Learning the Map, Not the Route: Curriculum, Capacity, and Transfer in Tool-Navigating Language Models

Anonymous authors
Paper under double-blind review

Abstract

We compare what language models learn from knowledge-graph navigation traces, which query a graph with tools and end in a machine-verifiable evidence path, and from free-form chain-of-thought distilled from a teacher (free-CoT). In a pre-registered study of 0.6B to 14B models on a synthetic, SHACL-governed kinship benchmark with strict format isolation, facts are withheld from the prompt and every edge in an emitted evidence path must have been retrieved during the episode, so a recited path scores zero. Each finding is tied to a registered criterion or labelled post hoc. (1) On the original corpus, navigation traces lose to free-CoT by 21.0 points at 12B. The decisive split inverts an ordering direction the trace compiler had held constant, and the navigation model's failures recite that constant: it learned the route, not the map. (2) When both corpora are regenerated with that direction varied, navigation leads by 23.7 points (Gemma-12B; seed-level interval +5.2 to +42.3), and the result replicates at Qwen3-14B under a rule registered before training (+13.9; +5.1 to +22.8). Varying it improves navigation by 25 to 29 points in every family tested; free-CoT drops in Gemma but not in Qwen3-14B. (3) Training on prose-only reasoning traces suppresses tool use (tool-call rates of 1.3% to 58.6% at 0.6B to 4B). Adding the tool calls back to the same prose restores tool use to 100%, and the gain from varying the direction holds for prose traces too (+23.8). At 4B the compiled notation adds 6.1 points on the original curriculum and 9.1 on the repaired one, a smaller effect that comes mostly from how the evidence path is written and does not carry over to transfer. (4) On a bijectively relabeled second domain, every model evaluated without in-context examples scores close to zero on the verified metric, for opposite reasons at the two sizes: the 4B model keeps the notation but loses the task, and the 12B model does the opposite. Two in-context examples fully restore the notation, and the navigation-trained 12B model then produces verified evidence on 65.1 ± 5.9% of the new domain, 13.8 points above the untrained base model given the same examples (below the registered 15-point margin; the seed-level interval includes zero) and 5.8 points below the free-CoT model. The ability transfers, but no better than after free-CoT training. We report the null result with its repair, seed-level uncertainty next to every item-level interval, and three comparisons that would have favored our hypothesis but that our registered gates refused.

1 Introduction

The project is named for the structure that gives cognitive science its term for a cognitive map (O'Keefe & Nadel, 1978): the representation an animal navigates by, as opposed to the routes it has taken. We use three terms. The map is a task's relational structure—entities, relations, constraints, transitions, selection rules, termination conditions. A route is one policy instantiated over the map. A generator can emit thousands of superficially different trajectories and still teach a single route if some decision is held constant across all of them. The legend is the lexical layer through which the map is read and written: relation names and the forms in which they are expressed. Our transfer benchmark preserves the map

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and bijectively changes the legend (biologicalParentOf→directManagerOf), so what we test there is lexical relabeling of relational structure, not structural transfer.

Navigation supervision—training traces that walk a knowledge graph with tools (Thought:/Action:/Observation:)—promises to teach a task and to teach navigating as a skill that survives a change of domain. We test both under pre-registration, against the strongest comparator we could build (free-form chain-of-thought distilled from a teacher that saw the gold answer), on an instrument in which every claimed evidence edge must have been retrieved in the episode. Each answer reversed at least once, and the reversals are the paper:

1. The first promise initially failed (§4.1): compiled navigation traces lost to free-CoT distillation by 21.0 points at 12B, published as a null under the registered NO-GO clause.
2. The failure was curricular and diagnosable per template (§4.2). On question types seen in training the two formats tie (75.8 vs 76.8); the decisive split inverts the ordering direction the compiler had kept constant, and the navigator's failures recite it. Varying it in 788 training items moves navigation +24.9 and free-CoT −19.8 on every template; the matched-curriculum contrast is +23.7, and it replicates at Qwen3-14B under a pre-registered rule (+13.9), where the navigation-side gain is +29.1 and 78% of the frozen model's failures recite the constant.
3. The free-CoT collapse is real, uniform, not a recitation of the wrong rule—and family-specific (§4.3): at Qwen3-14B the same swap helps free-CoT by 6.0.
4. What navigation supervision provides is participation, curriculum, and a secondary notation margin (§4.5): prose-only traces train the tools away below 14B; restoring them restores participation completely; the curriculum gain is format-independent; the notation margin (+6.1 frozen, +9.1 repaired) sits in serialization and vanishes in transfer.
5. Transfer at 12B: the competence moves, the legend does not, and navigation training moves it no better than free-CoT does (§4.6): capacity inverts which half of training survives; two exemplars repair the legend; at three seeds the exemplar-matched gain over the untrained base is +13.8 (registered: +15) and free-CoT-trained models transfer at least as well.

Thresholds, falsifiers and decision rules were committed in dated addenda before the data they govern (Appendix A). Undaggered numbers are registered metrics; numbers marked † are post-hoc or labelled sensitivity runs and never substitute for a registered metric.

2 Related work

Agent fine-tuning. ReAct-style trajectory tuning (Yao et al., 2023; Chen et al., 2023; Zeng et al., 2024), self-supervised API-call insertion (Schick et al., 2023), API instruction tuning (Patil et al., 2024; Qin et al., 2024) and browse-and-cite agents (Nakano et al., 2021) all teach tool use from demonstrations. Closest to our decomposition is Agent-FLAN (Chen et al., 2024), which separates format-following from reasoning content and shows that entangling them hurts; our B′ condition is the complementary manipulation—hold the reasoning fixed, add or remove the tool protocol—and isolates participation as the dominant factor at small scale. FireAct reports that trajectory diversity helps and that the best mixture depends on the base model; we measure one specific kind of diversity per template and show it interacts with format in sign. AgentTuning preserves general ability by mixing in instruction data; §4.5 names one mechanism mixture must guard against: training text that omits the protocol changes whether the model uses tools at all, invisibly to accuracy. LLMs navigating knowledge graphs. Think-on-Graph (Sun et al., 2024), Reasoning-on-Graphs (Luo et al., 2024) and Graph-CoT (Jin et al., 2024) use an LLM as an interactive explorer over a KG at inference; our task is theirs in miniature and under control, and the question we add is what training on such traces buys, at what scale, and what of it survives a change of legend; KG-Agent (Jiang et al., 2024) synthesizes tool-use instructions from reasoning programs over KGs — our contribution is orthogonal: an evaluation contract and controlled recipe comparisons rather than an agent framework. Cognitive maps and shortcuts. CogEval (Momennejad et al., 2023) asks whether LLMs hold cognitive maps at all; we ask the narrower,

trainable question of when supervision installs a map versus a route. "Decision exposure" (§4.2) is a curriculum-level statement of shortcut learning (Geirhos et al., 2020); the compositional held-out cell follows CLUTRR's design (Sinha et al., 2019); that transformers compose learned relations only under specific regimes (Wang et al., 2024) and dissociate form from content (Lake & Baroni, 2018) is consistent with what we measure. Distillation, forgetting, benchmarks, pre-registration. Condition B instantiates chain-of-thought distillation (Ho et al., 2023; Magister et al., 2023) with an oracle gate in place of self-consistency filtering (Zelikman et al., 2022); our enactment check verifies steps against tool observations rather than rewarding them (Lightman et al., 2024). Suppression of a protocol behavior by ordinary SFT is a behavioral cousin of catastrophic interference (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017; Luo et al., 2025). bAbI (Weston et al., 2015) and CLUTRR are the synthetic ancestors; KGQA datasets (Yih et al., 2016; Zhang et al., 2018) share the multi-hop structure without the governed schema. Our dated-addenda protocol follows the pre-registration workshop (ber, 2021); the operational addition is registered gates executed on every arm with contrast refusal. Prior work on structure-guided small-model agents (Volpini & Carrozza, 2026), inference-time KG navigation (Volpini & Raad, 2026) and linked data as agent memory (Volpini et al., 2026) motivated the study.

3 Benchmark, conditions and instrument

Kinship graph. A synthetic KG of 43,390 triples over 3,980 entities, governed by a SHACL shapes file that is the single source of truth for gold generation, path scoring and distractor authoring. Frozen eval: test_entity (1,099 items; question types seen in training) and the decisive test_structural split (4,437 items: 3,328 answerable, 1,109 gold abstentions). Training corpora (4,246 items, 1,061 abstentions) cover four templates, every one ordering by eldest; the decisive split consists entirely of the youngest-direction counterparts (youngest_living_bio_child, 1,951; …_grandchild, 1,106; youngest_bio_grandparent, 1,380)—same graph, same relations, selection rule inverted. This near-transfer, compositional design was registered as the decisive population before any training.

The bidirectional curriculum (Experiment Y). 788 of the 1,522 eldest_living_bio_child training items are re-minted as youngest_living_bio_child over the same start entities, for both formats: the compiler re-renders the traces; the teacher regenerates the prose (788/788 accepted by the oracle gate). Corpus size, item set and abstention count are unchanged; the other 3,458 items are byte-identical. youngest_living_bio_child becomes seen; …_grandchild ("youngest" seen only on children, "grandchild" only with "eldest") and youngest_bio_grandparent stay held out.

Transfer benchmark (Experiment X). A bijective relabeling Φ of a freshly generated graph into a corporate-organization domain (birthDate→hireDate, "youngest living grandchild" → "most recently hired active second-line report"), applied uniformly to graph, shapes, questions, prompts and exemplars. Its generator emits canonical phrasings only, so the two benchmarks are isomorphic in task structure but not surface difficulty (§4.6).

Conditions. All conditions train on identical items with identical treatment (LoRA r16, 3 epochs, batch 4×4; observation tokens masked; supervised tokens equalized ±2%) and differ only in the text of the records. B: question + Facts: block, teacher-distilled free-CoT prose (Gemini-2.5-flash, T=1.0, oracle-gated, teacher saw the gold) + terminal block. D: bare question, ontology-compiled Thought:/Action: navigation + terminal. B′: bare question, teacher prose thoughts + D's tool calls, observations and terminal blocks byte-identical. F (D with Facts:) is in §4.1; A, C, E ran only in the in-context pilot (Appendix B). Terminal contract: Evidence path: e → r → … then Answer:, or NONE/ABSTAIN.

Protocol and metrics. The facts-withheld TOOL protocol: no facts in the prompt; get_neighbors/get_attributes, ≤ 8 turns; thinking modes pinned off; one deterministic renderer per family. Arms whose training did not contain the tool grammar (BASE, B) receive two pinned exemplars; D, F, B′ do not—an asymmetry that favors BASE and B in every comparison and is quantified in §4.5. Enacted conjunction on gold-answerable items = path_valid ∧ grounding_ok ∧ conforms ∧ enacted: the chain parses, exists in the graph and ends at the emitted Answer:; the SHACL grounding gate and full template shape hold,

including the ordering constraint checked against graph dates (so a conforming path ends at a semantically correct entity); and every edge appeared in a tool observation of the same episode (a recited-but-correct path scores zero; measured recitation is 0.0% in every arm). Scope of the certificate: for the child and grandchild templates the shape pins the globally correct selection (conjunction with a non-gold answer: 0.0%); the grandparent shape validates the exhibited chain without pinning the global optimum (up to 52.4% of one arm's grandparent items conform yet miss gold — a quantified frozen-scorer gap), so conjunction is verified evidence-path success, not by itself an end-to-end certificate. Exact-gold answer accuracy is the strict selection observable, and strict full success = conjunction ∧ exact-gold preserves every headline direction (Gemma bidir D−B +24.3; Qwen3-14B +11.1; Appendix K). The records' answer_correct field measures candidate-pool membership (the exporter emits the full valid set; gold is its ordered first element) — a leniency discovered under review and logged as bug #8. Every answer observable (pool-member, exact-gold, strict full success) uses the identical denominator as conjunction: the same gold-answerable items, records and runs, with nothing filtered by model behavior. Abstention recall on gold-abstention items; the two domains are always partitioned on the gold label and never combined (Appendix C). Participation floor: an arm is valid only if its pooled tool-call rate on answerable items is $\geq 90\%$; the analysis code refuses any contrast whose either side fails. Uncertainty is reported two ways: the registered paired bootstrap over items (each item's 0/1 outcome averaged over the three seeds' checkpoints, items resampled with replacement, 10,000 draws, fixed seed; checkpoints held fixed, so the interval is conditional item-sampling uncertainty (Bouthillier et al., 2021)) and, because it does not capture between-seed variance (Reimers & Gurevych, 2017), a Welch t interval over three seed means beside it; ± is the SD across seeds (per-seed values in Appendix D).

4 Results

4.1 The pre-registered null and its diagnosis (C2)

Given a sufficiency-gated fact block, an untrained 12B model clears 90% on the decisive split (Appendix B); hence facts are withheld. Gemma-4-12B, 3 seeds, all floors pass:

Arm	conjunction ↑	abstention ↑	tool-call
BASE-nav (untrained + exemplars)	48.2	10.5	100%
B-nav (free-CoT)	78.4 ± 5.5	99.0 ± 0.5	97.3%
D-nav (navigation)	57.4 ± 6.1	37.2 ± 13.8	98.6%
F-nav	56.1 ± 7.2	80.6 ± 7.7	95.6%
RAG-linked / RAG-BM25 (untrained, open-book)	58.1 / 53.3	96.5 / 96.3	—

$D - B = -21.0\,pp$ (item CI (-22.3, -19.7); Welch (-36.1, -5.9)). The registered rule fails and inverts; the null is published. Two facts locate it. On test_entity the formats tie (D 75.8, B 76.8). And 30.6% of D's failed answerable items contain "eldest"/"oldest" and 38.1% "ascending"—the phrases of its training traces—against 2.5% "youngest": the model executes the trained route against a question that inverted it.

4.2 Decision exposure: one swap, opposite signs on every template (C3)

Both corpora were regenerated with the ordering rule varied in 788 items (§3). Table 1 and Fig. 1a give the effect per template; the interaction is the central number — one intervention moves the two recipes 44.7 points apart.

The navigator's largest gain is in the held-out compositional cell (+31.6), where "youngest" and "grandchild" were never trained together; the free-CoT arm loses on every template, including the one it now sees; the sign of D−B flips on every template. Curriculum-matched contrast: $D - B = +23.7\,pp$ (item CI (22.6, 24.9); Welch (5.2, 42.3)). The repaired split mixes 1,651 now-seen child items with 1,677 still-held-out items; restricted to the held-out templates the contrast is +24.5 (item CI (22.7, 26.3)) and the D-side repair +27.4 (25.6, 29.2) — larger than on the full split (Appendix K). Abstention moves the other way (D-bidir 51.0 ± 7.0 vs B-bidir 98.6 ± 1.0) and is reported wherever the margin appears.

Table 1: Per-template enacted conjunction, Gemma-12B, 3 seeds (± seed SD; Δ: paired item bootstrap, 95% CI). Registered falsifier (gain only in the now-seen cell) did not fire.

Template (n)	under bidir	B frz	B bidir	ΔB	D frz	D bidir	ΔD	D−B frz	D−B bidir
child (1,651)	now seen	86.6	75.1	−11.5 (-13.0, -10.0)	75.7	98.1	+22.4 (21.0, 23.8)	−10.9	+23.0
grandchild (681)	held out, comp.	54.8	32.6	−22.2 (-24.7, -19.8)	32.7	64.3	+31.6 (28.5, 34.6)	−22.1	+31.7
grandparent (996)	held out	81.1	49.1	−32.0 (-34.1, -29.8)	44.1	68.7	+24.6 (22.4, 26.8)	−37.0	+19.6
all (3,328)		78.4	58.6	−19.8 (-21.0, -18.6)	57.4	82.4	+24.9 (23.7, 26.1)	−21.0	+23.7 (22.6, 24.9)

Fig 2 Decision exposure: one curriculum swap, per template and across families (C3, Y2)

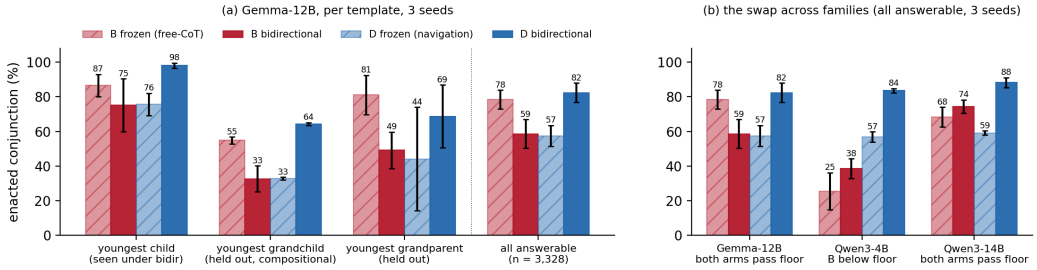


Figure 1: (a) Gemma-12B, per template, 3 seeds. (b) The same swap across families (all answerable items, 3 seeds): the navigation gain is general; the free-CoT collapse is Gemma-specific; at Qwen3-4B the free-CoT arms are below the participation floor.

Decision exposure. A trace can expose the structure of a solution without exposing the decisions that generated it. If a decision variable is held constant throughout the curriculum, the learner has no evidence that it is a variable: it can learn $\pi(\text{state}) \to \text{eldest}$ where the task requires $\pi(\text{goal}, \text{state}) \to \{\text{eldest} \mid \text{youngest}\}$, and the trace is perfectly structured in both cases. Reusable navigation requires exposure of the decisions the model does not already carry: pretraining and instructions can support some decisions without fine-tuning variation, but a decision this fine-tune held constant was learned as a constant. The phrase audit makes this concrete: after the swap, 0.2% of D's residual failures contain the constant (vs 30.6%); at Qwen3-14B the corresponding figures are 78.0% before and 0.0% after. The answer-level audit is sharper: among wrong-in-pool answers on items with ≥ 3 dated candidates, D-frozen selects the eldest — the formerly fixed rule's answer — 59.8%/72.0% of the time at Gemma-12B and 82.4%/85.9% at Qwen3-14B, against 28–35% for B and 0% for D-bidir (Appendix K): the route is executed at the answer, not merely recited. This is shortcut learning (Geirhos et al., 2020) at the level of a generator's design choice.

4.3 What happened to the free-CoT arm

Why would exposing more variety cost a free-CoT arm 20 points? We can rule out the corpus and characterize the behavior; we cannot yet isolate the mechanism. The regenerated prose passed the oracle gate 788/788 with zero retries; the swapped items are the same length as the originals (1,345 vs 1,359 characters) and differ in ordering phrasing only. The collapse is consistent across seeds and uniform across templates. Its anatomy (Appendix E): B-bidir does not pick the wrong person more often once it answers; it stops answering in the contract—unparseable terminations double (10.7% → 21.2% of answerable items), abstention on answerable items rises 2.4× (5.0% → 12.2%), and the arm runs ahead of the tool loop, emitting several Action: lines before any tool returns, in 18.8% of episodes versus 5.3%; answer-line accuracy falls 83.6 → 65.5. B is trained with a complete Facts: block and evaluated without one: its prose was distilled for reading, not for waiting on tools. The registered replication (§4.4) bounds the effect: at Qwen3-14B the same swap moves B by +6.0 (item CI (5.0, 7.0)), so the collapse is Gemma-specific.

Best-of-each and the pooled view. The matched-curriculum contrast is the right comparison for the recipe question: the corpora differ only in training-record text, but the recipes differ

in more than notation (fact availability, tool-protocol exposure, evaluation-time exemplars), so the margin belongs to the navigation-supervision recipe, not to notation alone (§4.5). It is also not the comparison a practitioner would make. Against Gemma's strongest free-CoT arm (B on the original corpus), the repaired navigator is +3.9 pp on answerable items (item CI (3.0, 4.9); Welch (-10.5, 18.4)) and −48.0 pp on abstention; pooled over all 4,437 items (a secondary view), B-frozen is that family's best arm (83.6 ± 4.2 vs D-bidir 74.5 ± 6.0). At Qwen3-14B the ordering differs: the repaired navigator leads on every view (pooled 87.4 ± 3.3 vs B-bidir 79.3 ± 3.3, few-shot base 74.7), because its abstention deficit there is 9 points, not 48. The claim is therefore not "navigation dominates"; it is that with the decision exposed, navigation supervision wins the answerable domain at matched curriculum in both families where the contrast is valid, at one-thirteenth the generation cost (Appendix H), with a family-dependent abstention deficit that is the registered target of a deferred reward-tuning experiment.

4.4 Replication in a second family and at three seeds (Experiment Y2)

The 4B replication had moved D from 56.8 ± 3.0 (3 seeds, frozen) to 83.4 (seed 0, bidirectional), but B-bidir at Qwen3-4B issues tool calls on only 70.1% of answerable items, so no matched-curriculum contrast existed outside Gemma. Y2 was registered before training with three rules: P7, C3 replicates iff both Qwen3-14B bidirectional arms pass the floor and D−B has item CI-low > +5 and a seed-level interval excluding zero; P8, per-arm curriculum effects at 14B reported two-sided; P9, the 4B D-side effect at three seeds (CI-low > +10), the 4B contrast reported only if B-bidir's floor passes.

Table 2: Experiment Y2 (3 seeds each). Floors pooled over seeds.

Cell	conjunction	abstention	tool-call	floor
Qwen3-14B B frozen	68.3 ± 5.7	90.7 ± 4.7	98.4%	pass
Qwen3-14B B bidir	74.3 ± 3.9	94.2 ± 1.6	100.0%	pass
Qwen3-14B D frozen	59.1 ± 1.2	75.8 ± 10.4	98.6%	pass
Qwen3-14B D bidir	88.3 ± 2.8	84.8 ± 11.2	98.1%	pass
Qwen3-14B BASE (few-shot)	75.7	71.9	94.4%	pass
Qwen3-4B B bidir	38.5 ± 5.7	95.6 ± 1.3	73.5%	fail
Qwen3-4B D bidir	83.6 ± 1.2	76.7 ± 8.5	100.0%	pass

P7 passes: D-bidir − B-bidir = +13.9 *pp* (item CI (12.9, 15.0); Welch (5.1, 22.8)); C3 is a two-family result (Fig. 1b). P8: the navigation gain replicates, +29.1 (Welch (21.5, 36.8); Gemma +24.9, Qwen3-4B +26.8); the free-CoT collapse does not, +6.0 (Welch (-6.7, 18.7); Gemma −19.8). "One intervention, opposite signs" is therefore Gemma-specific; what holds everywhere is that exposing the decision lifts navigation by 25–29 points. P9 passes (+26.8, Welch (18.7, 34.9)); the 4B contrast is refused (intent-to-treat, labelled: +45.1). Held-out view: the aggregate +29.1 mixes a +50.5 jump on the now-seen child template with +8.1 (6.4, 9.8) on the held-out templates (grandchild +21.2; grandparent at a 90.5 frozen ceiling), while the matched contrast restricted to held-out templates is +24.2 (22.3, 26.2) — larger than the full-split +13.9, because B-bidir's gains concentrate on the now-seen template (Appendix K). Two consequences: at Qwen3-14B the repaired navigator leads the answerable and pooled views, is within 9 points on abstention, and exceeds the few-shot base by +12.6, which revises the frozen-corpus finding of §4.5 that navigation training "adds nothing to a large model"—what adds nothing at 14B is a curriculum that teaches a route.

4.5 Prose-only traces train the tools away; three contrasts (C4, C5, C7, C9)

A hardening rollout returned exactly 0.0%; the free-CoT arm on Llama-Nemotron-8B had made zero tool calls in 4,437 items at two seeds while BASE (28.7%) and D (99.9%) navigated on the same renderer. Auditing every arm against the registered floor—which, we record as a process failure, had only ever been applied to the confirmatory arms—gives the Qwen3 ladder (frozen corpus, 3 seeds; full table in Appendix F): B's pooled tool-call rate is 1.3% / 16.9% / 58.6% / 84.7% / 98.4% at 0.6B / 1.7B / 4B / 8B / 14B while D's is flat at 86–100%. By the ladder's pre-committed rule (fewer than three informative sizes

⇒ untestable) the capacity-crossover hypothesis is untestable (C4). Three things are established instead. Prose-only supervision suppresses the behavior the evaluation requires, graded with capacity within family (C5†), with base-model disposition dominating across families. Two qualifications: prose-only SFT does not uniformly reduce tool use against the untrained base (at 4B, B's 58.6% exceeds BASE's 35.3%), and the smallest navigation arms also fail the floor (86.3%/86.2% at 0.6B/1.7B); the accurate claim is that prose-only supervision does not reliably install participation where trajectory supervision installs it, and on one family actively suppresses it (0.0% vs the base's 28.7%). At Qwen3-14B, the only ladder size with a valid contrast, free-CoT leads on the frozen corpus by 9.2 (item CI (-11.0, -7.4); Welch (-23.7, 5.3); C7), replicating the direction of the 12B null—and §4.4 shows the swap reverses it. On the frozen corpus, navigation training transforms a small model and adds nothing to a large one (registered T1): D − BASE = +41.1 at 4B, −10.3 at 8B, −16.5 at 14B, against a baseline carrying exemplar help; under the repaired curriculum D-bidir-14B reaches 88.3, +12.6 over the same base.

Three contrasts, on both corpora (C9, C13). B′ rewrites every Thought: of the D episodes into free prose with the same teacher; actions, observations and terminals are byte-identical (frozen corpus: 4,246/4,246 through gates, 2,085 items kept at 1.020×; bidirectional corpus, registered before construction: 4,246/4,246, 2,090 at 1.020×; the retained sets are seeded uniform subsamples under the token match — a disclosed coverage confound whose direction we do not assert). Participation (P2, P10): B′ tool-call 100.0% on both corpora where B pooled 58.6% / 84.7%—the absence of the protocol from the trace is the dominant causal factor, cured by construction. Curriculum: the decision-exposure gain is format-independent—B′-bidir − B′-frozen = +23.8 (item CI (22.5, 25.0)), the fourth replication of the swap (+24.9 / +26.8 / +29.1 on D across families). Notation (P3, P11): frozen, D − B′ = +6.1 at 4B (item CI (4.7, 7.4), under the registered +5 bar; −0.2 at 8B); repaired, +**9.1** (item CI (7.9, 10.3), clearing the bar; Welch (-2.2, 20.5), still not seed-significant). Its anatomy: B′-bidir answers better (answer-acc 89.1 vs 85.7) and abstains worse (64.9 vs 76.7); D-bidir 83.6 ± 1.2 vs B′-bidir 74.5 ± 4.4 on conjunction—the compiled notation buys the terminal discipline that turns a correct answer into a valid, grounded, conformant, enacted path, not the reasoning. And none of it buys transfer: curriculum-matched with exemplars on the org domain, D − B′ = −6.3 (P14, §4.6). The operative reading: a small model needs traces that carry the tool protocol and a curriculum that varies the decisions; the compiled notation adds a real but secondary in-domain margin, concentrated in serialization and abstention, and contributes nothing to transfer. These experiments do not establish a robust notation advantage or notation equivalence: the D−B′ margin is item-CI-real, seed-fragile, home-domain-only and coverage-confounded. Free-CoT supervision fails small models not because prose is too hard for them but because prose-only traces silently train the tools away.

4.6 Transfer: competence moves (no better than free-CoT), legend does not (C10, C12)

Org domain (4,313 items). The Gemma-12B arms and the 4B exemplar-forced arm are reported at three seeds (Addenda 12.2–12.3, registered in advance); BASE is seed-invariant; B′-4B zero-shot is single seed. answer-acc† (Answer: line only) and labeled† (answered items whose path names a relation) are post-hoc observables the first zero forced into view.

Table 3: Experiment X. † post-hoc or labelled sensitivity; ± over three seeds where given.

Arm	org conj	tool-call	org answer-acc†	kinship answer-acc†	labeled†
BASE 12B (exemplars)	51.3	100%	51.1	48.2	100%
BASE 4B (exemplars)	53.3	82.7%	56.4	25.7	96%
D-frozen 12B (3 seeds)	0.0 ± 0.0	92.1% (76.9–100)	67.0 ± 16.7	69.5 ± 8.4	0%
D-bidir 12B (3 seeds)	2.0 ± 3.5	93.7% (81.3–100)	65.4 ± 4.5	81.1 ± 4.7	2%
D-bidir 4B (seed 0)	0.7	100%	3.2	85.5	75%
B 12B (free-CoT, exemplars; 3 seeds)	70.9 ± 6.6	96.9%	74.8 ± 6.7	83.6 ± 5.5	92%
B′ 4B (seed 0)	4.1	100%	32.6	83.8 ± 0.7	23%
D-bidir 4B + 2 exemplars† (3 seeds)	40.2 ± 3.7	100%	39.6 ± 3.4	85.5	100%
D-bidir 12B + 2 exemplars† (3 seeds)	65.1 ± 5.9	100%	62.5 ± 5.4	81.1 ± 4.7	99%

Prompt regimes: BASE and B rows carry the two pinned fairness exemplars (registered policy for arms whose training lacked the tool grammar); zero-shot D/B′ rows carry none; "+ 2 exemplars" rows force the same two pinned exemplars —the fixed fictional pair of §3

mapped through Φ, not selected per run (Appendix K), demonstrating vocabulary, output structure and loop shape together. Registered verdicts. P4 (a lexical component exists) passes trivially—the drop is to zero. P5 (does relation-abstraction transfer better?) is floor-dominated and uninformative. P6 (the 4B contrast) is refused: BASE-4B floors at 82.7%. P12 (three seeds) is not supported: the exemplar-matched capability contrast D-bidir+ex − BASE at 12B is +13.8 (item CI (12.4, 15.2)), short of the registered +15 margin, with a seed-level interval (-0.9, 28.5). P13 (two-sided): D-bidir+ex − B(ex) = −5.8 (item CI (-7.1, -4.5); Welch (-22.1, 10.5); per seed 69.5 / 58.4 / 67.5 against 63.8 / 72.0 / 76.8): the free-CoT-trained model transfers at least as well on the verified metric and abstains far better on the new domain (95–98% vs 74–90%).

What the artifacts show at three seeds†. A form–competence dissociation, inverted by capacity. The 4B navigator reproduces the trained legend (75% labeled paths) and cannot do the task (3.2% correct against 85.5% in-domain). The 12B navigators do the task—D-frozen answers 67.0 ± 16.7% of org questions, D-bidir 65.4 ± 4.5%, against the untrained base's 51.1%—and abandon the legend: zero relation labels in every D-frozen seed. Verbatim, D-frozen-12B emits Evidence path: Nayeli Okonkwo -> Wafa Okonkwo -> Fanta Okonkwo (correct answer, path_valid false) where the exemplar-carrying base writes …-> directManagerOf -> …. The seed spread is the other half of the finding: D-frozen's answer accuracy ranges 48.1–80.1 across seeds and one seed navigates on only 76.9% of items with 43.6% unparseable terminations—zero-shot participation on a new legend is itself fragile. Final-answer accuracy conflates not solving the task with no longer speaking the evaluator's protocol; agent evaluation needs the layers this experiment was forced to separate: capability → participation → execution → serialization → verification. Two in-context examples repair the legend completely (labeled 2% → 99.4 ± 0.8%) and convert the transferred competence into verified evidence: 65.1 ± 5.9% on the never-seen domain. What the repair does not buy: with the legend restored, navigation training is +13.8 over no training (below the registered margin) and −5.8 against free-CoT training—the competence that transfers is not specific to navigation supervision. At 4B the exemplar-forced arm scores 40.2 ± 3.7 against the sub-floor base's 53.3 — reported descriptively, with P6's refusal preserved; the org benchmark is easier at 4B (BASE-4B +37.6), voiding 4B level comparisons; the 4B/12B transfer arms also differ in family (Qwen3 vs Gemma), and the construction pairs a fresh graph seed with the relabeling, so the shift is graph and lexicon jointly; and the curriculum-matched contrast goes the other way: P14: D-bidir+ex − B′-bidir+ex = −6.3 (item CI (-7.4, -5.2); Welch (-16.4, 3.7); both floors and labeled 100%)—the prose-format arm transfers better at 4B, matching P13's direction. The table's answer-acc† is exact-gold string match — the strict selection observable (records agree: 62.5 ± 5.4); the records' answer_correct field is lenient pool membership (71.2 ± 5.0; §3). Under the strict end-to-end metric (conjunction ∧ exact-gold) D-bidir-12B+ex scores 62.0 ± 5.2 against 65.1 ± 5.9 conjunction, and every transfer comparison's direction is unchanged (Appendix K). The precise statement: at 12B the competence learned on one domain moves to another and becomes verified evidence once two examples of the new legend are supplied, but free-CoT-trained competence moves at least as well, so what the exemplar unlocks is a property of capable models, not of navigation traces; zero-shot (arms evaluated without exemplars; B never was), nothing transfers on a verified-evidence metric; at 4B the model memorized the legend, not the map.

5 Discussion

What went wrong, briefly. Eight defects were found (Appendix I); four produced perfect zeros indistinguishable from capability results and one cost a headline claim (an 8B↔14B crossover that existed at an improvised 50% participation bar and vanished at the registered 90%). The rules that came out of them—a zero is not a result until the instrument is proven able to produce a non-zero; a registered gate that does not run on every arm is not a gate; a reconstruction is an instrument too—are offered as practice, not contribution.

Structure substitutes for capacity, conditionally. The frozen-corpus capability shape, the suppression gradient and the transfer inversion fit one reading: structured trajectories provide scaffolding a small model cannot rebuild through free-form reasoning. Y2 bounds it from above: with the decision exposed, navigation supervision still lifts a 14B model 12.6

points over its few-shot base, so what a capable model does not need from the traces is a route, not a map. The post-training question is therefore conditional—value of supervision = f(capacity, curriculum diversity, trajectory representation, tool participation, domain shift)—and each argument was measured here in isolation. The recipe that lifted a 4B model into a 12B model's in-domain range made its cross-domain transfer worse than no training.

The map may not need to live in the weights. If a capable model transfers its competence within an isomorphic lexical shift whatever the supervision that produced it, loses only the legend, and recovers that from two examples, then for capable models the engineering leverage moves from teaching the map to exposing it: an entity graph (the map), a lexical-semantic layer that names it (the legend), an action graph of executable transitions (the routes) and a governance layer of permitted transitions (the boundaries)—four views of one runtime context. This study does not test that architecture; it supplies one empirical motivation: navigation policy, domain vocabulary and capacity behave as separable factors, so for small models, compile more of the map into the curriculum; for capable models, keep more of it outside the weights and make it navigable at runtime. An untrained capable model reads a served map essentially perfectly (Appendix B) and carries its competence across a legend change (§4.6); what it lacks, untrained, is the boundary—abstention recall 10.5% at Gemma-12B and 71.9% at Qwen3-14B—which is exactly the layer a runtime can enforce rather than a curriculum teach.

6 Limitations

The central contrast is two-family and ≥12B; at 4B and 8B prose-only arms fail the participation floor, so no valid free-CoT control exists and the 4B evidence is the D-side curriculum effect plus a labelled intent-to-treat number. Seed-level uncertainty is wide (three seeds; Welch (5.2, 42.3) and (5.1, 22.8) on the matched contrasts). The free-CoT collapse is Gemma-specific and unexplained. The suppression ladder ran on the frozen corpus; the notation contrast is item-CI- but not seed-significant on both corpora. Two synthetic domains, one isomorphism; transfer findings are post-hoc at three seeds, the registered exemplar-matched margin was missed (+13.8 vs +15), and exemplar-free transfer is ≈0. The abstention gap is the navigators' open failure. Exemplar asymmetry favors BASE and B (quantified, not removed); B′ covers 49% of D's items under token matching, a confound whose direction we do not assert; one family was substituted (tooling); token costs are proxy-tokenizer counts.

7 Conclusion

The study set out to show that navigation traces beat chain-of-thought for small models, and learned instead: that its first result was an artifact of its own trace compiler; that with the decision exposed, navigation wins the answerable domain at matched curriculum in both families where the contrast is valid and lifts a 4B model into a 12B model's range, while abstention remains its open failure; that the deepest difference between the supervisions is whether the tools survive fine-tuning; that navigation supervision contributes participation, curriculum and a secondary, home-domain notation margin; and that competence trained at 12B crosses a change of legend no further than free-CoT training carries it, while at 4B only the legend makes the trip. The practitioner's line: put the tools in the traces, vary every decision the agent is expected to learn, show two examples of the new domain's legend—and run your gates on every arm, because every important number in this paper was, at some point, a zero that meant something else.

Reproducibility statement

Code (generator, formatters, harness, scorer), both benchmarks, every training corpus and variant, the pre-registration with all addenda and the frozen claim ledger are released (anonymized supplementary material); per-item scored records for every arm are provided with them in the supplement, with a SHA-256 list and a label→run_id manifest, so every table and figure regenerates from the supplement alone, without GPU access or network. To meet the submission system's size limit the supplement omits two record fields that no re-

leased analysis reads—the full episode log and the retrieved-edge list, with complete rollout records retained for a sample of transfer items—and the emitted answer text of the 0.6B–8B ladder arms, whose tables use only scored fields; scripts/verify_supplement.py re-runs every released analysis against both the supplement and the untrimmed archive and diffs the outputs, which are identical, figures included. The complete archive, with all rollout transcripts and emitted text, is released with the paper. Appendix J binds each table to its artifacts. Metrics are reproducible from the archived records with the versioned scorer; generations are not bit-reproducible across vLLM batch sizes.

AI use statement

In this work, we used generative AI tools for the following tasks that require disclosure. Implementing methods; cleaning and reformatting datasets: Claude Code (Anthropic), running as an agent in the VS Code IDE, wrote and refactored the benchmark generator, formatters, training and evaluation harness, scorer, and the corpus-assembly, analysis, figure and release tooling. Designing or giving feedback on experiments: it ran a simulated pre-submission review that proposed the revision's three follow-up experiments (the second-family replication, the format contrast on the repaired curriculum, and multi-seed transfer), and drafted their registration addenda, including predictions and decision rules that apply the core pre-registration's thresholds; the authors approved each registration before any of its data existed and made every go/no-go decision. Supporting qualitative and thematic data analysis; interpreting results: it drafted the failure-mode, abstention and mechanism analyses and first-draft interpretations of results, each checked by the authors against the per-item records. Generating synthetic data sets: a teacher model (Gemini-2.5-flash) generates the chain-of-thought training corpora for conditions B and B′; this is part of the experimental design and an object of study (§3), while the benchmarks themselves are produced by deterministic code from fixed seeds. We have not used generative AI tools to develop the thesis or conceptual framework; to propose the core pre-registration's hypotheses, confirmatory thresholds, falsifiers or decision rules; or to generate, alter or select any evaluation record or reported number: records come from the evaluated models and a versioned deterministic scorer, and every number is recomputed by audited scripts (Appendix J). Formulating mathematical claims, proving them, and translation are not applicable to this work. Additionally, we used generative AI tools for drafting and editing prose, including the text of later pre-registration addenda written under the authors' direction, and for literature search. We have reviewed all AI-assisted work: code is covered by the repository's test suite and re-run against archived records, every citation was verified against its publication venue or arXiv record, and the supplement's regeneration of every table and figure was verified byte for byte against the full archive. We take full responsibility for the final content of this work, including text, claims, code and artifacts produced with the aid of generative AI.

Ethics statement

All data are synthetic: the kinship and organization graphs, names, dates and questions are generated by the released code from fixed seeds and describe no real person or organization. No human subjects were involved.

References

NeurIPS 2020 workshop on pre-registration in machine learning. In Luca Bertinetto, João F. Henriques, Samuel Albanie, Michela Paganini, and Gül Varol (eds.), Proceedings of Machine Learning Research, volume 148, 2021.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks. Proceedings of Machine Learning and Systems, 3, 2021.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. FireAct: Toward language agent fine-tuning. arXiv preprint arXiv:2310.05915, 2023.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. Agent-FLAN: Designing data and methods of effective agent tuning for large language models. In Findings of the Association for Computational Linguistics: ACL 2024, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2:665–673, 2020.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Proceedings of ACL 2023, pp. 14852–14882, 2023.

Jinhao Jiang, Kun Zhou, Wayne Xin Zhao, Yang Song, Chen Zhu, Hengshu Zhu, and Ji-Rong Wen. KG-Agent: An efficient autonomous agent framework for complex reasoning over knowledge graph. In arXiv preprint arXiv:2402.11163, 2024.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, and Jiawei Han. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In Findings of the Association for Computational Linguistics: ACL 2024, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13):3521–3526, 2017.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Proceedings of ICML 2018, pp. 2873–2882, 2018.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In International Conference on Learning Representations, 2024.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In International Conference on Learning Representations, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2025.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In Proceedings of ACL 2023 (Short Papers), pp. 1773–1781, 2023.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. Psychology of Learning and Motivation, 24:109–165, 1989.

Ida Momennejad, Hosein Hasanbeig, Felipe Vieira Frujeri, Hiteshi Sharma, Robert Osazuwa Ness, Nebojsa Jojic, Hamid Palangi, and Jonathan Larson. Evaluating cognitive maps and planning in large language models with CogEval. In Advances in Neural Information Processing Systems, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332, 2021.

John O'Keefe and Lynn Nadel. The Hippocampus as a Cognitive Map. Oxford University Press, 1978.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. In Advances in Neural Information Processing Systems, 2024.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In International Conference on Learning Representations, 2024.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In Proceedings of EMNLP 2017, pp. 338–348, 2017.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In Advances in Neural Information Processing Systems, 2023.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Proceedings of EMNLP-IJCNLP 2019, pp. 4506–4515, 2019.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In International Conference on Learning Representations, 2024.

Andrea Volpini and Chiara Carrozza. Sangue e grafi: Teaching a small model to read the bloodline. WordLift project report, June 12, 2026. https://wordlift.io/blog/en/sangue-e-grafi/, 2026.

Andrea Volpini and Elie Raad. RLM-on-KG: Heuristics first, LLMs when needed: Adaptive retrieval control over mention graphs for scattered evidence. arXiv preprint arXiv:2604.17056, 2026.

Andrea Volpini, Elie Raad, Bruno Gamba, and Davide Riccitelli. Structured linked data as a memory layer for agent-orchestrated retrieval. arXiv preprint arXiv:2603.10700, 2026.

Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization. In Advances in Neural Information Processing Systems, 2024.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards AI-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698, 2015.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In International Conference on Learning Representations, 2023.

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In Proceedings of ACL 2016 (Short Papers), pp. 201–206, 2016.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In Advances in Neural Information Processing Systems, 2022.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. AgentTuning: Enabling generalized agent abilities for LLMs. In Findings of the Association for Computational Linguistics: ACL 2024, pp. 3053–3077, 2024.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In Proceedings of AAAI 2018, 2018.

A Pre-registration and gates

Thresholds, falsifiers and decision rules were committed in dated addenda (1–12) before the data they govern; retractions are annotated in place, never rewritten; outcomes are computed mechanically by scripts written before the data existed. The claim ledger is the single source of what the paper asserts; a claim changes only when an artifact contradicts it. Registered rules used here: the structural contrast (D vs B, item CI-low $> +5\,pp$), the capability contrast (D vs BASE, CI-low $> +15\,pp$ and absolute $\geq 40\%$), enactment integrity (recitation $< 2\%$), the participation floor (pooled tool-call $\geq 90\%$, contrast refused otherwise), the ladder informativeness rule (a size counts only if the better arm reaches $\geq 10\%$ and both arms clear the floor; fewer than three informative sizes $\Rightarrow$ the capacity hypothesis is untestable), the Experiment Y bound (held-out compositional cell, CI-low $> +10\,pp$), and the Y2 rules P7–P9 (§4.4). The pre-registration and ledger are released verbatim.

B In-context reasoning is solved at 12B (C1); conditions A, C, E

Untrained Gemma-4-12B with a sufficiency-gated fact block clears the pre-committed $\geq 90\%$ ceiling on the structural split; every registered difficulty dial closed below the $\geq 25\%$ genuine-failure bar (hop depth: $0.06\% \rightarrow 1.61\% \rightarrow 1.19\%$ failure across hops 1–3). Fine-tuning on any of A–F added no measurable signal there, and two conditions (A, F) scored below the untrained base. This preliminary forced the facts-withheld protocol and is the only place conditions A (terminal-only in trace context), C (terminal-only with Facts:) and E (rejected-candidate enumeration) were run.

C Metric definitions

For a gold-answerable item, path_valid: the emitted chain parses and every edge exists in the graph; grounding_ok: it starts at the question entity and ends at the answer; conforms: it satisfies the template's SHACL shape (right relations, right order, ordering constraint respected); enacted: every edge appeared in a tool observation of the same episode. Enacted conjunction is their conjunction. For a gold-abstention item, abstention recall is whether the model abstained. Unparseable outputs score as failures in their gold domain. answer-acc compares the Answer: line to the gold candidate set. Pooled accuracy (one outcome per item over all 4,437) is a secondary, non-registered view.

D Per-seed values and intervals

Arm	conjunction s0 / s1 / s2	abstention s0 / s1 / s2	tool-call (worst seed)
Gemma-12B B frozen	73.2 / 84.2 / 77.9	98.6 / 99.5 / 98.9	92.4%
Gemma-12B D frozen	60.8 / 61.1 / 50.4	39.5 / 22.5 / 49.7	96.3%
Gemma-12B F frozen	62.8 / 57.2 / 48.4	87.2 / 82.4 / 72.1	94.0%
Gemma-12B B bidir	49.5 / 60.3 / 66.0	98.5 / 97.7 / 99.7	96.4%
Gemma-12B D bidir	87.3 / 83.5 / 76.3	56.5 / 53.3 / 43.1	100.0%
Gemma-12B D anonymized	28.2 / 43.5 / 41.4	41.2 / 29.2 / 59.2	93.9%
Qwen3-14B B frozen	66.0 / 64.1 / 74.8	—	95.9%
Qwen3-14B B bidir	70.0 / 77.6 / 75.3	—	100.0%
Qwen3-14B D frozen	57.8 / 60.1 / 59.5	—	95.7%
Qwen3-14B D bidir	85.8 / 91.4 / 87.6	—	96.5%
Qwen3-4B D bidir	83.4 / 84.9 / 82.5	—	100.0%
Qwen3-4B B bidir	35.0 / 45.0 / 35.4	—	70.1%

Welch t 95% intervals (pp): Gemma D−B frozen (-36.1, -5.9); D−B bidir (5.2, 42.3); D-bidir−D-frozen (9.6, 40.2); B-bidir−B-frozen (-38.2, -1.4); D-bidir−B-frozen (-10.5, 18.4); F−B (-39.0, -5.6); Qwen3-14B D-bidir−B-bidir (5.1, 22.8); D-bidir−D-frozen (21.5, 36.8); B-bidir−B-frozen (-6.7, 18.7); Qwen3-4B D-bidir−D-frozen (18.7, 34.9); Qwen3-4B D−B′ (-1.6, 13.7); Qwen3-8B D−B′ (-17.4, 17.0); Qwen3-14B D−B frozen (-23.7, 5.3).

E Anatomy of the free-CoT collapse (Gemma-12B)

Arm	failure rate	unparseable	abstained on answerable	answer right, path fails	multi-Action: turn	answer-acc†
B frozen	21.6%	10.7%	5.0%	5.2%	5.3%	83.6 ± 5.5
B bidir	41.4%	21.2%	12.2%	7.0%	18.8%	65.5 ± 10.5
D frozen	42.6%	15.4%	9.3%	12.7%	0.1%	69.5 ± 8.4
D bidir	17.6%	10.4%	2.6%	1.1%	0.0%	81.1 ± 4.7

Phrase audit (share of failed answerable items whose output contains the phrase): D frozen "eldest"/"oldest" 30.6%, "ascending" 38.1%, "youngest" 2.5%; D bidir 0.2% / 0.0% / 22.6% (and "descending" 25.2%); B frozen 0.8% / 0.0% / 45.7%; B bidir 0.5% / 0.0% / 33.4%. Answer-line accuracy per template (B frozen → B bidir): child 90.5 → 82.3, grandchild 65.3 → 45.2, grandparent 84.6 → 51.7.

F Capacity ladder

Table 4: Qwen3 ladder and Llama-Nemotron-8B, frozen corpus, decisive split (3 seeds; BASE single-run, exemplars). Tool-call pooled over seeds; contrasts refused when either arm is below 90%.

Family	arm	conjunction	abstention	tool-call	floor
Qwen3-14B	BASE / B / D	75.7 / 68.3 ± 5.7 / 59.1 ± 1.2	71.9 / 90.7 / 75.8	94.4 / 98.4 / 98.6%	pass / pass / pass
Qwen3-8B	BASE / B / D / B′	64.5 / 30.8 ± 2.9 / 54.2 ± 6.8 / 54.4 ± 1.2	76.3 / 79.8 / 65.5 / 65.2	88.5 / 84.7 / 100 / 100%	fail / fail / pass / pass
Qwen3-4B	BASE / B / D / B′	15.7 / 25.4 ± 10.7 / 56.8 ± 3.0 / 50.7 ± 0.5	23.7 / 96.3 / 82.3 / 64.3	35.3 / 58.6 / 100 / 100%	fail / fail / pass / pass
Qwen3-1.7B	BASE / B / D	8.9 / 2.1 ± 2.7 / 4.3 ± 2.0	15.8 / 73.8 / 61.8	40.6 / 16.9 / 86.2%	fail / fail / fail
Qwen3-0.6B	BASE / B / D	0.0 / 0.1 ± 0.1 / 12.6 ± 3.8	14.4 / 95.8 / 28.3	3.7 / 1.3 / 86.3%	fail / fail / fail
Nemotron-8B	BASE / B / D	5.0 / 0.0 / 27.9 ± 5.2	10.9 / 85.0 / 31.9	28.7 / 0.0 / 99.9%	fail / fail / pass

Contrasts: Qwen3-14B D−B = −9.2 (item CI (-11.0, -7.4); Welch (-23.7, 5.3)); Qwen3-8B and 4B D−B refused (B below floor; unregistered intent-to-treat +23.5 and +31.3); Qwen3-4B D−B′ = +6.1 (4.7, 7.4); Qwen3-8B D−B′ = −0.2 (-1.6, 1.2). With exemplars restored to D (a sensitivity run), the 14B deficit against the few-shot base is 15.2 points and the 8B deficit 3.3.

G Transfer details and verbatim outputs

Full output of D-frozen-12B on org item 10ff4314891a (correct answer, no relation labels): Thought: Compare hireDates of active candidates [Danai Okonkwo, Fanta Okonkwo, Habiba Okonkwo, Rehema Okonkwo, Xiomara Okonkwo] and find the most recent hireDate. …The most recent hireDate is 2009-12-17. / Evidence path: Nayeli Okonkwo -> Wafa Okonkwo -> Fanta Okonkwo / Answer: Fanta Okonkwo. Of D-frozen-12B's 3,235 answerable org items, 2,771 parsed and 464 (14%) were unparseable; 2,669 carried the correct answer. The opaque-relation-trained arm (Experiment Z, −20.3 in-domain) does not navigate under the new vocabulary (28.4% tool-call), which is why P5 is uninformative.

H Cost

A verified answer costs free-CoT ≈2,368 generated tokens and navigation ≈179 (≈13×) at equal observation tokens and tool turns. Accounting: "tokens per verified answer" sums whole-transcript tokens (prompt including exemplars where carried, generated text, observations) over all answerable episodes including failures, divided by enacted-verified answers; the 179/2,368 figures count generated tokens only under the same denominator. Frontier (Gemma-12B, seed 0 of the 82.4 ± 5.6 arm): D-bidir 87.3% at 670 tokens per verified answer, open-book RAG 58.1% at 257, B 78.4% at 3,124. Proxy tokenizer applied identically across arms; single seed; reconstructed from transcripts since the record fields were unpopulated; exemplar text inflates the B and BASE totals.

I Bug ledger

1. Tool-grammar mismatch — caught by 0 tool calls in a clean smoke.
2. Train/inference render mismatch silently dropping observations — trained arms at 0 calls while untrained made 2–7.
3. Corpus contract drift — printing the corpus.
4. Under-specified run_id + trusted cache — counting records per arm.
5. Corpus variant absent from run_id, beside a comment asserting the opposite — recomputing run_ids against logs.
6. Registered gate run on a subset of arms, then dropped — one exact 0.0%; reading the artifact.
7. Untested reconstruction inverting an efficiency conclusion — its implication contradicted a measured error rate.
8. answer_correct measured candidate-pool membership while its docstring promised multi-answer correctness; the grandparent shape under-constrains the global optimum — caught by the second external reviewer's denominator question and the Appendix K audit.

The floor audit itself first applied an improvised 50% bar instead of the registered 90%—a threshold violation caught and recorded, whose correction retracted the capacity crossover. vLLM is not bit-deterministic across batch sizes at temperature 0.

J Hash chain and provenance

Benchmark (kinship, frozen) 9ced3f8bfcc52a7e; benchmark (org) 3bce1a47a587612d; corpus (frozen, A–F) 6fac2354a8f745a5; bidirectional variant run identity 0f31833f53d4d07b (Gemma B-bidir and every ladder cell) / e5f9024dbb81ac4d (Gemma D-bidir); anonymized relations ee60c6eb7ed2770c; B′ b78b36e02ef99463; registered treatment config 48a293afc6accee2; scorer 0.4.0. Every cell is addressed by a nav run_id through the manifest written by the released record collector; the run_ids for the Gemma-12B and Y2 arms are listed in the released draft's Appendix C.

K Held-out views, metric cross-tabs, abstention anatomy, exemplars (labelled)

All from archived records (Addendum 13); no registered metric changed. Held-out = grandchild + grandparent templates (n = 1,677 of 3,328); child (n = 1,651) is seen under the bidirectional curriculum.

Table 5: Enacted conjunction by evidence kind (mean ± SD over 3 seeds).

Arm	full split	held-out only	now-seen child
Gemma B frozen / B bidir	78.4 ± 5.5 / 58.6 ± 8.4	70.4 ± 6.1 / 42.4 ± 3.3	86.6 ± 6.5 / 75.1 ± 15.3
Gemma D frozen / D bidir	57.4 ± 6.1 / 82.4 ± 5.6	39.5 ± 17.7 / 66.9 ± 11.1	75.7 ± 6.4 / 98.1 ± 1.6
Qwen3-14B B frozen / B bidir	68.3 ± 5.7 / 74.3 ± 3.9	49.5 ± 11.3 / 54.1 ± 8.5	87.5 ± 4.3 / 94.9 ± 0.8
Qwen3-14B D frozen / D bidir	59.1 ± 1.2 / 88.3 ± 2.8	70.2 ± 2.7 / 78.3 ± 5.6	47.9 ± 0.7 / 98.4 ± 2.0

Held-out-only contrasts (pp; item CI; full-split beside): Gemma D-repair +27.4 (25.6, 29.2) | +24.9; Gemma D−B bidir +24.5 (22.7, 26.3) | +23.7; Gemma B-repair −28.0 | −19.8; Qwen3-14B D-repair +8.1 (6.4, 9.8) | +29.1; Qwen3-14B D−B bidir +24.2 (22.3, 26.2) | +13.9; Qwen3-14B B-repair +4.6 | +6.0. Per-template Qwen3-14B (child/grandchild/grandparent, 3-seed means): B 87.5/55.5/45.3; B-bidir 94.9/62.3/48.5; D 47.9/40.6/90.5; D-bidir 98.4/61.8/89.6; the grandparent column is the most seed-volatile cell in the study (per-seed matrix in the released draft's Appendix F).

Cross-tab, conjunction × answer_correct (answerable items, seeds pooled): conj∧¬ans = 0.03% (Gemma B), 0.05% (B-bidir), 2.27% (Gemma D-bidir), 0.25% / 0.14% (Qwen3-14B

B-/D-bidir); ¬conj∧ans = 5.2 / 7.0 / 1.1 / 2.9 / 0.3%. Full success (conj∧ans, mean ± SD): 78.4 ± 5.5, 58.6 ± 8.4, 80.1 ± 5.2, 74.1 ± 4.1, 88.1 ± 3.0 — every headline contrast keeps its sign and approximate size (Gemma bidir D−B +21.5; Qwen3-14B +14.0).

Abstention anatomy (gold-abstention items, mean ± SD over seeds; mutually exclusive): Gemma D-bidir — abstained 51.0 ± 7.0, answered with a parseable path claim (unsupported answers) 36.6 ± 1.8, unparseable (incl. empty generations) 12.4 ± 8.3; Gemma B-bidir 98.6 / 0.2 / 1.2; Qwen3-14B D-bidir 84.8 / 10.6 / 4.7; B-bidir 94.2 / 2.0 / 3.9. Representative failure: Thought: Living candidates: […]. Order by birthDate descending and take the youngest …/ Answer: Uma Eriksson on a gold-ABSTAIN item. Turn-limit terminations are not separable in the archived records and are not claimed.

Exact-gold selection accuracy (Addendum 14). The records' answer_correct is candidate-pool membership (bug #8); strict observables (mean ± SD over seeds): exact-gold accuracy and strict full success (conjunction ∧ exact-gold) — Gemma B frozen 75.4 / 73.1 ± 5.8; B bidir 57.0 / 54.4 ± 7.4; D frozen 51.7 / 51.0 ± 1.9; D bidir 79.4 / 78.7 ± 4.7; Qwen3-14B B frozen 65.4 / 64.0 ± 4.2; B bidir 73.3 / 72.2 ± 3.6; D frozen 43.5 / 43.4 ± 0.5; D bidir 83.4 / 83.3 ± 3.4. Every headline direction is preserved (Gemma bidir D−B +24.3; Qwen3-14B +11.1; the Qwen3-14B frozen deficit widens to −20.6 because D-frozen's conjunction was shape-inflated on grandparent items). Conjunction ∧¬exact-gold sits entirely in the grandparent template (Qwen3-14B D-frozen 52.4%; Gemma D-frozen 21.4%; D-bidir 12–16%; child/grandchild 0.0%): that shape validates the exhibited chain without pinning the global optimum. The §4.2 answer-level mechanism audit uses items with ≥ 3 dated candidates.

The pinned exemplars. Two fixed worked examples (one answerable, one abstention) over a closed fictional surname disjoint from every generated entity, hashed into every run_id; the org pair is the same two examples mapped through Φ ("Ada/Uri Sample", directManagerOf/hireDate). Verbatim text in the released draft's Appendix H.

L Addendum 12 outcomes

Addendum 12 registered, before construction: (i) B′ on the bidirectional curriculum at Qwen3-4B (3 seeds); (ii) Experiment X at seeds 1–2 for the Gemma-12B arms and a curriculum-matched 4B transfer contrast. All cells ran the same day; outcomes under the registered rules: extbfP10 pass (B′-bidir tool-call 100%); extbfP11 +9.1 (item CI (7.9, 10.3), clearing the +5 bar; Welch (-2.2, 20.5)), with the curriculum gain on B′ itself +23.8; extbfP12 not supported (+13.8 vs the +15 margin; seed-level interval includes zero); extbfP13 −5.8 (item CI (-7.1, -4.5)); extbfP14 −6.3 (item CI (-7.4, -5.2)). Reported in §4.5 and §4.6.