1}) &\le m_q^{i-(2N-2),k}+\sum_{3N-1+i\le j\le 4N-2+i}\bar{s}_j^{i,k}(F_{-3N-i+j+1}+F_{-2N-i+j+1})\\ &\overset{(a)}{\le} m_q^{i-(2N-2),k}+s_v\sum_{l\in[2N]}F_l\\ &\overset{(b)}{\le} m_q^{i-(2N-2),k}+\sum_{j\in\{j\,|\,\boldsymbol{b}_j=\bar{\boldsymbol{x}}_l^{i,k},l\in[2N]\}}\bar{m}_j^{1,k}\end{aligned}$$

where (a) comes from $\boldsymbol{v}=\arg\max_{\boldsymbol{x}\in\mathcal{B}_N,\boldsymbol{x}\neq\boldsymbol{q}}\boldsymbol{x}^\top\boldsymbol{W}\boldsymbol{q}$ and (b) comes from the attention from the first $\boldsymbol{v}$ to itself and its previous $2N-1$ terms. Let $i=2N-2, k=N$, combining the inequalities above, we get

$$1-O(N\varepsilon_0)\le\bar{m}_q^{2N-2,N}\le 2s_q(F_{-1}+F_{L-5N+2})+m_q^{2N-2,N}+m_q^{3N-2,N}+...+m_q^{0,N}+\sum_{j\in\{j\,|\,\boldsymbol{b}_j=\bar{\boldsymbol{x}}_l^{i,k},l\in[2N]\}}\bar{m}_j^{1,N} \tag{54}$$

$$\le 2s_q(F_{-1}+F_{L-5N+2})+O(N\varepsilon_0) \tag{55}$$

Combining these we get $s_q(F_{-1}+F_{L-5N+2})\ge 1/2-O(N\varepsilon_0)$. Moreover, from (52), we have $s_q(F_{-1}+F_{L-5N+2})\le\bar{m}_q^{i,k}/2\le 1/2+O(N\varepsilon_0)$, which results in

$$|s_q(F_{-1}+F_{L-5N+2})-1/2|\le O(N\varepsilon_0) \tag{56}$$

Based on this, we wish to show that s_v is small. Substituting $l\in\{4N-4,L-N\}$ into (51), we have $s_v(F_{-1}+F_{L-5N+2})\le m_q^{2N-2,N}\le\varepsilon_0$, which implies that

$$\frac{s_v}{s_q}\le\Gamma=\frac{O(\varepsilon_0)}{1/2-O(N\varepsilon_0)} \tag{57}$$

Using $2s_q+(L-2)s_v\ge 1$, we have

$$1\le 2s_q+(L-2)s_v\le(2+(L-2)\Gamma)s_q\implies s_q\ge\frac{1}{2+(L-2)\Gamma}\ge\frac{1-O(N\varepsilon_0)}{2+(L-2N-2)O(\varepsilon_0)} \tag{58}$$

Combining this with $2s_q\le 1$, we get

$$\frac{(1+LO(\varepsilon_0))(1+O(N\varepsilon_0))}{1-O(N\varepsilon_0)}\ge\frac{(1+O(N\varepsilon_0))(1+(L/2-N-1)O(\varepsilon_0))}{1-O(N\varepsilon_0)} \tag{59}$$

$$\ge F_{-1}+F_{L-5N+2}\ge 1-O(N\varepsilon_0) \tag{60}$$

At this stage, we have already proved that when $N\varepsilon_0\le O(1)$, $|F_{-1}+F_{L-5N+2}-1|\le O(L\varepsilon_0)$ and $s_v/s_q\le O(\varepsilon_0)$. The primary remaining proof is to prove $|F_l|<O(\varepsilon_0)$ for all $l\neq -1$. Since $\bar{m}_j^{i,k}\le\varepsilon_0$ for $\boldsymbol{b}_j\neq\boldsymbol{q}_0$, by tracking the contribution of the last $\boldsymbol{q}$ on $\bar{m}_j^{i,k}$, we have

$$s_q\sum_{0\le l\le L-N,l\notin\{N-1,L-2N,L-3N\}}F_l\le\sum_{j\in\{j\,|\,\boldsymbol{b}_j\in\mathcal{B},\boldsymbol{b}_j\neq\boldsymbol{q}_0\}}\bar{m}_j^{0,0}\le O(N\varepsilon_0) \tag{61}$$

$$s_q\sum_{l\in\{L-2N,L-3N\}}F_l\le\sum_{j\in\{j\,|\,\boldsymbol{b}_j\in\{\boldsymbol{q}',\boldsymbol{v}'\}\}}\bar{m}_j^{1,0}\le\varepsilon_0 \tag{62}$$

$$s_qF_{N-1}\le m_q^{0,0}\le\varepsilon_0 \tag{63}$$

Combining these three inequalities and $W = L-N$, we get $\sum_{l=0}^{W} F_l \le O(N\varepsilon_0)/s_q \le O(N\varepsilon_0)\frac{2+(L-2N-2)O(\varepsilon_0)}{1-O(N\varepsilon_0)} \le O(N\varepsilon_0(1+L\varepsilon_0))$. Additionally, since $F_{L-5N+2} \le O(N\varepsilon_0(1+L\varepsilon_0)) \le O(NL\varepsilon_0)$, we get

$$|F_{-1} - 1| < O(NL\varepsilon_0) \tag{64}$$

Finally, we want to bound $\sum_{-L+N\le j\le -2} F_j$. By tracking the contribution of the first $\boldsymbol{q}$ to $\bar{m}_j^{0,0}$ for any $j \in \{j \mid \boldsymbol{b}_j \in \mathcal{B}, \boldsymbol{b}_j \neq \boldsymbol{q}_0\}$, we have $s_q F_l \le \bar{m}_j^{0,0}$ where $\boldsymbol{b}_j = \boldsymbol{x}_{-l+3N-2}^{i,k}$ for any $-L+3N-1 \le l \le -2$. Summing over l we get,

$$s_q \sum_{-L+3N-1\le l\le -2, l\neq W-4N+3} F_l \le \sum_{j\in\{j \mid \boldsymbol{b}_j\in\mathcal{B}, \boldsymbol{b}_j\neq \boldsymbol{q}_0\}} \bar{m}_j^{0,0} \le O(N\varepsilon_0) \tag{65}$$

Note that $F_{-L+4N-2}$ touches $\boldsymbol{q}_0$ for $\bar{\boldsymbol{X}}^{0,0}$ so we need to handle it separately. We can consider $\bar{m}_j^{1,0}$ instead and get

$$s_q F_{-L+4N-2} \le \bar{m}_{L-N+1}^{1,0} \le \varepsilon_0 \tag{66}$$

For $\sum_{-L+N\le j\le -L+3N-2} F_j$, we consider the following example:

$$\hat{\boldsymbol{X}}^{i,k} = \begin{bmatrix} \boldsymbol{Z}_q & \boldsymbol{v}_0 & \boldsymbol{Z}_{v'}^{N-1+i} & \boldsymbol{Z}_v & \boldsymbol{Z}_{v'}^{N-1+k} & \boldsymbol{Z}_q' & \boldsymbol{Z}_{v'}^{N-1} & \boldsymbol{Z}_{v'}^{n_{i,k}} & \boldsymbol{Z}_q \end{bmatrix} \in \mathbb{R}^{L\times d} \tag{67}$$

Similarly we define $\hat{\boldsymbol{s}}^{i,k} = \mathbb{S}(\hat{\boldsymbol{X}}^{i,k}\boldsymbol{W}\boldsymbol{q})$ and the probability that selects $\boldsymbol{q}$ and $\boldsymbol{v}$ as $\hat{s}_q$ and $\hat{s}_v$. Note that $\hat{s}_v/\hat{s}_q = s_v/s_q \le O(\varepsilon_0)$ and $(L-2)\hat{s}_v + \hat{s}_q \ge 1$. We can then obtain the exact lower bound as s_q for $\hat{s}_q$, i.e., $\hat{s}_q \ge \frac{1-O(N\varepsilon_0)}{2+(L-2N-2)O(\varepsilon_0)}$. Note that the output of $f(\hat{\boldsymbol{X}}^{i,k})$ can also be written as $f(\hat{\boldsymbol{X}}^{i,k}) = \sum_{\boldsymbol{b}_j\in\mathcal{B}} \hat{m}_j^{i,k}\boldsymbol{b}_j$ where $\{\hat{m}_j^{i,k}\}_{\boldsymbol{b}_j\in\mathcal{B}}$ is the corresponding coefficients. Moving forward, by tracking the attendance of the first $\boldsymbol{q}$ to the last $2N-1$ terms, we have $\hat{s}_q F_l \le \hat{m}_j^{0,0}$ where $\boldsymbol{b}_j = \hat{\boldsymbol{x}}_{-l+i+N-1}^{i,k}$ for any $-L+N \le j \le -L+3N-2$. Repeating the same argument based on Assumption 1, we get $\hat{m}_j^{0,0} \le \varepsilon_0$ for any $j \in \{j \mid \boldsymbol{b}_j \in \mathcal{B}, \boldsymbol{b}_j \neq \boldsymbol{v}_0\}$, which leads to

$$\hat{s}_q \sum_{-L+N\le l\le -L+3N-2} F_l \le \hat{s}_q \sum_{j\in\{j \mid \boldsymbol{b}_j\in\mathcal{B}, \boldsymbol{b}_j\neq \boldsymbol{v}_0\}} \hat{m}_j^{0,0} \le O(N\varepsilon_0) \tag{68}$$

Combining (65), (66) and (68), we have

$$\sum_{-L+N\le l\le -2} F_l \le O(N\varepsilon_0)(1/\hat{s}_q + 1/s_q) \le O(N\varepsilon_0)\frac{2+(L-2N-2)O(\varepsilon_0)}{1-O(N\varepsilon_0)} \le O(N\varepsilon_0(1+L\varepsilon_0)) \tag{69}$$

In summary, we get

$$\|\boldsymbol{F} - \boldsymbol{D}_{-1}\|_{\ell_1} \le O(N\varepsilon_0(1+L\varepsilon_0) + N\varepsilon_0) \le O(L\varepsilon_0(1+N\varepsilon_0)) \tag{70}$$

$$\|\boldsymbol{F}_{\ge 0}\|_{\ell_1} = \sum_{i=0}^{W} F_i \le O(N\varepsilon_0(1+L\varepsilon_0)) \tag{71}$$

$$o_q = \frac{s_v}{s_q} \le \Gamma = \frac{O(\varepsilon_0)}{1-O(N\varepsilon_0)} \le O(N\varepsilon_0) \tag{72}$$

where $\boldsymbol{v}$ is chosen to be the most similar tokens to $\boldsymbol{q}$ in terms of attention probabilities. Now we discuss the scenario where $W > L-N$. First note that the output $f(\boldsymbol{X})$ can be written as $f(\boldsymbol{X}) = 2\sum_{l\in[L]} s_l \sum_{j=-W}^{W} F_j \boldsymbol{x}_{l-j}$ where s_l is the softmax value at l-th position. We are interested in s_q in particular, which is proven to converge to $1/2$ when ε diminishes. Also, note that the smallest possible index of $\boldsymbol{q}$ is $N-1$ since it's the last token of an N-gram. Then, when $W > L-N$, the left end of the convolutional filter never interacts with s_q since the index of $\boldsymbol{x}_{i-j}$ is out of bound, i.e., $i-j = N-1+W > L-1$ □

Using the results from Lemma 3, we can establish the length generalization on N-AR task.

Proof of Proposition 3. Given a sequence $\boldsymbol{X}$ of length L', let $\boldsymbol{q}$ be the last token of $\boldsymbol{X}_q = \boldsymbol{X}_k = \texttt{norm}(\boldsymbol{X} * \bar{\boldsymbol{F}})$ and we define s_q as the attention probability that selects $\boldsymbol{q}$. Assume the first occurrence of $\boldsymbol{q}$ in $\boldsymbol{X}_k$ is i and $\boldsymbol{q}' = \boldsymbol{x}_i$. By definition, the target

vector $\boldsymbol{v}$ is the token following $\boldsymbol{q}'$ in $\boldsymbol{X}$, i.e., $\boldsymbol{v} = \boldsymbol{x}_{i+1}$. Let $\mathcal{I} = [L'] - \{i, L'-1\}$. Let $\boldsymbol{a} = \mathbb{S}(\boldsymbol{X}_k \boldsymbol{W}\boldsymbol{q}) \in \mathbb{R}^{L'}$ be the softmax probabilities where $a_i = a_{L'-1} = s_q$

$$f(\boldsymbol{X}) = \sum_{j\in\mathcal{I}} a_j \boldsymbol{z}_j + s_q(\boldsymbol{z}_i + \boldsymbol{z}_{L'-1}) \tag{73}$$

where $\boldsymbol{z}_j = 2\sum_{i=-W}^{W} F_i \boldsymbol{x}_{j-i}$. We define R as a universal constant and $\Xi = RL\varepsilon_0(1+N\varepsilon_0)$ such that $\|\boldsymbol{F}\|_{\ell_1} \le 1+\Xi$ from Lemma 3. Then we get $\|\boldsymbol{z}_j\|_{\ell_2} \le 2\|\boldsymbol{F}\|_{\ell_1} \le 2(1+\Xi)$ for all $j \in [L']$. Note that $a_j/s_q \le s_v/s_q = \Gamma = \frac{O(\varepsilon_0)}{1-O(N\varepsilon_0)}$ for all $j \in \mathcal{I}$ and that $2s_q + \sum_{j\in\mathcal{I}} a_j = 1$. As a result, there exists some constant $R_0 > 0$ such that

$$\frac{1}{2} \ge s_q \ge \frac{1}{2+(L'-2)\Gamma} = \frac{1-O(N\varepsilon_0)}{2+(L'-2N-2)O(\varepsilon_0)} \implies |2s_q - 1| \le R_0 L'\varepsilon_0 \tag{74}$$

and

$$\sum_{j\in\mathcal{I}} a_j = 1 - 2s_q \le R_0 L'\varepsilon_0 \tag{75}$$

Moreover, due to right-clipped convolution, we have $\|\boldsymbol{z}_{L'-1}\|_{\ell_2} = 2\|\sum_{i=0}^{W} F_i\|_{\ell_1} \le 2\Xi$. Next, according to the value retrieval at i-th position, we have

$$\|\boldsymbol{z}_i - 2\boldsymbol{v}\|_{\ell_2} \le |2F_{-1} - 2|\|\boldsymbol{v}\|_{\ell_2} + 2|\sum_{j\ne -1} F_j| \le 2\Xi \tag{76}$$

Utilizing these findings above, we get

$$\begin{aligned}
\|f(\boldsymbol{X}) - \boldsymbol{v}\|_{\ell_2} &\le \|\sum_{j\in\mathcal{I}} a_j \boldsymbol{z}_j\|_{\ell_2} + \|s_q(\boldsymbol{z}_i + \boldsymbol{z}_{L'-1}) - 2s_q\boldsymbol{v}\|_{\ell_2} + |2s_q - 1|\|\boldsymbol{v}\|_{\ell_2} && (77)\\
&\le |\sum_{j\in\mathcal{I}} a_j| \max_j \|\boldsymbol{z}_j\|_{\ell_2} + s_q(\|\boldsymbol{z}_i - 2\boldsymbol{v}\|_{\ell_2} + \|\boldsymbol{z}_{L'-1}\|_{\ell_2}) + |2s_q - 1| && (78)\\
&\le 2R_0 L'\varepsilon_0(1+\Xi) + 2\Xi + R_0 L'\varepsilon_0 && (79)\\
&\le 3R_0 L'\varepsilon_0 + 2\Xi + 2R_0\Xi L'\varepsilon_0 && (80)\\
&\le 3\varepsilon_0(R_0 L' + 2RL(1+N\varepsilon_0)(1+R_0 L'\varepsilon_0)) && (81)
\end{aligned}$$

Let c_0, c_1 be absolute constants to be determined. Assuming $N\varepsilon_0 \le O(1)$ (i.e. bounded by constant), we have that

$$\|f(\boldsymbol{X}) - \boldsymbol{v}\|_{\ell_2} \le c_0\varepsilon_0(L' + L + LL'\varepsilon_0)$$

where $c_0 \ge 3\max\{R_0, 2R(1+N\varepsilon_0), 2R_0R(1+N\varepsilon_0)\}$. Assuming the stronger bound $L\varepsilon_0 \le O(1)$ and $c_1 \ge c_0(1 + L/L' + L\varepsilon_0)$, we have that

$$\|f(\boldsymbol{X}) - \boldsymbol{v}\|_{\ell_2} \le c_1\varepsilon_0 L'$$

This concludes the advertised results. □

E.4. Proof of Theorem 4

Below we state a generalization of Theorem 4 which distinguishes two scenarios: Short convolution with PE and Long convolutions with no PE.

Theorem 4 (Selective Copy)**.** *Consider the setting of Def. 4. There is a 1-layer CAT f using exponential-decay query-convolution $\boldsymbol{F}_q$ as follows:*

- *Suppose $\boldsymbol{F}_q$ is infinitely long (namely parameterized as an SSM with state matrix $\boldsymbol{A} = \rho$ for some decay parameter $\rho < 1$). Then, using $d = |\mathcal{S}| + 3$ dimensional token embeddings, f solves unique selective copying.*
- *Suppose $\boldsymbol{F}_q \in \mathbb{R}^N$ and input sequences contain at most N signal tokens. Using $d = |\mathcal{S}| + 4$ and 1-D positional encodings, f solves unique selective copying.*

Proof. Let T be the maximum context length the model encounters. Specifically, $T = L + N + 1$ where L is the maximum length of the input sequence X that precedes the special token $\perp$ and N is the maximum number of signal tokens in X. Recall that the cardinality of the signal vocabulary $\mathcal{S}$ is allowed to be larger than N and we resume generation until outputting all signal tokens. Let $Z = [X \perp z_{L+2} \ \dots \ z_t]$ denote the current input sequence where $[X \perp]$ is the initial input that kickstarts decoding. Denote boldface $\boldsymbol{Z}, \boldsymbol{X}$ to be the embedded sequences of Z, X. We use query convolution thus the CAT model is given by $f(\boldsymbol{Z}) = \text{nearest\_token\_embedding}(\boldsymbol{Z}^\top \mathbb{S}(\boldsymbol{Z}\boldsymbol{W}\boldsymbol{z}_t^*))$ where $\boldsymbol{Z}^* = \boldsymbol{F}_q * \boldsymbol{Z}$ is the convolved sequence and $\boldsymbol{z}_t^*$ is the last token of the convolved query sequence for $L + 1 \le t \le T$. We set convolution to be $F_{q,i} = \rho^i$ for $0 \le i < W$ for a suitable $\rho \le 1$ to be determined where W is the window size of the convolution. This choice aggregates the current token and the $W - 1$ most recent tokens and allows for all all-ones filter as a special case. For the first statement of the theorem $W = \infty$ whereas for the second statement $W = N$.

The choice of token embeddings. We construct the token embeddings as follows:

- Token embedding of the ith token has the form $\boldsymbol{x}_i = [\boldsymbol{x}_i', s_i, p_i]$. Here
  - **Base embedding.** $\boldsymbol{x}_i'$ is the *base embedding* vector associated to the discrete token value x_i. We choose these $\boldsymbol{x}_i'$ embeddings to have unit Euclidean norm.
  - **Signal indicator.** $s_i \in \mathbb{R}$ is an indicator of whether the token is a signal token or not. We set $s_i = 1$ for signal tokens and the $\perp$ token and $s_i = 0$ for noise tokens.
  - **Position encoding.** $p_i \in \mathbb{R}$ is the positional encoding of the i'th token. We simply set $p_i = i/T$ where $T = L+N+1$. p_i is only required for short convolution i.e. when $W = N$.
- The base embeddings of noise tokens $\mathcal{N}$ are orthogonal to that of signal tokens and $\perp$ token.
- The base embeddings of signal tokens $\mathcal{S}$ and $\perp$ are also orthogonal to each other.

Let D_{noisy} be the dimension of the subspace spanned by the base embeddings of noise tokens. We can choose $D_{\text{noisy}} = 1$ by setting all base embeddings for the noise tokens to be identical. The signal tokens and $\perp$ token occupies an orthogonal $|\mathcal{S}| + 1$ dimensional subspace. Together, this recovers the embedding dimensions advertised in the theorem statement, namely

- **With positional encoding and $W = N$:** We need an embedding dimension of $d = |\mathcal{S}| + D_{\text{noisy}} + 3 \ge |\mathcal{S}| + 4$ where two additional dimension is due to s_i and p_i.
- **Without positional encoding and $W = \infty$:** We need an embedding dimension of $d = |\mathcal{S}| + D_{\text{noisy}} + 2 \ge |\mathcal{S}| + 3$ where the additional dimension is due to s_i.

• **Construction of the CAT model.** We construct a one layer CAT model with the following criteria in the order of priority:

1. The model should always select signal tokens.
2. The model should select a signal token not previously selected.
3. The model should select the farthest signal token from the current/last token (i.e. generates signal tokens that are closer to the start of the sequence).

To satisfy the three criteria above, we pick the attention weights $\boldsymbol{W}$ as follows when $W = N$:

$$\boldsymbol{W} = \begin{bmatrix} -\alpha \boldsymbol{I}_{N+1+D_{\text{noisy}}} & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & -\theta \end{bmatrix}. \tag{82}$$

The choice for $W = \infty$ is same except that we do not have the positional encoding coefficient θ. Recall that we also choose the convolutional filter as $F_{q,i} = \rho^i$ for $0 \le i < W$ for $\rho < 1$. Specifically, we choose $\rho = 2^{-1/T}$ so that $\rho^T = 1/2$. This choice guarantees that $\rho^i - \rho^{i+1} \ge c/T$ for all $0 \le i < T$ for some absolute constant $c > 0$.

We will accomplish the proof inductively. Suppose that X contains N' unique signal tokens and that until time t for some $L + 1 \le t \le L + N' + 1$, the model outputs the correct sequence of $t' = t - L - 1$ unique signal tokens. We will prove that it

will accurately output the next signal token in line with suitable choice of α, β, θ. To this end, we state the following lemma regarding the output of the query-convolution z_t^*.

Note that $z_t^* = \sum_{i=1}^{t} \rho^i z_{t-i}$. Recall that z_{L+1} to z_t are unique correctly-ordered signal tokens where we set $z_0 = \perp$. Denote the rest of the $N' - t'$ signal tokens with correct order by q_1 to $q_{N'-t'}$. Here q_1 is the left most such token in X and the token we wish to output next. We can write z_t^* in terms of signal tokens and noise tokens as follows:

$$z_t^* = \sum_{i=1}^{t'+1} b_i z_{t-i} + \boldsymbol{n} + \sum_{j=1}^{N'-t'} a_j \boldsymbol{q}_j, \tag{83}$$

where we set $b_i := \rho^i$. Here the first term $\sum_{i=1}^{t'+1} b_i z_{t-i}$ is due to last $t'+1$ tokens (including $\perp$) that are already generated. The $\boldsymbol{n}$ term denotes the aggregated contribution of the noise tokens to the convolution. $\sum_{j=1}^{N'-t'} a_j \boldsymbol{q}_j$ is the contributions of the signal tokens that are yet to be generated. Crucially note that,

- If $W = \infty$, a_i is strictly increasing because convolution coefficients $F_{q,j}$ are strictly decreasing (with a gap lower bounded by c/T).
- Whether $W = N$ or $W = \infty$, $b_i = \rho^i$ is strictly decreasing and $b_{t'+1} \geq a_{N'-t'} + c/T$. That is, the contribution of any token already generated is larger than any token that is yet to be generated.

Let us write $z_t^* = [z_t'^*\ s\ p]$. Note that $s \geq 1/2$ because $\perp$ token is involved in the convolution and $\rho^T = 1/2$. Similarly, if we employ PE, we have that $p \geq (L+1)/2T \geq 1/4$ for the same reason. Given a token $\boldsymbol{x}_i = [\boldsymbol{x}_i'\ s_i\ p_i]$, through (82), we have that

$$\text{score}_i = \boldsymbol{x}_i^\top \boldsymbol{W} z_t^* = -\alpha \langle z_t'^*, \boldsymbol{x}_i' \rangle + \beta s s_i - \theta p p_i. \tag{84}$$

We now proceed with the proof which relies on choosing $\alpha, \beta, \theta > 0$ in a suitable fashion. Specifically, we will choose their relative ratios $\beta/\alpha, \alpha/\theta$ suitably to ensure the desired token q_1 receives the highest score. After ensuring this, we can suitably scale up α, β, θ in a proportional fashion, which will also scale up the scores of each token. Thanks to softmax attention, this will ensure that the model precisely retrieves the token with the highest score.

Scenario 1: $W = \infty$**.** In this scenario, we don't use PE, thus, effectively $\theta = 0$. We need to satisfy aforementioned criteria: First, we want the highest score to be a signal token. We will guarantee this by observing $s_i = 0$ for noise tokens, $s > 0$ and by setting $\beta/\alpha \gg 1$. Second, we want the highest score to be q_1, the left most signal token that has not been output yet. Now, since $W = \infty$, q_1 receives the lowest coefficient of a_1 in (83). Using orthogonality and unit Euclidean norm, this implies that $\left\langle z_t'^*, \boldsymbol{q}_1' \right\rangle = a_1$. In contrast, any other signal token has a larger inner product by at least c/T. Choosing $\alpha = 1$ (and then suitably scaling it up together with β), this implies that, q_1 is indeed the token with the highest score that will be generated next.

Scenario 2: $W = N$ **and we employ PE.** We again follow the score decomposition (84). Observe that $\left\langle z_t'^*, \boldsymbol{x}_i' \right\rangle, ss_i, pp_i$ are all bounded by 1 in absolute value. Thus, by controlling the relative ratios of the scalar knobs $\beta > \alpha > \theta = 1$, we can enforce the three criteria listed above. Recall that q_1 denotes the next signal token we wish to output next. We will prove that q_1 achieves the strictly highest score amoung the tokens of Z. To proceed, set $\beta/\alpha \gg 1$ and $\alpha/\theta \gg 1$.

- Since β dominates α and θ, following the same argument in Scenario 1, noise tokens will have strictly lower scores than signal tokens, thus cannot be generated next.
- Following (84), the signal tokens have a score contribution of $-\alpha \cdot b_i$ or $-\alpha \cdot a_j$ from the inner product term $\left\langle z_t'^*, \boldsymbol{x}_i' \right\rangle$. Here b_i denoted the coefficient of a generated signal token whereas a_j denoted the coefficient of a missing signal token. Next recall from (83) that $b_i \geq c/T > 0$ and $b_i \geq a_j + c/T$ thanks to the $\boldsymbol{F}_q$ choice. Since α dominates θ, this implies that the generated signal tokens have strictly less score than the missing signal tokens.
- Finally, we wish to show that q_1 has the highest score among missing signal tokens. First, recall from (83) that a_1 is the smallest coefficient among the missing signal tokens. As a result, it achieves the largest inner product score $-\alpha \cdot \left\langle z_t'^*, \boldsymbol{x}_i' \right\rangle$. To complete the proof, we use positional encoding to break any score ties. Since q_1 is the left most missing signal token, any other missing signal token will achieve a strictly worse position encoding score $-\theta p p_i$ as $p \geq 1/4$, $\theta = 1$, and $p_i = i/T$ is strictly increasing. This guarantees that q_1 achieves the strictly highest score as desired.

To summarize, by choosing suitable $\beta \gg \alpha \gg \theta = 1$ and proportionally scaling up α, β, θ sufficiently, we conclude with the proof. □

F. Proofs for Section B – Convolution-Attention Tradeoff

F.1. Proof of Proposition 1

The key to the proof is establishing the detection threshold of the correct block ID β in (LCAT) i.e. we wish to guarantee $b = \beta$. Once correct block is retrieved, the rest of the argument is identical to AR over dense attention as we retrieve the correct blocks. Observe that, we have $\bar{L} - 1$ blocks in total (not counting the local/final block). Note that $\boldsymbol{z}_i \sim \mathcal{N}(0, \sigma^2 B \boldsymbol{I}_d/d)$ for $i \neq \beta$ and $\boldsymbol{z}_\beta \sim \mathcal{N}(\boldsymbol{x}_L, \sigma^2(B-1)\boldsymbol{I}_d/d)$.

Set $g_i = \boldsymbol{z}_i^\top \boldsymbol{x}_L \cdot \sqrt{d/B}$ for $i < \bar{L}$ and g_β. Observe that g_i's and g_β are independent random variables. Additionally, $g_{i\neq\beta} \sim \mathcal{N}(0, \sigma^2)$, $g_\beta \sim \mathcal{N}(\sqrt{d/B}, (1-1/B)\sigma^2)$. Let $g_{\max} = \max_{i\neq\beta} g_i$. We have the following gaussian concentration inequalities

$$\mathbb{P}(g_{\max} \geq \sigma(\sqrt{2\log L'} + t)) \leq e^{-t^2/2} \tag{85}$$

$$\mathbb{P}(g_\beta \leq \sqrt{d/B} - \sigma t) \leq e^{-t^2/2}. \tag{86}$$

Combining these three, we find that, with probability $1 - 2e^{-t^2/2}$, whenever $\sqrt{d/B} \geq \sigma(\sqrt{2\log L'} + 2t)$, we have that

$$g_\beta > g_{\max} = \max_{i\neq b} g_i.$$

This condition is implied by $d \geq \sigma^2 B(\sqrt{2\log \bar{L}} + 2t)^2$. Applying change of variable on t, we conclude with the result.

Retrieving the value token. Once the correct block is identified, (query, value) pair is retrieved by applying full softmax attention with $\boldsymbol{W} = c\boldsymbol{I}$ with $c \to \infty$ within the selected two blocks. Recall that local attention retrieves the query and the choice of convolutional filter will return the value ahead of the query. To guarantee this, we only need to prove that $\boldsymbol{x}_L$ also has the largest correlation to itself within the two selected blocks we apply local attention. To this aim, we similarly use the fact that correlations between $\boldsymbol{x}_L$ and the other tokens in the selected blocks are IID $\mathcal{N}(0, \sigma^2/d)$ variables. There are at most $2B-2$ such other tokens. Consequently, the maximum local correlation $g^{\text{loc}}_{\max}$ obeys $\mathbb{P}(g^{\text{loc}}_{\max} \geq \sigma(\sqrt{2\log(2B)}+t)/\sqrt{d}) \leq e^{-t^2/2}$. We wish to guarantee that $g^{\text{loc}}_{\max} < 1$. This holds with $1 - e^{-t^2/2}$ probability whenever $d \geq \sigma^2(\sqrt{2\log(2B)} + t)^2$. This latter condition is implied by the original condition because $\sqrt{B}(\sqrt{2\log \bar{L}} + 2t) \geq \sqrt{2B\log 2} + 2t \geq \sqrt{2\log(2B)} + t$. Union bounding, we end up with a success probability of at least $1 - 3e^{-t^2/4}$.

Next, we wish to show the converse result. We recall that as the expectation of supremum of K IID $\mathcal{N}(0,1)$ become $(1 \pm o(1))\sqrt{2\log K}$ as K grows to infinity. Thus, for sufficiently large $\bar{L} \geq C_\varepsilon$, we have that $\mathbb{E}[g_{\max}] \geq \sqrt{(2-\varepsilon)\log \bar{L}}$. Consequently, we can write the reversed inequalities

$$\mathbb{P}(g_{\max} \leq \sigma(\sqrt{(2-\varepsilon)\log \bar{L}} - t)) \leq e^{-t^2/2} \tag{87}$$

$$\mathbb{P}(g_\beta \geq \sqrt{d/B} + \sigma t) \leq e^{-t^2/2}. \tag{88}$$

Combining these, we conclude with the advertised reverse inequality. As a result, with the same probability, we fail to identify the block containing the target query/value pair.

The next subsection proves the uniform AR guarantees via an application of Slepian's lemma.

F.2. Proof of Uniform Associative Recall via Slepian's Lemma (Proposition 1 continued)

Slepian's Lemma (Slepian, 1962) is an important gaussian comparison inequality. A specific variation is the following result that holds for a random gaussian matrix. We first introduce the Gaussian width definition that is important for assessing the complexity of a geometric set in a high-dimensional space.

Definition 7 (Gaussian width)**.** *Let $S \in \mathbb{R}^d$ and $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Gaussian width of S is defined as $\omega(S) = \mathbb{E}[\sup_{\boldsymbol{x}\in S} \boldsymbol{x}^\top \boldsymbol{g}]$*

Proposition 4 (Slepian's Lemma)**.** *Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be a matrix with IID $\mathcal{N}(0,1)$ entries. Let $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_n)$ and $h \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Given sets $A \in \mathbb{R}^n, B \in \mathbb{R}^d$, we have that*

$$\mathbb{E}[sup_{\boldsymbol{a}\in A, \boldsymbol{b}\in B} \boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{b}] \leq \mathbb{E}[sup_{\boldsymbol{a}\in A, \boldsymbol{b}\in B} \boldsymbol{a}^\top \boldsymbol{g}\|\boldsymbol{b}\|_{\ell_2} + \boldsymbol{b}^\top h\|\boldsymbol{a}\|_{\ell_2}].$$

We have the following application of Slepian's lemma.

Lemma 4. *Let $\boldsymbol{X} \in \mathbb{R}^{L\times d}$ be a matrix with IID $\mathcal{N}(0,1)$ entries. Let $g \sim \mathcal{N}(0, \boldsymbol{I}_d)$ be an independent vector. Fix a subset of unit sphere $S \in \mathbb{R}^d$. With probability $1 - e^{-t^2/2}$, we have that*

$$\sup_{\boldsymbol{\beta}\in S} \|\boldsymbol{X}\boldsymbol{\beta}\|_{\ell_\infty} \le \sqrt{2}(\sqrt{\log L} + \omega(S) + t).$$

Proof. Define the augmented matrix $\boldsymbol{X}' = \begin{bmatrix} \boldsymbol{X} \\ -\boldsymbol{g} \end{bmatrix}$. Define the set $A \in \mathbb{R}^{L+1}$ such that $\boldsymbol{a} \in A$ has the following form. $\boldsymbol{a}$ has two nonzero entries both of which are equal to 1. Additionally, last entry is nonzero i.e. $\boldsymbol{a}_{L+1} = 1$. Using $\|\boldsymbol{a}\|_{\ell_2} = \sqrt{2}$ and $\|\boldsymbol{\beta}\|_{\ell_2} = 1$, we now apply Slepian's lemma as follows

$$\mathbb{E}[sup_{\boldsymbol{a}\in A,\boldsymbol{\beta}\in S} \boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{\beta}] \le \mathbb{E}[sup_{\boldsymbol{a}\in A,\boldsymbol{b}\in B} \boldsymbol{a}^\top \boldsymbol{g}\|\boldsymbol{\beta}\|_{\ell_2} + \boldsymbol{\beta}^\top h\|\boldsymbol{a}\|_{\ell_2}] \tag{89}$$

$$\le \mathbb{E}[\|\boldsymbol{g}\|_{\ell_\infty} + \sqrt{2}\sup_{\boldsymbol{\beta}\in S} \boldsymbol{\beta}^\top h] \tag{90}$$

$$\le \sqrt{2}(\sqrt{\log L} + \omega(S)). \tag{91}$$

To proceed, observe that $\boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{\beta}$ is a $\sqrt{2}$-Lipschitz function of $\boldsymbol{X}$. This implies the statement of the lemma. □

Proposition 5. *Consider the setting of Proposition 1. Suppose we wish to solve AR for the worst query drawn from a set S which is subset of the unit sphere. If $d \ge 2\sigma^2 B(\sqrt{\log \bar{L}} + \omega(S) + t)^2$, (LCAT) solves AR with probability at least $1 - 2e^{-t^2/2}$ for all $\boldsymbol{x}_L \in S$.*

Proof. The proof follows the steps of Section F.1 with the following distinction. Note that, to determine the correct block, we now need to do a worst case analysis. Namely, let $z'_\beta = \boldsymbol{x}_L - z_\beta$, $z'_i = z_i$ for $i \neq \beta$, and set $\boldsymbol{Z}' = \begin{bmatrix} z'_1 \\ \vdots \\ z_{\bar{L}-1} \end{bmatrix}$. Also let $\boldsymbol{Z} = [z_1 \ \dots \ z_{\beta-1} \ z_{\beta+1} \ \dots \ z_{\bar{L}-1}]$. The accurate detection of the block β coincides with the following event

$$\inf_{\boldsymbol{x}_L\in S} \|\boldsymbol{Z}\boldsymbol{x}_L\|_{\ell_\infty} - z_\beta^\top \boldsymbol{x}_L > 0.$$

Using $z^\top \boldsymbol{x}_L = 1 - z'^\top \boldsymbol{x}_L$ and defining the set A to be the set of all vectors with exactly two 1s with one of the 1 appearing at position β, the above event can alternatively be written as

$$\sup_{\boldsymbol{a}\in A,\boldsymbol{x}_L\in S} \boldsymbol{a}^\top \boldsymbol{Z}'\boldsymbol{x}_L < 1.$$

Now applying Lemma 4 on the left hand side, we find that, with probability $1 - e^{-t^2/4}$,

$$\sup_{\boldsymbol{a}\in A,\boldsymbol{x}_L\in S} \boldsymbol{a}^\top \boldsymbol{Z}'\boldsymbol{x}_L \le \sqrt{\frac{2B}{d}}\sigma(\sqrt{\log L} + \omega(S) + t)$$

Consequently, whenever $d > 2\sigma^2 B(\log L + \omega(S) + t)^2$, we conclude with the result. Note that, when S is an r-dimensional subspace, we plug in the well-known bound $\omega(S) \le \sqrt{r}$. Finally, we need to union bound this event with the event that the query token can be identified through local attention by letting $\boldsymbol{W}_k = \boldsymbol{W}_q = \sqrt{c}\boldsymbol{I}$ and $c \to \infty$. To do so, we apply Lemma 4 over the $2B-2$ non-query tokens. Denoting these tokens by $\boldsymbol{X}^{loc} \in \mathbb{R}^{d\times(2B-2)}$, we have that $\mathbb{P}(\boldsymbol{X}^{loc}\boldsymbol{q} \ge \sqrt{2\sigma^2/d}\cdot(\sqrt{\log(2B)} + \omega(S) + t)) \le e^{-t^2/2}$. Consequently, $\boldsymbol{X}^{loc}\boldsymbol{q} < \boldsymbol{q}^\top\boldsymbol{q} = 1$ as soon as the same condition $d \ge 2\sigma^2 B(\sqrt{\log \bar{L}} + \omega(S) + t)^2$ holds. This introduces an additional $e^{-t^2/4}$ probability of failure. □

F.3. Proof of Proposition 2

We essentially follow the proof of Proposition (1). The only differences are that, the variance calculations, comparison of block correlations, and signal-to-noise ratio bounds will all slightly change due to exponential smoothing impacting the full context window. To proceed, let us observe the following scenarios for a block ID $1 \le i < \bar{L}$:

- **Scenario 1:** $i < \beta$**.** In this scenario, z_i is exponentially-smoothed sum of IID vectors with $\mathcal{N}(0,1)$ entries. Recalling $\rho = e^{-1/B}$, the variance σ_z^2 of entries of z_i is upper bounded by

$$\sigma_z^2 = \sum_{i=0}^{\infty} \rho^{2i} = \frac{\sigma^2}{1-\rho^2} \le 1.2B. \tag{92}$$

  Here, we used the fact that for $B = 1$, the bound holds and for $B \ge 2$, we have that $\rho^2 = e^{-2/B} \le 1 - \frac{1}{B}$. The latter implies $1 - \rho^2 \ge 1/B$ and $1/(1-\rho^2) \le B$.

  The above bound on σ_z^2 implies that, setting $g_i = z_i^\top x_L \cdot \sqrt{d/B}$, we have that $g_i \sim \mathcal{N}(0, \sigma_i^2)$ with $\sigma_i^2 \le 1.2\sigma^2$.

- **Scenario 2:** $i = \beta$**.** In this scenario, the variance upper bound σ_i^2 above is still applicable. The key is to estimate and lower bound the mean component similar to the proof of Proposition (1). Let the query token appear in the kth position of block β for $k \in [B]$. Define $p = (k-1)/B$. Observe that

$$\mathbb{E}[g_\beta] = \mathbb{E}[z_\beta^\top x_L \cdot \sqrt{d/B}] = e^{-p}\sqrt{d/B}.$$

- **Scenario 3:** $i > \beta$**.** This is essentially same as Scenario 2, because, thanks to the exponential smoothing, the signal token from block β will propagate to future z_i's. The coefficient of the propagation satisfies

$$\mathbb{E}[g_i] = \mathbb{E}[z_i^\top x_L \cdot \sqrt{d/B}] = e^{-(p+i-\beta)}\sqrt{d/B}.$$

Now that we have gathered these three scenarios, we can define $g_{\max} = \max_{i \ne \beta} g_i - \mathbb{E}[g_i])$ similar to above. $g_{\max}$ is a supremum of independent Gaussians of bounded variance controlled by (92). Through this, we have that

$$\mathbb{P}(g_{\max} \ge 1.6\sigma(\sqrt{\log \bar{L}} + t)) \le e^{-t^2} \tag{93}$$

$$\mathbb{P}(g_\beta - \mathbb{E}[g_\beta] \le 1.6\sigma t) \le e^{-t^2}. \tag{94}$$

Secondly, for $i \ne \beta$, using $p \le 1$, we have that

$$\mathbb{E}[g_\beta] - \mathbb{E}[g_i] \ge (e^{-p} - e^{-(p+i-\beta)})\sqrt{d/B} \ge (e^{-1} - e^{-2})\sqrt{d/B} \ge 0.23\sqrt{d/B}.$$

Consequently, we require $0.23\sqrt{d/B} > 1.6\sigma(\sqrt{\log \bar{L}} + 2t)$. Using $\tau = t/2$, this is guaranteed by $d \ge 50B\sigma^2(\sqrt{\log \bar{L}} + \tau)^2$ with probability at least $1 - 2e^{-\tau^2/4}$. Once the correct block is identified, (query, value) pair is retrieved by applying dense softmax attention with $W = cI$ with $c \to \infty$ over the selected blocks thanks to the choice of convolutional filter. This argument is identical to "Retrieving the value token" proof in Section F.1 and introduces an additional $e^{-\tau^2/2}$ probability of failure in the union bound.