IDENTITY BRIDGE: ENABLING IMPLICIT REASONING VIA SHARED LATENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite remarkable advances, large language models often fail at compositional reasoning tasks, a phenomenon exemplified by the "curse of two-hop reasoning". This paper introduces the Identity Bridge, a simple yet powerful mechanism that resolves this compositionality gap by supervising the model on a zero-hop identity task. We demonstrate empirically that this addition enables models to successfully perform out-of-distribution two-hop reasoning, a task they otherwise completely fail. To explain this phenomenon, we provide a theoretical analysis using a simplified Emb-MLP model, proving that identity supervision reshapes the model's latent geometry. We show this alignment is induced by an implicit nuclear-norm regularization during optimization, which favors low-rank solutions that share structure across tasks. For complex tasks, we use small initialization or weight decay to enhance the regularization effect, which enhances the latent space alignment effect and slows down the generalization decay. Finally, we extend our investigation to large-scale models, observing that they still achieve two-hop reasoning through the latent memory, which provides crucial inspiration for enhancing their implicit reasoning abilities.

1 INTRODUCTION

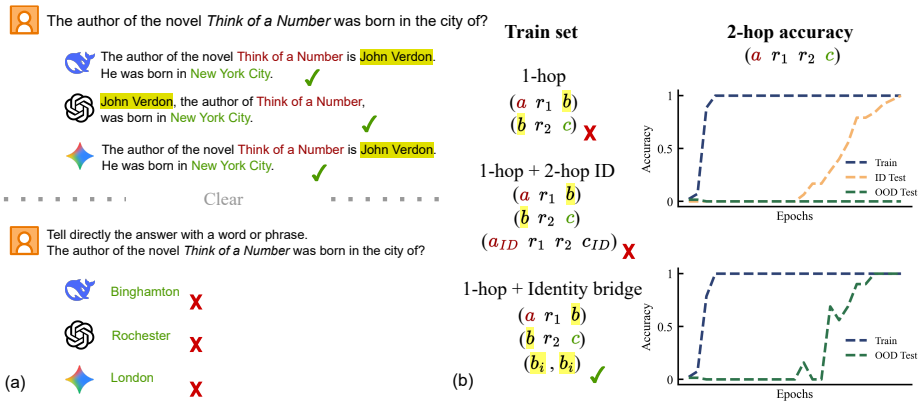

Figure 1: Brief description of the two-hop curse. (a) Large models have difficulty completing two-hop reasoning tasks without CoT assistance. (b) Single-hop tasks and in-distribution two-hop tasks cannot help generalize out-of-distribution two-hop tasks, but identity bridges can.

Large language models achieve strong performance across many tasks, yet they still stumble on behaviors that seem elementary to humans, including the reversal curse Berglund et al. (2024); Allen-Zhu & Li (2024; 2025) and the two-hop curse Balesni et al. (2024). The latter exposes a core limitation in current training and data: models often fail to compose two single-hop facts ("A to B" and "B to C") into the correct conclusion ("A to C") unless an explicit chain-of-thought is provided or the composition appears verbatim in training data Wang et al. (2024); Ye et al. (2025). The issue persists even in state-of-the-art systems Yang et al. (2024); Dziri et al. (2023).

We revisit this phenomenon and show that a minimal addition, an identity bridge, is sufficient to unlock robust out-of-distribution (OOD) two-hop reasoning. We augment training with a zero-hop task mapping each bridge token to itself. Although trivial by itself, this supervision reshapes the latent space so that the second hop can reliably latch onto the features produced by the first hop. Empirically, models that previously failed on OOD two-hop queries begin to compose once identity supervision is present, though the benefit diminishes as task complexity increases. This also clarifies a gap between synthetic two-hop settings and pretrained models: pretraining effectively endows models with a latent form of identity bridging (e.g., the ability to restate text), which partially supports composition.

To understand the mechanism, we analyze a simplified Emb–MLP model, a transformer layer with uniform attention. Gradient-based training induces an implicit nuclear-norm bias toward low-rank, structure-sharing solutions. With identity supervision, this bias promotes cross-task memory sharing and aligns the first-hop subject→bridge representation with the second-hop bridge→object mapping, yielding positive OOD margins on held-out compositions.

We further study high-complexity regimes with larger bridge vocabularies and more relation slices. In these settings, implicit regularization alone is insufficient: relying only on shared latent memory leaves subject states too weakly tied to the correct object. Prior work indicates that small initialization provides a useful implicit bias in large models Zhang et al. (2024; 2025); Yao et al. (2025). Building on this, we find that small initialization or weight decay strengthens regularization, tightens representation alignment across layers, and markedly improves OOD generalization. These results support the view that alignment quality tracks generalization.

Finally, we examine pretrained LLMs on real two-hop datasets. Even without explicit two-hop supervision, fine-tuning signals show that models increase probability mass on the correct tail when prompted with bridge-related cues, consistent with identity-bridge effects accrued during pretraining.

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

1. We introduce the Identity Bridge, a zero-hop supervision that reliably enables OOD two-hop composition (Figure 2).
2. We develop a uniform-attention theory (Theorem 1 and 2) showing how identity supervision, together with implicit nuclear-norm regularization, induces cross-task memory sharing and positive OOD margins which reflects the mechanisms at work within large language models.
3. We identify and address high-complexity failure modes, demonstrating that stronger regularization such as small initialization or weight decay improves alignment and OOD accuracy (Figure 6 and 7).
4. We present evidence on pretrained LLMs that aligns with our account and the existence of identity bridges (Figure 8).

2 Related Work

In this section, we discuss related work and recent progress on implicit reasoning.

Implicit reasoning failure on synthetic data There has been a line of works studying the failure of implicit reasoning on synthetic data. 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 large body of work has studied the evidence underlying reasoning in LLMs. Press et al. (2022); Xu et al. (2024) observed a significant gap between the accuracy of single-hop and double-hop tasks in large models, and this gap does not decrease as the model size increases. 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 golden inference. In order to analyze the possible reasons, Biran et al. (2024) analyzes the intermediate states in the transformer through circuit analysis and points out that one possible reason is that the first-hop task is completed too late, making the model unable to utilize the first-hop information.

Explicit vs. Implicit Reasoning. The failure of two-hop reasoning mainly comes from the model's inability to combine the information of single-hop data in the latent space. One solution is to trade time for space and use an explicit chain of thought (COT) to let the model reason step by step Wei et al. (2022). However, recent work Turpin et al. (2023) points out that CoT is not always faithful and may rely on superficial heuristics. An possible alternative to avoid the use of CoT is to use an explicit recurrence model to implement reasoning in the latent space Dehghani et al. (2019); Hutchins et al. (2022), although this is less common in real-world use. Another approach is to improve the model's reasoning ability during pre-training. A promising approach is to use small initialization and weight decay techniques Zhang et al. (2025), which have also been initially tried in pre-training Hang et al. (2025).

3 PRELIMINARIES

3.1 TWO-HOP REASONING TASK.

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 relations in $\mathcal{R}_1$:

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

Then define a deterministic (or sampling) map

$$g_1 :\; \mathcal{E}_1 \times \mathcal{R}_1 \to \mathcal{E}_2 \quad \text{such that} \quad g_1(e_1, r_i) \in \mathcal{E}_{2,i}.$$

The first-hop triple set is

$$\mathcal{T}_1 \;=\; \{\, (e_1, r_1, e_2) \;:\; e_1 \in \mathcal{E}_1,\; r_1 \in \mathcal{R}_1,\; e_2 = g_1(e_1, r_1) \,\}.$$

This partitioning of $\mathcal{E}_2$ ensures that each $r_i \in \mathcal{R}_1$ only co-occurs with bridge entities from its dedicated slice $\mathcal{E}_{2,i}$, 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 $(e_1, r_i, e_2) \in \mathcal{T}_1$ and $(e_2, r_j, e_3) \in \mathcal{T}_2$, the composed query is (e_1, r_i, r_j) with answer

$$(e_1, r_i, r_j) \;=\; g_2\big(g_1(e_1, r_i),\, r_j\big) \;=\; e_3.$$

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

$$(e_2) \;\to\; e_2,$$

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.

3.2 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:

$$\text{Complexity} = \max_{e_1 \in \mathcal{E}_1} \#\{e_3 = (e_1, r_i, r_j) \mid r_i \in \mathcal{R}_1, r_j \in \mathcal{R}_2\}.$$

3.3 Model Architecture

Transformer. We use a standard GPT-2 model. Let d_{vob}, d_m, d_k denote the vocabulary size, embedding space dimension, and query-key-value projection dimension, respectively. Then, each sequence $x_{1:L}$ is embedded to $\boldsymbol{X}$ through embedding matrix $\boldsymbol{E} \in \mathbb{R}^{d_{\text{vob}} \times d_m}$. Then use standard attention and MLP modules, and use residual connections and layer norm between each module. For details on the model implementation, please refer to appendix.

Embedding-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, denoted by $\mathcal{V}_{\text{in}}$ and $\mathcal{V}_{\text{out}}$. The embedding matrix is $\boldsymbol{E} \in \mathbb{R}^{|\mathcal{V}_{\text{in}}| \times d_m}$, with row $\boldsymbol{e}_s$ corresponding to token $s \in \mathcal{V}_{\text{in}}$, and the projection matrix is $\boldsymbol{W}_{\text{proj}} \in \mathbb{R}^{d_m \times |\mathcal{V}_{\text{out}}|}$. For a sequence $X = (s_1, \ldots, s_T)$, we define:

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

$$f_{\boldsymbol{\theta}}(X) = \left(\sum_{t=1}^{T} \boldsymbol{e}_{s_t} \right) \boldsymbol{W}_{\text{proj}} \in \mathbb{R}^{|\mathcal{V}_{out}|}.$$

3.4 Parameter Initialization

For any learnable weight matrix $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}$, with d_1 and d_2 denoting the input and output dimensions, we initialize each entry from a Gaussian distribution $\boldsymbol{W}_{i,j} \sim \mathcal{N}\big(0, \sigma^2\big)$. 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. (2024); Yao et al. (2025) shows that networks initialized in this way often have stronger reasoning capabilities.

4 Results

In this section, we construct different construct data sets of different complexity to demonstrate the role of identity bridge which unlocks out-of-distribution composition. To further elucidate the mechanism of the identity bridge, we use embedding-MLP as a simplified model on a dataset with complexity one to prove that the identity bridge, together with the regularization of the gradient descent algorithm, enables the model to share latent space memory, thereby completing two-hop reasoning. On high-complexity data, we use small initialization settings to enhance the implicit regularization effect brought by the gradient and achieve generalization.

4.1 Identity Matters for OOD Generalization

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

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

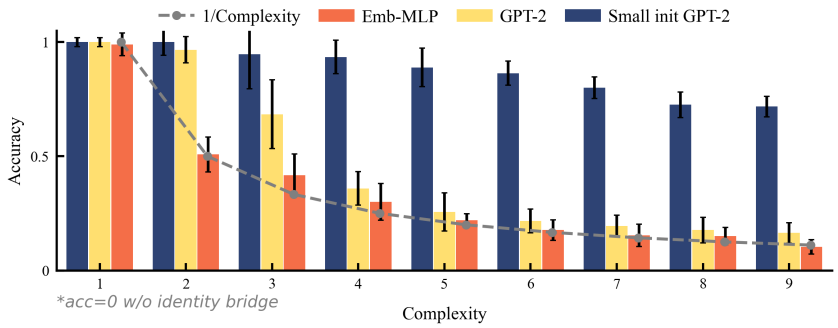

Figure 2: Test accuracy of different models on datasets with different complexities. The accuracy of the standard GPT2 and Emb-MLP models is well consistent at different complexity levels, and the small initialization model achieves excellent performance exceeding 1/C.

Partition the bridge set evenly as

$$\mathcal{E}_2 \;=\; \bigcup_{j=1}^{C} \mathcal{E}_{2,j}, \qquad \mathcal{E}_{2,j} \;=\; \{\, e_{2,k} \,:\, (j-1)N+1 \le k \le jN \,\}.$$

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

$$g_1(e_{1,i},\, r_{1,j}) \;=\; e_{2,(j-1)N+i}, \qquad g_2(e_{2,k},\, r_2) = e_{3,\lfloor \frac{k}{N} \rfloor + k \bmod N},$$

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:

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

with one-to-one correspondences $g_1(a_i, r_1) = b_i$ and $g_2(b_i, r_2) = c_i$, hence the composed answer for (a_i, r_1, r_2) is c_i.

Figure 2 reports test accuracy across models and complexity levels C. Identity bridges is effective throughout since it yields non-zero OOD two-hop generalization for all models regardless of whether there are nonlinearities in the model or using small initialization. In particular, at $C = 1$ the identity signal suffices for small-initialized GPT-2, standard GPT-2, and the simplified Emb–MLP to perform well, consistent with the implicit-regularization account developed in Secs. 4.2 and 4.3.

It should be noted that, as complexity C increases, accuracy for standard GPT-2 and Emb–MLP declines simultaneously, indicating that Emb–MLP captures the operative mechanism of the transformer and that gradient-descent implicit bias alone is insufficient for strong OOD composition at higher complexity. In contrast, the small-initialized GPT-2 degrades more gracefully, suggesting a stronger regularization effect: the initialization-induced bias further constrains the latent geometry and better preserves the bridge–object coupling needed for composition which will be discussed in Sec. 4.4.

4.2 Cross-task memory via implicit regularization

We now examine how identity bridge enables composition of one-hop tasks. To make the mechanism explicit, we use the Emb–MLP to solve task with one complexity and analyze the row-wise logit templates encoded by

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

where the i-th row of $\boldsymbol{W}$ is the (unnormalized) logit vector produced by input token i; specifically, $\boldsymbol{W}_{ij}$ is the logit assigned by token i to output token j. 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.

In this setup, Fig. 3(a) shows that relations act primarily as set selectors: r_1 boosts logits toward $\mathcal{E}_2$ while suppressing $\mathcal{E}_3$, and r_2 does the converse. Hence, the substantive computation of the two hops

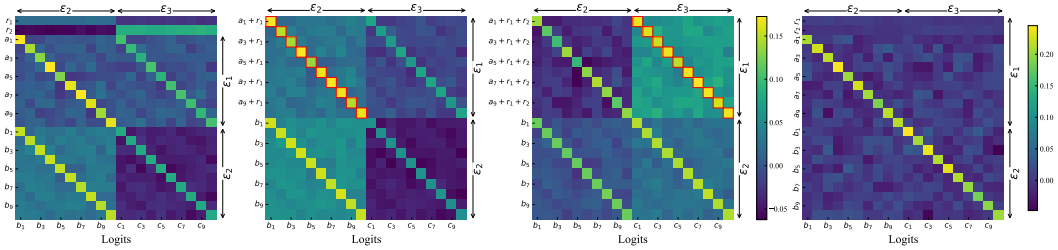

Figure 3: **Row-wise logit templates in Emb–MLP.** Panels (a–c) are trained with bridge identity supervision; panel (d) omits it. **(a)** Base logit matrix: the top two rows correspond to relation tokens r_1 and r_2; the remaining rows correspond to entity tokens in $\mathcal{E}_1$ and $\mathcal{E}_2$. **(b)** One-hop input (a_i, r_1), visualized as the row-wise sum of a_i and r_1. **(c)** 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. **(d)** Same visualization as in (c) but trained without identity supervision.

is carried by the entity rows of $\boldsymbol{W}$. The key question is whether the subject rows for a_i encode a discriminative bias toward the correct tail c_i.

With identity supervision, Fig. 3(a) 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) concentrates subject logits on the appropriate bridge slice; under the implicit nuclear-norm regularization induced by gradient-based training, as discussed in Sec. 4.3, 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. 3(b) and 3(c) 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.

4.3 Uniform-attention theory

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

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

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

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

In the Emb–MLP model, writing the logit matrix $\boldsymbol{W} = \boldsymbol{E}\,\boldsymbol{W}_{\text{proj}} \in \mathbb{R}^{|\mathcal{V}_{\text{in}}| \times |\mathcal{V}_{\text{out}}|}$, which leads to the following convex formulation in $\boldsymbol{W}$ (e.g., Huang et al. (2025)):

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

Problem equation 2 is convex since the objective function is convex and constraints are linear. While a KKT point of equation 1 does not, in general, certify optimality for equation 2 without additional structure, our empirical evidence in Sec. 4.2 shows that the optimizer of equation 2 closely matches the learned row-wise logit templates. We therefore use equation 2 to state margin consequences for two-hop generalization.

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

Theorem 1 (Positive OOD margin with identity supervision)**.** *Assume training includes zero-hop identity supervision over $\mathcal{E}_2$, and let $\boldsymbol{W}^\star$ (equivalently $\boldsymbol{\theta}^\star$) solve equation 2. 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 on held-out compositions.

Proof sketch. We first use the dataset's permutation symmetry to get a highly structured optimal solution. We then obtain a closed-form objective in this structure and, with a mild symmetry constraint, reduce the analysis to a few parameters. To avoid the discussion about subderivative, we add a simple slack variable and apply KKT; the KKT conditions force tight, linear ties between the subject, bridge, and object blocks. These ties shrink the OOD margin check to a one-dimensional inequality that feasibility makes strictly positive, yielding a positive margin on every held-out two-hop query. See appendix A.2 for details of the proof.

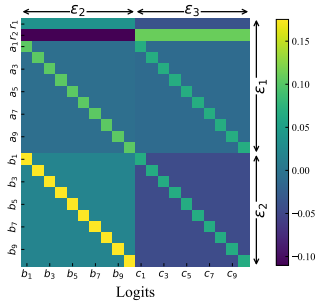

Figure 4: The optimal solution of optimization problem 2 with identity bridge.

Theorem 2 (Failure without identity supervision)**.** *If identity supervision is omitted, any solution of equation 2 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 held-out compositions.

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 A.3 for details of the proof.

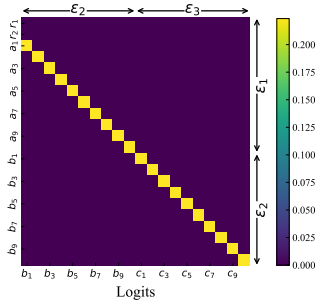

Figure 5: The optimal solution of optimization problem 2 without identity bridge.

4.4 High-complexity regimes: strengthening regularization and alignment

As dataset complexity increases with larger bridge vocabulary and more relation slices, the implicit regularization is no longer sufficient to solve two-hop reasoning task since relying solely on shared latent memory makes the model unable to distinguish the information of the object carried by the subject. We therefore strengthen regularization either by small initialization or by weight decay. Both interventions substantially recover OOD performance.

Figures 6 visualize the geometry of hidden states across layers using the polar alignment plots. As this figure illustrated ,models with small initialization or with weight decay exhibit tighter alignment between (e_1, r_1) and the corresponding bridge e_2 than the standard GPT-2 trained with the same dataset.

Furthermore, as Fig. 7 indicates, the emergence of shared latent memory of a_i to c_i 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.

Mechanistically, both small initialization and weight decay alleviates the pain of the model performing two-hop reasoning only through shared memory. Under higher complexities, We need to further use the information of the bridge entity to enable the model to correctly identify the corresponding object. The alignment of the hidden state allows the model to more directly use the second-hop data to restore positive OOD margins and two-hop generalization.

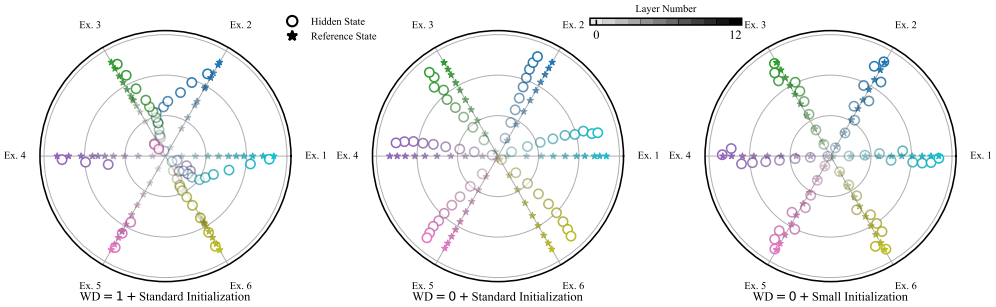

Figure 6: T-SNE visualization of hidden states with respect to one-hop data and corresponding bridge entity. Six samples are selected for each setting, where the star represents the hidden state of the bridge entity and the circle represents the hidden state of the first hop data (e_1, r_i) at position r_i. Colors from light to dark represent hidden states from shallow to deep.

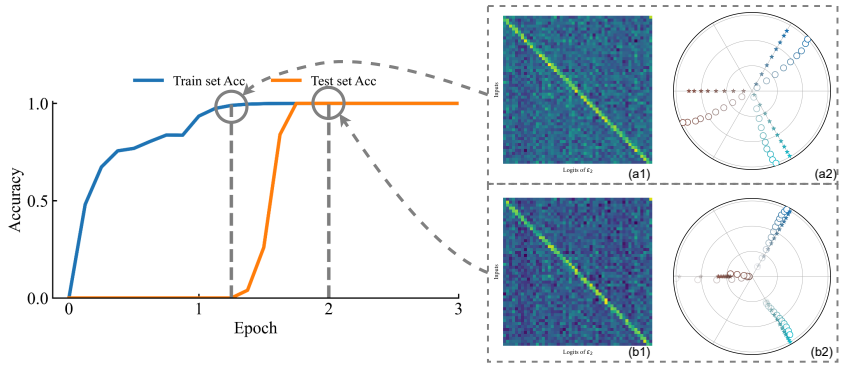

Figure 7: Alignment analysis along the training steps for small initialized GPT-2 with complexity= 2. The two images correspond to when the memorization on 1-hop ends but OOD generalization ability is missing, and the moment when OOD generalization is completed. Figure (a1) and (b1) shows the relationship between the input a_i and the output logits of c_i, similarly to Fig. 3. Figure (a2) and (b2) is the T-SNE visualization of hidden states like Fig. 6.

4.5 Real Task Analysis

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 which was introduced by 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. Although fine-tuning can improve the two-hop reasoning ability of the model to a certain extent, the accuracy improvement brought by the identity mapping is not significant. However, we can still verify the theoretical results on a large model.

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. 8 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

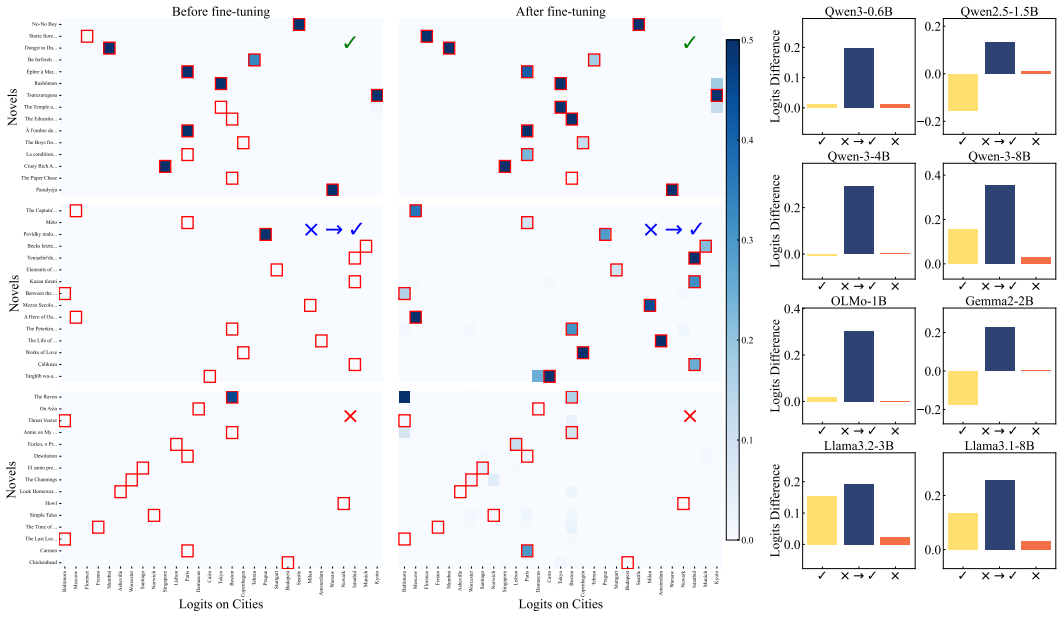

Figure 8: The probability of the model outputting the alternative city token when using the prompt corresponding to (e_1, r_2) before and after fine-tuning on datasets. The red box indicates the answer to the corresponding two-hop reasoning data. The bar chart shows the change in the probability of the corresponding two-hop answer when using the prompt (e_1, r_2) for different datasets.

the identity bridge during the pre-training process. 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.

5 Discussions

Conclusion. We revisited two-hop compositional generalization and showed that the identity bridge on bridge entities reliably lifts OOD two-hop accuracy across model families (GPT-2 variants and Emb–MLP) and dataset complexities. Empirically, identity supervision aligns the first-hop subject to bridge representation with the second hop, enabling composition; stronger regularization (small initialization or weight decay) further tightens this alignment and restores OOD performance in high-complexity regimes.

Limitations. Our theory is developed in a simplified Emb–MLP (uniform-attention) setting and does not model attention dynamics explicitly. Nevertheless, it captures the key phenomena observed in standard GPT-2—latent-space sharing, alignment under identity bridge, and failure without it—providing a sufficient explanatory account of the empirical trends we report.

LLM Usage

In this work, the LLMs are employed to correct grammatical errors and inappropriate words.

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.

Mikita Balesni, Tomek Korbak, and Owain Evans. The two-hop curse: Llms trained on a to b, b to c fail to learn a to c. *arXiv preprint arXiv:2411.16353*, 2024.

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/.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyzdRiR9Y7.

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.

Liangkai Hang, Junjie Yao, Zhiwei Bai, Tianyi Chen, Yang Chen, Rongjie Diao, Hezhou Li, Pengxiao Lin, Zhiwei Wang, Cheng Xu, et al. Scalable complexity control facilitates reasoning ability of llms. *arXiv preprint arXiv:2505.23013*, 2025.

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.

DeLesley Hutchins, Imanol Schlag, Yuhuai Wu, Ethan Dyer, and Behnam Neyshabur. Block-recurrent transformers. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=uloenYmLCAo.

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.

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

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.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=bzs4uPLXvi.

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.

Jiaran Ye, Zijun Yao, Zhidian Huang, Liangming Pan, Jinxin Liu, Yushi Bai, Amy Xin, Liu Weichuan, Xiaoyin Che, and Lei Hou. How does transformer learn implicit reasoning? 2025.

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

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*, 2024. URL https://openreview.net/forum?id=YOBGdVaYTS.

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 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.

A.1 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 (2), Huang et al. (2025) Lemma 3)**.** *Suppose $\boldsymbol{W}$ is the solution to the optimization problem (2) 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*

$$\boldsymbol{W}^\intercal = \begin{pmatrix} a_1\boldsymbol{I}_n + a_2\boldsymbol{E}_n & b_1\boldsymbol{I}_n + b_2\boldsymbol{E}_n & e\mathbf{1}_n & f\mathbf{1}_n \\ c_1\boldsymbol{I}_n + c_2\boldsymbol{E}_n & d_1\boldsymbol{I}_n + d_2\boldsymbol{E}_n & g\mathbf{1}_n & h\mathbf{1}_n \end{pmatrix}. \tag{3}$$

The parameters follow the following constraints:

$$\begin{aligned} &a_1, d_1 \geq 1, \\ &a_1 + a_2 + e \geq c_1 + c_2 + g + 1, \\ &d_1 + d_2 + h \geq b_1 + b_2 + f + 1. \end{aligned} \tag{4}$$

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

$$\begin{aligned} &b_1 \geq 1, \\ &b_1 + b_2 \geq d_1 + d_2 + 1. \end{aligned} \tag{5}$$

To analyze the optimal solution of optimization problem 2, we need an explicit formula of the nuclear norm. After block multiplication, $\boldsymbol{W}^\intercal\boldsymbol{W}$ can be written as:

$$\boldsymbol{W}^\intercal\boldsymbol{W} = \begin{pmatrix} C_{A1}\boldsymbol{I}_n + C_{A2}\boldsymbol{E}_n & C_{B1}\boldsymbol{I}_n + C_{B2}\boldsymbol{E}_n \\ C_{B1}\boldsymbol{I}_n + C_{B2}\boldsymbol{E}_n & C_{D1}\boldsymbol{I}_n + C_{D2}\boldsymbol{E}_n \end{pmatrix}, \tag{6}$$

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 can be found in Huang et al. (2025). Readers can also refer to the proof of Lemma 4, the proof ideas are similar.

Through direct calculation, we have the following explicit expression for the nuclear norm of $\boldsymbol{W}$.

Lemma 2. *The $\boldsymbol{W}$ in restricted form 7 has the nuclear norm $\|\boldsymbol{W}\|_*$.*

$$\begin{aligned} &(n-1)\sqrt{C_{A1} + C_{D_1} + 2|a_1d_1 - b_1c_1|} \\ &+\sqrt{C_{A1} + nC_{A2} + C_{D_1} + nC_{D2} + 2\sqrt{(C_{A1} + nC_{A2})(C_{D_1} + nC_{D2}) - (C_{B1} + nC_{B2})^2}}. \end{aligned} \tag{7}$$

To fully characterize the properties of the optimal solution to optimization problem 2, existence alone is not enough. Huang et al. (2025) introduces the following assumption and proves the uniqueness of the solution to the optimization problem.

Assumption 1. *Suppose that $\boldsymbol{W}$ is a solution to optimization problem (2), but takes a different form from (3). Then we assume that $\mathbf{1}_{2n}^\intercal\boldsymbol{W}^\intercal \neq \mathbf{0}_{nm+2}^\intercal$.*

A.2 Proof for Theorem 1

We now prove Theorem 1. The proof steps are as follows: First, with the help of Assumption 2, we add constraints and simplify the problem. Next, we introduce slack variables to transform the problem into an equivalent form. Finally, under certain assumptions, we analyze the local minimum of the relaxed problem and complete the proof.

On the basis of the observations from numerical experiments, we introduce the following assumption which can be seen as a supplement to Assumption 1.

Assumption 2. *Suppose that $\boldsymbol{W}$ is a solution to optimization problem (2) with restricted form (6). Then we assume that $\mathbf{1}_{2n}^{\intercal}\boldsymbol{W}^{\intercal} = \mathbf{0}_{2n+2}^{\intercal}$.*

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

Lemma 3. *Suppose Assumption 2 holds. Then, the optimization (2) problem is equivalent to the following optimization problem.*

$$\begin{aligned}
\min_{a_i,b_i,c_i,d_i,e,f,g,h} \quad & (n-1)\sqrt{M_1} + \sqrt{2M_2} \\
\textit{s.t.} \quad & a_1, b_1, d_1 \geq 1, \\
& a_1 + a_2 + e \geq c_1 + c_2 + g + 1, \\
& b_1 + b_2 \geq d_1 + d_2 + 1, \\
& d_1 + d_2 + h \geq b_1 + b_2 + f + 1, \\
& a_1 + c_1 = -n(a_2 + c_2), \quad b_1 + d_1 = -n(b_2 + d_2), \\
& e = -g, \quad f = -h,
\end{aligned} \tag{8}$$

where the terms M_1 and M_2 are defined as

$$\begin{aligned}
M_1 &= a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2|a_1 d_1 - b_1 c_1|, \\
M_2 &= (a_1 + n a_2)^2 + (b_1 + n b_2)^2 + n e^2 + n f^2.
\end{aligned}$$

Proof. The proof consists of two main steps. First, we establish several key identities among the problem's coefficients that arise from Assumption 2. Second, we substitute these identities into the nuclear norm expression from (7) to derive the simplified objective function.

Step 1: Deriving Identities from Assumption 2. The four equality constraints presented in the lemma statement are a direct consequence of the structural properties imposed by Assumption 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:

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

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 (2) 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 (8). The remaining constraints are carried over directly, which completes the proof. □

To avoid introducing a discussion of the subderivative, we consider an equivalent form of problem (8) which simplifies the proof process. We give the following proposition.

Proposition 1 (Equivalent Reformulation)**.** *Let the original optimization problem (8), denoted as* (P_{orig})*, be defined over the feasible set* $\mathcal{X} = \{x = (a_1, a_2, \ldots, g, h)\}$*. The objective function can be rewritten as:*

$$F_{\text{orig}}(x) = (n-1)\sqrt{M_1(x, |u(x)|)} + \sqrt{2M_2(x)},$$

where $u(x) := a_1 d_1 - b_1 c_1$*,* $M_1(x, v) := a_1^2 + b_1^2 + c_1^2 + d_1^2 + 2v$*, and* $M_2(x)$ *is a term independent of* $u(x)$*.*

We introduce an auxiliary variable t *to construct a reformulated problem, denoted as* (P_{ref})*. Its feasible set is* $\widetilde{\mathcal{X}} = \{(x, t) \mid x \in \mathcal{X}, t \geq |u(x)|\}$*, and its objective function is:*

$$F_{\text{ref}}(x, t) = (n-1)\sqrt{M_1(x, t)} + \sqrt{2M_2(x)}.$$

Assume that $n > 1$ *and* $M_1(x, t) > 0$ *over the feasible set* $\widetilde{\mathcal{X}}$*. Then, the two problems are equivalent in the following senses:*

1. *Their optimal values are equal:* $\inf_{x \in \mathcal{X}} F_{\text{orig}}(x) = \inf_{(x,t) \in \widetilde{\mathcal{X}}} F_{\text{ref}}(x, t)$.
2. *Their sets of optimizers correspond to each other via the mapping* $x^* \mapsto (x^*, |u(x^*)|)$*. Specifically, if* x^* *is an optimizer for* (P_{orig})*, then* $(x^*, |u(x^*)|)$ *is an optimizer for* (P_{ref})*. Conversely, if* (x^*, t^*) *is an optimizer for* (P_{ref})*, then it must hold that* $t^* = |u(x^*)|$*, and* x^* *is an optimizer for* (P_{orig})*.*

Proof. The proof proceeds in three steps. First, we show that the optimal value of (P_{ref}) is less than or equal to that of (P_{orig}). Second, we prove the reverse inequality. Finally, we establish the one-to-one correspondence between the sets of optimizers.

Step 1: Showing $\inf_{\widetilde{\mathcal{X}}} F_{\text{ref}} \leq \inf_{\mathcal{X}} F_{\text{orig}}$

Let x be an arbitrary feasible point in $\mathcal{X}$. We can construct a corresponding point in $\widetilde{\mathcal{X}}$ by setting $t_x := |u(x)|$. Since $x \in \mathcal{X}$ and $t_x = |u(x)| \geq |u(x)|$, the point (x, t_x) is feasible for (P_{ref}), i.e., $(x, t_x) \in \widetilde{\mathcal{X}}$.

By substituting t_x into the objective function of (P_{ref}), we find:

$$\begin{aligned} F_{\text{ref}}(x, t_x) &= (n-1)\sqrt{M_1(x, |u(x)|)} + \sqrt{2M_2(x)} \\ &= F_{\text{orig}}(x). \end{aligned}$$

Since the infimum of a function over a set is less than or equal to its value at any point in that set, we have:

$$\inf_{(x',t') \in \widetilde{\mathcal{X}}} F_{\text{ref}}(x', t') \leq F_{\text{ref}}(x, t_x) = F_{\text{orig}}(x).$$

This inequality holds for any arbitrary $x \in \mathcal{X}$. Therefore, by taking the infimum over all $x \in \mathcal{X}$ on the right-hand side, we obtain:

$$\inf_{(x,t) \in \widetilde{\mathcal{X}}} F_{\text{ref}}(x, t) \leq \inf_{x \in \mathcal{X}} F_{\text{orig}}(x). \tag{9}$$

Step 2: Showing $\inf_{\widetilde{\mathcal{X}}} F_{\text{ref}} \geq \inf_{\mathcal{X}} F_{\text{orig}}$

Now, let (x, t) be an arbitrary feasible point in $\widetilde{\mathcal{X}}$. By definition, $x \in \mathcal{X}$ and $t \geq |u(x)|$. The objective function $F_{\text{ref}}(x, t)$ depends on t only through the term $\sqrt{M_1(x, t)}$. Let us define a function

$\phi(v) := \sqrt{v}$. Since we assumed $M_1(x,t) > 0$, and the coefficient $n-1 > 0$, the function $v \mapsto (n-1)\phi(v)$ is strictly increasing for $v > 0$.

Given that $t \geq |u(x)| \geq 0$, we have:

$$\sqrt{M_1(x,t)} = \sqrt{a_1^2 + \cdots + 2t} \geq \sqrt{a_1^2 + \cdots + 2|u(x)|} = \sqrt{M_1(x,|u(x)|)}.$$

Multiplying by the positive constant $(n-1)$ and adding the non-negative term $\sqrt{2M_2(x)}$ to both sides preserves the inequality:

$$(n-1)\sqrt{M_1(x,t)} + \sqrt{2M_2(x)} \geq (n-1)\sqrt{M_1(x,|u(x)|)} + \sqrt{2M_2(x)}.$$

This is equivalent to:

$$F_{\text{ref}}(x,t) \geq F_{\text{orig}}(x).$$

This inequality holds for any arbitrary $(x,t) \in \widetilde{\mathcal{X}}$. The right-hand side, $F_{\text{orig}}(x)$, is the value of the original objective at a point in its feasible set $\mathcal{X}$. Therefore, its value must be greater than or equal to the infimum of the original problem:

$$F_{\text{ref}}(x,t) \geq F_{\text{orig}}(x) \geq \inf_{x' \in \mathcal{X}} F_{\text{orig}}(x').$$

By taking the infimum over all $(x,t) \in \widetilde{\mathcal{X}}$ on the left-hand side, we get:

$$\inf_{(x,t)\in\widetilde{\mathcal{X}}} F_{\text{ref}}(x,t) \geq \inf_{x\in\mathcal{X}} F_{\text{orig}}(x). \tag{10}$$

Combining inequalities (9) and (10), we conclude that the optimal values of the two problems are equal:

$$\inf_{(x,t)\in\widetilde{\mathcal{X}}} F_{\text{ref}}(x,t) = \inf_{x\in\mathcal{X}} F_{\text{orig}}(x).$$

Step 3: Correspondence of Optimizers

Let $p^* = \inf_{\mathcal{X}} F_{\text{orig}} = \inf_{\widetilde{\mathcal{X}}} F_{\text{ref}}$ be the common optimal value.

($\Rightarrow$) Suppose x^* is an optimizer for (P_{orig}), meaning $x^* \in \mathcal{X}$ and $F_{\text{orig}}(x^*) = p^*$. Let $t^* = |u(x^*)|$. As shown in Part 1, the point (x^*, t^*) is in $\widetilde{\mathcal{X}}$ and $F_{\text{ref}}(x^*, t^*) = F_{\text{orig}}(x^*) = p^*$. Since p^* is the infimum for (P_{ref}), the point (x^*, t^*) must be an optimizer for (P_{ref}).

($\Leftarrow$) Conversely, suppose (x^*, t^*) is an optimizer for (P_{ref}), meaning $(x^*, t^*) \in \widetilde{\mathcal{X}}$ and $F_{\text{ref}}(x^*, t^*) = p^*$. From the chain of inequalities derived in Part 2, we know that for any feasible point (x,t):

$$F_{\text{ref}}(x,t) \geq F_{\text{orig}}(x) \geq \inf_{x' \in \mathcal{X}} F_{\text{orig}}(x') = p^*.$$

Applying this to our optimizer (x^*, t^*):

$$p^* = F_{\text{ref}}(x^*, t^*) \geq F_{\text{orig}}(x^*) \geq p^*.$$

This forces all inequalities in the chain to hold with equality. Therefore, we must have $F_{\text{orig}}(x^*) = p^*$, which proves that x^* is an optimizer for (P_{orig}).

Furthermore, we must also have the first inequality hold with equality:

$$F_{\text{ref}}(x^*, t^*) = F_{\text{orig}}(x^*).$$

Substituting the definitions of the objective functions, this equality becomes:

$$(n-1)\sqrt{M_1(x^*, t^*)} = (n-1)\sqrt{M_1(x^*, |u(x^*)|)}.$$

Since the function $v \mapsto (n-1)\sqrt{M_1(x^*, v)}$ is strictly increasing (as $n > 1$ and $M_1 > 0$), the equality of function values implies the equality of their arguments. Thus, it must hold that:

$$t^* = |u(x^*)|.$$

In conclusion, the sets of optimizers for the two problems are in a one-to-one correspondence via the mapping $x^* \mapsto (x^*, |u(x^*)|)$. This completes the proof. □

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

$$\begin{aligned}\min_{a_i,b_i,c_i,d_i,e,f,g,h,t} \quad & (n-1)\sqrt{a_1^2+b_1^2+c_1^2+d_1^2+2t} \\ & +\sqrt{2}\sqrt{(a_1+na_2)^2+(b_1+nb_2)^2+ne^2+nf^2},\end{aligned} \tag{11}$$

The inequality constraints are

$$\begin{aligned}&g_1(x,t): a_1-1\geq 0,\\ &g_2(x,t): a_1+a_2+2e-c_1-c_2-1\geq 0,\\ &g_3(x,t): b_1-1\geq 0,\\ &g_4(x,t): b_1+b_2-d_1-d_2-1\geq 0,\\ &g_5(x,t): d_1-1\geq 0,\\ &g_6(x,t): d_1+d_2-b_1-b_2-2f-1\geq 0,\\ &g_7(x,t): t-(a_1d_1-b_1c_1)\geq 0,\\ &g_8(x,t): t+(a_1d_1-b_1c_1)\geq 0,\\ &g_9(x,t): t\geq 0.\end{aligned}$$

The equality constraints are

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

Having established an equivalent, continuously differentiable formulation of our problem in (11), we now give the proof of Theorem 1. We first give a proof of the theorem under the following conditions and then prove that this condition holds for the optimal solution of the optimization problem (11) in the following discussion. We assume that an optimal solution satisfies

$$a_1=1, \quad a_1+a_2+2e-c_1-c_2-1=0 \tag{12}$$

Proof of Theorem 1. Thanks to Assumption 2, 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

$$c_1+c_2+g+h>a_1+a_2+e+f. \tag{13}$$

This is because

$$\begin{aligned}&s_{(X,y),b_j}=c_1+c_2+g+h-\max\{a_1,0\}-a_2-e-f, \quad \forall j\in[N]\\ &s_{(X,y),c_j}=c_1, \quad \forall j\neq i.\end{aligned} \tag{14}$$

Based on Assumption 4 and inequality constraint $a_1\geq 0$, we just need to prove (13). Using the constraints $e+g=0$ and $f+h=0$, inequality (13) can be reformulated as

$$c_1+c_2-(a_1+a_2)>2e+2f. \tag{15}$$

Utilizing condition (12), The left side of the inequality is simplified to

$$\begin{aligned}c_1+c_2-(a_1+a_2)&=a_1+a_2+2e-1-(a_1+a_2)\\ &=2e-1.\end{aligned} \tag{16}$$

As a result, inequality (15) holds if and only if

$$f<-\frac{1}{2}. \tag{17}$$

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

$$\begin{aligned}b_1+b_2&\geq d_1+d_2+1\\ &\geq b_1+b_2+2f+2,\end{aligned} \tag{18}$$

which implies that

$$f\leq -1. \tag{19}$$

As a result, inequality (15) holds which implies $q(X,y)>0$ for all OOD query. □

Finally, we prove that condition (12) holds. Our approach is to analyze its solution structure using the Karush-Kuhn-Tucker (KKT) framework. First we review the definition of KKT conditions.

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

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

where f, g_n, h_m are continuously differentiable functions. We say that $\boldsymbol{x} \in \mathbb{R}^d$ is a feasible point of (P) if $\boldsymbol{x}$ satisfies $g_n(\boldsymbol{x}) \leq 0$ for all $n \in [N]$ and $h_m(\boldsymbol{x}) = 0$ for all $m \in [M]$.

Definition 2 (KKT point)**.** *A feasible point $\boldsymbol{x}$ of (P) is a KKT point if $\boldsymbol{x}$ satisfies KKT conditions: there exists $\lambda_1, \ldots, \lambda_N \geq 0$ and $\mu_1, \ldots, \mu_M \in \mathbb{R}$ such that*

1. *Stationarity: $\nabla f(\boldsymbol{x}) - \sum_{n=1}^{N} \lambda_n \nabla g_n(\boldsymbol{x}) - \sum_{m=1}^{M} \mu_m \nabla h_m(\boldsymbol{x}) = 0$.*
2. *Complementary slackness: $\forall n \in [N] : \lambda_n g_n(\boldsymbol{x}) = 0$.*

In general, global minimizers of (P) need not meet KKT condition. But with appropriate regularity assumption, the KKT conditions become necessary. To ensure the validity and specificity of following analysis, we introduce the following standard assumptions.

Assumption 3 (Regularity Assumption)**.** *Any optimal solution to the reformulated optimization problem (11) satisfies the Karush-Kuhn-Tucker (KKT) conditions.*

Since the optimization problem characterizes the behavior of the network's normalized parameters, we introduce the following assumptions to rule out degeneracies.

Assumption 4 (Solution Non-degeneracy)**.** *Any optimal solution $(a_1^*, \ldots, h^*, t^*)$ of the reformulated optimization problem (11) is non-degenerate, which means the coefficient c_1^* is strictly positive and solution does not degenerate as n increases, i.e., $a_1^*, \ldots, h^*, t^* = \boldsymbol{\Theta}(1)$.*

The non-degeneracy assumption is motivated by the underlying nuclear norm minimization objective. Intuitively, given the constraints $a_1, b_1, d_1 \geq 1$, a solution where $c_1 \leq 0$ would imply a significant imbalance in the matrix structure, likely leading to a suboptimal, larger nuclear norm. Our analysis therefore focuses on the more representative case where $c_1 > 0$.

With the help of the Assumption 3, we get the useful proposition which provides identity relationships between parameters.

Proposition 2. *Suppose Assumption 3 holds. All optimal solution (x^*, t^*) of reformulated optimization problem satisfy:*

$$a_1^* + n a_2^* = e^*, \quad c_1^* + n c_2^* = -e^* \tag{20}$$

Proof. Using stationarity property for parameter c_2, a_2 and e, we find that the optimal solution satisfies

$$c_2: \quad 0 + \lambda_2 - \mu_1 n = 0 \quad \Rightarrow \quad \lambda_2 = \mu_1 n. \tag{S1}$$

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

$$e: \quad \sqrt{2}\frac{e^*}{\sqrt{M_2}} - 2\lambda_2 = 0. \tag{S3}$$

Substitute equation S1 into equations S2 and S3 respectively, we get

$$a_1^* + n a_2^* = e^*.$$

Another equality comes from the equality constraint $a_1 + c_1 + n(a_2 + c_2) = 0$. □

To prove condition 12, we start from the following binary proposition and strengthen it at the end.

Proposition 3. *Suppose Assumption 3 and Assumption 4 hold, for all optimal solution (x^*, t^*) of reformulated optimization problem, at least one of the following two inequality constraints is tight:*

$$\begin{aligned} & g_1 : a_1 - 1 \geq 0, \\ & g_2 : a_1 + a_2 + 2e - c_1 - c_2 - 1 \geq 0. \end{aligned} \tag{21}$$

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

$$\Delta a_1 = -\alpha\varepsilon, \quad \Delta b_1, \Delta b_2, \Delta d_1, \Delta d_2 = 0, \quad \Delta e, \Delta f, \Delta t = 0 \tag{22}$$

By proposition 2, we take the following variables as

$$\Delta a_2 = \frac{\alpha}{n}\varepsilon, \quad \Delta c_1 = -\gamma\varepsilon, \quad \Delta c_2 = \frac{\gamma}{n}\varepsilon. \tag{23}$$

Since we take $\Delta t = 0$, we need to maintain $\Delta u = 0$. It establishes the following relation between α and γ:

$$\begin{aligned}\Delta u &= \Delta a_1 d_1 - b_1 \Delta c_1 \\ &= -(d_1\alpha - b_1\gamma)\varepsilon.\end{aligned}$$

It implies that $d_1\alpha - b_1\gamma = 0$. Next we show that by choosing appropriate parameter α, 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. We find that

$$\begin{aligned}\Delta M_1 &= 2a_1\Delta a_1 + 2c_1\Delta c_1 \\ &= -2\alpha\varepsilon(a_1 + \frac{c_1 d_1}{b_1}) \\ \Delta M_2 &= 0.\end{aligned} \tag{24}$$

As a result, it implies that

$$\Delta F = -(n-1)\varepsilon\alpha(a_1 + \frac{c_1 d_1}{b_1}). \tag{25}$$

Since $c_1 > 0$ by Assumption 4 and constraints $a_1, b_1, d_1 \geq 1$, we have $a_1 + \frac{c_1 d_1}{b_1} > 0$ which ensures that this 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. Inequality constraints g_7 to g_9 and 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.

□

We strengthen the above proposition by further analyzing the KKT conditions, thus proving that the conditions must hold for the optimal point.

Proposition 4. *Suppose Assumption 3 and Assumption 4 hold, for all optimal solution (x^*, t^*) of reformulated optimization problem, both of the following two constraints are tight:*

$$\begin{aligned}&g_1 : a_1 - 1 \geq 0, \\ &g_2 : a_1 + a_2 + 2e - c_1 - c_2 - 1 \geq 0.\end{aligned} \tag{26}$$

Proof. According to Proposition 7, we classify and discuss the following two situations.

Case 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 equation S1, we get

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

Furthermore, due to equation 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 Assumption 4:

$$a_1 + a_2 + 2e - c_1 - c_2 - 1 = 1 - \frac{1}{n} - c_1 + \frac{1}{n}c_1 - 1 < 0. \tag{27}$$

Case 2: g_2 is tight, but g_1 is not.

We write the stationary condition for a_1, a_2, c_1, c_2, t and note that complementary slackness of KKT condition implies $\lambda_1 = 0$ since g_1 is not tight.

$$\begin{aligned}
(\mathrm{S}_{c_2}) \quad & \lambda_2 - n\mu_1 \;=\; 0, \\
(\mathrm{S}_{c_1}) \quad & (n-1)\frac{c_1}{\sqrt{A}} + \lambda_2 - \mu_1 + b_1(\lambda_8 - \lambda_7) \;=\; 0, \\
(\mathrm{S}_{a_2}) \quad & \sqrt{2}\,\frac{n(a_1 + na_2)}{\sqrt{B}} - \lambda_2 - n\mu_1 \;=\; 0, \\
(\mathrm{S}_{a_1}) \quad & (n-1)\frac{a_1}{\sqrt{A}} + \sqrt{2}\,\frac{a_1 + na_2}{\sqrt{B}} - \lambda_2 - \mu_1 + d_1(\lambda_7 - \lambda_8) \;=\; 0, \\
(\mathrm{S}_{t}) \quad & (n-1)\frac{1}{\sqrt{A}} - (\lambda_7 + \lambda_8 + \lambda_9) \;=\; 0,
\end{aligned} \tag{28}$$

Substituting S_{c_2} into S_{c_1}, we get the first expression of $\lambda_7 - \lambda_8$.

$$\lambda_7 - \lambda_8 = \frac{n-1}{b_1}\left(\frac{c_1}{\sqrt{A}} + \frac{1}{n}\lambda_2\right). \tag{29}$$

Substituting S_{c_2} into S_{a_2}, we get

$$\lambda_2 = \frac{\sqrt{2}n(a_1 + na_2)}{2\sqrt{B}}. \tag{30}$$

So we can simplify condition about S_{a_2} and get

$$\lambda_7 - \lambda_8 = \frac{n-1}{d_1}(\frac{1}{n}\lambda_2 - \frac{a_1}{\sqrt{A}}). \tag{31}$$

Using two forms of $\lambda_7 - \lambda_8$, we get

$$\lambda_2 = \frac{n(d_1c_1 + a_1b_1)}{(b_1 - d_1)\sqrt{A}}. \tag{32}$$

However, we note that $\lambda_2 = \frac{n(d_1c_1+a_1b_1)}{(b_1-d_1)\sqrt{A}}$ and $\lambda_2 = \frac{\sqrt{2}n(a_1+na_2)}{2\sqrt{B}}$. It implies that

$$\sqrt{2}e\sqrt{A}(b_1 - d_1) = 2\sqrt{B}(d_1c_1 + a_1b_1). \tag{33}$$

Based on Assumption 4, the right hand is $\Omega(\sqrt{n})$ but the left hand is $O(1)$, it makes an contradiction.

□

A.3 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 $\boldsymbol{W}$ is the solution to the optimization problem 2 without identical task. There exists a solution with a_1, a_2, b_1, b_2 and α, β such that*

$$\boldsymbol{W}^\intercal = \begin{pmatrix} a_1\boldsymbol{I}_n + a_2\boldsymbol{E}_n & b_1\boldsymbol{I}_n + b_2\boldsymbol{E}_n & \alpha\mathbf{1}_n & \beta\mathbf{1}_n \\ b_1\boldsymbol{I}_n + b_2\boldsymbol{E}_n & a_1\boldsymbol{I}_n + a_2\boldsymbol{E}_n & \beta\mathbf{1}_n & \alpha\mathbf{1}_n \end{pmatrix}. \tag{34}$$

The parameters follow the following constraints:

$$\begin{aligned}
& a_1 \geq 1, \\
& a_1 + a_2 + \alpha \geq b_1 + b_2 + \beta + 1.
\end{aligned} \tag{35}$$

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

$$\sigma(\boldsymbol{W}^\intercal) = \begin{pmatrix} \boldsymbol{P}_\sigma & 0 \\ 0 & \boldsymbol{P}_\sigma \end{pmatrix} \boldsymbol{W}^\intercal \operatorname{diag}\{\boldsymbol{P}_\sigma, \boldsymbol{P}_\sigma, 1, 1\}.$$

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

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

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

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

We further exploit the symmetry and consider the following transformation

$$\tau(\boldsymbol{W}^\intercal) = \begin{pmatrix} 0 & \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{pmatrix} \boldsymbol{W}^\intercal \operatorname{diag}\left\{ \begin{pmatrix} 0 & \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}.$$

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

Under Assumption 2, we consider the problem

$$F \;=\; (n-1)\sqrt{2(a_1^2+b_1^2)+2|a_1^2-b_1^2|} \;+\; 2\sqrt{(a_1+na_2)^2+n\alpha^2} \tag{36}$$

subject to

$$\begin{cases} a_1 \geq 1, \\ a_1 + a_2 + \alpha \;\geq\; b_1 + b_2 - \alpha + 1, \\ a_1 + b_1 + n(a_2 + b_2) = 0. \end{cases} \tag{37}$$

To prove Theorem 2, we just need to prove the following condition holds for the optimal solution of optimization problem 36:

$$b_1 + b_2 < a_1 + a_2 \tag{38}$$

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 equation 36 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

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

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:

$$\begin{aligned} \lambda_2 - \mu_1 &= 0 \\ \lambda_2 - n\mu_1 &= 0. \end{aligned} \tag{39}$$

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

$$2n\frac{X}{\sqrt{X^2 + n\alpha^2}} - \lambda_2 - n\mu_1 = 0. \tag{40}$$

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

$$2(n-1) + 2\frac{X}{\sqrt{X^2 + n\alpha^2}} - \lambda_1 - \lambda_2 - \mu_1 = 0. \tag{41}$$

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

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

Consequently,

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

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

$$a_1 = 1, a_2 = -\frac{1}{n}, b_1 = -\frac{1}{n-1}, b_2 = \frac{1}{n(n-1)}, \alpha = 0. \tag{42}$$

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

$$a_1 + na_2 = 0, \alpha = 0. \tag{43}$$

Moreover, it implies that

$$a_2 = -\frac{1}{n}, b_1 + nb_2 = 0 \tag{44}$$

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 $a_1 + a_2 + \alpha \geq b_1 + b_2 - \alpha + 1$.

□

B Experimental Details

B.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 $\boldsymbol{E} \in \mathbb{R}^{d_{\text{vocab}} \times d_m}$ and summed with learned positional embeddings. Each layer $\ell = 1, \ldots, L$ applies

$$\boldsymbol{Z}^{(\ell)} = \boldsymbol{X}^{(\ell)} + \mathrm{MHA}\Big(\mathrm{LN}\Big(\boldsymbol{X}^{(\ell)}\Big)\Big),$$
$$\boldsymbol{X}^{(\ell+1)} = \boldsymbol{Z}^{(\ell)} + \mathrm{MLP}\Big(\mathrm{LN}\Big(\boldsymbol{Z}^{(\ell)}\Big)\Big),$$

where MHA is causal multi-head attention with H heads (queries/keys/values computed by linear maps in $\mathbb{R}^{d_m \times H d_k}$ and output projection in $\mathbb{R}^{H d_k \times d_m}$), 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 $\boldsymbol{E}$ (tied embeddings) and projects to logits in $\mathbb{R}^{d_{\text{vocab}}}$.

C 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