

G Training Details and Bigger Models

G.1 Training Details

Our training procedure largely follows the public implementation provided by Wang et al. [24], with a few modifications to accommodate our specific experimental setting.

The model is a decoder-only Transformer, identical in architecture to GPT-2, with 8 layers, 768 hidden dimensions, and 12 attention heads. Optimization is performed using AdamW with a learning rate of 1×10^{-4} , 2000 warm-up steps, weight decay of 0.1, and a batch size of 1024. All models are trained significantly beyond convergence to allow observation of late-stage generalization behavior (Section 2.2).

Training is conducted on NVIDIA RTX 3090 GPUs, and the maximum training duration is extended to 3 weeks to ensure stable cross-distributions generalization. All experiments are implemented using the same PyTorch and Huggingface Transformers framework as in the original codebase.

G.2 Scaling Analysis: Dynamics and Alignment in Larger Models

We extend our main experimental setup by training a larger model, Qwen2.5-1.5B, under the same base configuration. Our goal is to examine whether the developmental trajectory of multi-hop reasoning observed in smaller models persists at scale, and to further investigate the alignment between behavioral accuracy and representational metrics.

The training progresses successfully through Phase I (memorization) and Phase II (in-distribution generalization), and reaches Phase III (cross-distribution generalization). However, we observe increased instability during Phase III: although the model demonstrates the ability to generalize to Test-OI queries, the performance exhibits significant fluctuations. We report the Test-II and Test-OI accuracies, alongside the ID Cohesion and OOD Alignment metrics (Figure 13).

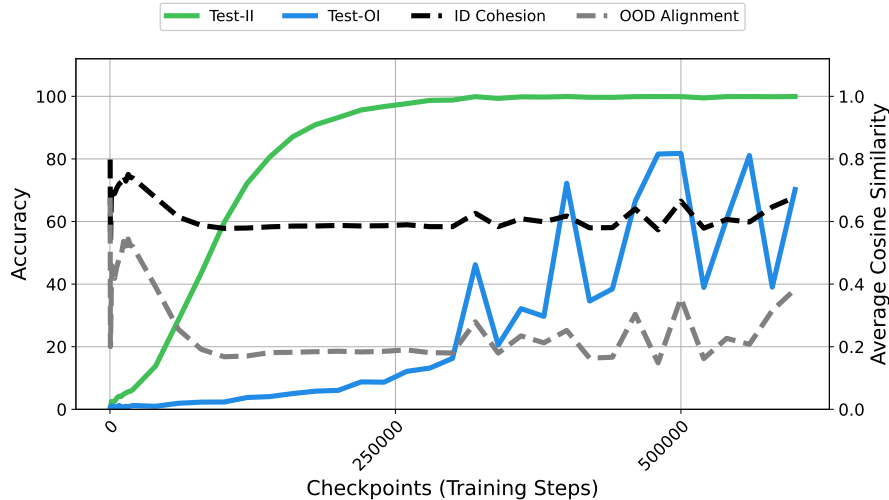


Figure 13: Test-II and Test-OI accuracy and ID Cohesion and OOD Alignment metrics over training steps in the Qwen2.5-1.5B model. Although the model reaches Phase III generalization, substantial variance is observed in both Test-OI accuracy and OOD Alignment, which nonetheless remain tightly coupled. In contrast, Test-II accuracy does not closely track the ID Cohesion metric, suggesting a representational bottleneck at the second relational step.

Interestingly, we find that the Test-II accuracy does not rise in lockstep with the ID Cohesion metric, in contrast to the strong correlation observed in smaller models (Figure 2). We interpret this decoupling as evidence that ID Cohesion is a necessary but not sufficient condition for successful Test-II generalization. While a coherent latent space is required to support compositional reasoning, achieving high Test-II accuracy also depends on the model’s ability to utilize these representations in **executing the second relational step**. In other words, beyond aligning representations, the model

must also learn to map from an aligned intermediate state to the correct final output via the second-hop relation.

In smaller models, these two aspects—representation alignment and second-hop reasoning—tend to emerge together as part of a single learning phase, leading to tight coupling between ID Cohesion and Test-II performance. In contrast, larger models appear to decouple these processes: representational clustering may occur early, while second-hop reasoning capabilities require additional training to fully mature. As a result, the second relational step becomes the **dominant bottleneck** in Phase II generalization.

This interpretation is further supported by the close alignment between the OOD Alignment metric and Test-OI accuracy. Because second-hop reasoning over ID triples is already well established by Phase II, generalization on Test-OI becomes predominantly constrained by whether OOD-derived intermediate representations have successfully aligned with the ID-centric latent space. This tight correlation holds across model scales: in both the main 2-hop setup with smaller models and the 3-hop results in Appendix C (e.g., alignment between Test-OII and OOD-derived clustering), the emergence of Phase III generalization closely tracks OOD Alignment. In our large-model experiment, although the Test-OI accuracy exhibits high variance, its fluctuations are closely mirrored by the OOD Alignment metric, reinforcing our hypothesis.

We leave the optimization of Phase III training strategies for larger models to future work. Our findings suggest that alignment-based representational diagnostics may serve as useful guides for tuning training schedules or data exposure in this regime, and we encourage future work to explore these directions further.

H Validation of Phase III Emergence via ID/OOD Ratio Ablation

To validate our hypothesis in Section 4.1 that the emergence of cross-distribution generalization (Phase III) depends on the dominance of in-distribution (ID) supervision, we conduct an ablation study by varying the ID/OOD ratio of atomic triples under the base configuration.

Experimental Setup. We begin with the full base configuration, which includes all atomic triples (both ID and OOD) and the complete set of Train-II queries. We fix the total number of atomic triples and vary the ID/OOD ratio while keeping all other components of training unchanged. Specifically, we test three settings: 80% ID / 20% OOD, 50% ID / 50% OOD, and 30% ID / 70% OOD. In all cases, the **Train-II / ID ratio** is held constant, as prior work Wang et al. [24] identifies this as a critical factor for Phase II generalization.

Results. All three ID/OOD configurations unsurprisingly reach Phase I and Phase II, to focus on the emergence of Phase III, we report the Test-OI accuracy and OOD Alignment Score, which capture the model’s ability to reason across distributions and align OOD-derived intermediate representations with the ID-induced cluster structure.

As shown in Figure 14, in the 0.8/0.2 setting, both Test-OI accuracy and OOD Alignment Score increase together during training, indicating that the model successfully assimilates OOD-derived intermediate representations into the ID-induced subspace and is able to reuse them for cross-distribution reasoning. In contrast, in the 0.5/0.5 and 0.3/0.7 settings, both metrics remain consistently low, suggesting that the model fails to form aligned representations for OOD triples and consequently cannot generalize to Test-OI queries. This divergence across configurations highlights that strong ID supervision is essential for enabling Phase III generalization.

I Evidence for Representation Clustering from the Decodable Subspace

To validate the representational mechanism discussed in Section 4.2 that ID triple supervision constrains the r_1 representation to lie in a decodable subspace, we design an ablation experiment to test whether held-out ID triples can be recovered solely through shared compositional contexts in Train-II queries.

Specially, We randomly select a subset of ID triples (e.g., $(e_A, r_1) \rightarrow e_B$) to exclude from the atomic triple training set. These held-out triples are removed from all atomic query contexts but remained involved in Train-II queries (e.g., $(e_A, r_1, r_2) \rightarrow e_C$, where e_B serves as the intermediate entity).

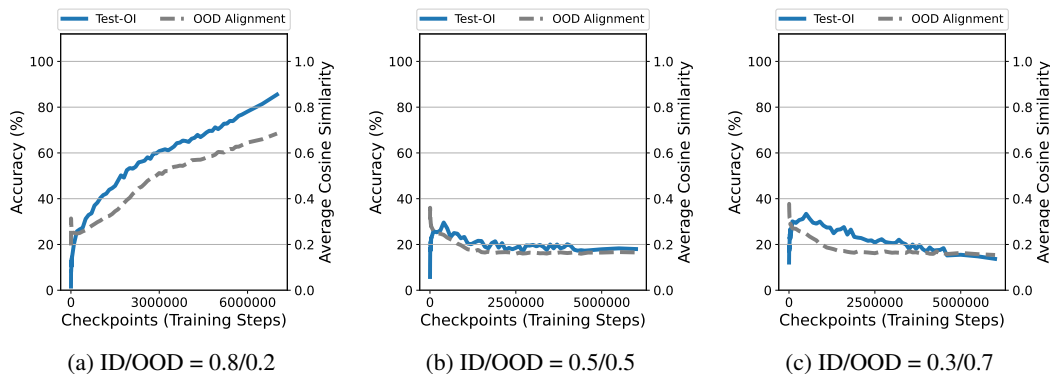
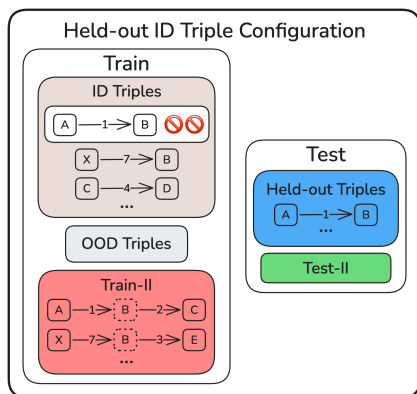


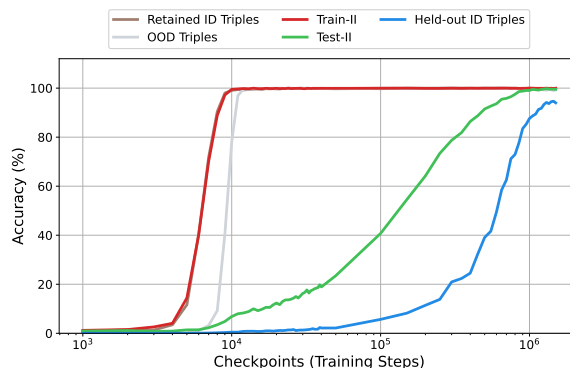
Figure 14: **Test-OI accuracy and OOD Alignment Score** under different ID/OOD splits. All configurations successfully reach Phase I and Phase II. Only the 0.8/0.2 setting supports Phase III generalization, as indicated by joint increases in Test-OI accuracy and OOD alignment. Lower-ID settings fail to align OOD-derived bridge entities, preventing cross-distribution reasoning.

Crucially, the model retains exposure to other atomic triples that sharing the same tail entity, such as $(e_X, r_7) \rightarrow e_B$, which appear in both atomic and corresponding compositional queries (e.g., $(e_X, r_7, r_3) \rightarrow e_Y$), Figure 15a illustrates the configuration. This configuration enables us to validate whether ID atomic triples supervision constrains the r_1 hidden state to a decodable region by testing whether these held-out triples could be recovered.

As illustrated in the Figure 15b, The model successfully recovers held-out triples despite their absence from atomic training. Crucially, this recovery capability emerges concurrently with Test-II generalization, confirming that the model leverages the same intermediate representation subspace for both atomic and compositional reasoning. The results indicate that ID triple supervision accelerates generalization not by providing explicit factual memorization, but by structurally constraining the model’s representational space to align atomic and compositional reasoning pathways.



(a) Illustration of the data construction.



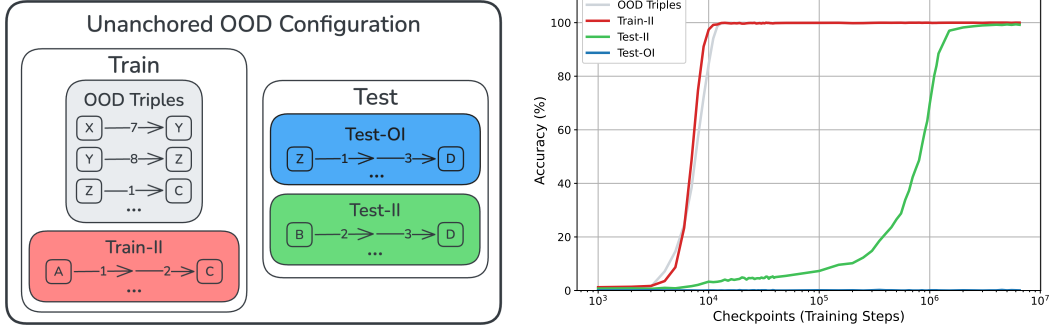
(b) Training curve showing accurate prediction of the held-out triples.

Figure 15: **Verification of ID triple constraint effect.** The model successfully recovers held-out ID triples by leveraging representational constraints from multi-hop supervision and structurally related retained triples, supporting the claim that ID triples accelerate generalization by shaping a decodable representational subspace.

J Removing ID Triples Breaks First-Hop OOD Generalization

To test whether ID supervision is necessary for cross-distribution (Test-OI) generalization, we constructed a simplified configuration that removes all ID triples from training. The model is trained only on OOD atomic triples and Train-II 2-hop queries, as illustrated in Figure 16a.

744 Despite having access to OOD facts and multi-hop supervision, the model fails to generalize to Test-OI
 745 queries where the first hop is from an OOD triple. As shown in Figure 16b, Test-OI accuracy remains
 746 near chance throughout training. This validates our claim in Section 4.3: without representational
 747 anchoring from ID triples, OOD-derived entities cannot support implicit multi-hop reasoning.



(a) Illustration of the Unanchored OOD Configuration. (b) Training dynamics: Test-OI accuracy fails to improve

Figure 16: Validation experiment under ID-removed configuration. Without ID triples, the model fails to reach Phase III, confirming that representational anchoring from ID supervision is essential for OOD-based generalization.