THE REACH OF INTERVENTIONS: WHAT CAN BE WRITTEN INTO A (SEMI-)FROZEN LANGUAGE MODEL?

Anonymous authors
Paper under double-blind review

ABSTRACT

Writing new knowledge into a partially frozen language model means changing an answer's reachability through one of three routes: carrying and placing content into weights, carrying and placing content through activations, or reweighting or tilting the base model's existing sampling support. Fine tuning, RL, steering, retrieval, and editing span these routes, and the route determines what each can reach. We develop a predictive theory of what each intervention can write. Using sampling support, verifiability, and a prior-support decomposition, the theory separates items into display, verifiable closure, and underdetermined layers, and derives an inheritance law for reachable answers. It predicts learning limits and per item learnability before training. Across synthetic and public benchmarks, the theory explains channel specific reach and failure modes, matches observed ceilings, and turns post training from trial and error into measurement.

1 INTRODUCTION

Modern language models are increasingly deployed as semi-frozen systems: a pre-trained model whose core weights are largely fixed, augmented by a growing set of mechanisms such as activation steering (Li et al., 2023), memory modules (Packer et al., 2023), retrieval (Lewis et al., 2020), post-training, and parameter-efficient fine-tuning that add facts, skills, or behaviors. These mechanisms are usually studied separately, yet they share a common abstraction: each writes information into a computation the model already has. They differ, however, in what they can write. This raises a question that practice answers by trial and error and theory has addressed only through isolated impossibility results (Petrov et al., 2024; Xu et al., 2024): *what can each channel write?* This paper builds the missing framework as a predictive theory of intervention reach: given a knowledge item, it predicts the causal effect of each channel family and identifies which channels are bounded by the model's existing support. The framework begins by classifying each knowledge item into one of three layers, according to what the base model can already support.

Consider a trained model and a user named Mira. If we tell the model that Mira was born in Osaka, the model correctly answers that Mira is Japanese. If we remove that statement from the prompt, the model can no longer answer the question about Mira's nationality. However, the model still answers the generic question "what nationality is someone born in Osaka?" fluently. The model therefore has the association between Osaka and Japan, and it can perform the one-step inference from birthplace to nationality. What the model lacks is the binding between Mira and Osaka. We call this state *verifiable closure*: the model does not sample the answer on its own, but the answer becomes available once the premise is supplied.

The second layer is more severe. Verifiable closure still relies on knowledge that the model already possesses. The underdetermined layer arises when the required fact lies outside the prior altogether, such as an unusual collection that the model has never encountered. In this case, the model can neither recall the fact nor verify it, although a multiple-choice test may still hit the correct answer by preference alone. We call this state the *underdetermined* layer.

The third layer is *display*. In this layer, the prior already supports the answer, so the model can often select it when the answer appears among options or when the prompt narrows the space of

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.

possible answers. However, direct prompting does not reliably elicit the answer (Kadavath et al., 2022), because the model's output distribution is influenced by other preferences.

These layers are not merely conceptual; they occur in real models and can be separated by measurement. We use Qwen3-8B as a running example because it is the main base in our controlled experiments, and similar patterns appear across families. For a closure item, Qwen3-8B never samples the mother of Sean Penn in 64 tries, but it answers correctly once the relevant premise is supplied. For an underdetermined item, it cannot answer about the newest NASA telescope because the fact is post-cutoff. Display items are those where the prior already supports the answer, but direct prompting does not reliably elicit it, and their identification requires the measurement procedure introduced below.

To assign items to these layers, we use three simple measurements: sampling support, verifiability, and a preference decomposition. Sampling support asks whether the model produces the correct answer on its own. Verifiability asks whether the model can complete the inference once a premise containing the answer is supplied. The preference decomposition separates multiple-choice accuracy that comes from general preference from accuracy that comes from item-specific knowledge. These measurements place each item into one of the three layers and connect each layer to a natural family of methods: elicitation for display (Burns et al., 2023; Mallen et al., 2023b), memory or retrieval for closure (Packer et al., 2023; Lewis et al., 2020), and injection or editing for underdetermined content (Meng et al., 2022a;b).

The central question is whether the reach of these channels can be predicted rather than discovered by trial and error. We propose a single principle: no channel can make the model answer correctly beyond the union of what the base model can already sample and what the channel explicitly carries. We call this the *inheritance law*, because each channel inherits its reach from the base model's existing support or from the content it supplies.

This principle yields different predictions for different channel families. Direction-based interventions, such as activation steering (Li et al., 2023), can only tilt existing preferences within a narrow range, so they cannot introduce new facts. Content-based channels, such as retrieval or memory injection (Lewis et al., 2020; Su et al., 2025), can supply new content and let the model compose it with its existing inference ability. Reward-only reinforcement learning (Chen et al., 2026; Wen et al., 2026) can only reweight answers the model already samples, so it cannot learn facts that never appear in its own generations. Gradient-based writing can bind specific facts into the weights (Meng et al., 2022a), but it also affects related knowledge (Cohen et al., 2024). These predictions match the three layers introduced above and are tested systematically in our experiments.

We test these predictions in a controlled study on Qwen3-8B, with replication on Qwen3-14B and four additional models. The study covers the full matrix of channels and layers introduced above. Two results frame the empirical picture. First, on items that require a binding the base model does not sample, content-based injection succeeds at roughly eleven times the rate of direction-based steering. Second, reward-only reinforcement learning improves accuracy only up to the level predicted by the answer distribution it can already sample, and it saturates there. These results are consistent with the inheritance law and with findings from five external research areas that report related limits (§4).

The paper makes four contributions. It states the inheritance law together with its indistinguishability theorem and chain rule. It introduces an audit and probe-cascade methodology for measuring what a channel can reach. It provides controlled cross-model evidence for the operator constants predicted by the theory. It releases ReachBench, a public-data benchmark for measuring sampling support and verifiability on natural items.

2 THEORY

The theory's goal is to make channel reach predictable rather than merely observed. Given a knowledge item and its layer, the theory should predict the causal effect of each channel family. Because knowledge states are not directly observable, the theory must begin with measurable quantities. We therefore define items, scores, sampling support, verifiability, and the prior-support decomposition before stating the inheritance law.

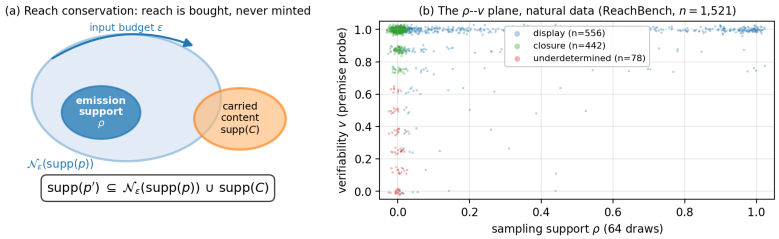


Figure 1: The reach conservation law. (a) Each intervention holds a budget ε and carried content C; the reachable answers stay within the budget-widened base support $\cup$ carried content. (b) The ρ–v plane of the 1,521 ReachBench items, colored by layer; v is ρ with the premise supplied.

2.1 Definitions

Definition 1 (Items, scores, deficit)**.** *A knowledge item is a pair $k = (q, a)$; the score $s_M(a|q)$ is the mean token log-probability of a given q; the deficit $d(k) = s_M(\text{gold}|q) - \max_{a' \in D(k)} s_M(a'|q)$ over the distractor set $D(k)$ is the signed margin over the best distractor, negative until the gold leads.*

Definition 2 (Sampling support and verifiability)**.** *Sampling support $\rho(k)$ is the probability that a single temperature-one sample hits a; verifiability $v(k)$ is the hit rate of a under the same sampler when a premise containing it is supplied. Both depend on the base model alone.*

Definition 3 (Prior-support decomposition)**.** *Closed-book multiple-choice accuracy decomposes into a marginal-preference floor, estimated by leave-one-out aggregation of per-answer scores, plus an item-specific lift (the residual).*

Definition 4 (Layers)**.** Display*: the prior supports the answer, but direct prompting does not reliably elicit it; operationally, a positive item-specific lift with deficit inside a narrow band.* Verifiable closure*: $\rho \approx 0$ and $v \approx 1$.* Underdetermined*: ρ and v are near zero at the measurement budget.*

Definition 5 (Content carriage and placement)**.** *A protocol* carries content *if its training signal includes the answer itself; it* places content *if its update raises the answer's probability through a likelihood term or content-bearing activations.*

A lexicon mapping every construct to its symbol and instrument is in Appendix Table 8.

2.2 The inheritance law

The inheritance law states that what a policy can come to answer correctly after any intervention or training is contained in the union of two sources: the base model's sampling support, widened by the intervention's budget, and the content the protocol carries and places.

The law can be stated as a conservation principle (Fig. 1a). Every intervention has two resources. The first is an input or activation budget ε, which widens the base model's support by allowing answers within ε of that support. The second is carried content C, which adds answers the protocol explicitly supplies. The reachable answer set therefore obeys

$$\text{supp}(p') \subseteq \mathcal{N}_\varepsilon(\text{supp}(p)) \cup \text{supp}(C),$$

where $\mathcal{N}_\varepsilon$ denotes the budget-widened support. No intervention adds answers beyond this union.

The two resources enter through different routes. Content that enters through sampling is learned at a rate proportional to its support. Content that enters through supervised likelihood is written as item bindings. Content that enters through activations participates in the base model's full computation. These routes are not interchangeable: they differ in what they can carry, how reliably they place it, and whether their reach depends on the base model's existing support.

Carriage and placement are independent conditions, and together they give a four-way protocol taxonomy (context/content injection, SFT, DPO, RLVR/steering; Table 2, appendix). Context and content injection carry and place via activations. Supervised fine-tuning (SFT) carries and places into weights. DPO carries but does not place Rafailov et al., 2023; Pal et al., 2024. RLVR and steering neither carry nor place. The reach of each row follows from which resource it uses.

2.3 PROVABLE CORE: SUPPORT-BOUNDED LEARNING

The provable core concerns on-policy reward-only updates with deterministic optimizers.

Assumption 1 (Item-symmetric policy class)**.** *The policy class cannot express per-item answer preferences beyond those induced by the marginal distribution. The assumption is exact for uninformative prompts and degrades to prompt-conditioned marginal overlap otherwise (tested in §3.6).*

Theorem 1 (Support-Bounded Learning)**.** *Consider GRPO with binary rewards and a leave-one-out baseline, $A_i = r_i - \frac{1}{G-1}\sum_{j\neq i} r_j$; updates are deterministic functions of sampled trajectories and rewards. Let $\rho(k)$ denote the probability that the correct answer is sampled for item k at initialization. The theorem has three parts.*

(i) Zero-support indistinguishability. *On the event that a^* is never sampled, the observed data is identically distributed for any two unsampled candidates. Hence no deterministic update can distinguish among them, and the trained policy's accuracy on such items is bounded by the largest posterior mass among never-sampled candidates.*

(ii) Rate law. *Writing ρ for the hit probability of a^* under the current policy (Definition 2's ρ at initialization; $\mathbb{E}_{\text{miss}}$ averages over samples other than a^*), the expected advantage of a hit is $1-\rho$ and of a miss $-\rho$, so $\mathbb{E}[\nabla\mathcal{J}] = \rho(1-\rho)\left[\nabla \log \pi(a^*|q) - \mathbb{E}_{\text{miss}}\nabla \log \pi\right]$. Separately, a group of size $G \geq 2$ is* informed *(at least one hit and one miss) with probability $1-(1-\rho)^G-\rho^G$. The gradient factor and the informed-group probability both vanish linearly in ρ (self-inclusive baselines: all advantages $\times(G-1)/G$; Appendix C).*

(iii) Marginal matching. *Under Assumption 1, the unique context-free maximizer of expected log-likelihood is the training marginal P_{train}; a context-free policy with answer distribution p attains (stochastic evaluation) $\sum_c p(c)\,P_{\text{test}}(c)$. Hence the marginal-matching policy attains the overlap $\sum_c P_{\text{train}}(c)\,P_{\text{test}}(c)$, while the context-free supremum is $\max_c P_{\text{test}}(c)$: the overlap ceilings likelihood-driven learning, not all context-free policies.*

Corollary 1 (Finite budget)**.** *If $\rho(a^*|q) \leq \rho_0$, no sample hits in n rollouts with probability at least $1-n\rho_0$, and the item is blind for the entire run with that probability; the expected rollouts to the first hit is $1/\rho$, so at budget n the effectively supported answers are $\{a : \rho(a) \gtrsim 1/n\}$.*

The zero-support and rate-law parts follow directly from the reward structure, and the marginal-matching part adds Assumption 1 (proofs in Appendix C). Corollary 1 is what the experiments measure: support means support at the stated sampling budget, with exact support as the $n \to \infty$ limit. GRPO test accuracy lands at the ceiling predicted by the marginal-matching result, and rejection-sampling fine-tuning reproduces the same ceiling, confirming the ceiling is not optimizer-specific.

2.4 CHANNEL PROPOSITIONS (P1–P4)

The inheritance law takes different forms for different channel families.

Proposition 1 (P1: direction channels are band-limited tilting)**.** *Direction-based interventions perturb activations by a bounded amount, so they can only change answers whose deficit lies within a narrow band. Items outside this band remain unreachable regardless of intervention strength. Moreover, for premise-type content, the readout signal is uncorrelated with the answer direction, so even in-band items cannot be reached (*P1$^+$*).*

Proposition 2 (P2: content channels superpose the stored set)**.** *A content channel writes a set of token embeddings, which we call the* stored set*. Injectable content lies in the union of this stored set and the prior support, independent of ρ. An injected token acts as a virtual in-context token, so composition with prior capabilities is nearly free. At single-item granularity, information is carried by token identity: truncating attention to one token loses nothing.*

Proposition 3 (P3: RLVR is support-bounded reweighting)**.** *Reward-only reinforcement learning reweights answers that the model already samples. The gradient expectation is proportional to $\rho(1-\rho)$, so zero support yields a pathwise zero gradient. With sparse support, learning recovers the answer marginal only, attaining the overlap under Assumption 1.*

Proposition 4 (P4: gradient writes implant spikes and drain neighbors)**.** *Gradient-based writing installs form-robust, bearer-specific bindings. Transfer to new bearers requires rule-type signal,*

which arbitrary facts lack. The same cross-entropy objective applies a probability-proportional negative drift to unwritten neighbors; the net effect is non-monotone in similarity (§3.2).

2.5 THE CHAIN-RULE DECOMPOSITION OF SUPPORT

The inheritance law says reach depends on the base model's sampling support, but not where that support comes from. We decompose it by treating the answer mode m (country, city, occupation, ...) as a latent variable. For a question q and answer a, the base conditional factorizes as

$$\rho(a|q) = \sum_m P(m|q) \cdot P(a|m, q). \tag{1}$$

The first factor, *mode mass*, is set by the question form: it says how much probability the model assigns to the relevant answer type. The second factor, *within-mode prior frequency*, is set by pretraining frequency and can be estimated externally, for example from Wikipedia. More generally, m can be any latent identifiable from the question alone whose groups predict answer frequency.

This decomposition connects two lines of work. On the storage side, capacity scales with exposure frequency (Allen-Zhu and Li, 2025; Lu et al., 2024; Feldman, 2020). On the extraction side, memory succeeds on popular entities and fails on the tail (Mallen et al., 2023a). The chain rule unifies them: *support follows the answer's own frequency, not the subject's fame*. Obscure-subject failures are therefore binding failures: the model knows the type of fact but not this instance.

On this frequency-structured prior, each channel acts as an operator. Direction-based tilting spends only a small fraction of its budget on the answer direction, so it changes probabilities within a narrow range. Reward-only reweighting conserves probability mass and is bounded by the overlap ceiling. Gradient writing installs item bindings while draining neighbors in proportion to their baseline mass, and it also sharpens the mode globally. Content injection passes through the base model's own readout, so injection is the bottleneck and composition is nearly free. These operator effects are measured in §3, and a logit-lens probe locates the closure stratum in network depth (Appendix B).

3 EXPERIMENTS

Each claim below is adjudicated by a dedicated control or countercheck, and training arms use three seeds. The base model is Qwen3-8B (bf16), with constants replicated on Qwen3-14B. Throughout, $\pm$ is the seed standard deviation, pooled proportions carry Wilson 95% CIs, and per-seed values are in the released log. We use two evidence bases. The synthetic beds isolate mechanism, following the synthetic-biography methodology of Allen-Zhu and Li (2025); their marginals are known by construction, with first-name-level splits and verified zero leakage (Appendix B). ReachBench then demonstrates the framework on natural public data (§3.6).

3.1 MEASUREMENT AND AUDITING

Testing the theory requires four instruments that separate what a channel writes from what the base model already supports. The *no-context identity audit* requires that any channel with learned components, when run with its memory cleared, reproduce the bare base; it catches distillation leakage in modulation adapters and extends to external methods (§4). A *three-stage probe cascade* localizes failures to write separability, retrieval, or read-value conductance, and paired oracle bounds separate access from readout. *Prior-support decomposition* assigns items to layers (Definition 3) and certifies the absence of item knowledge. Finally, four *testbed disciplines* guard the testbeds: fragment-level name partitioning, per-item distractors, exchangeability checking, and marginal shortcut auditing.

3.2 CHANNELS AS OPERATORS: BAND, INJECTION, BINDINGS

The direction channel requires no learned content, so adjudication starts there. Steering helps an item only when the base model is already close to the answer, and this band is narrow: increasing the steering strength barely moves it, and targeted interventions fail to widen it (Fig. 4, appendix). The band also cannot deliver new content. On derived-closure items, the direction channel reaches only a small fraction of what the content channel reaches. A control explains the gap: steering with

Table 1: Channel×layer reach matrix. In-band: steering helped-rate among in-band items (Fig. 4); row (c): premise-supplied context. Injection on the leak-free subset (Fig. 7); SFT row: train/paraphrase/test; RLVR: pass@64.

	In-band	Display	Closure	Underd.
(a) direction	32%@band	.013	.027±.010	≈0
(b) content	n/a	.333±.036	.311±.029	.33
(c) context	∼1	1.000	1.000	∼1
(d) SFT (closed-book)	n/a	1.00/.89/flat	.97/.30/flat	.94/.67/flat
(e) RLVR	0	0	.16@64	0

a *wrong* answer's direction helps almost as much as steering with the right one, so the apparent gains are variance compression rather than new content. Whatever effect the direction channel has stays inside the band, consistent with P1 and P1$^+$.

The content channel shows the opposite pattern. Supervised on where to read, it injects premise-type facts into unseen sessions across all seeds. Its readout concentrates essentially all attention mass on a single discriminative answer token (Fig. 5, appendix), explaining why attention truncation loses nothing and values transfer zero-shot. Injection is the bottleneck; composition is free, consistent with P2.

Gradient writes have the largest reach and the largest side effects, consistent with P4. Closed-book training fits trained items nearly perfectly and survives paraphrase across beds (row (d) of Table 1), so the bindings are form-robust. However, it does not transfer to new bearers, and it actively suppresses neighbors: after writing a city binding, the derived gold country often falls below distractors, non-monotonically in answer similarity, with drain depth proportional to baseline mass. Final-layer placement installs stronger bindings with heavier suppression, and high-budget weight writing degrades the in-context read path itself. Detailed numbers are in Table 1 and Fig. 4.

3.3 RLVR AND THE REACH MATRIX

Reward-only learning is where the theory has the most direct consequences, because practice relies on it most. We first examine the density of sampling support. The country form of our testbed question yields nonzero pass@64, but support is concentrated on a minority of items; the city form and novel strings produce zero hits across tens of thousands of samples. Support is therefore a property of the question form and answer type, not of training effort. GRPO (Shao et al., 2024) then behaves as Theorem 1 predicts: its signal fraction tracks measured support across cells. Under an extended budget, pooled test accuracy rises to just below the ceiling computed a priori from the split alone, with no fitted parameters and no training run, and the 14B base saturates in the same way. Learning exists, is item-blind, and stops where the answer marginal dictates (Fig. 6, appendix).

Table 1 assembles the channel×layer reach measured above. The remaining subsections test the matrix's boundary conditions, audit external methods against it, and extend the framework to natural public data.

3.4 BOUNDARIES AND THE POST-TRAINING BUDGET

Without changing the pretraining distribution, how much can post-training enhance a frozen model? (E-S2)[1] Table 7 (appendix) decomposes the base-to-oracle gap on closure items, from bare accuracy to premise-supplied accuracy. The best channel closes only a minority of the gap; the remaining majority is left unclosed by every tested channel, and closing it requires pretraining change or permanently present content. The RLVR share is predictable before training, and on the underdetermined layer it is zero at the measurement budget (Theorem 1).

The theory's boundary conditions are sharp, and each is quantified in the released log. The naive band model, the product of margin sensitivity and modulation budget, overpredicts the band by an order of magnitude and also overestimates sensitivity. The correct decomposition, band = budget × sensitivity × alignment, exposes a stable cross-model constant: alignment is nearly identical on 8B

[1]E-tags are pre-registration keys in the released experiment log.

and 14B, so roughly 95% of steering budget is spent in directions orthogonal to the decision margin (E-T1).

The remaining boundary results are negative. The strongest curriculum escape fails: adding hints to prompts makes nonzero reward appear, but after fading, reward and test accuracy return to zero, and closure items are slightly hurt (E-T2). DPO departs from SFT in both of its forms: budget-matched, it fits underdetermined items far less well than SFT, and its chosen-answer margin rises early and then collapses. Reintroducing a chosen-NLL term (Pal et al., 2024; Lambert et al., 2024) lifts train fit only weakly, with test accuracy still low (E-T3/T3b). Rejection-sampling fine-tuning shares GRPO's reach as predicted, and any protocol that samples its own training data inherits the sampling ceiling (E-T4). Finally, sampling at high temperature reduces support to trace levels without enabling learning at any practical budget (E-T5).

3.5 Chain-rule tests and public-benchmark validation

The chain rule's operators hold with the predicted signs. Mode mass is nonzero for the country form and measured zero for the city form (E-L1). Hit items' gold answers have higher prior frequency than miss items'. Trained items' derived answers show a two-component operator, local drain plus global mode sharpening, with the net sign flipping across baseline (E-L2).

The chain rule is also predictive: fitting on half the items and predicting the other half gives a positive rank correlation (E-P1). On PopQA (Mallen et al., 2023a), ρ correlates with *answer* popularity across all three bases, and the correlation survives controlling for subject popularity (E-P2); ReachBench upgrades both to prospective prediction. In a long run, GRPO saturates below the ceiling (E-S1; Fig. 8), and a ROME (Meng et al., 2022a) editing baseline reproduces the SFT signature (E-S3).

Finally, chain-rule-predicted support matches realized per-item gradient signal, and the bottom predicted quartile receives zero signal; the causal chain from frequency to support to signal closes.

3.6 ReachBench: the framework on natural public data

We assemble *ReachBench* from three public sources: EntityQuestions (Sciavolino et al., 2021), which supplies entity-centric questions over Wikidata relations; PopQA (Mallen et al., 2023a) (answer-entity pageviews as the external frequency axis); and FreshQA (Vu et al., 2024) (post-cutoff facts, year ≥ 2025, as the underdetermined pool). Every item gets a mode, an entity, and an entity-disjoint train/test split with zero overlap verified. Its layer is assigned *by measurement*, never by source label. On a stratified sample of Qwen3-8B, display, closure, and underdetermined items all appear in substantial proportions. Of the post-cutoff items, most have zero support at the measurement budget. Mean ρ rises across pageview quintiles: the chain rule on natural data.

The first test is prediction rather than fitting. We fit the within-mode frequency relation on *train* entities and predict the measured support of held-out *test* entities. The prediction achieves a positive rank correlation across held-out test items, with predicted-quintile means fully monotone; this is stronger than on our synthetic beds. The structure replicates across families: Llama-3.1-8B, GLM-4-9B, and DeepSeek-R1-Distill-7B agree with the Qwen assignment on most items, and layer assignment is invariant to question paraphrase. DeepSeek's reasoning-distilled variant shifts items from display to closure while preserving the structure. The attainable-rate ordering replicates on every family, including the 2026-generation Qwen3.5-9B under a thinking-adapted window (Table 4, appendix).

The law's rate prediction can be taken on-policy. We train GRPO on a mixed-layer natural pool. The per-item reward signal tracks measured ρ across seeds (Fig. 3c), and most signaled items had measured $\rho > 0$. The signaled minority with measured-zero support is the measurement floor, whose rate scales down with per-item training samples as predicted, and the FreshQA arm verifies at training time: its measured-zero items produce zero signal in every run. No layer shows reliable test gains: zero-support layers stay at the floor, and support does not transfer across entities.

What does the input budget change? Revealing a closure item's premise token by token traces the answer-hit spectrum of Fig. 2: hit rate stays flat until the answer token, then jumps sharply. Across Qwen, Llama, GLM, and DeepSeek-R1, the largest single-token transition almost always lands on the answer token or its immediate neighbor, while a scrambled premise of the same length recovers

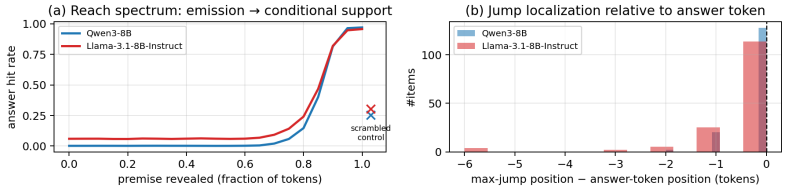


Figure 2: The reach spectrum (Qwen3-8B, Llama-3.1-8B; 150 closure items each). (a) Hit rate as the premise is revealed token by token; $\times$ marks the scrambled-premise control. (b) Max-jump position relative to the answer token (99%/93% within one token).

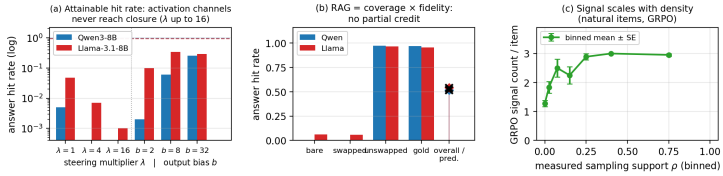


Figure 3: Three laws verified on natural items (a, b: Qwen3-8B, Llama-3.1-8B). (a) Attainable hit rate under steering λ and output bias b; dashed = premise input. (b) RAG under forced misses (53%): diamond = observed, $\times$ = mixture prediction. (c) Per-item GRPO signal vs. measured ρ; the ρ=0 bin is the k=64 floor.

only a fraction of it. Reach over the input channel is therefore determined by single discriminative tokens.

Other channels can attempt the same items (Fig. 3a), but they remain bounded. An output bias on the answer logits, the strongest answer-only calibration adjustment, recovers about a third of the items at its peak and declines beyond it, because the continuation must still come from the model. Mid-layer steering stays near zero at every magnitude tested, even at many times the contrastive strength, and no activation budget attains closure.

Retrieval follows the same decomposition and is robust to document noise. With retrieval misses forced at low coverage, accuracy matches the mixture prediction $\text{cov} \cdot v + (1 - \text{cov}) \cdot \rho$. Paraphrased documents that preserve the answer lose little, while documents whose discriminative token is masked return the item to bare performance. Detailed numbers appear in Fig. 3.

How much can prompt informativeness substitute for content? Constraint-style hints leave support nearly unchanged: a type name adds nothing, a first-letter hint lifts closure items only slightly, and a candidate list containing the answer still leaves hit rate near zero. The choice follows prior preference, not the binding. Few-shot in-context examples behave the same way: same-mode exemplars teach the task format, not the answer.

Under letter hints, GRPO learning tracks hint-conditioned support, and items still at zero support stay blind. The rate law and ceiling therefore pass to the prompt-conditioned overlap unchanged. Detailed numbers appear in the released log.

The mainstream methods land on the predicted walls. Inference-time intervention (ITI; Li et al., 2023) with a contrastive probe direction recovers almost nothing at any strength tested: the canonical steering method stays inside the direction band, so P1 is not specific to our formulation. Dense retrieval (embedding-similarity scoring instead of BM25) gives the same decomposition. Free-form generation (no choices shown) reproduces the layers: bare accuracy is near zero, premise-supplied accuracy is near one, and closure support stays negligible under top-p and top-k decoding.

4 External Corroboration

The observations in this section predate our theory; each matches a predicted limit. We first extend the NOCTX audit (§3.1) beyond our own channels, to δ-mem (Lei et al., 2026), a published delta-rule fast-weights adapter, and to BM25 retrieval, on the Qwen3-4B-Instruct-2507 base under our

4-choice protocol (LME/LOCOMO (Wu et al., 2025; Maharana et al., 2024)). Our content channel, run on LME under the same protocol, gains end-task accuracy while keeping a zero no-context gain. Full numbers are in Table 3 (appendix).

The audit separates the audited methods cleanly (Table 3). With its memory cleared, δ-mem flips zero predictions, the identity property our content channel was designed around. Whether prior contamination appears is therefore a property of the write path, not of the "parametric memory" label. Passing the audit, however, is not the same as helping. δ-mem's net gain under our instrument is zero on one dataset and within seed noise on the other. All of its fixes occur on low-margin errors, where the band law predicts small gains. Retrieval behaves differently: its gain flips sign across datasets, and when retrieval misses, the item returns to ρ (Fig. 3b). A no-context decomposition is a cheap audit, one pass per method, separating these cases.

P3 predicts a graded structure from sparse to dense support, and both sides of the RLVR boundary debate report it: the sparse symptom, where pass@1 rises while large-k coverage falls (Chen et al., 2026); the dense-rung extension theorem (Wen et al., 2026); the industrial sparse rung (Zhou, 2026); and the granularity/coverage decomposition (Clay et al., 2026). The KL-regularized optimum is an exponential *tilt* of the reference policy (GX-Chen et al., 2026) and preserves support at convergence; Theorem 1 extends support-boundedness of the learning signal to every finite-time update.

The same limits appear across editing, agents, steering, and injection. The layers anchor their literatures: elicitation for display (Kadavath et al., 2022; Burns et al., 2023; Mallen et al., 2023b), agent memory for closure (Packer et al., 2023), and editing, fine-tuning, and retrieval for underdetermined content (Meng et al., 2022b; Gekhman et al., 2024; Lewis et al., 2020; Khandelwal et al., 2020). Parameter editing breaks on multi-hop composition (Zhong et al., 2023), and edits ripple into related facts (Cohen et al., 2024). MemoryAgentBench's per-skill ceilings read as our matrix rows at industrial scale (Hu et al., 2026). Steering spans the formal precedent (Petrov et al., 2024), the canonical readout (Li et al., 2023), and automation (Cui and Chen, 2025); P1$^+$ reads any steering-based fact injection as prior-internal rebinding, not newly placed content. REMEDI (Hernandez et al., 2023) establishes content-shaped writing, parametric RAG (Su et al., 2025) routes retrieval into parameters, and AnyEdit's (Jiang et al., 2025) single-token barrier corroborates the injection bottleneck.

5 Practical Implications

The three measurements (ρ at 64 samples, premise probes, and the decomposition) are a fixed per-item compute budget. On the closure testbed, a frozen rule (content routes iff $v \geq 0.75$; never steer) picks the measured-best channel for almost every item, with small mean regret. Display-layer items (their majority-of-k sits above pass@1, as the layer's definition implies) need only elicitation: steering and reweighting suffice. Closure-layer items ($\rho \approx 0$, verification high) are where reward-only learning is most effective, reachable by retrieval and injection and closed to steering. Underdetermined items admit only content-shaped routes, retrieval or context (Table 5, appendix).

We recommend keeping session facts external, read out in content form the base can compose on. Run the support test before training; route zero-density items to retrieval; order closure items by density, also the learning-speed order. Both RL internalization and weight-writing fail on external facts (E-T2; Theorem 1); the reliable editing space is display-layer rebinding plus retrieval.

6 Limitations

Each arm uses one base (Qwen3-8B), with replication on four additional models, three outside the Qwen lineage (Table 4); the constants are base-specific. Assumptions: P1 local linearity; P1$^+$ readout–direction uncorrelation; P3 uninformative prompts; P4 first-order drift. The synthetic beds isolate mechanism; ReachBench tests natural data; caveat: the measurement floor; layer assignment is threshold-insensitive. Our GRPO runs use LoRA and short schedules; extrapolation rests on mechanism, not scale. What post-training can add to a frozen model is knowable in advance.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=FxNNiUgtfa.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2026.

Donovan Clay, Saket Gollapudi, Sankar Harilal, Min Jang, Jacob Morrison, Sewoong Oh, and Natasha Jaques. Demystifying reinforcement learning post-training of language models. *arXiv preprint arXiv:2608.24949*, 2026.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024.

Sasha Cui and Zhongren Chen. Painless activation steering: An automated, lightweight approach for post-training large language models. *arXiv preprint arXiv:2509.22739*, 2025.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In Konstantin Makarychev, Yury Makarychev, Madhur Tulsiani, Gautam Kamath, and Julia Chuzhoy, editors, *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*, pages 954–959. ACM, 2020. doi: 10.1145/3357713.3384290. URL https://doi.org/10.1145/3357713.3384290.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7765–7784, 2024.

Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. Kl-regularized reinforcement learning for generative modelling is designed to mode collapse. In *International Conference on Learning Representations*, volume 2026, pages 50665–50693, 2026.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

Geert Heyman and Frederik Vandeputte. Steer like the llm: Activation steering that mimics prompting. *arXiv preprint arXiv:2605.03907*, 2026.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pages 156259–156291, 2026.

Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.

Minjae Kang and Jaehyung Kim. Enhancing instruction following of LLMs via activation steering with dynamic rejection. *arXiv preprint arXiv:2603.06745*, 2026.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Jingdi Lei, Di Zhang, Junxian Li, Weida Wang, Kaixuan Fan, Xiang Liu, Qihan Liu, Xiaoteng Ma, Baian Chen, and Soujanya Poria. δ-mem: Efficient online memory for large language models. *arXiv preprint arXiv:2605.12357*, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Xingyu Lu, Xiaonan Li, Qinyuan Cheng, Kai Ding, Xuanjing Huang, and Xipeng Qiu. Scaling laws for fact memorization of large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 11263–11282. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.658. URL https://doi.org/10.18653/v1/2024.findings-emnlp.658.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9802–9822, 2023a.

Alex Mallen, Madeline Brumley, Julia Kharchenko, and Nora Belrose. Eliciting latent knowledge from quirky language models. *arXiv preprint arXiv:2312.01037*, 2023b.

Maxime Méloux and Christophe Cerisara. Novel-WD: Exploring acquisition of novel world knowledge in LLMs using prefix-tuning. *arXiv preprint arXiv:2408.17070*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022a.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 11048–11064, 2022.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 237–250, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with DPO-positive. *arXiv preprint arXiv:2402.13228*, 2024.

Aleksandar Petrov, Philip Torr, and Adel Bibi. When do prompting and prefix-tuning work? a theory of capabilities and limitations. In *International Conference on Learning Representations*, volume 2024, pages 6031–6054, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, 2021.

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

Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric retrieval augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1240–1250, 2025.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, 2024.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *International Conference on Learning Representations*, volume 2026, pages 49450–49483, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=pZiyCaVuti`.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

Shaochen Zhong, Yifan Louie Lu, Lize Shao, Bhargav Bhushanam, Xiaocong Du, Yixin Wan, Yucheng Shi, Daochen Zha, Yiwei Wang, Ninghao Liu, et al. MQuAKE-remastered: Multi-hop knowledge editing can only be advanced with reliable evaluations. In *International Conference on Learning Representations*, volume 2025, pages 99708–99736, 2025.

Zexuan Zhong, Zhengxuan Wu, Christopher D Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, 2023.

Todd Zhou. When RLVR shrinks the reasoning boundary: Diagnosing pass@ k inversion. *arXiv preprint arXiv:2607.20543*, 2026.

Table 2: Protocol taxonomy under the inheritance law (§2.2).

Protocol	Carries	Places	Reach
Context, content inj.	yes	yes (activ.)	stored set ∪ prior
Supervised FT	yes	yes (weights)	item bindings
DPO-style pref.	yes	no	ratios only
RLVR, steering	no	no	support / band

Table 3: No-context decomposition (§4). Δ = gain over the bare base; Δ_{noctx} = gain with memory emptied. δ-mem/RAG on Qwen3-4B-Instruct-2507, ours on Qwen3-8B.

Method (write path)	Dataset	Δ	Δ_{noctx}
modulation-with-distillation adapter (ours; dev)	LME	$+13.1$pp	$+13.1$pp
content channel (ours)	LME	$+3.0$pp	0.0pp
δ-mem (published; delta-rule)	LME / LOCOMO	0.0 / $+0.5$pp	0.0 / 0.0
BM25 RAG (retrieval)	LME / LOCOMO	$+5.3$ / -1.5pp	0.0

Table 4: Layer structure and attainable rates across five models (same items, same protocol). Disp./Clos./Und.: layer shares (%); Agr.: agreement with Qwen3-8B; Jump: spectrum gate on or next to the answer token; Prem./Bias pk./Steer max.: premise hit, peak output bias, best steering. †Qwen3.5-9B: thinking-adapted window (64 tokens, think-blocks stripped); its depressed premise hit is protocol-relative.

Model	Disp.	Clos.	Und.	Agr.	Jump	Prem.	Bias pk.	Steer max.
Qwen3-8B	46	29	25	100	99	.970	.311	.005
Llama-3.1-8B	60	22	19	75.5	93	.956	.343	.076
GLM-4-9B	50	24	26	76.1	95	.888	.387	.029
DeepSeek-R1-Distill-7B	31	40	29	74.0	100	.881	.203	.003
Qwen3.5-9B†	60	14	26	70.9	90	.296	.292	.015

Table 5: Diagnose-then-channel decision table (§5).

Diagnosis	Layer	Usable	Closed
In-band, maj@k≫pass@1	display	steering, reweighting	—
$\rho \approx 0$, verify high	closure	RLVR (sparse), retr., inj.	steering
$\rho = v = 0$	underd.	retrieval, context; SFT binds exact items only	RLVR, steering

Table 6: Adjudication log (selection; full log released).

Prediction	Adjudicating control	Refined view
Injection rises with attention rank	top-1 = full (0.290 vs. 0.295)	token identity carries
Undermined SFT collapses paraphrase	paraphrase survives (+30pp)	bindings form-robust, layer-blind
Display-layer items have positive density	0 hits in 20k samples	ρ set by (form, type)
Direction channel gains on display bed	0.013, all items out of band	read by deficit, not label
Band from budget × sensitivity	off by 10×	alignment constant 0.045
Naive hint curriculum	reward dies at fade	static reading holds
DPO behaves like SFT	fits 0.068 vs. 0.944	carries, does not place
DPO-positive restores placement	train 0.111, margins unstable (E-T3b)	placement is graded, soft cell boundary

Table 7: Post-training enhancement budget (closure layer; free-form scoring; base $= 0.013$, oracle $= 1.000$, gap $= 0.987$; Budget $=$ (reach$-$base)/gap).

Channel	Reach	Budget	Mechanism
direction	.027	1.4%	band-limited tilting
content inj.	.311	30.2%	stored-set superposition
SFT (test)	flat	$\approx$0	item bindings, no transfer
RLVR	.058	4.6%	support reweighting
RLVR ceiling	.0647	5.2%	marginal overlap
Unclosed by tested channels	—	*70%*	

Table 8: Lexicon. Each construct is measured, not assumed; the last column names the instrument.

Term	Symbol	Meaning	Measured by
Sampling support	$\rho(k)$	P(one T=1 sample hits a)	pass@k density
Verifiability	$v(k)$	confirm rate given premise	premise probes
Deficit	$d(k)$	gold $-$ best distractor score	4-choice scoring
Band	δ	max bounded activation shift	budget$\times$sens.$\times$align.
Stored set	B	token embeddings stored at write	write/read cascade
Carriage / placement	—	signal contains / raises $p(a)$	taxonomy (Tab. 2)
Marginal floor	—	leave-one-out prior-only accuracy	prior-support decomp.
Item lift	—	residual above floor	prior-support decomp.
Mode	m	shared answer form (e.g. country)	chain-rule fit
Drain	—	neighbor suppression under CE	never-reinforced arm

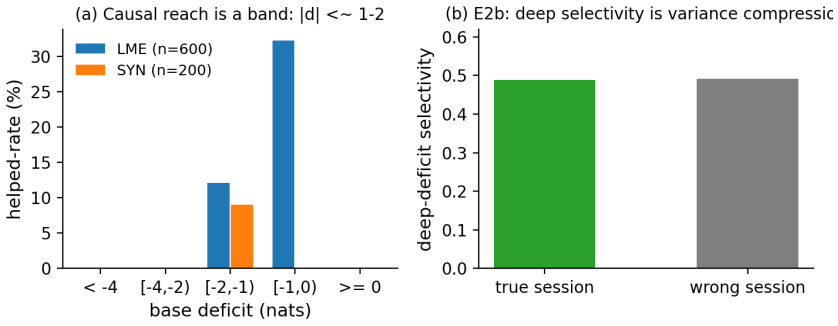


Figure 4: Direction channel (§3.2). **Left:** helped-rate vs. base deficit: reach is a band. **Right:** wrong-session control: identical selectivity (variance compression).

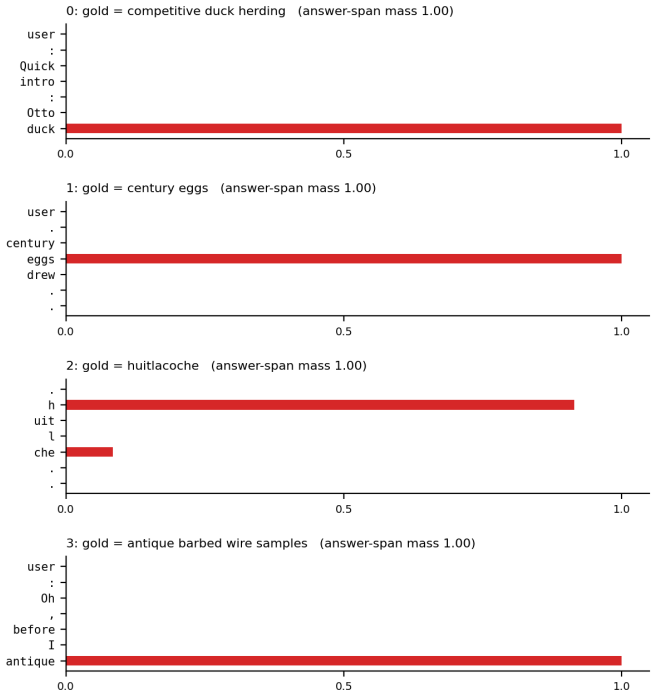


Figure 5: Readout attention on unseen sessions (content channel, §3.2): mass 0.999–1.0 on one discriminative answer token.

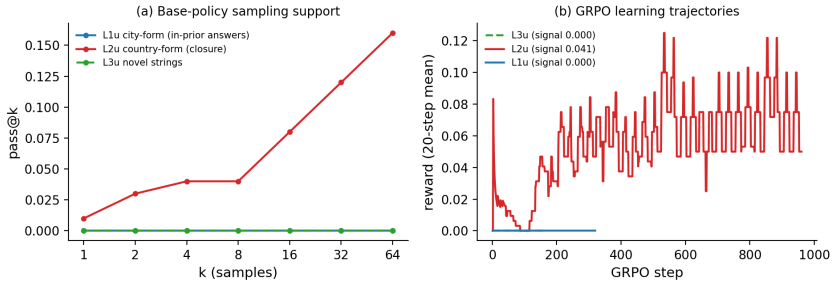


Figure 6: Sampling support (pass@k, left) and GRPO trajectories (right) on the RLVR arm (§3.3): signal fraction tracks density cell by cell.

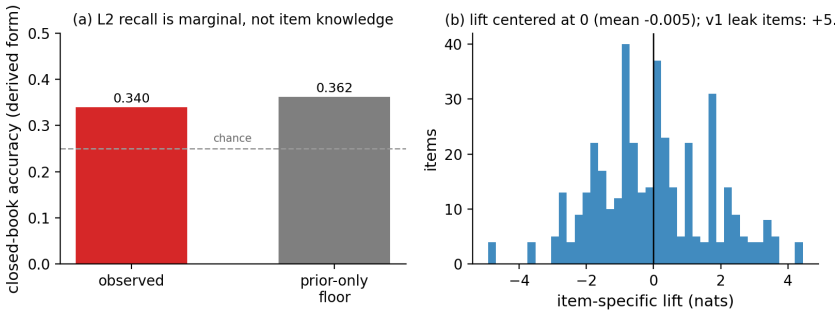


Figure 7: Prior-support decomposition (E-L1): **left**, accuracy below the prior-only floor; **right**, item lift centered at zero; five earlier leak items at $+5.5$ nats.

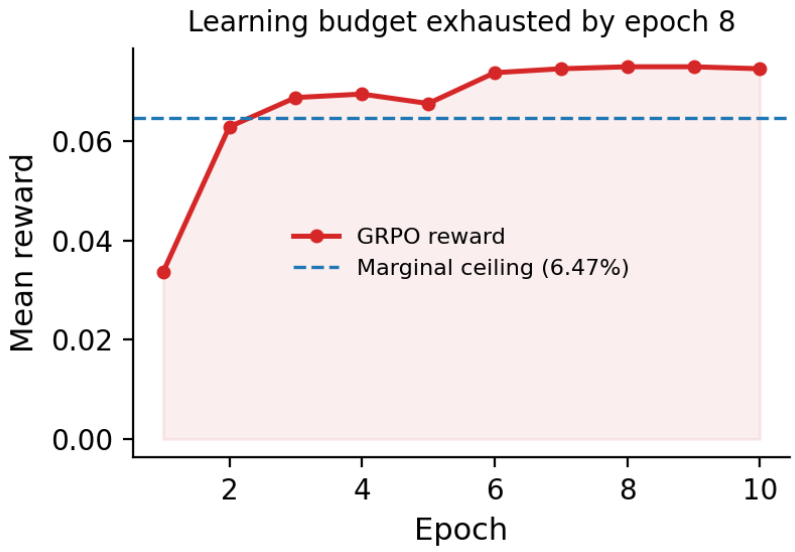


Figure 8: GRPO saturation (E-S1): reward approaches its asymptote by epoch 8 and flattens; test accuracy stays below the marginal ceiling.

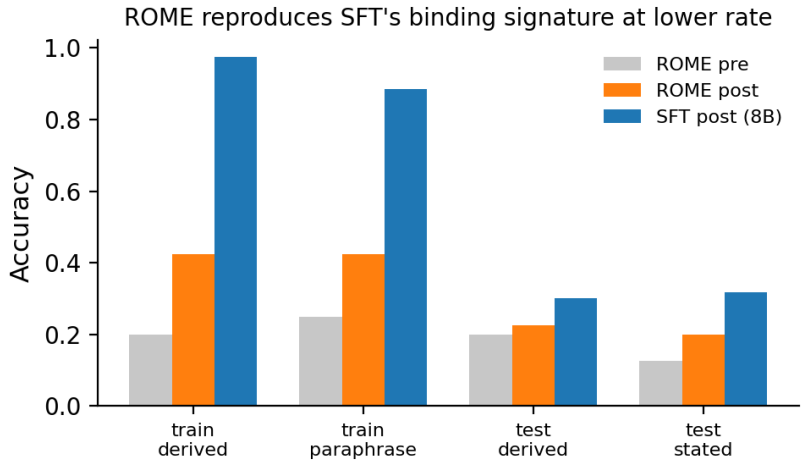


Figure 9: ROME vs. SFT structural comparison (E-S3): same binding signature (train rise, paraphrase robust, test flat), lower injection rate.

A Supplementary tables and figures

B Corroboration details and additional channel evidence

Testbeds (full specification). Two deterministic synthetic families (no RNG; index arithmetic). *Repeated-name*: 18 names, 30 cities, with marginal structure, for marginal-learning and suppression evidence. *Unique-pair*: one fact per unique full name, first-name-level partitioning across the split (zero leak, verified). Item classes mirror the introduction: display (current city), closure (stated birth city, derived country question; verification 1.000, recall at the floor), and underdetermined (novel facts).

Knowledge editing. Parameter-side multi-hop collapse is the field's earliest wall (GPT-J $40.5\% \rightarrow 7\%$ after editing; Zhong et al., 2023); MQuAKE Remastered (Zhong et al., 2025) audited 33–76% label corruption, echoing our testbed disciplines; 2026's high scores route through agents or retrieval (the context channel), while the parameter side stays inside row (d) of our matrix.

Memory agents. MemoryAgentBench (Hu et al., 2026): retrieval 71.8 (read side; our controlled injection rate 0.311, Table 1), test-time learning 53.9 (write-side wall), selective forgetting at most 7.0 (drain), and commercial memory agents below their own backbone (21.1 vs. 42.2): the gradient channel's dose effect (§3.2) at industrial scale.

Direction- and content-shaped interventions. In-context learning analyses (Xie et al., 2021; Min et al., 2022) support the elicitation reading of the display layer; 2026 work moved to effectiveness conditions (Heyman and Vandeputte, 2026; Kang and Kim, 2026). Novel-WD (Méloux and Cerisara, 2024) stores new facts in long free prefixes (stored-set capacity, outside the bounded-direction qualifier); channel comparisons remain rare (Ovadia et al., 2024), part of the gap our factorial fills; hallucination analyses (Kalai et al., 2025; Xu et al., 2024) bound the prior from the other side.

E-S3 (ROME editing baseline). Forty sequential rank-one MLP edits at layer 18 (unique-pair testbed, closed-book 4-choice, Qwen3-4B-Instruct-2507): train rises to 0.425 (SFT on the same base: 0.974) with paraphrase robustness, and test stays near its base (0.225 edited vs. 0.200 unedited; Fig. 9; matrix row (d) is the 8B base). ROME reproduces SFT's structural signature at lower injection rate: P4 is a property of gradient writes generally, not of SFT specifically. The 40-edit sequence also reproduces the progressive accuracy loss characteristic of sequential editing collapse, connecting the probability-proportional drift to forgetting curves at edit scale.

Depth anchor (E-L3). A logit-lens probe on the closure testbed (Qwen3-8B) reads answer-token mass at each depth: country-form questions accumulate 31.7% of mass on country tokens by layer 32, and the final layers dissipate it to 1.0%, close to the measured closure sampling rate (1.3%, Table 7).

E-S4 (channels on natural knowledge). PopQA stratified by measured support (4-choice accuracy, 64-sample budget); "zero-support" here means zero observed hits at the measurement budget, not true zero. GRPO raises the zero-support bin $0.55 \rightarrow 0.95$ and the mid bin $0.70 \rightarrow 0.80$ (display-layer items with sampling dispersion being sharpened); SFT degrades both bins to 0.000 (overfitting plus neighbor drain). The measurement floor ($1/k$) distinguishes "observed zero" from "true zero"; the theorem applies only to the latter.

C Proofs and Derivations

C.0 Notation. Autoregressive transformer M with L layers; h_l the answer-position representation at layer l; u_t the unembedding row of token t; J the Jacobian from layer l to logits with $\|J\| \leq J_{\max}$; policy π_θ.

C.1 Theorem 1 (proof). Part (i). On the event that a^* is never sampled, every observed reward is zero regardless of which unsampled candidate is correct, and no trajectory contains an unsampled candidate; hence the likelihood of the observed data is identical under $a^* = c_1$ and $a^* = c_2$ for

unsampled c_1, c_2. Since the update is a deterministic function of the data, the induced policy distribution is identical, so its accuracy at selecting among unsampled candidates equals guessing under the posterior, which is unchanged among unsampled candidates by the same symmetry.

Part (ii). With the leave-one-out baseline $A_i = r_i - \frac{1}{G-1}\sum_{j\neq i} r_j$, the other $G-1$ draws are independent of r_i, so the per-item expected advantage of a correct sample is $\mathbb{E}[A|a=a^*] = 1-\rho$ and of a miss is $\mathbb{E}[A|a\neq a^*] = -\rho$. The expected gradient is

$$\mathbb{E}[\nabla\mathcal{J}] = \rho(1-\rho)\,\nabla\log\pi(a^*|q) + (1-\rho)(-\rho)\,\mathbb{E}_{\text{miss}}[\nabla\log\pi] = \rho(1-\rho)\,[\nabla\log\pi(a^*|q) - \mathbb{E}_{\text{miss}}\nabla\log\pi].$$

With the self-inclusive mean $r_i - \bar{r}$, the baseline retains a $1/G$ fraction of r_i, and every conditional advantage is scaled by $(G-1)/G$: sign and ρ-dependence are unchanged. The probability that a group of size G is informed (nonzero advantage variance, i.e. at least one hit and at least one miss) is $1-(1-\rho)^G - \rho^G > 0$ iff $\rho \in (0,1)$; for small ρ this is $G\rho + O(\rho^2)$. The two quantities, expected advantage $\propto \rho(1-\rho)$ and informed-group probability $\approx G\rho$, vanish linearly in ρ; measured per-item signal counts track the latter.

Part (iii). Under Assumption 1 the policy class contains no per-item signal; by part (i), items with $\rho = 0$ contribute no item-specific gradient; items with $\rho > 0$ contribute gradients proportional to $\rho(1-\rho)$, but the only item-independent quantity the aggregate gradient can maximize is the marginal log-likelihood, whose optimum is P_{train}. For unseen items under the same prompts, accuracy is $\sum_c P_{\text{train}}(c)\cdot P_{\text{test}}(c)$ by construction.

Weight sharing induces item-blind normalization effects (marginal drift) but no item-level distinction. Protocols with process rewards or answer-bearing hints carry content and leave the theorem's premises.

C.2 Proposition P1. For $h_l \mapsto h_l + Wg(h_l)$ with $\text{rank}(W) \leq k$, $\|W\| \leq B$, $\|g\| \leq G$ and locally linear downstream, the first-token logit shift is $\langle u_t, JWg(h_l)\rangle + O(B^2)$, bounded by $J_{\max}\|u_t\|BG$; mean-scored multi-token answers divide the bound by answer length, giving the band. P1$^+$: when the correlation between g and $J^\top u_{\text{gold}}$ vanishes, the expected shift is zero at any deficit (second-order concentration of the direction cosine).

C.3 Proposition P2. With $h' = h + \sum_i \alpha_i(h)e_i$ over the stored set (e_i the stored-token embeddings, α_i the injection gates, h' the post-injection representation), the reachable answers lie in the union of the stored set and the prior support; treating the injection site as a virtual context token gives composition at context-level cost; token-identity carriage is an empirical proposition verified by the truncation experiment.

C.4 Proposition P3. With P_{train} and P_{test} the answer marginals over the 16-country candidate set (320 train items; 80 test items, evaluated as 240 item-seed pairs over three seeds), the marginal-matching policy attains $\sum_c P_{\text{train}}(c)\,P_{\text{test}}(c) = 6.47\%$, while the best fixed single guess (a degenerate context-free policy) attains $\max_c P_{\text{test}}(c) = 8.75\%$; no fitted parameter enters, and any train/test distribution shift rescales the sum directly. Per-country item counts (train/test), from which both numbers recompute: Poland 25/7, Austria 22/6, United Kingdom 22/6, Spain 18/6, Belgium 26/6, Japan 21/6, Norway 18/6, Sweden 18/6, Netherlands 21/5, Germany 19/5, Italy 20/5, Switzerland 24/5, New Zealand 24/4, Canada 11/3, France 21/2, Portugal 10/2.

C.4b Measurement protocol for ρ (§3.3). ρ is defined at $T{=}1$ and estimated with 64 samples (resolution $1/64$); items with zero hits in n samples carry the standard rule-of-three bound $\rho \leq 3/n$ at 95% (e.g. 1.5×10^{-4} for $n{=}$20k). E-T5 measures the temperature dependence (ρ contracts but stays strictly positive for reachable strings at $T{=}1.5$). Train-time signal uses the same $T{=}1$ sampler, so the ρ estimate and the learning prediction share one measurement protocol. For group-relative advantages $A_i = r_i - \bar{r}$ with binary rewards, $\mathbb{E}[\nabla\mathcal{J}] = (G{-}1)/G\cdot\rho(1-\rho)[\nabla\log\pi(a^*|q) - \mathbb{E}_{\text{miss}}\nabla\log\pi]$; the zero case is the pathwise identity of Theorem 1; the KL-regularized optimum is an exponential tilt toward the training marginal, whose stochastic-evaluation accuracy is the overlap.

C.5 Proposition P4. For cross-entropy on a target y, the first-order gradient step moves the logit of each non-target token t' by $-\eta\,p_\theta(t'|x)$: the drift is probability-proportional (largest for tokens similar to the target under softmax competition; net effect empirically non-monotone in similarity,

§3.2). Binding form-robustness follows from gradient penetration across the paraphrase class, and bearer-specificity from the absence of rule signal in arbitrary facts.

REPRODUCIBILITY STATEMENT

All experiments follow pre-registered protocols; the full adjudication log (Table 6) is released, and per-item records accompany the supplementary material. Testbed generators are deterministic (index arithmetic, no RNG); every measurement names its seed, sample budget, and extraction rule. Base models are public (Qwen3-8B, Qwen3-14B, Qwen3-4B-Instruct-2507, Llama-3.1-8B, GLM-4-9B, DeepSeek-R1-Distill-7B, Qwen3.5-9B); the public PopQA benchmark is used without modification for E-P2. Code, per-item records for every figure and table, and the full experiment log are released at the anonymized repository `https://anonymous.4open.science/r/reach-of-interventions-repro-C501`; the trained direction- and content-channel adapters (three seeds each) accompany them, and the RLVR arms regenerate from the released code with the recorded seeds.

AI-USE STATEMENT

AI assistants were used for literature search, experiment scripting, data analysis, drafting and editing of text, and assistance with proof exposition, under direct author supervision; all scientific claims, experimental designs, pre-registrations, mathematical content, and the final text were determined and verified by the authors.