

Composition Collapse: Stable Factual Knowledge Does Not Imply Compositional Reasoning

Zhe Yu^{1,*} Wenpeng Xing^{1,2,*} Yunzhao Wei^{1,7} Jie Chen³

Hongzhi Wang⁴ Xuyang Teng⁵ Meng Han^{2,6,†}

¹Binjiang Institute of Zhejiang University, Hangzhou, China

²Zhejiang University, Hangzhou, China

³Hong Kong Baptist University, Hong Kong SAR, China

⁴Harbin Institute of Technology, Harbin, China

⁵Hangzhou Dianzi University, Hangzhou, China

⁶GenTel.io, Hangzhou, China ⁷East China Normal University, Shanghai, China

zyu@zju-if.com wpxing@zju.edu.cn mhan@zju.edu.cn *Equal contribution; †corresponding author.

Abstract

A model can know when two events occurred yet still answer which came first incorrectly. Multi-hop QA accuracy conflates such composition errors with missing facts. We introduce a double-gate protocol for temporal factual chains and evaluate a case for composition only if the model (i) answers every constituent fact correctly across paraphrased probes and (ii) answers every aligned sub-question correctly in isolation. If the original multi-hop answer remains incorrect, the case is a *residual composition failure*; we call the systematic occurrence of such failures *composition collapse*. On D4V2, a benchmark of 390 temporal chains, two post-trained models differ by only 0.3 percentage points in atomic stability yet by 47.4 points in depth-2 (two-fact) residual failure. In a controlled same-base comparison, outcome-verified reinforcement learning reduces depth-4 residual failure by 54 points relative to reasoning-trace SFT under matched data and LoRA budget. On 80 examples that pass both gates for every GRPO variant, failure decreases monotonically, showing that model-specific gate subsets are unlikely to drive the trend. Finally, chain-of-thought decoding recovers 70–75% of gate-passing failures, consistent with an important role for inference-time computation. These results motivate reporting atomic stability and gate-controlled composition separately when evaluating multi-hop reasoning.

1 Introduction

When a language model fails a multi-hop question such as *which came first, the telephone or powered flight?*, two explanations are possible. The model may not reliably know the dates, or it may know both dates yet fail to compare them. These failures call for different remedies, but end-task accuracy counts them identically.

Multi-hop QA benchmarks such as HotpotQA, 2WikiMultiHopQA, and MuSiQue evaluate questions that require combining multiple pieces of evidence (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). Aggregate accuracy cannot reveal why an answer failed. Press et al. (2023) formalise the *compositionality gap* as the fraction of multi-hop cases answered incorrectly even when their subquestions are answered correctly, and relate this gap to confidence in the subanswers. Our concern is complementary: one correct response does not establish that the model can reliably access the same fact under surface variation. If that response changes under paraphrase, a later multi-hop error may reflect unstable knowledge rather than failed composition.

We call correctness across paraphrased probes of the same constituent fact *atomic stability*. Figure 1 previews the central result: among post-trained models with similar atomic stability, residual composition failure differs by 47.4 percentage points at depth 2 and rises sharply with chain depth. We call this *composition collapse*: systematic failure to assemble stably-known facts into a correct answer.

We therefore measure composition only after two checks. Gate 1 verifies each constituent fact across multiple paraphrases, and Gate 2 verifies each aligned sub-question in isolation. A main-question error that survives both gates is a *residual composition failure*: the model retrieves the parts stably in isolation but answers the composed question incorrectly. This *double-gate protocol* replaces an aggregate gap with an error rate conditioned on stable atomic access. We report three complementary axes: Δ_{atom} (atomic stability), Δ_{comp} (residual failure at matched atomic stability), and Δ_{depth} (a coarse summary of where residual failure crosses 50%).

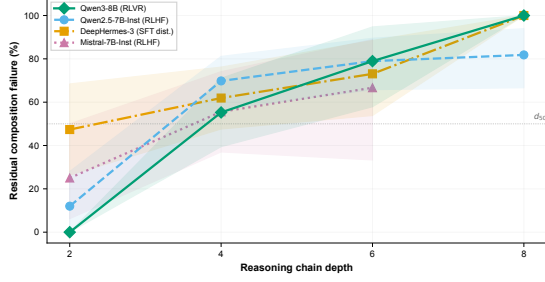


Figure 1: **Composition collapse in temporal factual chains.** Residual composition failure on D4v2 after both gates pass, with bootstrap 95% CIs. Closely matched atomic stability does not imply matched composition, and failure generally increases with depth. Deeper estimates have wider intervals because fewer cases pass every atomic check; all depth-8 denominators are at most 33.

We instantiate the protocol on D4v2, a 390-question temporal benchmark spanning depths 2–11 across four task families. D4v2 focuses on dates and event order because these support short, verifiable atomic answers and deterministic decompositions. Our main claims concern this temporal setting. In-context synthetic tasks and a mixed-domain pilot provide preliminary evidence beyond it, but do not establish domain generality.

We evaluate seven open-weight checkpoints at 7–13B scale spanning base models, instruction and preference post-training, reasoning-trace distillation, and multi-stage RL with distillation. We also run controlled same-base LoRA interventions comparing SFT and GRPO (Shao et al., 2024).

Three findings emerge. First, untuned base checkpoints rarely pass the atomic gate, so their multi-hop errors cannot be cleanly attributed to composition. Second, post-trained models with closely matched atomic stability differ by 47.4 points at depth 2, while same-base GRPO reduces depth-4 residual failure by 54 points relative to reasoning-trace SFT. Third, training gains are largest at covered depths and transfer only weakly to held-out deeper chains. Chain-of-thought decoding recovers 70–75% of gate-passing failures, consistent with inference-time computation being an important contributor. Same-case intersections, expanded depth-6/8 candidate pools, and human validation support the main measurement assumptions.

Our contributions are: (i) an instance-level double gate that separates stable atomic access from composition in temporal factual chains; (ii) D4v2

and reporting axes for comparing post-training recipes; and (iii) controlled GRPO/SFT interventions showing that residual composition can improve while atomic stability remains similar, with limited transfer to held-out depths.

2 Related Work

Multi-hop QA and composition measurement. HotpotQA provides supporting facts for multi-document questions (Yang et al., 2018); 2WikiMultiHopQA adds explicit reasoning paths (Ho et al., 2020); and MuSiQue constructs connected multi-hop questions from composable single-hop questions (Trivedi et al., 2022). These datasets improve control over evidence and reasoning structure, but end-task accuracy does not by itself verify that a particular model stably knows every required fact. The *compositionality gap*—the population-level rate of multi-hop failures among cases whose sub-questions are answered correctly—was formalised by Press et al. (2023). Dziri et al. (2023) instead study computation-graph depth; our chain depth counts required atomic facts. Press et al. also report that lower perplexity on correct subanswers is associated with greater composition success. Yang et al. (2025b) introduced latent multi-hop probes that reduce surface shortcuts and show markedly different success for country and year intermediate answers.

Our protocol adds two per-case checks. First, every required fact must be answered correctly across multiple paraphrased probes, rather than once. Second, every aligned sub-question must be answered correctly in isolation before the composed question is scored. Empirically, the stability gate removes 15–44% of cases admitted by the sub-question-only baseline.

Composition collapse is therefore distinct from the *compositionality gap* of Press et al. (2023). Their metric conditions on correct subquestion answers and aggregates over the evaluated population; ours measures residual error only on instances that pass two model-specific checks: stable correctness across paraphrased atomic probes and correctness on aligned sub-questions in isolation. Perplexity and paraphrase stability are also distinct: the former measures sequence-level confidence, whereas the latter tests reproducible access to the same fact.

Post-training and reasoning. Relevant approaches include RLHF (Ouyang et al., 2022), chain-of-thought prompting (Wei et al., 2022),

reasoning-trace distillation (DeepSeek-AI, 2025), and outcome-verified RL (Shao et al., 2024). Our reporting axes distinguish changes in atomic stability from changes in residual composition. In the controlled comparison, reasoning-trace SFT and GRPO use the same base model, training questions, and LoRA budget, allowing the training objective to be compared without the base-model differences present in off-the-shelf checkpoints.

Inference-time computation and composition. Biran et al. (2024) study where multi-hop information is integrated across layers, while Allen-Zhu and Li (2023) identify systematic limits in extracting facts represented in training data. TriLens detects hallucinations from per-layer logit-lens entropy trajectories over attention, feed-forward, and residual-stream outputs (Yang et al., 2026). It targets answer-level detection; we instead condition on stable atomic access and test composition. Recent work on large reasoning models diagnoses failures within generated reasoning trajectories: Jiang et al. (2026) find that exploring alternative solutions can accumulate reasoning-path errors and propose pruning them, whereas Dong et al. (2026) separate intra-step, inter-step, and instance-level failures and trigger targeted corrections during decoding. These analyses concern overt reasoning trajectories; our CoT diagnostic (§6) complements these findings: additional inference-time computation recovers 70–75% of gate-passing failures. Together with prompt-end patching probes (Appendix J), this result identifies generation-time computation as a plausible contributor while leaving the sufficiency of the underlying representations unresolved.

Agentic decomposition and memory. Adjacent work distributes complex reasoning across external system components. Unified-MAS synthesizes and optimizes domain-specific nodes separately from multi-agent topology design (Lin et al., 2026a), whereas Skill-MAS retains orchestration experience as an evolvable, transferable meta-skill (Lin et al., 2026b). T-Mem addresses a different bottleneck by using write-time triggers to support associative as well as surface-similar recall in conversational memory (Guo et al., 2026). These systems do not evaluate the closed-book factual-chain setting studied here, but they motivate extending gate-controlled evaluation to distinguish reliable component access from successful composition across components.

Cross-lingual reasoning. Multilingual work exposes an additional distinction between factual

access and the reasoning path used to express it. AdaMCoT adaptively selects intermediary thinking languages for multilingual factual reasoning (Weihua et al., 2026), while CCL-XCoT combines cross-lingual contrastive alignment with high-resource-language CoT to improve factual knowledge transfer and reduce hallucination in low-resource languages (Weihua et al., 2025). MMAC evaluates cultural grounding with input-aligned text, image, and speech questions, measuring cross-lingual and cross-modal consistency alongside knowledge generalization and grounding validity (Zheng et al., 2026). These settings mix language or modality transfer with reasoning and do not impose our closed-book double gate, but they motivate extending gate-controlled evaluation across both axes.

3 Method

Figure 2 summarises the pipeline. A composition instance consists of a main question q whose answer requires k atomic facts $a_1, \dots, a_k$; k is the chain depth. For each atom, the benchmark provides m_i paraphrased probes and an aligned sub-question s_i that matches its role in the chain. An instance *passes the atomic gate* iff the model answers every atom correctly across all paraphrases under our answer-equivalence adjudication. It *passes the sub-question gate* iff the model answers every s_i correctly in isolation.

We then compute the *residual composition failure rate*: the fraction of cases that pass both gates but still receive an incorrect main answer. This rate is an operational measure, not a direct observation of a latent reasoning state. Format and scoring artifacts can inflate it, while the validated adjudicator’s remaining false positives for correctness can slightly reduce it (Appendix O).

Given a recipe R and a comparator R' , three reporting axes ask whether post-training changes atomic stability, residual composition failure, or the depth profile:

$$\begin{aligned}\Delta_{\text{atom}}(R) &= s(M_R) - s(M_{R'}), \\ \Delta_{\text{comp}}(R) &= f(M_R \mid \text{gate}) - f(M_{R'} \mid \text{gate}), \\ \Delta_{\text{depth}}(R) &= d_{50}(M_R) - d_{50}(M_{R'}),\end{aligned}$$

where s is atomic stability, f is residual composition failure, and d_{50} is the depth at which f first exceeds 50%. We use d_{50} only as a coarse diagnostic because it is interpolated from four bins and deeper gate-passing samples are smaller. Matching aggregate atomic stability within 2 pp reduces

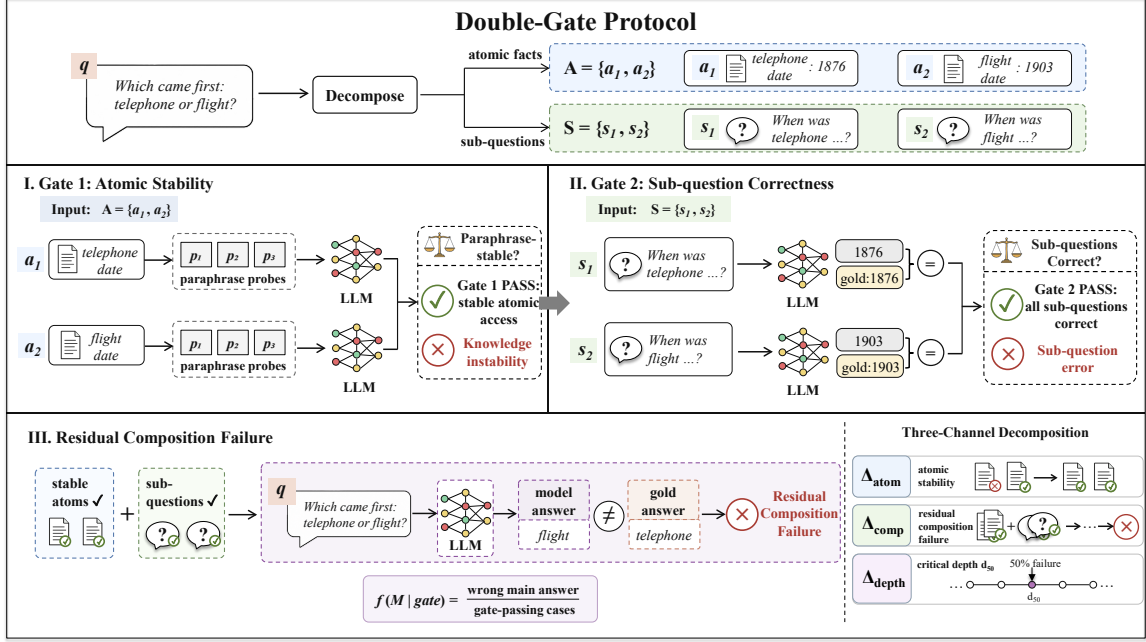


Figure 2: **Double-gate protocol for measuring residual composition failure.** A multi-hop question is decomposed into atomic facts and aligned sub-questions. **I.** Atomic Stability verifies stable and correct access to each atom across paraphrases. **II.** Sub-question Correctness verifies that each sub-question is answered correctly in isolation. **III.** Residual Composition Failure measures whether the original multi-hop question is still answered incorrectly after both gates are passed. This separates composition errors from unstable knowledge access and supports reporting along Δ_{atom} , Δ_{comp} , and Δ_{depth} .

knowledge-level confounding but does not guarantee that the gates admit identical cases. We therefore also report same-case intersections (Appendix F).

Benchmark. D4v2 contains 390 main questions and 2 490 atomic sub-questions across four task families: a depth-2 pair-ordering control and three families evaluated at depths 4, 6, and 8 (TEMPORAL_RANK, TEMPORAL_SUCCESSOR, TEMPORAL_INTERVAL_DECOY). Three distractor-augmented interval variants add three required facts, yielding depths $d=7, 9, 11$ ($n=30$ each). Task families differ in required chain structure—ranking a set, stepping forward in time, and disambiguating overlapping intervals—so same-depth cases are not interchangeable in difficulty. Headline cross-recipe curves use the shared exact-depth cells $d \in \{2, 4, 6, 8\}$; the task-family analysis additionally groups each distractor-augmented interval variant by its core depth and states this explicitly. See Appendix B for the full breakdown. Three additional synthetic task families (KINSHIP, NUMERICAL, SPATIAL) built from fictional entities form a separate evaluation suite. As discussed in §4, the closed-book analysis excludes them because no cases pass the atomic-knowledge gate.

Recipe coverage and evaluation. We evaluate seven open-weight checkpoints at the 7–13B scale: three base models and four post-trained models whose disclosed pipelines include instruction tuning, preference optimization, reasoning-trace distillation, and multi-stage RL with distillation (checkpoints in Appendix C). Gate 2 follows the aligned-subquestion check of Press et al. (2023); Gate 1 adds the stricter requirement of stable correctness across paraphrased atomic probes. Full adjudication tiers and vLLM vs. HuggingFace cross-validation are reported in Appendix L and I. All evaluations use greedy decoding ($T=0$) and XML-structured outputs.

4 Experiments and Results

We first compare residual failure across post-trained checkpoints. We then test whether the observed differences come from model-specific gate-passing subsets or small high-depth samples. Base models rarely pass the atomic gate, closely matched post-trained models diverge at shallow depth, and task structure changes residual failure by up to 54 percentage points at fixed depth.

Figure 1 shows residual composition failure as a function of chain depth, with bootstrap 95% CIs

Table 1: Residual composition failure rate (%) by depth on D4v2, restricted to gate-passing post-trained models. “Atom stab.” is the per-probe atomic-stability rate; the atomic gate imposes a stricter per-fact criterion, so the effective fact-level pass rate is lower. Parenthesised values are n , the number of double-gate-passing cases. We omit $n < 5$.

Model	Stab. (%)	$d=2$	$d=4$	$d=6$	$d=8$
Qwen3-8B (Yang et al., 2025a) (RL+distill)	87.7	0.0 (17)	55.3 (38)	78.9 (19)	100.0 (11) [†]
Qwen2.5-7B-Inst (Yang et al., 2024) (SFT+DPO+GRPO)	89.9	12.0 (25)	69.8 (53)	78.9 (38)	81.8 (33)
DeepHermes-3 (Teknium et al., 2025) (R1 distill)	88.0	47.4 (19)	61.9 (42)	73.1 (26)	100.0 (15) [†]
Mistral-7B-Inst (Jiang et al., 2023) (instruction FT)	78.8	25.0 (16)	55.6 (27)	66.7 (9) [†]	—

[†] $n \leq 15$ gate-passing cases; point estimates in these cells should be interpreted with caution—bootstrap CIs exceed ± 25 pp.

over double-gate passing cases.

The parenthetical tags summarize only training stages disclosed by the model providers: Qwen3 combines multi-stage RL with strong-to-weak distillation, Qwen2.5-Instruct uses SFT, DPO, and GRPO, DeepHermes is distilled from R1, and Mistral-Instruct is instruction-fine-tuned. Thus, cross-model contrasts are correlational rather than controlled recipe ablations.

Base models are filtered by knowledge, not composition. All three base checkpoints we test (Llama-3-8B (Llama Team, AI @ Meta, 2024), Mistral-7B-v0.1 (Jiang et al., 2023), Qwen2.5-7B (Yang et al., 2024)) have atomic stability below 15%, so their joint stability at depth k is near zero and the atomic gate admits fewer than five cases at any depth. Their errors are therefore filtered at the knowledge stage, leaving too few cases to estimate residual composition failure. Full per-model atomic rates are reported in Appendix C.

Synthetic task families are excluded by design. Three synthetic families (KINSHIP, NUMERICAL, SPATIAL) built from fictional entities are part of the full suite but are excluded because no model can possess internal knowledge of entities absent from its training data (atomic accuracy = 0%); the protocol correctly prevents composition measurement when atoms are absent. We therefore restrict analysis to temporal facts where models demonstrably possess the requisite knowledge.

4.1 Composition with in-context evidence

To test whether composition difficulty can persist without closed-book knowledge access, we evaluated Qwen2.5-7B-Instruct on the same synthetic families with all facts supplied in the prompt as [Evidence] blocks. Under this in-context proto-

Table 2: In-context composition on synthetic families (Qwen2.5-7B-Instruct). Facts are supplied in the prompt. “Atom pass” = all sub-questions correct. Residual failure = main-question error given atom pass.

Task family	n	Atom pass	Residual failure (%)
Kinship (matrilineal)	33	8	62.5
Kinship (patrilineal)	33	9	66.7
Spatial (eastward)	33	12	91.7
Spatial (northward)	33	6	83.3
Numerical (offset)	33	1	100.0
Aggregate	165	36	88.9

col, only 21.8% of cases survive the sub-question gate. Among those gate-passing cases, residual composition failure is 88.9% (Table 2). Supplying all facts therefore does not eliminate multi-step errors. This experiment has a different estimand from the closed-book study and does not identify the dominant cause of closed-book failure, but it provides convergent evidence that retrieval is not the only source.

Quantifying the conflation. We recompute residual failure using only the sub-question gate (the standard single-gate protocol). Across models, this rate exceeds the double-gate rate by 2.5–9.8 pp, with cell-level inflation up to 11.4 pp. The atomic gate also removes 15–44% of single-gate-passing cases whose constituent facts are not stably known. Without this gate, the measured composition error includes failures of stable knowledge access.

Matched atoms, divergent composition: the composition collapse. Three checkpoints cluster tightly on aggregate atomic stability—Qwen3 87.7%, DeepHermes 88.0%, and Qwen2.5-Instruct 89.9%—yet yield 0%, 47%, and 12% residual composition failure at depth 2 (Table 1). The 47-point gap between Qwen3’s multi-stage RL/distillation pipeline and DeepHermes’ reasoning-trace distillation is the clearest cross-model correlational signal of *composition collapse*. Because matched aggregate stability does not imply identical gate-passing cases, we next repeat the comparison on shared subsets.

Same-case intersection. Among 26 cases that pass both gates for all four main post-trained models, residual failure remains separated: Qwen3 30.8%, Mistral 34.6%, Qwen2.5-Instruct 42.3%, and DeepHermes 53.8%. The stronger controlled check uses 80 cases shared by the Qwen2.5 baseline and all four GRPO variants: failure decreases

monotonically from 60.0% (baseline) to 47.5%, 43.8%, 33.8%, and 30.0%. Thus the principal training trend persists when every compared model is scored on exactly the same gate-passing cases (Appendix F).

Expanded high-depth candidates. We doubled each depth-6 and depth-8 candidate pool from 30 to 60 cases per family. The larger pools increase gate-passing samples and narrow confidence intervals without changing the qualitative depth trend or checkpoint comparisons. For example, for Qwen2.5-Instruct on TEMPORAL_INTERVAL_DECOY, the depth-6 denominator increases from 12 to 25 and the depth-8 denominator from 8 to 17, with stable point estimates. Mistral at depth 8 remains the principal small-sample exception (Appendix F).

Limited depth transfer under current recipes.

Table 1 shows that all four gate-passing recipes sit at or above 67% residual failure by depth 6, and every reportable measurement at depth 8 is 75% or higher; Qwen3 and DeepHermes both reach 100%. The coarse interpolated 50%-crossing lies no higher than 5–6 hops for the best recipe. Across the tested checkpoints—two multi-stage Qwen models, a reasoning-trace-distilled model, and an instruction-fine-tuned model—the high-depth profile changes little. In this recipe coverage, post-training improves shallow composition more reliably than it supports depth extrapolation.

Structure matters alongside depth. At fixed depth 4, Qwen3’s residual failure ranges from 38.5% on TEMPORAL_SUCCESOR to 92.9% on TEMPORAL_RANK—a 54-point gap at identical chain length (Table 3). The within-depth spread even exceeds the within-task depth progression on TEMPORAL_INTERVAL_DECOY (41% at $d=4 \rightarrow 57\%$ at $d=6 \rightarrow 80\%$ at $d=8$). The ordering of models by residual failure reverses across task families, showing that chain length alone does not describe task difficulty.

4.2 Three reporting axes

Having established that composition failure survives the atomic gate and varies across recipes, we summarise post-training differences using the three axes of §3: atomic stability (Δ_{atom}), residual composition failure at matched atomic stability (Δ_{comp}), and critical depth (Δ_{depth}).

Table 3: Residual failure (%) by task family and core depth for Qwen3-8B (double gate). Interval-decoy cells pool the corresponding three-distractor variant ($d+3$); other rows use exact depth. At fixed core depth, task structure changes failure by up to 54 pp. Parenthesised values are n . “–” = $n < 5$.

Task family	$d=2$	$d=4$	$d=6$	$d=8$
Pair control	0.0 (17)	–	–	–
Temp. successor	–	38.5 (13)	88.9 (9)	–
Temp. int. decoy	–	41.2 (17)	57.1 (7)	80.0 (5)
Temp. rank	–	92.9 (14)	83.3 (6)	100.0 (6)

4.2.1 Δ_{atom} : post-training sharply improves atomic stability

The base-to-post-trained transition raises atomic stability by 74–88 pp, whereas differences among post-trained models are only a few percentage points. For models at this scale, composition can be measured cleanly only after post-training makes the constituent facts sufficiently stable.

4.2.2 Δ_{comp} : similar atomic stability can hide 40+pp gaps

At matched aggregate atomic stability (87–90%), the three checkpoints diverge by 47 pp at depth 2 (Table 1). Using Qwen3 (RL+distill) as R and each alternative as R' , Δ_{comp} is –47.4 pp against DeepHermes and –12.0 pp against Qwen2.5-Inst. Two recipes with nearly identical aggregate atomic stability differ by 35 pp in residual composition failure. This cross-model contrast associates different post-training pipelines with different compositional behaviour; it does not isolate a single training-stage effect. Base-model confounds for DeepHermes (Llama-3 vs. Qwen2) are partly addressed via R1-distill comparisons in Appendix C.

4.2.3 Δ_{depth} : a coarse depth diagnostic

Across post-trained models, the estimated 50%-crossing generally falls between depths 3 and 5 (Table 4). Because these values interpolate four bins with smaller deep samples, we use them to summarise the observed curves rather than as precise capability limits. Within this resolution, post-training improves shallow composition more reliably than held-out-depth extrapolation.

5 Causal intervention: training reduces seen-depth collapse with limited transfer

The cross-recipe analysis in §4 is correlational. To hold the base model fixed, we train four con-

Table 4: Coarse 50%-crossing diagnostic (d_{50}) per task family, double gate. d_{50} is estimated by linear interpolation between the two depth bins ($\{2, 4, 6, 8\}$) bracketing 50% residual failure. Ranges reflect the uncertainty band from the 95% Clopper–Pearson CI on the underlying binomial rates. All values are approximate due to the coarse 4-bin resolution and small gate-passing sample sizes at deeper depths.

Model	RANK	SUCC.	INT_DECOY	Agg.
Qwen3 (RL+distill)	<4	[4,6]	[4,6]	[3,5]
Qwen2.5-Inst (multi-stage)	<4	[4,6]	[4,6]	[3,5]
DeepHermes (R1 distill)	<4	<4	<4	< 3
Mistral-Inst (instruction FT)	<4	<4	[4,5]	< 3

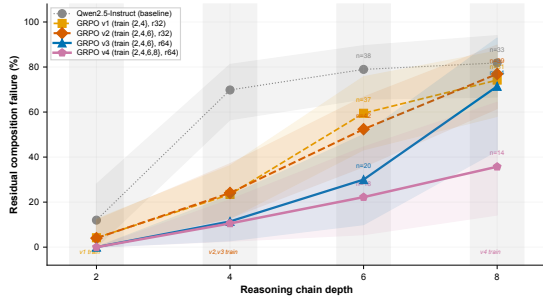


Figure 3: Baseline (Qwen2.5-Instruct, dotted) vs. four LoRA-GRPO variants. Shaded bands mark training-depth regions. v1/v2 (rank 32, 4 epochs) show that adding depth-6 training improves $d=6$ but not $d=8$; v3 (rank 64, 8 epochs) further reduces $d=6$ to 30.0% with minimal $d=8$ transfer; v4 (train $\{2, 4, 6, 8\}$, rank 64, 8 epochs) cuts $d=8$ to 35.7%, showing that the observed depth profile changes when depth 8 enters the training mix. All CIs are bootstrap 95%.

trolled LoRA-GRPO variants on Qwen2.5-7B-Instruct (Yang et al., 2024)¹: v1 trains on depths $\{2, 4\}$; v2 adds depth 6; v3 uses the same depths as v2 with a larger LoRA rank and twice as many epochs; and v4 adds depth 8 to the heavier-budget setup. Thus v1 holds out depths $\{6, 8\}$, v2 and v3 hold out depth 8, and v4 evaluates depth 8 in-distribution. Evaluation follows the protocol in §4.

Training improves covered depths without changing atoms. Across the four variants, training cuts depth-4 residual failure by 46–59 pp while atomic stability moves by only 1–3 pp. This provides a controlled Δ_{comp} estimate in the present setup; at depths 4 and 6, the baseline and GRPO v1 bootstrap CIs are disjoint (Figure 3). When depth 8

¹All variants use learning rate 10^{-5} , 8 generations per prompt, reward = consistency match + XML format + anti-template-echo, Unsloth (Han et al., 2023) (memory-efficient LoRA (Hu et al., 2022)) with patched TRL. v1/v2: LoRA rank 32, 4 epochs, ~ 1 h. v3/v4: LoRA rank 64, 8 epochs, ~ 3.2 h (v3) / ~ 5 h (v4). Details in Appendix G.

Table 5: Residual failure (%) for four GRPO variants vs. baseline. v2 adds depth-6 to training (7 pp improvement at $d=6$); v3 uses heavier budget (rank 64, 8 epochs), cutting $d=4$ and $d=6$ residual failure by an additional 13–22 pp. v4 (train $\{2, 4, 6, 8\}$, rank 64, 8 epochs) reduces $d=8$ by 46 pp vs. baseline and by 36 pp vs. v3, showing that training on depth 8 largely closes the depth-8 gap. Atomic stability remains within 1 pp of baseline (89.9%) for all four.

	$d=2$	$d=4$	$d=6$	$d=8$
Qwen2.5-Inst baseline	12.0	69.8	78.9	81.8
GRPO v1 (train $\{2, 4\}$, r32)	4.2	23.4	59.5	74.2
GRPO v2 (train $\{2, 4, 6\}$, r32)	4.0	24.1	52.4	76.9
GRPO v3 (train $\{2, 4, 6\}$, r64)	0.0	11.4	30.0	71.4
GRPO v4 (train $\{2, 4, 6, 8\}$, r64)	0.0	10.5	22.2	35.7
Δ v1	−7.8	−46.4	−19.5	−7.6
Δ v2	−8.0	−45.7	−26.6	−4.9
Δ v3	−12.0	−58.4	−48.9	−10.4
Δ v4	−12.0	−59.3	−56.7	−46.1

is added to training, v4 reduces its residual failure to 35.7%, a 46 pp drop from baseline and a 36 pp drop from v3. Depth 6 also improves from 30.0% to 22.2%.

Transfer to unseen depths remains weak.

Training on $\{2, 4\}$ improves depth 6 by 19 pp but depth 8 by only 8 pp. Adding depth 6 in v2 yields a further 7 pp gain at depth 6 but does not improve depth 8 relative to v1. A heavier budget without depth-8 data (v3) cuts depth-6 failure to 30.0% while leaving depth 8 at 71.4%, only 5 pp below v2. The ratio of out-of-distribution gain to in-distribution gain is 0.21. Because v4 evaluates depth 8 in-distribution, its larger gain reflects expanded training coverage, not depth extrapolation. Across variants, training closes much of the failure gap at covered depths but transfers only weakly to held-out depths.

Cross-model replication. The same GRPO recipe (train $\{2, 4, 6\}$, LoRA rank 32, 4 epochs) applied to Llama3-8B-Instruct (Llama Team, AI @ Meta, 2024) reduces $d=4$ residual failure to 17.9% and $d=6$ to 42.9%, outperforming the matched Qwen GRPO v2 by 13 pp at $d=4$ and 15 pp at $d=6$ (Table 6). The Δ_{comp} direction also appears under a second base-model family, reducing concern that the result is Qwen-specific. On Mistral-7B-v0.1 (Jiang et al., 2023), the same recipe raises atomic stability enough for 34 cases to survive the double gate at $d=4$, with 0% residual failure. Thus GRPO can improve both atomic knowledge and composition from a low-stability base in this setting, although the gate-passing sample remains small.

Table 6: Same GRPO recipe (train $\{2, 4, 6\}$, rank 32, 4 epochs) on two base-model families. The Δ_{comp} direction is consistent across both. Parenthesised values are n (double-gate-passing cases).

Model	Residual failure (%)		
	$d=2$	$d=4$	$d=6$
Qwen2.5-Inst (baseline)	12.0 (25)	66.7 (66)	78.7 (47)
Qwen2.5-Inst + GRPO v2	4.0 (25)	30.8 (65)	58.2 (55)
Llama3-8B-Inst + GRPO	0.0 (19)	17.9 (39)	42.9 (28)

Table 7: Same base model, same data ($d \in \{2, 4\}$), same LoRA budget. Under the double gate, GRPO cuts residual failure by 40–54 pp relative to SFT.

Recipe	Atom (%)	$d=2$	$d=4$
Baseline (untrained)	88.6	12.0	69.8
SFT-answer (answer only)	85.3	19.0	63.8
SFT-trace (chain + answer)	85.8	23.5	76.9
GRPO v1 (outcome RL)	87.1	4.2	23.4

Training objective vs. base model. To isolate the training objective from depth, we trained two additional variants on the same base model (Qwen2.5-7B-Instruct), same data ($d \in \{2, 4\}$), and same LoRA budget: SFT-answer (gold answer only) and SFT-trace (reasoning chain + answer). Under the double gate, GRPO’s residual failure at depth 4 (23.4%) is 54 pp lower than SFT-trace (76.9%) and 40 pp lower than SFT-answer (63.8%; Table 7). SFT-trace also exceeds the untrained baseline (69.8%) and SFT-answer in residual failure. In this controlled setting, imitating reasoning chains does not reliably transfer to novel composition questions.

6 Diagnostics

We use two inference-time diagnostics to test whether additional generation-time computation reduces residual failure. Additional probes (prompt-end patching, contamination control, and hidden-state trajectories) are in Appendix J and K.

Thinking mode as a generation-process probe. If insufficient generation-time computation contributes to residual failure, allocating more inference compute should reduce the measured rate. We evaluated Qwen3-8B on D4V2 with native CoT reasoning enabled (thinking=true, up to 16K internal tokens). Atomic stability is 82.3%, compared with 87.7% in standard mode. Among gate-passing failures, thinking recovers 72.2% at $d=4$ (13/18), 70.0% at $d=6$ (7/10), and 75.0% at $d=8$ (3/4).

Residual failure is also substantially flatter across depths 4, 6, and 8 (28.6%, 18.8%, and 23.1%) than in standard mode (56.8%→92.3%). This reduction is consistent with generation-time computation contributing to the observed failures. The task-family response is heterogeneous: TEMPORAL_RANK drops to 4.0%, while TEMPORAL_SUCCESSOR remains at 53.1%.

Convergent evidence from explicit CoT. On Qwen2.5-7B-Instruct with explicit CoT prompting, atomic stability drops 10 points to 79.7%, while residual failure falls to 47.1% overall. By depth, it is 42.3% at $d=4$, 73.3% at $d=6$, and 71.4% at $d=8$ (Appendix T). The reduction is strongest at depth 4, while the deeper estimates have small denominators. This second model shows a qualitatively similar reduction even as atomic stability moves in the opposite direction, further supporting generation-time computation as an important contributor.

Implication. Changing inference-time computation alone recovers many observed failures. The controlled GRPO intervention (§5) separately shows that residual failure can be reduced by training while atomic stability remains similar. Together, these results show sensitivity to both training and decoding in the tested settings, while leaving representational bottlenecks unresolved.

Failure taxonomy. We adjudicated 251 double-gate failure candidates across models and task families (Appendix P). Of these, 181 (72%) receive a structural composition label, such as temporal boundary confusion, intermediate-variable loss, or a binding shift; 70 (28%) are scoring or format artefacts where the reasoning is substantially correct. Format and scoring therefore account for a minority, but a non-negligible one, of candidate residual failures.

7 Discussion

7.1 What the protocol changes

The central empirical lesson is that post-training can change atomic stability and gate-controlled composition differently. Post-trained models with closely matched aggregate atomic stability differ by more than 40 percentage points in shallow residual failure. The same-case analysis shows that the principal training trend remains when every variant is evaluated on the same gate-passing examples.

The strongest controlled contrast is between SFT and outcome-verified RL. With the same

Qwen2.5 base model, depth-2/4 training questions, and LoRA budget, GRPO reaches 23.4% residual failure at depth 4, compared with 76.9% for SFT-trace and 63.8% for SFT-answer (Table 7). SFT-trace also underperforms the untrained baseline. In this controlled setting, trace imitation did not reliably transfer to held-out composition questions; the result does not establish a general disadvantage of reasoning-trace distillation.

7.2 Depth, structure, and inference

Training gains are strongly aligned with covered depths. Adding depth-6 training has little effect at depth 8, whereas including depth 8 reduces its residual failure from 81.8% to 35.7% (Table 5). Because depth 8 becomes in-distribution, this contrast is consistent with training-depth coverage shaping the gain and leaves architectural limits unresolved. Task family also changes failure by up to 54 points at the same depth, so the interpolated d_{50} is best read as a coarse curve summary rather than a scalar capability score.

Two inference diagnostics provide complementary evidence. Residual failure remains high on synthetic tasks when facts are explicitly supplied (Table 2), showing that retrieval is not the only possible source of multi-step error. Separately, chain-of-thought inference recovers 70–75% of Qwen3 gate-passing failures (§6). This recovery is consistent with generation-time computation contributing substantially, while leaving open whether every required representation is already available.

7.3 Scope and use

Our main evidence concerns closed-book temporal factual chains. A mixed-domain pilot (Appendix M) preserves non-zero residual failure and the broad model ordering, providing preliminary evidence beyond the primary domain. The in-context synthetic experiment studies assembly over supplied facts and therefore has a different estimand from the closed-book analysis.

The practical recommendation follows directly: when a benchmark can identify its constituent facts, report stable atomic access and composition on the gate-passing cases separately. Same-case intersections should accompany cross-model comparisons when gate-passing subsets differ, and deep cells should report denominators and uncertainty.

8 Conclusion

In temporal factual chains, stable atomic knowledge does not guarantee correct composition. We introduced a double-gate protocol that admits a multi-hop case only after every required fact is stable across paraphrases and every aligned sub-question is correct in isolation. On D4v2, closely matched post-trained models still differ by 47.4 percentage points at depth 2, and controlled GRPO training reduces residual failure relative to matched SFT while transferring weakly to held-out depths. Same-case intersections preserve the training trend, and CoT recovers many failures by changing inference-time computation. These results support evaluating atomic stability and gate-controlled composition as separate quantities, with explicit uncertainty at larger depths.

9 Limitations

1. Residual failure is operational. Format and scoring artefacts can inflate it: 70/251 candidate failures receive this label in our taxonomy. Conversely, the validated adjudicator’s remaining errors are false positives for correctness and can slightly underestimate failure. The metric is therefore not a direct observation of a latent reasoning state.

2. Cross-recipe comparisons remain correlational. Off-the-shelf models differ in architecture and training data. The same-base GRPO/SFT interventions and same-case intersections isolate some of these factors, but conclusions across recipes remain limited.

3. Deep estimates remain less precise. Doubling the depth-6/8 candidate pools increases gate-passing denominators and preserves the qualitative trend, but some cells, especially Mistral at depth 8, remain small. The interpolated d_{50} values are coarse summaries, not precise capability limits.

4. Domain scope is narrow. The main benchmark covers closed-book temporal facts. The in-context synthetic study uses a different estimand, and the mixed-domain pilot is too small to establish broad domain generality.

5. Validation is sampled. The adjudication study uses three author annotators and covers 225 temporal decisions; the paraphrase and decomposition checks cover 50 samples each. High agreement supports the measurement pipeline but cannot guarantee that every benchmark item or automatic decision is error-free.

Acknowledgments

This work was supported by the Platform-Oriented “Open Competition” Project of Hangzhou High-tech Zone (Binjiang) (Grant No. 2025JBGS-PT004).

Generative AI assistance was used for language polishing and organizational revisions during preparation of the final manuscript. The authors reviewed and verified all resulting changes and take full responsibility for the paper.

References

- Zeyuan Allen-Zhu and Yuanzhi Li. 2023. [Physics of language models: Part 3.1, knowledge storage and extraction](#). *arXiv preprint arXiv:2309.14316*.
- Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. [Hopping too late: Exploring the limitations of large language models on multi-hop queries](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Haonan Dong, Kehan Jiang, Haoran Ye, Wenhao Zhu, Zhaolu Kang, and Guojie Song. 2026. [NeuReasoner: Towards explainable, controllable, and unified reasoning via mixture-of-neurons](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22534–22558, San Diego, California, United States. Association for Computational Linguistics.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. [Faith and fate: Limits of transformers on compositionality](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 70293–70332.
- Weidong Guo, Dakai Wang, Zixuan Wang, Hui Liu, and Yu Xu. 2026. [T-Mem: Memory that anticipates, not archives](#). *Preprint*, arXiv:2606.15405.
- Daniel Han, Michael Han, and Unsloth team. 2023. [Unsloth](#).
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7B](#). *arXiv preprint arXiv:2310.06825*.
- Kehan Jiang, Haonan Dong, Zhaolu Kang, Zhengzhou Zhu, and Guojie Song. 2026. [FoE: Forest of errors makes the first solution the best in large reasoning models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24591–24629, San Diego, California, United States. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Hehai Lin, Yu Yan, Zixuan Wang, Bo Xu, Sudong Wang, Weiquan Huang, Ruochen Zhao, Minzhi Li, and Chengwei Qin. 2026a. [Unified-MAS: Universally generating domain-specific nodes for empowering automatic multi-agent systems](#). *Preprint*, arXiv:2603.21475.
- Hehai Lin, Qi Yang, and Chengwei Qin. 2026b. [Skill-MAS: Evolving meta-skill for automatic multi-agent systems](#). *Preprint*, arXiv:2606.18837.
- Llama Team, AI @ Meta. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. [Measuring and narrowing the compositionality gap in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical](#)

- reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Teknium, Roger Jin, Chen Guang, Jai Suphavadeeprasit, and Jeffrey Quesnelle. 2025. [DeepHermes 3 preview](#). Hugging Face model card, NousResearch, DeepHermes-3-Llama-3-8B-Preview.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. [MuSiQue: Multi-hop questions via single-hop question composition](#). *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Zheng Weihua, Xin Huang, Zhengyuan Liu, Tarun Kumar Vangani, Bowei Zou, Xiyan Tao, Yuhao Wu, AiTi Aw, Nancy F. Chen, and Roy Ka-Wei Lee. 2026. [AdaMCoT: Rethinking cross-lingual factual reasoning through adaptive multilingual chain-of-thought](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):33863–33871.
- Zheng Weihua, Roy Ka-Wei Lee, Zhengyuan Liu, Wu Kui, AiTi Aw, and Bowei Zou. 2025. [CCL-XCoT: An efficient cross-lingual knowledge transfer method for mitigating hallucination generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1768–1788, Suzhou, China. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Bohan Yang, Yijun Gong, Zhi Zhang, Ge Zhang, Wenpeng Xing, and Meng Han. 2026. [TriLens: Per-layer logit-lens entropy for white-box hallucination detection](#). *Preprint*, arXiv:2606.01033.
- Sohee Yang, Nora Kassner, Elena Gribovskaya, Sebastian Riedel, and Mor Geva. 2025b. [Do large language models perform latent multi-hop reasoning without exploiting shortcuts?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3971–3992.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Weihua Zheng, Zhengyuan Liu, Tanmoy Chakraborty, Weiwen Xu, Xiaoxue Gao, Bryan Chen Zhengyu Tan, Bowei Zou, Chang Liu, Yujia Hu, Xing Xie, Xiaoyuan Yi, Jing Yao, Chaojun Wang, Long Li, Rui Liu, Huiyao Liu, Koji Inoue, Ryuichi Sumida, Tatsuya Kawahara, and 16 others. 2026. [MMAC: A multilingual, multimodal alignment framework for cultural grounding evaluation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21667–21709, San Diego, California, United States. Association for Computational Linguistics.

A Prompt template and inference details

All models receive the following XML-structured prompt via their native chat template with `enable_thinking=false` for Qwen3:

```
You are a careful research assistant.
No external evidence is provided.
Answer from your internal knowledge
only. If you are genuinely uncertain,
answer exactly INSUFFICIENT_EVIDENCE.

Respond using exactly this XML format
and nothing else after the closing
answer tag:

<reasoning>Use 1-4 concise sentences
based only on your internal
knowledge.</reasoning>

<answer>final answer here</answer>

Do not use <think> tags.

<question>{QUESTION}</question>

If you are genuinely uncertain,
the final answer must be
<answer>INSUFFICIENT_EVIDENCE</answer>.
```

We use greedy decoding ($T=0$), a maximum of 512 output tokens, and `</answer>` as the stop string. Across all runs, $\geq 99\%$ of completions emit the closing tag. Inference uses vLLM (Kwon et al., 2023) 0.19 with each model’s native chat template; cross-validation against HuggingFace transformers is reported in Appendix I.

B D4v2 benchmark construction

D4v2 is the long-horizon composition benchmark used for all main-text tables and figures. It was constructed by the authors over approximately three weeks: one week for fact curation

and verification, one for algorithm implementation and question generation, and one for manual spot-checking and refinement. Its four task families use temporal facts about historical events, including dates, causal order, and interval membership. These facts support unambiguous atomic verification and deterministic composition checks. The benchmark, sub-question decompositions, and evaluation protocol are included in the supplementary artifact accompanying this paper.

Atom generation. Atomic facts are single temporal propositions independently verified by the authors against Wikipedia and standard reference timelines (e.g., Britannica, Oxford Reference). Each fact is cross-checked by at least two sources. Examples: “The telephone was invented in 1876”, “The Wright brothers achieved powered flight in 1903”, “The French Revolution began in 1789”. The atom pool spans five centuries (15th–20th) of political, scientific, and cultural history to reduce domain-specific confounds. Template-based generation expands each atom into four surface variants that change syntax, question framing, and entity mentions. We manually reviewed the templates and conducted the sampled validation reported below. The full set contains 79 temporal facts and 316 atomic probes (79×4 variants).

Sub-question decomposition. For each composition case, sub-questions are generated deterministically from its constituent atomic facts using type-specific templates (e.g., “In what year did [EVENT] occur?” for date facts, “When was [ENTITY] [VERB]?” for construction/invention facts). The mapping from atom to sub-question is deterministic: given a fact f_i with relation type r_i , the template for r_i produces a unique sub-question. This removes LLM-generation variability; we assess decomposition quality on a sample below.

Composition question construction. From the atom pool, we algorithmically generate multi-hop composition questions at depths $d \in \{2, 4, 6, 7, 8, 9, 11\}$ by sampling combinations of d facts and applying the family-specific composition template. Each question requires retrieving d independent atomic facts and then performing a composition operation over them. The four families instantiate different operations:

TEMPORAL_RANK. Sort d events by date. Requires retrieving all d dates and then order-

ing them. Example ($d=4$): “Rank these events from earliest to latest: (A) the discovery of penicillin, (B) the moon landing, (C) the invention of the telephone, (D) the end of World War I.”

TEMPORAL_SUCCESSOR. Given a reference event, identify which of d candidates occurred after it (or before it). Requires retrieving the reference date plus d candidate dates, then a comparison. Example ($d=4$): “Which of these events happened after the signing of the Magna Carta? (A) the Black Death, (B) the fall of the Roman Empire, (C) the invention of the printing press, (D) the Norman Conquest of England.”

TEMPORAL_INTERVAL_DECOY. Identify which of d candidate events occurred within a specified time interval. The interval is defined by two boundary events whose dates must be retrieved; $d - 1$ decoy events lie outside the interval. Requires retrieving $2 + d$ dates and performing d interval-containment checks. Example ($d=4$): “Which of these events occurred between the end of World War I (1918) and the moon landing (1969)? (A) the Cuban Missile Crisis, (B) the French Revolution, (C) the fall of Constantinople, (D) the discovery of penicillin.”

PAIR_FAR_CONTROL. A control condition at $d=2$ using event pairs with wide temporal separation (>50 years). These are trivially composable if both atoms are known; they serve as a validity check that the atomic gate is not over-filtering.

Case counts and depth coverage. Table 8 reports the case count per task family and depth. TEMPORAL_RANK and TEMPORAL_SUCCESSOR cover depths 4, 6, 8 (30 cases each); TEMPORAL_INTERVAL_DECOY covers the full depth ladder 4–11 (30 cases each, plus odd depths); PAIR_FAR_CONTROL only appears at depth 2 (30 cases). The total benchmark contains 390 composition cases.

Sampled construction validation. We validated the two generated components separately. For atomic paraphrases, we sampled 50 fact groups and asked whether all variants express the same underlying fact and preserve the same gold answer. All 50 groups passed both checks.

Table 8: Case count per task family and depth in D4v2.

Task family	$d=2$	$d=4$	$d=6$	$d=7$	$d=8$	$d=9$	$d=11$
TEMPORAL_RANK	—	30	30	—	30	—	—
TEMPORAL_SUCCESOR	—	30	30	—	30	—	—
TEMPORAL_INTERVAL_DECOY	—	30	30	30	30	30	30
PAIR_FAR_CONTROL	30	—	—	—	—	—	—

For decompositions, we sampled 50 composition cases across all four task families and depths 2–11. Annotators judged whether the sub-question set was necessary and sufficient to derive the main answer and whether any sub-question was ambiguous. All 50 sets were judged necessary and sufficient, and none was marked ambiguous. These results validate the sampled outputs while leaving the unsampled items subject to possible construction errors.

C Atomic stability by model

Table 9 reports per-model atomic stability and joint stability at depth $k=2$. Base models sit below 15% individual atomic stability, making joint stability at any depth ≥ 2 effectively zero—the atomic gate removes them entirely. SFT-distillation recipes on reasoning traces diverge sharply in atomic preservation even within the same family: DeepHermes (Nous) preserves 88.0%, while DeepSeek R1-Distill (DeepSeek-AI, 2025) drops to 40.5% on a Qwen2.5 base and 76.9% on a Llama-3 base. The gate excludes the Qwen R1-Distill variant entirely (joint stability at $k=2$ is $\approx 16\%$); the Llama R1-Distill variant survives but has insufficient n at depth ≥ 6 for stable estimates.

Table 9: Atomic stability (per-probe consistency-match rate) and joint stability at $k=2$ for all evaluated models. Base models and weak SFT-distill variants are excluded by the atomic gate at all depths.

Model	Recipe	Atom stab.	$P(\text{all stable})@k=2$
Llama-3-8B	Base	0.0%	≈ 0
Mistral-7B-v0.1	Base	5.1%	≈ 0
Qwen2.5-7B	Base	14.2%	$\approx 2\%$
R1-Distill-Qwen-7B	SFT distill (R1)	40.5%	$\approx 16\%$
R1-Distill-Llama-8B	SFT distill (R1)	76.9%	$\approx 59\%$
Mistral-7B-Inst	Instruction FT	78.8%	$\approx 62\%$
Qwen3-8B	RL + distillation	87.7%	$\approx 77\%$
DeepHermes-3-8B	R1 distillation (Nous)	88.0%	$\approx 77\%$
Qwen2.5-7B-Inst	SFT + DPO + GRPO	89.9%	$\approx 81\%$

D Within-task-family depth curves

Figure 6 plots residual failure vs. depth for each of the three depth-covering task families (TEMPORAL_RANK, TEMPORAL_SUCCESOR,

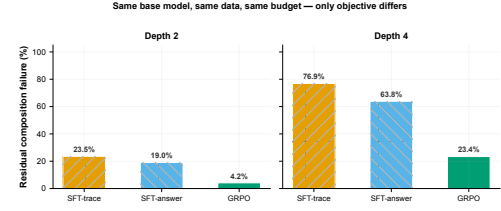


Figure 4: Controlled comparison of residual composition failure for SFT-trace, SFT-answer, and GRPO at depths 2 and 4. The models share the same Qwen2.5 base, training data, and LoRA budget; only the training objective differs.

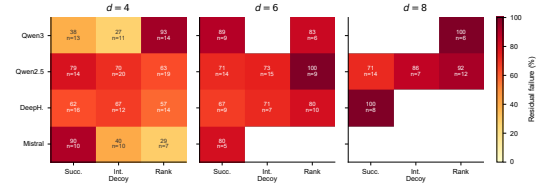


Figure 5: Residual composition failure (%) across {recipe} $\times$ {task family} $\times$ {depth} on D4v2 (reproduced from main text §4). Within a fixed depth, task family changes failure by up to 66 pp.

TEMPORAL_INTERVAL_DECOY) across the four post-trained models. TEMPORAL_RANK is consistently the hardest, reaching 93% residual failure at depth 4 for Qwen3. Model ordering by residual failure differs across task families, confirming that Δ_{depth} is entangled with task choice: a model that is strong on ranking may be weak on successor inference, and vice versa. This interaction precludes a single “depth ceiling” number and motivates our per-task-family reporting in Table 4.

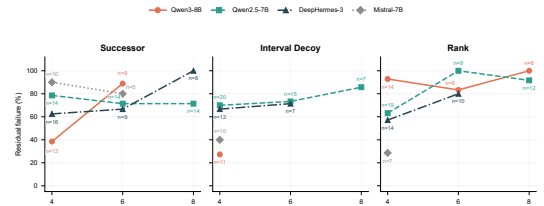


Figure 6: Residual composition failure vs. depth, per task family and model. TEMPORAL_RANK (left) is the hardest family across all models; model ordering reverses between TEMPORAL_SUCCESOR (centre) and TEMPORAL_INTERVAL_DECOY (right).

E Double-gate protocol definition

We formalise the double-gate protocol used throughout the paper. Consider a composition case with k sub-questions $q_1, \dots, q_k$. The model

answers each sub-question independently (single-hop prompt, same XML template) and the main question (multi-hop prompt).

Gate 1: atomic stability. For a sub-question q_i with gold answer a_i^* and model answer $\hat{a}_i$, we define two matching tiers:

- **Exact match:** $\hat{a}_i$ matches a_i^* after normalisation (lowercasing, punctuation stripping, date format canonicalisation).
- **Consistency match:** $\hat{a}_i$ is semantically equivalent to a_i^* under an LLM-based adjudicator (Gemini 2.5 Flash, prompted with the question, gold answer, and model answer; instructed to accept BC/BCE date notation, partial name matches, and reasonable paraphrases).

A fact is *stable* if the model answers correctly across all paraphrase variants under the consistency tier. An atom passes Gate 1 iff all its constituent facts are stable.

Gate 2: sub-question correctness in isolation.

The main question passes Gate 2 iff every sub-question in the case is individually answered correctly by the model (consistency tier) *when asked in isolation*. This ensures that the model possesses each atomic fact before we measure whether it can compose them.

Residual composition failure. A case is a *residual composition failure* iff it passes both gates but the main-question answer is incorrect under the consistency matcher. Let G_d be the gate-passing cases at depth d , and let $E_d \subseteq G_d$ be those with an incorrect main answer. The residual failure rate is $R(d) = |E_d|/|G_d|$.

Why two gates? Prior work commonly checks whether aligned single-hop questions are answered correctly before scoring their composition (Press et al., 2023). Gate 2 retains this check; Gate 1 further requires correct answers to every fact across paraphrased probes. As reported in §4, adding Gate 1 removes 15–44% of cases that the sub-question-only baseline would admit. The combined protocol therefore conditions residual failure on both reproducible atomic access (Gate 1) and correct isolated retrieval in the chain’s own wording (Gate 2).

F Robustness to admitted cases and depth sample size

Same-case intersections. Model-specific gates can admit different examples even when aggregate atomic stability is similar. We therefore intersect case identifiers and evaluate each comparison on examples that pass both gates for every model in that comparison. Table 10 reports the two central checks. The four-model intersection is small ($n=26$) and is used only as qualitative evidence. The controlled Qwen2.5/GRPO intersection is larger ($n=80$) and preserves the monotonic reduction from 60.0% to 30.0% residual failure.

Table 10: Residual failure on identical double-gate-passing cases. Wilson 95% CIs are computed on the shared subset.

Comparison	Model	Shared n	Failure	95% CI
Main models	Qwen3-8B	26	30.8%	[16.5, 50.0]
	Mistral-7B-Inst	26	34.6%	[19.4, 53.8]
	Qwen2.5-7B-Inst	26	42.3%	[25.5, 61.1]
	DeepHermes-3	26	53.8%	[35.5, 71.2]
Same-base GRPO	Baseline	80	60.0%	[49.0, 70.0]
	v1: train {2, 4}, r32	80	47.5%	[36.9, 58.3]
	v2: train {2, 4, 6}, r32	80	43.8%	[33.4, 54.7]
	v3: train {2, 4, 6}, r64	80	33.8%	[24.3, 44.6]
	v4: train {2, 4, 6, 8}, r64	80	30.0%	[21.1, 40.8]

The same-case analysis also preserves the controlled objective comparison. Across 95 cases shared by the baseline, SFT-answer, SFT-trace, and GRPO v1, residual failures are 64.2%, 63.2%, 68.4%, and 45.3%, respectively. Thus differences in model-specific admitted subsets do not explain the GRPO advantage.

Expanded depth-6/8 analysis. The original high-depth cells use 30 candidates per task family. We repeated the analysis after increasing the depth-6 and depth-8 pools to 60 candidates per family. Across the evaluated post-trained models and GRPO/SFT variants, the qualitative depth trend and relative recipe conclusions are unchanged, while larger gate-passing denominators narrow the intervals. For Qwen2.5-Instruct on TEMPORAL_INTERVAL_DECOY, depth 6 increases from 12 to 25 gate-passing cases and depth 8 from 8 to 17; both point estimates remain stable. Mistral depth 8 is the remaining recurrent small- n case. We therefore use this expansion as a robustness check and continue to treat individual deep cells as imprecise when their samples are small.

G GRPO training details

Setup. We train Qwen2.5-7B-Instruct with LoRA adapters on all projection layers (q/k/v/o_proj, gate/up/down_proj). v1 and v2 use rank 32, alpha 64, dropout 0.05, 4 epochs (~ 1 GPU-hour per variant); v3 and v4 use rank 64, alpha 128, dropout 0.05, 8 epochs (~ 3.2 –5 GPU-hours per variant). All variants use AdamW, learning rate 10^{-5} , per-device batch size 1, gradient accumulation 8, KL coefficient $\beta=0.04$. Training uses Unsloth’s patched TRL 0.24 to work around TRL-1.2 sampling issues on Qwen2+LoRA. Each run uses one RTX 4090. Total compute across training and all evaluations (6 off-the-shelf models at 7–13B scale, SFT baselines, cross-model GRPO replication, chain-of-thought, in-context, prompt-robustness, thinking-mode, and mechanism diagnostics) on $6\times$ RTX 4090 is approximately 60 GPU-hours.

Training data. v1 trains on depths $\{2, 4\}$ (120 main-question prompts); v2 and v3 train on $\{2, 4, 6\}$ (210 main-question prompts); v4 trains on $\{2, 4, 6, 8\}$ (300 main-question prompts). For v1–v3, depth 8 is strictly held out; for v1, depth 6 is also held out.

Reward function. The reward combines consistency match (1.0), an XML-format bonus (2×0.1 for opening and closing tags), and an anti-template-echo penalty (-0.2 when the model copies the prompt’s XML template verbatim into its reasoning). We sample eight generations per prompt and compute the reward independently for each generation.

Evaluation protocol. Evaluation uses greedy decoding ($T=0$) and consistency-tier adjudication, as in the main experiments. Atomic stability moves <2 pp in all four variants relative to the Qwen2.5-7B-Instruct baseline (89.9%).

H Cross-model GRPO replication

To test whether the GRPO composition benefit generalises, we applied the same GRPO recipe—train $\{2, 4, 6\}$, LoRA rank 32, 4 epochs, learning rate 10^{-5} —to Llama3-8B-Instruct and to Mistral-7B-v0.1 (base). The Llama3 instruct variant uses the same prompt template as the Qwen GRPO v2 evaluation; the Mistral base variant uses a ChatML template injected at merge time for compatibility with the instruct-style eval-

uation prompts. Both are evaluated under the identical double-gate protocol.

Table 11: Same GRPO recipe (train $\{2, 4, 6\}$, rank 32, 4 epochs) on three base models. Mistral base has a small gate-passing sample at depth 8 ($n=8$) due to base-model atomic instability; its numbers are included for completeness but should be interpreted with caution.

Base model	$d=2$	$d=4$	$d=6$	$d=8$
Qwen2.5-7B-Instruct (baseline)	12.0 (25)	66.7 (66)	78.7 (47)	80.0 (40)
Qwen2.5-7B-Instruct + GRPO v2	4.0 (25)	30.8 (65)	58.2 (55)	80.9 (47)
Llama3-8B-Instruct + GRPO	0.0 (19)	17.9 (39)	42.9 (28)	85.7 (14)
Mistral-7B-v0.1 + GRPO v5	5.9 (17)	0.0 (34)	0.0 (26)	0.0 (8 [†])

[†] $n < 15$ gate-passing cases; point estimate should be interpreted cautiously.

The GRPO composition benefit replicates across base-model families. Llama3 GRPO outperforms the matched Qwen GRPO v2 by 13–15 pp at both $d=4$ and $d=6$, supporting the conclusion that outcome RL on composition data improves composition net of atoms under a different architecture and pretraining distribution. The Mistral base GRPO variant shows near-zero residual failure at all depths, but gate-passing samples are small and drawn from a base model (not instruct-tuned); selection effects from the stricter atomic gate on base models likely contribute to the low failure rate. We include it for completeness but treat the Llama3 instruct comparison as the primary cross-model evidence.

I vLLM vs. HuggingFace cross-validation

We re-ran Qwen3-8B on a 10-probe atomic split with the HuggingFace transformers reference and vLLM 0.19. Final-answer agreement under the consistency matcher is 9/10; the single disagreement is a BC/BCE notational difference that both matchers accept. All main-text numbers use the vLLM pipeline (~ 2 minutes per model for the full 3196-row benchmark).

J Mechanism probe details

Boundary case set. The set contains 16 Qwen3-8B depth-4 success cases and 10 depth-6 failure cases, all of which pass the atomic gate. Cases come from TEMPORAL_SUCCESSOR and TEMPORAL_INTERVAL_DECOY. The HuggingFace runner saves hidden states at 10 layers: $\{0, 4, 8, 12, 16, 20, 24, 28, 32, 35\}$.

Prompt-end patching sweep. We pair successful donors with failure recipients, capped at 16

pairs. For each layer, we replace the recipient’s prompt-end hidden state with the donor state and resume generation. Across $16 \times 10 = 160$ patches, 0 cases flip to the donor’s correct answer; at layers 8–16, at most 2 of 16 recipients shift to an abstention category rather than the correct answer.

Cosine similarity of prompt-end states. Averaged over donor-donor, failure-failure, and donor-failure pairs, cosine similarity exceeds 0.95 at every tracked layer (layer 0: 0.9999, layer 35: 0.959). The gap between in-group and cross-group similarity never exceeds 0.01. Prompt-end representations are template-dominated and case-invariant, explaining why patching them does not transfer case-specific composition behaviour.

Step-to-step cosine trajectory. Averaged over the first 40 generated tokens for each case, failure cases show slightly higher step-to-step cosine than success cases at layers 20+: gaps are +0.002 (layer 20), +0.013 (layer 24), +0.016 (layer 28), +0.015 (layer 32), +0.031 (layer 35). Activation norms are within 5% between groups at every layer. The elevated step-to-step similarity in failures is consistent with a model that settles into a repetitive or “collapsed” generation trajectory, but the effect size is too small to be definitive without a larger mechanistic sample.

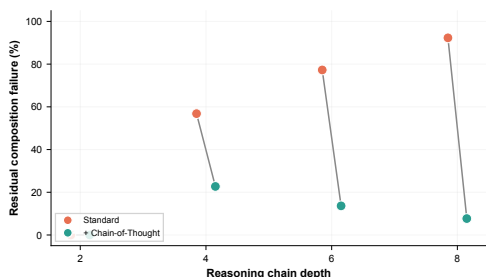


Figure 7: Residual composition failure for Qwen3-8B with standard decoding and native thinking mode. Thinking reduces failure at depths 4, 6, and 8; annotations report the absolute reduction and the standard-mode double-gate denominator.

K Pollution-control (post-cutoff) details

Dataset construction. Gemini 2.5 Flash generated atomic facts in batches of 10 for each of five categories: politics, science, sports, culture, and technology product launches. We retained events from November 2024 onward. The filtered set contains 31 atoms with four paraphrases each

(124 atomic probes), from which we algorithmically built 30 depth-2 and 20 depth-4 composition cases. We manually verified that every fact post-dates the training cutoff of each evaluated model.

Results. Table 12 reports atomic stability on D4v2 vs. post-cutoff data. The 55–72 pp drop is consistent across all recipes and precludes gate passage at any depth.

Table 12: Atomic stability on D4v2 vs. post-cutoff events. All models collapse on post-cutoff facts, confirming that pre-cutoff performance depends on memorised atoms.

Model	D4v2	Post-cutoff	Δ	Pass
Qwen3-8B	87.7%	33.1%	−54.6	0
Qwen2.5-7B-Inst	89.9%	29.0%	−60.9	0
DeepHermes-3-8B	88.0%	17.7%	−70.3	0
Mistral-7B-Inst	78.8%	7.3%	−71.5	0

Interpretation. If the D4v2 failures were chain-level memorisation of pre-cutoff facts, atomic stability should hold across the cutoff (memorisation should affect composition and single-hop retrieval equally). Instead, atomic stability drops sharply on post-cutoff events, showing that stable atomic access is largely confined to pre-cutoff facts. The pre-cutoff failures we report therefore occur on atoms that the gate verifies as stably accessible but that the model still fails to compose. Direct measurement of post-cutoff composition would require an in-context RAG protocol where all facts are supplied in the prompt; this is a natural extension of our framework.

L E2: Short-chain composition on jointly stable facts

E2 is the most directly interpretable short-chain benchmark in our suite. We take the intersection of facts that are stable across Qwen2.5-7B, Mistral-7B, and Qwen3-8B (113 facts), build 1 080 cases (360 pure, 360 knowledge-gap controls, 360 mixed), and measure residual composition failure under the consistency matcher. All controls (knowledge-gap and mixed) pass at $\geq 99.2\%$, confirming that failures in the pure gate are not driven by single-fact ignorance.

M E3: Cross-domain pilot

E3 extends the protocol to scientific-fact composition (event betweenness, timeline binding, clas-

Table 13: E2 residual composition failure on jointly stable facts. Qwen3’s 0% on this shallow set motivated the harder short-chain and long-chain D4v2 benchmarks.

Model	Atoms	Ex. stable	Cons. stable	Gate	Cons. fail
Qwen2.5-7B	1 400	162/350	194/350	360	39/360 = 10.8%
Mistral-7B	1 400	154/350	202/350	360	60/360 = 16.7%
DeepHermes-3-Llama-3-8B	1 400	168/350	216/350	360	59/360 = 16.4%
Qwen3-8B	1 400	181/350	220/350	360	0/360 = 0.0%

sification chains) alongside temporal reasoning. It contains 411 atomic facts and 188 composition cases, yielding 1 644 atomic rows and 623 composition rows per model. Because the underlying candidate set was generated by Gemini 2.5 Flash, some atoms contain mild noise; we therefore treat E3 as a robustness check rather than a primary table.

Table 14: E3 residual composition failure on cross-domain (scientific + temporal) facts. Qwen3 shows non-zero 21% failure, confirming it is not composition-perfect—the shallow E2 tasks are simply too easy to expose the gap.

Model	Ex. stable	Cons. stable	Gate	Cons. fail
Qwen2.5-7B	180/411	287/411	98	26/98 = 26.5%
Mistral-7B	114/411	251/411	77	25/77 = 32.5%
DeepHermes-3-8B	132/411	278/411	88	21/88 = 23.9%
Qwen3-8B	151/411	293/411	100	21/100 = 21.0%

N Harder-set: cross-model pressure test

The harder set was designed after observing Qwen3’s 0% failure on E2. It contains 160 short-chain temporal-binding and multi-fact ranking cases (340 atomic rows, 751 composition rows per model). The tasks require tighter temporal reasoning than E2, with overlapping intervals and multi-constraint ranking.

Table 15: Harder-set residual failure across six models, four evaluation tiers. “Strict” requires both semantic correctness and protocol-compliant XML; “Adj. residual” adds manual adjudication of format ambiguities. Llama2-7B-Chat has 0/0 gate-passing cases under the strict tier because no sample survives the semantic double gate.

Model	Raw	Semantic	Strict	Adj. res.
Qwen3-8B	14.3%	2.8%	3.4%	3.4%
Qwen2.5-7B	47.6%	50.4%	51.3%	50.4%
Mistral-7B	100.0%	47.9%	47.9%	47.9%
DeepHermes-3-8B	71.4%	46.7%	46.7%	40.2%
Llama2-7B-Chat	—	—	0/0	0/0
Llama2-13B-Chat	76.0%	76.0%	0/0	0/0

Under the adjudicated-residual tier (our most

conservative main-text analogue), Qwen3 still shows a non-zero 3.4% failure on 88 gate-passing cases, while Qwen2.5, Mistral, and DeepHermes sit at 40–50%. The gap between Qwen3 and the other post-trained checkpoints replicates at short depth on harder structure. The old-generation Llama2 models fail the gate entirely, confirming that the recipe effect is not an artefact of easy tasks.

O LLM adjudication validation against human judgment

All main-text composition results use Gemini 2.5 Flash to determine whether a model answer matches the gold answer under the consistency tier. We validate this automatic adjudicator against three members of the author team, who annotated a stratified sample of 225 temporal D4V2 decisions independently.

Sample design. We drew 225 cases from the full D4V2 evaluation pool, stratified across three models (Qwen2.5-7B-Instruct, DeepHermes-3-8B, Mistral-7B-Instruct; 75 cases each), balanced across Gemini-correct and Gemini-incorrect predictions with deliberate over-sampling of disagreement-prone cases (date-notation differences, partial name matches, and INSUFFICIENT_EVIDENCE responses).

Annotation protocol. The three author annotators independently judged all 225 cases while blind to Gemini’s judgment and to one another. Each annotator saw the gold answer and the model’s full generated text; the original question was not used for the match decision. Annotators judged only whether the extracted final answer semantically matched the gold answer and ignored the reasoning text. The consistency-tier guidelines accept BC/BCE notation variants, unambiguous partial name matches, reasonable answer paraphrases, and a one-year tolerance for calendar-boundary ambiguity; they reject factual contradictions and instruct annotators to mark uncertain cases incorrect. The released audit table restores each question by case identifier for inspection, without changing the original labels.

Inter-annotator agreement. The three annotators achieved Fleiss’ $\kappa = 0.875$ (near-perfect agreement), with 90.7% unanimous decisions (204/225) and 9.3% 2-1 splits (21/225). Pairwise

Cohen’s κ ranged from 0.812 to 0.911, reflecting natural individual variation on borderline cases. Majority vote across the three annotators was used to produce gold labels for evaluating Gemini.

Gemini vs. human gold. Tables 16 and 17 report per-model metrics and the combined confusion matrix.

Table 16: Human vs. Gemini adjudication on 225 temporal D4V2 cases (3 author annotators working independently, majority-vote gold labels, Fleiss’ $\kappa = 0.875$). Gemini achieves 94.6% precision and 100% recall when the correct-answer label is the positive class.

Model	n	Agree.	κ	Prec.	Rec.
Qwen2.5-7B-Inst	75	100.0%	1.000	1.000	1.000
DeepHermes-3-8B	75	100.0%	1.000	1.000	1.000
Mistral-7B-Inst	75	92.0%	0.839	0.838	1.000
Combined (temporal)	225	97.3%	0.947	0.946	1.000

Table 17: Gemini judgments against three-annotator majority vote. A positive label means that the model answer is correct.

	Gemini +	Gemini −
Human correct	106	0
Human incorrect	6	113

Interpretation. Against majority-vote human labels, Gemini achieves 94.6% precision and 100% recall. No majority-human-correct answer is marked incorrect, so the adjudicator creates no spurious residual failures in this sample. Its six errors are false positives for correctness, concentrated in Mistral-7B-Inst date-precision edge cases (e.g., “early 17th century” vs. “1608”). These errors slightly lower measured residual failure and therefore cannot explain the observed failure gaps.

The 75 Qwen3 cases (excluded from the primary analysis above) contain a mix of D4V2 temporal questions and synthetic fictional-entity questions from the KINSHIP and SPATIAL families. On the synthetic subset, the model correctly answers INSUFFICIENT_EVIDENCE (it cannot know fictional facts), but Gemini marks this as incorrect because the gold answer is a fictional entity name. This yields 46 of 50 synthetic cases where Gemini says False but the human says True—a systematic artifact of applying the consistency

matcher to questions whose gold answers refer to entities outside the model’s training data. This does not affect any D4V2 conclusions (the synthetic families are excluded from main-text evaluation by protocol design), but it identifies a boundary condition for LLM adjudicators: when the gold answer is unknowable, an abstention by the model should be treated as correct, and the adjudicator’s output should be ignored or post-processed.

P Failure taxonomy (adjudicated sample)

We took all 251 double-gate failure candidates across E2 and E3 (39 Qwen2.5 + 60 Mistral + 59 DeepHermes + 26 E3-Qwen2.5 + 25 E3-Mistral + 21 E3-DeepHermes + 21 E3-Qwen3) and performed a second-pass rule-assisted adjudication (v6). Each failure was classified into one of the categories defined below.

Category definitions.

Temporal order / boundary confusion. The model retrieves individual dates correctly but misorders events chronologically or misidentifies interval boundaries. This is the signature failure mode of compositional temporal reasoning: the atoms are known but their relative ordering is wrong.

Intermediate variable loss. The model correctly retrieves boundary dates early in the reasoning chain but loses or swaps one of them in later reasoning steps, leading to an incorrect final comparison.

Binding shift. The model correctly retrieves facts about entity A and entity B but assigns a property of A to B (or vice versa) during composition.

Answer type confusion. The model produces a factually relevant answer of the wrong type (e.g., names a person when asked for an event, or lists events when asked for a single date).

Over-abstention. The model answers INSUFFICIENT_EVIDENCE to the composition question despite correctly answering every sub-question in isolation. This reflects a calibration failure: the model possesses the

knowledge but retreats to uncertainty when asked to compose.

Evaluation format issue. The model’s reasoning chain is substantially correct (all atomic facts are identified and the logical structure is valid), but the final answer is malformed (XML tag errors, answer-type mismatch, unparseable format) or mis-scored by the adjudicator.

Ambiguity audit flags. We separately flag temporal overlap and question ambiguity when a strict ordering is contestable or multiple answers are defensible. These are benchmark-design diagnostics, not structural failure categories, and are not used as illustrative evidence that a model failed to compose known facts.

Table 18: Primary-label taxonomy over 251 double-gate failure candidates. “Structural residual” denotes cases assigned a composition-error label; “Scoring/format artefact” denotes cases where the model’s reasoning is substantially correct but the final answer is malformed or mis-scored.

Label	Count
Temporal order / boundary confusion	88
Evaluation format issue	70
Answer type confusion	31
Over-abstention	27
Intermediate variable loss	25
Binding shift	10
Structural residual	181
Scoring or format artefact	70

Of the 251 adjudicated failure candidates, 181 (72.1%) receive a structural composition label—the model answers the constituent facts but fails to assemble them correctly. Temporal order and boundary confusion (88/181, 48.6%) is the dominant structural class: the model retrieves individual dates correctly but misorders them or misplaces an event relative to interval boundaries. This is precisely the failure mode our composition protocol is designed to isolate. A further 10 cases exhibit binding shifts between entities, and 25 lose intermediate variables mid-reasoning. The 70 format-artefact cases (27.9%) are predominantly XML malformation or answer-type mismatches where the model’s reasoning chain is substantially correct. These cases comprise 27.9% of candidate failures; we report the primary metric without subtracting them because the

taxonomy is a post-hoc diagnostic rather than a replacement scoring rule.

Illustrative failure cases. Table 19 provides concrete examples of the major structural modes. Ambiguous or overlapping questions are excluded from this illustrative evidence.

Table 19: Illustrative structural failure cases (consistency tier). All cases pass both gates; ambiguity-flagged cases are excluded.

Mode	Example	M.
Temporal order confusion	Q: Rank from earliest to latest: (A) the invention of the printing press, (B) the fall of Constantinople, (C) the discovery of penicillin, (D) the end of WWI.	Q1.
M: “B, A, D, C” G: B, A, C, D (C and D swapped)		
Interval boundary confusion	Q: Which event occurred between the signing of the Magna Carta (1215) and the discovery of penicillin (1928)?	DH.
M: “French Revolution (1789)” G: End of WWI (1918)		
Intermediate variable loss	Q: Which event occurred between the fall of the Roman Empire (476) and the moon landing (1969)?	Q3.
M: “WWI (1914)” G: Discovery of penicillin (1928)		
Over-abstention	Q: What came first, the discovery of penicillin or the end of WWI?	Q1.
M: INSUFFICIENT_EVIDENCE G: End of WWI		
Binding shift	Q: Which was invented first, the telephone or the light bulb?	Q1.
M: “The telephone was invented in 1876; the light bulb in 1879. Therefore: the light bulb.” G: Telephone		

Excluded ambiguity example. The Franco-Prussian War example illustrates a *temporal-overlap* difficulty: the war concluded in the same year as German unification (1871), making the BEFORE/AFTER classification ambiguous without month-level precision. The model’s answer is historically defensible (the war largely preceded unification), but the gold answer treats unification as a strict temporal boundary. This category highlights a limitation of purely ordinal temporal benchmarks: many real-world events overlap, and forcing a strict-before/after judgment can produce contestable gold labels. We report it as a benchmark-design warning, not as evidence for a structural composition failure.

Similarly, several questions in the benchmark admit multiple valid answers (e.g., “What happened after Sputnik 1 but before the moon landing?” encompasses many historical events). When the model produces a factually correct answer that differs from the gold, the adjudicator marks it incorrect. We flag these cases as *question-ambiguity* diagnostics (distinct from structural composition errors) and recommend that future benchmark construction use exhaustively enumerated answer sets for open-ended temporal intervals.

Structural examples are drawn from the consistency-tier adjudicated sample across E2 and E3. Model abbreviations: QI = Qwen2.5-7B-Instruct, DH = DeepHermes-3-8B, Q3 = Qwen3-8B.

Q Synthetic task families: design and gate behaviour

Beyond the temporal-reasoning families used in D4v2, our full evaluation suite includes three synthetic task families designed to measure *pure* compositional reasoning when all required facts are supplied in the prompt (in-context protocol):

KINSHIP. Facts describe fictional family relations (e.g., “Alice is the mother of Bob”). Composition requires transitive closure over multiple relations.

NUMERICAL. Facts assign numeric values to fictional entities. Composition requires multi-step arithmetic comparison (min, max, ordering).

SPATIAL. Facts place fictional objects at named locations. Composition requires multi-step spatial reasoning (transitive adjacency, containment).

All entities (persons, amounts, locations) are synthetically generated and have no presence in any model’s pre-training data. This is by design: the synthetic families target a different research question (how models chain externally provided facts) from D4v2 (how models compose internal knowledge).

Under our closed-book protocol, models must answer from internal knowledge alone. Every evaluated model correctly abstains on the synthetic families—atomic accuracy is 0% for all models because no model can possess internal knowledge of entities absent from its training data. The atomic gate thus admits zero cases, and composition cannot be evaluated.

In-context evaluation. To assess how the same model handles composition when facts *are* explicitly provided, we ran Qwen2.5-7B-Instruct on the synthetic families using the evidence prompt style (facts supplied as [Evidence N] blocks in the prompt). The results reveal a sharp domain split not visible under the closed-book protocol (Table 20):

Table 20: In-context synthetic composition: Qwen2.5-7B-Instruct with facts supplied in the prompt. “Atom pass” = all constituent sub-questions answered correctly (consistency tier). “Residual failure” = main-question error given all atoms correct.

Task family	<i>n</i>	Main acc.	Atom pass	Resid. fail
KINSHIP (matrilineal)	33	72.7%	8	5/8 = 62.5%
KINSHIP (patrilineal)	33	78.8%	9	6/9 = 66.7%
NUMERICAL (offset chain)	33	0.0%	1	1/1
SPATIAL (eastward)	33	97.0%	12	11/12 = 91.7%
SPATIAL (northward)	33	93.9%	6	5/6 = 83.3%
Total	165	68.5%	36	32/36 = 88.9%

Two patterns stand out. First, *difficulty varies sharply by domain*: spatial transitive-reasoning chains are near-ceiling (94–97%), kinship chains are moderate (73–79%), and numerical offset chains yield 0% accuracy—the model cannot perform even two-step arithmetic over provided numeric facts. Second, *atomic extraction is already a severe bottleneck*: only 36 of 165 cases (21.8%) survive the sub-question gate even though all facts are in the prompt. On those 36 gate-passing cases, residual composition failure is 88.9%, indicating that even when Qwen2.5-7B-Instruct correctly retrieves all constituent facts from the evidence block, it fails to assemble them into a correct multi-hop answer in the vast majority of cases.

These in-context results complement the closed-book D4v2 findings. The closed-book protocol measures composition over internal knowledge and excludes synthetic entities that the model cannot know. The in-context variant supplies every fact, then uses the gate to verify that the model can extract those facts before measuring composition. The same double-gate logic therefore applies in both regimes, although atomic access comes from internal knowledge for D4v2 and from prompted evidence for the synthetic families.

R Prompt-format robustness

To assess whether residual composition failure depends sensitively on the prompt template, we evaluated Qwen2.5-7B-Instruct on the same 122 atomic-gate-passing D4v2 composition cases under three prompt variants, using the same model, inference parameters, and adjudication pipeline:

V1 (XML + reasoning). Our standard protocol: the model is instructed to output <reasoning> and <answer> tags (prompt style question_only).

V2 (XML answer-only). The model is instructed to output only an `<answer>` tag, with no reasoning section (prompt style `question_only_no_reasoning`).

V3 (plain text). No XML formatting instruction; the prompt asks “Question: ... Answer the question using only your internal knowledge” (prompt style `question_only_plain`).

Table 21: Prompt-format robustness on 122 gate-passing D4V2 composition cases (Qwen2.5-7B-Instruct). V1 is our standard protocol. “Accuracy” is consistency-tier correctness.

Variant	Format	Accuracy	Abstention
V1 (XML + reasoning)	100%	26.2%	3.3%
V2 (XML answer-only)	100%	19.7%	25.4%
V3 (plain text)	0%	15.6%	45.9%

Three findings are relevant to the interpretation of our main results.

First, *format compliance is near-perfect* with explicit XML instructions (100% of V1 and V2 completions contain the closing `</answer>` tag) but drops to 0% when no XML format is requested (V3). Since our adjudication pipeline parses `<answer>` tags for answer extraction, V3 answers were extracted via a last-line heuristic; this introduces additional measurement noise in the plain-text condition.

Second, *removing the reasoning instruction sharply increases abstention*. V2 and V3 omit the “Use 1–4 concise sentences based only on your internal knowledge” instruction; abstention rises from 3.3% (V1) to 25.4% (V2) to 45.9% (V3). The model retreats to INSUFFICIENT_EVIDENCE when not explicitly directed to reason, even though these 122 cases demonstrably pass the atomic gate (the model knows the constituent facts when asked in isolation).

Third, *cross-variant answer agreement is low*. Only 24 of 122 questions (19.7%) receive identical final answers across V1, V2, and V3. Pairwise agreement is 32.0% (V1 vs. V2), 23.8% (V1 vs. V3), and 58.2% (V2 vs. V3). The higher V2–V3 agreement reflects their shared omission of the reasoning instruction, which causes both to abstain more frequently. The low V1–V2 agreement (32%) shows that even a seemingly minor change—removing a 1–4 sentence reasoning

prompt—shifts the model’s final answer on the majority of questions.

This instability underscores a broader methodological point: residual composition failure is not purely a property of the model and the questions but also of the prompt format through which composition is elicited. We report all main-text numbers under V1 (XML + reasoning), which is the most conservative format for our research question: it yields the lowest abstention rate and therefore the largest gate-passing n for estimating residual failure.

S Annotation guidelines

All human validation in this paper follows a single annotation protocol. We document the complete guidelines here for reproducibility.

Task definition. The annotation task is a binary decision: given a gold answer and a model’s full generated output, does the model’s extracted final answer semantically match the gold answer? Annotators saw the gold answer (a short fact, date, event name, or entity) and the complete generation, including any `<reasoning>` and `<answer>` XML blocks. They judged only the extracted final answer and ignored the reasoning text. The original question was not used for this match decision; the released audit table restores it by case identifier for inspection.

Consistency-tier matching rules. The following rules define what constitutes a semantic match:

- Exact match after normalisation.** Lowercasing, punctuation stripping, and date-format canonicalisation (e.g., “1876”, “1876 AD”, and “the year 1876” all reduce to 1876). If the normalised forms are identical, the answer matches.
- BC/BCE equivalence.** “200 BC” and “200 BCE” are treated as identical.
- Partial name matches.** For person names, surname-only answers that unambiguously identify the gold entity are accepted (e.g., “Einstein” matches “Albert Einstein”). For event names, standard short forms are accepted (e.g., “WWI” matches “World War I”).

4. **Reasonable paraphrases.** If the model’s answer expresses the same factual content as the gold answer in different words, it is accepted. This applies to *answer surface forms*, not to question forms. Example: “the year 1876” matches “1876” when the gold is “1876”; similarly, “1876 AD” and “the year 1876” are accepted surface variants of the same date answer. The annotation guidelines do *not* classify question forms (e.g., “In what year was the telephone invented?”) as paraphrases of answers—these are distinct probe types used in the atomic stability gate (Gate 1), where the model is asked to answer the same underlying fact through different question surface forms.
5. **Numerical tolerance.** For year answers, ± 1 year is accepted to account for calendar-era boundary ambiguity.
6. **Abstention handling.** INSUFFICIENT_EVIDENCE is treated as a valid model answer, not as a missing answer. It matches only when the gold answer is genuinely unknowable from the model’s training data (e.g., synthetic entities). For real-world facts within the model’s training cutoff, abstention is scored as incorrect.

Rejection rules. The following are scored as incorrect regardless of the model’s reasoning quality:

1. **Factual contradiction.** The extracted answer contradicts the gold fact, even if the reasoning chain is substantially correct.
2. **Wrong entity.** The answer names a different event, person, or date than the gold (e.g., “Franco-Prussian War” instead of “Spanish-American War”).
3. **Incomplete answer.** The answer omits a required component (e.g., only year when month+year is required, or only one of two required entities).
4. **Format violation.** The model’s output contains no extractable answer (no `<answer>` tag or equivalent). In the XML protocol, $\geq 99\%$ of completions emit the closing tag; the rare format-violation cases are scored as incorrect.

Annotation procedure. The three annotators judged each case independently while blind to the automatic adjudicator’s decision and to one another. We then used majority vote as the human reference label. The automatic adjudicator was counted as correct only when its binary decision matched this majority label; the released per-annotator files preserve the original labels and notes.

T Chain-of-thought baseline

To test whether explicit step-by-step reasoning instructions reduce residual composition failure, we evaluated Qwen2.5-7B-Instruct on the full D4V2 benchmark with a modified system prompt that instructs the model to think through each question step by step. The system prompt was: “You are a careful research assistant. No external evidence is provided. Answer from your internal knowledge only. Think through each question step by step: first identify each relevant fact, then reason about their relationships, and finally combine them. If you are genuinely uncertain, answer exactly INSUFFICIENT_EVIDENCE.”

Inference used vLLM 0.19 with the model’s native chat template, greedy decoding ($T=0$), and `max_new_tokens = 1024` (doubled from the standard 512 to accommodate reasoning chains). All other inference parameters and the adjudication pipeline were identical to the main evaluation (Appendix A). The format instruction required `<reasoning>` and `<answer>` XML tags, with stop-string `</answer>`.

Under CoT, atomic stability (consistency gate) is 79.7% (63/79 facts stable), compared to 89.9% for standard Qwen2.5-7B-Instruct. Despite the 10 pp drop in atomic stability, residual composition failure falls to 47.1% overall (32/68 double-gate-passing cases). By depth:

Table 22: CoT vs. standard Qwen2.5-7B-Instruct on D4V2 by depth. n is the number of double-gate-passing cases.

	$d=2$	$d=4$	$d=6$	$d=8$
Standard (residual failure %)	12.0 (25)	69.8 (53)	78.9 (38)	81.8 (33)
CoT (residual failure %)	12.5 (16)	42.3 (26)	73.3 (15)	71.4 (7)

At depths 4 and 8, CoT reduces residual failure by 28 pp and 10 pp, respectively, despite smaller gate-passing n . Atomic stability drops while composition improves, showing that Δ_{atom} and Δ_{comp} can move in opposite directions when only

the generation process changes.

U Artifact License and Availability

The D4v2 benchmark, evaluation data (case-level records in `results/processed/`), and annotation guidelines are released under the Creative Commons Attribution 4.0 International license (CC BY 4.0). The model-generation and analysis scripts are released under the MIT License. All artifacts are available at the supplementary material accompanying this submission. The models evaluated are publicly available open-weights models (Qwen3, Qwen2.5, Mistral, Llama 3) and community-trained variants (DeepHermes-3); no proprietary model access is required to reproduce the results.