FROM ATOMIC TO COMPOSITE: REINFORCEMENT LEARNING ENABLES GENERALIZATION IN COMPLEMENTARY REASONING

Sitao Cheng[1†], **Xunjian Yin**[2], **Ruiwen Zhou**[3], **Yuxuan Li**[1], **Xinyi Wang**[4], **Liangming Pan**[5†], **William Yang Wang**[6], **Victor Zhong**[1†]

[1]University of Waterloo [2]Duke University [3]National University of Singapore
[4]Princeton University [5]Peking University [6]University of California, Santa Barbara

{sitao.cheng, victor.zhong}@uwaterloo.ca liangmingpan@pku.edu.cn

ABSTRACT

Does Reinforcement Learning (RL) merely amplify existing skills, or synthesize novel skills? We investigate this open question through the lens of *Complementary Reasoning*: the critical practical capability of integrating internal knowledge with external context, a prerequisite for reliable Retrieval-Augmented Generation. Using a controlled synthetic dataset of open-domain biographies to avoid contamination, we decompose this capability into two atomic skills: *Parametric Reasoning* (retrieving facts encoded in model weights) and *Contextual Reasoning* (processing novel information in the context window). We present two key findings. First, models supervised directly on the composite task achieve high accuracy on seen facts and reasoning paths (90%) but collapse on novel facts and reasoning paths (18%), indicating that Supervised Fine-Tuning (SFT) relies on rote memorization rather than true skill integration. Second, RL acts as a reasoning skill *synthesizer* rather than a mere amplifier, successfully bridging this generalization gap. However, we uncover a prerequisite: RL can only synthesize new composite strategy if the base model has first mastered the independent atomic skills via SFT. These results suggest that decoupled atomic training followed by RL offers a scalable path to synthesizing complex novel skills.

1 INTRODUCTION

The evolution of Large Language Models (LLMs) is driven by Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL) (Team et al., 2024; Achiam et al., 2023). While SFT imparts foundational knowledge, its maximum likelihood nature favors *memorization*, often limiting out-of-distribution generalization (Chu et al., 2025). Conversely, RL is hypothesized to transform the distribution towards goal-oriented problem-solving (Schulman et al., 2017). While recent "Thinking" models (Guo et al., 2025) show that extensive RL improves reasoning via test-time computation, the mechanism of this improvement remains debated. We investigate by addressing two questions: ***1)*** *Does RL incentivize the acquisition of new reasoning skills, or merely amplify existing skills found in the SFT distribution?* ***2)*** *What training strategies and data conditions are strictly necessary for an LLM to generalize to complex reasoning?*

Studying these questions requires **1)** disentangling skills before and after training, and **2)** a "clean" environment where skill sufficiency is verifiable without pre-training contamination. Prior attempts yield conclusions: some suggest RL synthesizes complex skills (Yuan et al., 2025; Liu et al., 2025), while others characterize RL as a probability amplifier (Wu & Choi, 2025; Yue et al., 2025). Resolving this requires a controlled setting. Prior studies relying on open-domain math or coding benchmarks suffer from data contamination and an inability to disentangle knowledge retrieval from reasoning mechanics. Furthermore, these domains often define "new skills" as the extrapolation of

[†]Corresponding author.

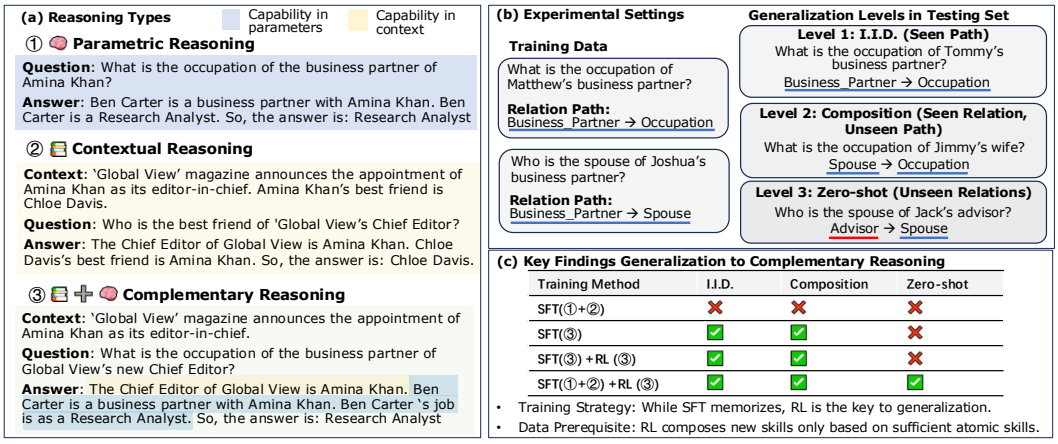

Figure 1: Our settings and findings. (a) Examples of Complementary Reasoning requiring both Parametric and Contextual skills. (b) Evaluation protocol across three levels of difficulty. __ and __ denotes seen and unseen pattern, respectively. (c) The SFT Generalization Paradox: Models trained on atomic skills generalize better via RL than models trained directly on the composite task.

existing ones (*e.g.,* longer reasoning chains) and rely on a narrow set of relations (*e.g.,*$+, -, *$) — they do not test the model's ability to synthesize answers from truly novel relational patterns.

We propose **Complementary Reasoning (COMP)** as an ideal testbed to address these. Beyond its utility for Retrieval-Augmented Generation, COMP is a rigorous proxy for reasoning capability: it forces the model to bridge internal parametric knowledge with external contextual information to form novel logical connections. In Figure 1(a), asking *"What is the occupation of the business partner of Global View's chief editor"* requires integrating external information (Amina Khan is the chief...) with parametric knowledge (Ben Carter is a business...). Unlike math, where rules (*e.g.,* the distributive property) are fixed, this task demands complementary composition of seen parametric and novel contextual knowledge simultaneously, challenging for LLMs (Cheng et al., 2024a). To investigate this, we decouple Complementary Reasoning into the composition of two atomic skills: **Parametric Reasoning** ($\mathcal{C}_{\text{MEM}}$), which relies on facts encoded within model weights, and **Contextual Reasoning** ($\mathcal{C}_{\text{CTX}}$), which relies on novel information provided in context windows.

Investigating COMP on open-domain benchmarks (*e.g.* HotpotQA (Yang et al., 2018)) is limited by data contamination: in web-scale corpora, distinguishing reasoning from recalling pre-training shortcuts is impossible. We address this by constructing a dataset of biographies underpinned by a knowledge graph of real-world relations (*e.g.,* 'Father') but populated with synthetic entities. This prevents reliance on shortcuts, enforcing the boundary between Parametric and Contextual knowledge (Allen-Zhu & Li, 2023a;b). We evaluate across three generalization levels (Figure 1(b)): *IID* (seen paths), *Composition* (unseen paths of seen relations), and *Zero-shot* (unseen relations). Using controlled multi-hop QA pairs, we train models under varied data combinations and strategies (*e.g.,* SFT, RL) to identify the optimal settings for generalization.

Our experiments uncover the SFT Generalization Paradox (Figure 1(c)): contrary to expectation, training directly on the composite target task via SFT fails to generalize out-of-distribution despite high in-distribution performance. However, RL catalyzes generalization — particularly in structural zero-shot settings where relation paths are entirely novel — provided the model first captures the necessary atomic skills. This challenges the view of RL as merely a probability amplifier. Our findings suggest a scalable path: rather than collecting expensive complex reasoning traces, one can efficiently teach atomic skills via SFT, then leverage RL to unlock generalization for complex reasoning. We summarize our key contributions:

• We define Complementary Reasoning and introduce a controlled dataset decoupling the complex ability into atomic Parametric and Contextual skills and enabling rigorous generalization evaluation.

• We provide empirical evidence that while SFT is sufficient for memorizing distributions, RL is essential for generalization, specifically enabling the model to tackle structural zero-shot combinations that SFT fails to resolve.

• We uncover a prerequisite for RL generalization: RL synthesizes new complex skills **only when the base model possesses sufficient atomic skills**. A model with atomic skills gains significantly more from RL than from direct SFT on composite data regardless of data amount.

2 PROBLEM DEFINITION: COMPLEMENTARY REASONING

To systematically investigate the limitations of SFT and the mechanism of RL-driven generalization, we must first formalize the mechanical components of reasoning. In this section, we define the atomic skills, the composite task, and the specific gradations of difficulty used to stress-test the model's capabilities beyond the training distribution.

Reasoning Types We define ***Complementary Reasoning*** (**COMP**) as a task requiring both parametric and contextual skills, which complement each other. Accordingly, we have ***Parametric Reasoning*** **(MEM)** requiring skills stored in a model's weights and ***Contextual Reasoning*** **(CTX)** requiring skills in the context window. Formally, we define the capability requirement as a logical conjunction:

$$\mathcal{C}_{\text{COMP}} \iff \mathcal{C}_{\text{MEM}} \wedge \mathcal{C}_{\text{CTX}}.$$

This implies that a failure in either atomic skill ($\mathcal{C}_{\text{MEM}}$ or $\mathcal{C}_{\text{CTX}}$) necessitates a failure in complementary task. Intuitively, humans find it straightforward to tackle complementary tasks if the new (unknown) information is given. However, while fluent in $\mathcal{C}_{\text{MEM}}$ and $\mathcal{C}_{\text{CTX}}$ by training with large-scale Context-Question-Answer examples, LLMs struggle to generalize to $\mathcal{C}_{\text{COMP}}$ (Cheng et al., 2024a).

We study generalization to complementary reasoning through **multi-hop factual reasoning**. This task typically requires retrieval of knowledge and linking of several facts, formally defined as traversing a relation path $P = (r_1, r_2, \ldots, r_k)$ starting from a topic entity to the answer. For COMP, each relation r_i in P falls into either parametric or contextual knowledge, requiring the seamless integration of both sources. In Figure 1(a), "*What is the occupation of the business partner of Global View's new chief editor?*" traverses the path "*Global View – Chief Editor – Business Partner – Occupation – Answer*".

Generalization Levels Traditional random data split assumes that testing and training data is independent and identically distributed (IID). However, this assumption fails to capture the complexity of real-world reasoning. For instance, a web-agent often encounters infinite combinations of operations that cannot be exhaustively covered in a finite training set (Deng et al., 2023). To rigorously investigate whether models can transcend rote memorization to novel scenarios, we evaluate performance across *three levels of difficulty* based on the novelty of the relation path (Gu et al., 2021; Huang et al., 2023a). Given the relation path P, let $\mathcal{P}_{train}$ denote the set of relation paths in the training set, and $\mathcal{R}_{train}$ denote the set of individual relations in those paths. We categorize the generalization levels as follows (exemplified in Figure 1(b)):

• ***IID Generalization*** evaluates the ability to apply learned patterns. The relation path tested is fully observed during training ($P_{test} \in \mathcal{P}_{train}$). For example, if the model is trained on *Business_Partner→Job* (*e.g.,* "the job of [X]'s business partner?"), an IID test query would query the same structure for a different entity. Our experiments show that SFT is typically sufficient for this level as it merely requires recalling the observed path structure.

• ***Composition Generalization*** tests the ability to recombine in-distribution primitives into novel reasoning patterns. While every individual relation constituting the path has been observed in disjoint contexts during training ($\forall r \in P_{test}, r \in \mathcal{R}_{train}$), the specific sequence of relations is unseen ($P_{test} \notin \mathcal{P}_{train}$). For example, suppose the training set contains *Business_Partner→Job* and *Business_Partner→Spouse*. A Composition test might query *Spouse→Job*. Although the model knows both relations independently, it must synthesize them into a new compound reasoning path without explicit prior demonstration.

• ***Structural Zero-shot Generalization*** is the most challenging. Unlike standard zero-shot prompting, it tests unseen relational primitives during instruction-tuning. The relation path involves at least one relation never seen in any QA pair in train set ($\exists r \in P_{test}, r \notin \mathcal{R}_{train}$). While the model may semantically understand the relation "*Advisor*" from pre-training, it has never been explicitly instructed to retrieve or reasoning with it in the task format (i.e. "the spouse of [X]'s advisor").

3 EXPERIMENT SETUP

3.1 SYNTHETIC HUMAN BIOGRAPHIES

To avoid data contamination and control knowledge sufficiency, we construct a synthetic dataset of human biographies grounded in a relational Knowledge Graph (KG) (Allen-Zhu & Li, 2023a;b).

We specify 39 relations, randomly partitioned into disjoint Parametric and Contextual sets. We populate the KG with meaningful synthetic facts (*e.g.,* name, birthday) using deterministic tools to ensure no overlap with pre-training corpora. We generate 10k biographies for MEM and CTX, with 5k shared characters to enable bridging, simulating real-world scenarios where new facts attach to known entities (details in Appendix B).

QA Construction from KG To construct multi-hop questions, we sample relation paths consisting of 2–5 hops, ensured to be distributed identically for each reasoning type. We enforce a constraint that intermediate nodes must be entities rather than generic values (*e.g.,* dates, emails) to ensure multi-step dependency. For COMP, we incorporate relations from both MEM and CTX. Then, we generate three linguistically diverse question and answer (with CoT) templates to ensure syntactic variety without relying on repetitive patterns with an LLM. Details in Appendix B.

Data Split for Generalization Levels To ensure the validity of generalization levels, we implement a rigorous multi-stage splitting protocol at the *relation path level*: **1) IID**: we reserve a subset of relation paths $\mathcal{P}_{\text{IID}}$ for both train and test. To prevent rote memorization, paths in $\mathcal{P}_{\text{IID}}$ are identical between train and test, but *entities* are disjoint. **2) Structural Zero-shot**: We designate a subset of relations as $\mathcal{R}_{\text{unseen}}$. *Every* path containing a relation $r \in \mathcal{R}_{\text{unseen}}$ is assigned to the zero-shot test set. This guarantees that the model never encounters these unseen relations during SFT. **3) Composition**: We randomly partition the remaining paths (*i.e.,* not IID, not Structural Zero-shot) into train set $\mathcal{P}_{\text{train\_comp}}$ and test set $\mathcal{P}_{\text{test\_comp}}$. Crucially, we verify that the sets of single relations present in the training and test splits are identical ($\mathcal{R}_{\text{test}} \equiv \mathcal{R}_{\text{train}}$), ensuring that the difficulty arises solely from novel *combination* of known relations.

Through this protocol, we guarantee that IID tests seen paths, Composition tests unseen paths composed of seen relations, and Zero-shot tests paths with unseen relations. We then construct QA pairs via random-walk over the KG. Details in Appendix B.

3.2 TRAINING PROTOCOL

We adopt a standard SFT followed by RL pipeline using Qwen-2.5-1.5B. **Input Formulation**: For MEM, the input is only the *Question*. For CTX and COMP, inputs include the *Context* (new facts) and *Question*. **Training**: SFT uses standard next-token prediction on QA pairs and parametric biographies. RL uses Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with binary outcome rewards. Evaluation uses exact matching on the COMP test set.

3.3 BASELINE ANALYSIS: THE LIMITS OF SUPERVISED FINE-TUNING

To demonstrate why SFT alone is insufficient for robust Complementary Reasoning, we analyze task difficulty and the generalization capabilities of two SFT baselines: a model trained on atomic skills ($\text{SFT}_{\text{MEM+CTX}}$) and one trained directly on the target composite task (SFT_{COMP}).

Complementary reasoning is data-hungry and inherently difficult. Table 1 shows that achieving high IID performance on COMP requires significantly more data (∼180k) than on atomic tasks MEM (∼88k) and CTX (∼3k). This validates that integrating internal and external knowledge is structurally more complex than applying skills in isolation.

Table 1: Statistics of train and test set.

Group	Train	IID	Com.	0-shot
MEM	88,031	1,921	1,141	782
CTX	2,651	1,910	1,320	453
COMP	180,919	2,135	1,415	918

Atomic skills do not spontaneously compose. A key question is whether teaching a model the necessary components (MEM and CTX) is sufficient to infer the composite skill. Table 2 shows that $\text{SFT}_{\text{MEM+CTX}}$ shows poor performance on COMP test set. While non-zero (indicating some transfer), it lags far behind the explicit training baseline (90.26% IID), confirming that possessing atomic skills does not guarantee their integration without further guidance (Cheng et al., 2024a; Yin et al., 2023). Analysis details in Appendix F.1.

Table 2: SFT results on COMP test set.

Train Data	IID	Com.	0-shot
MEM + CTX	35.18	28.20	24.07
COMP	90.30	76.25	18.41

SFT Memorizes rather than Generalizes. While COMP achieves a near-perfect 90.26% on IID, it collapses to 18.41% on Zero-shot (Table 2). This indicates that SFT incentivizes memorizing

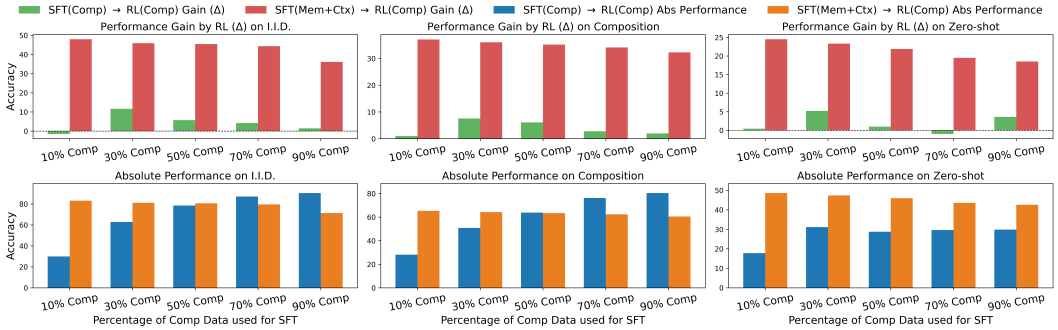

Figure 2: Comparison of training with different complementary data proportions. The top row shows the gain from RL, while the bottom row presents the absolute performance after RL. RL generalizes to Complementary Reasoning only from the model with both Parametric and Contextual skills.

relational patterns seen during training; When faced with novel relations where memorization is impossible, the superficial reasoning learned via SFT fails.

4 Reinforcement Learning Enables Generalization

We hypothesize that generalization requires a specific curriculum: *establishing sufficient atomic abilities via SFT, followed by synthesizing complex skills via RL.* Formally, we propose the recipe $\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$ as the optimal path to generalization. We validate this recipe by **sufficiency** (consistent performance) and **necessity** (data and training prerequisites for generalization).

4.1 Sufficiency: RL as a Skill Synthesizer

To validate if our recipe consistently yields superior generalization compared to direct tuning on the target task, we conduct a controlled experiment by varying the scale of COMP data. More comparisons are in Appendix E.

Settings. A common practice for complex tasks is SFT cold-start followed by extensively RL (Guo et al., 2025). However, it is unclear how much data is needed to SFT or RL for generalization. To study this, we partition COMP training data into two subsets: $x\%$ for SFT, and $(100-x)\%$ for RL. We compare two recipes using identical RL data: **1)** $\mathbf{SFT_{COMP} \rightarrow RL_{COMP}}$: SFT on the $x\%$ data, then RL on the remaining $(100-x)\%$; **2)** $\mathbf{SFT_{MEM+CTX} \rightarrow RL_{COMP}}$: SFT on all atomic MEM +CTX data (about 50% of COMP data in amount) and then RL on the same $(100-x)\%$ COMP data. We vary $x\%$ from 10% to 90%. Figure 2 illustrates relative performance gain derived specifically from the RL stage (top row) and final absolute performance (bottom row).

RL efficiently synthesizes from atomic skills. The top row in Figure 2 reveals a striking contrast. Regardless of data volume (10-90%), $\text{SFT}_{\text{MEM+CTX}}$ (red bars) achieves massive performance gains from RL with COMP data across all generalization levels. In contrast, SFT_{COMP} (green bars) shows inability to improve, gaining almost nothing from RL. This indicates that without atomic foundations, RL merely optimizes within an existing distribution rather than effectively exploring reasoning paths. Conversely, given atomic skills, RL acts as a powerful synthesizer, actively composing known facts into new reasoning skills.

Zero-shot generalization of RL over atomic skills is consistently superior. In the critical Structural Zero-shot setting (Bottom Row, 3rd Column) where memorization is impossible, $\text{SFT}_{\text{MEM+CTX}}$ (yellow bars) consistently and significantly outperforms SFT_{COMP} (blue bars) regardless of data scale. This confirms that our recipe fosters genuine generalization to unseen relations, whereas direct training on complementary data fails to extrapolate beyond the training distribution.

The SFT Generalization Paradox: Distinguishing Generalization from Memorization. We observe a nuanced result in the IID and Composition settings (Bottom Row, 1st & 2nd Columns). As the data portion x exceeds 70%, absolute performance of SFT_{COMP} (blue bars) surpasses $\text{SFT}_{\text{MEM+CTX}}$ (yellow bars) on IID tasks but fails on Zero-shot. We term this the **memorization trap**: SFT on composite data encourages overfitting to specific training relation paths rather than learning the underlying reasoning algorithm. RL on atomic skills avoids this trap by forcing the model to derive the path dynamically, resulting in robust Zero-shot performance.

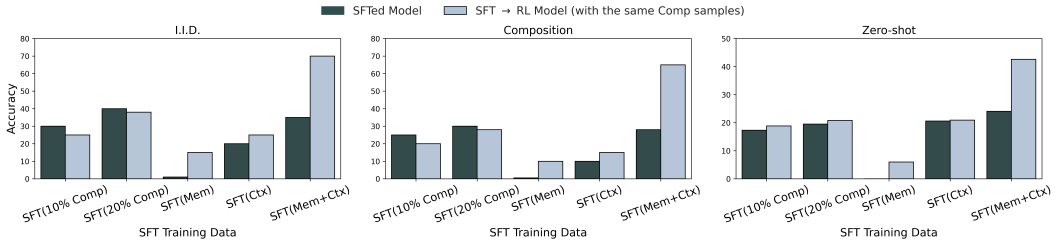

Figure 3: Necessity of atomic skills for RL generalization. We conduct RL with the same amount of COMP data from different SFT trained models. Only $\text{SFT}_{\text{MEM+CTX}}$ generalizes well in all levels.

4.2 NECESSITY: ATOMIC SKILLS AS PREREQUISITES FOR GENERALIZATION

While Section 4.1 establishes "$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$" as sufficient for generalization, questions regarding the boundary conditions remain: **1)** Are both atomic skills strictly necessary, or can the model generalize from partial skills? **2)** Is RL uniquely required, or can other strategies suffice given the same base model?

Necessity of Sufficient Atomic Skills (Data Condition). To determine if complete atomic capabilities are prerequisite for generalization, we compare $\text{SFT}_{\text{MEM+CTX}}$ against three baselines with similar initial performance but deficient atomic foundations: **1)** SFT_{MEM} (SFT on MEM only); **2)** SFT_{CTX} (SFT on CTX only); **3)** $\text{SFT}_{10\%\text{COMP}}$ and $\text{SFT}_{20\%\text{COMP}}$ (SFT on COMP data with comparable initial performance). We apply identical RL training with a fixed subset of 12.8k randomly sampled COMP data to all models. Figure 3 shows COMP results before and after RL.

Removing any atomic skill collapses generalization. Models trained via SFT solely on COMP or CTX fail to generalize significantly after RL, demonstrating that Complementary Reasoning is not merely an additive task but a synthesis requiring the sufficiency of atomic skills. **Generalization potential is driven by foundational capability, not initial metrics.** Notably, while $\text{SFT}_{10\%\text{COMP}}$ and $\text{SFT}_{20\%\text{COMP}}$ exhibit match $\text{SFT}_{\text{MEM+CTX}}$'s initial performance, their post-RL gain is negligible. In contrast, $\text{SFT}_{\text{MEM+CTX}}$—with both parametric and contextual reasoning skills—nearly doubling its performance in all generalization levels. Furthermore, even initially low-performing SFT_{MEM} and SFT_{CTX} models gain more from RL than the model SFTed with COMP data. This confirms that underlying atomic skills, not initial performance, are the true predictor of RL success.

Necessity of RL for Generalization (Training Strategy). Given sufficient atomic skills, we compare training strategies (SFT, LoRA (rank=256), and RL) with identical COMP samples. Figure 4 shows that all strategies significantly improves the model $\text{SFT}_{\text{MEM+CTX}}$, yet reveals a dichotomy between memorization and generalization. While further SFT (blue) yields the highest IID performance (indicating memorization of seen patterns), it lags significantly behind RL (green) in Zero-shot. RL outperforms both SFT and LoRA on unseen relational combinations, confirming that while SFT excels at pattern matching, **RL is uniquely necessary to incentivize the active composition of skills required for out-of-distribution reasoning**.

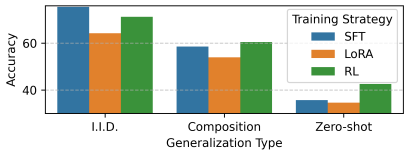

Figure 4: Performance of different training strategies.

5 ANALYSIS OF THE MODEL WITH SUFFICIENT ATOMIC ABILITIES

Having established atomic skills as the prerequisite for RL generalization, we analyze the features of base model before RL (*i.e.,* $\text{SFT}_{\text{MEM+CTX}}$ and SFT_{COMP}) from the perspective of sample efficiency and pass@k performance. See Appendix F for additional analyses on training dynamics, impact of training loss, PCA analysis of latent space, and uncertainty analysis.

5.1 SAMPLE EFFICIENCY COMPARISON

We examine **whether learning atomic skills induces better sample efficiency than learning the composite task directly**, both before and after RL.

Atomic skill learning requires less SFT data to prime RL generalization. To compare sample efficiency, we reserve a fixed subset (∼90k) of COMP data exclusively for the RL stage. For the SFT

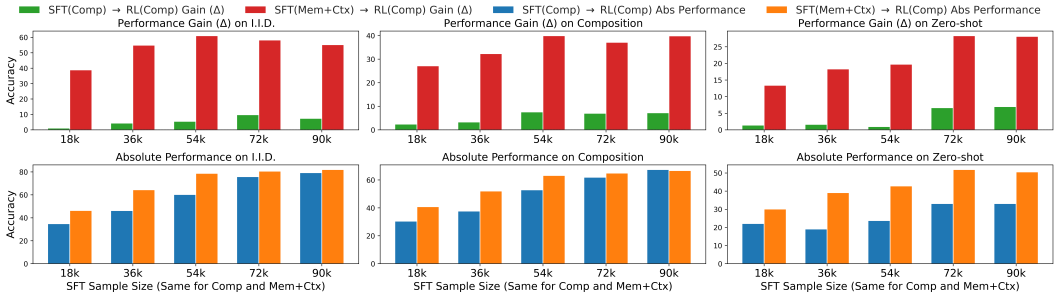

Figure 5: Performance with the same amount of SFT and RL data. We compare $\text{SFT}_{\text{MEM+CTX}}$ and SFT_{COMP} by relative (the top row) and absolute performance (the bottom row) after RL_{COMP}.

stage, we define a data budget (sweeping 20%–100% of the non-reserved volume) to construct two strictly comparable datasets: **1) SFT_{COMP}:** composite samples; **2) $\text{SFT}_{\text{MEM+CTX}}$:** atomic samples (balanced MEM and CTX). Crucially, for a fair comparison, we control not just *number of samples*, but also the *information content*: the distribution of reasoning steps (hop counts) is identical across SFT_{COMP} and $\text{SFT}_{\text{MEM+CTX}}$, ensuring that the models are exposed to equivalent reasoning complexities. We then apply RL using the reserved set. Figure 5 shows that $\text{SFT}_{\text{MEM+CTX}}$ (red/orange) consistently outperforms SFT_{COMP} (green/blue) across all data scales. Even with minimal SFT data ($\sim$18k), $\text{SFT}_{\text{MEM+CTX}}$ is successfully "primed" for RL, whereas SFT_{COMP} struggles, confirming that atomic skills provide a significantly more efficient foundation for complex reasoning.

Sufficient atomic skills enable few-shot adaptation. We investigate how much COMP data is needed to "trigger" generalization once sufficient atomic skills are established. Fixing $\text{SFT}_{\text{MEM+CTX}}$ as base model, we further train using RL, SFT, or LoRA with varying sizes of data, ranging from a tiny set (50 samples) to a medium set (12.8k, $< 10\%$ of total). We compare this against an upperbound baseline: SFT on 100% COMP data. In Figure 6 reveals **rapid adaptation:** with just 50 samples, $\text{SFT}_{\text{MEM+CTX}}$ significantly improces in complementary reasoning, regardless of training strategy. In addition, we also observe **data efficiency:** with $<$10% of data, $\text{SFT}_{\text{MEM+CTX}}$ effectively matches the upperbound of SFT_{COMP} trained on the entire dataset (purple dotted line). This demonstrates that once atomic skills are acquired, the cost of assembling them into a complex reasoning skill is remarkably low.

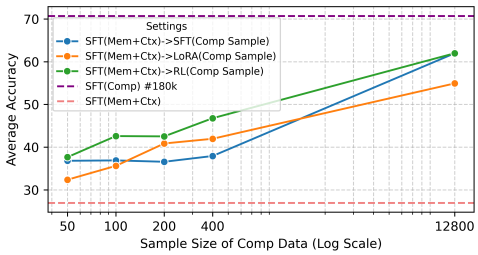

Figure 6: Few-shot adaptation of $\text{SFT}_{\text{MEM+CTX}}$ with different training strategies.

5.2 THE ROLE OF RL: SYNTHESIZER VS. AMPLIFIER

We now address the debate: Does RL incentivize genuinely new reasoning capabilities, or merely re-weight the probability of existing skills already present in the base distribution? We analyze *pass@k* performance before and after RL (Yue et al., 2025), varying k from 2^0 to 2^9 to distinguish these roles: **1) *Synthesis:*** Divergent curves at large k imply that RL synthesizes novel skills that SFT effectively does not discover via sampling. **2) *Amplification:*** Merging curves imply that skills are latent in the SFT distribution, and that RL merely amplifies their generation probability. We compare our $\text{SFT}_{\text{MEM+CTX}}$ against a strong baseline SFT_{COMP} using identical RL data. Figure 7 presents the results over different generalization levels.

RL synthesizes new pathways for models with sufficient atomic skills. The top row shows parallel scaling: RL performance (orange) remains significantly higher than the SFT baseline (blue) even at $k = 2^9$. The persistent gap of pass@k curves suggests a fundamental shift in mechanism. If RL were merely amplifying latent behaviors, the SFT model (given enough attempts k) would eventually find the solution. The persistent gap indicates that RL has synthesized a robust reasoning mechanism—*specifically, the mechanism to bridge context and memory*—that is effectively absent from the SFT distribution. This provides empirical evidence that, given sufficient atomic priors, RL creates new capabilities rather than just re-weighting existing ones. Notably, this discovery of new reasoning mechanisms occurs not only for zero-shot but also on other settings, suggesting that RL optimizes the logic of combination itself.

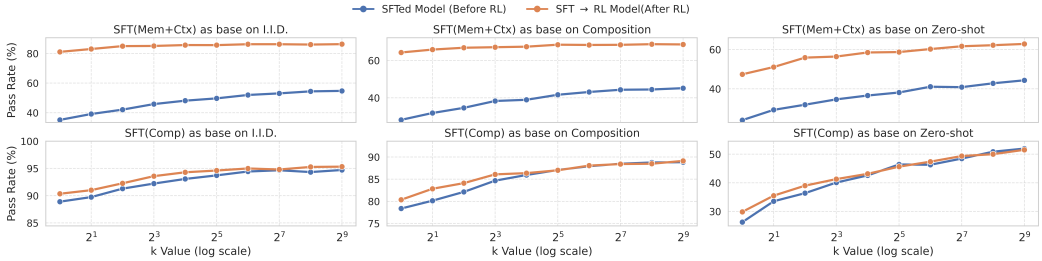

Figure 7: Pass@k comparison for $\text{SFT}_{\text{MEM+CTX}}$ and SFT_{COMP}. It shows that RL synthesizes new compositional skills only based on models with sufficient atomic skills.

Without sufficient atomic skills, RL merely amplifies existing behaviors from composite skills. In contrast, the bottom row shows that for SFT_{COMP}, the curves rapidly converge as k increases. By $k = 2^5$, the SFT model nearly matches the RL model. This indicates that as the model trained on composite data memorized the target distribution during SFT, RL serves primarily as an amplifier—boosting the likelihood of the optimal reasoning paths without discovering fundamentally new ones. This dichotomy confirms that sufficient atomic skills is the prerequisite for RL to function as a synthesizer. Without them, RL degenerates into a mere probability amplifier.

5.3 Case Study

We analyze failure modes by examining the **intersection of incorrect samples** before and after RL to identify how RL changes detailed reasoning behavior. For each persistent error, we align the generated CoT with the ground truth to pinpoint the exact step of deviation from the correct CoT. We then classify the error source (MEM vs. CTX) and calculate **Progress** (the normalized position of the failure step within the path). Table 3 summarizes the resulting patterns.

Table 3: Error analysis. **MEM** and **CTX** denotes error occurring at parametric and contextual relation, respectively. **Prog.** represents the normalized position of the first error step within the reasoning path.

Model	CTX	MEM	Prog.
SFT_{COMP}	90%	10%	54.5%
$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$	86%	14%	45.0%
$\text{SFT}_{\text{MEM+CTX}}$	86%	14%	18.5%
$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$	30%	70%	71.8%

$\text{SFT}_{\text{MEM+CTX}}$, **$\text{SFT}_{\text{COMP}}$**, and **$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$** share failure modes, predominantly characterized by a high prevalence of CTX ($> 85\%$) and early-stage failures (Progress $\leq 54.5\%$). Qualitatively, $\text{SFT}_{\text{MEM+CTX}}$ tends to hallucinate when retrieving information from the provided context, similar to its imitation learning process (*i.e.,* contextual reasoning). Conversely, SFT_{COMP} frequently fail to identify the correct relation when bridging from MEM to CTX, with 62% of errors in SFT_{COMP} getting stuck at the first hop. In contrast, **$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$** flips the distribution: 70% of errors are from MEM, and the average error position occurs much later (71.8%), mostly at the final hop. This indicates that *RL effectively optimizes the bridging logic, pushing failure modes from early contextual retrieval to late parametric recall*.

6 Conclusions

While current LLMs are proficient in multi-hop reasoning with either parametric or contextual knowledge, they struggle to handle questions that require the integration of both knowledge sources. In this paper, we study how LLMs can generalize to Complementary Reasoning with post-training strategies (*i.e.,* SFT and RL). We conduct strictly controlled experiments with our synthetic human biographies based on a relational knowledge graph, constructing QA pairs for atomic skills (*i.e.,* Parametric and Contextual Reasoning) and the compound skill (*i.e.,* Complementary Reasoning). We also split the dataset into different levels of generalization difficulties I.I.D., Composition and Zero-shot. Based on this, we study the effect of training strategies and find that only by firstly SFT with sufficient atomic skills can LLMs generalize via RL. This finding suggests a scalable path for training reasoning LLMs: focusing on teaching the model sufficient fundamental atomic skills via SFT, and then leveraging RL for generalization to OOD complex tasks.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023a.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. *arXiv preprint arXiv:2309.14402*, 2023b.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. Understanding the interplay between parametric and contextual knowledge for large language models. *arXiv preprint arXiv:2410.08414*, 2024a.

Sitao Cheng, Ziyuan Zhuang, Yong Xu, Fangkai Yang, Chaoyun Zhang, Xiaoting Qin, Xiang Huang, Ling Chen, Qingwei Lin, Dongmei Zhang, et al. Call me when necessary: Llms can efficiently and faithfully reason over structured environments. *arXiv preprint arXiv:2403.08593*, 2024b.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the web conference 2021*, pp. 3477–3488, 2021.

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

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*, 2020.

Xiang Huang, Sitao Cheng, Yuheng Bao, Shanshan Huang, and Yuzhong Qu. Markqa: A large scale kbqa dataset with numerical reasoning. *arXiv preprint arXiv:2310.15517*, 2023a.

Xiang Huang, Sitao Cheng, Yiheng Shu, Yuheng Bao, and Yuzhong Qu. Question decomposition tree for answering complex questions over knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12924–12932, 2023b.

Xiang Huang, Sitao Cheng, Shanshan Huang, Jiayu Shen, Yong Xu, Chaoyun Zhang, and Yuzhong Qu. Queryagent: A reliable and efficient reasoning framework with environmental feedback-based self-correction. *arXiv preprint arXiv:2403.11886*, 2024.

Xiang Huang, Jiayu Shen, Shanshan Huang, Sitao Cheng, Xiaxia Wang, and Yuzhong Qu. Targa: Targeted synthetic data generation for practical reasoning over structured data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2704–2726, 2025.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. Disentangling memory and reasoning ability in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1681–1701, 2025.

Minsung Kim, Dong-Kyum Kim, Jea Kwon, Nakyeong Yang, Kyomin Jung, and Meeyoung Cha. Training dynamics of parametric and in-context knowledge utilization in language models. *arXiv preprint arXiv:2510.02370*, 2025.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Amrith Setlur, Matthew Y. R. Yang, Charlie Snell, Jeremy Greer, Ian Wu, Virginia Smith, Max Simchowitz, and Aviral Kumar. e3: Learning to explore enables extrapolation of test-time compute for llms, 2025. URL https://arxiv.org/abs/2506.09026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. *arXiv preprint arXiv:2405.15071*, 2024.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IQxBDLmVpT.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Fang Wu and Yejin Choi. The invisible leash: Why rlvr may not escape its origin. In *2nd AI for Math Workshop@ ICML 2025*, 2025.

Shiming Yang, Yuxuan Tong, Xinyao Niu, Graham Neubig, and Xiang Yue. Demystifying long chain-of-thought reasoning. In *Forty-second International Conference on Machine Learning*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Xunjian Yin, Baizhou Huang, and Xiaojun Wan. Alcuna: Large language models meet new knowledge. *arXiv preprint arXiv:2310.14820*, 2023.

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. From $f(x)$ and $g(x)$ to $f(g(x))$: Llms learn new skills in rl by composing old ones, 2025. URL https://arxiv.org/abs/2509.25123.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and rl on reasoning language models. *arXiv preprint arXiv:2512.07783*, 2025.

LIMITATIONS

Although we have validated our conclusion that only "$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$" generalizes, which well addresses the research questions, we still have some limitations: 1) Although we show from several distinct angles that our finding holds, we only test on Qwen models. While evaluated on the Qwen family, our decomposition of reasoning into atomic priors suggests these findings are likely architecture-agnostic. 2) As we investigate from the sample efficiency, the pass@k performance, the training dynamic, the embedding distributions and case study, it is still difficult to find the mechanisms that enable the $\text{SFT}_{\text{MEM+CTX}}$ model to outperform. Future work should investigate the mechanistic interpretability of how RL circuits recruit atomic attention heads. 3) While our study relies on synthetic biographies, this was a necessary design choice to strictly enforce the boundary between Parametric (known) and Contextual (new) information—a boundary that is impossible to guarantee in web-scale corpora due to pre-training contamination. Future work should validate these findings on controlled splits of real-time benchmarks (*e.g.,* news QA) where the "new" information is strictly dated post-training. Moreover, future work should study how to mix COMP data with MEM and CTX data to improve the overall I.I.D., Composition and Zero-shot performance. 4) It would be interesting to adapt our findings into real-world knowledge-intensive benchmarks, especially to check the OOD testing performance.

A RELATED WORK

Roles of Post-training Strategies The power of LLMs is unleashed fundamentally via SFT followed by RL. While SFT establishes foundational knowledge, its maximum likelihood estimation inherently favors *memorization* of the training distribution, leading to poor generalization (Wang et al., 2025; Chu et al., 2025). In contrast, RL, driven by reward signals, transforms the model's distribution into goal-oriented problem-solving (Schulman et al., 2017). However, whether RL combines new skills (Yuan et al., 2025; Liu et al., 2025), or just amplifying from existing distributions (Wu & Choi, 2025; Yue et al., 2025; Yang et al.; Setlur et al., 2025) is still debated. Our findings stand for RL incentivizes new skills, but only under some prerequisites.

Knowledge-intensive Reasoning A foundational capability for intelligence is to reason with knowledge. Multi-Hop Reasoning, which requires multiple facts to answer, serves as an ideal testbed (Yang et al., 2018; Ho et al., 2020; Huang et al., 2025). Inherently, this task requires the ability to retrieve knowledge from parametric or external bases and logically *compose* intermediate facts (Huang et al., 2023b; Cheng et al., 2024b; Huang et al., 2024; Jin et al., 2025). Recent studies identified that LLMs struggle to generalize when using both new and known knowledge for reasoning (Cheng et al., 2024a; Yin et al., 2023). However, relying solely on recent benchmarks risks data contamination, making it hard to distinguish between new and known knowledge. This necessitates moving beyond standard benchmarks to controlled experimental setups.

Behavioral Study with Synthetic Data To ease the problem of data contamination, *synthetic data* has become essential for evaluation (Allen-Zhu & Li, 2023a;b). With novel facts, entities, or narratives (*e.g.,* human biographies), researchers can strictly control the sufficiency of knowledge and the for the task (Yin et al., 2023; Kim et al., 2025; Yuan et al., 2025; Wang et al., 2024; Zhang et al., 2025). This also allows for rigorous evaluation of the model's capability under controlled settings. Our study adopts a relational knowledge base with templates to study knowledge-intensive reasoning under varying post-training strategies, where we are able to strictly control the parametric and new knowledge. With the synthetic relational patterns, we can systematically examine how LLMs could generalize to *OOD* patterns.

B DETAILS OF DATA CONSTRUCTION

To systematically construct our synthetic human biography, we build a relational knowledge graph with fake information by Python Faker Library [1] and the help of GPT-4o.

[1] https://faker.readthedocs.io/en/master/

Table 4: Templates for Relations in our Knowledge Graph.

Relation	Statement Template	Question Template
address	{e1} resides at {e2}.	What is {e1}'s address?
awards	{e1} won the {e2} award.	What awards has {e1} won?
best_friend	{e1}'s best friend is {e2}.	Who is {e1}'s closest friend?
birth_date	{e1} was born on {e2}.	When was {e1} born?
birth_place	{e1} hails from {e2}.	Where was {e1} born?
boss	{e1} works under {e2}.	Who is {e1}'s boss?
boss_of	{e1} manages {e2}.	Who works under {e1}?
child	{e2} is the child of {e1}.	Who is {e1}'s child?
classmate	{e1} studied alongside {e2}.	Who attended school with {e1}?
colleague	{e1} works alongside {e2}.	Who are {e1}'s colleagues?
died_in	{e1} passed away in {e2}.	Where did {e1} die?
died_on	{e1} died on {e2}.	When did {e1} pass away?
email	You can reach {e1} at {e2}.	What is {e1}'s email address?
favorite_food	{e1}'s favorite dish was {e2}.	What food does {e1} enjoy the most?
first_language	{e1}'s native language was {e2}.	What is {e1}'s first language?
hobby	A favorite activity of {e1} is {e2}.	What does {e1} enjoy doing?
influence	{e1} shaped the career of {e2}.	Who was influenced by {e1}?
influenced_by	{e1} looked up to {e2}.	Who inspired {e1}?
known_for	{e1} gained recognition for {e2}.	What is {e1} famous for?
leader_of	{e1} was the leader of {e2}.	Which group was {e1} in charge of?
lived_in	{e1} resided in {e2}.	Where has {e1} lived?
major	{e1} majored in {e2}.	What did {e1} specialize in?
mentored_by	{e1} received guidance from {e2}.	Who mentored {e1}?
mentoring	{e2} is a student of {e1}.	Who does {e1} mentor?
nationality	{e1} is a citizen of {e2}.	What is {e1}'s nationality?
neighbor	{e1} lives next to {e2}.	Who resides beside {e1}?
occupation	{e1} is employed as {e2}.	What does {e1} do for a living?
parent	{e1}'s parent is {e2}.	Who is the parent of {e1}?
pet	{e1} owns a pet called {e2}.	What is the name of {e1}'s pet?
philanthropy	{e1} donated to {e2}.	Which causes did {e1} support?
phone	{e1} can be reached at {e2}.	What is {e1}'s phone number?
rival	{e1} had a rivalry with {e2}.	Who did {e1} compete with?
roommate	{e1} shared a room with {e2}.	Who lived with {e1}?
service	{e1} was a member of {e2}.	Which organization did {e1} serve in?
sibling	{e1} and {e2} are siblings.	Who are {e1}'s siblings?
spouse	{e1} is married to {e2}.	Who is {e1}'s spouse?
university	{e1} went to {e2}.	Which university did {e1} attend?
worked_at	{e1} held a position at {e2}.	Where did {e1} work?
wrote	{e1} authored the book {e2}.	Which book did {e1} write?

We finally synthesize 39 relations including eight symmetric relations (*e.g., spouse, sibling*), and eight pairs of inverse relations (*e.g., child and parent*) to mimic real-world complexity. For each relation, we adopt an LLM to construct and validate three natural language templates. We adopt heuristic rules to constrain the tailed entity of the relation (*e.g.,* birthday should be a date). Table 4 shows the relations and templates.

For QA pairs construction, after sampling the relation paths/combinations, we adopt GPT-4o to generate question templates. Table 5 shows an example question template.

After we split the relation combinations based on the generalization levels 2, we can synthesize human biographies based on the knowledge graph (KG). We firstly translate the dict-formed biographies of each 'character' entity from KG into natural language paragraphs with the relation templates. Table 6 shows an example of the biography dict and corresponding natural language paragraph.

Then, we construct QA pairs via random-walk over the KG: given a relation path, we sample an entity as the starting point and traverse the path to the answer. The final answer is then appended

Table 5: An example of Question and CoT answer templates based on a relation path (4-hop). {e1} denotes the topic entity in the question.

Component	Content / Examples
Relation Path	`sibling` → `boss_of` → `mentored_by` → `best_friend`
Question Templates	1. Who is the best friend of the person mentoring the employee of the sibling of {e1}? 2. Can you tell me the best friend of the person who mentored the employee of {e1}'s sibling? 3. Who is the best friend of the person mentoring the employee of the sibling that {e1} has?
CoT Answer Templates	1. {e2} is {e1}'s brother/sister. {e3} works under {e2}. {e3} was trained by {e4}. {e4}'s best friend is {e5}. So, the answer is: {e5} 2. {e1} and {e2} are siblings. {e2} is the boss of {e3}. {e3} was trained by {e4}. {e4}'s best friend is {e5}. So, the answer is: {e5} 3. {e2} is {e1}'s brother/sister. {e2} is the boss of {e3}. {e3} received guidance from {e4}. {e5} is {e4}'s closest friend. So, the answer is: {e5}

after "So, the answer is:". Crucially, we incorporate Chain-of-Thought (Wei et al., 2022) into the target answer for three strategic reasons: 1) to facilitate decomposition of complex queries; 2) to facilitate retrieval and application of factual knowledge, suggested by Allen-Zhu & Li (2023a); 3) to precisely evaluate whether the model is retrieving facts from the parameter or context. As construction is scalable, we provide sufficient samples to ensure excellent IID SFT performance ($> 90\%$).

C Details of SFT Training Baselines

As discussed in Section 3.1, we are able to synthesize as much data as needed. However, in real-world scenarios, while LLMs can easily handle either Parametric or Contextual Reasoning, probably through post-training with sufficient data, they struggles in Complementary Reasoning task, where it is hard to collect ample data. We show in Table 1 that Complementary Reasoning is data-hungry and most difficult compared to parametric and contextual reasoning. Here we show empirical results training with SFT and evaluating on the corresponding testing set in Table 7.

It further shows that Contextual Reasoning is the easiest. While LLMs learns to adopt new knowledge during SFT training, they manage to handle most of the unseen knowledge in the context window. We hypothesize that this ability is essential for further generalization to new reasoning patterns. Moreover, SFT training involving parametric knowledge would be hard to generalize on Zero-shot settings. Both parametric and complementary reasoning, while performing well on I.I.D. setting, drop significantly on Composition and almost fail on Zero-shot setting.

This further highlights our motivations: we try to figure out the recipe of training strategies and mix of training data required for generalization to complementary reasoning on all difficulty levels.

D Model Scaling

We study whether our findings that "RL enables generalization in Complementary Reasoning from sufficient atomic skills" persist as models scale. We compare Qwen-2.5-0.5B, Qwen-2.5-1.5B and Qwen-2.5-3B due to limited compute. Figure 8 illustrates the performance trends across model sizes. We observe consistent behaviors that strongly support the effectiveness of our proposed recipe.

As shown by the blue lines, $\text{SFT}_{\text{MEM+CTX}}$ leads to substantial performance jumps after RL_{COMP}. This improvement is particularly pronounced in the Zero-shot setting, indicating that the model with sufficient atomic skills successfully generalizes the reasoning capabilities acquired during RL. In contrast, SFT_{COMP} (orange dashed line), while starting with decent performance, exhibits limited

Table 6: An example of human biography in the form of Python Dict and Natural Language Document.

Python Dictionary	**Natural Language Document**

"Allison Hill": {
"name": "Allison Hill",
"birth_date": "1942-04-29",
"occupation": "Civil engineer, consulting",
"email": "garzaanthony@example.org",
"phone": "538.990.8386",
"new": true,
"died_on": "2024-11-01",
"child": "Donald Marsh",
"pet": "Whiskers",
"wrote": "Baby administration",
"influenced_by": "Matthew Cooper",
"mentoring": "Daniel Watkins",
"hobby": "painting",
"classmate": "Adam Villanueva",
"first_language": "Finnish",
"roommate": "Shannon Krause",
"university": "University of Chicago",
"service": "Habitat for Humanity",
"known_for": "painting",
"died_in": "Brownbury",
"boss": "Lindsey Johnson",
"favorite_food": "tacos"
}
	Allison Hill has a pet named Whiskers. Allison Hill spoke Finnish as their first language. A favorite activity of Allison Hill is painting. Lindsey Johnson is the boss of Allison Hill. Allison Hill was born on 1942-04-29. Allison Hill died on 2024-11-01. Allison Hill shared a room with Shannon Krause. Allison Hill penned Baby administration. Allison Hill was inspired by Matthew Cooper. The contact email for Allison Hill is garzaanthony@example.org. Allison Hill was famous for painting. Allison Hill was a member of Habitat for Humanity. Allison Hill mentors Daniel Watkins. Allison Hill's place of death was Brownbury. Allison Hill's phone number is 538.990.8386. Allison Hill works as a Civil engineer, consulting. Allison Hill is the parent of Donald Marsh. Allison Hill was a classmate of Adam Villanueva. Allison Hill loved eating tacos. Allison Hill went to University of Chicago.

Table 7: Empirical performance. For each reasoning type, we training on the training set and test on the corresponding test set. Note that this is different from other experiments that focus on evaluation over complementary reasoning test set.

Training Data	**I.I.D**	**Comp**	**Zero-shot**
Parametric Reasoning	93.96	82.30	3.71
Contextual Reasoning	98.53	95.53	69.53
Complementary Reasoning	90.26	76.61	7.76

growth after RL training (orange solid line). The gap between the pre-RL and post-RL performance is marginal, suggesting that SFT with composite data may limit the model's potential for further generalization.

Crucially, *these trends hold true across all model sizes*. When the model size increases to 3B, the superiority of $\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$ over $\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$ remains significant (*e.g.,* a gap of approximately 13% in Zero-shot accuracy for the 3B model). This confirms that our conclusions are robust to model scaling and implies that $\text{SFT}_{\text{MEM+CTX}}$ is a more effective foundation for scaling up RL-based reasoning.

Table 8: Empirical results on Complementary Reasoning test set.

Setting	I.I.D	Comp	Zero-shot
$\text{SFT}_{\text{MEM+CTX}}$	35.18	28.20	24.07
$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{MEM+CTX}}$	28.43	28.34	27.89
$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$	74.00	62.47	49.56
$\text{SFT}_{\text{MEM+CTX+COMP}}$	**80.14**	**62.90**	43.25
$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$	73.11	60.85	**50.87**

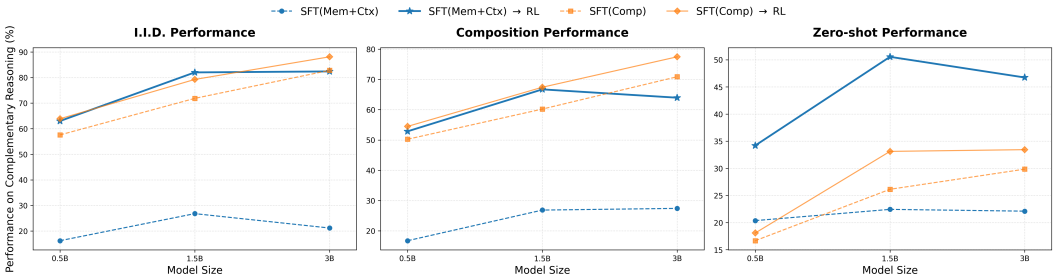

Figure 8: Model scaling analysis of Qwen model family across 0.5B, 1.5B, and 3B parameters.

E Details on Sufficiency of the Finding

In Section 4.1, we focus on whether our proposed recipe "$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$" consistently yields superior generalization, compared to direct training on the target task "$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$" which is the common practice in real-world complex tasks. Here, we compare "$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$" with other possible baselines: 1) "$\text{SFT}_{\text{MEM+CTX}}$": SFT with the entire MEM and CTX data, which is the same as in Table 2; 2) "$\mathbf{SFT_{MEM+CTX} \rightarrow RL_{MEM+CTX}}$": SFT with 80% of MEM and CTX data and then RL with the rest 20% of MEM and CTX data. The portion is based on our empirical results of splitting COMP data; 3) "$\mathbf{SFT_{MEM+CTX} \rightarrow RL_{MEM+CTX} \rightarrow RL_{COMP}}$": furthur RL with COMP based on "$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{MEM+CTX}}$"; 4) "$\mathbf{SFT_{MEM+CTX+COMP}}$": SFT with the mix of MEM, CTX and COMP data. Specifically, for RL_{COMP}, we adopt the 12.8k random samples used in Section 4.2. We show the performance on COMP test set in Table 8.

It shows that $\mathbf{SFT_{MEM+CTX} \rightarrow RL_{MEM+CTX} \rightarrow RL_{COMP}}$ has very similar performance compared with our proposed $\mathbf{SFT_{MEM+CTX} \rightarrow RL_{COMP}}$, especially for Zero-shot setting. However, this does not contradict our conclusion, as the capture of sufficient atomic skills and RL training are essential for Zero-shot generalization and sufficient atomic skills are still the prerequisites. In addition, comparing $\text{SFT}_{\text{MEM+CTX}}$ with $\mathbf{SFT_{MEM+CTX} \rightarrow RL_{MEM+CTX}}$, we find that RL with atomic skills may not be beneficial to composite tasks. Note that for MEM and COMP data, every test sample in COMP would be either Composition or Zero-shot level (these two baselines have never seen any COMP data during training. Moreover, Table 8 further enhances the sufficiency of our findings that our recipe $\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$ is the key to generalization.

It is interesting that $\mathbf{SFT_{MEM+CTX+COMP}}$ also manages to generalize to Zero-shot scenarios to some extent and achieves comparable results with our $\mathbf{SFT_{MEM+CTX} \rightarrow RL_{COMP}}$. This shows that SFT can also generalize with careful mix of training data. However, this does not challenge our SFT Memorization Paradox that SFT directly with composite data fails to generalize. Furthermore, the setting itself is not as realistic-it is not trivial to obtain ample composite (COMP) data encompassing both MEM and CTX skills. Also, it is difficult to mix COMP data to SFT training with all atomic skills. However, our $\mathbf{SFT_{MEM+CTX} \rightarrow RL_{COMP}}$ assumes a common post-training practice with an initial SFT and a following RL stage. While the base model grasp the atomic skills (*i.e.,* today's available LLMs), we focus on the training strategies and data mix when we have some composite data. And we demonstrate the condition of RL generalization.

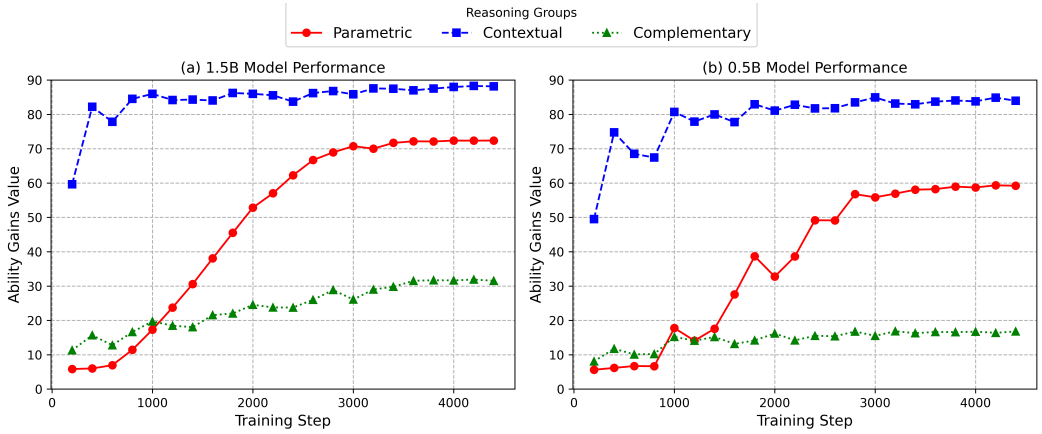

Figure 9: Training dynamics of SFT with parametric and contextual reasoning data over training steps. Ability gains are calculated over MEM (Red Line), CTX (Blue Line) and COMP (Green Line), respectively. As the training progresses, complementary reasoning ability emerges to some extent.

F DETAILS ON OTHER FACTORS THAT AFFECT GENERALIZATION

There are many factors affecting models'performance or generalizability. We showcase the sample efficiency and pass@k performance as some features to see why the model with atomic skills are essentials for RL generalization in composite tasks. Here, we investigate from the perspective of training dynamics of $\text{SFT}_{\text{MEM+CTX}}$, the impact of training loss of the base model for RL, the embedding distributions of the SFT and RL model and the uncertainty of the model.

F.1 THE TRAINING DYNAMICS OF $\text{SFT}_{\text{MEM+CTX}}$

Settings To check when and how the ability of complementary reasoning ability (COMP) emerges through the training of parametric and contextual reasoning *i.e.,* $\text{SFT}_{\text{MEM+CTX}}$, we showcase the performance of MEM, CTX and COMP during the SFT training over MEM and CTX. We study Qwen-1.5B-Base and Qwen-0.5B-Base.

The ability of complementary reasoning emerges to some extent with the progress of both parametric and contextual reasoning. Figure 9 shows the training dynamic over training steps. First, it shows that as the SFT training with both MEM and CTX data progresses, the ability of COMP somehow emerges with MEM and CTX to some extent. Second, the conclusion is consistent over different model sizes, with larger models generalizes better to COMP with MEM and CTX data. Third, it further demonstrates that contextual reasoning is the easiest (learned from relatively early stage), while the parametric and complementary reasoning is relatively difficult.

F.2 THE IMPACT OF TRAINING LOSS FOR RL GENERALIZATION

We investigate a critical question regarding the interplay between Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL): *Does the degree of SFT convergence dictate the model's potential for RL-based generalization?* Specifically, we aim to determine if a "grokking-like" phenomenon exists–where continuous optimization of SFT loss, even after apparent metric saturation, further unlocks the model's reasoning capabilities during the RL stage. Also, we study at which checkpoint (*i.e.,* training loss) may the model emerge the ability of generalization. We take four intermediate checkpoints of $\text{SFT}_{\text{MEM+CTX}}$ to further conduct RL_{COMP} and evaluate the performance on three levels of generalization.

Figure 10 illustrates the performance trajectories across SFT checkpoints (1k to 4k steps), revealing three distinct phases of capability emergence:

• **Insufficient Representation.** At the early stage (*i.e.,* 1k steps, Training Loss ≈ 0.44), the model has not yet internalized the necessary reasoning patterns. Consequently, it fails to generalize effec-

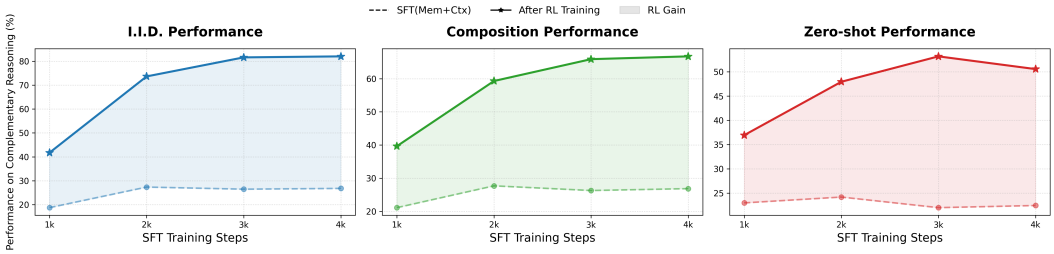

Figure 10: The impact of training steps (*i.e.,* training loss) of the SFT model on RL generalization.

tively during the RL stage, resulting in suboptimal performance across all metrics (e.g., Zero-shot accuracy is significantly lower compared to later stages).

• **Emergence of Capabilities.** As the SFT loss decreases to approximately 0.16 (*i.e.,* 2k steps), we observe a sharp "phase transition" in downstream RL performance. While the base SFT model's direct performance (dashed lines) shows only moderate improvements, its *latent potential* for RL adaptation increases dramatically. This suggests that the critical structures required for reasoning generalization are established during this interval.

• **Saturation and Robustness.** Performance peaks around 3k steps (Loss ≈ 0.05). Interestingly, further extending training to 4k steps—where the model nearly memorizes the training data (Loss ≈ 0.0004)—does not yield further significant gains, nor does it lead to performance degradation. This indicates that while "grokking" (delayed generalization) effectively occurs between 1k and 3k steps, the benefit saturates once the loss drops below a certain threshold (< 0.05). The model becomes robust, maintaining its high plasticity for RL even when deeply fitted to the SFT distribution.

In conclusion, **minimizing SFT loss is crucial up to a point**. The generalization capability for RL does not scale infinitely with lower loss but requires a sufficient "incubation" period (up to 3k steps in our setting) to fully emerge.

F.3 Analysis of Latent Space Dynamics via PCA

Setting: Layer-wise PCA Projection To investigate the internal mechanism behind the superior performance of $\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$ compared to $\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$, we conduct a Principal Component Analysis (PCA) on the hidden states of the models. Our goal is to visualize how different training strategies affect the representation of Parametric, Contextual, and Complementary Reasoning.

Formally, for a given model pair (Anchor Model M_{anc} and Target Model M_{tgt}) and a specific layer l, we extract the hidden states corresponding to the last token of the input queries. Let $\mathbf{H}_l^{anc} \in \mathbb{R}^{N \times D}$ and $\mathbf{H}_l^{tgt} \in \mathbb{R}^{N \times D}$ denote the hidden state matrices for N queries at layer l, where D is the hidden dimension. To capture the relative shift induced by training, we fit the PCA transformation on M_{anc}'s states $\mathbf{H}_l^{anc}$, which defines a 2D coordinate system based on the principal variations of the reference model. We then project M_{tgt}'s states $\mathbf{H}_l^{tgt}$ into this fixed coordinate system. The shift vector for layer l is calculated as the difference between the centroids of the projected target states and the anchor states. This process is repeated for all layers. Figure 11 shows the layer-wise shifts by scatter points. The large markers represent the global centroid shift (z^*) for each Reasoning type.

Disentanglement of atomic skills in $\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$ As shown in the top-left panel of Figure 11, $\text{SFT}_{\text{MEM+CTX}}$ exhibits a significant "disentanglement" of atomic reasoning types. The centroid for MEM data (Blue Circle) and CTX data (Orange Triangle) move in distinct directions and magnitudes within the principal component space. This suggests that the $\text{SFT}_{\text{MEM+CTX}}$ stage effectively separates the internal representations required for parametric recall versus contextual reasoning. Furthermore, in the subsequent RL stage (top-right panel), this separation is maintained and refined. Notably, the COMP data (Green Square) aligns closely with the established distributions of MEM and CTX data. This indicates that the model can effectively generalize the logic learned from MEM and CTX data to the composite COMP tasks.

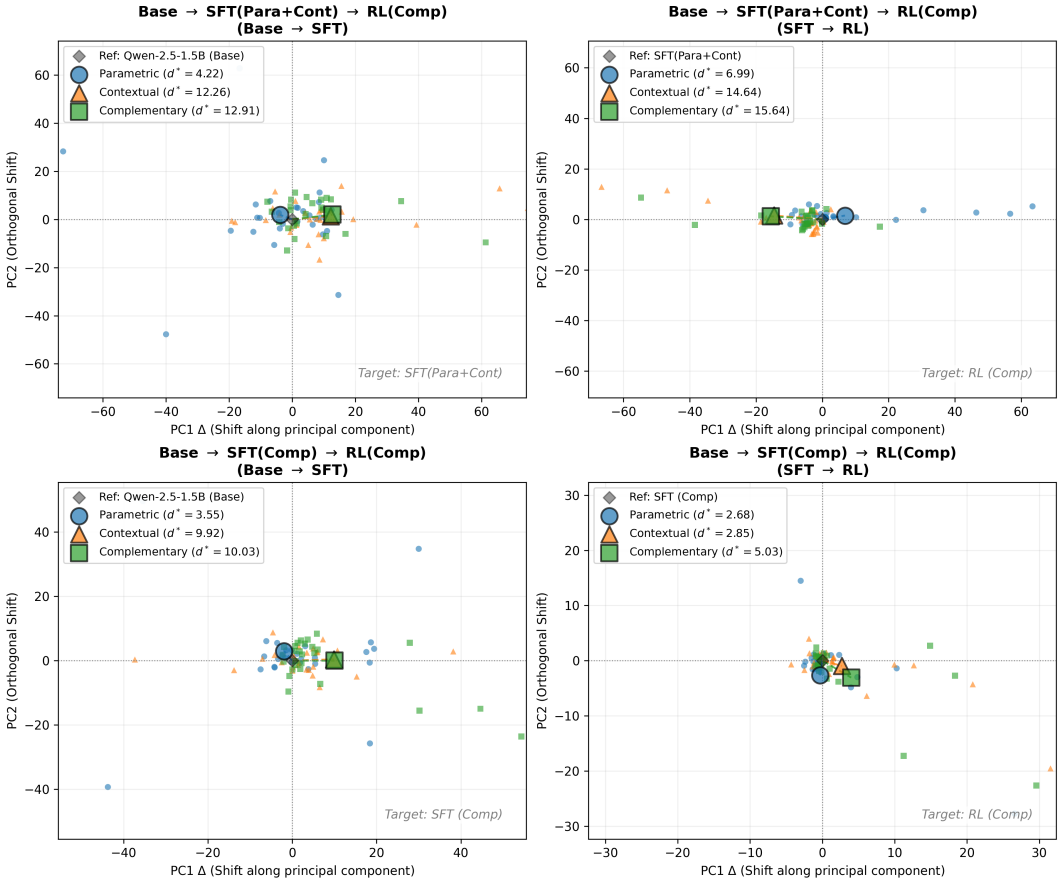

Figure 11: PCA Analysis of $\text{SFT}_{\text{MEM+CTX}}$ $\rightarrow\text{RL}_{\text{COMP}}$ and SFT_{COMP} $\rightarrow\text{RL}_{\text{COMP}}$. The scatter points represent the layer-wise shifts. The large markers represent the global centroid shift (z^*) for each reasoning type.

Entanglement of skills in SFT_{COMP} $\rightarrow\text{RL}_{\text{COMP}}$. In contrast, the bottom row reveals a phenomenon of "representation entanglement". Please note that the scale of the axis for the bottom row is lower than that for the top row. For the model trained only on COMP data, the embeddings for MEM, CTX, and COMP queries remain tightly clustered together, after both the SFT stage (bottom-left) and the RL stage (bottom-right). The centroids for all three data types are located close to the origin with overlapping distributions. This lack of separation implies that SFT_{COMP} fails to distinguish between the underlying mechanisms of parametric and contextual reasoning, instead learning a coupled representation.

We hypothesize that the superior generalization capability of $\text{SFT}_{\text{MEM+CTX}}$ $\rightarrow$ RL(Comp) stems from this structural disentanglement. By explicitly separating the latent representations of parametric and contextual capabilities during the SFT stage, the model establishes a robust basis that facilitates better adaptation during the RL phase. Conversely, the coupled representations in SFT_{COMP} limit the model's ability to distinctly apply these capabilities, leading to suboptimal performance.

F.4 MODEL UNCERTAINTY

We investigate the evolution of model uncertainty (quantified by the average prediction entropy $\times 100$) to understand the underlying dynamics of RL generalization. Table 9 presents the entropy metrics across I.I.D., Compositional, and Zero-shot subsets.

Uncertainty is correlated with SFT convergence, not necessarily RL potential. We first address whether high uncertainty is a prerequisite for effective RL exploration. Comparing the $\text{SFT}_{\text{MEM+CTX}}$ checkpoints, the uncertainty drops significantly from 3k steps (12.92) to 4k steps (7.13) as the loss

Table 9: Analysis of model uncertainty (measured by average entropy $\times 100$) across different training settings. **Overall** denotes average uncertainty on the test set, while the breakdown columns show uncertainty on I.I.D., Compositional (Comp.), and Zero-shot settings. Lower values indicate higher model confidence.

Setting	Overall	I.I.D.	Comp.	Zero-shot
Effect of Training Steps (SFT on MEM +CTX*)*				
$\text{SFT}_{\text{MEM+CTX}}$ 3k step	12.92	13.41	12.05	13.12
$\text{SFT}_{\text{MEM+CTX}}$ 4k step	7.13	6.33	7.45	8.46
10% COMP *data for* SFT_{COMP}, *90%* COMP *data for* RL_{COMP}				
SFT_{COMP}	8.93	6.56	9.89	12.96
$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$	8.99	8.49	8.97	10.17
$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$	2.58	1.63	2.54	4.84
30% COMP *data for* SFT_{COMP}, *70%* COMP *data for* RL_{COMP}				
SFT_{COMP}	7.18	5.24	7.98	10.46
$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$	6.11	4.85	6.36	8.65
$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$	2.90	1.96	2.84	5.17
90% COMP *data for* SFT_{COMP}, *10%* COMP *data for* RL_{COMP}				
SFT_{COMP}	4.14	2.38	4.05	8.36
$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$	4.11	3.27	3.97	6.25
$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$	4.10	3.23	3.88	6.46

minimizes. Despite this lower starting entropy, we previously observed that the 4k-step model sustains high performance after RL. This indicates that a "calibrated" and confident SFT model (lower entropy) does not hinder subsequent RL generalization.

Task Difficulty Indicator. Consistently across all settings, the uncertainty is highest in the **Zero-shot** subset (e.g., 12.96 for 10% SFT(G3)). This aligns with intuition, as the model exhibits lower confidence on unseen distributions.

$\text{SFT}_{\text{MEM+CTX}}$ Facilitates Efficient RL Adaptation. A key insight emerges when comparing the post-RL behaviors. In the 30% data setting, *30%* SFT_{COMP} (7.18) and $\text{SFT}_{\text{MEM+CTX}}$ (7.13) start at remarkably similar uncertainty levels. However, after applying identical RL training, $\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$ shows minimal entropy reduction ($7.18 \rightarrow 6.11$), suggesting the model struggles to find a more optimal, confident policy. In contrast, our recipe $\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$ achieves a drastic reduction in uncertainty ($7.13 \rightarrow \mathbf{2.90}$). This demonstrates that the model with sufficient atomic skills does not merely provide "randomness" for exploration; rather, it structures the latent space (see Figure 11) in a way that allows RL to efficiently converge to a high-confidence, correct solution.

G TERMINOLOGIES

Table 10: Glossary of Symbols and Abbreviations

Symbol/Abbreviation	**Description**
MEM	Parametric reasoning type.
CTX	Contextual reasoning type.
COMP	Complementary reasoning type.
RL	Reinforcement Learning, one of the core training strategies discussed in this work.
SFT	Supervised Fine-Tuning, one of the core training strategies discussed in this work.
LoRA	Low-Rank Adaptation, one of the training strategies discussed in this work.
$\text{SFT}_{\text{MEM+CTX}}$	The model obtained after performing SFT on the combined training set of MEM and CTX.
SFT_{COMP}	The model obtained after performing SFT on the training set of COMP.
$\text{SFT}_{10\%\text{COMP}}$	The model obtained after performing SFT on 10% random samples of COMP training set.
$\text{SFT}_{\text{MEM+CTX}} \rightarrow \text{RL}_{\text{COMP}}$	The model obtained after performing RL on the training set of COMP from $\text{SFT}_{\text{MEM+CTX}}$.
$\text{SFT}_{\text{COMP}} \rightarrow \text{RL}_{\text{COMP}}$	The model obtained after performing RL on the training set of COMP from SFT_{COMP}.
I.I.D	Independent and Identically Distributed generalization type in the test set.