WHEN THE RATIONALE DISAGREES WITH THE LOG: DIAGNOSING AND REPAIRING CHAIN-OF-THOUGHT FOR SEQUENTIAL RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought reasoning is critical to improve Large Language Models (LLM) capabilities in mathematics, code and logic, but has not be reliably successful in recommendation. We show that such a dichotomy is decided by a single quantity measurable *before* any reasoning supervision is trained, and we derive it in closed form. Because a decoder-only model emits the rationale and the answer through the same readout, the reasoning and ranking gradients satisfy an exact identity whose sign is governed by one indicator: whether the item the rationale argues for is the item that was logged. In mathematics the indicator equals one by construction and the two gradients reinforce. In *non-verifiable, multi-modal* prediction, where many continuations are legitimate but only one is recorded, it collapses to the overlap between the true continuation law and the teacher's, which we measure at 0.07–0.13 against a chance level of 0.05; the gradients then provably conflict. Three further measurements confirm the mechanism: the reasoning-ranking cosine turns negative exactly when predicted, deepens toward the layers at which the model commits to an answer, and leaves reasoning states nearly answer-agnostic under a linear probe. The identity is also constructive, naming token-level rationale imitation as the sole negative driver and predicting that removing it reverses the sign. It does: conflict-aware training (CARE) and outcome-reward reasoning restore answer decodability and turn chain-of-thought from a penalty into a gain over direct supervised fine-tuning. A corollary for practice is that the prevailing remedy, replacing discrete traces with latent ones, relocates the imitation target rather than removing it, which is why a change confined to the loss outperforms it. Reasoning is not the wrong tool for recommendation; imitating rationales is.

1 INTRODUCTION

Sequential recommendation predicts the next item a user will interact with from their history. Several properties of the task suggest that careful, step-by-step reasoning should help: intent is typically not stated, the plausible continuations are numerous, and the cue distinguishing the true next interaction from a near-miss is subtle and often compositional, combining a brand thread, a price band and a seasonal shift. Chain-of-thought (CoT), a foundational technique of LLM reasoning in mathematics, code and planning (Wei et al., 2022; Kojima et al., 2022), is a promising approach to making such deliberation to weigh competing hypotheses before committing to a recommendation decision (Pournemat et al., 2025; Hou et al., 2024; Bao et al., 2023; Lin et al., 2026; Anwaar et al., 2025; Li et al., 2026).

Such attempts have been met with mixed results. Independent studies report that textual CoT is neutral or harmful in this setting (Zhang et al., 2025; Li et al., 2025; Wang et al., 2025), and we observe similar results under controlled conditions (Tables 1 and 14). Additional computation, additional supervision and additional capacity all leave the model measurably worse. The cause has remained unexplained, and in its absence each proposed remedy rests on conjecture rather than on a diagnosis.

Figure 1 sets the two domains side by side and adds the quantity this paper argues is responsible. In both domains chain-of-thought is used in the same way: the model is trained and prompted to write a rationale before committing to an answer, and the comparison in each case is against an otherwise

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.

identical model that answers directly. The two domains are scored by different metrics, we therefore report the relative change with respect to learning to generate the answer without reasoning. The left column alone reproduces the puzzle: CoT yields $+30\%$ to $+273\%$ on mathematics and logic and results in a decrease of 7–11% on the four recommendation datasets of Table 1. The right column resolves it. It reports the cosine between the reasoning-token and answer-token gradients at epoch 1, the first epoch at which the answer head has specialized and the conflict has emerged (Appendix N): the two objectives are essentially orthogonal on the math tasks (-0.09 to $+0.06$) and strongly opposed on every recommendation dataset (-0.42 to -0.57). We draw from this the *sign* and the contrast between domains, not an ordering within either, since the per-dataset magnitudes are not separated by more than the probe's sampling error (Appendix N). The failure is therefore not diffuse but has a single measurable correlate, and the remainder of the paper derives that correlate in closed form, verifies it behaves as the derivation requires, and removes the term responsible.

Our position: the bug is in the loss, not in the trace. The prevailing response has been to change the *form* the reasoning trace takes, replacing discrete tokens with latent alternatives on the premise that tokens are too coarse to represent preference. We argue that this premise is mistaken. The difficulty lies in the objective against which the trace is trained, and a latent trace supervised by imitation inherits the same defect: it relocates the imitation target instead of removing it. The strongest latent baseline we evaluate at best ties direct supervision, while a change confined to the loss clears it (Section 5); representational sophistication is being expended on the wrong bottleneck.

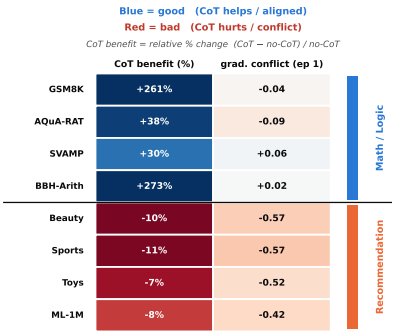


Figure 1: CoT benefit over direct supervision (left) and reasoning–answer gradient cosine (right), at epoch 1, for math and logic tasks (above the rule) and recommendation datasets (below). Blue favourable, red unfavourable. Cosines are estimated over a 2048-example probe (App. O).

Rationale imitation and answer ranking compete rather than cooperate. A decoder-only model emits both through the same readout, so the two losses act on the same parameters, and their interaction is governed by a single condition: whether the item the rationale argues for is the item the user selected. In mathematics this holds by construction, since a correct derivation terminates at the correct answer. In recommendation it typically fails, and it fails for a reason intrinsic to the task: the next-item distribution is *multi-modal*, so many continuations are plausible, an honest rationale commits to one of them, and the log records another. Imitating that rationale therefore trains the model to prefer an item the user did not select. Multi-modality, not the architecture, is what turns reasoning supervision against ranking.

The scope is a task structure, and it yields a test that runs before training. Recommendation is the clearest instance of a broader class, *non-verifiable, multi-modal prediction*, in which many outputs are legitimate, only one is logged, and no ground-truth rationale can exist even in principle. Dialogue evaluation, advertisement selection, content moderation and personalized education share this structure, and each is presently adopting CoT on the strength of results obtained in mathematics and code. We study recommendation because it is the setting in which every term of the analysis is observable , including the rationale–log agreement rate, which depends on the teacher and the task rather than the student and so can be estimated before any reasoning supervision is fitted. On our four benchmarks it is 0.07–0.13 against a chance level of 0.05, far enough below the ≈ 1 of verifiable tasks that the identity predicts harm. We summarize our contributions as follows:

1. **An explanation** (Section 2.1). A closed-form identity for the conflict between the reasoning and ranking gradients, whose sign is determined by whether the item the rationale argues for is the item that was logged. Because the next-item law is multi-modal, that condition fails in recommendation while holding by construction in mathematics, so one equation accounts both for the success of CoT there and its failure here. The identity is derived at the readout and shown to persist in a full multi-token Transformer.
2. **A quantity that predicts the sign in advance** (Sections 2.3, 3.1). The identity reduces the question to the rationale–log agreement rate, measurable without training the student. We measure it rather than assume it, and report the teacher's self-consistency beside it, which shows the disagreement to be multiplicity of valid continuations rather than teacher incompetence.

3. **Evidence that the explanation is correct** (Section 3). Three measurements establish that the predicted conflict arises at the point in training, in the region of the network, and to the degree the theory specifies, rather than only that a penalty is present.
4. **A remedy derived from the diagnosis** (Section 4). CARE, a family of conflict-aware training rules, together with outcome-reward reasoning, which removes the imitation target altogether. Both convert CoT from a liability into a gain over direct supervision, whereas methods that change the trace's representation rather than its supervision do not.

2 SETUP AND DIAGNOSIS

Task and pipeline. Let $\mathcal{I}$ be an item catalogue and let user u have interaction history $h_u = (i_1, \ldots, i_n)$, the items consumed in order, truncated to the $L{=}10$ most recent. Sequential recommendation scores every $i \in \mathcal{I}$ so that the held-out next item $i^\star$ is ranked first. A language model cannot score $\mathcal{I}$ directly: our catalogues contain 3.7K–18K items (Table 6). Standard practice factorizes the task into two stages. *Stage 1* retrieves a shortlist $\mathcal{C}_u \subset \mathcal{I}$ of K candidates using a computationnally cheap retriever (Appendix M). *Stage 2* reranks $\mathcal{C}_u$ with an LLM prompted over the history and the candidate list. The shortlist is what makes the second stage feasible at all: K candidates can be listed verbatim in a prompt, whereas the full catalogue cannot.

Our object of study is the reasoner, not the retriever. This paper asks a single question: *does chain-of-thought reasoning help an LLM decide which item a user will consume next?* Answering it requires isolating the reasoning stage from every other source of variation, and in a retrieve-then-rerank pipeline the dominant such source is Stage-1 recall. Table 8 makes the point quantitative: even the best training-free retriever places $i^\star$ in its top ten only 4.7-20.8% of the time, so reranking its raw output would measure retrieval recall far more than it measures reasoning. We therefore remove that confound by construction. Every method reranks the *same* oracle-complete shortlist, the retriever's top 19 candidates plus $i^\star$, shuffled into $K = 20$ slots under a fixed seed, so retrieval recall is exactly 1 and the candidate set is byte-identical across methods, backbones and training regimes. The target is always present and always reachable; the only remaining question is whether the model ranks it first. Any failure to do so is thus attributable to the reasoner alone, which is precisely the quantity our diagnosis and repair concern, c.f., Appendix L.

Notation. We write x for the prompt encoding $(h_u, \mathcal{C}_u)$, which renders the history as text and the K candidates $\mathcal{C}_u$ as a list lettered (A)–(T); y for the gold answer, the letter of $i^\star$ in that list; and z for a reasoning trace. Templates and a complete filled-in prompt are in Appendix B (Figure 2). By *reasoning* we mean any deliberate computation mapping x to a ranking using more than a single feedforward pass. Textual CoT realizes z as a token sequence and supervises both spans:

$$\mathcal{L}_{\mathrm{CoT}}(\theta) \;=\; \underbrace{-\log p_\theta(z \mid x)}_{\mathcal{L}_{\mathrm{reason}}} \;\; \underbrace{-\log p_\theta(y \mid z, x)}_{\mathcal{L}_{\mathrm{rank}}} . \tag{1}$$

Equation equation 1 is the object this paper analyses: a sum of two losses acting on shared parameters, only one of which has a verifiable target.

Two senses in which a trace can be useful. A trace may *resolve genuine uncertainty* about the answer, or *make an already-determined answer easier for a bounded model to compute*; much of the confusion surrounding CoT stems from conflating these. Separating them requires two models differing only in objective, $\hat{\theta}^{\mathrm{CoT}}$, minimizing Eq. equation 1 with traces from a teacher $z_n \sim \pi(\cdot \mid x_n)$, and $\hat{\theta}^{\text{No-CoT}}$, minimizing the answer span alone, identically initialized and given the same optimizer, data and step budget, so every difference between them is attributable to $\mathcal{L}_{\mathrm{reason}}$. The teacher induces a *target reasoning distribution* $p^*_\pi(z \mid x)$ the student must reproduce; the two policies used here are the answer-agnostic BLIND and answer-conditioned HINDSIGHT of Section 3.1. Four quantities govern what follows (Definitions 5–8, Appendix E; extended discussion in Appendix E.2): $\mathcal{I}_{\mathrm{true}} = \log \frac{p^*(y|x,z)}{p^*(y|x)}$, what the trace reveals *about the world*; $\mathcal{I}_{\mathrm{comp}} = \log \frac{p_{\hat{\theta}^{\mathrm{CoT}}}(y|x,z)}{p_{\hat{\theta}^{\text{No-CoT}}}(y|x)}$, how much it helps *this model* compute an answer it could not reach in one pass, with expectation I^{comp}_π under π; $\delta_{\mathrm{KL}} = \mathbb{E}_x[D_{\mathrm{KL}}(p^*_\pi(z \mid x) \,\|\, p_{\hat{\theta}^\pi}(z \mid x))]$, how closely the student reproduces that target, the term that penalizes HINDSIGHT-style supervision; and B_y, bounding the answer negative log-likelihood, small when the prompt determines the answer and of order $\log M$ when the next-item law is diffuse over M comparable continuations.

Why recommendation differs. Mathematics, code and logic sit at $\mathcal{I}_{\text{true}} = 0$ with $\mathcal{I}_{\text{comp}} \gg 0$: the answer is fixed by the statement but a bounded model cannot reach it in one pass, so the trace is a computational scaffold, made effective by *determinacy* (small B_y) and *verifiable compositionality*, i.e., checkable steps, so the trace is constructive and its target canonical. Recommendation has neither, with a consequence that separates it sharply: *no corpus of ground-truth reasoning exists, and none can exist even in principle*, since labelling *why* a user chose an item demands the counterfactual that is the prediction target. Every supervision trace must therefore be synthesized, so p^*_π is a modelling decision and δ_{KL} is a live term here but negligible in mathematics, c.f., Appendix I.

2.1 A Risk Bound, and the One Channel It Cannot Use

Decomposing the test risk of the CoT model gives a sufficient condition for it to beat direct prediction, trading the benefit I^{comp}_π against the penalty of failing to reproduce the target reasoning at test time, amplified by answer difficulty, as stated by our Theorem 6 in Appendix F:

$$I^{\text{comp}}_\pi \;>\; B_y \sqrt{\tfrac{\delta_{\text{KL}}}{2}}\,. \tag{2}$$

The next lemma closes one of the three quantities exactly, with no modelling assumption at all.

Lemma 1 (A prompt-derived trace carries no information about the log; proof in Appendix H)**.** *Suppose the trace is generated from the prompt alone, so that y is never revealed during generation. Then $Y - X - Z$ forms a Markov chain and, by the data-processing inequality, $I(Y; Z \mid X) = 0$, hence $p^*(y \mid x, z) = p^*(y \mid x)$ and $\mathcal{I}_{\text{true}}(z; x, y) = 0$.*

Whatever CoT accomplishes here, then, it cannot be informational. The question reduces to the sign of $\mathcal{I}_{\text{comp}}$, which, unlike $\mathcal{I}_{\text{true}}$, is a property of the trained models rather than of the task, and cannot be settled by an information-theoretic argument alone. Determining it is the subject of Section 2.1.

2.2 The Sign of I^{comp}_π Is a Gradient Alignment

As defined in Eq. 8, I^{comp}_π compares two fully trained models, which says nothing about where the harm comes from and offers no way to measure it short of training both. We therefore re-express it as a property of the optimization: since the two models differ only by the extra gradient the reasoning loss contributes, the damage must be visible in how that gradient sits relative to the ranking one. Write $g_{\text{a}}(\theta) = \nabla_\theta \mathcal{L}_{\text{rank}}(\theta)$ and $g_{\text{r}}(\theta) = \nabla_\theta \mathcal{L}_{\text{reason}}(\theta)$, each obtained by masking the labels of the complementary span (Appendix O), and run gradient descent from a shared initialization θ_0 for T steps at learning rate η: the no-CoT model on g_{a} alone, the CoT model on $g_{\text{a}} + g_{\text{r}}$ (Eq. 16, Appendix G). The two trajectories differ only by the additional term $-\eta\, g_{\text{r}}$, which is the entire object of study; we evaluate both gradients along the common no-CoT reference trajectory and abbreviate $g^t_{\text{a}} := g_{\text{a}}(\theta^N_t)$, $g^t_{\text{r}} := g_{\text{r}}(\theta^N_t)$.

Lemma 2 (First-order link between I^{comp}_π and gradient conflict; proof in Appendix G)**.** *Let $R_a(\theta) = \mathbb{E}_{(x,y)}\big[\mathbb{E}_{z\sim p^*_\pi(\cdot|x)}[-\log p_\theta(y \mid z, x)]\big]$ denote the population answer risk, so that $I^{\text{comp}}_\pi = R_a(\hat\theta^{\text{No-CoT}}) - R_a(\hat\theta^{\text{CoT}})$. Under standard smoothness and boundedness conditions (A1)-(A4), stated in Appendix G,*

$$I^{\text{comp}}_\pi = \eta \sum_{t=0}^{T-1} \big\langle g^t_{\text{a}}, g^t_{\text{r}} \big\rangle + O(\eta^2) = \eta \sum_{t=0}^{T-1} \big\|g^t_{\text{a}}\big\| \big\|g^t_{\text{r}}\big\| \cos\big(g^t_{\text{a}}, g^t_{\text{r}}\big) + O(\eta^2). \tag{3}$$

To leading order the sign of I^{comp}_π is therefore the accumulated reasoning–ranking cosine: sustained alignment means chain-of-thought helps, sustained conflict that it hurts. The measured cosine is therefore not a proxy for the penalty but its leading-order term. Appendix G relates Eq. 3 to influence-function diagnostics (Koh & Liang, 2017; Pruthi et al., 2020).

2.3 Why the Cosine Is Negative: An Exact Identity at the Shared Readout

Lemma 2 makes the question geometric but does not explain *why* the cosine drifts negative in recommendation and non-negative in mathematics. The explanation concerns not the domain but the place at which the two losses meet: in a decoder-only model the rationale tokens and the answer

token are decoded through the *same* unembedding, so both losses act on the same logits, which is also where the empirical conflict concentrates (Fig. 3). Index the target space by $k \in \{1, \dots, K\}$, let $\theta \in \mathbb{R}^K$ be the shared logits, and set $p = \mathrm{softmax}(\theta)$, so the ranking loss is $\mathcal{L}_{\mathrm{rank}}(\theta) = -\log p_y$. We adopt one assumption, discharged in §2.3.

Assumption 3 (Shared-readout reduction)**.** *An honest teacher trace commits to a single* rationale-implied *item $m \in \{1, \dots, K\}$. Since the trace is decoded through the same head as the answer, we model the effect of supervising it as a cross-entropy toward m, $\mathcal{L}_{\mathrm{reason}}(\theta) = -\log p_m$, discarding the answer-neutral remainder of the multi-token loss.*

This collapses a multi-token rationale to the single decision it argues for. Under it both gradients take the softmax form $g_{\mathrm{a}} = p - e_y$ and $g_{\mathrm{r}} = p - e_m$, with e_k the k-th standard basis vector, and their inner product admits a closed form. The reduction is for transparency rather than necessity: Section 2.3 proves the same sign without it, and Section 2.3 measures $\Pr(y{=}m)$ directly.

Lemma 4 (Readout conflict identity; proof in Appendix H)**.** *Under Assumption 3, for any logit vector θ,*

$$\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle \;=\; \|p\|^2 \;-\; p_y \;-\; p_m \;+\; \mathbf{1}\{y = m\}. \tag{4}$$

The identity is exact, holds for arbitrary p, and contains no free parameters. Every term but the last is symmetric in y and m, so the entire asymmetry, and with it the sign, resides in the indicator $\mathbf{1}\{y = m\}$: does the item the rationale argues for equal the item that was logged? This gives the two regimes directly (Corollary 10, Appendix H). If $m = y$, as for verifiable unimodal targets, $\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle = (1 - p_y)^2 + (\|p\|^2 - p_y^2) \geq 0$ and the two losses reinforce. If $m \neq y$, as for non-verifiable multi-modal ones, $\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle = \|p\|^2 - p_y - p_m$, strictly negative whenever $p_y + p_m > \|p\|^2$; in the balanced bimodal case $p_y = p_m = a$ it equals $-2a(1 - a) < 0$ for every $a \in (0, 1)$, peaking at $a = \frac{1}{2}$, so the conflict is worst exactly when the model is undecided between two modes.

From a per-example identity to a task-level criterion. Taking expectations over prompts, with $y \sim q(\cdot \mid x)$ the logged item and $m \sim \tilde{q}(\cdot \mid x)$ the rationale-implied one, conditionally independent given x because the teacher never observes y, the indicator becomes an agreement probability:

$$\mathbb{E}\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle \;=\; \mathbb{E}\|p\|^2 - \mathbb{E}[p_y] - \mathbb{E}[p_m] + \Pr(y{=}m), \qquad \Pr(y{=}m) = \mathbb{E}_x \langle q, \tilde{q} \rangle. \tag{5}$$

Verifiable unimodal tasks have $\Pr(y{=}m) = 1$ and the expectation is non-negative. Otherwise Cauchy–Schwarz gives $\Pr(y{=}m) \leq \sqrt{\sum_k q_k^2} \sqrt{\sum_k \tilde{q}_k^2}$, small whenever the true posterior is diffuse or the teacher concentrates away from its mass, and the expectation turns negative (Proposition 11, Appendix H.5). Near initialization p is diffuse and the sign is indeterminate at order $1/K$; the conflict switches on as mass concentrates on the two disagreeing modes, the falsifiable timing prediction Section 3.1 confirms.

Discharging the assumption: a multi-token Transformer. The conclusion does not depend on Assumption 3. Treating the rationale as it is actually supervised, token by token, the shared unembedding contributes $G_U = \sum_{t=1}^{T} \langle h_a, h_{r,t} \rangle F_t$ with $F_t = \langle p_a, p_{r,t} \rangle - p_{a,z_t} - p_{r,t,y} + \mathbf{1}\{y = z_t\}$, an exact factorization whose sign reproduces Corollary 10 (Theorem 12, Appendix H): $G_U \leq 0$ when answer and reasoning tokens occupy disjoint sub-vocabularies and the heads are specialized and confident, $G_U \geq 0$ when $z_t = y$ at the answer-carrying step. The driver of the negative sign is the single term $-p_{a,z_t}$, present only because of the imitation target e_{z_t}, which identifies that target as the object to remove. A concentration argument bounds the cosine at initialization by $O(1/\sqrt{d})$ (Lemma 20, Appendix N), predicting the milder conflict we observe at 7B than 3B; but it precedes head specialization, whereas Theorem 12 governs the trained regime, and the penalty persists at 7B. Scale mitigates the conflict without resolving it.

The remedy is determined by the analysis, not chosen heuristically. Of the three quantities governing Eq. 2, the informational term $\mathcal{I}_{\mathrm{true}}$ is identically zero and immovable (Lemma 1); the learnability penalty δ_{KL} is positive only because $p_\theta(z \mid x)$ is asked to match an external teacher; and $\mathcal{I}_{\mathrm{comp}} < 0$ arises solely from the imitation target e_{z_t} (Theorem 12). Removing that target, by decaying its weight, by projecting out its ranking-opposed component, or by replacing it with a reward on the ranking the trace produces, eliminates both negative terms simultaneously. Section 4 develops these remedies, and Section 5 shows that the sign of the effect duly reverses.

3 EMPIRICAL VALIDATION OF THE DIAGNOSIS

Section 2.1 makes three falsifiable predictions. **(P-i)** Forcing a textual trace should not help even an *untrained* reranker, since the mechanism concerns the trace's target and the task's multi-modality rather than fine-tuning. **(P-ii)** The cosine should be non-negative at initialization and turn negative once probability mass concentrates on two disagreeing modes. **(P-iii)** The conflict should be sharpest at the layers where the model commits to an answer, and the reasoning states there should be nearly answer-agnostic. Probe protocols and per-dataset tables are in Appendices Q, O and P.

(P-i) The effect precedes fine-tuning. Untrained rerankers at three scales already show it: forcing a trace is neutral or harmful in every instruct-model cell on NDCG@10 and in ten of twelve on NDCG@1, and neither scale nor reasoning-specific tuning rescues it (Table 14, Appendix Q). Setting $\hat{\theta}^{\mathrm{CoT}} = \hat{\theta}^{\text{NO-COT}}$ to the pretrained model collapses $\mathcal{I}_{\mathrm{comp}}$ to $\log[p_\theta(y \mid x, z)/p_\theta(y \mid x)]$, exactly the REASON-minus-DIRECT contrast tabulated there, so the sign is negative before a single gradient step and cannot be an artefact of overfitting a distilled corpus.

3.1 (P-II) THE CONFLICT TURNS ON EXACTLY WHEN PREDICTED

We fine-tune Qwen2.5-3B/7B on four datasets under direct (NO-COT) and CoT supervision with teacher traces from Qwen2.5-32B, the strongest zero-shot reranker in the study, probing gradient geometry post-hoc at epochs 0–3. Because recommendation supplies no ground-truth rationales, every trace is synthesized, and *how* it is synthesized is a free variable: we distill two teacher families from the same frozen model, differing *only* in what the teacher may see. BLIND reasons blind to the ground truth, so $\Pr(y{=}m) = \mathbb{E}_x\langle q, \tilde{q}\rangle$; HINDSIGHT writes the trace with the answer revealed, then scrubbed by strict prompt constraints, so $\Pr(y{=}m) \to 1$ by construction. The two instantiate, as data, the two corners of Proposition 11 (Appendices C.2–C.4, O).

The measured geometry follows Eq. 5 closely (Table 9, Appendix N). At epoch 0, where the distribution is diffuse and the sign is indeterminate at order $1/K$, every cell is *positive* ($+.11$ to $+.40$ across both scales). After one epoch of mass concentration BLIND swings sharply negative ($-.31$ to $-.71$ at 3B), while HINDSIGHT, whose $\Pr(y{=}m) \to 1$, stays near zero ($-.02$ to $-.30$): the theory specifies the sign, the timing, and which regime is spared, and all three hold. HINDSIGHT also carries roughly ten times the gradient magnitude of BLIND, whose share of the total norm collapses below 1%, so an honest trace is not merely misdirected but a comparatively weak update. Table 11 (Appendix O) collects the per-cell result, the paper's most direct empirical statement: $\cos(g_{\mathrm{r}}, g_{\mathrm{a}})$ is negative in all eight cells, at both scales and on every dataset, so by Lemma 2 I_π^{comp} is itself negative and the trace is not a neutral scaffold but an actively ranking-opposed one. Cells with more sustained conflict tend to pay the larger penalty, as the same lemma predicts, though one observation per cell is too small a sample to call a correlation.

The agreement term is measured, not assumed. Proposition 11 attributes the conflict to a small $\Pr(y{=}m)$, and that quantity is directly observable: the BLIND teacher ends its trace with its own ranking, so m is its top choice and $\Pr(y{=}m)$ its top-1 accuracy on the training split. It is 0.071–0.128 across the four datasets, 1.4 to 2.6 times chance, so the rationale is genuinely informative yet far from the ≈ 1 that verifiable tasks enjoy. Resampling the teacher gives a self-consistency $\Pr(m{=}m') = \mathbb{E}_x \sum_k \tilde{q}_k^2 = 0.44$–$0.62$, the collision probability of its own implied distribution and four to six times its accuracy, so $\tilde{q}$ has effective support $1/\sum_k \tilde{q}_k^2 \approx 2$: the teacher is concentrated on roughly two candidates and confidently wrong rather than diffuse. Two consequences favour the mechanism. The reasoning gradient is a *stable* commitment to one wrong vertex, so it accumulates across steps in Eq. 3 rather than cancelling, and the disagreement is multiplicity of valid continuations, not teacher incompetence (Appendix O.3).

3.2 (P-III) THE CONFLICT IS WHERE THE ANSWER IS DECIDED

Theorem 12 localizes the conflict at the shared readout, so it should be sharpest where the model commits to an answer. It is: the per-layer cosine is negative at *every* layer, so the conflict is broad, but it deepens toward the final layers (Fig. 3). If rationale imitation pulls the shared head away from the logged item, the model's best strategy is to stop relying on the rationale, and a frozen linear probe confirms exactly that. Decoding the model's *own committed answer* from the final reasoning token (chance 0.05, answer-position ceiling 0.67–0.76), full-loss CoT reaches only 0.11–0.19 in all eight

cells. Every conflict-resolving method raises decodability, in the order full-loss $<$ gproj $<$ anneal, reaching 0.41–0.69, and not as an accuracy artefact, since methods with near-identical top-1 differ by 0.19 here (Table 13, Appendix P).

4 REPAIR: REASONING WITHOUT AN IMITATION TARGET

Section 2.1 identifies one object as responsible: the imitation target e_{z_t} in $\mathcal{L}_{\text{reason}}$. It drives both fixable terms of Eq. 2, since the negative sign of $\mathcal{I}_{\text{comp}}$ arises from it (Theorem 12) and the penalty $B_y\sqrt{\delta_{\text{KL}}/2}$ is incurred only because imitation requires matching an external teacher, while the third term is immovable ($\mathcal{I}_{\text{true}} \equiv 0$, Lemma 1). Removing that one object therefore addresses both. We consider two ways of doing so: CARE keeps the imitation loss but neutralizes its ranking-opposed component (§4.1), and outcome-reward reasoning replaces $\mathcal{L}_{\text{reason}}$ with a reward on the ranking the trace induces (§4.2). Table 5 in Appendix J summarizes which term each controls.

4.1 CARE: CONFLICT-AWARE REASONING SUPERVISION

Write the CoT objective with an explicit weight on the reasoning term, $\mathcal{L}(\theta) = \mathcal{L}_{\text{rank}}(\theta) + \lambda_t\mathcal{L}_{\text{reason}}(\theta)$, and recall from Remark 9 that $I_\pi^{\text{comp}} = \eta\sum_t \lambda_t\langle g_{\text{a}}^t, g_{\text{r}}^t\rangle + O(\eta^2)$, so the harm lies entirely in the negative terms of that sum. The following rules remove them in different ways. The two CARE rules are alternatives to each other, not stages of a pipeline; outcome-reward reasoning (§4.2) is a subsequent stage that may be applied to either.

ANNEAL: retiring the target once it has served its purpose. Proposition 11 leaves the sign indeterminate at order $1/K$ at initialization and turns it negative only as mass concentrates on the two disagreeing modes, which suggests the rationale loss is useful early, when it teaches output format, candidate bookkeeping and the answer protocol, and harmful thereafter. ANNEAL therefore holds $\lambda_t = 1$ for E_{sw} epochs, then decays it linearly to zero:

$$\lambda_t \;=\; \min\{1,\; \max(0,\; 1 - \tfrac{t-E_{\text{sw}}}{E_{\text{dec}}})\}. \tag{6}$$

After the switch the reasoning positions are still *generated* but no longer *supervised*, free to specialize toward whatever assists the answer, the most complete removal of e_{z_t} available without leaving supervised learning, and consistent with ANNEAL being the most answer-decodable method in Table 13.

GPROJ: removing only the ranking-opposed component. GPROJ keeps the imitation target but discards the part of its gradient that opposes ranking, projecting g_{r} onto the orthogonal complement of g_{a} whenever the two disagree, after multi-task gradient surgery (Yu et al., 2020). This enforces $\langle g_{\text{a}}, \tilde{g}_{\text{r}}\rangle \geq 0$ at every step, so every term of Eq. 3 is non-negative and $I_\pi^{\text{comp}} \geq 0$ to first order: the penalty is rendered non-positive by construction. It is the more conservative rule, preserving whatever non-conflicting structure the rationale carries, consistent with its recovering less decodability than ANNEAL (update rule: Eq. 23, Appendix J).

4.2 OUTCOME-REWARD REASONING: REMOVING THE IMITATION TARGET

CARE repairs the conflicting gradient but leaves in place the object that creates it; the stronger intervention removes that object altogether. Let $\pi_\theta(z \mid x)$ generate the trace and $p_\theta(\cdot \mid z, x)$ be the answer readout, inducing a ranking $\sigma_\theta(x, z)$ over $\mathcal{C}_u$. We define a single verifiable reward equal to the ranking quality of that induced order against the logged item:

$$R(x, y, z) \;=\; \text{NDCG@}K\big(\sigma_\theta(x, z),\, y\big), \qquad z \sim \pi_\theta(\cdot \mid x). \tag{7}$$

This is the entire training signal for the stage: no $\mathcal{L}_{\text{reason}}$, no teacher distribution, and no target token e_{z_t}, so the term Theorem 12 identifies as the sole negative driver is absent from the objective by construction. We initialize from a CARE checkpoint rather than from the base model, so the imitation target is absent during outcome-reward training but was present in the initialization; Appendix K records which checkpoint is used per cell. Note that y enters only as a training-time score, since z is still sampled from x alone, so $Y - X - Z$ and hence $\mathcal{I}_{\text{true}} = 0$ are preserved with no test-time leakage; and that the reward is read from the same letter-logit ARGSORT used at evaluation, so the training signal and the test metric coincide.

The reward is not differentiable in z, so we optimize it with direct preference optimization (Rafailov et al., 2023), which needs only one round of offline sampling: per prompt we draw G traces, score them with Eq. 7, and form a preference pair from the best and worst. The objective (Eq. 25, Appendix K) carries an implicit trust region against the frozen supervised initialization, which is *not* the δ_{KL} of Theorem 6, that term measured the unlearnability of an external teacher and is absent here by construction.

Outcome-reward training therefore drives both fixable terms to their most favourable values simultaneously: the reward is a monotone functional of the same ranking quality $\mathcal{L}_{\mathrm{rank}}$ scores, so $\langle g_{\mathrm{a}}, g_{\mathrm{r}}^{\mathrm{RL}} \rangle \geq 0$ and hence $I_{\pi}^{\mathrm{comp}} \geq 0$, while no external trace distribution must be reproduced, so $\delta_{\mathrm{KL}} \to 0$ (Proposition 18, Appendix K). Since $\mathcal{I}_{\mathrm{true}} = 0$ is immovable within the textual trace, this is the most favourable setting of the two fixable terms available to a token-space procedure, which lends it a clean interpretation as a control: should it exceed direct supervision, the excess must be attributable to $\mathcal{I}_{\mathrm{comp}} > 0$, the other two channels being provably closed. Section 5 shows that it does, on every completed dataset, establishing the positive claim of this paper: reasoning assists recommendation provided it is supervised by outcomes rather than imitated from rationales.

5 Experiments

Setup. Four benchmarks spanning text-rich and collaboratively-dominated regimes (Amazon Beauty, Toys, Sports, ML-1M) under a strict leave-last-two-out split. We compare non-LLM recommenders, zero-shot and fine-tuned LLMs, alternative reasoning media, and ours. A block of controls separately closes the alternatives to our positive claim. Every row of Table 1 reorders the *identical* $K{=}20$ oracle-complete shortlist (Section 2), so retrieval recall is 1.0 and no result can be explained by candidate generation; comparisons are therefore *paired*, and we quote margins against a paired resampling error. Full protocol can be found in Appendices A, D and L.

(1) Rationale imitation is computationally expensive, and removing it reverses the sign. SFT with rationales trails direct supervision on N@1 in *every one of the eighteen completed backbone-dataset cells across both model families*, and within a scale the loss tracks the strength of the direct model, most sharply on ML-1M at 7B, where N@1 falls from .270 to .068: the signature Theorem 12 predicts, since more specialized heads have more to lose. Across scales the penalty shrinks with hidden width, as Lemma 20 anticipates, so this cell is the most extreme we observe rather than a trend (finding (4)). Both remedies then clear direct supervision: CARE-ANNEAL at 7B beats NO-COT on Beauty, Toys and Sports, and OUTCOME-DPO (init: CARE) beats it in every completed cell at both scales, with the largest lift where NO-COT is weakest, on ML-1M at 3B.

(2) Reasoning genuinely helps, but only after the conflict is removed. Outcome-reward training exceeds direct supervision in every completed cell, robustly on Beauty, Toys and Sports and marginally on ML-1M at 7B. By Proposition 18 the other two channels are closed ($\mathcal{I}_{\mathrm{true}} = 0$ provably, $\delta_{\mathrm{KL}} \to 0$ by construction), so the excess must be attributable to $\mathcal{I}_{\mathrm{comp}} > 0$. The sharpest statement about outcome supervision, however, is not that it beats direct supervision but that *it does not work on a conflicted checkpoint*. The two OUTCOME-DPO rows share reward, rollout compute budget and sampling, differing only in initialization: starting from CARE rather than SFT-CoT yields $+4.8$ to $+10.8$ NDCG@10 in all eight cells. On ML-1M at 7B the conflicted initialization reaches .399 against a NO-COT level of .494, the repaired one .500. Outcome reward is thus not an alternative to removing the imitation target but a beneficiary of it. Conversely the two removals *converge*: OUTCOME-DPO (init: CARE) and the better CARE rule differ by a median 0.001 NDCG@10 (range -0.012 to $+0.006$, inside resampling error), as a correctly identified cause predicts, so the free rule captures essentially all of the recoverable gain.

(3) Changing the form of the trace does not help; changing its supervision does. CoCoNuT, the strongest latent-reasoning baseline, at best ties NO-COT , on Qwen at 7B, and falls below it at 3B on three of four datasets and, on Llama, by 3.9 to 7.0 NDCG@1 points in every one of ten cells; Soft Thinking and Think-as-Hidden are worse still. We record the one exception rather than omit it: CoCoNuT at 3B does clear NO-COT on Sports (.435 against .366 NDCG@10, panel (a) of Table 1), the single cell in which a latent trace improves on direct supervision. It does not disturb the reading, since CARE-GPROJ at the same scale reaches .446 on that cell and the ordering elsewhere is unchanged, but the claim is that latent methods at best tie direct supervision *in aggregate*, not that they never exceed it. A change confined to the loss outperforms all of them, exactly as Table 5

Table 1: All methods rerank the identical oracle-complete K=20 shortlist (retrieval recall = 1.0). **Bold**: best fine-tuned LLM method per dataset, metric and panel. †Soft Thinking requires a backbone that natively emits <think> blocks, so it is run on DeepSeek-R1-Distill-Qwen-7B; see Appendix S.2.

	ML-1M			Amazon Beauty			Amazon Toys			Amazon Sports		
Method	N@1	N@10	R@10	N@1	N@10	R@10	N@1	N@10	R@10	N@1	N@10	R@10
Non-LLM sequential recommenders (backbone-free)												
SASRec	.082	.280	.558	.112	.397	.760	.132	.407	.760	.120	.377	.716
HSTU	.088	.313	.620	.044	.298	.672	.048	.306	.684	.022	.240	.566
HSTU-large	.122	.337	.638	.068	.347	.714	.080	.352	.740	.028	.268	.606
AlphaRec	.168	.438	.794	.166	.411	.736	.196	.422	.710	.150	.390	.712
(a) Qwen2.5-3B backbone												
Zero-shot LLM (no training)												
ZS-direct (3B)	.046	.198	.432	.102	.274	.517	.096	.279	.540	.090	.264	.510
ZS-reason (3B)	.052	.178	.351	.085	.256	.480	.089	.248	.448	.083	.253	.477
Fine-tuned LLM: imitated rationale												
SFT NO-COT (3B)	.106	.296	.582	.172	.369	.635	.160	.373	.656	.150	.366	.656
SFT COT (3B)	.053	.234	.508	.146	.374	.680	.158	.383	.677	.137	.382	.711
Alternative reasoning modalities: imitation target relocated												
COCONUT (3B)	.053	.229	.500	.179	.385	.655	.137	.339	.621	.192	.435	.747
Soft Thinking (R1-Distill-7B)†	.053	.197	.404	.096	.264	.489	.107	.277	.495	.093	.253	.461
Think-as-Hidden (3B)	.057	.196	.390	.069	.181	.332	.054	.137	.245	.108	.295	.532
Outcome reward, conflicted initialization (control)												
OUTCOME-DPO (init: SFT-COT, 3B)	.102	.307	.602	.146	.373	.677	.160	.385	.679	.139	.382	.710
OUTCOME-GRPO (init: SFT-COT, 3B)	.104	.313	.615	.144	.372	.679	.158	.384	.679	.138	.382	.709
Ours: imitation target removed												
CARE-ANNEAL (3B)	.193	.384	.650	.118	.329	.623	.148	.343	.618	.152	.358	.646
CARE-GPROJ (3B)	**.208**	**.427**	**.720**	**.205**	**.442**	.749	.205	.443	.749	.195	.446	.770
OUTCOME-DPO (init: CARE-GPROJ, 3B)	.204	.415	.698	.204	**.442**	**.751**	**.206**	**.445**	**.752**	**.197**	**.449**	**.775**
(b) Qwen2.5-7B backbone												
Zero-shot LLM (no training)												
ZS-direct (7B)	.054	.212	.446	.116	.314	.586	.121	.324	.594	.110	.319	.604
ZS-reason (7B)	.061	.211	.421	.106	.291	.530	.113	.311	.564	.103	.295	.547
Fine-tuned LLM: imitated rationale												
SFT NO-COT (7B)	.270	.494	.780	.202	.440	.750	.188	.433	.748	.179	.426	.752
SFT COT (7B)	.068	.267	.563	.165	.395	.700	.170	.399	.697	.157	.405	.737
Alternative reasoning modalities: imitation target relocated												
COCONUT (7B)	**.277**	**.501**	.784	.198	.439	.752	.202	.449	.764	.174	.414	.735
Think-as-Hidden (7B)	.054	.174	.342	.070	.184	.336	.066	.191	.358	.110	.300	.542
Outcome reward, conflicted initialization (control)												
OUTCOME-DPO (init: SFT-COT, 7B)	.169	.399	.707	.166	.397	.702	.171	.399	.696	.158	.405	.736
OUTCOME-GRPO (init: SFT-COT, 7B)	.179	.412	.720	.165	.395	.701	.171	.401	.700	.158	.405	.735
Ours: imitation target removed												
CARE-ANNEAL (7B)	.266	.487	.772	.224	**.470**	**.785**	**.214**	**.459**	**.773**	.215	.474	**.802**
CARE-GPROJ (7B)	.271	.494	.780	.193	.431	.743	.205	.448	.761	.171	.419	.745
OUTCOME-DPO (init: CARE-ANNEAL, 7B)	.273	.500	**.786**	**.226**	**.470**	**.785**	.205	.447	.760	**.220**	**.476**	.801

predicts, since latent methods relocate the imitation target rather than removing it: representational sophistication is being expended on the wrong bottleneck. The diagnosis also replicates across families. On Llama-3.2-1B and Llama-3.1-8B, with pipeline, shortlists and evaluation cases held fixed, the penalty reappears in every completed cell, so it is a property of the task rather than of the Qwen family. The *repair*, by contrast, is not scale-invariant: the two CARE rules cross over, ANNEAL favoured at 7B/8B and GPROJ at 3B, and half of this follows from Lemma 20, since GPROJ's intervention is smaller than ANNEAL's by exactly $|\cos(g_r, g_a)| = O(1/\sqrt{d})$ and so decays with width. CARE adds no inference cost over SFT-CoT and, for ANNEAL, no training cost at all; outcome-reward training requires G=6 offline rollouts per prompt, collected once.

6 CONCLUSION

CoT reliably helps in mathematics, code and logic, and often hurts in recommendation. The explanation is neither that recommendation is unsuited to deliberation nor that token space is the wrong form for a trace, but that the rationale and the answer are decoded through the same head: because the next-item distribution is multi-modal, an honest rationale commits to a different item than the one logged, and the two gradients pull the shared logits apart. That single fact yields an exact identity whose sign is decided by the rationale-log agreement probability, and it names the term responsible, removing the token-level imitation target, by annealing it, by projecting out its ranking-opposed component, or by replacing it with a reward on the ranking the trace induces, reverses the effect in every case. Two conclusions outlive the setting. The mechanism is a property of the task structure, so it should predict the sign of the CoT effect wherever prediction is non-verifiable and multi-modal, and the agreement rate it turns on is computationally cheap to measure in advance. And the failure is located in the objective, not the representation, which is why methods that relocate the imitation target into a latent space inherit the penalty while a change confined to the loss removes it.

AI USE STATEMENT

In this work, we used generative AI tools principally for paraphrasing and copy-editing of author-written prose, and to a lesser extent for debugging code. We have reviewed all AI-assisted work: every paraphrased passage was read and revised by the authors, and AI-assisted code was verified by rerunning the affected stage and confirming that all reported metrics were unchanged.

REPRODUCIBILITY STATEMENT

Every component needed to reproduce this paper is specified in the appendix: datasets, splits and candidate construction, the Stage-1 retriever ablation that fixes the shortlist, the exact teacher prompts and post-processing for both CoT regimes, the two ranking-extraction protocols, the gradient-conflict probe, the definitions and hyperparameters of every repair method, the outcome-reward training loop, the decodability probe with its positive and negative controls, and all proofs.

ETHICS STATEMENT

This work uses only public benchmark datasets (Amazon reviews, MovieLens). We train no user-identifying model and release no user data. One societal consideration is worth stating directly: our decodability probe shows that a model's stated rationale can be largely disconnected from the computation that produces its answer, so rationales emitted by reasoning-based rerankers should not be presented to users as explanations of why an item was recommended.

REFERENCES

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216, 1993.

Fahad Anwaar, Adil Mehmood Khan, Muhammad Khalid, Usman Zia, and Kezhi Wang. Reasoning-guided collaborative filtering with language models for explainable recommendation. *arXiv preprint arXiv:2602.05544*, 2025.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. TALLRec: An effective and efficient framework for LLM-based recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, pp. 1007–1014, 2023.

Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys)*, pp. 39–46, 2010.

Mukund Deshpande and George Karypis. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, 22(1):143–177, 2004.

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

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *Proceedings of the 46th European Conference on Information Retrieval (ECIR)*, pp. 364–381, 2024.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation: Benchmarking llms as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3251–3265, 2026.

Dietmar Jannach and Malte Ludewig. When recurrent neural networks meet the neighborhood for session-based recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems (RecSys)*, pp. 306–310, 2017.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 197–206, 2018.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1885–1894, 2017.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213, 2022.

Jiahao Li et al. Plr: Preference-based latent reasoning for generative recommendation. *arXiv preprint arXiv:2601.03153*, 2025.

Yufei Li, Zaiwei Zhang, Mingfu Liang, Kavosh Asadi, Jay Xu, Jimmy Kim, Chongyang Bai, Jieyi Zhang, Hongye Xie, Prachi Agrawal, et al. GR2 technical report. *arXiv preprint arXiv:2606.31984*, 2026.

Xinyu Lin, Hanqing Zeng, Hanchao Yu, Yinglong Xia, Jiang Zhang, Aashu Singh, Fei Liu, Wenjie Wang, Fuli Feng, Tat-Seng Chua, and Qifan Wang. Verifiable reasoning for llm-based generative recommendation. *arXiv preprint arXiv:2603.07725*, 2026.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Malte Ludewig, Noemi Mauro, Sara Latifi, and Dietmar Jannach. Empirical analysis of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, 31:149–181, 2021.

Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd International Workshop on Web Information and Data Management (WIDM)*, pp. 9–15, 2001.

Mobina Pournemat, Keivan Rezaei, Gaurang Sriramanan, Arman Zarei, Jiaxiang Fu, Yang Wang, Hamid Eghbalzadeh, and Soheil Feizi. Reasoning under uncertainty: Exploring probabilistic reasoning capabilities of llms. *arXiv preprint arXiv:2509.10739*, 2025.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19920–19930, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 811–820, 2010.

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW)*, pp. 285–295, 2001.

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

Leheng Sheng, An Zhang, Yi Zhang, Yuxin Chen, Xiang Wang, and Tat-Seng Chua. Language representations can be what recommenders need: Findings and potentials. In *ICLR*, 2025.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Xin Wang et al. Lasar: Latent sequential autoregressive reasoning for recommendation. *arXiv preprint arXiv:2605.10207*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, Yinghai Lu, and Yu Shi. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152*, 2024.

Yupeng Zhang et al. Openonerec: An open-source framework for unified generative recommendation. *arXiv preprint arXiv:2606.06260*, 2025.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *Advances in Neural Information Processing Systems*, 38:168990–169012, 2026.

CONTENTS OF THE APPENDIX

Part A – The task, the prompts, and the reasoning traces

A Experimental Setup . . . 13

B Prompt Templates . . . 14

C Teacher Chain-of-Thought: BLIND and HINDSIGHT . . . 14

D Ranking Extraction: Zero-Shot vs. Trained SFT . . . 17

Part B – Theory and proofs

E Formal Setup . . . 18

F Proof of Theorem 6 . . . 20

G Proof of Lemma 2 . . . 20

H Proofs for the Rationale–Log Disagreement . . . 22

I When CoT Helps: Task Regimes . . . 26

Part C – Method details

J CARE: Conflict-Aware Reasoning, Full Specification . . . 27

K Outcome-Reward Reasoning: Implementation . . . 28

Part D – Protocol and reproducibility

L Implementation Details . . . 30

M Retrieval Ablation: Choosing the Stage-1 Retriever . . . 31

Part E – Additional results and measurements

N Near-Orthogonality of Reasoning and Ranking Gradients at Scale 32
O Gradient-Conflict Probe: Protocol and Full Measurements 35
P Reasoning-State Answer Decodability 36
Q Zero-Shot LLM Reranking: Full Results and Analysis 40
R Teacher-Regime Ranking Results 42
S Additional Results and Ablations 42

Part A: The Task, the Prompts, and the Reasoning Traces

A Experimental Setup

Setup. Four benchmarks spanning text-rich and collaboratively-dominated regimes, Amazon Beauty, Toys, Sports and MovieLens-1M (Hou et al., 2026), under a strict leave-last-two-out split, with all corpus statistics fit on training prefixes only (Appendix L).

We compare four families: non-LLM recommenders (SASRec (Kang & McAuley, 2018), HSTU-large (Zhai et al., 2024), AlphaRec (Sheng et al., 2025)); zero-shot LLM; fine-tuned LLM (SFT NO-CoT, SFT CoT); and alternative reasoning modalities (CoCoNuT (Hao et al., 2025), Soft Thinking (Zhang et al., 2026), Think-as-Hidden). Ours are CARE-ANNEAL, CARE-GPROJ and outcome-reward DPO.

Controls. Because our positive claim is that reasoning helps *once outcome-supervised*, we evaluate a block of controls that separately close the alternatives (extra training, extra test-time computation, inference-time search, and the reward acting alone with the trace playing no role), alongside the closest outcome-supervised prior work (Lin et al., 2026). Their rows, and the alternative each is meant to close, are in Table 1.

This appendix gives the end-to-end experimental design in one place. Per-component detail lives in the appendices it points to; nothing substantive is stated only here.

A.1 What is held fixed across every experiment

The design principle is that exactly one thing varies between any two numbers we compare. Concretely, the following are byte-identical across all methods, backbones, and training regimes:

1. **The data split.** Strict leave-last-two-out at three distinct positions, so validation and test targets are never training labels, and all corpus statistics used for candidate construction are fit on training prefixes only (Appendix L.2).
2. **The candidate set.** One cached oracle-complete shortlist per case: the first-order Markov retriever's top-19 plus the inserted ground-truth target, shuffled into $K{=}20$ slots under seed 0 (Section 2, Appendix M). Retrieval recall is 1.0 for every method.
3. **The evaluation cases.** A fixed deterministic subsample per dataset (seed 0), reused by every method.
4. **The teacher.** All distilled traces come from one frozen Qwen2.5-32B-Instruct at temperature 0.7, top-p 0.9, seed 0, on an identical case set and ordering (Appendix C).
5. **Optimization.** Identical optimizer, learning rate, batch size, precision, and epoch count for NO-COT, CoT, and every CARE variant (Table 7); the variants differ only in how g_{r} is combined with g_{a}.

A.2 MODELS

Instruction-tuned Qwen2.5 backbones (Qwen Team, 2024) at 3B and 7B for all training experiments; 32B as the frozen teacher and as the largest zero-shot reference; DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025) as a reasoning-tuned zero-shot comparison. To test whether our findings are specific to one pretraining family we repeat the core comparison on instruction-tuned Llama-3.2-1B and Llama-3.1-8B (Llama Team, AI @ Meta, 2024), holding the pipeline, shortlists, prompts and evaluation cases fixed (Appendix S.1). Serving and decoding parameters are in Appendix L.3.

A.3 POINTERS

Dataset statistics and preprocessing: Table 6. Splits and candidate construction: Appendix L.2. Zero-shot serving stack and decoding: Appendix L.3. Supervised fine-tuning hyperparameters: Table 7. CARE rules and hyperparameters: Appendix J. Outcome-reward training: Appendix K. Prompt templates: Appendix B. Ranking extraction: Appendix D.

A.4 METRICS AND RANKING EXTRACTION

NDCG@k and Recall@k for $k \in \{1, 5, 10\}$; with one relevant item per case, Recall@k and HR@k coincide and NDCG@1 is top-1 accuracy (Appendix L.4). One methodological point affects how the two regimes may be compared: zero-shot models are prompted to emit a permutation as text, which we parse and which can fail, whereas trained models are scored by ARGSORT over the candidate letter logits, the readout they were trained on, which never fails. Comparisons *within* each regime are like-for-like; across the two, see Appendix D.

B PROMPT TEMPLATES

The reranking prompt has four parts: a system role, the user history (most recent L interactions as text), the $K{=}20$ candidate list labeled (A)–(T) in shuffled order, and an instruction to output `Final Ranking: ...`. This is the prompt referred to as x in Section 2; the answer y is the letter of the logged item within the candidate list.

C TEACHER CHAIN-OF-THOUGHT: BLIND AND HINDSIGHT

This appendix details the two teacher chain-of-thought (CoT) regimes introduced in Section 3.1: the answer-agnostic BLIND trace and the answer-conditioned HINDSIGHT trace. Both are distilled from the same frozen teacher and differ *only* in what the teacher may condition on when writing the trace.

C.1 SETUP AND NOTATION

For each training case i we have a user history H_i, a candidate set C_i of $K{=}20$ items presented in shuffled letter slots $A, \ldots, T$, and the gold next item y_i. A frozen teacher $\mathcal{T}$ (Qwen2.5-32B-Instruct, bf16, temperature 0.7, top-p 0.9, seed 0) produces a reasoning trace r_i, which is sentence-split into steps and used as the reasoning supervision for CoT-SFT and the Coconut curriculum. We use Qwen2.5-32B-Instruct as the teacher because it is the strongest zero-shot reranker in our study, consistently leading every smaller backbone across datasets (c.f. Table 14); it therefore provides the highest-quality reasoning traces available to distill from. The two regimes differ only in the teacher's conditioning information:

Table 2: The two teacher-CoT regimes. They share the teacher, decoding parameters, case set, and ordering; they differ only in whether the gold next item y_i is revealed to the teacher.

Regime	Teacher sees	Answer access	Role
BLIND	H_i, C_i	✗ (answer-agnostic)	deployable reasoning a student could plausibly reproduce at test time
HINDSIGHT	H_i, y_i	✓ (answer-conditioned, then scrubbed)	reasoning upper bound: "given the answer, what rationale justifies it?"

Zero-shot reranking prompt (Amazon Beauty)

```
SYSTEM:
You are a recommendation assistant for beauty products. Given a
user's interaction history and a list of candidate items, rank the
candidates from most to least likely to be the item the user will
purchase next.

USER HISTORY (oldest to newest):
1. Title: Neutrogena Hydro Boost Gel. Price: $15.99. Brand:
Neutrogena. Category: Moisturizers.
2. Title: CeraVe Foaming Cleanser. Price: $11.49. Brand: CeraVe.
Category: Cleansers.
3. Title: The Ordinary Niacinamide 10%. Price: $5.90. Brand: The
Ordinary. Category: Serums.

CANDIDATES (in random order):
(A) Title: Paula's Choice 2% BHA Exfoliant. Price: $32. Brand:
Paula's Choice. Category: Exfoliants.
(B) Title: La Roche-Posay Sunscreen SPF 50. Price: $24. Brand: La
Roche-Posay. Category: Sunscreen.
...
(T) Title: Maybelline Mascara Lash Sensational. Price: $8. Brand:
Maybelline. Category: Mascara.

INSTRUCTION:
Consider the user's history above. You may reason through your choice
first. End your response with exactly one line: Final Ranking: , ,
..., where each label is one of (A)-(T), listed exactly once, ordered
from most to least likely next purchase.
```

Figure 2: Example prompt for zero-shot LLM reranking on Amazon Beauty. The LLM receives the user's recent history as structured text, a shortlist of $K = 20$ candidates labeled (A)–(T), and an instruction to produce a permutation. For reasoning-tuned models (e.g., R1-distilled), the model emits a trace before the final ranking; for plain models, reasoning is optional but permitted by the instruction "you may reason through your choice first."

C.2 BLIND: STANDARD, ANSWER-AGNOSTIC REASONING

The BLIND trace is generated with the standard student "reason step-by-step, then rank" prompt, *not* conditioned on y_i:

$$r_i^{\text{must}} = \mathcal{T}(H_i, C_i;\ \texttt{reason=must}).$$

The teacher emits free reasoning followed by a `Final Ranking:` line; we keep all sentences preceding `Final Ranking:` as reasoning steps, *untruncated*. This is the distinction our code records as `must` vs. `mustfull`: an earlier pipeline capped the trace at 6 steps, whereas the version used throughout this paper retains every step (and additionally stores the raw completion, `cot_raw`). BLIND is the regime that consumes this full, untruncated trace. Because the teacher never sees y_i, the trace is an honest deliberation the student could plausibly reproduce at test time, but its logic is frequently misaligned with the logged target.

BLIND teacher prompt (answer-agnostic)

System: You are a recommendation assistant. Given a user's interaction history and a list of candidate items, reason step by step about the user's preferences, then rank the candidates from most to least likely to be purchased next.

User: USER HISTORY (oldest→newest): $\{H_i\}$
CANDIDATES (shuffled): (A) ... (T) ...
INSTRUCTION: Reason step by step about the user's tastes and needs, then end with exactly one line: Final Ranking: ⟨labels⟩.

C.3 HINDSIGHT: ANSWER-CONDITIONED RATIONALIZATION, LEAKAGE-GUARDED

The HINDSIGHT trace is answer-conditioned:

$$r_i^{\text{orc}} = \mathcal{T}(H_i, y_i; \texttt{profile prompt}).$$

The teacher is *privately* shown the ground-truth next item y_i and asked to write a short user-taste profile (tastes, brand affinities, price sensitivity, product needs) grounded in the history, using y_i only to decide which threads to emphasize. Strict prompt constraints prevent the answer from ever surfacing in the trace. The teacher prompt has the following structure (verbatim):

HINDSIGHT teacher prompt (answer-conditioned, leakage-guarded)

System: You are an expert consumer analyst. Given a shopper's purchase history, you infer who they are, their tastes, brand affinities, price sensitivity, and product needs, grounded strictly in the evidence of their history.

User: SHOPPER PURCHASE HISTORY (oldest→newest): $\{H_i\}$
PRIVATE, the shopper's ACTUAL next purchase (FOR YOUR EYES ONLY; NEVER reveal, name, describe, or hint at it): $\{y_i\}$
STRICT RULES: do not name, quote, or hint at the next item (name, brand, category, or price); no ``the actual next item is...''; stay at the level of tastes and needs, not specific products; no ranked list; 3--5 sentences.

C.4 WORKED EXAMPLE

We show a real case (`amazon_beauty`, row 0). Gold answer $y_i =$ (J) *Alba Botanica Pore Purifying Pineapple Enzyme Facial Cleanser*.

amazon_beauty, row 0, gold answer (J)

User history (oldest→newest): (1) Clarisonic Replacement Brush Head (Delicate); (2) Tree Hut Shea Body Butter; (3) Alba Botanica Very Emollient Body Lotion.
Candidates: (A) Versace men's EDT; (B) SHANY makeup train case; (C) Nivea body wash; … (J) Alba Botanica … Facial Cleanser; … (20 total).

BLIND (answer-agnostic, 58 steps, excerpt):
"Let's analyze the user's history and the candidates step by step. User History Analysis: the user has shown interest in skin-care products, particularly body moisturizers (body butter, lotion) … also purchased a replacement brush head for a face device, suggesting facial-skincare interest … No history of hair-care or fragrance products. … " → ends in `Final Ranking: …`

HINDSIGHT (answer-conditioned, leakage-scrubbed, 3 sentences):
"This shopper demonstrates a clear preference for skincare that prioritizes moisture and gentle care (Clarisonic Delicate head, Tree Hut Shea Body Butter, Alba Botanica Emollient Lotion). They show brand loyalty, particularly to Alba Botanica and Tree Hut. Given purchases for both face and body, they will likely seek additional gentle, natural skincare, possibly focusing on face care to complement their body-care routine."

What the example illustrates. The HINDSIGHT trace surfaces exactly the two threads that pick out (J), Alba Botanica brand loyalty and "face care to complement body care", yet never names, categorizes, or hints at the cleanser, satisfying the leakage guard. The BLIND trace is generic candidate bookkeeping with no such targeting. This is the intended contrast: HINDSIGHT is optimally targeted but answer-derived (an upper bound), while BLIND is honest, deployable reasoning.

C.5 METHODS SUMMARY

We distill two reasoning supervisions from a frozen Qwen2.5-32B-Instruct teacher over the training split (temperature 0.7, seed 0), on an identical case set and ordering. BLIND-CoT is answer-agnostic:

the teacher reasons step-by-step over the history and candidate list, and we retain the full rationale preceding its final ranking. HINDSIGHT-CoT is answer-conditioned: the teacher privately observes the ground-truth next item and writes a 3–5-sentence user-taste profile, instructed to use the answer only to prioritize which evidence to emphasize while being explicitly forbidden from naming, describing, or hinting at it (enforced by construction via the prompt constraints). Only training cases receive HINDSIGHT rationales; validation and test never do. HINDSIGHT-CoT thus provides an information-targeting upper bound on the reasoning, against which the deployable BLIND-CoT is compared.

D RANKING EXTRACTION: ZERO-SHOT VS. TRAINED SFT

Because we only ever observe a single label (the gold next item y_i) and never a ground-truth *ordering*, the ranking over the $K{=}20$ candidates is produced by two fundamentally different mechanisms in the zero-shot (ZS) and trained (SFT) settings. We make this explicit here, as it is relevant for interpreting the results.

Zero-shot: listwise, generated as text. The untrained model is prompted to emit a full ordered list, e.g. `Final Ranking: C, A, T, F, ...`, which a parser reads into an item order. The ranking is therefore whatever permutation the model writes (listwise generation). Being free-form, the output can be malformed (hallucinated labels, duplicates, missing items, wrong count); the parser handles these failures, and any item that is never surfaced is scored as a miss. ZS quality is thus entangled with the model's ability to *format* a coherent full permutation and is sensitive to prompt and decoding.

Trained SFT: pointwise, read from letter-logits. SFT is trained on single gold-letter supervision: the model learns to emit `###` ⟨`letter`⟩ for the next item. At evaluation we therefore do not parse a generated list. Instead we (i) bring the model to the answer position, for CoT, greedily generate the reasoning until the `###` marker; for No-CoT, simply append `###`; (ii) read the next-token distribution restricted to the 20 candidate letter tokens $A, \ldots, T$; and (iii) ARGSORT these 20 logits to obtain the ranking. The SFT ranking is thus

$$\pi_\theta(x) \;=\; \text{ARGSORT}\; p_\theta\big(\text{letter} \;\big|\; \text{prompt}\,[+\,\text{reasoning}]\big),$$

a *pointwise* scoring readout: each candidate is scored by the model's confidence that it is the next item, then sorted. An ARGSORT of 20 numbers is always a valid, complete permutation, so this protocol has no parsing step and no coverage failures.

Table 3: The ranking is produced differently in the two settings. The SFT logit-readout is the faithful evaluation of what the model was trained to do (pointwise next-item prediction) and is failure-free; the ZS listwise protocol elicits a ranking the model was never trained to produce and carries a parsing-failure floor.

	Zero-shot	Trained SFT
How ranking is formed	model generates a text permutation	ARGSORT of the 20 letter-logits
Style	listwise (writes the whole order)	pointwise (scores each candidate, then sorts)
Parsing	needed; can fail	none; ARGSORT is always valid
Robustness	fragile (malformed output → misses)	deterministic, always complete
What is optimized	nothing (prompted only)	exactly this readout (target is the gold letter)

Why this protocol. SFT is trained pointwise, so the logit readout is the faithful evaluation of what it learned, and it is clean (an ARGSORT of 20 scores is always a valid ranking, with no parsing errors). ZS models were never trained on this answer format, so a ranking can only be elicited by prompting for the list and parsing it. Consequently, the ZS numbers measure *elicited* listwise ranking under a fixed prompt, whereas the SFT numbers measure the *trained* pointwise scoring readout.

Part B: Theory and Proofs

E Formal Setup

Definition 5 (Informativeness, learnability, and answer difficulty)**.** *For a trace z and a pair (x, y), define the* true *and* computational *informativeness of the trace as*

$$\mathcal{I}_{\mathrm{true}}(z; x, y) \;=\; \log \frac{p^*(y \mid x, z)}{p^*(y \mid x)}, \qquad \mathcal{I}_{\mathrm{comp}}(z; x, y) \;=\; \log \frac{p_{\hat{\theta}^{\mathrm{CoT}}}(y \mid x, z)}{p_{\hat{\theta}^{\text{No-CoT}}}(y \mid x)}, \tag{8}$$

and write I_π^{comp} for the expectation of $\mathcal{I}_{\mathrm{comp}}$ under π. Define further the reasoning learnability *and the* answer-difficulty constant *B_y, the latter being the smallest value for which the stated inclusion holds:*

$$\delta_{\mathrm{KL}} \;=\; \mathbb{E}_x\Big[D_{\mathrm{KL}}\big(p_\pi^*(z \mid x) \,\big\|\, p_{\hat{\theta}^\pi}(z \mid x)\big)\Big], \qquad -\log p_{\hat{\theta}^\pi}(y \mid z, x) \;\in\; [0,\, B_y] \tag{9}$$

for all (x, y) and all z in the support of π.

$\mathcal{I}_{\mathrm{true}}$ is information the trace carries *about the world*, a property of the task; $\mathcal{I}_{\mathrm{comp}}$ is how much it helps *this model* compute an answer it could not compute in one pass, a property of the fitted models. δ_{KL} measures how closely the student's trace distribution approximates p_π^*, vanishing under exact reproduction and large when the target is unlearnable from the prompt, so it is the term that penalizes HINDSIGHT-style supervision, whose target depends on information the student never sees at test time. B_y is small when the prompt essentially determines the answer and large when the next-item law is diffuse: a predictor facing M comparable continuations incurs answer negative log-likelihood of order $\log M$ on whichever the log records, however well specified it is.

E.1 The chain-of-thought risk bound

The bound below decomposes the test risk of the chain-of-thought model into the risk it would attain under the target reasoning distribution, plus the downside of sampling traces from its own imperfect approximation to it.

Theorem 6 (When chain-of-thought improves over direct prediction)**.** *Let $\hat{\theta}^\pi, \hat{\theta}^{\text{No-CoT}}$ and B_y be as above, and let $R(\cdot)$ denote the population test risk with traces sampled from the model itself. Then $R(\hat{\theta}^\pi) \le R(\hat{\theta}^{\text{No-CoT}}) - \underbrace{I_\pi^{\mathrm{comp}}}_{\textit{benefit}} + \underbrace{B_y\sqrt{\delta_{\mathrm{KL}}/2}}_{\textit{cost}}$, and consequently chain-of-thought improves over direct prediction whenever Eq. equation 2 holds.*

Proof. See Appendix F. □

Condition equation 2 trades a benefit, I_π^{comp}, against a downside, namely the penalty for failing to reproduce the target reasoning at test time, amplified by answer difficulty. Being one-sided it is sufficient but not necessary: violating it renders the bound uninformative rather than establishing harm. That asymmetry does not weaken what follows, since the violation we measure is not marginal. The next lemma closes the informational channel exactly, without any modelling assumption.

This appendix states formally the two trained models and the trace policy that Section 2 introduces in prose.

Definition 7 (The two trained models)**.** *Let $\mathcal{D} = \{(x_i, z_i, y_i)\}_{i=1}^n$ be a training set of prompts, teacher traces, and logged items. Define the two empirical risk minimizers*

$$\hat{\theta}^{\mathrm{CoT}} \;\in\; \arg\min_\theta \; \frac{1}{n}\sum_{i=1}^n \Big[-\log p_\theta(z_i \mid x_i) \;-\; \log p_\theta(y_i \mid z_i, x_i)\Big], \tag{10}$$

$$\hat{\theta}^{\text{No-CoT}} \;\in\; \arg\min_\theta \; \frac{1}{n}\sum_{i=1}^n \Big[-\log p_\theta(y_i \mid x_i)\Big], \tag{11}$$

that is, the minimizer of the chain-of-thought objective equation 1 and the minimizer of the answer-span loss alone. Both are initialized at the same θ_0 and trained with the same optimizer, learning rate, batch size, and step budget, so that any difference between them is attributable to the reasoning supervision rather than to training effort. Lemma 2 relies on precisely this fact.

Definition 8 (Trace policy and target reasoning distribution)**.** *A* trace policy π *is the procedure that produces the supervision traces z_i in $\mathcal{D}$; in our experiments this is a frozen teacher model together with its prompt. A policy induces a* target reasoning distribution $p^*_\pi(z \mid x)$ *over traces, which is the distribution the student is asked to reproduce and against which the reasoning term of Eq. equation 1 is supervised. We write $\hat{\theta}^\pi$ for the chain-of-thought minimizer obtained under policy π. The two policies used in this paper are the answer-agnostic* BLIND *teacher and the answer-conditioned* HINDSIGHT *teacher of Section 3.1, whose target distributions we denote p^*_{BL} and p^*_{HS} respectively.*

Note that p^* without a subscript is the Bayes-optimal predictor over *answers* y, whereas p^*_π is a distribution over *traces* z; the two are distinct objects.

E.2 Extended discussion: the two senses of a useful trace

This is the long form of Section 2. A trace z may *resolve genuine uncertainty* about the answer, or *make an already-determined answer easier for a bounded model to compute*; much of the confusion surrounding chain-of-thought stems from conflating these. Separating them requires comparing two trained models that differ only in objective: $\hat{\theta}^{\mathrm{CoT}}$, which minimizes Eq. equation 1 over a training set $\{(x_n, y_n, z_n)\}_{n=1}^N$ with traces drawn from a teacher $z_n \sim \pi(\cdot \mid x_n)$, and $\hat{\theta}^{\text{No-CoT}}$, which minimizes the answer span alone, $-\log p_\theta(y \mid x)$. The two differ in exactly one respect: $\hat{\theta}^{\mathrm{CoT}}$ carries the extra term $\mathcal{L}_{\mathrm{reason}}$ and conditions its answer on the trace it has just written. They are otherwise identically initialized and given the same optimizer, data and step budget, so every difference we report between them is attributable to $\mathcal{L}_{\mathrm{reason}}$ (Definition 7, Appendix E).

The trace policy π, in our experiments a frozen teacher with its prompt, induces a *target reasoning distribution* $p^*_\pi(z \mid x)$ that the student must reproduce; we write $\hat{\theta}^\pi$ for the corresponding minimizer. The two policies used here are the answer-agnostic BLIND and answer-conditioned HINDSIGHT teachers of Section 3.1; their prompts, decoding settings, the leakage guard applied to HINDSIGHT, and side-by-side example traces are in Appendices C.2, C.3 and C.4. Note that p^* without a subscript is the Bayes-optimal predictor over *answers*, whereas p^*_π is a distribution over *traces* (Definition 8).

Three quantities govern what follows, defined formally as Definition 5 in Appendix E. The *true* informativeness,

$$\mathcal{I}_{\mathrm{true}} = \log \frac{p^*(y \mid x, z)}{p^*(y \mid x)},$$

is what the trace reveals *about the world*, a property of the task. The *computational* informativeness,

$$\mathcal{I}_{\mathrm{comp}} = \log \frac{p_{\hat{\theta}^{\mathrm{CoT}}}(y \mid x, z)}{p_{\hat{\theta}^{\text{No-CoT}}}(y \mid x)},$$

is how much it helps *this model* compute an answer it could not reach in one pass, a property of the fitted models; we write I^{comp}_π for its expectation under π. The *reasoning learnability* $\delta_{\mathrm{KL}} = \mathbb{E}_x[D_{\mathrm{KL}}(p^*_\pi(z \mid x) \,\|\, p_{\hat{\theta}^\pi}(z \mid x))]$ measures how closely the student reproduces the target, and is the term that penalizes HINDSIGHT-style supervision. Finally B_y bounds the answer negative log-likelihood: small when the prompt determines the answer, of order $\log M$ when the next-item distribution is diffuse over M comparable continuations.

Why recommendation differs from mathematics. Mathematics, code and logic sit at $\mathcal{I}_{\mathrm{true}} = 0$ with $\mathcal{I}_{\mathrm{comp}} \gg 0$: the answer is fixed by the statement but a bounded model cannot reach it in one pass, so the trace is a computational scaffold. Two properties make that work, *determinacy* (so B_y is small) and *verifiable compositionality* (checkable steps that compose, making the trace constructive and its target canonical). Recommendation has neither, with a consequence that separates it sharply from those domains: **no corpus of ground-truth reasoning exists, and none can exist even in principle**, since labelling *why* a user chose an item demands the counterfactual that is the prediction

target. Every supervision trace must therefore be synthesized, so p^*_π is a modelling decision and δ_{KL} is a live term here but negligible in mathematics (Table 4, Appendix I).

F Proof of Theorem 6

Proof. We decompose the test risk, bound the distribution-shift term via Pinsker's inequality, and re-express the remainder through computational informativeness.

Step 1 (Decomposition via oracle reasoning). Define the *oracle risk*, the risk attained if the target reasoning distribution p^*_π could be used at inference:

$$R_{\mathrm{oracle}}(\hat{\theta}^\pi) = \mathbb{E}_{(x,y)}\, \mathbb{E}_{z\sim p^*_\pi(\cdot|x)}\big[-\log p_{\hat{\theta}^\pi}(y \mid z, x)\big]\,, \tag{12}$$

and write $R(\hat{\theta}^\pi) = R_{\mathrm{oracle}}(\hat{\theta}^\pi) + \Delta_{\mathrm{shift}}$, where Δ_{shift} is the downside of sampling from the model's own $p_{\hat{\theta}^\pi}(z \mid x)$ instead of $p^*_\pi(z \mid x)$.

Step 2 (Bounding Δ_{shift}). Fix (x, y) and let $f(z) = -\log p_{\hat{\theta}^\pi}(y \mid z, x) \in [0, B_y]$. By the variational characterization of total variation, for any bounded f, $|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \le B_y\, \mathrm{TV}(P, Q)$. Applying Pinsker's inequality $\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(Q\|P)}$ with $P = p_{\hat{\theta}^\pi}$, $Q = p^*_\pi$, then taking expectation over x and using Jensen's inequality (concavity of $\sqrt{\cdot}$),

$$\Delta_{\mathrm{shift}} \;\le\; B_y\, \mathbb{E}_x\left[\sqrt{\tfrac{1}{2} D_{\mathrm{KL}}\big(p^*_\pi(z \mid x)\,\|\,p_{\hat{\theta}^\pi}(z \mid x)\big)}\right] \;\le\; B_y \sqrt{\tfrac{\delta_{\mathrm{KL}}}{2}}. \tag{13}$$

Step 3 (Re-expressing the oracle risk). Add and subtract $\log p_{\hat{\theta}^{\text{no-CoT}}}(y \mid x)$ inside R_{oracle}:

$$\begin{aligned} R_{\mathrm{oracle}}(\hat{\theta}^\pi) &= \mathbb{E}_{(x,y)}\big[-\log p_{\hat{\theta}^{\text{no-CoT}}}(y \mid x)\big] \\ &\quad - \mathbb{E}_{(x,y)}\, \mathbb{E}_{z\sim p^*_\pi}\big[\log p_{\hat{\theta}^\pi}(y \mid z, x) - \log p_{\hat{\theta}^{\text{no-CoT}}}(y \mid x)\big] \\ &= R(\hat{\theta}^{\text{no-CoT}}) - I^{\mathrm{comp}}_\pi, \end{aligned} \tag{14}$$

where the first term is the no-CoT risk by definition and the second is exactly the expected computational informativeness I^{comp}_π (Definition 5).

Step 4 (Combine).

$$R(\hat{\theta}^\pi) = R_{\mathrm{oracle}}(\hat{\theta}^\pi) + \Delta_{\mathrm{shift}} \;\le\; R(\hat{\theta}^{\text{no-CoT}}) - I^{\mathrm{comp}}_\pi + B_y \sqrt{\tfrac{\delta_{\mathrm{KL}}}{2}}, \tag{15}$$

which is the risk bound of Theorem 6. The equivalence $R(\hat{\theta}^\pi) < R(\hat{\theta}^{\text{no-CoT}}) \iff I^{\mathrm{comp}}_\pi > B_y \sqrt{\delta_{\mathrm{KL}}/2}$ follows immediately, giving Eq. equation 2. □

G Proof of Lemma 2

G.1 The two training trajectories

The construction Section 2.2 compresses is stated here in full.

As defined in Eq. equation 8, I^{comp}_π compares two fully trained models, which says nothing about where the harm comes from and offers no way to measure it short of training both. We therefore re-express it as a property of the optimization: since the two models differ only by the extra gradient the reasoning loss contributes, the damage must be visible in how that gradient sits relative to the ranking one. Write $g_{\mathrm{a}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{rank}}(\theta)$ and $g_{\mathrm{r}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{reason}}(\theta)$, each obtained by masking the labels of the complementary span (Appendix O), and train the two models of Definition 7 by gradient descent from a shared θ_0 for T steps at learning rate η. Writing θ^N_t, θ^C_t for the iterates,

$$\theta^N_{t+1} \;=\; \theta^N_t - \eta\, g_{\mathrm{a}}\big(\theta^N_t\big), \qquad \theta^C_{t+1} \;=\; \theta^C_t - \eta\Big(g_{\mathrm{a}}\big(\theta^C_t\big) + g_{\mathrm{r}}\big(\theta^C_t\big)\Big), \qquad \theta^N_0 = \theta^C_0 = \theta_0, \tag{16}$$

so that $\hat{\theta}^{\text{No-CoT}} = \theta^N_T$ and $\hat{\theta}^{\mathrm{CoT}} = \theta^C_T$. The trajectories differ only by the additional term $-\eta\, g_{\mathrm{r}}$, which is the entire object of study. We evaluate both gradients along a single common reference trajectory, taking the no-CoT one, and abbreviate $g^t_{\mathrm{a}} := g_{\mathrm{a}}(\theta^N_t)$ and $g^t_{\mathrm{r}} := g_{\mathrm{r}}(\theta^N_t)$.

Recall the two update rules equation 16, the shared initialization $\theta_0^N = \theta_0^C = \theta_0$, and the reference gradients $g_{\mathrm{a}}^t = g_{\mathrm{a}}(\theta_t^N)$, $g_{\mathrm{r}}^t = g_{\mathrm{r}}(\theta_t^N)$. Assumptions (A1)–(A4) of Lemma 2 enter at Steps 2, 2, 3 and 1 respectively.

Proof. Since $\theta_0^N = \theta_0^C$ and the two updates in equation 16 differ only by the extra term $-\eta\, g_{\mathrm{r}}(\theta_t^C)$, an induction on t gives $\theta_t^C - \theta_t^N = O(\eta)$ for every fixed $t \le T$: at $t = 0$ the difference is 0, and each step adds a term of order η plus an L-Lipschitz contraction/expansion of the existing $O(\eta)$ gap, which stays $O(\eta)$ over a finite horizon. By L-Lipschitzness of the gradients it follows that $g(\theta_t^C) = g(\theta_t^N) + O(\eta)$ for $g \in \{g_a, g_r\}$.

By a first-order Taylor expansion of R_a (valid by (A1)) and using $\nabla R_a = g_{\mathrm{a}}$ (A2), along the CoT trajectory

$$\begin{aligned} R_a(\theta_{t+1}^C) - R_a(\theta_t^C) &= \big\langle \nabla R_a(\theta_t^C),\ \theta_{t+1}^C - \theta_t^C \big\rangle + O(\eta^2) \\ &= -\eta \,\big\langle g_a(\theta_t^C),\ g_a(\theta_t^C) + g_r(\theta_t^C) \big\rangle + O(\eta^2), \end{aligned} \tag{17}$$

using $\nabla R_a = g_{\mathrm{a}}$ (A2) and $\theta_{t+1}^C - \theta_t^C = -\eta(g_{\mathrm{a}} + g_{\mathrm{r}})(\theta_t^C)$ from Eq. equation 16. Likewise, along the no-CoT trajectory,

$$R_a(\theta_{t+1}^N) - R_a(\theta_t^N) = -\eta \,\big\langle g_a(\theta_t^N),\ g_a(\theta_t^N) \big\rangle + O(\eta^2). \tag{18}$$

Each per-step change carries an explicit prefactor η, so perturbing the gradients inside it by $O(\eta)$ requires only $O(\eta^2)$. Using $g(\theta_t^C) = g(\theta_t^N) + O(\eta) = g^t + O(\eta)$ from Step 1 and the boundedness of the gradients,

$$\begin{aligned} R_a(\theta_{t+1}^C) - R_a(\theta_t^C) &= -\eta \big\langle g_a^t + O(\eta),\ (g_a^t + g_r^t) + O(\eta) \big\rangle + O(\eta^2) \\ &= -\eta \big(\|g_a^t\|^2 + \langle g_a^t, g_r^t \rangle \big) + O(\eta^2), \end{aligned} \tag{19}$$

where every cross term is $\eta \cdot (\text{bounded, by (A3)}) \cdot O(\eta) = O(\eta^2)$. Identically, $R_a(\theta_{t+1}^N) - R_a(\theta_t^N) = -\eta \|g_a^t\|^2 + O(\eta^2)$. The two trajectories are therefore never assumed *equal*: they are only $O(\eta)$-close (Step 1), and that discrepancy affects these first-order terms only at second order, which is why we may evaluate both on the common reference (no-CoT) trajectory $\{\theta_t^N\}$.

Summing each identity over $t = 0, \dots, T-1$ and using the shared initialization $R_a(\theta_0^C) = R_a(\theta_0^N)$,

$$\begin{aligned} R_a(\theta_T^C) - R_a(\theta_T^N) &= \sum_{t=0}^{T-1} \Big[\big(R_a(\theta_{t+1}^C) - R_a(\theta_t^C) \big) - \big(R_a(\theta_{t+1}^N) - R_a(\theta_t^N) \big) \Big] \\ &= -\eta \sum_{t=0}^{T-1} \langle g_a^t, g_r^t \rangle + O(\eta^2). \end{aligned} \tag{20}$$

Therefore $I_\pi^{\mathrm{comp}} = R_a(\theta_T^N) - R_a(\theta_T^C) = \eta \sum_{t=0}^{T-1} \langle g_a^t, g_r^t \rangle + O(\eta^2)$, and expanding the inner product as $\langle g_a^t, g_r^t \rangle = \|g_a^t\|\, \|g_r^t\| \cos(g_a^t, g_r^t)$ gives Eq. equation 3. The sign claim is immediate since the norms are non-negative. □

Remark 9 (Weighted form, and why ANNEAL works)**.** *If the reasoning loss carries a per-step weight λ_t, so that the CoT update in Eq. equation 16 becomes $\theta_{t+1}^C = \theta_t^C - \eta\big(g_{\mathrm{a}}(\theta_t^C) + \lambda_t\, g_{\mathrm{r}}(\theta_t^C)\big)$, the identical argument gives $I_\pi^{\mathrm{comp}} = \eta \sum_t \lambda_t \langle g_{\mathrm{a}}^t, g_{\mathrm{r}}^t \rangle + O(\eta^2)$. Two consequences used in Section 4: driving $\lambda_t \to 0$ after the epoch at which the cosine turns negative (*ANNEAL*, Eq. equation 6) removes precisely the negative terms of the sum, and replacing g_{r}^t by any $\tilde{g}_{\mathrm{r}}{}^t$ with $\langle g_{\mathrm{a}}^t, \tilde{g}_{\mathrm{r}}{}^t \rangle \ge 0$ (*GPROJ*, Eq. equation 23) makes every term non-negative, hence $I_\pi^{\mathrm{comp}} \ge 0$ to first order.*

G.2 PROOF OF REMARK 9

Suppose the reasoning loss carries a per-step weight $\lambda_t \ge 0$, so that the CoT update in Eq. equation 16 is replaced by

$$\theta_{t+1}^C = \theta_t^C - \eta\Big(g_{\mathrm{a}}\big(\theta_t^C\big) + \lambda_t\, g_{\mathrm{r}}\big(\theta_t^C\big) \Big), \qquad \theta_0^C = \theta_0.$$

Steps 1 and 3 are unchanged provided $\sup_t \lambda_t < \infty$: the per-step perturbation is still $O(\eta)$ over a fixed horizon. Step 2 becomes

$$R_a(\theta_{t+1}^C) - R_a(\theta_t^C) = -\eta \big(\|g_{\mathrm{a}}^t\|^2 + \lambda_t \langle g_{\mathrm{a}}^t, g_{\mathrm{r}}^t \rangle \big) + O(\eta^2),$$

and telescoping against the unchanged no-CoT recursion $R_a(\theta_{t+1}^N) - R_a(\theta_t^N) = -\eta\|g_{\mathrm{a}}^t\|^2 + O(\eta^2)$ gives

$$I_\pi^{\mathrm{comp}} \;=\; \eta \sum_{t=0}^{T-1} \lambda_t \left\langle g_{\mathrm{a}}^t, g_{\mathrm{r}}^t \right\rangle + O(\eta^2).$$

The same computation with g_{r}^t replaced by a modified reasoning gradient $\tilde{g_{\mathrm{r}}}^t$ (as in GPROJ, Appendix J) yields $I_\pi^{\mathrm{comp}} = \eta \sum_t \langle g_{\mathrm{a}}^t, \tilde{g_{\mathrm{r}}}^t \rangle + O(\eta^2)$, which is non-negative to first order whenever the modification enforces $\langle g_{\mathrm{a}}^t, \tilde{g_{\mathrm{r}}}^t \rangle \geq 0$ pointwise, the invariant maintained by Algorithm 1.

H PROOFS FOR THE RATIONALE–LOG DISAGREEMENT

Corollary 10 (Agreement implies alignment; disagreement implies conflict)**.** *Let $p = \mathrm{softmax}(\theta)$ be arbitrary.*

(i) Verifiable, unimodal targets *(mathematics, code, logic). If $m = y$, then $\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle = (1 - p_y)^2 + (\|p\|^2 - p_y^2) \geq 0$, with equality if and only if $p = e_y$.*

(ii) Non-verifiable, multi-modal targets *(recommendation). If $m \neq y$, then $\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle = \|p\|^2 - p_y - p_m$, strictly negative whenever $p_y + p_m > \|p\|^2$. In the balanced bimodal case $p_y = p_m = a$, with the remaining mass negligible,*

$$\left\langle g_{\mathrm{a}}, g_{\mathrm{r}} \right\rangle \;=\; -2a(1-a) \;<\; 0 \qquad \textit{for every } a \in (0,1). \tag{21}$$

This is the geometric core of our diagnosis. When rationale and log disagree, e_m and e_y are distinct vertices of the simplex and the two gradients pull the shared logits toward different corners, so no parameter configuration can both reproduce the rationale and rank the logged item first. The conflict is forced by the geometry of softmax cross-entropy on a shared head, not by a particular optimizer, backbone, or dataset. Note also that $2a(1-a)$ in Eq. equation 21 peaks at $a = \frac{1}{2}$: the conflict is most severe exactly when the model is undecided between two modes, the typical case in recommendation and the atypical one in mathematics.

From a per-example identity to a task-level criterion. Taking expectations converts the identity into a statement about the task. Fix a prompt x. The logged item is a draw $y \sim q(\cdot \mid x)$ from the true next-item distribution, and the rationale-implied item is a draw $m \sim \tilde{q}(\cdot \mid x)$ from the distribution the teacher induces; the two are conditionally independent given x, because the teacher never observes y (Lemma 1).

Proposition 11 (The agreement probability controls the conflict)**.** *Let the expectation be taken over prompts $x \sim \mathcal{D}$ together with $y \sim q(\cdot \mid x)$ and $m \sim \tilde{q}(\cdot \mid x)$, drawn independently given x. Under Assumption 3,*

$$\mathbb{E}_{x,y,m}\left\langle g_{\mathrm{a}}, g_{\mathrm{r}} \right\rangle \;=\; \mathbb{E}_x\|p\|^2 \;-\; \mathbb{E}_{x,y}[p_y] \;-\; \mathbb{E}_{x,m}[p_m] \;+\; \Pr(y{=}m), \tag{5}$$

where $\|p\|^2$ is deterministic given x and carries an expectation only because we average over prompts, and $\Pr(y{=}m) = \mathbb{E}_x\langle q(\cdot \mid x), \tilde{q}(\cdot \mid x)\rangle$. If the trace is verifiable and the posterior unimodal, so that $m = y$ almost surely, then $\Pr(y{=}m) = 1$ and $\mathbb{E}_{x,y,m}\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle \geq 0$. If instead y and m are drawn from distinct distributions, the true next-item distribution $q = p^*(\cdot \mid x)$ and the teacher's implied distribution $\tilde{q}$, then $\Pr(y{=}m) = \langle q, \tilde{q} \rangle \leq \sqrt{\sum_k q_k^2}\sqrt{\sum_k \tilde{q}_k^2}$ by Cauchy–Schwarz, which is small whenever the true posterior is diffuse, or the teacher concentrates away from its mass, or both; the special case $\tilde{q} = q$ recovers the collision probability $\sum_k q_k^2 \approx 1/M$. In either case $\mathbb{E}_{x,y,m}\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle < 0$ whenever $\mathbb{E}_{x,y}[p_y] + \mathbb{E}_{x,m}[p_m] > \mathbb{E}_x\|p\|^2 + \Pr(y{=}m)$.

Proof. See Appendix H. □

One point of interpretation matters for the measurement. The identity governs the *per-example* inner product, so its empirical counterpart is the average of per-example inner products, not the inner product of gradients accumulated over a batch: the latter adds cross-example terms $\langle g_{\mathrm{a}}^{(n)}, g_{\mathrm{r}}^{(n')} \rangle$ with $n \neq n'$ that the identity does not model, and which need not share its sign. Lemma 2, by contrast, concerns the batch gradient that actually drives training, so we report both quantities in Section 3.

This renders the two regimes exact: the only term capable of restoring alignment is $\Pr(y = m)$, one for verifiable unimodal targets and collapsing to the overlap $\langle q, \tilde{q} \rangle$ between the true continuation distribution and the teacher's when the next-item distribution is multi-modal. Section 3.1 measures both factors of that overlap directly. It also predicts the *dynamics*, not merely the eventual sign. Near initialization p is diffuse, giving $\|p\|^2 \approx 1/K$ and $\mathbb{E}[p_y]+\mathbb{E}[p_m] \approx 2/K$, so the sign is indeterminate at order $1/K$; the conflict then switches on as mass concentrates on the two disagreeing modes, until $p_y + p_m > \mathbb{E}\|p\|^2$. The falsifiable prediction is that the cosine should begin non-negative and turn negative after the first epoch, which Section 3.1 confirms.

H.1 Statement of the multi-token readout theorem

The following result assumes no single rationale-implied item, treating the multi-token rationale as it is actually supervised. Let W_U be the shared unembedding, h_a the hidden state at the answer position and $\{h_{r,t}\}_{t=1}^{T}$ those at the reasoning positions, with $p_a = \mathrm{softmax}(W_U h_a)$, $p_{r,t} = \mathrm{softmax}(W_U h_{r,t})$, answer target e_y, teacher-forced targets e_{z_t}, and $\mathcal{L}_{\mathrm{rank}} = -\log p_{a,y}$, $\mathcal{L}_{\mathrm{reason}} = -\sum_t \log p_{r,t,z_t}$.

Theorem 12 (Exact readout factorization and its sign)**.** *The contribution of the shared unembedding to $\langle g_{\mathrm{a}}, g_{\mathrm{r}} \rangle$ factorizes exactly as*

$$G_U \;=\; \sum_{t=1}^{T} \big\langle h_a, h_{r,t} \big\rangle \, F_t, \qquad F_t \;=\; \big\langle p_a, p_{r,t} \big\rangle - p_{a,z_t} - p_{r,t,y} + \mathbf{1}\{y = z_t\}. \tag{22}$$

When the answer and reasoning tokens occupy disjoint sub-vocabularies and the two heads are specialized and confident, $\mathbf{1}\{y = z_t\} = 0$ and $G_U \le 0$; this is the recommendation *regime. When instead $z_t = y$ at the answer-carrying step, $F_t \ge (1 - p_{a,y})(1 - p_{r,t,y}) \ge 0$ and $G_U \ge 0$; this is the* mathematics *regime. The precise conditions and constants are stated as Theorem 16 in Appendix H.*

This recovers Corollary 10 without Assumption 3 and sharpens the diagnosis in the way that matters for a remedy. Of the four terms in F_t, the driver of the negative sign is $-p_{a,z_t}$, present only because of the imitation target e_{z_t}: it is the requirement that the shared head reproduce specific rationale tokens, drawn from a sub-vocabulary disjoint from the answer, that forces $F_t \le 0$. The confidence condition also predicts *when* this takes hold: once each head has committed its opposite-vocabulary mass, which is exactly where the empirical cosine turns negative.

This appendix collects the proofs of the results in Section 2.1: Lemma 1 ($\mathcal{I}_{\mathrm{true}} \equiv 0$), Lemma 4 (the readout identity), Corollary 10 (the two regimes), Proposition 11 (the agreement probability), and Theorem 12 (the multi-token lift). We restate each statement before proving it so the appendix is self-contained.

H.2 Proof of Lemma 1

Proof. By construction the trace is a (possibly stochastic) function of the prompt with no access to the target: $z \sim p_\theta(\cdot \mid x)$, independent of y given x. Hence the joint factorizes as $p(x, y, z) = p(x)\, p^*(y \mid x)\, p_\theta(z \mid x)$, which is exactly the Markov chain $Y - X - Z$. The data-processing inequality gives $I(Y; Z \mid X) = 0$, equivalently $p^*(y \mid x, z) = p^*(y \mid x)$ for every z in the support of $p_\theta(\cdot \mid x)$. Substituting into Definition 5, $\mathcal{I}_{\mathrm{true}}(z; x, y) = \log p^*(y \mid x, z) - \log p^*(y \mid x) = 0$. □

Corollary 13 (A diffuse next-item distribution leaves nothing for reasoning to resolve)**.** *Since y depends on x only through $p^*(y \mid x)$ and z is a function of x, conditioning on z cannot sharpen the posterior over y. When $p^*(y \mid x)$ is multi-modal (large $H(Y \mid X)$: many items are valid continuations), the residual uncertainty is irreducible by any x-derived trace. Hence any benefit of chain-of-thought here must be* computational *($\mathcal{I}_{\mathrm{comp}}$), never* informational*: $\mathcal{I}_{\mathrm{true}} \equiv 0$ regardless of the trace.*

Remark 14 (What Lemma 1 does and does not establish)**.** *It establishes $\mathcal{I}_{\mathrm{true}} = 0$ from the dependence structure alone, with no modeling assumptions: CoT* cannot help informationally. *The stronger claim that CoT* hurts *($\mathcal{I}_{\mathrm{comp}} < 0$) is a separate, geometry-dependent statement about the shared readout and is* not *implied by $p^*(y \mid x)$ on its own, since math and code also satisfy $\mathcal{I}_{\mathrm{true}} \approx 0$ yet attain $\mathcal{I}_{\mathrm{comp}} \gg 0$. That is precisely what Lemma 4 and Theorem 12 supply.*

H.3 Proof of Lemma 4

Proof. Under Assumption 3 the two losses are $\mathcal{L}_{\text{rank}}(\theta) = -\log p_y$ and $\mathcal{L}_{\text{reason}}(\theta) = -\log p_m$ with $p = \text{softmax}(\theta)$. For a softmax cross-entropy with target class c, $\nabla_\theta(-\log p_c) = p - e_c$. Hence $g_{\text{a}} = p - e_y$ and $g_{\text{r}} = p - e_m$. By bilinearity,

$$\langle p - e_y,\, p - e_m\rangle = \langle p, p\rangle - \langle p, e_m\rangle - \langle e_y, p\rangle + \langle e_y, e_m\rangle = \|p\|^2 - p_m - p_y + \mathbf{1}\{y = m\},$$

using $\langle p, e_k\rangle = p_k$ and $\langle e_y, e_m\rangle = \mathbf{1}\{y = m\}$. □

H.4 Proof of Corollary 10

Proof. Substitute the indicator into Eq. equation 4.

(1) $m = y$. Then $\langle g_{\text{a}}, g_{\text{r}}\rangle = \|p\|^2 - 2p_y + 1 = (1-p_y)^2 + (\|p\|^2 - p_y^2)$. Both terms are non-negative: $(1 - p_y)^2 \geq 0$ trivially, and $\|p\|^2 = \sum_k p_k^2 \geq p_y^2$. Equality requires $p_y = 1$, i.e. $p = e_y$.

(2) $m \neq y$. Then $\langle g_{\text{a}}, g_{\text{r}}\rangle = \|p\|^2 - p_y - p_m$, which is negative precisely when $p_y + p_m > \|p\|^2$. In the balanced bimodal case $p_y = p_m = a$ with the remaining mass negligible, $\|p\|^2 = 2a^2 + o(1)$ and $p_y + p_m = 2a$, so $\langle g_{\text{a}}, g_{\text{r}}\rangle = 2a^2 - 2a = -2a(1 - a)$, which is strictly negative for every $a \in (0, 1)$ since $a^2 < a$ on that interval. □

H.5 Proof of Proposition 11

Proof. Take the expectation of Eq. equation 4 first over (y, m) at fixed x, then over x. At fixed x the term $\|p\|^2$ is deterministic, so it passes through unchanged; averaging over prompts turns it into $\mathbb{E}_x\|p\|^2$. Since $\mathbb{E}[\mathbf{1}\{y = m\}] = \Pr(y{=}m)$, this gives Eq. equation 5. Note that the left-hand side is the mean of the *per-example* inner products; the inner product of gradients accumulated over a batch differs from it by the cross-example terms $\langle g_{\text{a}}^{(n)}, g_{\text{r}}^{(n')}\rangle$, $n \neq n'$, which Eq. equation 4 does not describe.

The unimodal case is Corollary 10(1) in expectation: if $m = y$ almost surely then $\Pr(y = m) = 1$ and the per-example inner product is non-negative pointwise, hence so is its expectation.

For the multi-modal case, y and m are both functions of x (Lemma 1), hence conditionally independent given x, so $\Pr(y = m) = \sum_k \Pr(y{=}k)\Pr(m{=}k)$. With both drawn from the posterior $q = p^*(\cdot \mid x)$ this is $\sum_k q_k^2$, the collision probability of q. For M equal modes, $\sum_k q_k^2 = M \cdot (1/M)^2 = 1/M$; the collision probability equals 1 only for a point mass and is minimized by the uniform distribution over the support. The stated inequality is a rearrangement of Eq. equation 5. □

Remark 15 (The teacher does not sample the true posterior)**.** *An earlier version of this analysis modelled the answer-agnostic teacher's implied item as a draw from the same posterior that generated the log, which gives $\Pr(y = m) = \sum_k q_k^2$. We noted at the time that this was the* favourable *assumption for chain-of-thought, since a teacher drawing from a different distribution $\tilde{q}$ has an even lower agreement probability, $\Pr(y = m) = \sum_k q_k \tilde{q}_k \leq \sqrt{\sum_k q_k^2}\sqrt{\sum_k \tilde{q}_k^2}$ by Cauchy–Schwarz, with equality only when $\tilde{q} = q$.*

Measurement shows the general case is the operative one (Section 3.1). Resampling the BLIND *teacher gives $\sum_k \tilde{q}_k^2 = \Pr(m{=}m') \approx 0.44$–$0.62$, so the teacher's distribution is concentrated on roughly two candidates, whereas its agreement with the log is only 0.07–0.13. The teacher is therefore self-consistent but confidently misdirected: it does not sample the true continuation distribution. Two consequences follow. The conflict is* stronger *than the exchangeable model implies, since a stable commitment to one wrong vertex accumulates across steps in Eq. equation 3 rather than partially cancelling. And the diffuseness of the true posterior can be read off without a proxy model: with $\tilde{q}$ concentrated on about two candidates, $\langle q, \tilde{q}\rangle \approx 0.11$ implies that q places only about 22% of its mass on the candidates a 32B reranker most favours.*

H.6 THEOREM 12: FULL STATEMENT AND PROOF

We restate the theorem with the constants that Section 2.3 suppresses. Let $W_U \in \mathbb{R}^{|V|\times d}$ be the shared unembedding, h_a the hidden state at the answer position, $\{h_{r,t}\}_{t=1}^{T}$ those at the reasoning positions, $p_a = \mathrm{softmax}(W_U h_a)$, $p_{r,t} = \mathrm{softmax}(W_U h_{r,t})$, answer target e_y, teacher-forced reasoning target e_{z_t}, and $\mathcal{L}_{\mathrm{rank}} = -\log p_{a,y}$, $\mathcal{L}_{\mathrm{reason}} = -\sum_t \log p_{r,t,z_t}$.

Theorem 16 (Exact readout factorization and sign; full form)**.** $G_U := \langle \nabla_{W_U}\mathcal{L}_{\mathrm{rank}}, \nabla_{W_U}\mathcal{L}_{\mathrm{reason}}\rangle = \sum_{t=1}^{T}\langle h_a, h_{r,t}\rangle F_t$ *with* $F_t = \langle p_a, p_{r,t}\rangle - p_{a,z_t} - p_{r,t,y} + \mathbf{1}\{y = z_t\}$.

(i) ***Disjoint support (recommendation).*** *Assume* (A1) $y \in V_A$, $z_t \in V_R$, $V_A \cap V_R = \emptyset$; (A2) $\langle h_a, h_{r,t}\rangle \geq 0$; (A3) *each head places mass* $\geq 1-\delta$ *on its own sub-vocabulary and its committed cross-token dominates the residual opposite-vocabulary mass,* $p_{a,z_t} \geq p_a(V_R) - \eta$ *and* $p_{r,t,y} \geq p_{r,t}(V_A) - \eta$. *Then* $\mathbf{1}\{y = z_t\} = 0$ *and, with* $V_0 := V \setminus (V_A \cup V_R)$ *and* $c_0 := \min\{p_a(V_0), p_{r,t}(V_0)\}$,

$$F_t \;\leq\; 2\delta^2 + \delta c_0 - \big(p_a(V_R) + p_{r,t}(V_A)\big) + 2\eta,$$

so $F_t \leq 0$, *and hence* $G_U \leq 0$ *whenever* $p_a(V_R) + p_{r,t}(V_A) \geq 2\delta^2 + \delta c_0 + 2\eta$.

(ii) ***Shared support (math).*** *If* $z_t = y$ *at the answer-carrying step and the answer token lies in the reasoning support, then* $\mathbf{1}\{y = z_t\} = 1$ *and* $F_t = \langle p_a, p_{r,t}\rangle - p_{a,y} - p_{r,t,y} + 1 \geq (1 - p_{a,y})(1 - p_{r,t,y}) \geq 0$, *so with (A2) the contribution is non-negative.*

Proof. **Factorization equation 22.** For a cross-entropy loss $-\log(\mathrm{softmax}(W_U h))_c$, the gradient with respect to W_U is the rank-one matrix $(p - e_c)h^\top$ with $p = \mathrm{softmax}(W_U h)$. Hence $\nabla_{W_U}\mathcal{L}_{\mathrm{rank}} = (p_a - e_y)h_a^\top$ and, by additivity over positions, $\nabla_{W_U}\mathcal{L}_{\mathrm{reason}} = \sum_t (p_{r,t} - e_{z_t})h_{r,t}^\top$. For rank-one matrices $uh_1^\top$ and $vh_2^\top$ the Frobenius inner product is $\langle uh_1^\top, vh_2^\top\rangle_F = \mathrm{tr}(h_1 u^\top v h_2^\top) = \langle u, v\rangle\langle h_1, h_2\rangle$. Applying this termwise gives Eq. equation 22.

Expansion of F_t. Expand $\langle p_a - e_y,\, p_{r,t} - e_{z_t}\rangle$ by bilinearity, using $\langle p_a, e_{z_t}\rangle = p_{a,z_t}$, $\langle e_y, p_{r,t}\rangle = p_{r,t,y}$, and $\langle e_y, e_{z_t}\rangle = \mathbf{1}\{y = z_t\}$.

Case (i): disjoint support. (A1) gives $\mathbf{1}\{y = z_t\} = 0$. Partition $V = V_A \cup V_R \cup V_0$ and split the overlap $\langle p_a, p_{r,t}\rangle = \sum_{v\in V_A} + \sum_{v\in V_R} + \sum_{v\in V_0}$. On V_A we have $p_{r,t}(v) \leq p_{r,t}(V_A) \leq \delta$, so $\sum_{V_A} p_a(v)p_{r,t}(v) \leq \delta\sum_{V_A} p_a(v) \leq \delta$; a symmetric bound gives $\sum_{V_R} \leq \delta$; and $\sum_{V_0} p_a p_{r,t} \leq \min\{p_a(V_0), p_{r,t}(V_0)\} = c_0$. A tighter accounting using $p_a(V_A) \leq 1$ and $p_{r,t}(V_A) \leq \delta$ under the confident-head normalization ($p_a(v) \leq \delta$ for v outside its argmax) yields $\sum_{V_A} \leq \delta^2$, and symmetrically for V_R; combining, $\langle p_a, p_{r,t}\rangle \leq 2\delta^2 + \delta c_0$. By (A3), $p_{a,z_t} + p_{r,t,y} \geq p_a(V_R) + p_{r,t}(V_A) - 2\eta$. Substituting,

$$F_t \;\leq\; 2\delta^2 + \delta c_0 - \big(p_a(V_R) + p_{r,t}(V_A)\big) + 2\eta,$$

so $F_t \leq 0$ under the stated inequality, and $G_U = \sum_t \langle h_a, h_{r,t}\rangle F_t \leq 0$ by (A2).

Case (ii): shared support. With $\mathbf{1}\{y = z_t\} = 1$, $F_t = \langle p_a, p_{r,t}\rangle - p_{a,y} - p_{r,t,y} + 1$. Since $\langle p_a, p_{r,t}\rangle \geq p_{a,y}p_{r,t,y}$ (retaining only the $v = y$ term; all others are non-negative),

$$F_t \;\geq\; p_{a,y}p_{r,t,y} - p_{a,y} - p_{r,t,y} + 1 = (1 - p_{a,y})(1 - p_{r,t,y}) \;\geq\; 0.$$

With (A2), $G_U \geq 0$. □

Remark 17 (Scope of Theorem 12)**.** *The theorem proves the sign of the* readout *contribution G_U. The full inner product $\langle g_a, g_r\rangle$ also receives deeper-layer terms; these are controlled at initialization by Lemma 20 (Appendix N) and tracked empirically thereafter by the per-layer probe (Fig. 3), which finds them negative at every layer but shallower than the readout's. We do not bound them in closed form, and we flag this as a genuine limitation of the result. Condition (A3) formalizes "specialized heads": after a few epochs the answer head commits its opposite-vocabulary mass to the reasoning tokens it must still emit, and vice versa, so $p_a(V_R)$ and $p_{r,t}(V_A)$ exceed the $O(\delta^2)$ incidental overlap. This is exactly the regime in which the empirical cosine turns negative (Table 9), so (A3) is not a convenience assumption but the formal content of the epoch-0-to-epoch-1 transition we measure.*

H.7 Discussion: why the diagnosis is constructive

The three results above isolate the negative driver to a single algebraic object. In Eq. equation 22, the terms $\langle p_a, p_{r,t}\rangle$ and $\mathbf{1}\{y = z_t\}$ are either $O(\delta^2)$ or zero under (A1); the terms $-p_{a,z_t}$ and $-p_{r,t,y}$ are the drivers, and both arise from requiring the shared head to place mass on a *prescribed* token from the opposite sub-vocabulary. Remove the prescription and the drivers disappear:

- Setting $\lambda_t \to 0$ (ANNEAL, Eq. equation 6) removes $\nabla_{W_U}\mathcal{L}_{\text{reason}}$ entirely, so $G_U = 0$.
- Projecting out the conflicting component (GPROJ, Eq. equation 23) enforces $\langle g_{\text{a}}, \tilde{g}_{\text{r}}\rangle \geq 0$ at every step by construction, so the sum in Eq. equation 3 is non-negative.
- Replacing $\mathcal{L}_{\text{reason}}$ with an outcome reward (Section 4.2) removes the target e_{z_t} from the objective, so no term of the form $-p_{a,z_t}$ exists; the reasoning-side update is instead a policy gradient that ascends the same ranking quality $\mathcal{L}_{\text{rank}}$ scores (Proposition 18).

I When CoT Helps: Task Regimes

This appendix expands the argument compressed into Section 2: the two properties that make chain-of-thought work in math, code, and logic, a unique answer and locally verifiable steps, are exactly the two recommendation lacks, and each maps onto a distinct term of Eq. equation 2.

I.1 What makes mathematics, code, and logic amenable to CoT

Where the known successes of chain-of-thought lie. Mathematics, code generation, and logical deduction occupy the regime $\mathcal{I}_{\text{true}} = 0$ with $\mathcal{I}_{\text{comp}} \gg 0$: the answer is determined by the problem statement, but a bounded model cannot compute it in a single pass, so the trace functions as a *computational scaffold*. Two properties make that scaffold effective. The first is **determinacy**, $p^*(y \mid x) = 1$, under which B_y is small. The second is **verifiable compositionality**: the answer decomposes into steps that can be checked against fixed rules independently of the final answer. Together these render the trace constructive, so that $\mathcal{I}_{\text{comp}} \gg 0$, and its target canonical, so that δ_{KL} is small.

Recommendation possesses neither property. Its next-item law is stochastic and multi-modal, so that $p^*(y \mid x) < 1$ and B_y is correspondingly large. Preference is moreover neither compositional nor verifiable: a statement such as "the user prefers affordable skincare" is a hypothesis that can be checked only against the unknown outcome, and no corpus of ground-truth rationales can exist even in principle, since one cannot label *why* a user will choose a given item without the counterfactual that is itself the prediction target. Traces in this setting are therefore descriptive rather than constructive. Table 4 places recommendation against the domains where CoT reliably helps.

I.2 The three failures map one-to-one onto the three quantities

The failures are not a list of grievances; each corresponds to a distinct term of Eq. equation 2. Non-compositionality gives $\mathcal{I}_{\text{comp}} \lesssim 0$: a token-space trace cannot supply the computational scaffold that drives $\mathcal{I}_{\text{comp}} \gg 0$ in the verifiable regime, because the predictive signal (co-occurrence, recency, collaborative geometry) does not decompose into locally checkable steps. The absence of a canonical rationale gives large δ_{KL}: there is no low-KL target for $p_\theta(z \mid x)$ to reach. And stochastic multi-modal targets give large B_y, which *amplifies* the δ_{KL} penalty. Note that structure alone does not fix the sign of $\mathcal{I}_{\text{comp}}$: math and code also satisfy $\mathcal{I}_{\text{true}} \approx 0$ yet attain $\mathcal{I}_{\text{comp}} \gg 0$, which is exactly why Section 2.1 derives the sign rather than asserting it.

Table 4: Why CoT helps elsewhere but not here. Tasks with a unique answer and verifiable, compositional steps yield $\mathcal{I}_{\text{comp}} \gg 0$. Recommendation has neither, so $\mathcal{I}_{\text{true}} \approx 0$ (Lemma 1) and, as Section 2.1 shows, $\mathcal{I}_{\text{comp}} < 0$ under token-level imitation.

Task	Unique answer?	Verifiable steps?	$\mathcal{I}_{\text{true}}$	$\mathcal{I}_{\text{comp}}$	CoT helps?
Arithmetic / math	✓	✓	≈ 0	$\gg 0$	✓
Code generation	✓	✓	≈ 0	$\gg 0$	✓
Logical deduction	✓	✓	≈ 0	$\gg 0$	✓
Creative writing	✗	✗	> 0	≈ 0	partial
Recommendation	✗	✗	≈ 0	< 0 *(imitated)*	✗
Recommendation	✗	✗	≈ 0	≥ 0 *(outcome)*	✓

Table 5: Which term of Eq. equation 2 each method controls, and the resulting outcome. Only $\mathcal{I}_{\text{true}}$ is immovable; the other two are properties of the loss. The last column is measured in Section 5. †Averaged at 7B over the three datasets on which every method has completed (MovieLens-1M, Beauty, Sports); the per-cell values, including the -22.7 extreme on MovieLens-1M, are in Table 1.

Method	Imitation target e_{z_t}	$\langle g_{\text{a}}, g_{\text{r}} \rangle$	δ_{KL}	Outcome vs. NO-COT
SFT-CoT (BLIND)	present, honest	< 0 (measured)	large	hurts (-9.8 N@10)†
SFT-CoT (HINDSIGHT)	present, answer-aligned	≈ 0 (measured)	*very* large	hurts (unlearnable)
CARE-ANNEAL	retired after E_{sw}	$\to 0$	$\to 0$	**beats**
CARE-GPROJ	present, projected	≥ 0 *by constr.*	present	ties/beats
Outcome reward (DPO)	**absent**	≥ 0 *by constr.*	$\to 0$	**beats**
Latent baselines (CoCoNuT, ...)	present (curriculum)	unaddressed	present	ties

I.3 SUMMARY TABLE

Part C: Method Details

J CARE: CONFLICT-AWARE REASONING, FULL SPECIFICATION

This appendix gives the full specification of the two CARE rules introduced in Section 4.1, their algorithmic form, their computational cost, and their relationship to prior multi-task gradient methods.

J.1 WHICH TERM EACH METHOD CONTROLS

Table 5 summarizes the accounting of Section 4: only $\mathcal{I}_{\text{true}}$ is immovable; the other two terms of Eq. equation 2 are properties of the loss, and every method below is classified by which of them it neutralizes.

J.2 DESIGN RATIONALE

All three rules act on the same object: the component of the reasoning gradient that opposes the ranking gradient, which Lemma 2 identifies as the leading-order term of the CoT penalty and Theorem 12 attributes to the token-level imitation target. They differ in *how much* of the reasoning signal they are willing to sacrifice to remove it.

- ANNEAL is the most aggressive: after the switch epoch the reasoning loss is gone, so nothing of the teacher trace constrains the reasoning positions. Those positions are still *generated* at inference, the model still deliberates, but they are free to specialize toward whatever aids the answer. This is why it is the most answer-decodable method in Table 13, and it is the closest a supervised method can come to the outcome-reward regime without rollouts.
- GPROJ is the most conservative: it keeps every component of g_{r} that is not actively fighting g_{a}, so whatever genuine structure the rationale carries (format, candidate bookkeeping, category reasoning) is preserved. Its guarantee is exact but local, non-negative alignment at every step,

and it achieves that guarantee with the trade-off of leaving the imitation target in place, which is why its decodability gain is smaller.

Relation to multi-task gradient surgery. GPROJ is the two-task specialization of PCGrad (Yu et al., 2020), with an asymmetry that matters: we project the *reasoning* gradient onto the orthogonal complement of the *ranking* gradient and never the reverse, because the two objectives are not peers. Ranking is the objective; reasoning supervision is a means.

J.3 GPROJ: GRADIENT PROJECTION

GPROJ: removing only the conflicting component. A more conservative alternative retains the imitation target but discards the part of its gradient that opposes ranking. Following the projection idea from multi-task learning (Yu et al., 2020), whenever the two gradients disagree we project g_{r} onto the orthogonal complement of g_{a}:

$$\tilde{g}_{\mathrm{r}} \;=\; \begin{cases} g_{\mathrm{r}} - \dfrac{\langle g_{\mathrm{r}}, g_{\mathrm{a}} \rangle}{\|g_{\mathrm{a}}\|^2}\, g_{\mathrm{a}}, & \text{if } \langle g_{\mathrm{r}}, g_{\mathrm{a}} \rangle < 0, \\ g_{\mathrm{r}}, & \text{otherwise,} \end{cases} \qquad \theta \;\leftarrow\; \theta - \eta\big(g_{\mathrm{a}} + \lambda\, \tilde{g}_{\mathrm{r}}\big). \tag{23}$$

By construction $\langle g_{\mathrm{a}}, \tilde{g}_{\mathrm{r}} \rangle \geq 0$ at every step, so every term of Eq. equation 3 is non-negative and $I_\pi^{\mathrm{comp}} \geq 0$ to first order: the chain-of-thought penalty is rendered non-positive by construction. The rule is conservative in that it preserves whatever non-conflicting structure the rationale carries, which is consistent with its recovering less answer decodability than ANNEAL.

J.4 ALGORITHM

Algorithm 1 CARE: conflict-aware reasoning supervision (one optimizer step)

Require: batch (x, z, y); parameters θ; rule $\in \{\text{ANNEAL}, \text{GPROJ}\}$; epoch t; lr η; $E_{\mathrm{sw}}, E_{\mathrm{dec}}$

$g_{\mathrm{a}} \leftarrow \nabla_\theta \mathcal{L}_{\mathrm{rank}}(\theta; x, y)$ ▷ answer span only; reasoning labels masked to -100

$g_{\mathrm{r}} \leftarrow \nabla_\theta \mathcal{L}_{\mathrm{reason}}(\theta; x, z)$ ▷ reasoning span only; answer labels masked to -100

if rule $=$ ANNEAL **then**

  $\lambda_t \leftarrow$ Eq. equation 6; $\quad \tilde{g}_{\mathrm{r}} \leftarrow \lambda_t\, g_{\mathrm{r}}$

else if rule $=$ GPROJ **then**

  if $\langle g_{\mathrm{r}}, g_{\mathrm{a}} \rangle < 0$ **then**

    $\tilde{g}_{\mathrm{r}} \leftarrow g_{\mathrm{r}} - \frac{\langle g_{\mathrm{r}}, g_{\mathrm{a}} \rangle}{\|g_{\mathrm{a}}\|^2}\, g_{\mathrm{a}}$ ▷ remove the ranking-opposed component

  else $\tilde{g}_{\mathrm{r}} \leftarrow g_{\mathrm{r}}$

  end if

end if

$\theta \leftarrow \theta - \eta\,(g_{\mathrm{a}} + \tilde{g}_{\mathrm{r}})$

Invariant: for GPROJ, $\langle g_{\mathrm{a}}, \tilde{g}_{\mathrm{r}} \rangle \geq 0$ at every step, hence $I_\pi^{\mathrm{comp}} \geq 0$ to first order by Lemma 2.

Computational Cost. ANNEAL is computationally free: it changes a scalar. GPROJ requires the two span-masked gradients to be materialized separately (one extra backward pass, or one forward with two masked loss heads) plus one dot product and one axpy over the flat gradient, in our runs. Both add *zero* inference cost over SFT-CoT.

K OUTCOME-REWARD REASONING: IMPLEMENTATION

The following proposition records the effect of this substitution on the two fixable terms of the bound.

Proposition 18 (Outcome reward aligns the reasoning gradient)**.** *Let* $g_{\mathrm{r}}^{\mathrm{RL}} = \nabla_\theta\big(-\mathbb{E}_{z\sim\pi_\theta}[R(x, y, z)]\big)$ *denote the reasoning-side update induced by Eq. equation 7. Since* R *is a monotone functional of the same ranking quality that* $\mathcal{L}_{\mathrm{rank}}$ *scores,* $g_{\mathrm{r}}^{\mathrm{RL}}$ *is in expectation a descent direction of* $\mathcal{L}_{\mathrm{rank}}$*, so that*

$$\big\langle g_{\mathrm{a}},\, g_{\mathrm{r}}^{\mathrm{RL}} \big\rangle \;\geq\; 0 \qquad \Longrightarrow \qquad I_\pi^{\mathrm{comp}} \;\geq\; 0 \tag{24}$$

by Lemma 2, in contrast to the strictly negative I_π^{comp} of token-level chain-of-thought (Theorem 12). Moreover, no external distribution $p^*(z \mid x)$ must be reproduced, so that $\delta_{\mathrm{KL}} \to 0$ and the penalty term of Eq. equation 2 vanishes.

This appendix gives the implementation of the outcome-reward procedure of Section 4.2.

K.1 Reward computation

The candidate set $\mathcal{C}_u$, with $|\mathcal{C}_u| = 20$, is presented in shuffled order with single-token labels A through T. Given a sampled trace z, we append the answer marker and read the next-token distribution restricted to the twenty label tokens; sorting these logits yields the induced ranking $\sigma_\theta(x, z)$, from which the reward of Eq. equation 7 is computed against the gold label of y. The reward is therefore exactly the quantity used at evaluation, so that the training signal and the test metric coincide. We emphasize that the label y is used only to *score* σ_θ; the trace z is generated from x alone, which is what preserves the Markov structure invoked in Section 4.2.

K.2 Direct preference optimization

We warm-start π_θ and the reference policy π_{ref} from the best conflict-mitigated supervised checkpoint in each cell. For each prompt we sample $G{=}6$ traces, score them with Eq. equation 7, and retain the best–worst pair whenever their reward margin exceeds 0.1. Completions terminate at the answer marker, so that the preference signal shapes the *reasoning* rather than the answer label itself. The objective is

$$\mathcal{L}_{\mathrm{DPO}}(\theta) \;=\; -\,\mathbb{E}_{(x,z^+,z^-)} \log \sigma\!\left(\beta \log \frac{\pi_\theta(z^+ \mid x)}{\pi_{\mathrm{ref}}(z^+ \mid x)} \;-\; \beta \log \frac{\pi_\theta(z^- \mid x)}{\pi_{\mathrm{ref}}(z^- \mid x)}\right), \tag{25}$$

where π_{ref} is the frozen supervised initialization and β controls the strength of the implicit trust region. We optimize it with $\beta = 0.1$ using AdamW over the retained pairs.

Algorithm 2 Outcome-reward reasoning via direct preference optimization

initialize $\pi_\theta, \pi_{\mathrm{ref}} \leftarrow$ supervised checkpoint
for all prompts x with logged item y **do**
  sample $\{z_i\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid x)$
  $R_i \leftarrow \mathrm{NDCG@}K(\sigma_\theta(x, z_i),\, y)$ ▷ Eq. equation 7
  $z^+ \leftarrow z_{\arg\max_i R_i}, \quad z^- \leftarrow z_{\arg\min_i R_i}$
  if $\max_i R_i - \min_i R_i > 0.1$ **then**
    retain the pair (x, z^+, z^-)
  end if
end for
$\theta \leftarrow \arg\min_\theta \mathcal{L}_{\mathrm{DPO}}(\theta)$ over the retained pairs ▷ Eq. equation 25

Hyperparameters. $G = 6$ samples per prompt; sampling temperature 1.0 with top-p 0.9; maximum trace length 512 tokens; preference margin threshold 0.1; $\beta = 0.1$; AdamW. Remaining optimization settings follow Table 7. The initialization for each cell is the CARE rule with the higher *validation* NDCG@10 in that cell (GPROJ at 3B and ANNEAL at 7B/8B in the majority of cells, consistent with the crossover of Section 5); no test-split information enters this choice.

A note on the choice of estimator. An online estimator such as GRPO (Shao et al., 2024) optimizes the same reward and would be a natural alternative. We report the offline preference formulation because it requires a single round of sampling and was sufficient to establish the effect of interest; we make no claim that it is the stronger of the two. The measured comparison in fact favours GRPO slightly on the conflicted initialization (e.g. .412 against .399 NDCG@10 on MovieLens-1M at 7B, Table 1), and we report it rather than suppress it: the margin is small, it does not alter the conclusion that outcome reward fails on a conflicted checkpoint, since both estimators remain below the NO-COT level of .494 there, and the choice of estimator is orthogonal to the claim being tested.

Part D: Protocol and Reproducibility

L Implementation Details

This appendix collects the dataset statistics, splits, serving stack, decoding parameters, and fine-tuning hyperparameters used throughout the paper.

L.1 Datasets and statistics

We prepare four sequential-recommendation benchmarks with resolvable item text (usable for LLM reasoning): three Amazon review categories (Beauty, Toys, Sports) and MovieLens-1M. All four appear in the main tables. These span an order of magnitude in catalog size (3.7K–18K items) and a wide range of average sequence lengths ($\sim$8–166 interactions per user). Item text is rendered from each dataset's metadata (title, brand, category, price for Amazon; title/genre for MovieLens). Table 6 reports the full statistics.

Table 6: Dataset statistics. The four datasets above the rule have resolvable item text and are used throughout.

Dataset	#Users	#Items	#Actions	Avg. Len.	Sparsity	Item Text	Task Type
Amazon Beauty	22,363	12,101	198,502	8.88	99.93%	✓	Next-item (implicit)
Amazon Toys	19,412	11,924	167,597	8.63	99.93%	✓	Next-item (implicit)
Amazon Sports	35,598	18,357	296,337	8.32	99.95%	✓	Next-item (implicit)
MovieLens-1M	6,040	3,706	1,000,209	165.60	95.53%	✓	Next-item (explicit)

L.2 Splits and candidate construction

We adopt a strict leave-last-two-out temporal protocol. For each user sequence $S = (i_1, \ldots, i_n)$ we form three non-overlapping targets at distinct positions: train predicts i_{n-2} from $(i_1, \ldots, i_{n-3})$; validation predicts i_{n-1} from $(i_1, \ldots, i_{n-2})$; and test predicts i_n from $(i_1, \ldots, i_{n-1})$. Because the three targets occupy different positions, validation and test items are never used as training labels, eliminating leakage; all corpus statistics used for candidate construction are fit only on the training prefixes. Sequences with fewer than three interactions are discarded. Only the most recent $L{=}10$ interactions are shown to the model as history.

For every case we build an *oracle-complete* shortlist of $K{=}20$ candidates: the top-$(K{-}1)$ items from a first-order Markov-chain retriever (selected in Appendix M) plus the ground-truth target (inserted if not already retrieved), padded with popularity-sampled negatives when the retriever returns fewer than $K{-}1$ items. The K positions are then shuffled with a fixed seed so the target sits at an unpredictable slot. Unless otherwise noted, zero-shot evaluation uses a fixed subsample of $n{=}500$ deterministic test cases per dataset (seed 0); the same cases are used for all methods. The two ranking-extraction mechanisms (listwise text generation for zero-shot; letter-logit ARGSORT for trained models) are described in Appendix D.

L.3 Serving stack (zero-shot inference)

All zero-shot models are served through a unified wrapper around vLLM 0.8.5.post1 (V1 engine). Generation is fully batched: all 500 cases per (dataset, split) are dispatched in a single `llm.chat()` call, using vLLM's continuous batching and paged-attention scheduler. Runs use a single 80 GB A100 GPU (`tensor_parallel_size=1`); models are loaded sequentially. Shared engine configuration: FlashAttention backend; prefix caching enabled (shared system prompt cached across the batch); chunked prefill enabled (`max_num_batched_tokens` = 16384); CUDA graphs enabled (compilation level 3, capture sizes up to 512); KV-cache dtype `auto`; GPU memory utilization 0.85; seed 0 (engine and sampler).

Output parsing (zero-shot). The last `Final Ranking:` line is extracted by regex, tokenized into letter labels, and mapped to item IDs. Invalid, repeated, or omitted labels are dropped (this can only lower the target's rank, never fabricate one); if no valid ranking line is produced, the case contributes 0 to all metrics.

L.4 FINE-TUNING SETUP

For the supervised experiments we fine-tune Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct under direct (No-CoT) and CoT supervision. Reasoning traces are distilled from a frozen Qwen2.5-32B-Instruct teacher (the two teacher-CoT regimes are detailed in Appendix C); traces are produced only for the training split, and the training set is capped at $8{,}000$ distilled cases per (dataset, backbone) pair while the full validation and test sets are always evaluated. Table 7 lists the hyperparameters.

Table 7: Supervised fine-tuning hyperparameters (shared unless noted).

Setting	Value
Backbones	Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	0.01
Epochs (No-CoT, CoT)	3
Teacher (distillation)	Qwen2.5-32B-Instruct

Metrics. We report NDCG@k, Recall@k, and HR@k for $k \in \{1, 5, 10\}$. Because each case has exactly one relevant (ground-truth) item, Recall@k and HR@k coincide. Metrics are averaged over the test set of each (dataset, backbone) pair.

M RETRIEVAL ABLATION: CHOOSING THE STAGE-1 RETRIEVER

The first stage of the retrieve-then-rerank pipeline (Section 2) must shrink a 10^4–10^6-item catalog to a handful of candidates the reasoner can operate over. Because our study targets the reasoner (Stage 2), we require a Stage 1 that is *training-free*: no learned parameters, no task-specific adaptation, and no entanglement with the model under evaluation. This keeps the candidate generator fully auditable and ensures any change in reranking quality is attributable to the reasoner. This appendix reports the ablation used to pick that retriever.

M.1 RETRIEVERS

We benchmark five classical training-free retrievers:

- **MostPop** (Cremonesi et al., 2010): globally most popular items, non-personalized (the floor any context-aware method must beat).
- **ItemKNN** (Sarwar et al., 2001; Deshpande & Karypis, 2004): popularity-normalized (cosine) co-occurrence similarity to the user's history (window $= 5$, max context $= 50$).
- **SKNN** (Jannach & Ludewig, 2017; Ludewig et al., 2021): session-based KNN with Jaccard similarity (k=100, up to 1000 candidate sessions).
- **Markov Chain** (Rendle et al., 2010): first-order item-to-item transition probabilities, depending only on the most recent item.
- **Association Rules** (Agrawal et al., 1993; Mobasher et al., 2001): directional confidence scores over co-occurrence patterns (window $= 5$, max context $= 50$).

Evaluation is *full-catalog*: every positively scored item competes for rank, and a target that is never surfaced (score ≤ 0) is unranked and contributes 0 to all metrics, so coverage failures are counted rather than masked by a popularity backoff. We report NDCG@k and Recall@k for $k \in \{1, 5, 10\}$ on the test split.

Across all four datasets (Table 8), the first-order Markov chain is the strongest training-free retriever by a wide margin, four to seven times the Recall@1 of item-KNN, association rules, session-KNN, and popularity (e.g., Beauty test Recall@1 0.039 vs. 0.009 for item-KNN; ML-1M test Recall@10 0.208 vs. 0.046). The lesson is that, for next-item prediction, the single most recent interaction carries most of the computationally cheap signal: methods that blend a wide window (item-KNN, association rules) or ignore order (session-KNN) dilute it. We therefore adopt the first-order Markov chain as the Stage-1 retriever used to build the oracle-complete shortlists throughout the paper; it

Table 8: Training-free retriever ablation (full-catalog, test split). The first-order Markov chain dominates every dataset by a wide margin (more than fourfold the Recall@1 of the next-best method), so we adopt it as the Stage-1 retriever throughout. **Bold**: best per column within a dataset.

Dataset	Retriever	N@1	N@5	N@10	R@1	R@5	R@10
Beauty	MostPop	0.001	0.004	0.005	0.001	0.007	0.011
	ItemKNN	0.009	0.017	0.024	0.009	0.026	0.045
	SKNN	0.000	0.003	0.012	0.000	0.008	0.036
	AssociationRules	0.007	0.016	0.023	0.007	0.026	0.048
	MarkovChain	**0.039**	**0.055**	**0.058**	**0.039**	**0.069**	**0.078**
Toys	MostPop	0.002	0.004	0.005	0.002	0.006	0.009
	ItemKNN	0.011	0.023	0.030	0.011	0.034	0.058
	SKNN	0.000	0.005	0.015	0.000	0.012	0.044
	AssociationRules	0.009	0.022	0.029	0.009	0.035	0.058
	MarkovChain	**0.052**	**0.065**	**0.067**	**0.052**	**0.076**	**0.080**
Sports	MostPop	0.002	0.004	0.005	0.002	0.006	0.009
	ItemKNN	0.006	0.010	0.013	0.006	0.015	0.024
	SKNN	0.000	0.002	0.008	0.000	0.005	0.023
	AssociationRules	0.005	0.011	0.015	0.005	0.018	0.030
	MarkovChain	**0.026**	**0.034**	**0.036**	**0.026**	**0.042**	**0.047**
ML-1M	MostPop	0.003	0.005	0.008	0.003	0.008	0.016
	ItemKNN	0.007	0.017	0.023	0.007	0.027	0.046
	SKNN	0.000	0.002	0.006	0.000	0.005	0.015
	AssociationRules	0.005	0.012	0.020	0.005	0.020	0.045
	MarkovChain	**0.047**	**0.098**	**0.118**	**0.047**	**0.146**	**0.208**

also sets a strong, essentially computationally free baseline that any LLM-based reranker must clear to justify its computational cost.

Part E: Additional Results and Measurements

N Near-Orthogonality of Reasoning and Ranking Gradients at Scale

Section 2.1 showed the conflict is forced by the shared readout once the heads specialize. This appendix establishes the complementary fact: *at initialization*, and for any sufficiently wide backbone, the two gradients are near-orthogonal with high probability. Together the two results explain the observed dynamics, a cosine that starts near zero (or slightly positive), turns negative as specialization proceeds, and is milder at larger d, and they show that scale attenuates the penalty without removing it.

Section 2.3 summarizes this result in one paragraph; we give the argument in full here. Consider a Transformer with L layers, hidden dimension d, and $p = O(Ld^2)$ parameters. Write the full gradient of each objective as a concatenation of per-layer gradients,

$$g_{\mathrm{r}} = \big(g_{\mathrm{reason}}^{(1)\top}, \ldots, g_{\mathrm{reason}}^{(L)\top}\big)^{\top}, \qquad g_{\mathrm{a}} = \big(g_{\mathrm{rank}}^{(1)\top}, \ldots, g_{\mathrm{rank}}^{(L)\top}\big)^{\top},$$

where $g^{(\ell)} \in \mathbb{R}^{d^2}$ is the flattened gradient with respect to layer ℓ's weight matrix. For a single layer the gradient is an outer product $\nabla_{W^{(\ell)}}\mathcal{L} = \delta^{(\ell)}(h^{(\ell-1)})^{\top}$, i.e. $g^{(\ell)} = h^{(\ell-1)} \otimes \delta^{(\ell)}$, with $h^{(\ell-1)}$ the shared input activation and $\delta^{(\ell)}$ the back-propagated error signal.

Assumption 19 (Effective independence at initialization)**.** *At initialization θ_0, for each layer ℓ: (i) the error signals $\delta_{\mathrm{reason}}^{(\ell)}$ and $\delta_{\mathrm{rank}}^{(\ell)}$ lie in subspaces of dimension at most $r \ll d$; and (ii) conditioned on the shared activation $h^{(\ell-1)}$, their subspace bases are independent, Haar-uniformly oriented rotations in $\mathbb{R}^d$.*

Justification and scope. Before task-specific training, the reasoning span and the ranking span occupy disjoint token positions; their back-propagated errors traverse independent paths through randomly initialized attention weights, so the subspaces are effectively independent at epoch 0. The assumption weakens as training specializes the ranking head (epoch 1 onward), which is exactly where the empirical cosine departs from the bound, and exactly the regime Theorem 12 governs. The two results are therefore complementary rather than competing: Lemma 20 describes the initialization, and the readout identity describes what happens after.

N.1 Statement

Lemma 20 (Gradient near-orthogonality in overparameterized backbones)**.** *Under Assumption 19, for any $\epsilon > 0$,*

$$\big|\cos(g_{\mathrm{r}}, g_{\mathrm{a}})\big| \;\le\; \frac{r\sqrt{L}}{\sqrt{d}} + \epsilon$$

with probability at least $1 - 2L\exp(-c\,d\,\epsilon^2/L)$ for an absolute constant $c > 0$. In particular, for fixed r, L the cosine concentrates at zero as $d \to \infty$ at rate $O(1/\sqrt{d})$.

Proof. **Step 1 (per-layer inner product).** By the Kronecker structure,

$$\langle g^{(\ell)}_{\mathrm{reason}}, g^{(\ell)}_{\mathrm{rank}}\rangle = \big\langle h^{(\ell-1)} \otimes \delta^{(\ell)}_{\mathrm{reason}},\; h^{(\ell-1)} \otimes \delta^{(\ell)}_{\mathrm{rank}}\big\rangle = \|h^{(\ell-1)}\|^2\,\langle \delta^{(\ell)}_{\mathrm{reason}}, \delta^{(\ell)}_{\mathrm{rank}}\rangle.$$

Step 2 (random-subspace bound). Let $U_{\mathrm{reason}}, U_{\mathrm{rank}} \in \mathbb{R}^{d\times r}$ be orthonormal bases of the two error subspaces and write $\delta^{(\ell)}_{\mathrm{reason}} = U_{\mathrm{reason}}a$, $\delta^{(\ell)}_{\mathrm{rank}} = U_{\mathrm{rank}}b$. Then $\langle \delta^{(\ell)}_{\mathrm{reason}}, \delta^{(\ell)}_{\mathrm{rank}}\rangle = a^\top U_{\mathrm{reason}}^\top U_{\mathrm{rank}} b$. Under Assumption 19, $U_{\mathrm{reason}}^\top U_{\mathrm{rank}} \in \mathbb{R}^{r\times r}$ is the cross-Gram matrix of two independently, uniformly oriented r-dimensional subspaces, whose entries have mean 0 and variance $1/d$. By standard concentration for random subspace projections (Vershynin, 2018), $\|U_{\mathrm{reason}}^\top U_{\mathrm{rank}}\|_F \le r/\sqrt{d} + t$ with probability at least $1 - 2\exp(-ct^2 d)$. Cauchy–Schwarz on the bilinear form gives $|\langle \delta^{(\ell)}_{\mathrm{reason}}, \delta^{(\ell)}_{\mathrm{rank}}\rangle| \le \|\delta^{(\ell)}_{\mathrm{reason}}\|\,\|\delta^{(\ell)}_{\mathrm{rank}}\|(r/\sqrt{d} + t)$.

Step 3 (aggregate across layers). Substituting into the global cosine and using the triangle inequality,

$$|\cos(g_{\mathrm{r}}, g_{\mathrm{a}})| \le \frac{\sum_\ell \|h^{(\ell-1)}\|^2\,\|\delta^{(\ell)}_{\mathrm{reason}}\|\,\|\delta^{(\ell)}_{\mathrm{rank}}\|}{\|g_{\mathrm{r}}\|\,\|g_{\mathrm{a}}\|}\Big(\tfrac{r}{\sqrt{d}} + t\Big).$$

Since $\|g^{(\ell)}\| = \|h^{(\ell-1)}\|\,\|\delta^{(\ell)}\|$, the numerator equals $\sum_\ell \|g^{(\ell)}_{\mathrm{reason}}\|\,\|g^{(\ell)}_{\mathrm{rank}}\|$, and by Cauchy–Schwarz across layers this is $\le \|g_{\mathrm{r}}\|\,\|g_{\mathrm{a}}\|$ (prefactor ≤ 1). Under approximately balanced per-layer norms at initialization (RMSNorm), $\|g^{(\ell)}\| \approx \|g\|/\sqrt{L}$, so the prefactor is bounded by $\sqrt{L}$ in the worst case, giving $|\cos(g_{\mathrm{r}}, g_{\mathrm{a}})| \le \sqrt{L}\,(r/\sqrt{d} + t)$.

Step 4 (union bound). Set $t = \epsilon/\sqrt{L}$ and union-bound over the L layers:

$$|\cos(g_{\mathrm{r}}, g_{\mathrm{a}})| \le \sqrt{L}\Big(\tfrac{r}{\sqrt{d}} + \tfrac{\epsilon}{\sqrt{L}}\Big) = \frac{r\sqrt{L}}{\sqrt{d}} + \epsilon$$

with probability at least $1 - 2L\exp(-c\,d\,\epsilon^2/L)$. □

N.2 Numerical check

Model	d	L	Predicted $\lvert\cos\rvert \le$	Measured $\lvert\cos\rvert$ (ep. 0)
Qwen2.5-3B	2048	36	$r\sqrt{36}/\sqrt{2048} \approx 0.13\,r$	0.15
Qwen2.5-7B	3584	28	$r\sqrt{28}/\sqrt{3584} \approx 0.09\,r$	0.01

For 3B, $r \approx$ 1–2 matches $|\cos| = 0.15$; for 7B, $r < 1$ (the subspaces barely overlap), consistent with $|\cos| \approx 0$. The larger hidden dimension of the 7B backbone gives its reasoning and ranking gradients substantially more room to occupy orthogonal subspaces. This is a *consistency* check of the bound, not an independent test, since r is not measured directly.

Table 9: Gradient geometry under BLIND teacher CoT, both backbones. cos is $\cos(g_r, g_a)$; $f_N = \|g_r\|/(\|g_r\| + \|g_a\|)$ (%) is the reasoning gradient's share of the total norm. Every cell starts *positive* at epoch 0 and swings negative once mass concentrates, on every dataset and at both scales: the sign and the timing predicted by Proposition 11. Estimated over a probe of 1124–2048 examples per cell; magnitudes carry a sampling error of roughly ± 0.2 and the ordering of datasets by $|\cos|$ should not be read (see text). The HINDSIGHT counterpart is Table 10.

		Epoch 0		Epoch 1		Epoch 2		Epoch 3	
Dataset	Model	cos	f_N	cos	f_N	cos	f_N	cos	f_N
Amazon Beauty	3B	+0.341	2.9	−0.620	0.9	−0.376	0.9	−0.480	1.0
	7B	+0.216	2.4	−0.397	1.7	−0.219	1.5	−0.288	1.4
Amazon Sports	3B	+0.366	2.6	−0.408	0.8	−0.343	1.0	−0.145	1.6
	7B	+0.246	2.3	−0.273	1.3	−0.305	1.4	−0.246	1.4
Amazon Toys	3B	+0.351	2.4	−0.705	0.5	−0.548	0.4	−0.456	0.7
	7B	+0.203	2.1	−0.336	1.7	−0.217	1.2	−0.305	1.3
MovieLens-1M	3B	+0.399	2.4	−0.477	0.8	−0.317	1.1	−0.274	1.2
	7B	+0.244	2.2	−0.174	1.4	−0.219	1.4	−0.230	1.6
Goodreads	3B	+0.381	2.5	−0.314	0.7	−0.355	1.0	−0.310	1.0
	7B	+0.187	2.0	−0.322	1.1	−0.297	1.3	−0.323	1.4

Table 10: Gradient geometry under HINDSIGHT teacher CoT, both backbones; notation as in Table 9. Because the rationale agrees with the log by construction, the cosine stays close to zero throughout, and the reasoning gradient carries roughly ten times the norm share it does under BLIND. Informative reasoning is therefore not what conflicts; disagreeing reasoning is. Probe sizes and precision as in Table 9.

		Epoch 0		Epoch 1		Epoch 2		Epoch 3	
Dataset	Model	cos	f_N	cos	f_N	cos	f_N	cos	f_N
Amazon Beauty	3B	+0.347	7.6	−0.231	11.6	−0.254	8.5	−0.169	12.0
	7B	+0.118	6.5	−0.021	17.0	−0.002	13.8	−0.053	17.5
Amazon Sports	3B	+0.360	7.4	−0.091	12.2	−0.083	10.0	−0.060	11.5
	7B	+0.132	6.3	+0.011	18.6	−0.051	16.3	−0.003	15.5
Amazon Toys	3B	+0.377	6.9	−0.303	15.0	−0.244	20.3	−0.146	11.2
	7B	+0.126	6.3	−0.174	17.6	−0.153	25.2	−0.002	13.2
MovieLens-1M	3B	+0.363	5.8	−0.016	11.1	−0.027	7.3	−0.112	11.5
	7B	+0.110	5.2	−0.025	15.3	−0.027	16.7	−0.025	21.0
Goodreads	3B	+0.400	6.7	−0.298	17.3	−0.170	14.0	−0.086	14.2
	7B	+0.113	5.2	−0.078	16.1	−0.073	21.4	−0.003	18.4

Scale attenuates but does not remove the conflict. Modelling each layer's gradient as an outer product and assuming the two error signals occupy independent, Haar-oriented rank-r subspaces at initialization, concentration gives $|\cos(g_r, g_a)| \le r\sqrt{L}/\sqrt{d} + \epsilon$ w.h.p. (Lemma 20, Appendix N), so the cosine concentrates at zero as $O(1/\sqrt{d})$. This predicts, and we observe, a milder conflict at 7B than at 3B. But the bound holds *at initialization*, before the ranking head specializes, whereas Theorem 12 governs the trained regime, and the chain-of-thought penalty persists at 7B. Increasing scale therefore mitigates the problem without resolving it.

N.3 7B GRADIENT GEOMETRY

Scale attenuates the conflict but does not remove it. The conflict is markedly milder on the 7B backbone than on the 3B one: the worst-case $\cos(g_r, g_a)$ dips only to -0.17 to -0.40 (vs. -0.31 to -0.71 at 3B; Table 9). It does *not*, however, return to zero: at epoch 3 the 7B cosine still sits between -0.23 and -0.32 on every dataset. Lemma 20 explains why: a wider trunk has room to serve both objectives in near-orthogonal subspaces. Yet the CoT penalty persists at 7B in every dataset (Table 15) and is in fact *largest* there on MovieLens-1M, because the stronger NO-COT baseline leaves more to lose, consistent with Theorem 12, whose condition (A3) bites harder for more specialized heads. Attenuation of the conflict is therefore not equivalent to its removal.

Table 11: $I_\pi^{\mathrm{comp}} < 0$, measured. Reasoning and ranking gradients oppose each other in every cell, and the cells with more sustained opposition tend to pay the larger penalty, as Lemma 2 predicts. Cosines are measured under BLIND supervision; the penalty is SFT-CoT minus SFT-NO-COT. We give the values rather than a correlation: one observation per cell is too small a sample, and Eq. equation 3 weights the cosine by gradient norms that differ across scale, so cells at different scales are not exchangeable. Per-epoch traces and the 7B analogue are in Appendix N.

Backbone	Dataset	$\cos(g_{\mathrm{r}}, g_{\mathrm{a}})$ ep. 3	mean cos ep. 1–3	ΔN@1	ΔN@10
Qwen2.5-3B	Amazon Beauty	-0.48	-0.49	-0.026	$+0.005$
	Amazon Sports	-0.15	-0.30	-0.013	$+0.016$
	Amazon Toys	-0.46	-0.57	-0.002	$+0.010$
	MovieLens-1M	-0.27	-0.36	-0.053	-0.062
Qwen2.5-7B	Amazon Beauty	-0.29	-0.30	-0.037	-0.045
	Amazon Sports	-0.25	-0.27	-0.022	-0.021
	Amazon Toys	-0.30	-0.29	-0.018	-0.034
	MovieLens-1M	-0.23	-0.21	-0.202	-0.227

O GRADIENT-CONFLICT PROBE: PROTOCOL AND FULL MEASUREMENTS

This appendix specifies the gradient-conflict probe used throughout Section 3, its relationship to influence functions, and the full per-epoch measurements.

O.1 PROBE PROTOCOL

We measure gradient geometry *post hoc* along the SFT trajectory; no model is trained or modified for this analysis. For each (dataset, backbone) cell we take the CoT checkpoints at epochs $0, 1, 2, 3$ and, over a fixed probe set, accumulate token-weighted population gradients g_{r} and g_{a} by masking the labels of the complementary span (Algorithm 3). Masking with -100 (the `ignore_index`) rather than slicing the sequence is essential: it keeps the forward pass, attention pattern, and activations byte-identical between the two accumulations, so the only difference between g_{r} and g_{a} is which positions contribute to the loss. Both are normalized by their token count so the reported ratio $\rho = \|g_{\mathrm{r}}\|/\|g_{\mathrm{a}}\|$ is per-token and comparable across examples of different trace length.

Algorithm 3 Gradient conflict probe

Require: checkpoint θ_e; probe set $\{(x, r, a)\}$ of prompts, reasoning spans, and answer spans
1: **for** span $\in \{\mathrm{reason}, \mathrm{rank}\}$ **do**
2: mask labels outside the span with -100
3: $g_{\mathrm{span}} \leftarrow \sum_n \nabla_\theta \ell^{(n)}$, accumulated in fp32
4: normalize g_{span} by the number of contributing tokens
5: **end for**
6: **return** $\cos(g_{\mathrm{r}}, g_{\mathrm{a}})$, $\|g_{\mathrm{r}}\|$, $\|g_{\mathrm{a}}\|$, $\rho = \|g_{\mathrm{r}}\|/\|g_{\mathrm{a}}\|$

Per-layer variant. For Fig. 3 we compute the same quantity restricted to each transformer layer's parameters, $\rho(\ell) = \cos(g_{\mathrm{r}}^{(\ell)}, g_{\mathrm{a}}^{(\ell)})$, at the final full-loss checkpoint, and average over datasets per backbone.

Probe set. A fixed subsample of 1,124–2,048 training cases per cell (seed 0), identical across epochs, regimes and methods, so the epoch-to-epoch change in the cosine cannot be a sampling artefact. Where two logged runs differ in probe size they agree on every qualitative claim and disagree on magnitude by up to 0.19, which is why we read the cosine for its sign, timing and regime-selectivity only (Appendix N).

O.2 FULL MEASUREMENTS

Tables 9 and 10 report the BLIND and HINDSIGHT geometry at both scales; the 7B analogue; both are in Appendix N.

O.3 MEASURED RATIONALE–LOG AGREEMENT

Proposition 11 attributes the conflict to a small $\Pr(y{=}m)$, and that quantity is directly observable: the BLIND teacher ends its trace with its own ranking, so the rationale-implied item m is simply its top choice, and $\Pr(y{=}m)$ is its top-1 accuracy on the training split. It is 0.071–0.128 across the four datasets (Table 12, Appendix O.3): 1.4 to 2.6 times chance, so the rationale is genuinely informative, yet far from the ≈ 1 that verifiable tasks enjoy. The disagreement between rationale and log is therefore not incompetence but multiplicity of valid continuations.

Resampling the teacher a second time sharpens the picture and corrects the exchangeable model. Its self-consistency is $\Pr(m{=}m') = \sum_k \tilde{q}_k^2 \in [0.44, 0.62]$, four to six times its accuracy, so the teacher's implied distribution $\tilde{q}$ is concentrated on roughly two candidates and is emphatically *not* the true continuation distribution: the equality $\Pr(y{=}m) = \sum_k q_k^2$ of the exchangeable special case is rejected. Two consequences follow, and both favour the mechanism. The reasoning gradient is a *stable* commitment to one wrong vertex rather than a diffuse nudge, so it accumulates across steps in Eq. equation 3 instead of partially cancelling. And the diffuseness of the true posterior can now be read off without any proxy model: with $\tilde{q}$ concentrated on about two candidates, $\langle q, \tilde{q} \rangle \approx 0.11$ implies that the true distribution places only about 22% of its mass on the candidates a 32B reranker most favours.

Table 15 (Appendix R) closes the loop: neither regime clears NO-COT on N@1 in *any* of the eight cells, and the gap widens at 7B where the direct backbone is strongest. They fail for opposite reasons. BLIND fails on the *benefit* side, since $\mathcal{I}_{\text{comp}} < 0$ by conflict leaves nothing to gain. HINDSIGHT fails on the *penalty* side: its trace is informative and well behaved as an optimization target but depends on information the student never has at test time, so $p_\theta(z \mid x)$ cannot match $p^*_{\text{HS}}(z \mid x)$, δ_{KL} is large, $B_y\sqrt{\delta_{\text{KL}}/2}$ dominates, and it is in fact the worst of the three. Textual chain-of-thought under imitation is therefore squeezed from both sides: honest reasoning is uninformative *and* conflicting, while informative reasoning is out-of-distribution *and* unreproducible. The direct-readout compute term, the change in the gold answer's negative log-likelihood under the computationally free forward pass with no reasoning generated, is negative in every cell, so both regimes measurably degrade the weights' ability to answer directly.

Table 12 reports the two teacher-level quantities of Proposition 11. Both are properties of the frozen Qwen2.5-32B teacher and are therefore identical across student backbones. $\Pr(y{=}m)$ is the BLIND teacher's top-1 accuracy on the training split, i.e. how often the item its rationale argues for is the item that was actually logged; $\Pr(m{=}m')$ is its self-consistency across two independent samples at the same decoding configuration, which estimates $\sum_k \tilde{q}_k^2$, the concentration of the teacher's own implied distribution.

Table 12: Teacher-level agreement, computed over the full training split ($n = 7{,}999$ for the Amazon datasets, $n = 6{,}040$ for MovieLens-1M; parse-failure rate 0.01%). Chance is $1/K = 0.05$; Wilson 95% intervals. $\Pr(m{=}m') \gg \Pr(y{=}m)$ in every dataset: the teacher is self-consistent but confidently misdirected, so its implied distribution $\tilde{q}$ is *not* the true continuation distribution q. We therefore report the overlap $\langle q, \tilde{q} \rangle$ that Proposition 11 actually requires, rather than the collision probability $\sum_k q_k^2$ of the exchangeable special case.

Dataset	$\Pr(y{=}m) = \langle q, \tilde{q} \rangle$	$\Pr(m{=}m') = \sum_k \tilde{q}_k^2$	ratio
Amazon Beauty	0.120 $[0.113, 0.127]$	0.624 $[0.594, 0.653]$	$5.2\times$
Amazon Sports	0.105 $[0.098, 0.112]$	0.596 $[0.565, 0.626]$	$5.7\times$
Amazon Toys	0.128 $[0.121, 0.136]$	0.611 $[0.580, 0.640]$	$4.8\times$
MovieLens-1M	0.071 $[0.065, 0.078]$	0.444 $[0.413, 0.475]$	$6.3\times$

P REASONING-STATE ANSWER DECODABILITY

P.1 THE REMEDY OPERATES THROUGH THE PREDICTED MECHANISM

Summary. Under standard chain-of-thought training, the reasoning tokens encode the model's own answer at close to chance level. Every method that resolves the late-layer reasoning–ranking conflict makes that answer several times more readable from the same hidden states. This provides evidence that the conflict is what prevents the reasoning from carrying the decision, and that the remedies of Section 4 operate through the predicted mechanism rather than through unrelated regularization.

P.2 MOTIVATION

Figure 3 shows that the reasoning-token gradient and the ranking (answer) gradient conflict throughout SFT and across all layers, with the conflict *deepening toward the final layers*. A natural question is whether *resolving* that conflict changes what the reasoning tokens actually represent. We test this directly: if a method makes the reasoning useful for ranking, then the model's answer should become linearly readable from the hidden states at the reasoning-token positions. We therefore probe, layer by layer, how much of the eventual answer is already encoded in the reasoning states produced by each training method.

P.3 PROBE SETUP

All probes are computed *post hoc* on the frozen BLIND checkpoints no model is trained or fine-tuned for this analysis. For each (dataset, backbone) cell we build a fixed probe set of $N{=}1000$ examples from the full-CoT training file (the held-out test/validation splits store no teacher reasoning steps, which are generated only at inference). Each example is teacher-forced as the concatenation [question] [reasoning steps] ["### ⟨letter⟩"], identical across all methods, and sequences exceeding 2048 tokens are dropped.

For a trained model and each transformer layer ℓ, we extract the hidden state at the *final reasoning token* (the pre-answer summary position that feeds letter prediction), yielding a feature vector $x_i^{(\ell)} \in \mathbb{R}^d$ per example i. We fit a frozen linear probe $g_\ell(x) = W_\ell x + b_\ell$ mapping $x^{(\ell)}$ to one of the $C{=}20$ candidate letters. Features are standardized and projected onto their top $K{=}128$ training principal components before the probe (the components are fit on the training split only, so there is no leakage); this makes the probe well-posed in the $d \gg N$ regime. The probe is trained on a random 70% split and we report accuracy on the held-out 30%; chance is $1/C = 0.05$.

Decoding target: the model's own answer. We decode the letter the model *itself* commits to (the arg-max over the candidate-letter logits at the answer position), not the gold letter. The gold letter is essentially undecodable everywhere on this task because the models are only weakly accurate (top-1 accuracy ≈ 0.11–0.21 for full-loss CoT): even at the exact position where the model emits its answer, a probe for the *gold* letter performs at chance (0.07 vs. 0.05 on Amazon Beauty / 3B). Decoding the model's committed answer removes this correctness ceiling and isolates the quantity of interest: *how strongly the reasoning states drive the model's decision*, independent of whether that decision is correct.

Controls. Two controls guard against a broken or trivially-strong probe. (i) A *positive control* that probes a synthetic one-hot-plus-noise feature achieves 1.00 held-out accuracy in every cell, confirming the probe and optimizer are sound. (ii) An *answer-position upper bound* that probes the hidden state feeding letter prediction recovers the model's own answer at 0.67–0.76 (well above chance), confirming the signal exists where it should; the same feature recovers the gold letter only at ≈chance, quantifying the correctness ceiling above. Coconut is layer-agnostic: we probe its continuous thought z_k (mean-pooled over latent slots) as a single reference value.

P.4 FORMALIZATION

Let M_θ be a trained model with L transformer layers over a candidate set $\mathcal{C}$, $|\mathcal{C}| = C = 20$. For a probe example i, let r_i be the index of the final reasoning token and p_i the position whose next-token distribution is the answer letter, and let $h^{(\ell)}(x_i) \in \mathbb{R}^d$ denote the residual-stream hidden state at layer $\ell \in \{0, \dots, L\}$ and position r_i. Writing $z_{p_i} \in \mathbb{R}^{|V|}$ for the model logits at p_i, the model's *committed answer* is

$$\hat{y}_i \;=\; \arg\max_{c \in \mathcal{C}} \; z_{p_i,\,c}, \tag{26}$$

which is the target we decode (in contrast to the gold letter $y_i^\star$).

We standardize features and project them onto the top-K principal components $P_K^{(\ell)} \in \mathbb{R}^{d \times K}$ fit on the *training split only*, $\tilde{h}^{(\ell)}(x_i) = P_K^{(\ell)\top}\big(h^{(\ell)}(x_i) - \mu^{(\ell)}\big)/\sigma^{(\ell)}$, and fit a per-layer linear probe

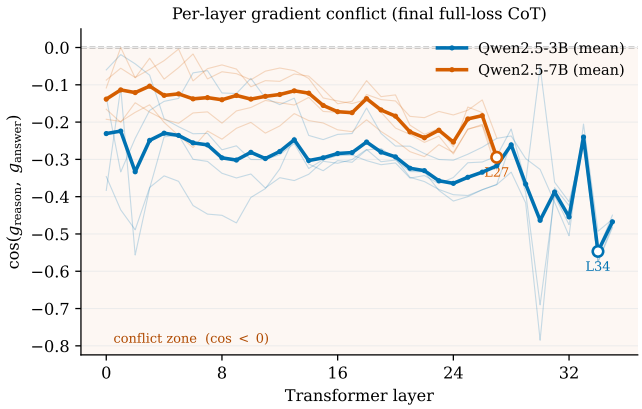


Figure 3: Per-layer directional conflict $\cos(g_{\text{reason}}, g_{\text{answer}})$ at the final full-loss (CoT) checkpoint, averaged over the four datasets per backbone (thin lines: individual datasets). The conflict is negative at every layer but deepens toward the final layers (minima marked), coinciding with the late-layer rise in answer decodability (Figure 4).

$\phi_\ell = (W_\ell, b_\ell)$ by regularized cross-entropy on the training split $\mathcal{D}_{\text{tr}}$:

$$\phi_\ell \;=\; \arg\min_{W,b} \sum_{i\in\mathcal{D}_{\text{tr}}} \text{CE}\Big(\text{softmax}\big(W\,\tilde{h}^{(\ell)}(x_i)+b\big),\, \hat{y}_i\Big) \;+\; \lambda\|W\|_F^2. \tag{27}$$

For a training method $m \in \{\text{full-loss}, \text{gproj}, \text{anneal}\}$ we report the held-out *decodability* at layer ℓ,

$$D_m(\ell) \;=\; \frac{1}{|\mathcal{D}_{\text{te}}|} \sum_{i\in\mathcal{D}_{\text{te}}} \mathbf{1}\Big[\arg\max_{c\in\mathcal{C}} \big(W_\ell\,\tilde{h}^{(\ell)}(x_i)+b_\ell\big)_c \;=\; \hat{y}_i\Big], \qquad \mathbb{E}[D]_{\text{chance}} = \tfrac{1}{C}, \tag{28}$$

and summarize each method by its best layer, $D^\star_m = \max_\ell D_m(\ell)$ (Table 13); Coconut is layer-agnostic, so D_{coco} is a single value obtained by replacing $\tilde{h}^{(\ell)}$ with the (standardized, PCA-projected) continuous thought z_k. The curves in Figure 4 are $\ell \mapsto D_m(\ell)$.

What the figure tests. The probe accuracy $D_m(\ell)$ is a lower bound on the linear (mutual) information between the layer-ℓ reasoning state and the model's decision: if the answer is not linearly encoded, $D_m(\ell) \approx 1/C$. The mechanism hypothesis is that resolving the reasoning/ranking conflict raises this quantity in the layers where the conflict is *sharpest*. The per-layer conflict $\rho(\ell) = \cos\big(g_r^{(\ell)}, g_a^{(\ell)}\big)$ (Figure 3) is negative at *every* layer, the conflict is broad, not confined to a band, but it deepens toward the final layers, reaching its minimum at $\ell^\star = \arg\min_\ell \rho(\ell)$ in the last few layers (e.g. $\ell^\star \approx 34/36$ for 3B and $27/28$ for 7B). Answer decodability mirrors this: it is at chance in the early and middle layers and rises only in the same final layers,

$$\rho(\ell) < 0 \;\; \forall \ell, \quad \ell^\star = \arg\min_\ell \rho(\ell) \text{ is late}; \qquad D_m(\ell) \approx \tfrac{1}{C} \text{ (early)}, \quad D_m(\ell) \gg \tfrac{1}{C} \text{ near } \ell^\star. \tag{29}$$

Within that late region the methods separate by how much they resolve the conflict,

$$D^\star_{\text{full-loss}} \;<\; D^\star_{\text{gproj}} \;<\; D^\star_{\text{anneal}}. \tag{30}$$

As a validity check, the answer-position feature $h(x_i)$ at p_i upper-bounds the achievable signal ($D \approx 0.67$–0.76 for $\hat{y}$), while decoding the *gold* letter $y^\star$ from the same feature stays at chance, so $D^{\text{gold}}(\ell) \approx 1/C$ everywhere, which is why Eq. equation 26 decodes $\hat{y}$ rather than $y^\star$.

P.5 Results

Figure 4 plots per-layer decodability for all eight cells, and Table 13 reports the best-layer value per method. Three findings hold across every cell.

1. **Answer commitment is a late-layer phenomenon.** Decodability sits at chance through the early and middle layers and rises sharply only in the final few layers, the same layers where the gradient conflict is *deepest* (Figure 3), tying the two analyses together.

Table 13: Best-layer decodability of the model's own answer (frozen linear probe, held-out accuracy; chance $= 0.05$). Bold marks the best training method per row.

Dataset	Backbone	full-loss	gproj	anneal	Coconut z_k
Amazon Beauty	Qwen2.5-3B	0.18	0.32	**0.51**	0.22
Amazon Sports	Qwen2.5-3B	0.14	0.25	**0.52**	0.11
Amazon Toys	Qwen2.5-3B	0.16	0.26	**0.58**	0.14
MovieLens-1M	Qwen2.5-3B	0.19	0.29	**0.41**	0.78
Amazon Beauty	Qwen2.5-7B	0.19	0.31	**0.65**	0.06
Amazon Sports	Qwen2.5-7B	0.13	0.23	**0.67**	0.39
Amazon Toys	Qwen2.5-7B	0.14	0.24	**0.67**	0.05
MovieLens-1M	Qwen2.5-7B	0.11	0.50	**0.69**	0.06

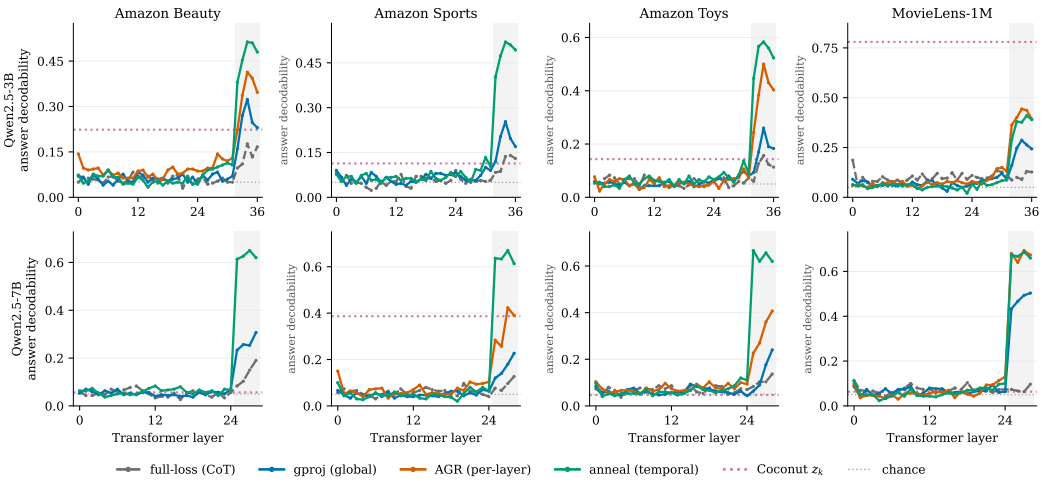


Figure 4: Per-layer decodability of the model's own answer from reasoning-token hidden states (frozen linear probe, held-out accuracy; chance $= 0.05$, dotted). Rows are backbones, columns are datasets; the shaded band marks the late-layer region where answer commitment emerges. In every cell decodability is flat until the final layers, then rises in the order full-loss (CoT) $<$ gproj $<$ anneal. The Coconut z_k reference (dashed) is layer-agnostic.

2. **Conflict resolution makes the reasoning carry the decision.** Full-loss CoT is the *least* answer-predictive in all eight cells (0.11–0.19): its reasoning states barely encode the answer, i.e. the model largely routes around them. Every conflict-resolving method raises decodability, in the consistent order full-loss $<$ gproj $<$ anneal, reaching 0.41–0.69 for ANNEAL, a two- to sixfold increase over full-loss CoT (and 8–$14\times$ chance). The ordering full-loss $<$ gproj $<$ anneal holds strictly in all eight cells.
3. **Decodability is not accuracy.** The probe measures how strongly the reasoning encodes the answer, not whether that answer is correct. Anneal and gproj have near-identical top-1 accuracy on several cells yet very different decodability (e.g. 0.51 vs. 0.32). The effect therefore reflects representational structure rather than a raw accuracy difference.

Coconut's z_k is typically near chance (0.05–0.14), consistent with its cross-example collapse, though it is occasionally high (e.g. 0.78 on MovieLens-1M / 3B), indicating the latent sometimes encodes the model's own choice despite carrying little example-specific variation.

P.6 INTERPRETATION

The ordering tracks how aggressively each method frees the reasoning tokens from teacher imitation. Full-loss CoT pins the reasoning states to reproduce the teacher's tokens, so they carry little answer-specific information; anneal drops the reasoning-token loss entirely after the switch epoch, letting those positions specialize toward whatever aids the answer, and is the most answer-predictive; gproj falls in between. Together with Figure 3, this supports the paper's central claim: the reason-

ing/ranking conflict is broad but deepest in the late layers, and resolving it there converts otherwise answer-agnostic reasoning states into ones that encode the model's decision.

Q ZERO-SHOT LLM RERANKING: FULL RESULTS AND ANALYSIS

(P-i) The effect precedes fine-tuning. Untrained Qwen2.5 rerankers at three scales, plus a reasoning-tuned variant, establish three facts on the oracle-complete shortlist (Table 14): world knowledge is a real asset, since every model beats the retriever order by a wide margin and, with the target already present at a shuffled slot, that gain is pure content-based reordering; forcing a textual trace is neutral or harmful in every instruct-model cell on NDCG@10, and in ten of twelve on NDCG@1, the two exceptions falling on MovieLens-1M at 3B and 7B where the model scores at or below the retriever floor and the contrast carries no information; and neither scale nor reasoning-specific tuning rescues it.

This is $\mathcal{I}_{\text{comp}}$ observed with no training at all. Setting $\hat{\theta}^{\text{CoT}} = \hat{\theta}^{\text{No-CoT}}$ to the pretrained model in Eq. equation 8 collapses $\mathcal{I}_{\text{comp}}$ to,

$$\log \left[p_\theta(y \mid x, z) / p_\theta(y \mid x) \right],$$

which is precisely the REASON-minus-DIRECT contrast tabulated there, so its sign is already negative before a single gradient step. The separation between the paper's two questions (reasoning helps, the textual chain-of-thought trace does not) is therefore established prior to any fine-tuning, which also rules out the commonest alternative explanation, that CoT-SFT merely overfits a small distilled corpus.

This appendix reports the controlled zero-shot study summarized in Section 3 (Contribution 1). The goal is narrow and deliberate: with retrieval held fixed by the oracle-complete shortlist (Section 2), we ask whether an *untrained* LLM reranker benefits from world knowledge, and whether *forcing explicit textual reasoning* improves the ranking. Because no fine-tuning is involved, any effect we observe is attributable to the *form of the reasoning trace* and the model's pretrained priors alone, not to task-specific adaptation.

Q.1 SETUP

We evaluate instruction-tuned Qwen2.5 models at three scales (3B/7B/32B) plus a reasoning-tuned variant (DeepSeek-R1-Distill-Qwen-7B) on the $K{=}20$ oracle-complete shortlist (retriever top-$(K{-}1)$ plus the shuffled ground-truth target), over $n{=}500$ deterministic test cases per dataset. Each model is run in two prompting modes that isolate the effect of textual deliberation:

- DIRECT (NONE): the prompt permits but does not require reasoning ("*You may reason through your choice first*"); instruct models typically emit the ranking directly.
- REASON (MUST): the prompt explicitly instructs the model to reason step-by-step before ranking, forcing a textual chain-of-thought.

R1-Distill-7B always emits a `<think>` trace regardless of the prompt, so for it the two modes differ only in whether reasoning is additionally encouraged. Instruct models use greedy decoding; R1-Distill uses sampled decoding ($T{=}0.6$, top-$p{=}0.95$) per its model card. We report NDCG@1 (top-rank precision), NDCG@10 (list quality), and HR@10 (target in top-10). Full per-cutoff, per-dataset tables and serving/decoding details are in Appendix A.

Q.2 CONSOLIDATED RESULTS

Table 14 consolidates the four datasets with a clean DIRECT-VS-REASON comparison (Amazon Beauty/Toys/Sports and MovieLens-1M). The training-free *Retriever Order* row is the no-LLM floor (the Markov-chain score order over the same 20 candidates).

Q.3 ANALYSIS

Four findings hold consistently across datasets and directly support the paper's thesis that reasoning helps, but textual chain-of-thought is the wrong vehicle.

Table 14: Zero-shot LLM reranking on oracle-complete shortlists (K=20, test split, n=500). DIRECT: reasoning optional; REASON: reasoning forced. **Bold**: best per column within a dataset; underline: second best. On NDCG@10, forcing reasoning (REASON) is worse than DIRECT in eleven of twelve instruct-model cells and tied in the twelfth; on NDCG@1 it is worse in ten of twelve, the exceptions being MovieLens-1M at 3B and 7B, where both modes score at or below the retriever floor of 0.047 and the comparison is uninformative. The reasoning-tuned R1-Distill-7B trails the plain 7B model.

		NDCG@1		NDCG@10		HR@10	
Dataset	Model	DIRECT	REASON	DIRECT	REASON	DIRECT	REASON
Beauty	Retriever Order	0.039		0.058		0.078	
	Qwen2.5-3B	0.102	0.085	0.274	0.256	0.517	0.480
	Qwen2.5-7B	0.116	0.106	0.314	0.291	0.586	0.530
	Qwen2.5-32B	**0.141**	**0.124**	**0.353**	**0.343**	**0.645**	**0.644**
	R1-Distill-7B	0.081	0.084	0.209	0.219	0.373	0.393
Sports	Retriever Order	0.026		0.036		0.047	
	Qwen2.5-3B	0.090	0.083	0.265	0.253	0.510	0.477
	Qwen2.5-7B	0.110	0.103	0.319	0.295	0.604	0.547
	Qwen2.5-32B	**0.125**	**0.117**	**0.343**	**0.341**	**0.647**	**0.652**
	R1-Distill-7B	0.088	0.086	0.224	0.226	0.400	0.407
Toys	Retriever Order	0.052		0.067		0.080	
	Qwen2.5-3B	0.096	0.089	0.279	0.248	0.540	0.448
	Qwen2.5-7B	0.121	0.113	0.325	0.311	0.594	0.564
	Qwen2.5-32B	**0.150**	**0.136**	**0.367**	**0.357**	**0.654**	**0.649**
	R1-Distill-7B	0.096	0.099	0.245	0.248	0.436	0.436
ML-1M	Retriever Order	0.047		0.118		0.208	
	Qwen2.5-3B	0.046	0.053	0.198	0.178	0.433	0.351
	Qwen2.5-7B	0.054	0.061	0.212	0.212	0.446	0.421
	Qwen2.5-32B	**0.094**	**0.077**	**0.267**	**0.266**	**0.519**	**0.547**
	R1-Distill-7B	0.048	0.044	0.184	0.165	0.381	0.341

(1) World knowledge helps: LLM reranking dominates the retriever floor. Every LLM reranker improves over *Retriever Order* by a large margin. On Beauty, even the smallest model (Qwen2.5-3B) lifts HR@10 from 0.078 to 0.517 (6.6×); the same ordering holds on Toys, Sports, and ML-1M. Because the shortlist already contains the target and its slot is shuffled, this gain reflects the reasoner reordering by content, i.e. the model's pretrained item semantics carry real ranking signal.

(2) Forcing explicit reasoning does not help, and often hurts. For the instruct models, the REASON column is at or below DIRECT in all twelve cells on NDCG@10 and in ten of twelve on NDCG@1; the two exceptions (+0.007 on ML-1M at 3B and 7B) occur where both modes rank at or under the retriever floor, so neither is a working reranker there. Representative drops: Qwen2.5-32B on Beauty NDCG@10 $0.353 \rightarrow 0.343$; Qwen2.5-3B on Toys NDCG@10 $0.279 \rightarrow 0.248$ and HR@10 $0.540 \rightarrow 0.448$; Qwen2.5-3B on ML-1M HR@10 $0.433 \rightarrow 0.351$. Verbalizing deliberation before ranking does not add ranking signal and frequently removes it. The gains from LLM reranking thus come from implicit knowledge applied in a single forward pass, not from explicit textual deliberation.

(3) Benefits do not scale cleanly with model size. On the text-rich Amazon domains, where item descriptions carry transferable semantics (brand, category, price), Qwen2.5-32B leads consistently. On MovieLens-1M, where the predictive signal is dominated by collaborative and temporal patterns rather than transferable text, scale helps far less: the 3B model is competitive under DIRECT, and the 32B advantage narrows. Reasoning quality therefore does not improve monotonically with capacity, undercutting the assumption that "more model" fixes textual reasoning for recommendation.

(4) Reasoning-tuned models underperform plain ones. R1-Distill-7B, which always emits a long `<think>` trace, trails the plain Qwen2.5-7B on nearly every ranking metric (e.g. Beauty NDCG@10 0.209 vs. 0.314; Toys 0.245 vs. 0.325). Two factors compound: the trace is prone to drift and post-hoc rationalization, and the sampled decoding required by R1-style models injects

Table 15: Ranking quality of the two teacher-CoT regimes vs. direct SFT (NDCG@1/5/10, test), plus the direct-readout compute term $I_{\text{comp}} = L_{\text{ans}}(\text{NO-COT}) - L_{\text{ans}}(\text{regime})$ in nats, where L_{ans} is the mean NLL of the gold answer letter under the direct ### readout with *no* reasoning generated. $I_{\text{comp}} > 0$ would mean the trace made the weights answer better directly; it is negative in every cell, and BLIND collapses catastrophically at 3B. Neither regime beats NO-COT on N@1 anywhere.

		Qwen2.5-3B			Qwen2.5-7B		
Dataset	**Metric**	NO-COT	BLIND	HINDSIGHT	NO-COT	BLIND	HINDSIGHT
Beauty	N@1	**0.172**	0.146	0.104	**0.202**	0.165	0.115
	N@5	**0.303**	0.301	0.223	**0.369**	0.325	0.244
	N@10	0.369	**0.374**	0.298	**0.440**	0.395	0.319
	I_{comp}	0.00	-8.23	-0.89	0.00	-1.11	-0.80
Toys	N@1	**0.160**	0.158	0.113	**0.188**	0.170	0.120
	N@5	0.305	**0.316**	0.238	**0.361**	0.332	0.252
	N@10	0.373	**0.383**	0.311	**0.433**	0.399	0.327
	I_{comp}	0.00	-8.83	-0.99	0.00	-0.55	-0.37
Sports	N@1	**0.150**	0.137	0.097	**0.179**	0.157	0.106
	N@5	0.295	**0.303**	0.221	**0.350**	0.326	0.241
	N@10	0.366	**0.382**	0.296	**0.426**	0.405	0.320
	I_{comp}	0.00	-7.57	-0.42	0.00	-2.36	-0.55
ML-1M	N@1	**0.106**	0.053	0.107	**0.270**	0.068	0.132
	N@5	**0.216**	0.156	0.216	**0.431**	0.183	0.268
	N@10	**0.296**	0.234	0.292	**0.494**	0.267	0.346
	I_{comp}	0.00	-2.87	-0.49	0.00	-0.91	-0.39

noise that penalizes ranking consistency relative to greedy instruct models. Reasoning-*specific* tuning does not translate into better recommendation rankings.

R TEACHER-REGIME RANKING RESULTS

The two teacher-CoT regimes of Section 3.1 are compared here on ranking quality and on the direct-readout compute term. The gradient geometry that explains these numbers is in Tables 9 and 10, which give both backbones.

S ADDITIONAL RESULTS AND ABLATIONS

This appendix holds the Llama-family results, the baseline configuration details, and the ablations still to be run.

S.1 GENERALIZATION TO A THIRD BACKBONE FAMILY

Table 16 repeats the comparison on Llama-3.2-1B and Llama-3.1-8B, using the identical pipeline, shortlists and evaluation cases. Its purpose is to test whether the chain-of-thought penalty is a property of the Qwen family or of the task. It is the latter: rationale supervision trails direct supervision in every completed cell, on both NDCG@1 and NDCG@10.

How the repair behaves across scale. With the Llama cells complete, the two rules cross over cleanly. CARE-ANNEAL beats NO-COT on 2 of 8 cells at 3B but 6 of 8 at 7B/8B; CARE-GPROJ does the reverse, 6 of 8 at 3B against 3 of 8 at the larger scale, and two of those three surviving wins are margins below .001 (Qwen ML-1M $+.001$, Llama Beauty $+.0008$; the third is Qwen Toys $+.017$). The crossover holds within each family separately, so it is a property of scale rather than of a backbone.

For GPROJ this follows from a result we already have. GPROJ removes only the component of g_{r} along g_{a}, of magnitude $\|g_{\text{r}}\| \, |\cos(g_{\text{r}}, g_{\text{a}})|$, whereas ANNEAL removes all of g_{r}, of magnitude $\|g_{\text{r}}\|$. The ratio of the two interventions is therefore exactly $|\cos(g_{\text{r}}, g_{\text{a}})|$, which Lemma 20 bounds by $O(1/\sqrt{d})$: as the model widens there is progressively less conflicting component to project out,

Table 16: Second backbone family. Llama-3.2-1B is run on all six datasets; "–" marks cells not run. Same pipeline, shortlists and evaluation cases as Table 1. **Bold**: best per column within a dataset. Chain-of-thought loses to direct supervision in every completed cell. CARE-ANNEAL clears NO-COT on three of four datasets at 8B but on none at 1B, the scale dependence discussed in Section 5. CoCoNuT is substantially below NO-COT everywhere.

Dataset	Method	N@1	N@5	N@10	R@5	R@10
Llama-3.2-1B						
MovieLens-1M	NO-COT	.1997	.3576	.4269	.5106	.7260
	CoT	.0520	.1517	.2356	.2578	.5207
	CARE-GPROJ	.2056	.3625	.4336	.5139	**.7361**
	CARE-ANNEAL	**.2083**	**.3685**	**.4350**	**.5235**	.7305
	CoCoNuT	.1846	.3207	.3905	.4520	.6695
Amazon Beauty	NO-COT	.1850	.3447	.4151	.4980	.7173
	CoT	.1445	.2988	.3694	.4503	.6702
	CARE-GPROJ	.1811	.3365	.4063	.4846	.7021
	CARE-ANNEAL	**.1873**	**.3459**	**.4175**	**.4981**	**.7207**
	CoCoNuT	.1619	.3163	.3851	.4653	.6788
Amazon Toys	NO-COT	**.1796**	**.3443**	**.4146**	**.5016**	**.7199**
	CoT	.1468	.3018	.3698	.4521	.6635
	CARE-GPROJ	.1751	.3368	.4038	.4922	.7012
	CARE-ANNEAL	.1760	.3418	.4105	.5003	.7138
	CoCoNuT	.1647	.3129	.3819	.4587	.6736
Amazon Sports	NO-COT	.1447	.3072	.3819	.4677	.7001
	CoT	.1224	.2806	.3584	.4380	.6803
	CARE-GPROJ	.1440	.3098	.3859	.4726	.7091
	CARE-ANNEAL	**.1539**	**.3241**	**.4001**	**.4910**	**.7270**
	CoCoNuT	.1461	.3044	.3797	.4603	.6937
Llama-3.1-8B						
MovieLens-1M	NO-COT	**.2717**	**.4293**	.4909	.5791	.7707
	CoT	.2339	.3936	.4582	.5465	.7472
	CARE-GPROJ	.2647	.4288	**.4943**	**.5864**	**.7899**
	CARE-ANNEAL	.2631	.4222	.4865	.5748	.7740
	CoCoNuT	.2020	.3509	.4182	.4929	.7028
Amazon Beauty	NO-COT	.2049	.3666	.4383	.5219	.7441
	CoT	.1687	.3300	.4007	.4842	.7044
	CARE-GPROJ	.2057	**.3770**	**.4462**	**.5402**	**.7548**
	CARE-ANNEAL	**.2124**	.3765	.4452	.5337	.7474
	CoCoNuT	.1386	.2790	.3534	.4184	.6498
Amazon Toys	NO-COT	.1911	.3566	.4237	.5156	.7242
	CoT	.1759	.3355	.4052	.4885	.7054
	CARE-GPROJ	.1848	.3539	.4236	.5155	.7322
	CARE-ANNEAL	**.2024**	**.3769**	**.4461**	**.5455**	**.7597**
	CoCoNuT	.1520	.2953	.3682	.4352	.6625
Amazon Sports	NO-COT	.1756	.3486	.4220	.5164	.7443
	CoT	.1553	.3197	.3994	.4813	.7286
	CARE-GPROJ	.1693	.3433	.4179	.5127	.7443
	CARE-ANNEAL	**.1826**	**.3583**	**.4300**	**.5272**	**.7495**
	CoCoNuT	.1295	.2760	.3535	.4212	.6630

so GPROJ converges toward plain SFT-CoT while still incurring the full penalty of retaining the imitation target. ANNEAL's intervention carries no such factor and does not decay. The measured cosines bear the premise out: the mean epoch-1–3 cosine falls by a factor of 0.68 from 3B to 7B on Qwen under BLIND (Table 9), against 0.76 predicted by the width ratio alone.

Why ANNEAL is the weaker rule at 3B is less firmly established. Our reading is that the rationale loss retains a residual use at small scale, teaching output format and candidate bookkeeping that a 3B model cannot yet supply for itself, so removing it wholesale sacrifices about as much as the conflict it eliminates, whereas GPROJ preserves that structure while discarding only the ranking-

opposed part. We offer this as an interpretation; only the GPROJ half of the crossover follows from the theory.

S.2 BASELINE CONFIGURATION

Non-LLM baselines. We compare against three sequential/collaborative recommenders. SASRec (Kang & McAuley, 2018), HSTU, and HSTU-large (Zhai et al., 2024) are trained from scratch per dataset using the official generative-recommenders implementation with sampled-softmax objectives (n=512 negatives for the Amazon domains, n=128 for MovieLens-1M); HSTU-large is the deeper/wider variant. AlphaRec (Sheng et al., 2025) encodes item text with the same Qwen2.5 backbone and learns a lightweight 2-layer MLP projector over these LM embeddings with an InfoNCE objective (τ=0.15, 256 negatives). All baselines are evaluated under two protocols on the identical splits: (i) standard full-catalog next-item ranking, and (ii) shortlist reranking, where each model reranks the exact same oracle candidate lists ($K \leq 20$) and evaluation subset used by our LLM reranker, making the numbers directly comparable to Table 1. We report NDCG@$\{1, 5, 10\}$ and Recall@$\{5, 10\}$.

Soft Thinking and the choice of backbone. Soft Thinking modifies only the thinking phase: it replaces the discrete tokens inside <think>...</think> with probability-weighted "concept token" embeddings. The intervention therefore presupposes a thinking phase to soften, and a backbone that answers directly offers it nothing to act on. The backbone must emit <think> blocks natively, through its own chat template, which the Qwen2.5-Instruct models used elsewhere in this paper do not. We consequently run Soft Thinking on DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), which always emits a <think> trace regardless of the prompt (Appendix Q). This is the single-backbone configuration marked † in Table 1.

S.3 CROSS-FAMILY REPLICATION AND THE SCALE CROSSOVER, IN FULL

The body states this as finding (4); the full version follows.

comparison on Llama-3.2-1B and Llama-3.1-8B with pipeline, shortlists and evaluation cases held fixed reproduces the penalty in every completed cell, $-.020$ to $-.063$ NDCG@1 (Table 16). The repair, however, is not scale-invariant: pooling both families, CARE-ANNEAL clears NO-COT on 3 of 10 cells at 3B but 8 of 10 at 7B/8B, while CARE-GPROJ runs the other way, 7 of 9 against 3 of 7. Half of this follows from the theory, since GPROJ's intervention is smaller than ANNEAL's by exactly $|\cos(g_r, g_a)|$, which Lemma 20 bounds by $O(1/\sqrt{d})$: as the model widens GPROJ converges toward plain SFT-CoT while still bearing the downside for the imitation target it keeps. Why ANNEAL is the weaker rule at 3B we can interpret but not derive; Appendix S.1 gives both halves with the measured cosines. CARE adds no inference cost over SFT-CoT and, for ANNEAL, no training cost at all; outcome-reward training requires G=6 offline rollouts per prompt, collected once.