BREAKING THE REVERSAL CURSE IN AUTOREGRESSIVE LANGUAGE MODELS VIA IDENTITY BRIDGE

Xutao Ma* **Yixiao Huang*** **Hanlin Zhu** **Somayeh Sojoudi**
UC Berkeley
{xutao_ma,yixiaoh,hanlinzhu,sojoudi}@berkeley.edu

ABSTRACT

Autoregressive large language models (LLMs) have achieved remarkable success in many complex tasks, yet they can still fail in very simple logical reasoning such as the "reversal curse" — when trained on forward knowledge data of the form "$A \to B$" (e.g., *Alice's husband is Bob*), the model is unable to deduce the reversal knowledge "$B \leftarrow A$" (e.g., *Bob's wife is Alice*) during test. Extensive prior research suggests that this failure is an inherent, fundamental limit of autoregressive causal LLMs, indicating that these models tend to memorize factual-level knowledge rather than capture higher-level rules. In this paper, we challenge this view by showing that this seemingly fundamental limit can be mitigated by slightly tweaking the training data with a simple regularization data recipe called the Identity Bridge of the form "$A \to A$" (e.g., *The name of Alice is Alice*). Theoretically, we prove that under this recipe, even a one-layer transformer can break the reversal curse by analyzing the implicit bias of gradient descent. Empirically, we show that a 1B pretrained language model finetuned with the proposed data recipe achieves a 40% success rate on reversal tasks, in stark contrast to a near-zero success rate when trained solely on forward-knowledge data. Our work provides a novel theoretical foundation for the reversal curse and offers a principled, low-cost path to encouraging LLMs to learn higher-level rules from data.

1 INTRODUCTION

Autoregressive large language models (LLMs) have demonstrated great capability in solving various complex tasks (Jaech et al., 2024; Guo et al., 2025a). However, they still struggle with some simple logical reasoning tasks, and one of the most well-known failure modes is the "reversal curse", which refers to the phenomenon that when LLMs have learned a forward relation "$A \to B$" (e.g., *Alice's husband is Bob.*) during training, they fail to answer the reverse question "$B \leftarrow A$" (e.g., *Who is Bob's wife?*) during test.

Extensive prior works have attempted to understand or resolve the reversal curse. Recent research (Zhu et al., 2024; Lin et al., 2024; Wang & Sun, 2025) suggests that this forward–reverse asymmetry is a fundamental generalization failure of autoregressive LLMs. Consequently, existing mitigation strategies generally follow two paths: (i) augmenting or reformatting training examples to explicitly show the reverse links, and (ii) adjusting the learning objective or training procedure to reduce directional bias. However, such interventions often result in substantial deviation from standard data pipelines or training recipes and can introduce trade-offs in overall model quality. Moreover, they commonly adopt the premise that, for causal autoregressive models, reliably answering the reversed queries ultimately requires that the specific direction be included in training.

In this paper, we challenge this view by showing that the reversal curse can be resolved by adding certain regularization to the training data, without being trained on reversal data. Specifically, we add the regularization by augmenting the training data with the "Identity Bridge", which was originally proposed by Lin et al. (2025) to solve two-hop reasoning tasks. The identity bridge refers to statements in the form of "$A \to A$" (e.g., *The name of Alice is Alice*, or *The wife of Alice's husband is Alice*). Semantically, it adds no new information to the dataset about the relations between entities, but it

*Equal contributions.

serves as a dataset-level regularization and can change the optimization landscape. In theory, we prove through the implicit bias of gradient descent that by adding the identity bridge regularization, even one-layer transformers can break the reversal curve on symbolic reasoning tasks. To further understand the mechanism of the identity bridge, we prove that the identity bridge regularized reversal task can be equivalently formulated as an out-of-context reasoning (OCR) problem (Cohen et al., 2024). Moreover, to validate our method in real-world scenarios, we conduct experiments with pretrained LLMs on real-world reversal tasks, and the trained model can achieve a pass rate as high as 40%, which is a significant improvement over the previous near-zero accuracy.

In summary, the main contributions of this paper include:

- We propose to use the identity bridge regularized data recipe to break the reversal curse in autoregressive LLMs. To the best of our knowledge, this is the first work that breaks the reversal curse in autoregressive LLMs without modifying the training paradigm (e.g., the model architecture or loss function) or reversing the training data.
- Theoretically, we prove through the implicit bias of gradient descent that even a one-layer transformer is able to break the reversal curse via identity bridge regularization, while without it, the reversal curse happens. We also prove that the identity bridge regularization is closely related to the OCR phenomenon.
- We conduct experiments on pretrained LLMs that achieve a 40% pass rate on real-world reversal tasks via the identity bridge regularization, a significant improvement over the previous near-zero accuracy.

2 PRELIMINARIES

Basic notations. Let $[N] = \{1, \cdots, N\}$. We denote $\boldsymbol{e}_i$ as one-hot vectors where only the i-th entry is non-zero and equals one, and denote $\boldsymbol{z}_x$ as the embedding for token x. Let $\mathcal{V}$ be the vocabulary set. We use $\mathbf{0}_{m\times n}$ to denote the $m \times n$ all-zero matrix and use $\mathbf{0}_d$ to denote the d-dimensional zero vector. $\mathcal{A}$ and $\mathcal{B}$ denote sets of entities and $\mathcal{R}$ denotes the set of relations.

Reversal reasoning task. Given two disjoint sets of N entities $\mathcal{A} := \{a_1, \cdots, a_N\}$ and $\mathcal{B} := \{b_1, \cdots, b_N\}$, a forward relation r_+ is a bijection mapping a_i to b_i for all $i \in [N]$, and the reverse relation r_- is defined as the inverse of r_+. The reversal reasoning task is to answer $r_-(b_i), i \in [N]$ when a model is only trained with the forward relations $r_+(a_i) = b_i,\ \forall i \in [N]$.

Dataset. To adjust to the language model input format, we represent each relational instance $r(s) = s'$ as a token sequence $[s, r|s']$, where s, s' are entities and r is a relation. We treat $[s, r]$ as the input sequence and s' as the target label. In this paper, we consider the following sets:

- Forward relation set: $\mathcal{D}_{r_+} = \{[a_i, r_+|b_i] : i \in [N]\}$;
- Reversal relation set: $\mathcal{D}_{r_-} = \{[b_i, r_-|a_i] : i \in [N]\}$;
- Identity bridge set: $\mathcal{D}_{\text{idn}} = \{[a_i, r_{\text{id}}|a_i] : i \in [N]\} \cup \{[b_i, r_{\text{id}}|b_i] : i \in [N]\}$;
- Relation set: $\mathcal{R} = \{r_+, r_-, r_{\text{id}}\}$.

The identity bridge dataset $\mathcal{D}_{\text{idn}}$, originally proposed by Lin et al. (2025) for two-hop reasoning tasks, contains an entity and an identity relation r_{id} as input, and maps them to the entity itself. The identity bridge dataset actually contains no information, but will serve as a regularization to help break the reversal curse.

One-layer transformer. We consider a one-layer decoder-only transformer TF, which takes a sequence $x_{1:T} := (x_1, \ldots, x_T) \in \mathcal{V}^{\mathrm{T}}$ as input and outputs a logit vector $\mathrm{TF}_{\boldsymbol{\theta}}(x_{1:T}) \in \mathbb{R}^d$ as follows:

$$\mathrm{TF}_{\boldsymbol{\theta}}(x_{1:T}) := \boldsymbol{W}_{\mathrm{O}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{T}}\boldsymbol{X}\mathrm{softmax}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}_{\mathrm{KQ}}\boldsymbol{x}_T), \tag{1}$$

where $\boldsymbol{X} = [\boldsymbol{z}_{x_1}, \cdots, \boldsymbol{z}_{x_T}] \in \mathbb{R}^{d\times T}$ is the embedding matrix of the tokenized sequence, $\boldsymbol{W}_{\mathrm{O}}, \boldsymbol{W}_{\mathrm{V}} \in \mathbb{R}^{d\times d_h}$ are the output and value matrices, respectively, $\boldsymbol{W}_{\mathrm{KQ}} = \boldsymbol{W}_{\mathrm{K}}\boldsymbol{W}_{\mathrm{Q}}^{\mathrm{T}} \in \mathbb{R}^{d\times d}$ is the reparameterized key-query matrix, and $\boldsymbol{\theta} = (\boldsymbol{W}_{\mathrm{O}}, \boldsymbol{W}_{\mathrm{V}}, \boldsymbol{W}_{\mathrm{KQ}})$ encodes all model parameters.

We further define the logit of token y as

$$\mathrm{TF}_{\boldsymbol{\theta}}(x_{1:T}; y) := \boldsymbol{z}_y^{\mathrm{T}} \mathrm{TF}_{\boldsymbol{\theta}}(x_{1:T}). \tag{2}$$

The next token probability $p(v|x_{1:T})$ is computed as the softmax of the logit vector:

$$p_{\boldsymbol{\theta}}(v|x_{1:T}) = \frac{\exp(\mathrm{TF}_{\boldsymbol{\theta}}(x_{1:T}; v))}{\sum_{v' \in \mathcal{V}} \exp(\mathrm{TF}_{\boldsymbol{\theta}}(x_{1:T}; v'))}. \tag{3}$$

Loss function. We use the cross-entropy loss

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{[x_{1:T}|y] \sim \mathcal{D}}[-\log p_{\boldsymbol{\theta}}(y|x_{1:T})], \tag{4}$$

where $\mathcal{D}$ is a dataset. The model parameter will be optimized by running the gradient flow $\dot{\boldsymbol{\theta}} = -\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta})$ or gradient descent with sufficiently small learning rate.

3 Theoretical Results

In this section, we study the reversal curse on a one-layer transformer theoretically.

We first provide an explanation for the reversal curse in Section 3.1. A recent paper (Zhu et al., 2024) also explains the reversal curse on a one-layer transformer by examining the learning dynamics, but their transformer model uses a non-factorized output and value matrix, *i.e.*, a single reparameterized $\boldsymbol{W}_{\mathrm{OV}}$ instead of the factorized form $\boldsymbol{W}_{\mathrm{O}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{T}}$. As a complement to this result, we analyze the case for a factorized transformer as Moreover, our analysis is based on the implicit bias of gradient descent, providing a different perspective to understand this phenomenon. Then, in Section 3.2, we propose a simple data recipe called the Identity Bridge and prove that it can help break the reversal curse in one-layer transformers. Finally, we provide insight in Section 3.3 to understand the mechanism of the identity bridge by relating to the out-of-context reasoning phenomenon (Huang et al., 2025a).

We take the embedding $\boldsymbol{z}_s$ for all entities $s \in \mathcal{A} \cup \mathcal{B}$ to be one-hot, and set the embedding of the forward relation $\boldsymbol{z}_{r_+}$ also to be a one-hot vector. For the embedding of the reversal relation $\boldsymbol{z}_{r_-}$, we set it to be the negation of $\boldsymbol{z}_{r_+}$, *i.e.*, $\boldsymbol{z}_{r_-} = -\boldsymbol{z}_{r_+}$. This choice is not arbitrary: it idealizes a widely observed geometric structure in learned embedding spaces, where relational transformations correspond to approximately linear directions. For instance, classical results in word embeddings show that relations can be captured by vector offsets (e.g., "king $-$ man $+$ woman $\approx$ queen"). Under this lens, $\boldsymbol{z}_{r_-} = -\boldsymbol{z}_{r_+}$ can be interpreted as an abstraction of the inductive bias learned in pre-training. Forward and inverse relations are expected to occupy opposite directions in representation space, reflecting their mutually inverse semantics (e.g., "husband" vs. "wife").

We fix the key-query matrix $\boldsymbol{W}_{\mathrm{KQ}}$ to $\mathbf{0}_{d \times d}$, so the trainable weights are the output and value matrices $\boldsymbol{W}_{\mathrm{O}}, \boldsymbol{W}_{\mathrm{V}}$, *i.e.*, $\boldsymbol{\theta} = (\boldsymbol{W}_{\mathrm{O}}, \boldsymbol{W}_{\mathrm{V}})$. This is reasonable since our prompt is in the form of $[s, r], s \in \mathcal{A} \cup \mathcal{B}, r \in \mathcal{R}$, which is clean and free of noise, so the attention should pick both tokens. In this spirit, by fixing $\boldsymbol{W}_{\mathrm{KQ}}$ to zero matrix, tokens s and r will always have equal attention weights $\frac{1}{2}$. Moreover, we also require the intrinsic dimension $d_h \geq d$, in order to derive non-degenerate solutions. We formalize this as the following assumption.

Assumption 3.1. The key-query matrix $\boldsymbol{W}_{\mathrm{KQ}}$ is fixed to $\mathbf{0}_{d \times d}$, and the intrinsic dimension d_h satisfies $d_h \geq d$.

Note that this assumption is also effectively adopted by Huang et al. (2025a). Under Assumption 3.1, the output logit is

$$\mathrm{TF}_{\boldsymbol{\theta}}([s, r]) = \frac{1}{2}\boldsymbol{W}_{\mathrm{O}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{T}}\boldsymbol{z}_s + \frac{1}{2}\boldsymbol{W}_{\mathrm{O}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{T}}\boldsymbol{z}_r. \tag{5}$$

Given a model parameter $\boldsymbol{\theta} = (\boldsymbol{W}_{\mathrm{O}}, \boldsymbol{W}_{\mathrm{V}})$, we further define the margin between the correct label $r(s)$ and incorrect label $s' \in \mathcal{A} \cup \mathcal{B} \backslash \{r(s)\}$ as their logit difference:

$$h_{[s,r],s'}(\boldsymbol{\theta}) = \mathrm{TF}_{\boldsymbol{\theta}}([s, r]; r(s)) - \mathrm{TF}_{\boldsymbol{\theta}}([s, r]; s'). \tag{6}$$

Since $\mathrm{TF}_{\boldsymbol{\theta}}([s, r])$ only depends on $\boldsymbol{W}_{\mathrm{O}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{T}}$, we can also write the margin as $h_{[s,r],s'}(\boldsymbol{W}_{\mathrm{O}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{T}})$.

Now, we are ready to state a fundamental lemma established in Huang et al. (2025a), which allows us to characterize $\boldsymbol{W}_{\mathrm{O}}\boldsymbol{W}_{\mathrm{V}}^{\mathrm{T}}$ under gradient descent as an SVM problem.

Lemma 3.2 (SVM form, Theorem 1 in Huang et al. (2025a))**.** *Suppose Assumption 3.1 holds. Consider running gradient descent with a small enough learning rate or gradient flow on loss* (4) *where the model is defined by* (3) *and* (5)*. If there exists a time t_0, such that $\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t_0)) < 1$, then any limiting point of $\boldsymbol{\theta}/\|\boldsymbol{\theta}\|$, where $\boldsymbol{\theta} = (\boldsymbol{W}_O, \boldsymbol{W}_V)$, is along the direction of a KKT point of a program, which has the same solutions for $\boldsymbol{W}_{OV}^F = \boldsymbol{W}_O\boldsymbol{W}_V^{\mathrm{T}}$ as the following nuclear norm minimization problem*

$$\min_{\boldsymbol{W}_{OV}^{\mathrm{F}}} \ \frac{1}{2}\|\boldsymbol{W}_{OV}^{\mathrm{F}}\|_\star^2 \quad \text{s.t. } h_{[s,r],s'}(\boldsymbol{W}_{OV}^{\mathrm{F}}) \geq 1, \forall [s,r] \in \mathcal{D},\ \forall s' \in \mathcal{V}\backslash\{r(s)\}, \tag{SVM}$$

where $\|\cdot\|_\star$ denotes the nuclear norm.

3.1 Implicit Bias Explains the Reversal Curse

In this part, we explain why the reversal curse happens from the lens of implicit bias by using Lemma 3.2.

Suppose the training dataset contains only the forward relation data. Then we can show that the trained model fails to generalize to the reversal task, which can be formally summarized in the following theorem.

Theorem 3.3. *Suppose Assumption 3.1 holds, $N \geq 2$, and problem* (SVM) *with forward relation training set $\mathcal{D} = \mathcal{D}_{r_+}$ admits a single optimal solution $\boldsymbol{W}_{OV}^+$. Then, $\forall [b, r_-] \in \mathcal{D}_{r_-}$, we have*

$$h_{[b,r_-],a}(\boldsymbol{W}_{OV}^+) = 0, \quad \forall a \in \mathcal{A}\backslash\{r_-(b)\}. \tag{7}$$

Thus, the model can not generalize to the reversal relation.

Proof. See Appendix B.1. □

We visualize the solution $\boldsymbol{W}_{\mathrm{OV}}^+$ in Theorem 3.3 in Figure 1. From Figure 1, the solution has a positive diagonal in the lower left block, which means the model has learned the forward knowledge. To solve the reversal task, the model must simultaneously store knowledge in the upper-right block. However, training only on the forward relation dataset yields a zero upper-right block, leading to a zero margin and the reversal curse.

3.2 Breaking the Reversal Curse via Identity Bridge

We have seen in Theorem 3.3 that training with the forward relation only dataset will suffer from the reversal curse. To resolve this problem, we propose to add regularization data to the training dataset, such that the model can encode the reversal knowledge in the weight via gradient.

The regularization data we use is called the "Identity Bridge", which was first introduced by Lin et al. (2025) to solve two-hop reasoning tasks. They showed that adding identity bridge data can enable the model to perform two-hop reasoning when only trained with one-hop data. In our paper, we take the spirit of this idea and use it to break the reversal curse.

The identity bridge dataset is in the form of

$$\mathcal{D}_{\text{idn}} = \{[a_i, r_{\text{id}}|a_i] : i \in [N]\} \cup \{[b_i, r_{\text{id}}|b_i] : i \in [N]\}, \tag{8}$$

where r_{id} is an identity relation.

The identity bridge dataset only contains training samples that map an entity to itself. This identity mapping reveals no information about the reversal knowledge, but it will affect the gradient, thus changing the optimization landscape.

We set the embedding of the identity relation to be a zero vector, *i.e.*, $\boldsymbol{z}_{r_{\text{id}}} = \mathbf{0}_d$. The reason for this choice is that the identity can be viewed as the composite of the forward relation and the reversal relation. Therefore, in this spirit, $\boldsymbol{z}_{r_{\text{id}}} = \boldsymbol{z}_{r_-} + \boldsymbol{z}_{r_+} = \mathbf{0}_d$, since we have set $\boldsymbol{z}_{r_-} = -\boldsymbol{z}_{r_+}$.

Then, we have the following theorem, showing that the identity bridge regularization can break the reversal curse.

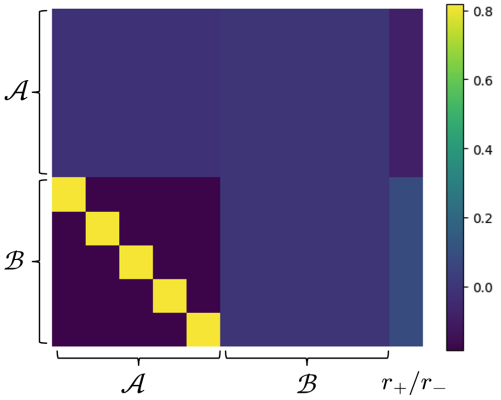


Figure 1: $\boldsymbol{W}_{OV}^{+}$ solution in Theorem 3.3 with forward relation dataset. In $\boldsymbol{W}_{OV}^{+}$, the diagonal weight of the upper right block is equal to the off-diagonal weight of it, which means when tested with the reversal data $[b_i, r_-]$, the trained model will output equal logits over all $a \in \mathcal{A}$. Thus, only training with the forward relation dataset will lead to the reversal curse.

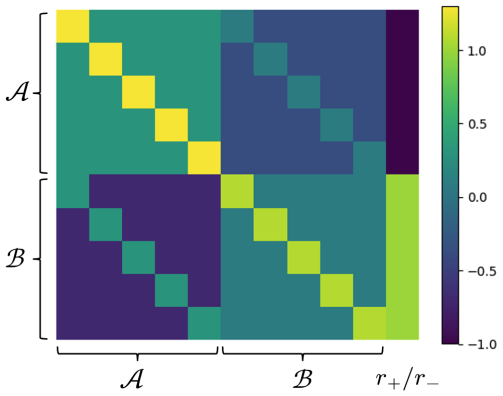


Figure 2: $\boldsymbol{W}_{OV}^{*}$ solution in Theorem 3.4 with identity bridge regularized dataset. In $\boldsymbol{W}_{OV}^{*}$, the diagonal weight of the upper right block is larger than the off-diagonal weight of it, so the logits of the correct reversal answer a_i will be larger than other answers when tested with $[b_i, r_-]$. Thus, the model can break the reversal curse with identity bridge regularized dataset.

Theorem 3.4. *Suppose Assumption 3.1 holds, $N \geq 2$, and the problem* (SVM) *with identity bridge regularized training set $\mathcal{D} = \mathcal{D}_{r_+} \cup \mathcal{D}_{idn}$ admits a single optimal solution $\boldsymbol{W}_{OV}^{*}$. Then, $\forall [b, r_-] \in \mathcal{D}_{r_-}$, we have*

$$h_{[b,r_-],a}(\boldsymbol{W}_{OV}^{*}) > 0, \forall a \in \mathcal{A} \cup \mathcal{B} \backslash \{r_-(b)\}. \tag{9}$$

Thus, the model can generalize to the reversal relation.

Proof. See Appendix B.2. □

We visualize the $\boldsymbol{W}_{\text{OV}}^{*}$ solution in Theorem 3.4 with the identity bridge regularized training dataset in Figure 2. We note that, in this case, the solution has a positive upper-right block, which means the model has encoded the reversal knowledge in the weights.

An intuitive explanation of why the identity bridge can help is that the identity training data forces the model to have positive weights at its two diagonal blocks, as observed in Figure 2, and to minimize the nuclear norm of $\boldsymbol{W}_{OV}^{*}$, the upper-right block is required to have positive diagonal values.

In addition to this explanation, we find that the identity bridge is closely related to the following out-of-context reasoning phenomenon, which can provide more insight into the working mechanism of our method.

3.3 Relation to Out-of-Context Reasoning

Out-of-context reasoning (OCR) (Cohen et al., 2024; Huang et al., 2025a) refers to a model's ability to deduce implications beyond the explicitly trained knowledge by drawing connections between different pieces of learned knowledge. Following Huang et al. (2025a), in an OCR task, one has a training subject set $\mathcal{S}_{\text{train}}$, a test subject set $\mathcal{S}_{\text{test}}$, two relations r_1, r_2, a set of facts $\mathcal{F}$, and a set of implications $\mathcal{I}$. Relation r_1 maps a subject to a fact in $\mathcal{F}$ and r_2 maps a subject to an implication in $\mathcal{I}$, such that $\forall s_1, s_2 \in \mathcal{S}_{\text{train}} \cup \mathcal{S}_{\text{test}}$, if they share the same fact $r_1(s_1) = r_1(s_2)$, then they have the same implication $r_2(s_1) = r_2(s_2)$. An OCR task is to train the model with the dataset $\mathcal{D}_{\text{train}} = \{[s, r_1|r_1(s)] : s \in \mathcal{S}_{\text{train}} \cup \mathcal{S}_{\text{test}}\} \cup \{[s, r_2|r_2(s)] : s \in \mathcal{S}_{\text{train}}\}$ which contains implications on $\mathcal{S}_{\text{train}}$ and all facts, and test the implications on $\mathcal{S}_{\text{test}}$, *i.e.*, $\mathcal{D}_{\text{test}} = \{[s, r_2|r_2(s)] : s \in \mathcal{S}_{\text{test}}\}$. For example, if we take r_1 to be "lives in", and r_2 to be "speaks", then the training dataset looks like { ["Jone" ,"lives in" | "Japan"], ["Jone", "speaks" | "Japanese"], ["Mike","lives in" | "Japan"] }, and we expect the model can deduce "Mike speaks Japanese".

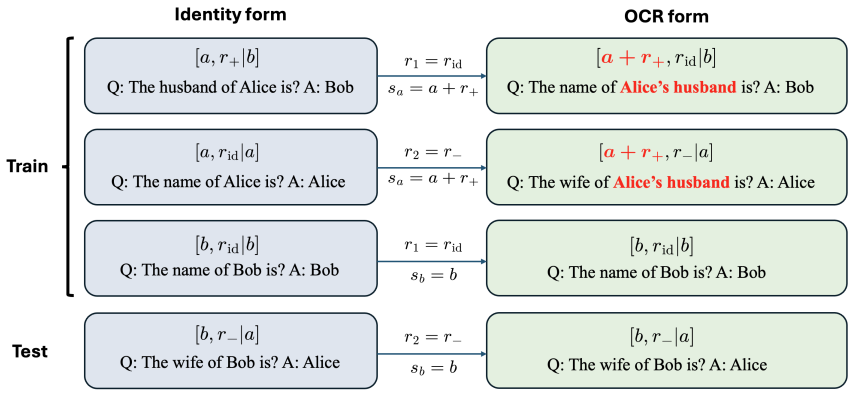


Figure 3: Illustration of Proposition 3.5 when $\boldsymbol{z}_{r_+} = -\boldsymbol{z}_{r_-}$. Identity form refers to the form of the identity regularized dataset, and OCR form refers to its OCR form given in Proposition 3.5. In the concrete "Husband-Wife" example in the Identity form, we set a to "Alice", b to "Bob", r_+ to "husband", r_- to "wife", and r_{id} to "name". In the corresponding OCR form, $s_a = a + r_+$ is "Alice's husband", $s_b = b$ is Bob, $r_1 = r_{\text{id}}$ is "name", and $r_2 = r_-$ is "wife". The left Identity form and the right OCR form are semantically equivalent and share the same test dataset.

We show in the following proposition that the identity bridge regularized reversal task is equivalent to an OCR task if we select an appropriate identity relation embedding.

Proposition 3.5. *For arbitrary embedding of entities $\boldsymbol{z}_s, s \in \mathcal{A} \cup \mathcal{B}$ and forward and reversal relations $\boldsymbol{z}_{r_+}, \boldsymbol{z}_{r_-}$, if the identity relation embedding is selected as $\boldsymbol{z}_{r_{id}} = (\boldsymbol{z}_{r_+} + \boldsymbol{z}_{r_-})/2$, then there exists an OCR task equivalent to the regularized reversal task in the sense that they have the same test dataset and their training datasets lead to the same SVM problem* (SVM).

Proof. The idea is to use the linearity of the one-layer transformer model (5) and rewrite the input sequence $[s, r]$ to an appropriate form. Specifically, we construct the OCR task as follows.

$$\mathcal{S}_{\text{train}} = \{s_a^{(i)} : i \in [N]\}, \boldsymbol{z}_{s_a^{(i)}} = \boldsymbol{z}_{a_i} + \frac{\boldsymbol{z}_{r_+} - \boldsymbol{z}_{r_-}}{2} \tag{10}$$

$$\mathcal{S}_{\text{test}} = \{s_b^{(i)} : i \in [N]\}, \boldsymbol{z}_{s_b^{(i)}} = \boldsymbol{z}_{b_i} \tag{11}$$

$$\boldsymbol{z}_{r_1} = \boldsymbol{z}_{r_{\text{id}}}, \boldsymbol{z}_{r_2} = \boldsymbol{z}_{r_-} \tag{12}$$

$$\mathcal{F} = \{u_b^{(i)} : i \in [N]\}, \boldsymbol{z}_{u_b^{(i)}} = \boldsymbol{z}_{b_i} \tag{13}$$

$$\mathcal{I} = \{u_a^{(i)} : i \in [N]\}, \boldsymbol{z}_{u_a^{(i)}} = \boldsymbol{z}_{a_i} \tag{14}$$

$$r_1(s_a^{(i)}) = u_b^{(i)}, r_2(s_a^{(i)}) = u_a^{(i)} \tag{15}$$

$$r_1(s_b^{(i)}) = u_b^{(i)}, r_2(s_b^{(i)}) = u_a^{(i)} \tag{16}$$

$$\mathcal{D}_{\text{train-OCR}} = \{[s_a^{(i)}, r_1|u_b^{(i)}], [s_a^{(i)}, r_2|u_a^{(i)}], [s_b^{(i)}, r_1|u_b^{(i)}] : i \in [N]\} \tag{17}$$

$$\mathcal{D}_{\text{test-OCR}} = \{[s_b^{(i)}, r_2|u_a^{(i)}] : i \in [N]\}. \tag{18}$$

In this construction, the test dataset $\mathcal{D}_{\text{test-OCR}}$ is the same as that of the reversal task. To verify the equivalence of the SVM form, we only need to check the sum of the subject embedding and the relation embedding. Specifically, we have

$$\boldsymbol{z}_{s_a^{(i)}} + \boldsymbol{z}_{r_1} = \boldsymbol{z}_{a_i} + \boldsymbol{z}_{r_+} \tag{19}$$

$$\boldsymbol{z}_{s_a^{(i)}} + \boldsymbol{z}_{r_2} = \boldsymbol{z}_{a_i} + \boldsymbol{z}_{r_{\text{id}}} \tag{20}$$

$$\boldsymbol{z}_{s_b^{(i)}} + \boldsymbol{z}_{r_1} = \boldsymbol{z}_{b_i} + \boldsymbol{z}_{r_{\text{id}}}. \tag{21}$$

Therefore, $\mathcal{D}_{\text{train-OCR}}$ leads to the same SVM problem as the identity bridge regularized dataset $\mathcal{D}_{r_+} \cup \mathcal{D}_{\text{idn}}$. □

We give an illustration of Proposition 3.5 in Figure 3, where we assume $\boldsymbol{z}_{r_+} = -\boldsymbol{z}_{r_-}$, so that $\boldsymbol{z}_{s_a^{(i)}} = \boldsymbol{z}_{a_i} + \frac{\boldsymbol{z}_{r_+} - \boldsymbol{z}_{r_-}}{2} = \boldsymbol{z}_{a_i} + \boldsymbol{z}_{r_+}$. In Figure 3, we give a concrete "Husband-Wife" example to show how this form equivalence works in real-world language data. In the example, we set a to "Alice", b to "Bob", r_+ to "husband", r_- to "wife", and r_{id} to "name". The trick in the form transformation is using the equality $r_{\text{id}} = r_+ + r_-$, *i.e.*, "name" = "wife" + "husband", and setting $s_a = a + r_+$, *i.e.*, "Alice's husband". For other parts in the corresponding OCR form, $s_b = b$ is Bob, $r_1 = r_{\text{id}}$ is "name", and $r_2 = r_-$ is "wife". The Identity form and the OCR form are semantically equivalent and share the same test dataset.

We note that the identity relation embedding used in Section 3.2 also satisfies the condition in Proposition 3.5, and Proposition 3.5 provides another explanation of why the identity bridge regularization can help break the reversal curse from the perspective of OCR.

Proposition 3.5 allows us to transform a reversal task into an OCR task, and we will show in the experiment that this transformation is critical to break the reversal curse in real LLMs.

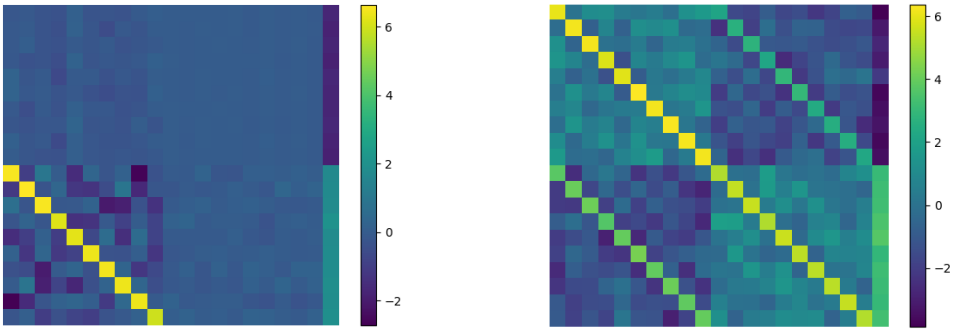
4 Experiments

In this section, we conduct experiments on both one-layer transformers and pretrained real-world LLMs to validate our method. In Section 4.1, we present experiments on one-layer transformers to verify the theoretical results in Section 3. Then, in Section 4.2, we conduct experiments on real LLMs to solve real-world reversal tasks. Detailed experiment setups and additional experimental results can be found in Appendix C.

4.1 One-Layer Transformer Experiments

We experiment with the reversal task with $N = 10$ pairs, and the embeddings are set as $\boldsymbol{z}_{a_i} = \boldsymbol{e}_i, \boldsymbol{z}_{b_i} = \boldsymbol{e}_{N+i}, \boldsymbol{z}_{r_+} = \boldsymbol{e}_{2N+1}, \boldsymbol{z}_{r_-} = -\boldsymbol{e}_{2N+1}, \boldsymbol{z}_{r_{\text{id}}} = \boldsymbol{0}_d$.

We present in Figure 4 the $\boldsymbol{W}_{\text{O}}\boldsymbol{W}_{\text{V}}^{\text{T}}$ weight after training with the forward relation dataset and the identity bridge regularized dataset, respectively. The trained $\boldsymbol{W}_{\text{O}}\boldsymbol{W}_{\text{V}}^{\text{T}}$ weights match the theoretical results in Theorem 3.3 and Theorem 3.4.



(a) (b)

Figure 4: (a). $\boldsymbol{W}_{\text{O}}\boldsymbol{W}_{\text{V}}^{\text{T}}$ weight after training with the forward relation dataset; (b) $\boldsymbol{W}_{\text{O}}\boldsymbol{W}_{\text{V}}^{\text{T}}$ weight after training with the identity bridge regularized dataset. The trained weights match the SVM solutions in Theorem 3.3 and Theorem 3.4.

4.2 Real Large Language Model Experiments

In this part, we consider two real-world reversal tasks, "Husband-wife" and "Parent-Child". In each task, the entity names are drawn from real-life names, and we randomly associate them to form 100 reversal pairs. We fine-tune a Llama-3.2-1B-Instruct model to solve these tasks.

The training dataset and test dataset contain prompts in the following format, where we take the "Husband-wife" task for an example.

Training dataset:

1. $[a, r_+|b]$: Q: The husband of Alice is? A: Bob.
2. $[a, r_{\text{id}}|a]$: Q: The name of Alice is? A: Alice.
3. $[b, r_-, r_+|b]$: Q: The husband of Bob's wife is? A: Bob.
4. $[a, r_+, r_{\text{id}}|b]$: Q: The name of Alice's husband is? A: Bob.

Test dataset:

- $[b, r_-, r_{\text{id}}|a]$: Q: The name of Bob's wife is? A: Alice.

In the identity bridge data for b, instead of using a prompt like "The name of Bob is?", we rewrite it by decomposing the identity relation "The name of" to be the composition of the forward and the reversal relation, *i.e.*, "The husband of b's wife". This rewriting trick transforms the reversal task into its OCR form (See Proposition 3.5). In this equivalent OCR task, the training subject "Alice" has two relations, "name" and "husband". The test subject is "Bob's wife", which has seen one of the relation "husband" in the training set, and the task is to generalize to the other relation "name".

Lastly, we add another data type $[a, r_+, r_{\text{id}}|b]$ in the training set, which is semantically equivalent to the forward relation data $[a, r_+|b]$. However, in training, it serves as an important regularization, preventing the model from learning a shortcut (see Section 4.2.1).

The reversal test results of the "Husband-Wife" task and the "Parent-Child" task are presented in Figure 5. For each task, the reversal test accuracy starts at 0% and increases after about 300 gradient steps, and finally can reach an accuracy of nearly 40%. In the first 300 gradient steps, the reversal test accuracy stays 0%, but the mean reciprocal rank (MRR) keeps increasing to around 0.7. The reason is that the answer (*e.g.*, "Alice.") usually consists of 4 to 5 tokens, and the first 300 gradient steps improve the accuracy of the last 3 to 4 tokens, leading to an MRR of about 0.7. Another interesting phenomenon is that as the reversal test accuracy increases from 0% to near 40%, the test loss only decreases marginally. This suggests that the model often places the correct token at rank one, but does so with low confidence—i.e., the logit gap between the correct token and competing tokens remains small. Such a small margin indicates that the prediction is sensitive to the gradient noise, which helps explain why the reversal test accuracy is fluctuating during training.

4.2.1 Ablation Experiments on Dataset

In this part, we focus on the "Husband-Wife" task and investigate the effect of the dataset on the model's reversal generalization ability. We consider the following three datasets.

- Identity (IDN) style dataset
  - Q: The husband of Alice is? A: Bob.
  - Q: The name of Alice is? A: Alice.
  - Q: The name of Bob is? A: Bob.
- OCR style dataset: 1 + 2 + 3.
- OCR+ dataset: 1 + 2 + 3 + 4.

In the IDN dataset, the identity relation of b is directly phrased as "The name of b is b", instead of using the rephrasing trick as in 3. The OCR dataset differs from the OCR+ dataset in the absence of the additional data 4.

Except for the reversal test data, we further consider the following shortcut test data

- $[b, r_-, r_{\text{id}}|b]$: Q: The name of Bob's wife is? A: Bob.

The shortcut test data measures whether the model learns a shortcut to copy the next token after "name" into its output, resulting in a prediction that contradicts the ground-truth knowledge.

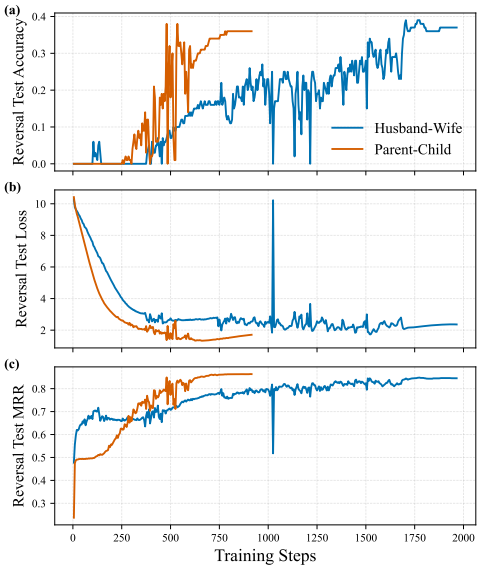


Figure 5: Reversal test results of the "Husband-Wife" task and the "Parent-Child" task: (a). Reversal test accuracy; (b). Reversal test loss; (c). Reversal test MRR. In both tasks, the model can break the reversal curse and reach near 40% reversal test accuracy.

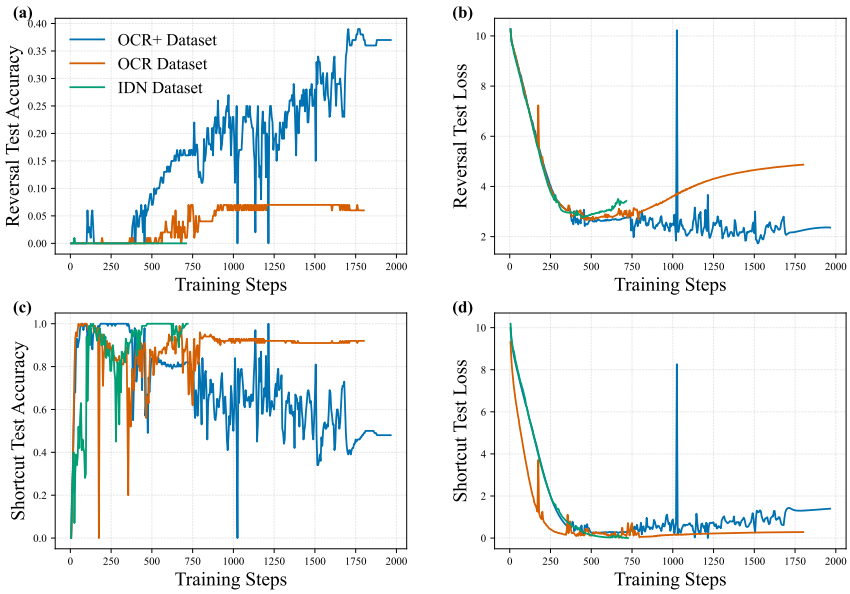


Figure 6: Test results for different datasets. (a). Reversal test accuracy; (b). Reversal test loss; (c). Shortcut test accuracy; (d). Shortcut test loss. In all cases, the model first quickly learns the shortcut (figure (c)), and then learn the reversal relation. The IDN dataset case stays trapped in the shortcut and fails to break the reversal curse. When using the identity bridge, the model can generalize to the reversal relation, as shown in the OCR dataset and OCR+ dataset cases. Adding the additional data 4 in OCR+ dataset can greatly help the model get rid of the shortcut trap.

The test results of these datasets are shown in Figure 6. Across all three datasets, the shortcut test accuracy quickly rises to 100% within a few gradient steps, indicating that the model learns the shortcut much faster than the reversal knowledge. With the IDN dataset, the model stays trapped in

this shortcut and never learns the reversal knowledge. While by rephrasing the identity relation to form 3, the model is able to break the reversal curse (as shown in the OCR dataset case), but is still significantly affected by the shortcut, only reaching a pass rate of 6%. Furthermore, by using the additional data 4, we can suppress this shortcut and greatly improve reversal test accuracy. These results suggest that the identity bridge does not work without reforming to its OCR form, which is ignored by other works studying the identity bridge (Lin et al., 2025).

4.2.2 Ablation Experiments on Entity Token Length

In this part, we investigate the influence of entity token length on the model's reversal generalization ability. We consider three types of names: number name (e.g., "34"), normal name (e.g., "Alice"), and long name (e.g., "Annaliese")[1]. The reversal test results are presented in Figure 7. From Figure 7, a surprising fact is that when the name is a short number, the reversal test accuracy can reach 100%. But when the name is long, the test accuracy decreases to 7%. This result suggests that entities with fewer tokens can be learned better in the reversal task, which is also observed in Wang & Sun (2025).

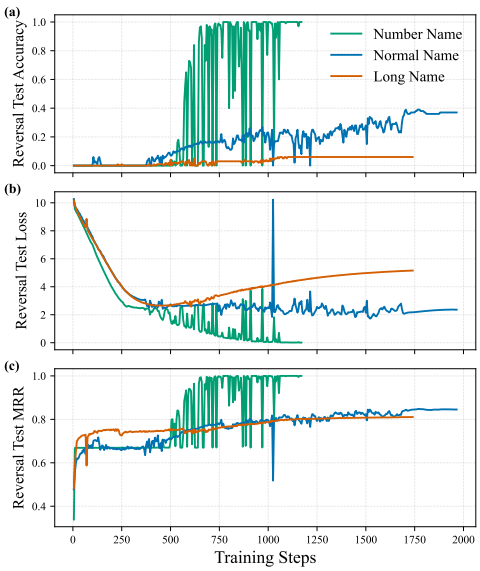


Figure 7: Reversal test results for different token lengths. (a). Reversal test accuracy; (b). Reversal test loss; (c). Reversal test MRR. The model generalizes better on shorter names, even reaching 100% accuracy on number names.

5 Conclusions

In this paper, we propose using the identity bridge regularized data recipe to break the reversal curse, which, to the best of our knowledge, is the first method that can work for autoregressive LLMs without modifying the training paradigm or reversing the training data. We prove in theory that this data recipe can help break the reversal curse even in a one-layer transformer, and relate it to the OCR phenomenon, which provides insight for understanding the identity bridge regularization. In experiments, a 1B pretrained language model finetuned with the proposed data recipe achieves a 40% success rate on reversal tasks, fully validating the effectiveness of the proposed method. The proposed identity bridge regularization method also admits limitations that the model can learn a shortcut, which is hard to eliminated, resulting in a gap between the 40% pass rate and the ideal 100% pass rate. For future directions, it would be helpful to understand the gap between one-token entity and multi-token entity, or symbolic data and textual token data, for real LLMs to solve the reversal tasks.

[1]The three types of names mainly differ in the number of tokens. E.g., number name like "34" is encoded in 1-2 tokens, normal name like "Alice" is encoded in 2-3 tokens, and long name like "Annaliese" is encoded in 3-5 tokens.

REFERENCES

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023.

Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems*, 36: 24519–24551, 2023.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024.

Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. *Advances in Neural Information Processing Systems*, 36: 11912–11951, 2023.

Olga Golovneva, Zeyuan Allen-Zhu, Jason Weston, and Sainbayar Sukhbaatar. Reverse training to nurse the reversal curse. *arXiv preprint arXiv:2403.13799*, 2024.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018a. URL `https://proceedings.mlr.press/v80/gunasekar18a.html`.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018b.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Qingyan Guo, Rui Wang, Junliang Guo, Xu Tan, Jiang Bian, and Yujiu Yang. Mitigating reversal curse in large language models via semantic-aware permutation training. *arXiv preprint arXiv:2403.00758*, 2024a.

Tianyu Guo, Druv Pai, Yu Bai, Jiantao Jiao, Michael I Jordan, and Song Mei. Active-dormant attention heads: Mechanistically demystifying extreme-token phenomena in llms. *arXiv preprint arXiv:2410.13835*, 2024b.

Tianyu Guo, Hanlin Zhu, Ruiqi Zhang, Jiantao Jiao, Song Mei, Michael I Jordan, and Stuart Russell. How do llms perform two-hop reasoning in context? *arXiv preprint arXiv:2502.13913*, 2025b.

Yixiao Huang, Hanlin Zhu, Tianyu Guo, Jiantao Jiao, Somayeh Sojoudi, Michael I Jordan, Stuart Russell, and Song Mei. Generalization or hallucination? understanding out-of-context reasoning in transformers. *arXiv preprint arXiv:2506.10887*, 2025a.

Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 19660–19722, 2024.

Yu Huang, Zixin Wen, Aarti Singh, Yuejie Chi, and Yuxin Chen. Transformers provably learn chain-of-thought reasoning with length generalization. *arXiv preprint arXiv:2511.07378*, 2025b.

M Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers. *arXiv preprint arXiv:2402.13512*, 2024.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1772–1798. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/ji19a.html.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ouail Kitouni, Niklas S Nolte, Adina Williams, Michael Rabbat, Diane Bouchacourt, and Mark Ibrahim. The factorization curse: Which tokens you predict underlie the reversal curse and more. *Advances in Neural Information Processing Systems*, 37:112329–112355, 2024.

Yingcong Li, Yixiao Huang, Muhammed E Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2024.

Pengxiao Lin, Zheng-An Chen, and Zhi-Qin John Xu. Identity bridge: Enabling implicit reasoning via shared latent memory. *arXiv preprint arXiv:2509.24653*, 2025.

Zhengkai Lin, Zhihang Fu, Kai Liu, Liang Xie, Binbin Lin, Wenxiao Wang, Deng Cai, Yue Wu, and Jieping Ye. Delving into the reversal curse: How far can large language models generalize? *Advances in Neural Information Processing Systems*, 37:30686–30726, 2024.

Zhicong Lu, Li Jin, Peiguang Li, Yu Tian, Linhao Zhang, Sirui Wang, Guangluan Xu, Changyuan Tian, and Xunliang Cai. Rethinking the reversal curse of llms: a prescription from human knowledge reversal. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7518–7530, 2024.

Ang Lv, Kaiyi Zhang, Shufang Xie, Quan Tu, Yuhan Chen, Ji-Rong Wen, and Rui Yan. An analysis and mitigation of the reversal curse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13603–13615, 2024.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

Mor Shpigel Nacson, Suriya Gunasekar, Jason Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *International Conference on Machine Learning*, pp. 4683–4692. PMLR, 2019a.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019b.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *International Conference on Machine Learning*, pp. 38018–38070. PMLR, 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Xu Pan, Ely Hahami, Jingxuan Fan, Ziqian Xie, and Haim Sompolinsky. Closing the data-efficiency gap between autoregressive and masked diffusion llms. *arXiv preprint arXiv:2510.09885*, 2025.

Heejune Sheen, Siyu Chen, Tianhao Wang, and Harrison H Zhou. Implicit regularization of gradient flow on one-layer softmax attention. *arXiv preprint arXiv:2403.08699*, 2024.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023a.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, 36:48314–48362, 2023b.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in neural information processing systems*, 36:71911–71947, 2023a.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*, 2023b.

Gal Vardi, Ohad Shamir, and Nati Srebro. On margin maximization in linear and relu networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022.

Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *arXiv preprint arXiv:2402.05738*, 2024.

Boshi Wang and Huan Sun. Is the reversal curse a binding problem? uncovering limitations of transformers from a basic generalization failure. *arXiv preprint arXiv:2504.01928*, 2025.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. *arXiv preprint arXiv:2505.23683*, 2025.

Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: unveiling how chain-of-thought enhances transformer sample efficiency. *arXiv preprint arXiv:2410.05459*, 2024.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Hanlin Zhu, Baihe Huang, Shaolun Zhang, Michael Jordan, Jiantao Jiao, Yuandong Tian, and Stuart J Russell. Towards a theoretical understanding of the'reversal curse'via training dynamics. *Advances in Neural Information Processing Systems*, 37:90473–90513, 2024.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Emergence of superposition: Unveiling the training dynamics of chain of continuous thought. *arXiv preprint arXiv:2509.23365*, 2025a.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. *arXiv preprint arXiv:2505.12514*, 2025b.

A Related Work

Reversal curse. The reversal curse phenomenon was first reported by Berglund et al. (2023). Extensive prior works have attempted to understand or resolve the reversal curse. Zhu et al. (2024) analyzed the training dynamics of different autoregressive models and showed that the reversal curse is caused by the model weights asymmetry. Lin et al. (2024) suggested that the reversal curse stems from the factual recall bias in LLMs. Wang & Sun (2025) conjectured that it is the inconsistency and entanglements of concept representations in LLMs incurs the reversal curse. To mitigate the reversal curse, several works (Golovneva et al., 2024; Guo et al., 2024a; Lu et al., 2024; Pan et al., 2025) proposed data-centric interventions such as reordering or paraphrasing training examples, or training on masked variants, while others (Lv et al., 2024; Kitouni et al., 2024) attributed the issue to the training objective itself and attempt to address it by modifying the learning formulation. However, these methods significantly change the existing training paradigms, which can harm the model performance. Separately, Nie et al. (2025) finds that diffusion language models trained with a

bidirectional denoising objective is able to break the reversal curse, further suggesting that the issue is closely tied to the autoregressive training paradigms. These works all failed to resolve the reversal curse in a reasonable way, leaving it a seemly impossible task. But our work challenges this view and breaks the reversal curse just by adding regularization to the training data.

Training dynamics of transformers. A growing body of literature examines the optimization of transformer-based models (Mahankali et al., 2023; Fu et al., 2023; Tian et al., 2023a;b; Zhang et al., 2024; Li et al., 2024; Huang et al., 2024; Guo et al., 2024b). In particular, recent works focus on understanding the transformer's behavior on various reasoning tasks through the lens of training dynamics. For example, previous studies have explored the emergence of induction heads (Boix-Adsera et al., 2023), factual recall (Nichani et al., 2024), chain-of-thought reasoning (Wen et al., 2024; Huang et al., 2025b), chain of continuous thought (Zhu et al., 2025b;a), multi-hop reasoning (Guo et al., 2025b; Wang et al., 2025), etc. Our theoretical analysis mainly builds on Huang et al. (2025a), which providedsss a theoretical understanding of out-of-context reasoning. Contrary to the negative result in Zhu et al. (2024), we show that even a one-layer transformer is able to learn the reversal reasoning with the identity bridge.

Implicit bias. The implicit bias of gradient descent has been extensively studied in classification tasks, which connects problems with logistic or exponentially-tailed loss to margin maximization (Soudry et al., 2018; Gunasekar et al., 2018b;a; Lyu & Li, 2019; Nacson et al., 2019b;a; Ji & Telgarsky, 2019; Vardi et al., 2022). There are also many works exploring this connection in attention-based models (Tarzanagh et al., 2023a;b; Li et al., 2024; Ildiz et al., 2024; Sheen et al., 2024; Vasudeva et al., 2024). Similar to Huang et al. (2025a), our work characterizes the solution to SVM programs to understand how the identity bridge helps break the reversal curse in transformers.

B Omitted Proof

B.1 Proof of Theorem 3.3

Without loss of generality, we assume the embedding of $a_i \in \mathcal{A}, b_i \in \mathcal{B}, r_+$ to be $\boldsymbol{z}_{a_i} = \boldsymbol{e}_i, \boldsymbol{z}_{b_i} = \boldsymbol{e}_{N+i}, \boldsymbol{z}_{r_+} = \boldsymbol{e}_{2N+1}$, respectively. Then, we can obtain the following lemma characterizing the symmetry of the solution $\boldsymbol{W}_{\mathrm{OV}}$.

Lemma B.1. *Suppose $\boldsymbol{W}_{OV}$ is the single solution to problem* (SVM) *with $\mathcal{D} = \mathcal{D}_{r_+}$. Then, $\boldsymbol{W}_{OV}$ admits the following form*

$$\boldsymbol{W}_{OV} = \begin{bmatrix} f_1\boldsymbol{I}_N + f_2\boldsymbol{E}_N & g_1\boldsymbol{I}_N + g_2\boldsymbol{E}_N & \beta\mathbf{1}_N \\ l_1\boldsymbol{I}_N + l_2\boldsymbol{E}_N & m_1\boldsymbol{I}_N + m_2\boldsymbol{E}_N & \alpha\mathbf{1}_N \end{bmatrix}, \tag{22}$$

where $f_1, f_2, g_1, g_2, l_1, l_2, m_1, m_2, \beta, \alpha$ are constants.

The singular values of $\boldsymbol{W}_{OV}$ in (22) *are*

$$\{\sigma_1^{(1)}, \sigma_1^{(2)}, \underbrace{\sigma_2^{(1)}, \cdots, \sigma_2^{(1)}}_{N-1}, \underbrace{\sigma_2^{(2)}, \cdots, \sigma_2^{(2)}}_{N-1}\}, \tag{23}$$

and they can be computed by

$$\sigma_1^{(1)} = \sqrt{\frac{(C_{A1} + NC_{A2} + C_{D1} + NC_{D2}) + \sqrt{(C_{A1} + NC_{A2} - (C_{D1} + NC_{D2}))^2 + 4(C_{B1} + NC_{B2})^2}}{2}}; \tag{24}$$

$$\sigma_1^{(2)} = \sqrt{\frac{(C_{A1} + NC_{A2} + C_{D1} + NC_{D2}) - \sqrt{(C_{A1} + NC_{A2} - (C_{D1} + NC_{D2}))^2 + 4(C_{B1} + NC_{B2})^2}}{2}}; \tag{25}$$

$$\sigma_2^{(1)} = \sqrt{\frac{(C_{A1} + C_{D1}) + \sqrt{(C_{A1} - C_{D1})^2 + 4C_{B1}^2}}{2}}; \tag{26}$$

$$\sigma_2^{(2)} = \sqrt{\frac{(C_{A1} + C_{D1}) - \sqrt{(C_{A1} - C_{D1})^2 + 4C_{B1}^2}}{2}}, \tag{27}$$

where

$$C_{A1} = f_1^2 + g_1^2 \tag{28}$$
$$C_{A2} = 2f_1f_2 + Nf_2^2 + 2g_1g_2 + Ng_2^2 + \beta^2 \tag{29}$$
$$C_{D1} = l_1^2 + m_1^2 \tag{30}$$
$$C_{D2} = 2l_1l_2 + Nl_2^2 + 2m_1m_2 + Nm_2^2 + \alpha^2 \tag{31}$$
$$C_{B1} = f_1l_1 + g_1m_1 \tag{32}$$
$$C_{B2} = f_1l_2 + f_2l_1 + Nf_2l_2 + g_1m_2 + g_2m_1 + Ng_2m_2 + \beta\alpha. \tag{33}$$

The proof of this lemma is almost the same as that of Lemmas 3, 4, and 5 in Huang et al. (2025a).

We can then give the proof of Theorem 3.3.

Proof. By (22), $\forall a \in \mathcal{A}\backslash\{r_-(b)\}, h_{[b,r_-],a}(\boldsymbol{W}_{\text{OV}}) = g_1$. Thus, it suffices to show that $g_1 = 0$ for the optimal solution.

By Lemma B.1, the nuclear norm of $\|\boldsymbol{W}_{\text{OV}}\|_\star$ is

$$\|\boldsymbol{W}_{\text{OV}}\|_\star = \sigma_1^{(1)} + \sigma_1^{(2)} + (N-1)\sigma_2^{(1)} + (N-1)\sigma_2^{(2)}. \tag{34}$$

$\sigma_1^{(1)} + \sigma_1^{(2)}$ is

$$\sigma_1^{(1)} + \sigma_1^{(2)} \tag{35}$$
$$= \sqrt{\frac{(C_{A1} + NC_{A2} + C_{D1} + NC_{D2}) + \sqrt{(C_{A1} + NC_{A2} - (C_{D1} + NC_{D2}))^2 + 4(C_{B1} + NC_{B2})^2}}{2}} + \sqrt{\frac{(C_{A1} + NC_{A2} + C_{D1} + NC_{D2}) - \sqrt{(C_{A1} + NC_{A2} - (C_{D1} + NC_{D2}))^2 + 4(C_{B1} + NC_{B2})^2}}{2}} \tag{36}$$
$$= \sqrt{C_{A1} + NC_{A2} + C_{D1} + NC_{D2} + 2\sqrt{(C_{A1} + NC_{A2})(C_{D1} + NC_{D2}) - (C_{B1} + NC_{B2})^2}} \tag{37}$$

$\sigma_2^{(1)} + \sigma_2^{(2)}$ is

$$\sigma_2^{(1)} + \sigma_2^{(2)} = \sqrt{\frac{(C_{A1} + C_{D1}) + \sqrt{(C_{A1} - C_{D1})^2 + 4C_{B1}^2}}{2}} + \sqrt{\frac{(C_{A1} + C_{D1}) - \sqrt{(C_{A1} - C_{D1})^2 + 4C_{B1}^2}}{2}} \tag{38}$$
$$= \sqrt{C_{A1} + C_{D1} + 2|f_1m_1 - g_1l_1|} \tag{39}$$

Therefore, the optimization problem (SVM) can be reformulated as

$$\begin{aligned} \min \|\boldsymbol{W}_{\text{OV}}\|_\star = & \sqrt{C_{A1} + NC_{A2} + C_{D1} + NC_{D2} + 2\sqrt{(C_{A1} + NC_{A2})(C_{D1} + NC_{D2}) - (C_{B1} + NC_{B2})^2}} \\ & + (N-1)\sqrt{C_{A1} + C_{D1} + 2|f_1m_1 - g_1l_1|} \\ \text{s.t.} \quad & l_1 \geq 1, \quad l_1 + l_2 + \alpha \geq f_1 + f_2 + \beta + 1, \quad l_1 + l_2 + \alpha \geq f_2 + \beta + 1. \end{aligned} \tag{40}$$

Notice that by the Cauchy-Schwarz inequality, we have

$$2\sqrt{(C_{A1} + NC_{A2})(C_{D1} + NC_{D2}) - (C_{B1} + NC_{B2})^2} \tag{41}$$
$$\geq 0 \tag{42}$$
$$\geq -\frac{1}{2}[(f_1 + Nf_2 + l_1 + Nl_2)^2 + (g_1 + Ng_2 + m_1 + Nm_2) + N(\beta + \alpha)^2], \tag{43}$$

where the first equality condition is $|f_1 + Nf_2| = |l_1 + Nl_2|, |g_1 + Ng_2| = |m_1 + Nm_2|$, and $|\beta| = |\alpha|$.

Therefore,

$$\sqrt{C_{A1} + NC_{A2} + C_{D1} + NC_{D2} + 2\sqrt{(C_{A1} + NC_{A2})(C_{D1} + NC_{D2}) - (C_{B1} + NC_{B2})^2}} \tag{44}$$

$$\geq \sqrt{C_{A1} + NC_{A2} + C_{D1} + NC_{D2} - \frac{1}{2}[(f_1 + Nf_2 + l_1 + Nl_2)^2 + (g_1 + Ng_2 + m_1 + Nm_2) + N(\beta + \alpha)^2]} \tag{45}$$

$$= \sqrt{\frac{(f_1 + Nf_2 - l_1 - Nl_2)^2 + (m_1 + Nm_2 - g_1 - Ng_2)^2 + N(\beta - \alpha)^2}{2}}, \tag{46}$$

where equality holds if and only if $f_1 + Nf_2 + l_1 + Nl_2 = g_1 + Ng_2 + m_1 + Nm_2 = \beta + \alpha = 0$.

Inspired by (46), we define the following new variables c_{f-l}, c_{m-g} and $c_{\beta-\alpha}$

$$c_{f-l} = f_1 + Nf_2 - l_1 - Nl_2, \quad c_{m-g} = m_1 + Nm_2 - g_1 - Ng_2, \quad c_{\beta-\alpha} = \beta - \alpha. \tag{47}$$

Then, the condition in (40) can be written as

$$l_1 \geq 1, \quad (N-1)l_1 \geq Nc_{\boldsymbol{\beta-\alpha}} + c_{f-l} + (N-1)f_1 + N, \quad (N-1)l_1 \geq Nc_{\boldsymbol{\beta-\alpha}} + c_{f-l} - f_1 + N. \tag{48}$$

Combining these observations, we can derive a lower bound problem for problem (40).

$$\begin{aligned} \min \|\boldsymbol{W}_{\text{OV}}\|_\star = & \sqrt{\frac{c_{f-l}^2 + c_{m-g}^2 + Nc_{\boldsymbol{\beta-\alpha}}^2}{2}} + (N-1)\sqrt{(f_1 - m_1)^2 + (g_1 - l_1)^2 + 4t} \\ \text{s.t.} \quad & q_1 : l_1 \geq 1, \\ & q_2 : (N-1)l_1 \geq (N-1)f_1 + c_{f-l} + Nc_{\boldsymbol{\beta-\alpha}} + N, \\ & q_3 : (N-1)l_1 \geq -f_1 + c_{f-l} + Nc_{\boldsymbol{\beta-\alpha}} + N, \\ & q_4 : t \geq f_1 m_1, \\ & q_5 : t \geq g_1 l_1. \end{aligned} \tag{49}$$

Notice that the lower bound problem (49) is optimized over variables $f_1, g_1, l_1, m_1, c_{f-l}, c_{m-g}$, $c_{\boldsymbol{\beta-\alpha}}$, and t. When these parameters are fixed, we can always find a set of other parameters such that the equality condition $f_1 + Nf_2 + l_1 + Nl_2 = g_1 + Ng_2 + m_1 + Nm_2 = \beta + \alpha = 0$ holds. Therefore, the lower bound problem (49) has the same optimal solution and optimal value as problem (40).

We notice that $l_1 = 1, f_1 = m_1 = g_1 = 0, c_{f-l} = c_{\boldsymbol{\beta-\alpha}} = -\frac{1}{N+1}, c_{m-g} = 0, t = 0$ is a feasible solution with objective value $N - 1 + \sqrt{\frac{1}{2(N+1)}}$. It serves as a baseline solution to help us eliminate suboptimal solutions (actually, it is the optimal solution).

Let us denote $M_1 = \frac{c_{f-l}^2 + c_{m-g}^2 + N c_{\boldsymbol{\beta}-\boldsymbol{\alpha}}^2}{2}$ and $M_2 = (f_1 - m_1)^2 + (g_1 - l_1)^2 + 4t$ for notation convenience. Then, the KKT condition for problem (49) is

$$\gamma_i \geq 0, i = 1, 2, 3, 4, 5 \tag{50}$$

$$l_1 : (N-1)\frac{l_1 - g_1}{\sqrt{M_2}} - \gamma_1 - (N-1)\gamma_2 - (N-1)\gamma_3 + g_1\gamma_5 = 0 \tag{51}$$

$$f_1 : (N-1)\frac{f_1 - m_1}{\sqrt{M_2}} + (N-1)\gamma_2 - \gamma_3 + m_1\gamma_4 = 0 \tag{52}$$

$$g_1 : (N-1)\frac{g_1 - l_1}{\sqrt{M_2}} + l_1\gamma_5 = 0 \tag{53}$$

$$m_1 : (N-1)\frac{m_1 - f_1}{\sqrt{M_2}} + f_1\gamma_4 = 0 \tag{54}$$

$$c_{f-l} : \frac{c_{f-l}}{2\sqrt{M_1}} + \gamma_2 + \gamma_3 = 0 \tag{55}$$

$$c_{\boldsymbol{\beta}-\boldsymbol{\alpha}} : N\frac{c_{\boldsymbol{\beta}-\boldsymbol{\alpha}}}{2\sqrt{M_1}} + N\gamma_2 + N\gamma_3 = 0 \tag{56}$$

$$c_{m-g} : \frac{c_{m-g}}{2\sqrt{M_1}} = 0 \tag{57}$$

$$t : (N-1)\frac{2}{\sqrt{M_2}} - \gamma_4 - \gamma_5 = 0, \tag{58}$$

where $\gamma_i, i \in [5]$ are the dual variables corresponding to the constraints $q_i, i \in [5]$.

By Equation (57), $c_{m-g} = 0$. By Equation (55) and Equation (56), $c_{\boldsymbol{\beta}-\boldsymbol{\alpha}} = c_{f-l}$. We then prove that $f_1 = 0$ by case analysis.

Case 1: $f_1 > 0$

In this case, constraint q_2 is strictly stronger than q_3, so q_3 can not be tight, which indicates that $\gamma_3 = 0$ by complementary slackness.

By computing (52) − (54), we have

$$0 = (52) - (54) = (f_1 - m_1)(\frac{2(N-1)}{\sqrt{M_2}} - \gamma_4) + (N-1)\gamma_2 \tag{59}$$

By Equation (54) and Equation (58),

$$f_1 - m_1 = f_1\gamma_4\frac{\sqrt{M_2}}{N-1} \geq 0 \tag{60}$$

$$\frac{2(N-1)}{\sqrt{M_2}} - \gamma_4 = \gamma_5 \geq 0 \tag{61}$$

Therefore, in Equation (59), both $(f_1 - m_1)(\frac{2(N-1)}{\sqrt{M_2}} - \gamma_4)$ and $(N-1)\gamma_2$ are non-negative, which means at least one of $f_1 - m_1$ and $\frac{2(N-1)}{\sqrt{M_2}} - \gamma_4$ is 0.

If $f_1 - m_1 = 0$, then by (54), $\gamma_4 = 0$, so by (58), $\gamma_5 = \frac{2(N-1)}{\sqrt{M_2}}$. Plugging $\gamma_5 = \frac{2(N-1)}{\sqrt{M_2}}$ into (53), we have $g_1 = -l_1 \leq -1$. Then, the objective value $\|\boldsymbol{W}_{\text{OV}}\|_\star \geq (N-1)\sqrt{(f_1 - m_1)^2 + (g_1 - l_1)^2 + 4t} \geq (N-1)\sqrt{(g_1 - l_1)^2} \geq 2(N-1) > N - 1 + \sqrt{\frac{1}{2(N+1)}}$. Thus, it can not be the optimal solution.

If $\frac{2(N-1)}{\sqrt{M_2}} - \gamma_4 = 0$, then by (58), $\gamma_5 = 0$. According to (53), $g_1 = l_1 \geq 1$. Then, the objective value $\|\boldsymbol{W}_{\text{OV}}\|_\star \geq (N-1)\sqrt{(f_1 - m_1)^2 + (g_1 - l_1)^2 + 4t} \geq (N-1)\sqrt{(g_1 + l_1)^2} \geq 2(N-1) > N - 1 + \sqrt{\frac{1}{2(N+1)}}$. Thus, it also can not be the optimal solution.

Therefore, there does not exist an optimal solution with $f_1 > 0$.

Case 2: $f_1 < 0$

Similarly, by complementary slackness, $\gamma_2 = 0$ and we have

$$0 = (52) - \ (54) = (f_1 - m_1)(\frac{2(N-1)}{\sqrt{M_2}} - \gamma_4) - \gamma_3 \tag{62}$$

Similarly, by Equation (54) and Equation (58), $f_1 - m_1 \leq 0$, and $(\frac{2(N-1)}{\sqrt{M_2}} - \gamma_4) \geq 0$, so we have $(f_1 - m_1)(\frac{2(N-1)}{\sqrt{M_2}} - \gamma_4) = 0$. By the same argument as in Case 1, there does not exist an optimal solution with $f_1 < 0$.

Therefore, the optimal solution must have $f_1 = 0$, so $f_1 m_1 = 0$.

We next show that the optimal $g_1 = 0$ by using the tightness of constraints q_4 and q_5.

If $g_1 > 0$, then $g_1 l_1 > 0 = f_1 m_1$, so constraint q_4 is not tight. Therefore, $\gamma_4 = 0$ and $\gamma_5 = \frac{2(N-1)}{\sqrt{M_2}}$. This indicates $g_1 = -l_1 \leq -1$ by (53), which contradict with $g_1 > 0$.

If $g_1 < 0$, then constraint q_5 is not tight, so $\gamma_5 = 0$, so $g_1 = l_1 \geq 1$ by (53), which contradict with $g_1 < 0$.

Therefore, the optimal g_1 is 0. So $h_{[b,r_-],a}(\boldsymbol{W}_{\text{OV}}^{+}) = 0, \forall a \in \mathcal{A} \backslash \{r_-(b)\}$. □

B.2 Proof of Theorem 3.4

Proof. We first note that Lemma B.1 also hold with $\mathcal{D} = \mathcal{D}_{r_+} \cup \mathcal{D}_{\text{idn}}$, so by (22), $\forall a \in \mathcal{A} \backslash \{r_-(b)\}$, it holds that $h_{[b,r_-],a}(\boldsymbol{W}_{\text{OV}}) = g_1$ and $\forall a \in \mathcal{B}$, we have $h_{[b,r_-],a}(\boldsymbol{W}_{\text{OV}}) = \min\{g_1 + g_2 - \beta - (m_1 + m_2 - \alpha), g_1 + g_2 - \beta - (m_2 - \alpha)\}$. Since the identity bridge data requires $m_1 \geq 1$ (as shown in (64)), it suffices to prove the following two inequalities

$$g_1 > 0, \quad g_1 + g_2 - \beta - (m_1 + m_2 - \alpha) > 0. \tag{63}$$

According to Lemma B.1, we can rewrite the problem Equation (SVM) as

$$\begin{aligned}
&\min \|\boldsymbol{W}_{\text{OV}}\|_\star \\
=&\sqrt{C_{A1} + NC_{A2} + C_{D1} + NC_{D2} + 2\sqrt{(C_{A1} + NC_{A2})(C_{D1} + NC_{D2}) - (C_{B1} + NC_{B2})^2}} \\
&+ (N-1)\sqrt{C_{A1} + C_{D1} + 2|f_1 m_1 - g_1 l_1|}, \\
&\text{s.t.} \quad f_1 \geq 1, l_1 \geq 1, m_1 \geq 1, \\
&\qquad l_1 + l_2 + \alpha \geq f_1 + f_2 + \beta + 1, \\
&\qquad f_1 + f_2 \geq l_1 + l_2 + 1, \\
&\qquad m_1 + m_2 \geq g_1 + g_2 + 1, \\
&\qquad m_1 + m_2 \geq g_2 + 1.
\end{aligned} \tag{64}$$

Similar to the proof of Theorem 3.3, we define $c_{f-l} = f_1 + Nf_2 - l_1 - Nl_2, c_{m-g} = m_1 + Nm_2 - g_1 - Ng_2, c_{\beta-\alpha} = \beta - \alpha$ and have the following lower bound problem

$$\begin{aligned}
&\min \|\boldsymbol{W}_{\text{OV}}\|_\star = \sqrt{\frac{c_{f-l}^2 + c_{m-g}^2 + Nc_{\boldsymbol{\beta-\alpha}}^2}{2}} + (N-1)\sqrt{(f_1 - m_1)^2 + (g_1 - l_1)^2 + 4t} \\
&\text{s.t.} \quad q_1 : f_1 \geq 1, \\
&\qquad q_2 : l_1 \geq 1, \\
&\qquad q_3 : m_1 \geq 1, \\
&\qquad q_4 : (N-1)l_1 \geq (N-1)f_1 + c_{f-l} + Nc_{\boldsymbol{\beta-\alpha}} + N, \\
&\qquad q_5 : (N-1)f_1 + c_{f-l} \geq (N-1)l_1 + N, \\
&\qquad q_6 : (N-1)m_1 + c_{m-g} \geq (N-1)g_1 + N, \\
&\qquad q_7 : (N-1)m_1 + c_{m-g} \geq -g_1 + N, \\
&\qquad q_8 : t \geq f_1 m_1, \\
&\qquad q_9 : t \geq g_1 l_1.
\end{aligned} \tag{65}$$

The lower bound problem (65) is optimized over variables $f_1, g_1, l_1, m_1, c_{f-l}, c_{m-g}, c_{\boldsymbol{\beta}-\boldsymbol{\alpha}}, t$. When these parameters are fixed, we can always find a set of other parameters such that the equality condition $f_1 + Nf_2 + l_1 + Nl_2 = g_1 + Ng_2 + m_1 + Nm_2 = \beta + \alpha = 0$ holds. Therefore, the lower bound problem (65) has the same optimal solution and optimal value as problem (64).

Notice that

$$f_1 = m_1 = l_1 = g_1 = 1, c_{f-l} = c_{m-g} = N, c_{\boldsymbol{\beta}-\boldsymbol{\alpha}} = -2, t = 1 \tag{66}$$

is a feasible solution with objective value

$$\sqrt{\frac{2N^2 + 4N}{2}} + 2(N-1) \leq (N+1) + 2(N-1) = 3N - 1, \tag{67}$$

which means any optimal solution must have an objective value no larger than $3N - 1$.

Let $M_1 = \frac{c_{f-l}^2 + c_{m-g}^2 + Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}}^2}{2}$ and $M_2 = (f_1 - m_1)^2 + (g_1 - l_1)^2 + 4t$. The KKT condition of problem (65) is

$$\gamma_i \geq 0, i \in [9] \tag{68}$$

$$l_1 : (N-1)\frac{l_1 - g_1}{\sqrt{M_2}} - \gamma_2 - (N-1)\gamma_4 + (N-1)\gamma_5 + g_1\gamma_9 = 0 \tag{69}$$

$$f_1 : (N-1)\frac{f_1 - m_1}{\sqrt{M_2}} - \gamma_1 + (N-1)\gamma_4 - (N-1)\gamma_5 + m_1\gamma_8 = 0 \tag{70}$$

$$g_1 : (N-1)\frac{g_1 - l_1}{\sqrt{M_2}} + (N-1)\gamma_6 - \gamma_7 + l_1\gamma_9 = 0 \tag{71}$$

$$m_1 : (N-1)\frac{m_1 - f_1}{\sqrt{M_2}} - \gamma_3 - (N-1)\gamma_6 - (N-1)\gamma_7 + f_1\gamma_8 = 0 \tag{72}$$

$$c_{f-l} : \frac{c_{f-l}}{2\sqrt{M_1}} + \gamma_4 - \gamma_5 = 0 \tag{73}$$

$$c_{m-g} : \frac{c_{m-g}}{2\sqrt{M_1}} - \gamma_6 - \gamma_7 = 0 \tag{74}$$

$$c_{\boldsymbol{\beta}-\boldsymbol{\alpha}} : N\frac{c_{\boldsymbol{\beta}-\boldsymbol{\alpha}}}{2\sqrt{M_1}} + N\gamma_4 = 0 \tag{75}$$

$$t : (N-1)\frac{2}{\sqrt{M_2}} - \gamma_8 - \gamma_9 = 0, \tag{76}$$

where $\gamma_i, i \in [9]$ are the dual variables corresponding to the constraints $q_i, i \in [9]$.

First, from constraints q_4 and q_5, we have

$$(N-1)l_1 \geq (N-1)f_1 + c_{f-l} + Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}} + N \geq (N-1)l_1 + N + Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}} + N \Rightarrow c_{\boldsymbol{\beta}-\boldsymbol{\alpha}} \leq -2, \tag{77}$$

so by (75), $\gamma_4 > 0$, which indicates constraint q_4 is tight.

We then prove the two inequalities.

Part 1: $g_1 > 0$

We prove by contradiction. Suppose $g_1 \leq 0$. Then, constraint q_9 is not tight, so $\gamma_9 = 0$. By (71)

$$\gamma_6 = \frac{1}{N-1}\gamma_7 + \frac{l_1 - g_1}{\sqrt{M_2}} > 0, \tag{78}$$

which means constraint q_6 is tight.

Plugging (78) into (74), we have

$$\sqrt{M_2} = \frac{l_1 - g_1}{\frac{c_{m-g}}{2\sqrt{M_1}} - \frac{N}{N-1}\gamma_7} \geq \frac{l_1 - g_1}{\frac{c_{m-g}}{2\sqrt{M_1}}} \geq \frac{2\sqrt{M_1}}{c_{m-g}}. \tag{79}$$

By constraint q_6,

$$c_{m-g} = (N-1)g_1 + N - (N-1)m_1 \leq 1. \tag{80}$$

Therefore,

$$\sqrt{M_2} \geq \frac{2\sqrt{M_1}}{c_{m-g}} \geq 2\sqrt{M_1}. \tag{81}$$

So the objective is lower bounded by

$$\begin{aligned}\|\boldsymbol{W}_{\text{OV}}\|_\star =& \sqrt{M_1} + (N-1)\sqrt{M_2} \geq \sqrt{M_1}(2N-1) \\ \geq& \sqrt{\frac{Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}}^2}{2}}(2N-1) \geq \sqrt{2N}(2N-1) \geq 4N-2 > 3N-1,\end{aligned}$$

and thus it is suboptimal to the known feasible solution (66).

Therefore, we must have $g_1 > 0$

Part 2: $g_1 + g_2 - \beta - (m_1 + m_2 - \alpha) > 0$

This inequality is equivalent to

$$(N-1)g_1 - (N-1)m_1 - c_{m-g} - Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}} > 0 \tag{82}$$

To prove this inequality, we discuss two cases.

Case 1: constraint $q_6 : (N-1)m_1 + c_{m-g} \geq (N-1)g_1 + N$ is tight. Then,

$$(N-1)g_1 - (N-1)m_1 - c_{m-g} = -N, \tag{83}$$

so

$$(N-1)g_1 - (N-1)m_1 - c_{m-g} - Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}} = -N - Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}} \geq -N - N \cdot (-2) = N. \tag{84}$$

Case 2: constraint q_6 is not tight. Then, $\gamma_6 = 0$. From part 1, we know $g_1 > 0$, so q_7 can not be tight (otherwise q_6 can not hold), which means $\gamma_7 = 0$. Combining these with (74), we conclude that $c_{m-g} = 0$. Therefore,

$$(N-1)g_1 - (N-1)m_1 - c_{m-g} - Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}} = (N-1)g_1 - (N-1)m_1 - Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}} > -(N-1)m_1 + 2N, \tag{85}$$

where last inequality uses $g_1 > 0$ and$c_{\boldsymbol{\beta}-\boldsymbol{\alpha}} \leq 2$.

By (85), it suffices to show $m_1 \leq 2$. Supposing the opposite, the objective is lower bounded by

$$\|\boldsymbol{W}_{\text{OV}}\|_\star = \sqrt{M_1} + (N-1)\sqrt{M_2} \geq 2 + (N-1)\sqrt{(m_1+f_1)^2} > 2 + (N-1)(2+1) = 3N-1, \tag{86}$$

which is suboptimal to the known feasible solution (66).

Therefore, in both cases, we have $(N-1)g_1 - (N-1)m_1 - c_{m-g} - Nc_{\boldsymbol{\beta}-\boldsymbol{\alpha}} > 0$. □

C Experiment Details

We provide implementation details on experiments in Section 4.1 and Section 4.2.

C.1 Experiment details of Section 4.1

In the one-layer transformer experiments, we use $N = 10$ reversal pairs, and use uniform initialization for $\boldsymbol{W}_O$ and $\boldsymbol{W}_V$ over $[-0.1, 0.1]$. In training, we use full-batch gradient descent with a learning rate of 0.001 to optimize the model.

The reversal test loss and mean reciprocal rank (MRR) of forward relation only training data and identity bridge regularized data are presented in Figure 8. From Figure 8, we observe that, without the identity bridge, the reversal test MRR stays around the initialization level, while adding identity bridge regularization, the model can generalize to all the reversal tests.

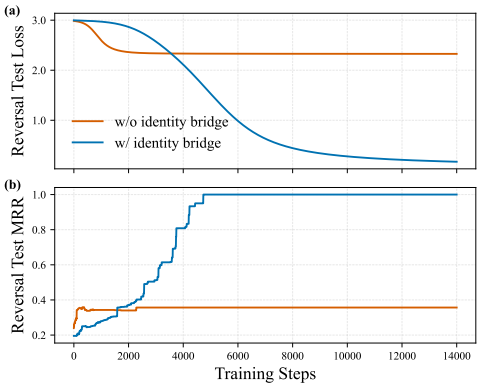


Figure 8: **Reversal test loss and mean reciprocal rank (MRR) of forward relation only training data vs. identity bridge regularized data.** Without the identity bridge, the model stays around the initialization level, while the model can generalize to all the reversal tests after adding the identity bridge regularization.

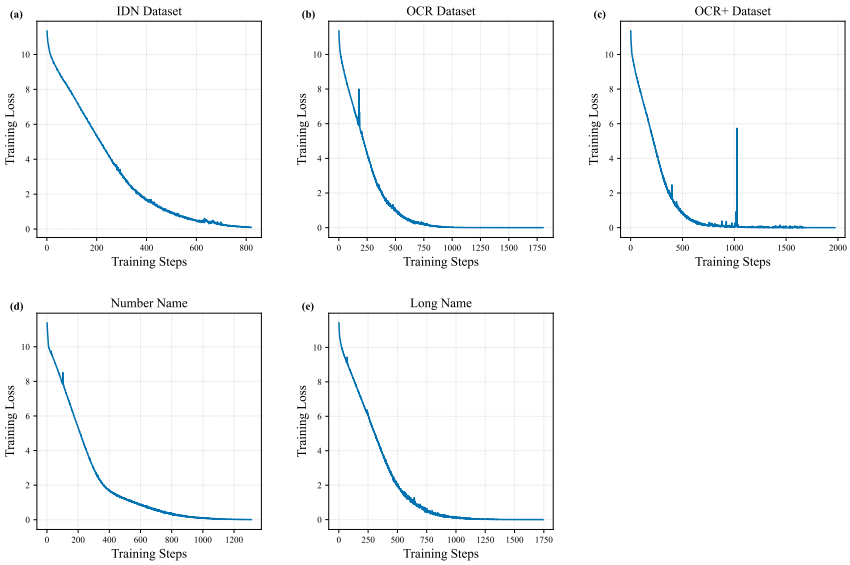


Figure 9: Training loss of different datasets and name lengths.

C.2 Experiment details of Section 4.2

For real LLM experiments, we use the cross-entropy loss and optimize it with the AdamW optimizer (Kingma, 2014). In finetuning, we use a temperature of 40, and search the best learning rate among $\{5 \cdot 10^{-5}, 10^{-4}, 2 \cdot 10^{-4}, 4 \cdot 10^{-4}\}$. The batch size is set to $1/5$ of the total data size. We also adopt a weight decay of 0.3 during training.

The training loss of the “Husband-Wife" task is presented in Figure 9, and the evolution of the corresponding reversal test first-token MRR is shown in Figure 10.

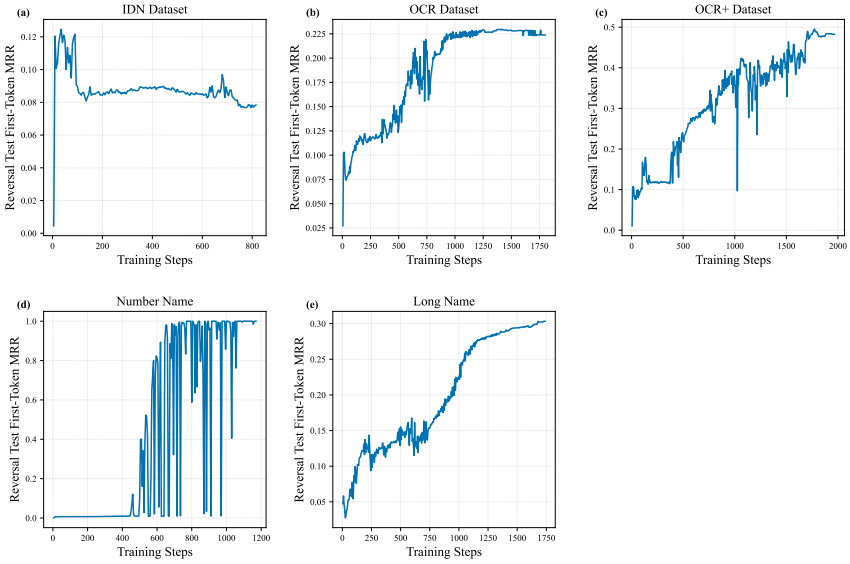


Figure 10: Reversal test first-token MRR of different datasets and name lengths.