

Unveiling the Mechanisms of Multi-Hop Reasoning in Transformers via Identity Bridge

Pengxiao Lin^{1,2,*} Zheng-An Chen^{1,*} Zhi-Qin John Xu^{1,2,3 †}

¹School of Mathematical Sciences, Shanghai Jiao Tong University

²Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University

³Shanghai Seres Information Technology Co., Ltd, Shanghai 200040, China.

Abstract

Large Language Models (LLMs) excel at multi-hop reasoning in distribution, yet fail on unseen compositions, a phenomenon known as the “curse of two-hop reasoning”. In this work, we argue that this phenomenon can be attributed to a missing supervision on the bridge entity. We formalize this gap by introducing identity bridge, a minimal supervision that enforces a identity mapping on bridge tokens. Under this supervision, even a one-layer transformer with uniform attention (Emb-MLP) can achieve out-of-distribution (OOD) two-hop generalization. We provide a theoretical analysis demonstrating that identity bridge induces an implicit regularization effect, leading the model to establish a direct subject-to-answer association. From an empirical perspective, the performance of standard GPT-2 models aligns closely with simple Emb-MLP models across varying levels of problem complexity. Finally, analyses of fine-tuned mainstream LLMs indicate that correct two-hop predictions consistently coincide with the establishment of a subject-to-answer relationship, extending our findings to realistic settings.

1 Introduction

Large language models (LLMs) achieve strong performance on a wide range of multi-step reasoning tasks, especially when assisted by chain-of-thought (CoT) [Wei et al. \(2022\)](#); [Yao et al. \(2023\)](#); [Kojima et al. \(2022\)](#). Despite their strong CoT-assisted performance, LLMs exhibit very fragile reasoning abilities when forced to produce the final answer directly [Bai et al. \(2025\)](#). On the one hand, models sometimes exhibit signatures consistent with latent multi-step processing, yet such evidence is suggestive rather than definitive and is often difficult to disentangle from shortcut strategies [Ding et al. \(2024\)](#); [Lin et al. \(2025\)](#). On the other hand, controlled synthetic settings reveal striking and systematic failures, including the reversal curse [Berglund et al. \(2024\)](#); [Allen-Zhu & Li \(2024; 2025\)](#); [Qi et al. \(2023\)](#) and the two-hop curse [Balesni et al. \(2025\)](#); [Dziri et al. \(2023\)](#). These phenomena suggest that seemingly minor design choices in synthetic data and training objectives can crucially interact with optimization-induced inductive biases, leading to brittle multi-hop inference. While recent work has offered partial analyses and empirical observations [Yang et al. \(2024\)](#); [Biran et al. \(2024\)](#); [Wang et al. \(2024\)](#); [Ye et al. \(2025\)](#); [Liu et al. \(2026\)](#), a theoretically grounded explanation of mechanisms behind multi-hop reasoning remains incomplete. This motivates the following questions.

1. Why is OOD compositional generalization in two-hop reasoning so brittle, and what overlooked factors cause this brittleness in a way that admits rigorous characterization?
2. When a transformer appears to generalize in two-hop settings, what mechanism underlies this behavior: a genuine two-step implicit inference, or a shortcut solution that mimics composition?

*Equal contribution.

†Corresponding author: xuzhiqin@sjtu.edu.cn.

In this work, we revisit the two-hop curse, where models fit two single-hop relations ($a \rightarrow b$ and $b \rightarrow c$ in the training set) yet fail to answer the composed query ($a \rightarrow c$ at evaluation), following the setup and definitions of Wang et al. (2024). We highlight an often overlooked factor that standard synthetic setups do not enforce the bridge entity to be identical in the input and the output. Without this constraint, training can fit the two single-hop tasks using disjoint features, causing composition to break down. We introduce identity bridge, a minimal fix that adds a zero-hop self-mapping for each bridge token, unlocking OOD two-hop generalization. This also clarifies a gap between synthetic two-hop settings and pretrained models. Pretraining effectively endows models with a latent form of identity bridging, such as the ability to copy or restate text, which can partially support composition.

To better understand the mechanism, we analyze a simplified Emb-MLP model, namely a one-layer transformer with uniform attention that has been widely used to study transformer reasoning Yao et al. (2025); Huang et al. (2025); Mahankali et al. (2024); Zhang et al. (2024a). We show that gradient-descent training in Emb-MLP exhibits an implicit nuclear-norm bias, favoring low-rank parameters that share structure across tasks, and we prove that this bias is sufficient to transfer from one-hop training to OOD two-hop generalization even in a single layer.

At the same time, the one-layer solution we uncover for two-hop reasoning is clearly a form of shortcut pattern, rather than the kind of step-by-step implicit reasoning. Therefore, we further investigated a high-complexity setting with more relationships and empirically demonstrated that the performance degradation of standard GPT-2 is similar to that of the Emb-MLP model, suggesting that Emb-MLP is a good mechanism proxy. We additionally observe the same degradation across LLaMA-style, Qwen3-style, and Gemma-style models: all three are near-perfect at complexity $C = 1$, but their OOD accuracy decreases steadily as the number of compositional branches grows. Moreover, we noticed the small initialization technique Zhang et al. (2024b; 2025); Yao et al. (2025) or weight decay is useful to improve OOD generalization in high-complexity regimes. Our analysis links this improvement to tight representation alignment across layers, thereby confirming and extending the explanation in Biran et al. (2024).

To sum up, our contribution can be summarized as follows.

1. We introduce the identity bridge, a minimal modification to the training data that adds a zero-hop task and reliably enables OOD two-hop composition in synthetic settings (Fig. 1).
2. Theoretically, We develop a uniform-attention theory (Theorems 1 and 2) showing how identity bridge induces memorization that links a to c , and why the standard setup without identity supervision fails.
3. Empirically, We show that Emb-MLP is a good proxy to standard transformer model, and that small initialization improves cross-layer alignment and OOD generalization (Fig. 4b). Controlled experiments with modern transformer models further show that the failure pattern persists across modern architectural choices (Fig. 1). We also provide evidence that modern LLMs exhibit widespread shortcut-like memorization even without explicit two-hop supervision (Fig. 3).

2 Related Work

Implicit reasoning failures on 2-hop data Allen-Zhu & Li (2025) studies a phenomenon in which transformers face difficulty in manipulating already learned knowledge. Press et al. (2022) constructs a two-hop reasoning task on a knowledge graph, demonstrating that transformers can generalize to in-distribution data through long-term training, a phenomenon known as grokking. However, out-of-distribution data cannot be generalized. Ye et al. (2025) further investigated this phenomenon, demonstrating that in-distribution generalization stems from the presence of bridge entities in the two-hop task in the training set, thereby inducing alignment. However, these works did not propose methods to alleviate the generalization difficulties of OOD.

Compositionality gap in LLMs A growing literature documents a robust compositionality gap in LLMs on multi-hop tasks. [Press et al. \(2022\)](#); [Xu et al. \(2024\)](#) report a substantial accuracy drop from single-hop to two-hop queries, which does not vanish with increased model scale. [Yang et al. \(2024\)](#) found limited evidence for implicit reasoning in large models by eliminating shortcuts, and [Yu \(2024\)](#) found a similar phenomenon in fine-tuning. [Kazemi et al. \(2023\)](#) also found that the model is more able to utilize popular knowledge rather than lesser-known knowledge during fine-tuning, which affects the model’s reasoning performance. These works have demonstrated that it is possible for models to exploit shortcuts instead of genuine inference. To mechanistically explain these failures, [Biran et al. \(2024\)](#) analyze intermediate transformer states via circuit methods and suggest that first-hop information may be formed too late to be utilized for composition. In contrast, [Liu et al. \(2026\)](#) challenge hop-aligned circuit explanations by uncovering layer-order inversion, where answer information can become decodable earlier than the bridge entity.

3 Preliminaries

3.1 two-hop Reasoning Task Setup

We study two-hop reasoning as the composition of two mappings over disjoint entity sets based on ([Wang et al., 2024](#); [Ye et al., 2025](#)). Let $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ denote the subject, bridge, and answer entity sets, respectively, and let R_1 and R_2 denote the first-hop and second-hop relation sets. We define two hop maps and associated two-hop compositional mapping as

$$g_1 : \mathcal{E}_1 \times R_1 \rightarrow \mathcal{E}_2, \quad g_2 : \mathcal{E}_2 \times R_2 \rightarrow \mathcal{E}_3 \quad g(e_1, r_1, r_2) := g_2(g_1(e_1, r_1), r_2).$$

Accordingly, the synthetic dataset contains the following three types of supervision: (i) first-hop facts $(e_1, r_1) \mapsto g_1(e_1, r_1)$, (ii) second-hop facts $(e_2, r_2) \mapsto g_2(e_2, r_2)$, and (iii) identity-bridge facts $e_2 \mapsto e_2$ for bridge entities $e_2 \in \mathcal{E}_2$. The OOD evaluation task consists of two-hop queries

$$(e_1, r_1, r_2) \mapsto g(e_1, r_1, r_2),$$

which are never provided as supervised training targets unless otherwise specified.

For intuition, one may think of $(a, r_1) \mapsto b$ as the first hop, $(b, r_2) \mapsto c$ as the second hop, and $(a, r_1, r_2) \mapsto c$ as the composed query. The identity bridge $b \mapsto b$ explicitly enforces alignment between the bridge token as an output of the first hop and as an input to the second hop.

Data complexity. We define the complexity of the synthetic two-hop task as the number of first-hop relation types. Concretely, we fix $N \in \mathbb{N}$ as the number of entities, and set $|\mathcal{E}_1| = |\mathcal{E}_3| = N$, $|R_2| = 1$, $|R_1| = C$, and partition the bridge set into C disjoint slices. This definition captures the branching factor of the compositional task: a larger complexity means that each subject participates in more possible two-hop compositions.

Each first-hop relation $r_{1,j}$ selects the j -th bridge slice, while the shared second-hop relation r_2 maps bridge entities to answer entities. Therefore, for each subject entity, different choices of $r_{1,j}$ induce C distinct two-hop compositions, so the dataset complexity is exactly C .

Given the symmetry between the two hops, we maintain an identical configuration for r_2 and focus our complexity analysis on r_1 . Unless otherwise specified, training uses only one-hop and identity-bridge supervision, while all two-hop queries are reserved for OOD evaluation.

3.2 Model Architecture

Let d_{vocab} , d_m , d_k denote the vocabulary size, embedding dimension, and query-key projection dimension, respectively. The input vocabulary size and output are denoted by $\mathcal{V}_{\text{in}}$ and $\mathcal{V}_{\text{out}}$. The embedding matrix is $E \in \mathbb{R}^{|\mathcal{V}_{\text{in}}| \times d_m}$, with row E_x corresponding to token $x \in \mathcal{V}_{\text{in}}$, and the projection matrix is $W_{\text{proj}} \in \mathbb{R}^{d_m \times |\mathcal{V}_{\text{out}}|}$.

Embedding-MLP (Emb-MLP). For tasks where attention serves only as an information-mixing mechanism, we adopt the Embedding-MLP model Yao et al. (2025); Huang et al. (2025), which can be viewed as a transformer layer with uniform attention. This formulation allows flexible handling of the input and output vocabularies. For a sequence $X = (x_1, \dots, x_s)$, we define:

Definition 1 (Embedding-MLP (Emb-MLP)). *Given parameters $\theta = (E, W_{\text{proj}})$, the model outputs logits is*

$$f_{\theta}(X) = \left(\sum_{t=1}^s E_{x_t} \right) W_{\text{proj}} \in \mathbb{R}^{|\mathcal{V}_{\text{out}}|}.$$

Modern decoder variants. To test whether the observed behavior depends on the GPT-2 architecture, we additionally evaluate three decoder-only variants: LLaMA-style (Pre-RMSNorm, SwiGLU feed-forward layers, and RoPE), Qwen3-style (Pre-RMSNorm, SwiGLU, RoPE, and grouped-query attention), and Gemma-style (Pre-RMSNorm, GeGLU, RoPE, and multi-query attention). Each model uses 12 layers, 12 query heads, $d_{\text{model}} = 768$, and $d_{\text{ff}} = 1024$. We train each setting for 20 epochs using a learning rate of 10^{-4} , batch size 500, weight decay 0.1, and cosine annealing with warmup, and report mean and standard deviation over three seeds.

For any learnable weight matrix $W \in \mathbb{R}^{d_1 \times d_2}$, with d_1 and d_2 denoting the input and output dimensions, we initialize each entry from a Gaussian distribution $W_{i,j} \sim \mathcal{N}(0, \sigma^2)$. The standard GPT-2 model take $\sigma = 0.02$. When we use a small initialization, we let $\sigma = d_1^{-\gamma}$, where $\gamma > 0.5$ since related work (Zhang et al., 2024b; Yao et al., 2025) shows that networks initialized in this way often have stronger reasoning capabilities.

4 Main Results

We identify an overlooked input-output inconsistency and introduce explicit bridging entity supervision, identity bridge, to recover implicit two-hop reasoning in transformer. To explain this, we analyze Emb-MLP logits and provide a rigorous theoretical account of the underlying mechanism. Finally, we extend these insights to two-hop reasoning in LLMs.

4.1 Motivation: the necessity of the identity bridge

We begin by illustrating the core intuition with a minimal example. The key observation is that, in the absence of additional constraints, the model does not naturally align its outputs with its inputs.

To make this issue concrete, we utilize activation patching Meng et al. (2022); Lv et al. (2024); Heimersheim & Nanda (2024) to examine the efficacy of bridge token b . We consider the GPT-2 model at complexity = 2. We feed the sequence (a, r_1) into the model and extract the hidden states at the r_1 position across different layers. These extracted states are then injected into the position of b within the sequence (b, r_2) , allowing the model to proceed with the subsequent inference. If the model recognizes that the extracted hidden state is equivalent to the input b , it should be able to further decode the corresponding answer $(b, r_2) \rightarrow c$. We report the activation patching accuracy in Fig. 1(left). The result shows that the model completely fails to decode c on OOD two-hop reasoning, which indicates token b doesn't bridge the gap between first hop and second hop.

We attribute this failure to a contextual disconnect between the input and output spaces. Under a simple two-hop task setup, the model is not explicitly required to establish an equivalence between the input token b and the output token b , which is a trivial capability for well-trained LLMs. Consequently, a straightforward solution to address this issue is to augment the training data with $b \rightarrow b$ "zero-hop" sequences, which we term as identity bridge.

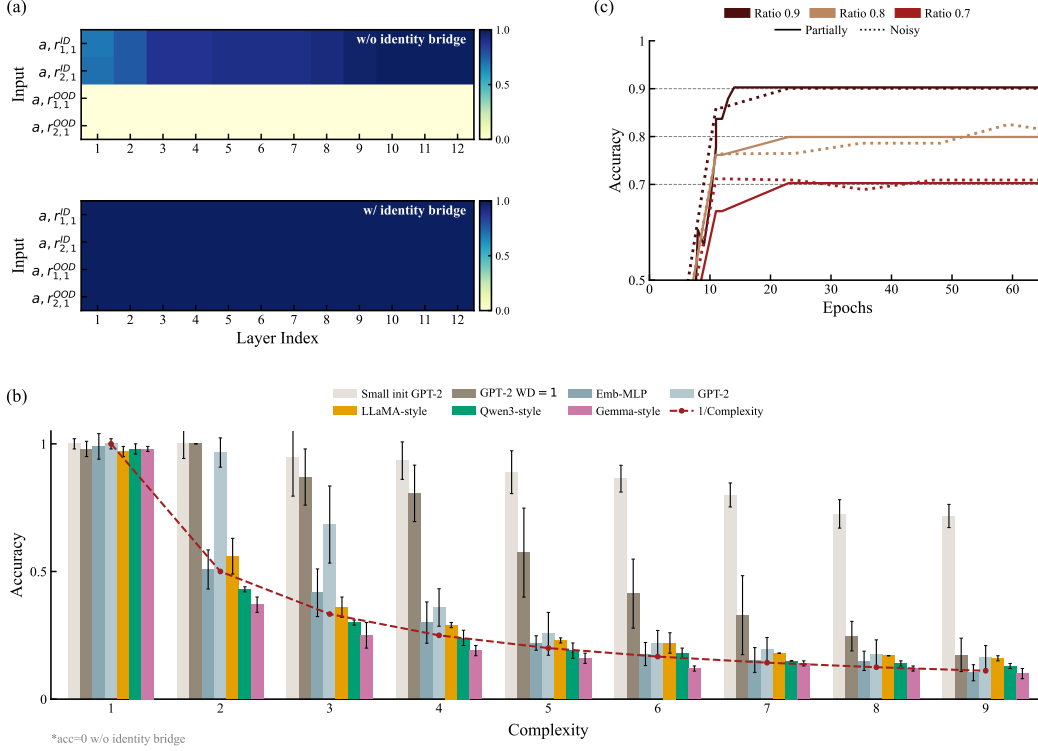


Figure 1: Left: Activation patching accuracy without (top) and with (bottom) the identity bridge. We report performance for both in-distribution (ID) relations (r^{ID}) and out-of-distribution (OOD) relations (r^{OOD}). Center: OOD two-hop accuracy from complexity $C = 1$ to $C = 9$ for Emb-MLP, standard and regularized GPT-2, and LLaMA-style, Qwen3-style, and Gemma-style models. For the three modern decoder variants, error bars denote standard deviation over three seeds. Right: The effect of retaining a certain proportion of identity bridge, while masking or adding label noise to the remaining. The model’s accuracy closely aligns with the proportion of retained data.

4.2 Identity bridge Works: Emb-MLP as a proxy for GPT-2

Fig. 1 reports test accuracy across models and complexity levels C . The identity bridge proves consistently effective, yielding non-zero OOD two-hop generalization for all models, irrespective of architectural nonlinearities or initialization schemes. Identity supervision suffices for the simplified Emb-MLP to achieve high performance, aligning with the implicit-regularization framework developed in Secs. 4.3 and 4.4.

The architecture comparison in the center panel strengthens this conclusion. We set the LLaMA-style, Qwen3-style, and Gemma-style models at the same size. At $C = 1$, models achieve nearly 1 OOD accuracy, respectively. As complexity increases to $C = 9$, their accuracies decrease. The shared trend across different architectures indicates that the phenomenon is robust to modern decoder design choices.

Notably, as complexity C increases, the accuracy of standard GPT-2 and Emb-MLP declines in tandem. This suggests that Emb-MLP effectively captures the mechanisms of the transformer, but also reveals that the implicit bias of gradient descent alone is insufficient for robust OOD composition at higher complexities. In contrast, the small-initialized GPT-2 degrades more gracefully, indicating a stronger regularization effect: the initialization-induced bias further constrains the latent geometry and better preserves the bridge-object coupling required for composition, a phenomenon we further discuss in Sec. 5.

4.3 Mechanism: how Emb-MLP solves two-hop reasoning

We now examine how identity bridge enables composition of one-hop tasks. Let the input and output vocabularies be $\mathcal{V}_{\text{in}} = \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{R}$ and $\mathcal{V}_{\text{out}} = \mathcal{E}_2 \cup \mathcal{E}_3$, respectively. To make the mechanism explicit, we use the Emb-MLP as the proxy model and analyze the row-wise logit templates encoded by

$$W = E W_{\text{proj}} \in \mathbb{R}^{|\mathcal{V}_{\text{in}}| \times |\mathcal{V}_{\text{out}}|},$$

where the i -th row of W is the logit vector produced by input token i . W_{ij} is the logit assigned by token i to output token j . Furthermore, since Emb-MLP is a uniform attention proxy, to see the model’s output logit for a sequence X , we only need to sum the corresponding rows of W . This facilitates our observation of the composite mechanism in two-hop tasks. We also drew the output logit of standard GPT-2 model to show a similar structure.

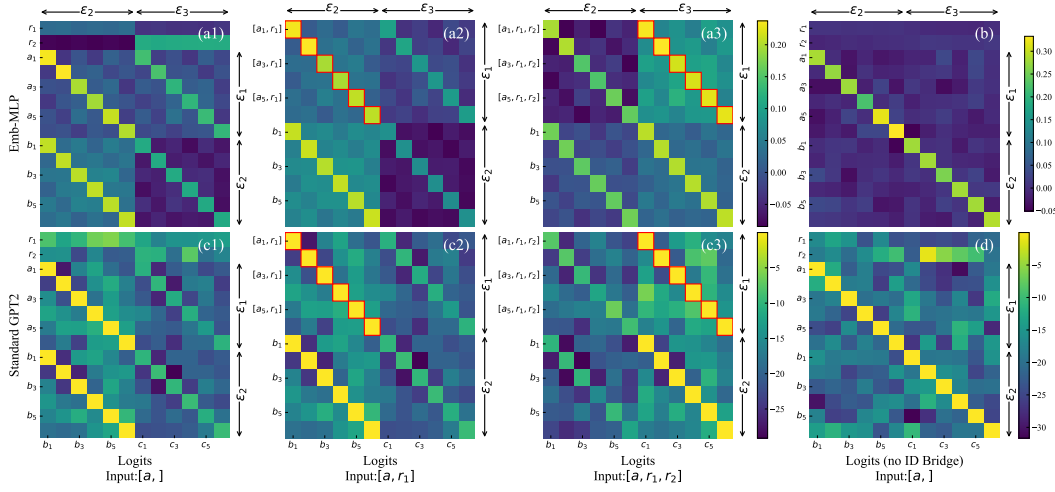


Figure 2: **Row-wise logit templates in Emb-MLP and comparison with Standard GPT-2 at complexity = 1 regime.** Panels (a1-a3) and (c1-c3) are trained with identity bridge; panel (b, d) omits it. **(a1)** Base logit matrix: the top two rows correspond to relation tokens r_1 and r_2 ; the remaining rows correspond to entity tokens in $\mathcal{E}_1$ and $\mathcal{E}_2$. **(a2)** One-hop input (a_i, r_1) , visualized as the row-wise sum of a_i and r_1 . **(a3)** Two-hop query (a_i, r_1, r_2) , visualized as the row-wise sum of a_i , r_1 , and r_2 . Red boxes indicate the current argmax output token. **(b, d)** Same visualization as in (a1) and (c1) but trained without identity supervision.

Fig. 2(a1) shows that relations act primarily as set selectors: r_1 boosts logits toward $\mathcal{E}_2$ while r_2 does the converse. Hence, the substantive computation of the two hops is carried by the entity rows of W . The key question is whether the subject rows for a_i encode a discriminative bias toward the correct tail c_i .

Fig. 2(a1) further indicates that each bridge token b_i is both self-peaked and object-aligned, exhibiting high logit on b_i and on its paired c_i . Training on (a_i, r_1) would only concentrate the subject’s logits on the corresponding bridging slices. However, under the implicit nuclear norm regularization introduced by gradient-based training, as discussed in Sec. 4.4, the lowest-rank way to satisfy all constraints shares this structure across blocks, effectively transferring the bridge’s object-aligned peak to the subject rows. Consequently, the rows associated with a_i inherit a tail-directed bias via their linkage to b_i , supplying the cross-task memory needed for composition. Figs. 2(a2) and 2(a3) show that the model then completes both the one-hop task and the two-hop generalization by combining subject entities with relations.

In contrast, without identity, non-label logits within a block tend to equalize, yielding a nearly diagonal-dominant pattern that conveys little information about the correct object c_i for a given subject a_i , thereby leading to failure of two-hop reasoning.

To further demonstrate that Emb-MLP is a good approximation of the standard transformer model, we visualized the logit of the standard GPT-2 model. The logit structures for single tokens, one-hop data, and two-hop data all show consistency with Emb-MLP, indicating that although there are differences between the Emb-MLP model and the transformer model, the Emb-MLP model reproduces mechanism of the transformer well in this task.

4.4 Theory: uniform-attention analysis of two-hop generalization

We analyze how identity bridge enables two-hop generalization in the Emb-MLP model on the dataset with complexity one, where attention acts only as uniform mixing. Previous experimental evidence has shown that this model contains similar mechanisms to the standard GPT-2 model.

Before presenting the main results, we introduce some necessary concepts and related results. For a labeled example (X, y) with $y \in \mathcal{V}_{\text{out}}$, define the pairwise logit gap and multiclass margin by

$$s_{(X,y),y'} = f_{\theta}(X)_y - f_{\theta}(X)_{y'} \quad q(X, y) = \min_{y' \in \mathcal{V}_{\text{out}} \setminus \{y\}} s_{(X,y),y'}. \quad (1)$$

Because Emb-MLP model is positively homogeneous in its parameters, and under standard separability with cross-entropy training, the normalized direction $\theta / \|\theta\|_2$ converges to a KKT point of the margin-maximization program established by [Lyu & Li \(2019\)](#):

$$\min_{\theta} \frac{1}{2} \|\theta\|_2^2, \quad \text{s.t. } s_{(X,y),y'} \geq 1, \quad \forall (X, y) \in \mathcal{D}_{\text{train}}, \quad \forall y' \in \mathcal{V}_{\text{out}} \setminus \{y\}. \quad (2)$$

In the Emb-MLP model, the logit matrix W can be viewed as a reparameterization of E and W_{proj} , which leads to the following equivalent convex reformulation in W introduced by [Huang et al. \(2025\)](#):

$$\min_W \frac{1}{2} \|W\|_*^2, \quad \text{s.t. } s_{(X,y),y'} \geq 1, \quad \forall (X, y) \in \mathcal{D}_{\text{train}}, \quad \forall y' \in \mathcal{V}_{\text{out}} \setminus \{y\}. \quad (3)$$

Problem (3) is equivalent to problem (2), as established by Lemma 1 in [Huang et al. \(2025\)](#), whose proof follows an idea similar to that of [Recht et al. \(2010\)](#). We briefly sketch the main argument here and refer readers to [Huang et al. \(2025\)](#) for full details. For Emb-MLP, the objective of problem (2) takes the form

$$\min_{E, W_{\text{proj}}} \frac{1}{2} \left(\|E\|_F^2 + \|W_{\text{proj}}\|_F^2 \right).$$

The equivalence follows from the well-known variational characterization linking the nuclear norm and the Frobenius norm,

$$\|W\|_*^2 = \min_{\{EW_{\text{proj}}=W\}} \frac{1}{2} \left(\|E\|_F^2 + \|W_{\text{proj}}\|_F^2 \right)$$

which completes the argument.

This equivalent convex reformulation greatly facilitates the analysis. The problem is convex, as the objective function is convex and the constraints are affine. We therefore work with problem (3) to derive margin-based consequences for two-hop generalization.

We now state the formal consequences for two-hop generalization.

Theorem 1 (Positive OOD margin with identity supervision). *Given Assumption 1 and large N . Assume training includes zero-hop identity supervision over $\mathcal{E}_2$, and let W^* (equivalently θ^*) solve problem (3). Then for every OOD query $X = (a_i, r_1, r_2)$ with label $y = c_i$, the multiclass margin is positive:*

$$q(X, y) > 0.$$

Hence, the composed mapping $g_2 \circ g_1$ is recovered in the OOD two-hop task.

Proof sketch. We first use the permutation symmetry of the dataset to get a highly structured optimal solution and obtain a closed-form objective function in this structure. By further

utilizing symmetry, we proved a set of symmetry conditions (Proposition 2) that an optimal solution must satisfy, further simplifying the problem. Next, using the construction method and proof by contradiction, we strengthened the conditions (Proposition 5 and Corollary B.1) that the optimal solution must satisfy. These links shrink the OOD margin check to a one-dimensional inequality that feasibility makes strictly positive, yielding a positive margin on every two-hop query and the success of OOD generalization. See appendix B.3 for details of the proof.

Theorem 2 (Failure without identity supervision). *Given Assumption 1. If identity supervision is omitted, any problem of (3) satisfies, for each OOD query $X = (a_i, r_1, r_2)$ with label $y = c_i$, the multiclass margin is negative:*

$$q(X, y) < 0.$$

Thus, the composed mapping fails on the OOD two-hop task.

Proof sketch. Without identity supervision, the training constraints are perfectly symmetric inside each block. Because the objective is convex and permutation-invariant, we can average any optimal solution over all within-block permutations and obtain an equally optimal, fully symmetric one. A short KKT check then shows the non-label logits equalize within blocks, so subjects carry no preference toward their true objects. When we test a held-out two-hop composition, the signal from the subject does not point to the correct tail and the margin becomes negative. Hence the margins are negative for all OOD two-hop queries and composition fails. See Appendix B.4 for details of the proof.

4.5 Real Task Analysis

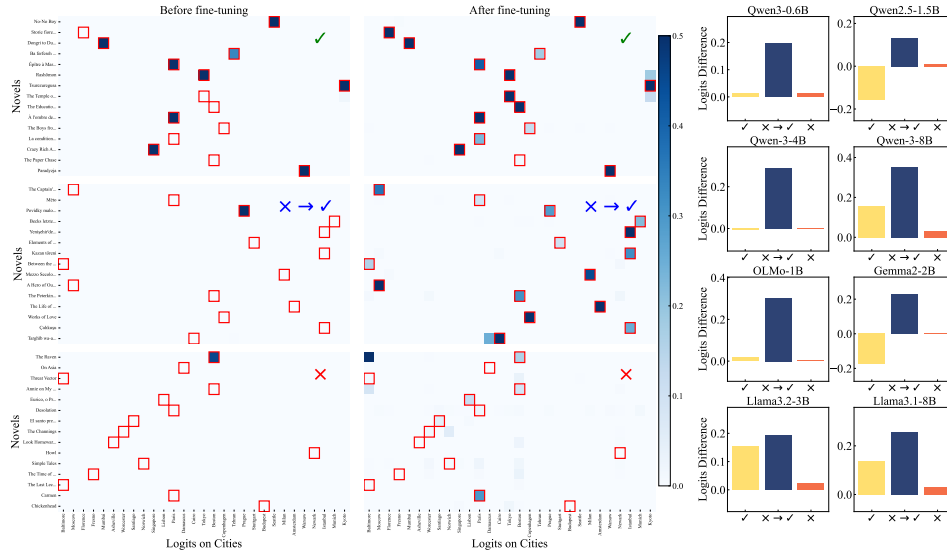


Figure 3: The probability of the model outputting the alternative city token when using the prompt corresponding to (e_1, r_2) before and after fine-tuning on datasets. The red box indicates the answer to the corresponding two-hop reasoning data. The bar chart shows the change in the probability of the corresponding two-hop answer when using the prompt (e_1, r_2) for different datasets.

In this section, we consider the mechanism of two-hop reasoning in real large models. As pointed out in Press et al. (2022), despite the existence of a combinatorial gap, real large models still have two-hop reasoning capabilities. We finetune pretrained models on TWOHOPFACT dataset Yang et al. (2024) to verifying our theoretical results. We filtered the data based on the bridge entity to ensure the OOD characteristics of the test data.

LLMs learn identity bridge in pretraining We perform a repeat task to monitor for identity bridges that may appear in the language model. Conceptually, the core requirement

for the identity bridge is the model’s ability to map the input and output of the bridge token b . Thus, we use the prompt “Repeat the following token $\{b\}$ directly” with 1-shot example to analyze whether the model possesses an identity bridge-like capability during pretraining. We utilize the Olmo2-0425 [OLMo et al. \(2024\)](#) version checkpoints for test and results are shown in Fig. 6.

LLMs learn $a \rightarrow c$ after fine-tuning For large models, since the parameters have been fully pre-trained, the alignment phenomenon is unlikely to be observed from the hidden state. The model should basically rely on the implicit regularization induced by the gradient descent algorithm to complete the task. In order to observe this mechanism, we extract the following three datasets based on the two-hop results for analysis: $\mathcal{D}_{\text{correct}}$ consists of data with correct two-hop reasoning whether fine-tuning or not, $\mathcal{D}_{\text{partial}}$ contains the data that are wrong in the first two hops of fine-tuning but correct after fine-tuning, and $\mathcal{D}_{\text{incorrect}}$ contains the data of fine-tuning the errors of the two-hop task.

As Fig. 3 shown, after training on the single-hop task, even without seeing the corresponding two-hop data, for the correct data after training, when we use prompts such as (e_1, r_2) , such as “the novel was born in the city of” but omit the “author” relation, the model still establishes a strong correlation between the subject and the object, suggesting that the model actually implicitly establishes the identity bridge during the pre-training process which utilizes similar implicit regularization.

We calculated the probability change of labels corresponding to two-hop data using this type of prompt. The results from different models consistently show that completing two-hop reasoning depends on improving the probability of the subject to the object. For the data that still got it wrong after fine-tuning, we found that the probability of the corresponding object was slightly improved. The experiments on other tasks such as MuSiQue dataset [Trivedi et al. \(2022\)](#) could be referred at Sec. C.3.

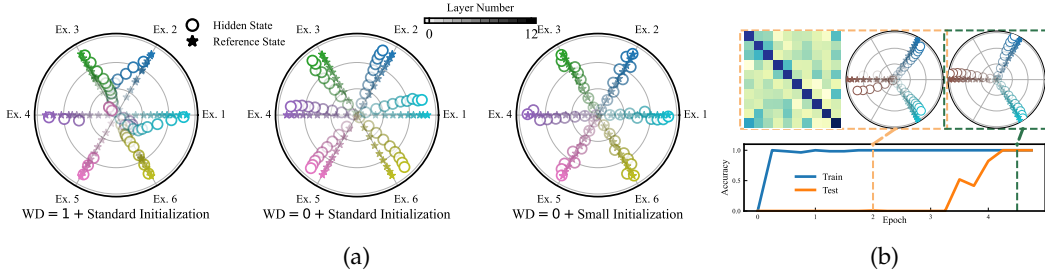


Figure 4: (a) T-SNE visualization of hidden states with respect to one-hop data and corresponding bridge entity. Six samples are selected for each setting, where the star represents the hidden state of the bridge entity and the circle represents the hidden state of the first hop data (e_1, r_i) at position r_i . Colors from light to dark represent hidden state from shallow to deep. (b) Alignment analysis along the training steps for small initialized GPT-2 with complexity= 2. The two images correspond to when the memorization on 1-hop ends but OOD generalization ability is missing, and the moment when OOD generalization is completed. Figure (a1) and (b1) shows the relationship between the input a_i and the output logits of c_i , similarly to Fig. 2.

5 Another direction: strengthening regularization and alignment

As dataset complexity increases, the implicit regularization is no longer sufficient to solve two-hop reasoning task since relying solely on identity bridge mechanism. We therefore strengthen regularization either by small initialization or by weight decay. Both interventions substantially recover OOD performance.

To further investigate whether the model truly learns implicit reasoning rather than merely adopting shortcut patterns, we analyze the alignment of hidden states at the last token

position across different layers using t-SNE dimensionality reduction for inputs (e_1, r_1) and e_2 . As illustrated in Figure 4(a), taking the hidden state of e_2 as a baseline, we observe that the hidden states of (e_1, r_1) increasingly approach those of e_2 with large weight decay or small initialization. This trend suggests that the network progressively learns to associate the (e_1, r_1) pair with the bridge entity e_2 . In contrast, standard GPT-2 models do not exhibit such alignment, indicating that without the aid of proper regularization, the model fails to recognize the intrinsic connection between entities.

Furthermore, as Fig. 4b(b) indicates, the emergence of identity bridge effects precedes the growth of generalization ability. With the alignment of hidden states improves during training, the test accuracy rises in step with this alignment trend. In other words, when the representations for one-hop data and the target bridge collapse into the same subspace, the second hop can reliably latch onto the correct features and composition succeeds.

6 Discussion

Conclusion In this work, we revisit the two-hop curse phenomenon. We show that a minimal modification to the training data, the identity bridge, is sufficient to unlock OOD two-hop composition. Through a rigorous analysis of a simplified Emb-MLP model, we prove that identity bridge provably improves OOD two-hop generalization, while its absence inevitably leads to failure.

Empirically, we find that Emb-MLP closely approximates the behavior of standard GPT-2, indicating that standard transformers also tend to solve two-hop tasks via shortcut rather than genuine step-by-step reasoning. Experiments with modern transformer models reproduce the same accuracy-complexity trade-off, broadening the evidence beyond GPT-2. Experiments on real language tasks suggest that pretraining already endows modern LLMs with implicit identity-bridge-like capabilities and help establish a subject-to-answer relationship. To move toward more faithful implicit reasoning, we investigate additional regularization methods.

Limitations Our theoretical analysis is conducted within a simplified Emb-MLP model with a uniform attention mechanism and does not explicitly capture the full attention dynamics of transformers. Furthermore, it focuses on two-hop reasoning in synthetic environments. Nevertheless, experiments show that this model tracks GPT-2 well across various complexity levels and reproduces its success and failure patterns with and without identity supervision. The additional architecture comparison reduces, but does not eliminate, this limitation: the same trend holds for three modern decoder variants, while substantially larger pretrained models and non-decoder architectures remain outside the controlled analysis.

Acknowledgements

This work is sponsored by the National Natural Science Foundation of China Grant No. 12371511, 12422119, 2025 Key Technology R&D Program “New Generation Information Technology” Project 25511103100 of Shanghai Municipal Science and Technology Commission and the Fundamental Research Funds for the Central Universities (project number YG2024ZD03)

References

- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML/24. JMLR.org, 2024.
- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=oDbil9CLoS>.

- Xiaoyan Bai, Itamar Pres, Yuntian Deng, Chenhao Tan, Stuart Shieber, Fernanda Viégas, Martin Wattenberg, and Andrew Lee. Why can't transformers learn multiplication? reverse-engineering reveals long-range dependency pitfalls, 2025. URL <https://arxiv.org/abs/2510.00184>.
- Mikita Balesni, Tomek Korbak, and Owain Evans. Lessons from studying two-hop latent reasoning, 2025. URL <https://arxiv.org/abs/2411.16353>.
- Lukas Berglund, Meg Tong, Maximilian 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". In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=GPKTIktA0k>.
- Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 14113–14130, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.781. URL <https://aclanthology.org/2024.emnlp-main.781/>.
- Mengru Ding, Hanmeng Liu, Zhizhang Fu, Jian Song, Wenbo Xie, and Yue Zhang. Break the chain: Large language models can be shortcut reasoners. *CoRR*, abs/2406.06580, 2024. doi: 10.48550/ARXIV.2406.06580. URL <https://doi.org/10.48550/arXiv.2406.06580>.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, and Jena D. Hwang. Faith and fate: Limits of transformers on compositionality. *ArXiv*, abs/2305.18654, 2023.
- Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL <https://arxiv.org/abs/2404.15255>.
- 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*, 2025.
- Mehran Kazemi, Sid Mittal, and Deepak Ramachandran. Understanding finetuning for factual knowledge extraction from language models. *CoRR*, abs/2301.11293, 2023. URL <https://doi.org/10.48550/arXiv.2301.11293>.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. Implicit reasoning in transformers is reasoning through shortcuts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9470–9487, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.493. URL <https://aclanthology.org/2025.findings-acl.493/>.
- Xukai Liu, Ye Liu, Jipeng Zhang, Yanghai Zhang, Kai Zhang, and Qi Liu. Layer-order inversion: Rethinking latent multi-hop reasoning in large language models. *arXiv preprint arXiv:2601.03542*, 2026.
- Ang Lv, Yuhan Chen, Kaiyi Zhang, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. Interpreting key mechanisms of factual recall in transformer-based language models, 2024. URL <https://arxiv.org/abs/2403.19521>.
- Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

- Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=8p3fu56lKc>.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.
- Chengwen Qi, Bowen Li, Binyuan Hui, Bailin Wang, Jinyang Li, Jinwang Wu, and Yuanjun Laili. An investigation of llms’ inefficacy in understanding converse relations. *arXiv preprint arXiv:2310.05163*, 2023.
- Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 2022.
- Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=D4QgSWxiOb>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
- Zhuoyan Xu, Zhenmei Shi, and Yingyu Liang. Do large language models have compositional ability? an investigation into limitations and scalability. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. URL <https://openreview.net/forum?id=4XPeF0SbJs>.
- Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*, 2024.
- Junjie Yao, Zhongwang Zhang, and Zhi-Qin John Xu. An analysis for reasoning bias of language models with small initialization, 2025. URL <https://arxiv.org/abs/2502.04375>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.
- Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Liu Weichuan, Xiaoyin Che, Lei Hou, and Juanzi Li. How do transformers learn implicit reasoning? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=19ygs48n0a>.

Yijiong Yu. Do llms really think step-by-step in implicit reasoning? 2024.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *J. Mach. Learn. Res.*, 25:49:1–49:55, 2024a. URL <https://jmlr.org/papers/v25/23-1042.html>.

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, and Zhi-Qin John Xu. Initialization is critical to whether transformers fit composite functions by reasoning or memorizing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL <https://openreview.net/forum?id=Y0BGdVaYTS>.

Zhongwang Zhang, Pengxiao Lin, Zhiwei Wang, Yaoyu Zhang, and Zhi-Qin John Xu. Complexity control facilitates reasoning-based compositional generalization in transformers. *arXiv preprint arXiv:2501.08537*, 2025.

A Notations and Definitions

We adopt the following notation conventions. All tokens in our synthetic tasks are encoded as positive integers. Calligraphic capital letters (e.g., $\mathcal{E}, \mathcal{R}$) denote sets of tokens, roman capital letters (e.g., N, C) denote scalar parameters such as dataset size and complexity, and lower-case letters denote individual tokens. The entity set $\mathcal{E}$ is partitioned into subject, bridge, and object subsets $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$, with typical elements $e_{1,\cdot} \in \mathcal{E}_1, e_{2,\cdot} \in \mathcal{E}_2$, and $e_{3,\cdot} \in \mathcal{E}_3$. The relation set $\mathcal{R}$ is split into first- and second-hop families $\mathcal{R}_1$ and $\mathcal{R}_2$, with elements $r_{1,\cdot} \in \mathcal{R}_1$ and $r_{2,\cdot} \in \mathcal{R}_2$. For any set S , we write $|S|$ or $\#S$ for its cardinality.

A.1 Two-Hop Reasoning Task Setup

To study the mechanism behind compositionality gap, we introduce the synthetic dataset in which all tokens are positive integers partitioned into a disjoint set of entities ($\mathcal{E}$) and relations ($\mathcal{R}$). Entities are split into subject, bridge, and object subsets:

$$\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3, \quad \text{with} \quad \mathcal{E}_i \cap \mathcal{E}_j = \emptyset \ (i \neq j).$$

Relations are divided into two disjoint families for the two hops:

$$\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2, \quad \text{with} \quad \mathcal{R}_1 \cap \mathcal{R}_2 = \emptyset.$$

One-hop tasks. We instantiate two one-hop tasks. For the first hop, we first partition the bridge entities $\mathcal{E}_2$ according to complexity determined by the number of relations in $\mathcal{R}_1$:

$$\mathcal{E}_2 = \bigcup_{j_1=1}^{|\mathcal{R}_1|} \mathcal{E}_{2,j_1}.$$

Then define a deterministic (or sampling) map

$$g_1 : \mathcal{E}_1 \times \mathcal{R}_1 \rightarrow \mathcal{E}_2 \quad \text{such that} \quad g_1(e_1, r_{1,j_1}) \in \mathcal{E}_{2,j_1}.$$

The first-hop triple set is

$$\mathcal{T}_1 = \{ (e_1, r_{1,j_1}, e_2) : e_1 \in \mathcal{E}_1, r_{1,j_1} \in \mathcal{R}_1, e_2 = g_1(e_1, r_{1,j_1}) \}.$$

This partitioning of $\mathcal{E}_2$ ensures that each $r_{1,j_1} \in \mathcal{R}_1$ only co-occurs with bridge entities from its dedicated slice $\mathcal{E}_{2,j_1}$, reducing spurious shortcuts across relations. Second-hop construction follows analogous principles.

Two-hop composition. A two-hop instance composes the two one-hop maps. Given $(e_1, r_{1,j_1}, e_2) \in \mathcal{T}_1$ and $(e_2, r_{2,j_2}, e_3) \in \mathcal{T}_2$, the composed query is $(e_1, r_{1,j_1}, r_{2,j_2})$ with answer

$$(e_1, r_{1,j_1}, r_{2,j_2}) = g_2(g_1(e_1, r_{1,j_1}), r_{2,j_2}) = e_3.$$

Identity Bridge. Unlike prior work, we also include a zero-hop task over bridge entities called identity bridge to establish the connection between two one-hop tasks and shape the model’s latent space. For each bridge entity $e_2 \in \mathcal{E}_2$ we add a training pair of the form

$$(e_2) \rightarrow e_2,$$

encouraging the model to implement an identity transformation $f(e_2) = e_2$ on the bridge entities. This task is not introduced for its standalone difficulty, but to regularize representations so that the composed mapping $g_2 \circ g_1$ can be more reliably recovered during two-hop generalization.

A.2 Dataset Setup

OOD two-hop reasoning. After determine the g_1 and g_2 maps, we can construct training dataset $\mathcal{D}_{\text{train}}$ and test dataset $\mathcal{D}_{\text{test}}$ where $\mathcal{D}_{\text{train}}$ contains all one-hop data and partial two-hop data and $\mathcal{D}_{\text{test}}$ contains only the two-hop data. To investigate OOD composition ability,

a two-hop (e_1, r_1, r_2) data is called out-of-distribution if the corresponding bridge entity e_2 has never appeared in the two-hop data of the training set. In this work, unless otherwise specified, the training set is restricted to contain only single-hop data, ensuring that all two-hop data are out-of-distribution.

Dataset Complexity By adjusting the configurations of g_1, g_2 , and the number of relations, we can control the complexity of the dataset. The dataset complexity is defined precisely as the number of object entities associated with each subject as follows:

$$\text{Complexity} = \max_{e_1 \in \mathcal{E}_1} \#\{e_3 = (e_1, r_{1,j_1}, r_{2,j_2}) \mid r_{1,j_1} \in \mathcal{R}_1, r_{2,j_2} \in \mathcal{R}_2\}.$$

To construct different complexity data, we instantiate families of datasets with controlled complexity. Fix $N \in \mathbb{N}$ and set $|\mathcal{E}_1| = |\mathcal{E}_3| = N$. Let $C \in \mathbb{N}$ denote the complexity parameter and take

$$|\mathcal{E}_2| = CN, \quad |\mathcal{R}_1| = C, \quad |\mathcal{R}_2| = 1.$$

Partition the bridge set evenly as

$$\mathcal{E}_2 = \bigcup_{j_1=1}^C \mathcal{E}_{2,j_1}, \quad \mathcal{E}_{2,j_1} = \{e_{2,k} : (j_1 - 1)N + 1 \leq k \leq j_1 N\}.$$

Write $\mathcal{R}_1 = \{r_{1,1}, \dots, r_{1,C}\}$ and $\mathcal{R}_2 = \{r_2\}$. The hop maps are defined by

$$g_1(e_{1,i_1}, r_{1,j_1}) = e_{2,(j_1-1)N+i_1}, \quad g_2(e_{2,k}, r_2) = e_{3, \lfloor \frac{k}{N} \rfloor + k \bmod N},$$

so that each $r_{1,j}$ selects the j -th bridge slice and r_2 collapses the bridge index modulo N onto $\mathcal{E}_3$. When $C = 1$, we construct the structured setting used for analysis:

$$\mathcal{E}_1 = \{a_1, \dots, a_N\}, \quad \mathcal{E}_2 = \{b_1, \dots, b_N\}, \quad \mathcal{E}_3 = \{c_1, \dots, c_N\}, \quad \mathcal{R}_1 = \{r_1\}, \quad \mathcal{R}_2 = \{r_2\},$$

with one-to-one correspondences $g_1(a_{i_1}, r_1) = b_{i_1}$ and $g_2(b_{i_2}, r_2) = c_{i_2}$, hence the composed answer for (a_{i_1}, r_1, r_2) is c_{i_1} .

B Theoretical Details

This appendix completes the proofs of Theorems 1 and 2. We first collect several standard tools and assumptions from the literature that will be used throughout the arguments, and then give the proofs via a detailed analysis of the nuclear-norm program—combining a constructive step with a contradiction argument.

B.1 Preliminaries for KKT Conditions

First we recall the notion of subdifferentials for convex functions and then state the KKT conditions in subdifferential form.

Let $\varphi : \mathbb{R}^d \rightarrow (-\infty, +\infty]$ be a proper convex function. The subdifferential of φ at $x \in \mathbb{R}^d$ is defined by

$$\partial\varphi(x) := \left\{s \in \mathbb{R}^d : \varphi(y) \geq \varphi(x) + \langle s, y - x \rangle \quad \forall y \in \mathbb{R}^d\right\}.$$

Consider the following optimization problem (P) for $x \in \mathbb{R}^d$:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g_n(x) \geq 0 \quad \forall n \in [N], \\ & h_m(x) = 0 \quad \forall m \in [M], \end{aligned}$$

where f, g_n, h_m are convex and finite-valued on their domains. We say that $x \in \mathbb{R}^d$ is a feasible point of (P) if x satisfies $g_n(x) \geq 0$ for all $n \in [N]$ and $h_m(x) = 0$ for all $m \in [M]$.

Definition 2 (KKT point). A feasible point x of (P) is called a KKT point if there exist multipliers $\lambda_1, \dots, \lambda_N \geq 0$ and $\mu_1, \dots, \mu_M \in \mathbb{R}$ such that

1. (Subdifferential stationarity) There exist subgradients $\mathbf{s}_f \in \partial f(\mathbf{x})$, $\mathbf{s}_{g_n} \in \partial g_n(\mathbf{x})$ for all $n \in [N]$, and $\mathbf{s}_{h_m} \in \partial h_m(\mathbf{x})$ for all $m \in [M]$ such that

$$\mathbf{s}_f - \sum_{n=1}^N \lambda_n \mathbf{s}_{g_n} - \sum_{m=1}^M \mu_m \mathbf{s}_{h_m} = \mathbf{0}.$$

Equivalently,

$$0 \in \partial f(\mathbf{x}) - \sum_{n=1}^N \lambda_n \partial g_n(\mathbf{x}) - \sum_{m=1}^M \mu_m \partial h_m(\mathbf{x}).$$

2. (Complementary slackness) For all $n \in [N]$,

$$\lambda_n g_n(\mathbf{x}) = 0.$$

In general, global minimizers of (P) do not necessarily satisfy the KKT conditions. However, under suitable regularity assumptions (such as a Slater-type constraint qualification), the KKT conditions become necessary for optimality. To ensure the validity and clarity of the subsequent analysis, we impose the following standard assumptions.

Next we recall a standard constraint qualification for convex optimization, known as Slater's condition, which guarantees that KKT conditions are satisfied at optimal solutions.

Definition 3 (Slater's condition). Consider problem (P) above and assume that f and g_n are convex functions on $\mathbb{R}^d$ for all $n \in [N]$, and each h_m is an affine function on $\mathbb{R}^d$ for all $m \in [M]$. We say that Slater's condition holds for (P) if there exists a point $\bar{\mathbf{x}} \in \mathbb{R}^d$ such that

$$g_n(\bar{\mathbf{x}}) > 0 \quad \forall n \in [N], \quad h_m(\bar{\mathbf{x}}) = 0 \quad \forall m \in [M].$$

Such a point $\bar{\mathbf{x}}$ is called a strictly feasible point.

Under Slater's condition, the KKT conditions become necessary (and, in the convex setting, also sufficient) for optimality.

Proposition 1 (KKT conditions under Slater's condition). Suppose that (P) is a convex optimization problem in the sense that f and g_n are convex and h_m are affine, and that Slater's condition holds. Let $\mathbf{x}^*$ be a global minimizer of (P). Then $\mathbf{x}^*$ is a KKT point; that is, there exist multipliers $\lambda_1^*, \dots, \lambda_N^* \geq 0$ and $\mu_1^*, \dots, \mu_M^* \in \mathbb{R}$ such that $\mathbf{x}^*$ and (λ^*, μ^*) satisfy the subdifferential KKT conditions above. Conversely, any KKT point of (P) is a global minimizer.

In other words, for convex problems satisfying Slater's condition, the KKT conditions provide an exact characterization of optimal solutions.

B.2 Auxiliary Results from the Literature

We restate the external lemmas and assumptions needed in our proofs, in formulations specialized to our notation. Proofs are omitted and can be found in the cited references.

Lemma 1 (Existence of a restricted form solution to (3), Huang et al. (2025) Lemma 3). Suppose $\mathbf{W}$ is the solution to the optimization problem (3) with identity task. There exists a solution with parameter $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, e, f, g, h$ such that

$$\mathbf{W}^\top = \begin{pmatrix} a_1 \mathbf{I}_n + a_2 \mathbf{E}_n & b_1 \mathbf{I}_n + b_2 \mathbf{E}_n & e \mathbf{1}_n & f \mathbf{1}_n \\ c_1 \mathbf{I}_n + c_2 \mathbf{E}_n & d_1 \mathbf{I}_n + d_2 \mathbf{E}_n & g \mathbf{1}_n & h \mathbf{1}_n \end{pmatrix}. \quad (4)$$

The parameters follow the following constraints:

$$\begin{aligned} a_1, d_1 &\geq 1, \\ a_1 + a_2 + e &\geq c_1 + c_2 + g + 1, \\ d_1 + d_2 + h &\geq b_1 + b_2 + f + 1. \end{aligned} \quad (5)$$

In addition, after introducing the identity mapping, the problem gains additional constraints:

$$\begin{aligned} b_1 &\geq 1, \\ b_1 + b_2 &\geq d_1 + d_2 + 1. \end{aligned} \quad (6)$$

To analyze the optimal solution of optimization problem 3, we need an explicit formula of the nuclear norm. After block multiplication, $\mathbf{W}^\top \mathbf{W}$ can be written as:

$$\mathbf{W}^\top \mathbf{W} = \begin{pmatrix} C_{A1} \mathbf{I}_n + C_{A2} \mathbf{E}_n & C_{B1} \mathbf{I}_n + C_{B2} \mathbf{E}_n \\ C_{B1} \mathbf{I}_n + C_{B2} \mathbf{E}_n & C_{D1} \mathbf{I}_n + C_{D2} \mathbf{E}_n \end{pmatrix}, \quad (7)$$

where the coefficients are:

$$\begin{aligned} C_{A1} &= a_1^2 + b_1^2 \\ C_{A2} &= 2a_1a_2 + na_2^2 + 2b_1b_2 + nb_2^2 + e^2 + f^2 \\ C_{D1} &= c_1^2 + d_1^2 \\ C_{D2} &= 2c_1c_2 + nc_2^2 + 2d_1d_2 + nd_2^2 + g^2 + h^2 \\ C_{B1} &= a_1c_1 + b_1d_1 \\ C_{B2} &= a_1c_2 + a_2c_1 + na_2c_2 + b_1d_2 + b_2d_1 + nb_2d_2 + eg + fh. \end{aligned}$$

The proof of Lemma 1 is provided in Huang et al. (2025). Readers may also consult the proof of Lemma 4, as the underlying ideas are closely related. Through direct calculation, we have an explicit expression for the nuclear norm of $\mathbf{W}$. We introduce the following notation to simplify the formula.

$$\mathbf{u} = (a_1 + na_2, b_1 + nb_2, \sqrt{ne}, \sqrt{nf})^\top, \quad \mathbf{v} = (c_1 + nc_2, d_1 + nd_2, \sqrt{ng}, \sqrt{nh})^\top \quad (8)$$

and

$$A = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}, \quad B = (\mathbf{u}, \mathbf{v}) \quad (9)$$

Lemma 2 (Closed-form of $\|\mathbf{W}\|_*$ under the restricted form). *Let $\mathbf{W}$ be in the restricted form Eq. (4). Then its nuclear norm admits the closed form*

$$\begin{aligned} \|\mathbf{W}\|_* &= (n-1) \sqrt{C_{A1} + C_{D1} + 2|a_1d_1 - b_1c_1|} \\ &\quad + \sqrt{C_{A1} + nC_{A2} + C_{D1} + nC_{D2} + 2\sqrt{(C_{A1} + nC_{A2})(C_{D1} + nC_{D2}) - (C_{B1} + nC_{B2})^2}}. \end{aligned} \quad (10)$$

In particular, Eq. (10) can be reformulated as

$$\|\mathbf{W}\|_* = (n-1)\|A\|_* + \|B\|_*. \quad (11)$$

Proof. Eq. (10) follows directly from the Lemma 6 in Huang et al. (2025). Therefore, we focus on Eq. (11). For the first term, we expand the expression according to the definition.

$$\sqrt{C_{A1} + C_{D1} + 2|a_1d_1 - b_1c_1|} = \sqrt{a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2|a_1d_1 - b_1c_1|} \quad (12)$$

Then, we check it is just $\|A\|_*$. Let $X \in \mathbb{R}^{n \times 2}$ and σ_1, σ_2 be singular values, we get $\|X\|_* = \sigma_1 + \sigma_2$. We use the following identity

$$\sigma_1 + \sigma_2 = \sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2}. \quad (13)$$

Moreover, we use

$$\sigma_1^2 + \sigma_2^2 = \|X\|_F^2, \quad \sigma_1\sigma_2 = \sqrt{\det(X^\top X)} \quad (14)$$

Substitute A into above equation, we get the result. The proof for the second term is similar. $\square$

To fully characterize the properties of the optimal solution to the optimization problem in (3), the mere existence of a solution is insufficient. To address this limitation, Huang et al. (2025) introduces the following assumption and establishes the uniqueness of the solution.

Assumption 1. *Suppose that $\mathbf{W}$ is a solution to optimization problem (3) with (or without) identity mapping, but takes a different form from (4) (or (62)). Then we assume that $\mathbf{1}_{2n}^\top \mathbf{W}^\top \neq \mathbf{0}_{2n+2}^\top$.*

Based on Assumption 1, the uniqueness theorem is stated as follows.

Theorem 3 (Uniqueness, rephrased from Huang et al. (2025) Theorem 5). *Given Assumption 1, the solution of optimization problem (3) with (or without) identity mapping has restricted form (4) (or (62)).*

B.3 Proof for Theorem 1

We start with the following Proposition which characterize the optimal solution of optimization problem (3) with restricted form.

Proposition 2. *Suppose that W is a optimal solution of optimization problem (3) with restricted form (4). Then, W satisfies $\mathbf{1}_{2n}^\top W^\top = \mathbf{0}_{2n+2}^\top$.*

Proof. Using the notation introduced in Eqs. (8) and (9) and the Lemma 2, the optimization problem can be reformulated as

$$\begin{aligned} \min_{a_i, b_i, c_i, d_i, e, f, g, h} \quad & (n-1)\|A\|_* + \|B\|_* \\ \text{s.t.} \quad & g_1 : a_1 \geq 1, \\ & g_2 : b_1 \geq 1, \\ & g_3 : d_1 \geq 1, \\ & g_4 : a_1 + a_2 + e \geq c_1 + c_2 + g + 1, \\ & g_5 : b_1 + b_2 \geq d_1 + d_2 + 1, \\ & g_6 : d_1 + d_2 + h \geq b_1 + b_2 + f + 1. \end{aligned} \tag{15}$$

We assert that the optimal solution to the optimization problem satisfies the KKT conditions. To this end, we only need to verify the convexity and the Slater condition.

Convexity. Since the nuclear norm is convex, the $\|A\|_*$ is convex with respect to a_1, b_1, c_1, d_1 . For the term $\|B\|_*$, it is also convex since it is a composition of convex function and affine transformation. Moreover, the optimization problem is convex since the objective function is convex and the constraints are affine.

Slater condition. We can construct a solution which is strictly feasible. Here, we give an example. Let $a_1 = 2, b_1 = 3, c_1 = 0.5, d_1 = 1.5, a_2 = 0, b_2 = 0, c_2 = 0, d_2 = 0, e = 0, g = 0, f = 0, h = 4$. Through direct testing, we can see that this solution is strictly feasible. Thus, we can use the KKT condition to characterize the optimal solution. Observing that $\|A\|_*$ is independent of $a_2, b_2, c_2, d_2, e, f, g, h$, we consider the stationary condition with respect to them since the symmetry of B with respect to u and v will be helpful. Using chain rule, we get

$$\begin{aligned} (a_2) : 0 &\in n(\partial_u \|B\|_*)_1 - \lambda_4, & (c_2) : 0 &\in n(\partial_v \|B\|_*)_1 + \lambda_4 \\ (b_2) : 0 &\in n(\partial_u \|B\|_*)_2 - \lambda_5 + \lambda_6, & (d_2) : 0 &\in n(\partial_v \|B\|_*)_2 + \lambda_5 - \lambda_6 \\ (e) : 0 &\in (\partial_u \|B\|_*)_3 - \lambda_4, & (g) : 0 &\in (\partial_v \|B\|_*)_3 + \lambda_4, \\ (f) : 0 &\in (\partial_u \|B\|_*)_4 + \lambda_6, & (h) : 0 &\in (\partial_v \|B\|_*)_4 - \lambda_6, \end{aligned} \tag{16}$$

Utilizing the symmetry, we find that there exist $s_u \in \partial_u \|B\|_*$ and $s_v \in \partial_v \|B\|_*$ such that

$$s_u + s_v = 0. \tag{17}$$

Next, we proceed by case analysis.

1. $u = v = 0$. In this case, we get $f = h = 0$ by the definition of u and v . Then the inequality constraints g_5 and g_6 will be reformulated as

$$b_1 + b_2 \geq d_1 + d_2 + 1, \quad d_1 + d_2 \geq b_1 + b_2 + 1. \tag{18}$$

It makes a contradiction and implies that $u = v = 0$ is not a solution.

2. $u = 0$ or $v = 0$. We just need to consider the $u = 0$ case and the proof of $v = 0$ case is similar. First, we recall the definition of the subderivative of nuclear norm. Let $X \in \mathbb{R}^{m \times n}$ with singular value decomposition $X = U\Sigma V^\top$, the subderivative of nuclear norm with respect to X is

$$\partial \|X\|_* = \{UV^\top + M : U^\top M = 0, MV = 0, \|M\|_2 \leq 1\} \tag{19}$$

Substitute $B = (0, v)$ into the definition, we find

$$v^\top M = 0, \quad Me_2 = 0. \tag{20}$$

It implies that

$$\partial\|B\|_* = \{(\mathbf{w}, \mathbf{v}) : \mathbf{w} \perp \mathbf{v}, \|\mathbf{w}\| \leq 1\}. \quad (21)$$

Then the condition (17) implies that there exists $\mathbf{w}$ such $\mathbf{w} + \mathbf{v} = 0$. It makes a contradiction since $\mathbf{w} \perp \mathbf{v}$.

3. $\mathbf{u} \neq 0$ and $\mathbf{v} \neq 0$. We will further discuss two scenarios in this case.

(a) $\mathbf{u}$ and $\mathbf{v}$ are linearly independent. In this case, the subgradient equals the gradient because there are no singularities. By direct computation, we get

$$\begin{aligned} \nabla_{\mathbf{u}}\|B\|_* &= \frac{1}{\|B\|_*} \left(\mathbf{u} + \frac{\|\mathbf{v}\|_2^2 \mathbf{u} - \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{v}}{\sqrt{\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2}} \right) \\ \nabla_{\mathbf{v}}\|B\|_* &= \frac{1}{\|B\|_*} \left(\mathbf{v} + \frac{\|\mathbf{u}\|_2^2 \mathbf{v} - \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u}}{\sqrt{\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2}} \right) \end{aligned} \quad (22)$$

Thus, the condition (17) implies that

$$\frac{1}{\|B\|_*} \left(1 + \frac{\|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle}{\sqrt{\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2}} \right) \mathbf{u} + \frac{1}{\|B\|_*} \left(1 + \frac{\|\mathbf{u}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle}{\sqrt{\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2}} \right) \mathbf{v} = 0. \quad (23)$$

Since $\mathbf{u}$ and $\mathbf{v}$ are linearly independent, the coefficient must be zero. It implies

$$\sqrt{\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2} = \langle \mathbf{u}, \mathbf{v} \rangle - \|\mathbf{v}\|^2 \quad \text{and} \quad \sqrt{\|\mathbf{u}\|_2^2 \|\mathbf{v}\|_2^2 - \langle \mathbf{u}, \mathbf{v} \rangle^2} = \langle \mathbf{u}, \mathbf{v} \rangle - \|\mathbf{u}\|^2 \quad (24)$$

which implies $\|\mathbf{u}\|_2^2 = \|\mathbf{v}\|_2^2$. Then by Cauchy-Schwarz inequality, we get $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|$. However, we find $\langle \mathbf{u}, \mathbf{v} \rangle \geq \|\mathbf{u}\| \|\mathbf{v}\|$ by Eq. (24) and $\|\mathbf{u}\| = \|\mathbf{v}\|$. As a result, we find $\mathbf{u}$ is parallel to $\mathbf{v}$ which makes a contradiction.

(b) $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent. In this case, we prove $\mathbf{u} = -\mathbf{v}$ which finishes the proof. Let $\mathbf{v} = \lambda \mathbf{u}$ and the problem turns into whether $\lambda = -1$. By the definition of the subgradient of the nuclear norm, we find

$$\partial\|B\|_* = \{\mathbf{u}(1, \lambda) + M : \mathbf{u}^\top M = 0, M(1, \lambda)^\top = 0, \|M\|_2 \leq 1\}. \quad (25)$$

Let $M = (\mathbf{w}_1, \mathbf{w}_2)$, we will find

$$\mathbf{u}^\top \mathbf{w}_1 = 0, \quad \mathbf{u}^\top \mathbf{w}_2 = 0 \quad (26)$$

Based on condition (17), we have

$$\mathbf{u}(1 + \lambda) + (\mathbf{w}_1 + \mathbf{w}_2) = 0. \quad (27)$$

Since $\mathbf{u}$ is perpendicular $\mathbf{w}_1, \mathbf{w}_2$, we get $1 + \lambda = 0$ and finish the proof.

□

Using Proposition 2, the nuclear norm minimization problem (3) can be simplified to a more tractable form. The following lemma provides this equivalent formulation.

Lemma 3. The optimization (3) problem is equivalent to the following optimization problem.

$$\begin{aligned} \min_{a_i, b_i, c_i, d_i, e, f, g, h} \quad & (n-1)\sqrt{M_1} + \sqrt{2M_2} \\ \text{s.t.} \quad & a_1, b_1, d_1 \geq 1, \\ & a_1 + a_2 + e \geq c_1 + c_2 + g + 1, \\ & b_1 + b_2 \geq d_1 + d_2 + 1, \\ & d_1 + d_2 + h \geq b_1 + b_2 + f + 1, \\ & a_1 + c_1 = -n(a_2 + c_2), \quad b_1 + d_1 = -n(b_2 + d_2), \\ & e = -g, \quad f = -h, \end{aligned} \quad (28)$$

where the terms M_1 and M_2 are defined as

$$\begin{aligned} M_1 &= a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2|a_1d_1 - b_1c_1|, \\ M_2 &= (a_1 + na_2)^2 + (b_1 + nb_2)^2 + ne^2 + nf^2. \end{aligned}$$

Proof. The proof consists of two main steps. First, we establish several key identities among the problem's coefficients that arise from Proposition 2. Second, we substitute these identities into the nuclear norm expression from (10) to derive the simplified objective function.

Step 1: Deriving Identities from Proposition 2. The four equality constraints presented in the lemma statement are a direct consequence of the structural properties imposed by Proposition 2. Their derivation involves straightforward algebraic manipulation and is omitted for brevity.

These equalities lead to two crucial identities for the aggregated coefficients. First, we show that $C_{A1} + nC_{A2} = C_{D1} + nC_{D2}$. By expanding the terms, we have:

$$\begin{aligned} C_{A1} + nC_{A2} &= a_1^2 + b_1^2 + n(2a_1a_2 + na_2^2 + 2b_1b_2 + nb_2^2 + e^2 + f^2) \\ &= (a_1 + na_2)^2 + (b_1 + nb_2)^2 + ne^2 + nf^2. \end{aligned}$$

Using the equality constraints like $a_1 + c_1 = -n(a_2 + c_2)$, the expression above is equivalent to $(c_1 + nc_2)^2 + (d_1 + nd_2)^2 + ng^2 + nh^2$, which is precisely the expansion of $C_{D1} + nC_{D2}$.

Second, we analyze the cross-term $C_{B1} + nC_{B2}$:

$$\begin{aligned} C_{B1} + nC_{B2} &= a_1c_1 + b_1d_1 + n(a_1c_2 + a_2c_1 + na_2c_2 + b_1d_2 + b_2d_1 + nb_2d_2 + eg + fh) \\ &= (a_1 + na_2)(c_1 + nc_2) + (b_1 + nb_2)(d_1 + nd_2) + neg + nfh. \end{aligned}$$

Applying the equality constraints again, this simplifies to:

$$C_{B1} + nC_{B2} = -(a_1 + na_2)^2 - (b_1 + nb_2)^2 - ne^2 - nf^2 = -M_2.$$

Step 2: Simplifying the Nuclear Norm. The second term of original objective function can be simplified based on extra equality constraints. We introduce coefficients $A = C_{A1} + nC_{A2}$, $D = C_{D1} + nC_{D2}$, and $B = C_{B1} + nC_{B2}$. From Step 1, we have established that $A = D = M_2$ and $B = -M_2$.

Consequently, the discriminant term $AD - B^2$ becomes:

$$AD - B^2 = (M_2)(M_2) - (-M_2)^2 = M_2^2 - M_2^2 = 0.$$

When the discriminant is zero, the original nuclear norm expression (likely involving square roots of eigenvalues) simplifies significantly, yielding the objective function stated in (28). The remaining constraints are carried over directly, which completes the proof. $\square$

We restate the optimization problem after reformulate and label each constraint to prepare for the optimal solution later.

$$\begin{aligned} \min_{a_i, b_i, c_i, d_i, e, f, g, h} \quad & (n-1)\sqrt{a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2|a_1d_1 - b_1c_1|} \\ & + \sqrt{2}\sqrt{(a_1 + na_2)^2 + (b_1 + nb_2)^2 + ne^2 + nf^2}, \end{aligned} \quad (29)$$

The inequality constraints are

$$\begin{aligned} g_1 : a_1 - 1 &\geq 0, \\ g_2 : a_1 + a_2 + 2e - c_1 - c_2 - 1 &\geq 0, \\ g_3 : b_1 - 1 &\geq 0, \\ g_4 : b_1 + b_2 - d_1 - d_2 - 1 &\geq 0, \\ g_5 : d_1 - 1 &\geq 0, \\ g_6 : d_1 + d_2 - b_1 - b_2 - 2f - 1 &\geq 0. \end{aligned}$$

The equality constraints are

$$\begin{aligned} h_1 : a_1 + c_1 + n(a_2 + c_2) &= 0, \\ h_2 : b_1 + d_1 + n(b_2 + d_2) &= 0. \end{aligned}$$

To use KKT condition to characterize the solution, we can check that the optimization problem (29) is a convex problem and satisfies the Slater condition which is similar to the proof in Proposition 2. Specific details are omitted here.

Before formally proving Theorem 1, we will prove several propositions that will be helpful in the proof.

Proposition 3. *Every optimal solution of reformulated optimization problem (29) satisfy:*

$$a_1 + na_2 = e, \quad c_1 + nc_2 = -e \quad (30)$$

in which $e \geq 0$.

Proof. Using stationarity property for parameter c_2, a_2 and e and the fact $u \neq 0$, we find that the optimal solution satisfies

$$c_2: \quad 0 + \lambda_2 - \mu_1 n = 0 \quad \Rightarrow \quad \lambda_2 = \mu_1 n. \quad (S1)$$

$$a_2: \quad \sqrt{2} \frac{n(a_1 + na_2)}{\sqrt{M_2}} - \lambda_2 - \mu_1 n = 0. \quad (S2)$$

$$e: \quad \sqrt{2} \frac{ne}{\sqrt{M_2}} - 2\lambda_2 = 0. \quad (S3)$$

Substitute Eq. (S1) into Eqs. (S2) and (S3) respectively, we get

$$a_1 + na_2 = e.$$

Another equality comes from the equality constraint $a_1 + c_1 + n(a_2 + c_2) = 0$. Moreover, $e \geq 0$ from Eq. (S3) and $\lambda_2 \geq 0$. $\square$

Proposition 4. *For every optimal solution of reformulated optimization problem (29), at least one of the following two inequality constraints is tight:*

$$\begin{aligned} g_1 : a_1 - 1 &\geq 0, \\ g_2 : a_1 + a_2 + 2e - c_1 - c_2 - 1 &\geq 0. \end{aligned} \quad (31)$$

Proof. We prove by contradiction. We show that if both two constraints are not tight, there is a feasible perturbation of this optimal point such that object value is strictly smaller. We take

$$\Delta a_1 = -\alpha\epsilon, \quad \Delta b_1, \Delta b_2, \Delta d_1, \Delta d_2 = 0, \quad \Delta e, \Delta f = 0 \quad (32)$$

By proposition 3, we take the following variables as

$$\Delta a_2 = \frac{\alpha}{n}\epsilon, \quad \Delta c_1 = -\gamma\epsilon, \quad \Delta c_2 = \frac{\gamma}{n}\epsilon. \quad (33)$$

Next we show that by choosing appropriate parameters α and γ , we can construct a feasible descent direction.

Step 1: Descent direction.

To prove that this direction is actually a descent direction, we consider the first-order approximation of the object function. First, We find that $\Delta M_2 = 0$ by our choice. Then we consider case by case:

1. $a_1 d_1 - b_1 c_1 = 0$. In this case, we let $-\alpha\epsilon d_1 - b_1(-\gamma\epsilon) = 0$ which implies $\alpha d_1 = b_1 \gamma$. As a result, we have

$$\begin{aligned} \Delta M_1 &= 2a_1 \Delta a_1 + 2c_1 \Delta c_1 \\ &= -2\alpha\epsilon(a_1 + \frac{c_1 d_1}{b_1}) \end{aligned} \quad (34)$$

It is a descent direction since $c_1 = \frac{a_1 d_1}{b_1} > 0$ and we can take $\alpha > 0$.

2. $a_1d_1 - b_1c_1 > 0$. In this case $M_1 = a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2(a_1d_1 - b_1c_1)$. We take $\gamma = 0$ and α sufficiently small and get $\Delta M_1 = -2(a_1 + d_1)\alpha\epsilon$ which implies that it is a descent direction.
3. $a_1d_1 - b_1c_1 < 0$. In this case, we have $M_1 = a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2(b_1c_1 - a_1d_1)$ and $c_1 > 0$. We take γ sufficiently small and $\alpha = 0$. It is a descent direction.

Step 2: Feasible direction.

For inequality constraints g_1 and g_2 , we can take ϵ small enough to ensure that they can't hit the boundary. Inequality constraints g_3 to g_6 and equality constraint h_2 still hold since the relevant variables have not changed. Equality constraint h_1 still hold due to our choice for a_1, a_2, c_1, c_2 . As a result, the construction is a feasible direction. □

Then, we claim that both two inequality constraints must be tight by further analysis. Before we prove this result, we first prove the following proposition.

Proposition 5. *There exists $N > 0$ such that, for all $n > N$, every optimal solution of reformulated optimization problem (29) satisfy $c_1^* > 0$.*

Proof. For the case of $a_1d_1 - b_1c_1 \leq 0$, the result is trivial since $c_1 \geq \frac{a_1d_1}{b_1} > 0$. And for the case of $a_1d_1 - b_1c_1 > 0$, we discuss case by case using Proposition 4.

1. g_1 is tight, but g_2 is not. Based on complementary slackness of KKT condition, we know that $\lambda_2 = 0$. However, based on Eq. (S1), we get

$$\mu_1 = \frac{\lambda_2}{n} = 0.$$

Then, using the KKT condition with respect to c_1 :

$$(n-1)\frac{c_1 - b_1}{\sqrt{M_1}} + \lambda_2 - \mu_1 = 0. \quad (35)$$

It implies that $c_1 = b_1 > 0$.

2. g_2 is tight, but g_1 is not. First, we consider the perturbation of the optimal solution. To maintain the constraints, we let

$$\Delta a_1 = \epsilon, \Delta a_2 = -\frac{\epsilon}{n}, \Delta c_1 = \epsilon, \Delta c_2 = -\frac{\epsilon}{n}. \quad (36)$$

It is a feasible direction and we consider

$$\Delta M_1 = 2a_1\epsilon + 2c_1\epsilon + 2d_1\epsilon - 2b_1\epsilon. \quad (37)$$

It implies that

$$a_1 + c_1 + d_1 - b_1 = 0. \quad (38)$$

Otherwise, it will be a descent direction. Using Proposition 3, we get

$$a_2 = \frac{e - a_1}{n}, c_2 = \frac{-e - c_1}{n}.$$

Combine this with the constraint g_2 , we get

$$\frac{n-1}{n}a_1 + \frac{2n+2}{n}e - \frac{n-1}{n}c_1 = 1. \quad (39)$$

Then, we have

$$1 \leq a_1 \leq \frac{n}{n-1}, 0 \leq e \leq \frac{1}{2n+2}, -\frac{1}{n-1} \leq c_1 \leq 0. \quad (40)$$

Using Eq. (38), we have

$$b_1 - d_1 = a_1 + c_1 \geq 1 - \frac{1}{n-1} \quad (41)$$

We consider

$$\begin{aligned} M_1 &= a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2(a_1d_1 - b_1c_1) \\ &\geq 1 + (2 - \frac{1}{n-1})^2 + 1 + 2 = 8 - O(\frac{1}{n}). \end{aligned} \quad (42)$$

Thus the leading term of objective function in this case is at least $2n\sqrt{2}$. Therefore, if a specific construction can be given whose objective function is strictly smaller than the above cases, the proof is complete. The specific construction is given below. We let

$$a_1 = 1, b_1 = 2, c_1 = 0.5, d_1 = 1, e = 0.5, f = -1. \quad (43)$$

And parameters are yet to be determined. Due to our choice, we check constraints g_2, g_4, g_6 and h_1, h_2 to determine them.

$$\begin{aligned} g_2 : a_2 - c_2 + 0.5 &\geq 0 \\ g_4 : b_2 - d_2 &\geq 0 \\ g_6 : d_2 - b_2 &\geq 0 \\ h_1 : 1.5 + n(a_2 + c_2) &= 0 \\ h_2 : 3 + n(b_2 + d_2) &= 0. \end{aligned} \quad (44)$$

Then we take

$$a_2 = c_2 = -\frac{3}{4n}, b_2 = d_2 = -\frac{3}{2n} \quad (45)$$

Then we compute the objective function

$$M_1 = 6.25, \quad M_2 = 0.75 + 1.25n \quad (46)$$

As a result, the objective function

$$(n-1)\sqrt{6.25} + \sqrt{2}\sqrt{3.75 + 1.25n} < 2\sqrt{2}n. \quad (47)$$

Thus, for large n , it makes a contradiction.

3. g_1 is tight and g_2 is tight. We consider the following perturbation:

$$\Delta a_1 = 0, \Delta c_1 = \varepsilon, \Delta a_2 = 0, \Delta c_2 = -\frac{\varepsilon}{n}, \Delta e = \frac{\varepsilon - \frac{1}{n}\varepsilon}{2} \quad (48)$$

Then we will find that

$$\Delta g_1 = 0, \Delta g_2 = 2\Delta e - \Delta c_1 - \Delta c_2 = 0, \Delta h_1 = 0. \quad (49)$$

Then we consider the change of the objective function:

$$\Delta((n-1)\sqrt{M_1}) = (n-1)\frac{c_1 - b_1}{\sqrt{M_1}}\varepsilon \leq -(n-1)\frac{1}{\sqrt{M_1}}\varepsilon \leq -\frac{n-1}{2\sqrt{2}}\varepsilon \quad (50)$$

Here we use $\sqrt{M_1} \leq 2\sqrt{2}$. If not, we can construct the solution with smaller leading term like the proof in case 2 which implies that the solution is not optimal. Then, we consider the second term:

$$\Delta(\sqrt{2}\|u\|_2) = \sqrt{2}\frac{ne}{\|u\|}\Delta e \quad (51)$$

Using inequality constraint g_4 and g_6 , we get $f \leq -1$ which implies $\|u\| \geq \sqrt{n}$. Moreover, we get $e \leq \frac{1}{2(n+1)}$ because of $c_1 \leq 0$ and g_1, g_2 are tight. As a result, we have

$$\Delta(\sqrt{2}\|u\|) \leq \frac{\sqrt{2}}{4} \frac{n-1}{(n+1)\sqrt{n}}\varepsilon \quad (52)$$

It implies that

$$\Delta((n-1)\sqrt{M_1} + \sqrt{2}\|u\|) \leq -\frac{n-1}{2\sqrt{2}}\varepsilon + O(n^{-\frac{1}{2}})\varepsilon < 0. \quad (53)$$

Thus, we have finished the proof.

□

Corollary B.1. *For all optimal solution of reformulated optimization problem, the inequality constraint g_2 must be tight.*

Proof. According to Proposition 4, we consider the case g_1 is tight but g_2 is not. Based on complementary slackness of KKT condition, we know that $\lambda_2 = 0$. However, based on Eq. (S1), we get

$$\mu_1 = \frac{\lambda_2}{n} = 0.$$

Furthermore, due to Eq. (S2), we get

$$e = \frac{n\sqrt{2}}{\sqrt{B}}(\lambda_2 + \mu_1 n) = 0$$

However, it makes an contradiction since the following constraints are violated due to Proposition 5:

$$a_1 + a_2 + 2e - c_1 - c_2 - 1 = 1 - \frac{1}{n} - c_1 + \frac{1}{n}c_1 - 1 < 0. \quad (54)$$

□

we now give the proof of Theorem 1.

Proof of Theorem 1. Thanks to Proposition 5, to prove $q(X, y) > 0$ for all OOD query $(X, y) = ((a_i, r_1, r_2), c_i)$, we just need to show that

$$c_1 + c_2 + g + h > a_1 + a_2 + e + f. \quad (55)$$

This is because

$$\begin{aligned} s_{(X,y),b_j} &= c_1 + c_2 + g + h - \max\{a_1, 0\} - a_2 - e - f, \quad \forall j \in [N] \\ s_{(X,y),c_j} &= c_1, \quad \forall j \neq i. \end{aligned} \quad (56)$$

Here, $c_1 > 0$ by Proposition 5. So we only need to prove $s_{(X,y),b_j} > 0$. Using the constraints $e + g = 0$ and $f + h = 0$, inequality (55) can be reformulated as

$$c_1 + c_2 - (a_1 + a_2) > 2e + 2f. \quad (57)$$

Utilizing Corollary B.1, The left side of the inequality is simplified to

$$\begin{aligned} c_1 + c_2 - (a_1 + a_2) &= a_1 + a_2 + 2e - 1 - (a_1 + a_2) \\ &= 2e - 1. \end{aligned} \quad (58)$$

As a result, inequality (57) holds if and only if

$$f < -\frac{1}{2}. \quad (59)$$

However, we get a better upper bound by combining inequality constraints g_4 with g_6

$$\begin{aligned} b_1 + b_2 &\geq d_1 + d_2 + 1 \\ &\geq b_1 + b_2 + 2f + 2, \end{aligned} \quad (60)$$

which implies that

$$f \leq -1. \quad (61)$$

As a result, inequality (57) holds which implies $q(X, y) > 0$ for all OOD query. □

B.4 Proof for Theorem 2

We begin our proof with the following lemma, which makes fuller use of the symmetry of the problem.

Lemma 4 (Existence of symmetry solution). *Suppose $\mathbf{W}$ is the solution to the optimization problem 3 without identical task. There exists a solution with a_1, a_2, b_1, b_2 and α, β such that*

$$\mathbf{W}^\top = \begin{pmatrix} a_1 \mathbf{I}_n + a_2 \mathbf{E}_n & b_1 \mathbf{I}_n + b_2 \mathbf{E}_n & \alpha \mathbf{1}_n & \beta \mathbf{1}_n \\ b_1 \mathbf{I}_n + b_2 \mathbf{E}_n & a_1 \mathbf{I}_n + a_2 \mathbf{E}_n & \beta \mathbf{1}_n & \alpha \mathbf{1}_n \end{pmatrix}. \quad (62)$$

The parameters follow the following constraints:

$$\begin{aligned} a_1 &\geq 1, \\ a_1 + a_2 + \alpha &\geq b_1 + b_2 + \beta + 1. \end{aligned} \quad (63)$$

Proof. We show that some orthogonal transformation of $\mathbf{W}^\top$ remains a solution to the optimization problem. Let σ be an arbitrary permutation of $1, \dots, n$, and let $\mathbf{P}_\sigma \in \mathbb{R}^{n \times n}$ denote the associated permutation matrix. Now, consider a permutation of the logit matrix.

$$\sigma(\mathbf{W}^\top) = \begin{pmatrix} \mathbf{P}_\sigma & 0 \\ 0 & \mathbf{P}_\sigma \end{pmatrix} \mathbf{W}^\top \text{diag}\{\mathbf{P}_\sigma, \mathbf{P}_\sigma, 1, 1\}.$$

We find that $\sigma(\mathbf{W}^\top)$ is still an optimal solution of the optimization problem. To verify this, we consider

$$\begin{aligned} s_{(a_i, r_1), b_j}(\sigma(\mathbf{W}^\top)) &= s_{(a_{\sigma^{-1}(i)}, r_1), b_{\sigma^{-1}(j)}}(\mathbf{W}^\top) \geq 1, \quad \forall j \in [n] - \{i\}, \\ s_{(a_i, r_1), c_j}(\sigma(\mathbf{W}^\top)) &= s_{(a_{\sigma^{-1}(i)}, r_1), c_{\sigma^{-1}(j)}}(\mathbf{W}^\top) \geq 1, \quad \forall j \in [n], \\ s_{(b_i, r_1), b_j}(\sigma(\mathbf{W}^\top)) &= s_{(b_{\sigma^{-1}(i)}, r_1), b_{\sigma^{-1}(j)}}(\mathbf{W}^\top) \geq 1, \quad \forall j \in [n], \\ s_{(b_i, r_1), c_j}(\sigma(\mathbf{W}^\top)) &= s_{(b_{\sigma^{-1}(i)}, r_1), c_{\sigma^{-1}(j)}}(\mathbf{W}^\top) \geq 1, \quad \forall j \in [n] - \{i\}. \end{aligned}$$

Moreover, $\sigma(\mathbf{W}^\top)$ is another solution since orthogonal transformation does not change the nuclear norm. Consider the average over all possible permutations:

$$\frac{\sum_{\sigma} \sigma(\mathbf{W}^\top)}{n!} = \begin{pmatrix} a_1 \mathbf{I}_n + a_2 \mathbf{E}_n & b_1 \mathbf{I}_n + b_2 \mathbf{E}_n & e \mathbf{1}_n & f \mathbf{1}_n \\ c_1 \mathbf{I}_n + c_2 \mathbf{E}_n & d_1 \mathbf{I}_n + d_2 \mathbf{E}_n & g \mathbf{1}_n & h \mathbf{1}_n \end{pmatrix}.$$

We further exploit the symmetry and consider the following transformation

$$\tau(\mathbf{W}^\top) = \begin{pmatrix} 0 & \mathbf{I} \\ \mathbf{I} & 0 \end{pmatrix} \mathbf{W}^\top \text{diag} \left\{ \begin{pmatrix} 0 & \mathbf{I} \\ \mathbf{I} & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}.$$

Similar verification shows that $\tau(\mathbf{W}^\top)$ is still the solution to the optimization problem. Taking the sum of $\mathbf{W}^\top$ and $\tau(\mathbf{W}^\top)$, we finish the proof. $\square$

We have $\mathbf{1}^\top \mathbf{W}^\top = 0$ similar to the proof of Proposition 2 and consider the problem

$$F = (n-1) \sqrt{2(a_1^2 + b_1^2) + 2|a_1^2 - b_1^2|} + 2 \sqrt{(a_1 + na_2)^2 + n\alpha^2} \quad (64)$$

subject to

$$\begin{cases} a_1 \geq 1, \\ a_1 + a_2 + \alpha \geq b_1 + b_2 - \alpha + 1, \\ a_1 + b_1 + n(a_2 + b_2) = 0. \end{cases} \quad (65)$$

To prove Theorem 2, we just need to prove the following condition holds for the optimal solution of optimization problem (64):

$$b_1 + b_2 < a_1 + a_2 \quad (66)$$

Proof of Theorem 2. Step 0: Structural simplification. Using the identity

$$a^2 + b^2 + |a^2 - b^2| = 2 \max\{a^2, b^2\},$$

the first square root in Eq. (64) reduces to

$$\sqrt{2(a_1^2 + b_1^2) + 2|a_1^2 - b_1^2|} = 2 \max\{|a_1|, |b_1|\}.$$

Since $a_1 \geq 1 > 0$, the objective becomes

$$F = 2(n-1) \max\{a_1, |b_1|\} + 2\sqrt{X^2 + n\alpha^2}, \quad X = a_1 + na_2.$$

Meanwhile, the linear constraint rewrites as

$$X = -(b_1 + nb_2).$$

Step 1. Necessary condition at optimum. Fixing (a_1, a_2) (so X is fixed), we can vary (b_1, b_2) while preserving $b_1 + nb_2$; this leaves X and the second term unchanged. If $|b_1| > a_1$, then decreasing $|b_1|$ towards a_1 reduces the first term and relaxes or maintains the inequality constraint, hence cannot be optimal. Therefore,

$$|b_1| \leq a_1,$$

and the first term simplifies to $2(n-1)a_1$.

Step 2. KKT analysis in the strict case $|b_1| < a_1$. In this regime, the objective is independent of b_1, b_2 . We write down the stationary condition for b_1, b_2 :

$$\begin{aligned} \lambda_2 - \mu_1 &= 0 \\ \lambda_2 - n\mu_1 &= 0. \end{aligned} \tag{67}$$

Thus, we get $\lambda_1 = \mu_1 = 0$. Then we consider stationary condition for a_2 :

$$2n \frac{X}{\sqrt{X^2 + n\alpha^2}} - \lambda_2 - n\mu_1 = 0. \tag{68}$$

It implies $X = 0$. Finally, we consider a_1 :

$$2(n-1) + 2 \frac{X}{\sqrt{X^2 + n\alpha^2}} - \lambda_1 - \lambda_2 - \mu_1 = 0. \tag{69}$$

As a result $\lambda_1 = 2(n-1)$. Based on complementary slackness, we have $a_1 = 1$. Thus,

$$a_2 = -\frac{1}{n}, \quad b_2 = -\frac{b_1}{n}, \quad b_1 + nb_2 = 0.$$

Consequently,

$$a_1 + a_2 = 1 - \frac{1}{n}, \quad b_1 + b_2 = (1 - \frac{1}{n})b_1,$$

and since $|b_1| < 1$, we obtain

$$b_1 + b_2 < 1 - \frac{1}{n} = a_1 + a_2.$$

Step 3. Boundary case $|b_1| = a_1$. We first find that $a_1 = 1$. Otherwise, we can perturb a_1, a_2, b_1, b_2 to lower the objective function while keeping X fixed. The first term of the optimization object is $2(n-1)$. Moreover, we can consider another solution

$$a_1 = 1, a_2 = -\frac{1}{n}, b_1 = -\frac{1}{n-1}, b_2 = \frac{1}{n(n-1)}, \alpha = 0. \tag{70}$$

Direct verification shows that the constraints are satisfied and the value of the objective function is $2(n-1)$. Thus, if we can find optimal solution in this case, the second term of objective function must be zero. As a result, we have

$$a_1 + na_2 = 0, \alpha = 0. \tag{71}$$

Moreover, it implies that

$$a_2 = -\frac{1}{n}, b_1 + nb_2 = 0 \tag{72}$$

Then we consider the following cases:

- (i) $b_1 = -a_1 = -1$. We have $b_2 = \frac{1}{n}$ and $a_1 + a_2 = 1 - \frac{1}{n} > b_1 + b_2$.
- (ii) $b_1 = a_1 = 1$. It contradicts with the constraint $a_1 + a_2 + \alpha \geq b_1 + b_2 - \alpha + 1$.

□

C Experimental Details

C.1 Architecture Details

Transformer (GPT-2 style). We use a standard decoder-only transformer with pre-norm residual blocks. Let d_{vocab} be the vocabulary size, d_m the model width, d_k the per-head query/key dimension, H the number of heads, L the number of layers, and T the context length. Tokens $x_{1:T}$ are embedded by a lookup matrix $E \in \mathbb{R}^{d_{\text{vocab}} \times d_m}$ and summed with learned positional embeddings. Each layer $\ell = 1, \dots, L$ applies

$$\begin{aligned} Z^{(\ell)} &= X^{(\ell)} + \text{MHA}\left(\text{LN}\left(X^{(\ell)}\right)\right), \\ X^{(\ell+1)} &= Z^{(\ell)} + \text{MLP}\left(\text{LN}\left(Z^{(\ell)}\right)\right), \end{aligned}$$

where MHA is causal multi-head attention with H heads (queries/keys/values computed by linear maps in $\mathbb{R}^{d_m \times Hd_k}$ and output projection in $\mathbb{R}^{Hd_k \times d_m}$), and MLP is a two-layer feed-forward network with GELU activation and hidden size $4d_m$. LayerNorm (LN) is applied in the pre-norm configuration; dropout is disabled unless stated. The language-model head shares weights with E (tied embeddings) and projects to logits in $\mathbb{R}^{d_{\text{vocab}}}$.

C.2 Synthetic Dataset Setup Details

C.3 More real world tasks

As a supplement, similar evaluations on the MuSiQue dataset exhibit the same subject-to-answer association observed in TWOHOPFACT. In particular, when models produce correct multi-hop predictions, their internal representations and decoding behavior consistently reflect a direct dependency between the subject and the final answer, despite differences in dataset construction and question formats. This consistency suggests that the emergence of subject-to-answer associations is not specific to a particular benchmark, but instead reflects a more general property of models that successfully perform two-hop reasoning.

C.4 LLMs learn identity bridge in pretraining

To further examine whether pretrained language models already encode an identity-consistent bridge mechanism, we analyze the repeat accuracy of OLMo checkpoints on a controlled probing task. In this experiment, the repeated entity is set to be the bridge token b in the TWOHOPFACT dataset. High repeat accuracy indicates that the model can reliably preserve and reproduce the bridge entity across positions, which is consistent with an identity mapping on the bridge token. Empirically, we find that OLMo checkpoints achieve consistently high repeat accuracy, suggesting that pretrained representations already support an implicit identity bridge prior to task-specific fine-tuning.

C.5 Discussion about multi-hop reasoning and reverse curse

Extension to 3-Hop Reasoning Although our primary analysis focuses on two-hop reasoning, the same framework naturally extends to 3-hop reasoning with the single-layer Emb-MLP model. For simplicity, we consider the following chain of 3-hop reasoning:

$$a \xrightarrow{r_1} b \xrightarrow{r_2} c \xrightarrow{r_3} d$$

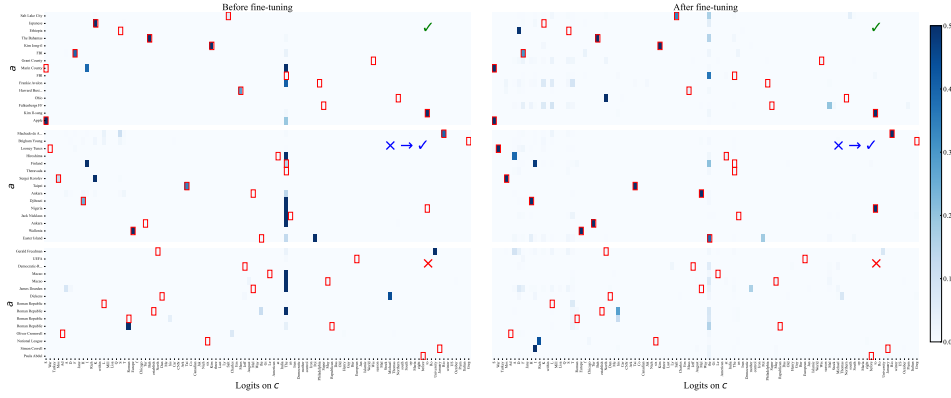


Figure 5: The probability of the model outputting the alternative city token when using the prompt corresponding to (e_1, r_2) before and after fine-tuning on datasets. The red box indicates the answer to the corresponding two-hop reasoning data. The bar chart shows the change in the probability of the corresponding two-hop answer when using the prompt (e_1, r_2) for different datasets.

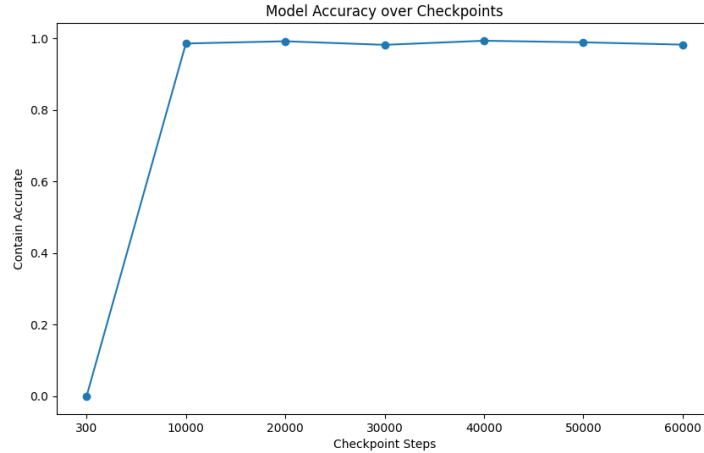


Figure 6: The repeat accuracy of Olmo checkpoints. The repeat entity are set as the bridge token b of TWOHOPFACT dataset.

Based on our theoretical findings, the Emb-MLP model can learn 3-hop reasoning only if it successfully identifies the relationships from a to d , which it can extract via the critical token r_3 . In Fig. 7, we present four different training setups: (a) only 1-hop pairs (a, r_1, b) , (b, r_2, c) , (c, r_3, d) ; (b) 1-hop pairs with identity bridges $b \rightarrow b, c \rightarrow c$; (c) 1-hop pairs combined with two-hop pairs (a, r_1, r_2, c) , (b, r_2, r_3, d) ; and (d) a combination of 1-hop, two-hop pairs, and identity bridges.

In Fig. 7(a) and (b), the model successfully captures the first and second hop patterns but fails to establish the 3-hop relationship. Training with two-hop data, however, strengthens the logits from $a \rightarrow c$ and from $b \rightarrow d$, which helps the model successfully complete the $a \rightarrow d$ connection, as shown in Fig. 7(c) and (d).

Connection to the ‘Reversal Curse’. Some studies have observed that autoregressive large models fail to recognize the reversal of relations, such as not understanding that $a = b$ implies $b = a$. Several works have proposed transformer improvements to address this issue. In Fig. 7(d), we observe that when the output is c , there is a clear activation on the diagonal of the group $\mathcal{E}_3 \rightarrow \mathcal{E}_2$ where the output is b . This indicates that with the help of the

identity bridge, the model can establish connections such as $b \rightarrow b, c \rightarrow c, b \rightarrow c$, thereby facilitating the model’s awareness of the relationship from c to b . This suggests that the identity bridge may also offer a potential solution to the Reversal Curse problem.

For completeness, we include additional results on 3-hop reasoning with the single-layer Emb-MLP model. Fig. 7 visualizes row-wise logit patterns under different training setups, illustrating how two-hop supervision and identity bridges facilitate the formation of the final subject-to-answer association.

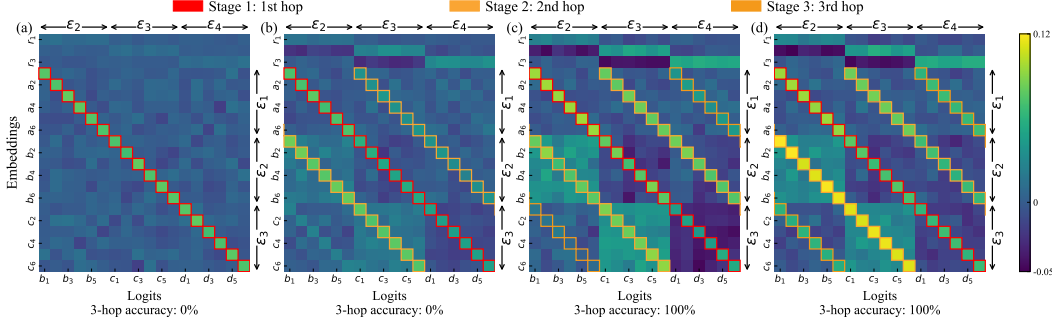


Figure 7: Row-wise logit templates in 3-hop reasoning task (a, r_1, r_2, r_3, d) of Emb-MLP model in different situations: (a) train on 1-hop (b) train on 1-hop and identity bridge (c) train on 1-hop and two-hop (d) train on 1-hop, two-hop and identity bridge. two-hop and identity bridge help the model to construct the relation between a to d highlighted by ‘Stage 3’.

D Experiments Compute Resources

The experiments were conducted on a server with the following configuration:

- 48 AMD EPYC 7352 24-Core Processors, each with 512KB of cache
- 251GB of total system memory
- 8 NVIDIA GeForce RTX 4080 GPUs with 16GB of video memory each
- The experiments were run using Ubuntu 22.04 LTS operating system