CHARACTERIZE THEN DISTILL: MECHANISTIC REASONING IN LARGE OUTPUT SPACES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning-trained language models can perform, zero-shot, multi-label tasks that require selecting a small set of relevant labels from a universe of thousands to hundreds of thousands of candidates. We ask how they do it mechanistically, and whether the mechanism can be distilled. We make the question measurable by treating each decision as a token-level event scored by the model's own decision margin: the token that picks a coarse region of the label space, the tokens that pick a label within it, and the token where the output departs from a close alternative (a near-miss) named earlier in the reasoning. Attribution, exact mean-ablation, knock-in into another example's context, and a null calibration that discounts generic heads then give individual attention heads causal standing. On clinical coding of hospital discharge summaries (MIMIC-IV), with all 5,651 candidate diagnosis codes in context, a small, global, phase-structured set of heads is necessary and sufficient, by ablation and knock-in, on essentially every summary; distinct head families attend to the candidate region and back to the near-miss named earlier; and, for the mentions decided in the reasoning, the region can already be elicited several tokens before the code, from a disjoint mid-layer set that reads the input. We introduce MISTILL: unlike chain-of-thought distillation, which transfers only the teacher's reasoning text, it also supervises the student's pooled attention at exactly these decision events. Read on heads found after training, it nearly doubles the causal recovery of the contrastive decision in a cross-family student and adds a small, seed-stable gain in one that already carries most of it, with no detected task difference when both objectives train bf16 weights and a task cost with fp32 master weights.
Code: `anonymous.4open.science/r/mistill-code-anon-3D07`

1 INTRODUCTION

Many practical labelling tasks ask for a small, unordered set of labels chosen from a fixed universe of thousands to hundreds of thousands, scored by micro-F1, macro-F1 and propensity-scored metrics that up-weight rare labels (Schultheis et al., 2023; Ullah et al., 2024; Zhang et al., 2025; Ullah et al., 2026). Clinical coding is the canonical case: a discharge summary receives a handful of the tens of thousands of ICD-10-CM codes (Mullenbach et al., 2018; Edin et al., 2023; Soroush et al., 2024; Motzfeldt et al., 2025). Reasoning-trained language models now attempt such tasks zero-shot with the candidates in context, writing long traces that narrow to a region of the label space and then argue between close alternatives (Guo et al., 2025). How do they make those decisions internally, and can a smaller model be taught the mechanism rather than only the text?

We answer the first with an instrument built for large label spaces. With the full inventory of 5,651 ICD-10-CM codes in the prompt, `Qwen3-14B` writes each code token by token, so every coding decision is the choice of one token: the chapter letter, each later character, and the character at which a code diverges from a sibling the trace named earlier (Figure 3). We score each decision by the model's own margin and give individual heads causal standing by attribution, exact mean-ablation, knock-in into another patient's context and a null calibration (Section 2). A small, global, phase-structured set of heads carries each decision on nearly every note: eight of 1,600 heads remove 0.62 of the chapter margin when ablated and restore 0.67 of it when knocked into another patient's context. Region readers attend to the emitted chapter's inventory lines, a contrastive head re-reads the near-miss named earlier, and, for the 36% of chapter mentions decided in the reasoning, the chapter can be

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.

elicited a median of 7 tokens before the letter, from a mid-layer set that shares none of the emission set's top eight (Section 3). The picture is largely preserved on a retrieved shortlist.

The characterization supplies what a distillation objective needs: where the decisions are made, which heads carry them, and what those heads read. MISTILL adds to chain-of-thought distillation a loss on the student's pooled attention at the teacher's decision events, over the student's own causally ranked heads and tokenizer-invariant span buckets (Section 4). Measured causally on heads found in the student after training, it raises the recovery of the contrastive decision: on `SmolLM3-3B`, whose undistilled recovery of it is 0.416, MISTILL raises it to 0.789, and on `Qwen3-4B`, which already carries most of it, it adds a small, seed-stable gain at the contrastive events over chain-of-thought distillation trained at the same precision (+0.032), though not at character-level decisions. Controls locate the transfer: the student's own top heads supervised at random positions fall below the untrained student, and the gain lands in the student's natural top heads whichever heads are supervised. Its task cost depends on the training precision: with bf16 weights, as all its ablations train, MISTILL is not distinguishable from chain-of-thought distillation trained the same way; with fp32 master weights it trails it (per-note micro-F1 difference −0.024 [−0.039, −0.009]).

Contributions. (1) An instrument that turns decisions over a large label space into token-level events with causal, calibrated, per-note head effects (Section 2). (2) A characterization of a reasoning-trained coder: global phase-structured head sets, what they read, and where in the reasoning the chapter can first be elicited (Section 3). (3) MISTILL, and evidence of what it transfers: the ranked heads need the teacher's decision positions, and the gain lands, reproducibly across seeds, in the student's own top heads; its task cost depends on the training precision (Section 4). (4) A metric set reported with its failures: three mechanism readouts do not separate the arms, and output validity and task score come apart. Causal claims rest on interventions on the model's own margin; attention patterns are described, not intervened on; the study covers one teacher family and one clinical task.

2 SETTING AND INSTRUMENT: DECISIONS ARE TOKENS

2.1 TASK INSTANCE, TEACHER, CORPUS

To ask which heads carry a coding decision, we fix the task, locate the decision in the trace and define a head's effect on it. The task is ICD-10-CM coding of `MIMIC-IV` discharge summaries (Johnson et al., 2023b) with the full inventory of 5,651 diagnosis codes and their descriptions in the prompt (about 106k tokens with the note); retrieval enters only for distillation (Sec. 3.5). The teacher is `Qwen3-14B` (Yang et al., 2025) in thinking mode, with 40 layers of 40 heads. The corpus is 100 traces over 94 notes spanning 21 of the 22 ICD-10-CM chapters, with teacher micro-F1 0.168; we call each trace a note. Appendices A and B give data access, prompt and decoding.

2.2 DECISIONS ARE TOKENS

The teacher's tokenizer writes a code as a chapter-letter token followed by one token per character (N18.9 as `␣N`, `1`, `8`, `.`, `9`), so each coding decision is the choice of one token (Figure 3). A P1 event is the chapter letter at a code's first mention; the inventory lines sharing that letter are the code's region (the 25 letters map imperfectly onto the 22 ICD-10-CM chapters). A P2 event is each later character; a P2C event is the character at which a code diverges from its near-miss, a sibling code named earlier in the trace (N18.9 after N18.1). Phase 1 is P1; Phase 2 is P2 with its contrastive case P2C. NULL events, prose positions with a random letter as target, serve for calibration. The 100 traces hold 2,064 P1, 7,688 P2 and 3,855 P2C events (Appendix C). At the position q whose output is the decided token, the metric is the model's own decision margin, $\text{margin}(q) = \text{logit}(\text{target}) - \text{logsumexp}(\text{competitors})$, a multi-way form of the usual patching metric (Zhang & Nanda, 2024).

2.3 MEASURING A HEAD'S EFFECT ON A DECISION

A hook at the input of each layer's output projection replaces one head's output at the event, either with its mean over the note (mean-ablation) or with a value from another run (knock-in; Heimersheim & Nanda, 2024). A head's importance is the margin it removes, $I_h(q) = \text{margin}(q) - \text{margin}(q \mid h \text{ mean-ablated})$; a set's *necessity* is the share of the clean margin its ablation destroys (above 1, the decision flips). First-order attribution for all 1,600 heads from one backward pass (Syed et al., 2024)

Table 1: Global per-phase sets: necessity (share of the clean margin destroyed by mean-ablation; > 1: flipped) and knock-in recovery in another patient's context, against random and bottom-k sets. Medians over notes; "notes": where the global set beats random (of 100; P2C: of 92). [R-C3]

	P1 (chapter)				P2 (character)				P2C (contrastive)			
k	global	random	bottom	notes	global	random	bottom	notes	global	random	bottom	notes
Necessity: share of the clean margin destroyed by mean-ablating the set (> 1: decision flipped)												
3	0.24	0.00	−0.15	92	0.41	0.01	−0.04	95	0.26	0.00	−0.06	90
8	0.62	0.00	−0.36	95	0.78	0.00	−0.18	98	0.63	0.00	−0.06	90
40	1.24	0.04	0.89	100	1.61	0.06	0.76	97	0.86	0.04	−0.50	90
100	1.60	0.09	1.57	100	2.21	0.26	1.41	99	0.89	0.07	−0.16	90
Sufficiency by knock-in: share of the clean decision restored in another patient's context												
3	0.16	0.00	−0.16	100	0.24	0.00	−0.45	100	0.05	0.00	−0.33	80
8	0.67	0.00	−0.28	100	0.48	0.01	−0.54	100	0.69	0.00	−0.54	92
40	0.86	0.04	−0.30	100	0.79	0.02	−0.81	100	0.85	0.02	−0.63	92
100	0.96	0.09	−0.33	100	0.89	0.05	−0.80	100	0.94	0.05	−0.78	92

nominates the top-40 at each event, plus the running global top-16, for exact ablation. Its correlation with exact effects is 0.36 over all heads under the hook used at the code token, which detached the graph at every layer, and 0.84 with the graph intact, as used at the locus; every set is therefore judged by exact intervention. Forward passes vary by 0.046 nats (s.d.), so we report means over notes.

For *sufficiency*, the corrupted run is the next note in the corpus at a matched event with a different emitted choice; the clean token at q is run in that context with the set's clean outputs knocked in. Knocking in all 1,600 heads returns the clean margin (median difference 0.000 nats), which scales recovery, $(\text{margin}_{\text{set}} - \text{margin}_{\text{corr}})/(\text{margin}_{\text{all}} - \text{margin}_{\text{corr}})$, from 0 (the other decision stands) to 1. At NULL events, where nothing is decided, the ranking procedure declares 309 heads positive and 282 negative (Benjamini–Hochberg, FDR 0.05; Benjamini & Hochberg, 1995); decision-specific importance $I_c = I - I_{NULL}$, averaged per note, makes 0 discoveries between two halves of the NULL events. A phase's ranking by I_c is its global set (top-k), tested against count-matched random heads, the note's own ranking and the global bottom-k. Appendix C validates the hook, and Appendix B discloses a retired instrument. Every head thus has a calibrated causal effect on every event.

3 Which heads carry the decisions?

3.1 A small global set is necessary and sufficient on nearly every note

Which heads carry each coding decision? First, does one fixed set of heads per decision type carry the decision on every note? Nearly (Table 1, Appendix D). Ablating the global top-8 of 1,600 heads destroys 0.62, 0.78 and 0.63 of the chapter, character and contrastive margins, and the top-40 flips chapter and character decisions; random sets of the same size remove almost nothing. In another patient's context, which favours a different chapter by a median of 8.7 nats, the chapter top-8's clean outputs recover 0.67 of the clean decision and the top-40 0.86; random sets recover at most 0.09. For chapter and character decisions the global set beats random on 92–100 of 100 notes for necessity and on 100 for knock-in (contrastive: 90 and 92 of 92; 80 at $k = 3$), and its bottom-k oppose the decision (Table 1). As a carrier it beats the note's own ranking only for the chapter decision at $k \geq 8$ (0.67 against 0.45); for characters the note's own ranking, fitted to the events it is tested on, is better at every k. We keep global sets because per-note sets are not identifiable at this event count (Appendix B), their top-40 knock-in is within 0.04 of the note's own for character and contrastive decisions, and their top-40 is largely shared across candidate presentations (Sec. 3.5).

3.2 The set is phase-structured

Do the decision types share heads? Partly (Appendix D, Table 8). `L39H38` is a shared core, first by I_c in every phase (first-order at P2C). Around it are chapter-specific heads (`L26H36`, `L26H37`, `L27H11`, `L27H13`), character-specific heads (`L39H37`, `L39H12`, `L38H22`, `L36H10`) and contrastive-specific heads (`L30H11`, `L38H39`), which rank near the bottom for chapter and character decisions. The

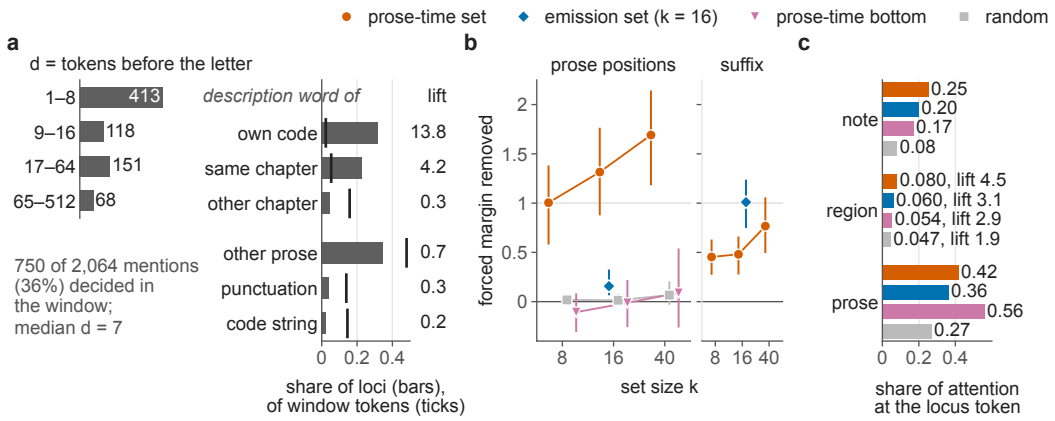


Figure 1: Where the chapter can be elicited, and by which heads. (a) Mentions decided in the reasoning: distance d from locus to chapter letter, and the classes of locus token (bars: share of loci; ticks: share of window tokens; right, lift). (b) Share of the forced chapter margin removed when a fixed set is mean-ablated only at the window positions (left) or only at the probe suffix (right): prose-time top-k, emission top-16, prose-time bottom-k, random-k; medians, whiskers 10th–90th percentile. (c) Attention at the locus token (natural context), four 16-head panels; descriptive. [R-C6]

chapter top-40 has 0 heads in layers 0–11 and 24 in layers 24–31: at the code token, the chapter decision is executed in middle and late layers. The contrastive ranking is first-order (the attribution pass made no exact patches at contrastive events); exact ablation of its 56 nominated heads reorders its top, in raw importance $\boldsymbol{I}$: `L39H12` 0.64, `L39H10` 0.36, `L30H11` 0.28, then `L39H38` 0.22.

3.3 What the heads read

What do these heads attend to? These readings are descriptive (attention is not intervened on). We split each attention row at the event by key-position type (Appendix E, Figure 4). In the chapter set, *region readers*, mid-layer heads led by `L26H36` and `L28H21`, put up to 0.47 of their attention on the emitted chapter's inventory lines, a region lift of up to 9.4 (attention share over token share), and they matter more on events where they read more of them (correlation +0.44); *prose readers* at the code token (`L39H38`, `L39H13`, `L38H24`, `L27H11`) put 0.73 to 0.84 of theirs on the reasoning prose. At contrastive events `L30H11` puts 0.33 of its attention on the earlier mentions of the near-miss (random heads 0.02; the code's own 0.13), more than on the code's own on 91 of 92 notes: a near-miss loop-back onto the sibling named earlier (the read-to-write link is correlational).

3.4 Where the chapter can be elicited, and by which heads

At the code token the chapter margin is usually already decisive (median P1 margin +6.02 nats). How early can it be elicited? A forcing probe cuts the context at an earlier position and makes the model name a code; the locus is the earliest position from which the forced chapter margin stays at least 2 nats (Appendix F). For 36% of the 2,064 chapter mentions the locus lies in the reasoning window since the previous code: a median of 7 tokens before the letter, reached in one step (median jump 8.3 nats), and in 32% of these cases at a word of the code's own inventory description (lift 14; Figure 1a). Another 29% inherit the chapter from an earlier code, and for 32% the probe never becomes decisive; neither group is attributed. Attribution over the 64 tokens ending at the locus (the attribution window) yields the *prose-time set*, led by `L24H19`, `L22H16` and `L24H1`: mid-layer (31 of its top-40 in layers 12–23), it shares 0 of its top-8 and 9 of its top-40 heads with the emission set, the chapter global set. Ablated only at the window positions (Figure 1b), its top-8 removes 1.00 of the forced margin, above random (0.02) on 100 of 100 notes, and its top-16 removes 1.32 against 0.16 for the emission top-16; at the suffix positions the order reverses (emission top-16 1.01, prose-time top-16 0.48). At the locus the prose-time heads read the note (0.25 of their attention; random 0.08; Figure 1c). A mid-layer set that reads the note thus makes the chapter elicitable a few tokens before the letter; the emission set largely transports that choice into the output (the transport ranking, Appendix F).

3.5 The mechanism on a retrieved shortlist, and what distillation takes from it

Does this hold on the shortlist the students see (Sec. 4.2)? Rerun unchanged on new teacher traces for the same notes with the shortlist in the prompt (Appendix G), the characterization is largely preserved: I_c rankings agree at Spearman 0.55/0.63/0.55 (chapter/character/contrastive), with 31/31/32 of the top-40 shared; top-40 knock-in recovers 0.92/0.82/0.85; and the chapter can be elicited a median of 9 tokens before the letter from the same prose-time set (Spearman 0.79). What shifts is the chapter top-8: `L39H38` leaves it, and it recovers less (0.47 against 0.67). Distillation takes three objects from the shortlist run: *where* to supervise (the events, and the 8 tokens before each chapter letter, where the chapter is already elicitable), *which heads* to pool (each phase's global top-40, plus the prose-time set on that window: the teacher's pooled weights w_h^T), and *what they read* (the read-side categories, as span buckets). In sum, a small, global, phase-structured head set carries each coding decision, a mid-layer set makes the chapter elicitable before the emission set writes it, and both largely hold on the shortlist.

4 Distilling the mechanism

4.1 Mistill: supervising the student's pooled heads at the teacher's decision events

Mistill adds to chain-of-thought distillation a loss on the student's attention at the positions where the teacher makes its decisions. Both models are teacher-forced on the teacher's trace $z^T(x) = [\text{prompt}(x); \tau^T(x)]$ (the teacher's side precomputed once), and the student minimises

$$\mathcal{L} = \mathcal{L}_{\text{CoT}} + \lambda_{\text{P1}}\,\mathcal{L}^{\text{P1}} + \lambda_{\text{P2}}\,\mathcal{L}^{\text{P2}}, \tag{1}$$

where $\mathcal{L}_{\text{CoT}}$ is exactly the Vanilla CoT-KD loss (cross-entropy on the trace tokens plus a KL term to the teacher's stored top-32 logits for `Qwen3-4B`; a universal logit distillation term for `SmolLM3-3B`, whose vocabulary differs) and $\lambda_{\text{P1}} = \lambda_{\text{P2}} = 0.1$, selected on the validation notes (Appendix H.3).

Pooled attention over span buckets. The two models have different heads and, for `SmolLM3-3B`, different tokens, so both sides are pooled. Over heads, each side uses a causally ranked top-40 with weights $w_h^M = \text{softmax}(I_c/\tau)$: for the teacher ($M{=}T$) its global sets on the shortlist (Sec. 3.5); for the student ($M{=}S$) the same instrument run once on the untrained student, fixed for the run. Over tokens, attention is summed within span buckets that exist in both tokenizations — one per shortlist line, per 32-word window of the note and per sentence of the trace:

$$\bar{\alpha}^M_{q,b} = \sum_h w_h^M \sum_{j \in b} \alpha^{M,h}_{q,j}. \tag{2}$$

At P2 and P2C events the code's own inventory line, its category siblings' lines, its earlier mentions (S), the near-miss's earlier mentions (N) and the code's emitted prefix are separate buckets.

Two terms per phase. At each supervised position q,

$$\ell_q = \text{KL}\big(\bar{\alpha}^T_q \,\big\|\, \bar{\alpha}^S_q\big) + \big(m^S_q c^S_q - m^T_q c^T_q\big)^2, \tag{3}$$

where m_q is the pooled mass on the decision's region (the emitted chapter's lines at P1; the code's own line, S and its prefix at P2 and P2C) and c_q is the pooled heads' direct contribution to the event's margin — their writes projected on the margin's gradient direction in the unembedding, divided by $|\,\text{margin}_q\,|$ (floored at one nat). $\mathcal{L}^{\text{P1}}$ averages ℓ_q over every P1 event and its KL term over the $W{=}8$ tokens before it (Sec. 3.4); there the teacher is pooled over its prose-time set and the student over its own P1 set, as no student prose-time set was measured. $\mathcal{L}^{\text{P2}}$ averages ℓ_q over P2 and P2C events, located by character offsets and mapped to each tokenizer (Appendix H.1).

A pre-registered third term per phase, a squared error on c alone, was harmful at pilot size and was removed, and λ's range was extended below its pre-registered floor — both post-hoc (Appendix H.2); the window term was never ablated (Appendix L). Feature-KD also matches bucket-pooled attention, but over all heads of four mapped layer pairs at sampled positions; Mistill differs in which heads it pools, where the loss acts and in its interaction term, and Section 4.5 removes each choice in turn.

4.2 SETUP

Candidates, corpus and teacher. Distillation runs on a retrieved shortlist instead of the full inventory: 1,000 codes plus the category siblings of the top 200, from a linear head on frozen chunk embeddings fused with zero-shot description similarity (Appendix A). Its recall, 0.903 on the validation notes and 0.936 on the test notes, is a ceiling every arm shares; prompts shrink to about 30k tokens, and w^T comes from the teacher's replication on this shortlist (Sec. 3.5). The corpus is 500 training notes (493 with a valid teacher trace) and 150 test notes, new and one per patient; the 100 characterization notes are the validation set. The teacher scores 0.227 micro-F1 on the test notes.

Students. `Qwen3-4B` shares the teacher's family and tokenizer; `SmolLM3-3B` (Bakouch et al., 2025) comes from another family, and its tokenizer merges digit runs, so some of its P2 events are chunk-level. Every arm samples in thinking mode with the teacher's settings and a 16,384-token cap; `Qwen3-4B` runs under YaRN (Peng et al., 2024) ×2 for up to 52k tokens (Appendix I.5).

Baselines. One strong method per family, from official code where it exists, at its published settings: Vanilla CoT-KD (cross-entropy and KL on the teacher trace), DistiLLM-2 (off-policy logit distillation; `Qwen3-4B` only), on-policy distillation (OPD; for `SmolLM3-3B` GOLD, its cross-tokenizer form), and Feature-KD, the generic attention and representation control. Every arm sees the same 493 notes for two epochs at batch 8, with the same optimiser, targets and evaluation (GOLD: one epoch, rollouts capped at 6,144 tokens); tuning budgets and wall-clock are not matched (Appendix I). MISTILL, OPD and GOLD train the student's weights in bf16 without an fp32 master copy, while Vanilla CoT-KD, Feature-KD and DistiLLM-2 keep fp32 master weights; since this alone changes how much the sequence-level objective learns, we report Vanilla CoT-KD at both precisions.

Metrics and statistics. External: micro-F1, macro-F1, P@8, P@15, PSP@15, tail macro-F1. Fidelity: trace PPL, BLEU-4 and LAS (a frozen `Qwen3-1.7B` predicting the codes with and without the rationale). Mechanism: M1, decision-margin agreement at the teacher's events (the quantity the loss supervises), and, on each arm's own heads, re-selected after training by the instrument of Sec. 2, so not tautological with the objective, M4 causal recovery (knock-in of its top-40 into another note's context, against a random set), M2 region lift, M3 near-miss loop-back and M5 locus distance (Appendix J). On the 150 paired test notes, `Qwen3-4B`'s Vanilla CoT-KD and MISTILL (both precisions) and Feature-KD have three seeds, `SmolLM3-3B`'s Vanilla CoT-KD and MISTILL two, every other cell one (marked). Task and fidelity differences carry paired bootstrap 95% intervals and Wilcoxon tests; mechanism differences, per-note sign tests, seed-matched and pooled over pairs.

4.3 THE TASK: NO DETECTED COST IN BF16, A COST IN FP32

Does supervising the decision events change what the student gets right? It depends on the training precision (Table 2). Trained like MISTILL and its ablations, with bf16 weights and no fp32 master copy, Vanilla CoT-KD scores 0.143 micro-F1 on `Qwen3-4B` against MISTILL's 0.149 (per note +0.007 [−0.005, +0.019]) and 0.055 on `SmolLM3-3B` against 0.051, and the two objectives also tie at 100 and 250 training notes (Appendix L.4). With fp32 master weights Vanilla CoT-KD reaches 0.187, a difference of +0.041 from the same objective in bf16, so, by the reading fixed before the control ran, the gap to fp32 Vanilla CoT-KD is largely the precision's (Appendix K.3). The objective does not share that gain: with fp32 master weights MISTILL scores 0.165 (+0.010 [−0.005, +0.025] from its bf16 runs) and trails fp32 Vanilla CoT-KD on the task (−0.024 [−0.039, −0.009]) and on LAS (−0.039 [−0.058, −0.021]). Its λ was selected with bf16 weights and not re-selected (Appendix H). What MISTILL does change is the readout it supervises: at the teacher's contrastive events the student's margin favours the teacher's choice more often (M1 P2C 0.803 against 0.773; higher on 348 of 441 note-pairs over three seed pairs), a lead that fp32 training removes (0.779 against fp32 Vanilla CoT-KD's 0.791). OPD, trained bf16 on the student's own rollouts, reaches 0.179 (one seed), level with fp32 Vanilla CoT-KD (−0.005 [−0.025, +0.015]), so the bf16 cost is claimed for off-policy objectives only. GOLD's lists parse (valid rate 0.613) yet score zero, naming categories rather than billable codes (Appendix K.2): the metric tiers must be kept apart. Supervising the teacher's decision events therefore shows no detected task cost when both objectives train bf16 weights, and a cost when both keep fp32 master weights; whether it transfers the mechanism is next.

Table 2: Distillation on MIMIC-IV (150 test notes; mean ± sd over seeds; single-seed rows marked). MISTILL, OPD, GOLD and (bf16) rows train bf16 weights, no fp32 master copy; the others keep fp32 master weights. M1: share of the teacher's contrastive events at which the student's margin favours the teacher's choice. Paired 95% intervals: Appendix K. [R-D1][R-D2][R-D3][R-D4][R-D5][R-D17]

		external (task)					fidelity		mech.
		F1							
Arm	Seeds	micro	macro	tail	PSP@15	valid	PPL	LAS	M1 P2C
Qwen3-4B									
Base[1 seed]	1	0.104	0.036	0.029	0.074	0.987	1.749	0.093	0.759
Vanilla CoT-KD	3	0.187 ±.007	0.072 ±.005	0.059 ±.010	0.138 ±.005	0.824 ±.028	1.516 ±.002	0.267 ±.004	0.791 ±.002
Vanilla CoT-KD (bf16)	3	0.143 ±.007	0.051 ±.002	0.041 ±.001	0.101 ±.003	0.898 ±.010	1.581 ±.002	0.183 ±.008	0.773 ±.000
Feature-KD	3	0.153 ±.010	0.054 ±.006	0.042 ±.003	0.104 ±.008	0.780 ±.046	1.586 ±.004	0.213 ±.017	0.771 ±.003
DistiLLM-2 [1 seed]	1	0.187	0.067	0.050	0.139	0.880	1.556	0.271	0.788
OPD [1 seed]	1	0.179	0.068	0.063	0.131	0.973	1.638	0.229	0.770
MISTILL	3	0.149 ±.002	0.054 ±.002	0.043 ±.002	0.108 ±.006	0.893 ±.012	1.590 ±.001	0.192 ±.008	0.803 ±.001
MISTILL (fp32)	3	0.165 ±.004	0.059 ±.005	0.046 ±.005	0.115 ±.003	0.824 ±.054	1.524 ±.001	0.228 ±.014	0.779 ±.001
SmolLM3-3B									
Base[1 seed]	1	0.003	0.002	0.000	0.002	0.153	1.947	−0.019	0.721
Vanilla CoT-KD	2	0.130 ±.025	0.048 ±.002	0.027 ±.012	0.085 ±.022	0.547 ±.066	1.621 ±.001	0.226 ±.016	0.787 ±.002
Vanilla CoT-KD (bf16)[1 seed]	1	0.055	0.016	0.009	0.031	0.380	1.703	0.081	0.756
Feature-KD [1 seed]	1	0.062	0.023	0.012	0.040	0.360	1.760	0.117	0.746
GOLD (6,144 cap)[1 seed]	1	0.000	0.000	0.000	0.000	0.613	2.132	−0.051	0.732
MISTILL	2	0.051 ±.013	0.017 ±.005	0.009 ±.001	0.030 ±.009	0.297 ±.033	1.708 ±.001	0.080 ±.000	0.741 ±.000

4.4 THE STUDENT'S OWN HEADS RECOVER MORE OF THE CONTRASTIVE DECISION

Does the student come to carry the teacher's contrastive decision in its own heads? We rerun the instrument of Section 2 on every trained student and knock its own top-40 heads into another note's context (M4; Table 3; per seed, Appendix K.1). These heads are re-selected after training and the readout is a causal knock-in effect, not the pooled attention the loss matches. Where recovery is low, MISTILL raises it: SmolLM3-3B's undistilled P2C recovery is 0.416, and MISTILL raises it to 0.789 and 0.774 on two seeds, against 0.541 and 0.523 for fp32 Vanilla CoT-KD (higher on 257 of 300 seed-matched note-pairs) and 0.476 for bf16 Vanilla CoT-KD (higher on 130 of 150). Where it is present — undistilled Qwen3-4B already restores 93% of the teacher's 0.847 — the gain is small but consistent: +0.011 over fp32 Vanilla CoT-KD pooled over three seed pairs (higher on 250, lower on 200) and +0.032 over bf16 Vanilla CoT-KD (higher on 304, lower on 146 of 450), the size of the gain on the supervised readout at matched precision. At the character-level P2 decisions the matched comparison runs the other way (−0.019, lower on 321 of 450), although MISTILL is above fp32 Vanilla CoT-KD there (+0.011): the gain is at the contrastive decisions, not at every decision the loss supervises. Three readouts do not separate the arms (Appendix J): P1 recovery is at ceiling in every arm, M3 barely moves, and every student's chapter locus is about as early as the teacher's (M5).

4.5 WHERE THE TRANSFER IS LOCATED: POSITIONS, HEADS, PHASES

Every arm here trains like MISTILL (bf16 weights, 493 notes, two epochs), so none of this depends on the precision question; an effect counts only if it holds per note against every MISTILL seed.

The ranked heads need the teacher's positions. Supervising the student's own top heads at random positions (RandPos) pushes P2C recovery *below the untrained student* on both students (0.631 against 0.788; 0.304 against 0.416), and the heads' reading of the code's prefix collapses with

Table 3: Mechanism on each arm's own causally ranked heads, found after training (150 test notes; medians over notes, then mean ± sd over seeds; single-seed rows marked; random-40 floors in parentheses): knock-in recovery of the own top-40 per phase; region lift and emitted-prefix share of the own top-16 P2C heads. Per seed: Table 29. [R-D6][R-D12][R-D15][R-D17]

		M4 knock-in, own top-40			read side (P2C)	
Arm	Seeds	P2C (random)	P2	P1	M2 lift (random)	prefix share
Teacher (`Qwen3-14B`; HD-002)		0.847 (0.015)	0.821	0.915	3.38 (1.95)	0.260
`Qwen3-4B`						
Base[1 seed]	1	0.788 (0.030)	0.815	0.946	2.56 (1.17)	0.185
Vanilla CoT-KD	3	0.859±.028 (0.036)	0.725±.053	0.946±.002	3.44±.32 (1.33)	0.208±.024
Vanilla CoT-KD (bf16)	3	0.839±.007 (0.037)	0.758±.007	0.952±.000	3.48±.12 (1.58)	0.188±.001
Feature-KD [1 seed]	1	0.844 (0.048)	0.888	0.962	3.05 (1.57)	0.228
MISTILL	3	0.872±.002 (0.025)	0.735±.001	0.956±.002	3.63±.06 (1.60)	0.269±.007
RandPos [1 seed]	1	0.631 (0.042)	0.771	0.907	2.93 (1.45)	0.076
`SmolLM3-3B`						
Base[1 seed]	1	0.416 (0.107)	0.809	0.985	3.57 (1.51)	0.121
Vanilla CoT-KD	2	0.532±.012 (0.101)	0.724±.015	0.976±.003	3.81±.08 (1.49)	0.172±.011
Vanilla CoT-KD (bf16)[1 seed]	1	0.476 (0.098)	0.867	0.985	3.29 (1.39)	0.151
Feature-KD [1 seed]	1	0.628 (0.099)	0.933	0.988	2.74 (1.37)	0.097
MISTILL	2	0.782±.011 (0.124)	0.867±.001	1.005±.001	2.65±.00 (1.34)	0.440±.001
RandPos [1 seed]	1	0.304 (0.095)	0.891	0.993	2.52 (1.36)	0.050

it. Random heads at random positions do no such harm (`Qwen3-4B` RandBoth 0.863, level with every MISTILL seed): the harm comes from the student's own ranked heads at random positions (event-weighted Vanilla CoT-KD: not run).

The gain lands in the student's own top heads, whichever heads are supervised. On `SmolLM3-3B` (undistilled P2C recovery 0.416, random floor 0.107), which heads are supervised sets how much is gained: the base's 40 most negatively attributed heads, which oppose the decision, yield at least as much as its top-40 (0.841 and 0.835 against 0.789 and 0.774), its top-8 less (0.671), and 40 random heads least (0.585). Yet the supervised heads are not where the mechanism ends up: in every arm, on both students, the own top-40 after training shares 27–38 heads with the untrained student's, and the bottom-40 arm's holds at most one of the 40 heads it supervised (Appendix L.2). On `Qwen3-4B`, whose base restores 93% of the teacher's decision, the choice barely matters — eight, forty or a hundred top heads reproduce MISTILL, as do forty random heads on one of two seeds — except that the most negatively attributed heads damage it (0.717, below the base).

The recovery is reproducible across seeds. MISTILL's three `Qwen3-4B` seeds agree within 0.004 and do not separate per note. Two seeds on random heads (RandHeads) differ (−0.033, lower on 111 of 150 notes), and Vanilla CoT-KD's seeds spread 0.052 with fp32 master weights and 0.013 with bf16 weights, two of the three bf16 pairs separating per note (p 10^{-4} and 0.04): seed stability appears to track the student's ranked heads.

Phases, terms and dose. The P2C gain is the Phase-2 supervision's: supervising only P2/P2C events keeps about nine-tenths of it and only P1 events loses most of it, while a third of the dose or either term alone keeps most of it (Figure 5). Neither M2 nor the prefix-reading share follows recovery across these ablations, so the claims rest on M4 (Appendix L).

5 RELATED WORK

Distillation. Training students on a teacher's rationales transfers multi-step reasoning (Hsieh et al., 2023; Ho et al., 2023; Magister et al., 2023; Shridhar et al., 2023); students can match a teacher's hidden states, attention maps or attention relations (Romero et al., 2015; Zagoruyko & Komodakis, 2017; Jiao et al., 2020; Wang et al., 2020), and LLM logit distillation has contrastive, on-policy and cross-tokenizer forms (Ko et al., 2025; Agarwal et al., 2024; Patiño et al., 2025). These transfer text,

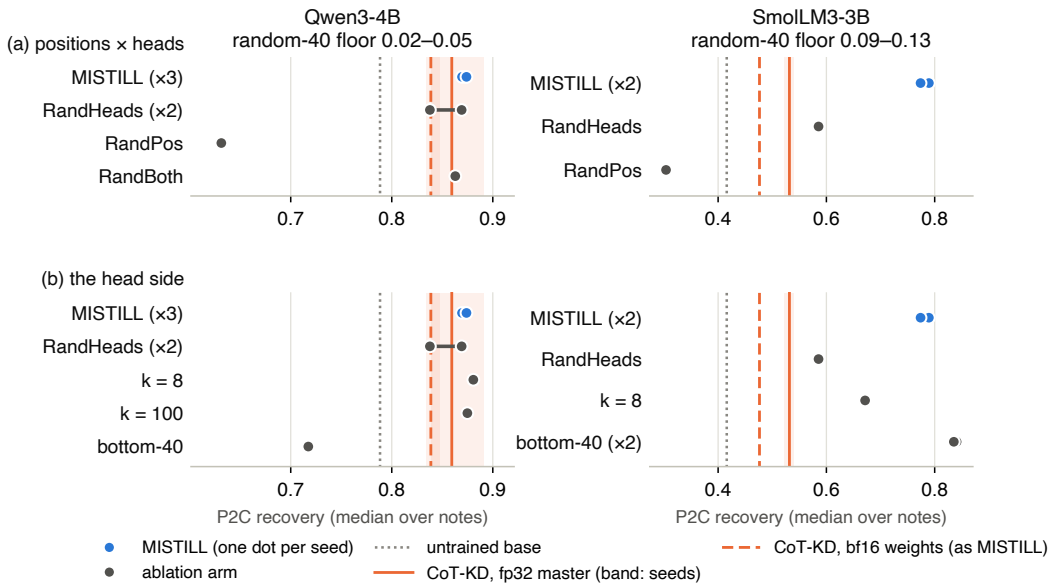


Figure 2: Where the transfer is located. (a) Own top heads supervised at random positions (RandPos) fall below the untrained student; random heads there do not. (b) On `SmolLM3-3B` the supervised heads set how much is gained; on `Qwen3-4B` only the most negatively attributed heads matter, and they hurt. Phases, terms, dose: Figure 5. bf16; one seed unless marked. [R-D7][R-D13][R-D14]

or act at every position on fixed layers. MISTILL keeps the sequence-level loss (Kim & Rush, 2016) and adds a target at the teacher's decision positions on each model's own causally ranked heads, so unlike circuit distillation (Wadhwa et al., 2025) it needs no component correspondence (Appendix N).

Interpretability and coding. Activation patching and circuit analyses localize behavior to components (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023a; Conmy et al., 2023), attribution patching screens components by gradient (Syed et al., 2024; Kramár et al., 2024), and metric and corruption choices matter (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). ICD coders are supervised classifiers with label-wise attention (Mullenbach et al., 2018; Edin et al., 2023), LLMs generate or retrieve candidate labels in context (Zhu & Zamani, 2024; D'Oosterlinck et al., 2024), and interpretability in coding has targeted trained classifiers (Wu et al., 2025a; Edin et al., 2024). New here: a reasoning model choosing from the full inventory in context, decisions as token-level events scored by the model's margin, and per-note necessity and sufficiency against a null calibration.

6 LIMITATIONS AND CONCLUSION

Limitations. The study covers one teacher family, one tokenizer whose code structure makes the decisions mechanical, and one clinical task. The read side is descriptive, the character ranking is first-order, and a locus is found only for the 36% of chapter mentions decided in the reasoning. A P2C event is defined by mention order, not by an annotated rejection. Distillation shares a retriever ceiling across arms and needs white-box access to both models. MISTILL and its controls train bf16 weights; with fp32 master weights Vanilla CoT-KD gains more than MISTILL, which then trails it (λ was selected in bf16), and the fp32 MISTILL checkpoints' own heads were not read. The baselines run at published settings, most controls and ablations have one seed, the window, τ and symmetric-window ablations were not run (Appendix L), and MISTILL trains slower than Vanilla CoT-KD (Appendix H).

Conclusion. In a reasoning-trained coder, each decision over a large label space is carried by a small, global, phase-structured head set; for mentions decided in the reasoning, the chapter can already be elicited in the prose, from a different, mid-layer set. Supervising a student's attention at those events raises the contrastive decision's causal recovery in its own top heads, whichever heads are supervised (ranked heads only at the teacher's positions), with no detected task cost in bf16 and a cost in fp32. Whether what the heads read is causal, and other teachers and label spaces, come next.

AI USE STATEMENT

In this work, we used generative AI tools (a large language model coding assistant) for implementing and debugging the experiment code (event extraction, attribution and patching, training and evaluation scripts, readers and figure scripts), for launching and monitoring runs on our clusters and rented GPUs, for analysing results against pre-registered readings, for maintaining the experiment logbook and planning documents, and for drafting and editing parts of the paper, including its tables, figures and bibliography entries. We have not used generative AI tools to generate data, labels or results: every number in the paper is produced by scripts from the experiment artifacts, and the tables and macros are generated rather than typed. All AI-assisted work was reviewed by the author: code was tested (unit tests, hook validation, instrument self-tests) and its outputs cross-checked against independent reads of the artifacts, every citation was checked against its primary source, and every claim was checked against the artifacts it cites. We take responsibility for the final content of this work.

ETHICS STATEMENT

`MIMIC-IV` is used under the PhysioNet credentialed-access data use agreement. No note text, reasoning-trace text or other protected content appears in the paper, its figures, the supplementary material or the released code: the paper reports only public ICD-10-CM codes, attention-head names and statistics, and the worked example (Figure 3) uses public codes, their inventory descriptions and placeholder text. Automated clinical coding is a decision-support setting; the models here are research artifacts, not validated for clinical use, and their errors (Section 4.3) would require human review in any deployment.

REPRODUCIBILITY STATEMENT

Appendix M holds one reproduction entry per reported result (identifiers [R-C1]–[R-C7] and [R-D0]–[R-D17]), each with its data, code, configuration, seeds, compute, command, expected artifact and tolerance. Every number in the paper is generated from the experiment artifacts by the readers named there (Appendix M). The data are public under credentialed access: any researcher can obtain `MIMIC-IV` from PhysioNet, and public code rebuilds the splits we use (Appendix A). Reruns reproduce the reported statistics within the stated intervals but not bit-exactly: decoding is stochastic and some kernels are nondeterministic (Appendix M). An anonymised snapshot of the code repository is at `https://anonymous.4open.science/r/mistill-code-anon-3D07`: the code, configurations and tests, the planning and walkthrough documents that record every run and result, and an index mapping each result identifier to the script that produces it. Its `results/` folder lists every reported number with the statistic it is and the artifact it was computed from, so each can be traced without access to the data. The run-by-run experiment logbook is withheld, since it records per-note details that fall under the data use agreement.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. The extreme classification repository: Multi-label datasets and code, 2016. URL http://manikvarma.org/downloads/XC/XMLRepository.html.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for LLMs. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=bwRxXiGO9A.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. MCC-KD: Multi-CoT Consistent Knowledge Distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6805–6820, 2023. doi: 10.18653/v1/2023.findings-emnlp.454. URL https://aclanthology.org/2023.findings-emnlp.454/.

Xiaoshu Chen, Sihang Zhou, Ke Liang, Xiaoyu Sun, and Xinwang Liu. Skip-Thinking: Chunk-wise Chain-of-Thought Distillation Enable Smaller Language Models to Reason Better and Faster. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12142–12157, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.610. URL https://aclanthology.org/2025.emnlp-main.610/.

Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. Unveiling the Key Factors for Distilling Chain-of-Thought Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15094–15119, 2025b. doi: 10.18653/v1/2025.findings-acl.782. URL https://aclanthology.org/2025.findings-acl.782/.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0719. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Karel D'Oosterlinck, Omar Khattab, François Remy, Thomas Demeester, Chris Develder, and Christopher Potts. In-Context Learning for Extreme Multi-Label Classification, January 2024. URL http://arxiv.org/abs/2401.12178. arXiv:2401.12178 [cs].

Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2572–2582. ACM, 2023. doi: 10.1145/3539618.3591918. URL https://doi.org/10.1145/3539618.3591918.

Joakim Edin, Maria Maistro, Lars Maaløe, Lasse Borgholt, Jakob Drachmann Havtorn, and Tuukka Ruotsalo. An unsupervised approach to achieve supervised-level explainability in healthcare records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4869–4890. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.280. URL https://aclanthology.org/2024.emnlp-main.280/.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Nilesh Gupta, Fnu Devvrit, Ankit Singh Rawat, Srinadh Bhojanapalli, Prateek Jain, and Inderjit Dhillon. Dual-encoders for extreme multi-label classification. In *International Conference on Learning Representations*, pp. 52858–52884, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e97d1081481a4017df96b51be31001d3-Abstract-Conference.html.

Yinhan He, Wendy Zheng, Yaochen Zhu, Zaiyi Zheng, Lin Su, Sriram Vasudevan, Qi Guo, Liangjie Hong, and Jundong Li. SemCoT: Accelerating Chain-of-Thought Reasoning through Semantically-Aligned Implicit Tokens. In *Advances in Neural Information Processing Systems*, volume 38, pp. 43455–43485, 2025. doi: 10.52202/085713-1445. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3ddbd473456a57e3cafb1ee51ddf8ff6-Abstract-Conference.html.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, April 2024. URL http://arxiv.org/abs/2404.15255. arXiv:2404.15255 [cs].

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. URL http://arxiv.org/abs/1503.02531. arXiv:1503.02531 [stat].

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830/.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pp. 10–20, Seattle, WA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.clinicalnlp-1.2. URL https://aclanthology.org/2022.clinicalnlp-1.2/.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 935–944. ACM, 2016. doi: 10.1145/2939672.2939756. URL https://doi.org/10.1145/2939672.2939756.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372/.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. *PhysioNet*, 2023a. doi: 10.13026/1n74-ne17. URL https://doi.org/10.13026/1n74-ne17. Version 2.2.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023b.

Minki Kang, Jongwon Jeong, Seanie Lee, Jaewoong Cho, and Sung Ju Hwang. Distilling LLM Agent into Small Models with Retrieval and Code Tools. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106501–106538, 2025. doi: 10.52202/085713-3553. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/99263f9bf46874e2b8b7f104b5063864-Abstract-Conference.html.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31044–31062. PMLR, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. AtP*: An efficient and scalable method for localizing LLM behaviour to components, March 2024. URL https://arxiv.org/abs/2403.00745v1.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. doi: 10.48550/arXiv.2308.03281. URL https://arxiv.org/abs/2308.03281.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.151. URL https://aclanthology.org/2023.acl-short.151/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Andreas Geert Motzfeldt, Joakim Edin, Casper L. Christensen, Christian Hardmeier, Lars Maaløe, and Anna Rogers. Code like humans: A multi-agent solution for medical coding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22612–22627. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1231. URL https://aclanthology.org/2025.findings-emnlp.1231/.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1101–1111, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1100. URL https://aclanthology.org/N18-1100/.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Kashyap, Stefan Winkler, Shao-Syuan Huang, Jie-Jyun Liu, and Chih-Jen Lin. Mimic-IV-ICD: A new benchmark for eXtreme MultiLabel classification, 2023. URL https://arxiv.org/abs/2304.13998.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Carlos Miguel Patiño, Kashif Rasul, Quentin Gallouédec, Ben Burtenshaw, Sergio Paniego, Vaibhav Srivastav, Thibaud Frere, Ed Beeching, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Unlocking on-policy distillation for any model family. Hugging Face article, October 2025. URL https://huggingface.co/spaces/HuggingFaceH4/on-policy-distillation.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations*, pp. 31932–31951, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/874a4d89f2d04b4bcf9a2c19545cf040-Abstract-Conference.html.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Erik Schultheis, Marek Wydmuch, Wojciech Kotłowski, Rohit Babbar, and Krzysztof Dembczyński. Generalized test utilities for long-tail performance in extreme multi-label classification. In *Advances in Neural Information Processing Systems*, volume 36, pp. 22269–22303. Curran Associates, Inc., 2023. doi: 10.52202/075280-0978. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/46994b3d6dd0fd5fca5f780af6259db5-Abstract-Conference.html.

Erik Schultheis, Wojciech Kotłowski, Marek Wydmuch, Rohit Babbar, Strom Borman, and Krzysztof Dembczyński. Consistent algorithms for multi-label classification with macro-at-k metrics. In *International Conference on Learning Representations*, pp. 13440–13469, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/3a02da3fdfd592d9a5273101c3546611-Abstract-Conference.html.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling Reasoning Capabilities into Smaller Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023. doi: 10.18653/v1/2023.findings-acl.441. URL https://aclanthology.org/2023.findings-acl.441/.

Ali Soroush, Benjamin S. Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W. Charney, Girish N. Nadkarni, and Eyal Klang. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI*, 1(5), 2024. doi: 10.1056/AIdbp2300040.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution Patching Outperforms Automated Circuit Discovery. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Nasib Ullah, Erik Schultheis, Mike Lasby, Yani Ioannou, and Rohit Babbar. Navigating extremes: Dynamic sparsity in large output spaces. In *Advances in Neural Information Processing Systems*, volume 37, pp. 117210–117236. Curran Associates, Inc., 2024. doi: 10.52202/079017-3721. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d4bdeed749a437de2cbe2e2c7e5a6a8a-Abstract-Conference.html.

Nasib Ullah, Jinbin Zhang, Jean Lucien Randrianantenaina, Erik Schultheis, and Rohit Babbar. HASTE: Hardware-aware dynamic sparse training for large output spaces. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.01117. To appear; arXiv:2606.01117.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning. GitHub repository, 2020. URL https://github.com/huggingface/trl.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3335–3341. International Joint Conferences on Artificial Intelligence Organization, 2020. doi: 10.24963/ijcai.2020/461. URL https://doi.org/10.24963/ijcai.2020/461.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. Investigating mysteries of CoT-augmented distillation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6071–6086, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.349. URL https://aclanthology.org/2024.emnlp-main.349/.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. Circuit distillation. *arXiv preprint arXiv:2509.25002*, 2025.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. SCOTT: Self-Consistent Chain-of-Thought Distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5546–5558, 2023b. doi: 10.18653/v1/2023.acl-long.304. URL https://aclanthology.org/2023.acl-long.304/.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

John Wu, David Wu, and Jimeng Sun. DILA: Dictionary Label Attention for Mechanistic Interpretability in High-dimensional Multi-label Medical Coding Prediction. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 1014–1038. PMLR, 2025a. URL https://proceedings.mlr.press/v259/wu25a.html.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval Head Mechanistically Explains Long-Context Factuality. In *International Conference on Learning Representations*, pp. 62143–62156, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9b77f07301b1ef1fe810aae96c12cb7b-Abstract-Conference.html.

JianZhi Yan, Le Liu, Youcheng Pan, Shiwei Chen, Yang Xiang, and Buzhou Tang. Towards Efficient CoT Distillation: Self-Guided Rationale Selector for Better Performance with Fewer Rationales. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7818–7835, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.413. URL https://aclanthology.org/2025.findings-emnlp.413/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1612.03928.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Jinbin Zhang, Nasib Ullah, Erik Schultheis, and Rohit Babbar. ELMO: Efficiency via low-precision and peak memory optimization in large output spaces. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76159–76174. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25bv.html.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1010. URL https://aclanthology.org/2024.emnlp-main.1010/.

Yaxin Zhu and Hamed Zamani. ICXML: An In-Context Learning Framework for Zero-Shot Extreme Multi-Label Classification. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2086–2098, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.134. URL https://aclanthology.org/2024.findings-naacl.134/.

Xianwei Zhuang, Zhihong Zhu, Zhichang Wang, Xuxin Cheng, and Yuexian Zou. UniCoTT: A Unified Framework for Structural Chain-of-Thought Distillation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3baOKeI2EU.

A TASK, DATA AND CANDIDATE PRESENTATION

Supports: Sec. 2.1, Sec. 4.2; [R-C1], [R-D0]

Data and access. The data are MIMIC-IV discharge summaries, from its MIMIC-IV-Note module, with their ICD-10 codes (Johnson et al., 2023a;b). Both are public under credentialed access: any researcher can obtain them from PhysioNet after completing its credentialing, which includes a course in human-subjects research, and signing the data use agreement; we use them under that agreement. ICD coding on MIMIC-III and MIMIC-IV is a standard clinical NLP benchmark for multi-label classification over a large label space (Mullenbach et al., 2018; Vu et al., 2020; Huang et al., 2022; Edin et al., 2023), and ICD-10 coding on MIMIC-IV has been proposed as a benchmark for extreme multi-label classification (Nguyen et al., 2023). We use the preprocessed ICD-10 release and splits of Edin et al. (2023), which their public code rebuilds from the PhysioNet files (MIMIC-IV and MIMIC-IV-Note v2.2), so anyone with access can recreate them: 122,279 summaries from 65,659 patients, with predefined train, validation and test splits of 89,099, 13,378 and 19,802 summaries and text that is lower-cased and stripped of punctuation. Its label universe of 7,942 codes mixes 5,651 ICD-10-CM diagnosis codes and 2,291 ICD-10-PCS procedure codes. We assign diagnoses only, so the inventory is the 5,651 diagnosis codes with their descriptions. No note text, reasoning-trace text or other protected content appears in the paper, its figures, the supplementary material or the code snapshot: we report public ICD-10-CM codes, attention-head names and statistics, and the one example code pair (N18.9 named after N18.1) is public.

Characterization corpus (TR-006). Notes come from the test split. A note is eligible if it has 800–2,500 words and 5–25 diagnosis codes in the inventory. Selection is stratified over the four quartiles of that code count and, within a stratum, greedy on ICD-10 blocks not yet covered, with at most one note per patient. The 94 notes cover 235 ICD-10 blocks and 21 of the 22 ICD-10-CM chapters, with 5–25 gold codes each (median 13). A note enters the prompt without its discharge-diagnosis section, which names diagnoses in words; six traces are pilot variants of six of the notes with that section kept, hence 100 traces over 94 notes. The trace is the unit of analysis, and we call each trace a note; the characterization on these traces is HD-001, and its replication on the retrieved shortlist is HD-002 (Appendix G). On these traces Qwen3-14B scores micro-F1 0.168; all 100 are valid (Appendix B), each several thousand tokens of reasoning followed by a JSON answer.

The full inventory in context. The prompt lists all 5,651 diagnosis codes, one line per code with its description, in a single order shuffled once with a fixed seed and shared by every note: the shared order makes the inventory a cacheable prefix, and the shuffle removes ICD order as a positional cue. The note and the instruction follow (Appendix B). A prompt is about 106k tokens (median over the 100 prompts), of which the inventory is about 105k, about 18.5 tokens per line. *Chapter* in this paper means a code's first character, the chapter letter; the inventory lines sharing the emitted letter are the emitted chapter, or region. The inventory uses 25 letters, which map imperfectly onto the 22 ICD-10-CM chapters: A–B, S–T and V–Y each form one chapter, and D and H are each split across two. Where a true ICD-10-CM chapter is meant, as in "21 of the 22 chapters", we say so.

The retrieved shortlist (distillation half). The students are distilled on a retrieved shortlist rather than the full inventory (Sec. 4.2), and the teacher is re-characterized on it (HD-002, Appendix G). The retriever (RT-003) embeds note chunks with a frozen gte-large encoder (Li et al., 2023). A supervised multi-label linear head, logistic regression on [max; mean]-pooled chunk embeddings trained on the `MIMIC-IV` train split with the patients of both corpora excluded, scores every code, and its ranking is fused with the zero-shot similarity between chunks and code descriptions. The top $K = 1{,}000$ codes are kept, and the category siblings (same three-character category) of the top 200 are added, so that the near-misses P2C events need stay in the list. The construction is deterministic and identical at training, validation and test. A pre-registered rule fixed K: the smallest K whose completed recall reaches 0.9 on both the validation notes and the distillation corpus. The corpus condition never bound: at every K of the sweep the corpus recalls more than the validation notes (at $K = 1{,}000$, 0.92 against 0.89 without completion, 0.939 against 0.903 with it), so the validation notes alone select the same K and the test notes' labels did not affect it. The sweep (code-level macro recall of the fused retriever on the validation notes and the 650-note corpus):

K / completion	validation	corpus
500 / none	0.80	0.84
1,000 / none	0.89	0.92
1,000 / siblings of the top 200 (chosen)	0.903	0.939

A zero-shot encoder alone recalls 0.45 of the gold codes at $K = 500$, hence the supervised stage. Macro recall over notes is 0.903 on the 100 validation notes, which are the characterization notes, and 0.936 on the 150 test notes; it is a ceiling every distilled arm shares. A shortlist has about 1,550 lines, and its prompts are 27–34k tokens (median 30k), about 3.5 times shorter than the full inventory's; header, line format, note and instruction are otherwise byte-identical. The teacher's traces on these prompts (TR-007 for the validation notes, TR-008 for the 650-note distillation corpus) score micro-F1 0.199 on the validation notes (the 99 with a valid trace) and 0.227 on the test notes.

B TEACHER, PROMPTS, DECODING; THE RETIRED INSTRUMENT

Supports: Sec. 2.1, Sec. 2.3; [R-C1], [R-C2]

Teacher and decoding. The teacher is `Qwen3-14B` (Yang et al., 2025) (`Qwen/Qwen3-14B`; revision in Appendix I.1), 40 layers of 40 attention heads, run in bf16 in its native thinking mode. Static YaRN rope scaling (Peng et al., 2024) with factor 4 extends its native context of 40,960 tokens to 163,840; the positional scaling is part of the characterized system. Traces are sampled at temperature 0.6, top-p 0.95 and top-k 20, with a fixed seed and a cap of 16,384 new tokens. The shortlist traces of Appendix G use the same settings.

Prompt. We describe the prompt rather than print it. It consists of a header naming the list as the assignable ICD-10-CM diagnosis codes, the candidate lines (the full inventory, or the shortlist), a header introducing the discharge summary, the note, and a short instruction, all in one user turn of the model's chat template with thinking enabled. The instruction gives the role of a clinical coder and the task of coding the summary; it points to the complete list above and asks the model to narrow first to the clinically plausible region and then decide among the specific codes in it; and it asks for the final answer as a JSON list of code strings. The instruction therefore invites a coarse-to-fine procedure.

The phases we measure are defined by the type of token-level decision (Section 2.2), not by this wording, and we make no claim about the teacher under other instructions.

Validity and parsing. A generation is valid when it closes its reasoning block before the cap. Its answer is the last JSON list after that block, restricted to inventory codes and de-duplicated; a truncated generation is invalid, is never scored, and its reasoning is never scraped for codes. A guard also flags degenerate generations: at least 10 codes of which more than 90% follow in ascending code order (nothing in the shuffled list suggests sorting), more than 50 codes, or a generation that reaches the cap. All 100 characterization traces are valid and none is degenerate. Sampling is seeded, but forward passes are not bit-exact on our hardware (Appendix C), so a regenerated trace can differ; every measurement re-scores the stored traces by teacher forcing.

The retired anchor-cosine instrument. Our first Phase-1 instrument asked whether a head's write points toward a direction fixed in advance. It scored a head by the cosine between its output at an event and an anchor, the centroid of the embeddings of the note's gold codes, weighted by how concentrated the head's attention is and averaged over events. Four runs on the same 100 traces (CS-001 to CS-004) retired it. (i) The anchor is not a well-formed object on real notes. Every note's gold codes span more than one ICD-10-CM chapter (median 8), and the score passes 6 of 100 notes against 5 expected by chance. A per-concept maximum over the gold codes passes 19 of 100 notes, but diffusely: 209 heads shift together and none recurs across notes. (ii) Every head has a stable, idiosyncratic alignment with any fixed direction, so tests on the cosine cannot be calibrated. Under a complete null, the per-note test (Benjamini–Hochberg at FDR 0.10 over heads) returned a non-empty set in 800 of 800 trials, and a head's null statistic kept its sign across a note's mentions 81.5% of the time. (iii) The positive control, an anchor on the code being emitted, passes 44 of 100 notes against a 200-draw null but 73 against the earlier 20-draw null: the size of the null matters. (iv) The one robust readout, recurrence of a head in the notes' top-20, found two heads: `L27H17`, in 42 of 100 notes, and `L11H10`. By chapter I_c they rank 128 and 364 of 1,600. Their recurrence is real as a correlation, but it is not about the decision. The replacement measures each head's effect on the model's own decision margin and validates every set by intervention (Section 2.3, Appendix C).

Identifiability of sets fitted to one note. A note has a median of 19 chapter decisions, one per first mention. CS-004 asked whether a head set can be fitted to a single note from that many events. Under the anchor-cosine statistic, sets fitted to one note (Benjamini–Hochberg at FDR 0.10 over 1,600 heads) were empty on the median note and did not reproduce between the two halves of a note's mentions (median Jaccard index 0.000), and the most recurrent head cleared the first Benjamini–Hochberg threshold on 2 of 100 notes: with about 19 events, a head must lie several standard errors from zero to survive a search over 1,600 heads. The causal instrument agrees where it can be checked. A ranking fitted to a note's own chapter events carries that note's chapter decision into another context less well than the global ranking at $k \geq 8$, whereas for character decisions, with about 70 events per note, the note's own ranking is the better carrier (Appendix D). We therefore use global sets and validate them on every note.

C THE INSTRUMENT IN DETAIL

Supports: Sec. 2.2, Sec. 2.3; [R-C1], [R-C2]

Figure 3 shows, on the code pair of Section 2.2, how a coding decision becomes a token-level event, what each event type is, and where the margin that scores it is read.

Event extraction. Codes are located in a trace by pattern and kept if they are inventory codes. `Qwen3-14B`'s tokenizer writes a code as a chapter-letter token followed by one token per character; each of the 25 letters and 36 characters that inventory codes use is a single token both with a leading space and without, and a letter token may carry the space. A mention is scored only if each of its characters is its own token in context, which excludes 67 of the 2,131 first mentions (median 19 per note). At every event, q is the input position whose output is the decided token, the target is the emitted token, and the competitors depend on the kind (Table 4):

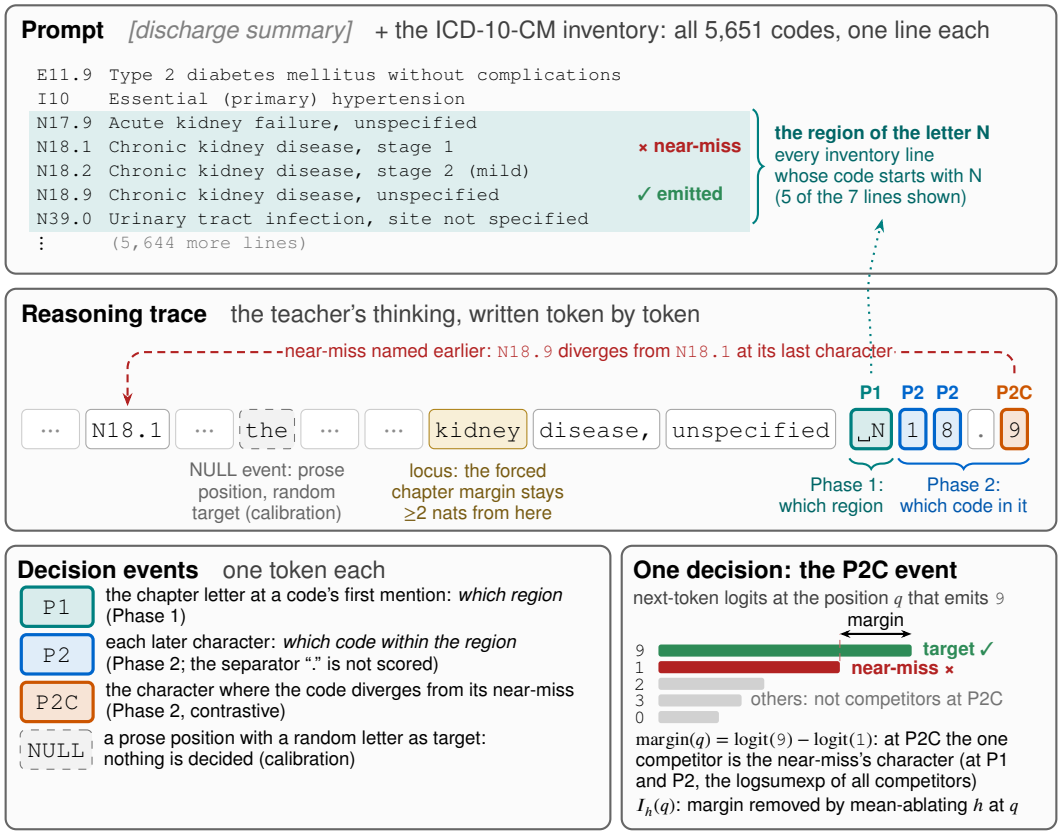


Figure 3: How a coding decision becomes a token-level event (illustrative: public ICD-10-CM codes and their inventory descriptions; the note and the reasoning prose are placeholders). The prompt holds the note and all 5,651 inventory lines; the region of a chapter letter is every line whose code starts with it. Writing `N18.9` after naming `N18.1`, the teacher decides one token at a time: the chapter letter (P1, Phase 1), each later character but the separator (P2, Phase 2), and the character at which the code diverges from its near-miss (P2C). NULL events, prose positions with a random target letter, calibrate the head rankings. Each event is scored by the model's own margin against its competitors (at P2C, the near-miss's character alone) at the position q that emits the token; a head's importance is the margin it removes when mean-ablated there (Sec. 2.3). The locus (Appendix F) is the earliest position from which the forced chapter margin stays at least 2 nats. [R-C1]

- P1: the chapter letter at the first mention of each inventory code; the competitors are the other inventory letters in the same spacing variant. 30% of first mentions are siblings of an earlier mention.
- P2: each later character of a first mention (the separator "." is not scored); the competitors are the characters that continue the prefix to another inventory code, and a character without any is skipped.
- P2C: at a mention of a code whose sibling (same three-character category) the trace named earlier, the first character at which the two codes differ. The single competitor is the sibling's character, so the margin is the log-odds against the near-miss. The event records the positions of the earlier mentions of the code (S) and of the near-miss (N). Events are capped at two per (code, near-miss) pair, the first mention and the first repeat that has earlier mentions of the code: uncapped, repeats gave 9,474 events (1,098 in one note), repeats of the same decision rather than new ones; capped, 3,855, of which 2,208 have earlier mentions of the code.
- NULL: 20 positions per trace drawn at random from the prose at least 8 tokens from any code token, each with a random letter as the target and the other letters as competitors.

The faithfulness pass uses every P1 event and a seeded subset of at most 50 P2 and 40 P2C events per note; 92 notes have a P2C event.

Table 4: Decision events over the 100 characterization traces: all events, events in the faithfulness pass (with the number of notes that have them), and the median decision margin (nats). P1 and P2 events are taken at the first mention of each inventory code, P2C events at the first mention and the first repeat of each (code, near-miss) pair. [R-C1]

Event	Decision	Events (100 traces)	Faithfulness pass (notes)	Median margin
P1	chapter letter (region of the label space)	2,064	2,064 (100)	+6.02
P2	each character (code within the prefix)	7,688	4,587 (100)	+2.76
P2C	character diverging from a named sibling	3,855	1,959 (92)	+3.38
NULL	prose position, random target (calibration)	2,000	—	−4.95

Decision margin, and the audit at the code token. $\text{margin}(q) = \text{logit}(\text{target}) - \text{logsumexp}(\text{competitors})$ is the log-odds the model gives the emitted token against its alternatives, a multi-way form of the usual patching metric (Wang et al., 2023a; Zhang & Nanda, 2024); taken over all other letters, the P1 margin is the log-odds of this chapter against any other. The entropy over the target and its competitors says how close the call was. Before any attribution, an audit read the margins at every event (Table 5). At the code token the chapter margin is usually already decisive; character decisions are far more often live, and some are sampled against the model's favourite; at NULL events a random letter sits far below zero at high entropy, as it should.

Table 5: Decision margins at the events of all 100 traces (nats): 10th percentile, median and 90th percentile; median entropy over the target and its competitors; share of events with a margin below 5 nats and below zero. [R-C1]

Event	p10	median	p90	entropy	margin < 5	margin < 0
P1	+0.74	+6.02	+10.85	0.02	42%	6%
P2	−0.49	+2.76	+9.19	0.26	68%	15%
P2C	0.00	+3.38	+8.00	0.15	66%	10%
NULL	−10.44	−4.95	−2.68	1.80	100%	99%

The hook and mean-ablation. Each attention layer's output projection takes the concatenation of its 40 heads' outputs. One forward pre-hook on that input can (i) accumulate each head's mean output, (ii) expose every head's output as a gradient leaf, (iii) replace one head's slice at chosen positions by its mean (mean-ablation), or (iv) replace it by values recorded in another run (knock-in). The unit of intervention is one head's output at one position. The mean μ_h is taken over every position of the note's own prompt and trace, during the single prefill that also builds the key–value cache; replacing

a head's output by its mean removes the head's information while keeping its typical contribution to the residual stream, whereas zeroing would move the model off its distribution (Wang et al., 2023a; Heimersheim & Nanda, 2024). Events are visited in decreasing position, so the cache is only ever cropped, and each measurement is a single-token forward, or a forward and a backward at one position. Before any run the hook was validated: with no intervention active the hooked forward is byte-identical to a plain one; zeroing all heads of a layer through the hook equals zeroing that layer's attention output; on a small randomly initialized model with the same code path, knocking in all heads reproduces the clean margin at every event; and in the locus pass the gradient-enabled forward reproduced the plain forward's margins exactly. $I_h(q) = \text{margin}(q) - \text{margin}(q \mid h \text{ mean-ablated})$ is positive when the head supports the emitted choice and negative when it opposes it. A set's necessity is the share of the clean margin destroyed when the set is mean-ablated at q (above 1, the decision flips).

Attribution as a screen, and its fidelity. First-order attribution estimates the effect of mean-ablating head h as $A_h = (\mu_h - z_h) \cdot \partial\,\text{margin}/\partial z_h$ for all 1,600 heads from one backward pass (Syed et al., 2024; Kramár et al., 2024). We measured its fidelity on the task before relying on it: at ten chapter decisions of one note, every head was also patched exactly (Table 6). The hook of the attribution pass detached the gradient leaf from the graph at every layer. That cut every path through later attention layers (a head's effect through later queries, keys and values) and kept only each head's direct path. Under that hook, attribution correlates 0.36 with the exact effect over all heads and 0.47 over heads above the noise floor, and its top-20 recovers 48% of the exact top-20's effect: a screen, not a measurement. The defect surfaced when the window attribution of Appendix F came back zero away from the last position, and it was fixed before that pass. The same self-test with the graph intact gives 0.84 (0.85 above noise), Spearman 0.51, 16 of the exact top-20, and 92% of its effect. Integrated gradients is worse under both hooks (0.63 with the graph intact), presumably because interpolating every layer at once moves the model off-distribution. Direct-logit attribution (Elhage et al., 2021) is uncorrelated with the exact effect in the self-test and anticorrelated with first-order attribution in the full run (median per-event Pearson −0.32): these heads act through later layers, not through their direct path.

What each ranking rests on: P1 and NULL importances are exact for the nominated heads (next paragraph) and first-order elsewhere; P2 importances are first-order under the detached hook (a re-rank with the intact graph was not run); the P2C global set is first-order under the detached hook and was refined by exact patching of 56 nominated heads at every P2C event of the faithfulness pass, where first-order and exact effects correlate at a per-event median of 0.72 (Appendix D); the prose-time ranking is first-order under the intact hook (Appendix F). Every set the paper reports is judged by exact intervention.

Table 6: Fidelity of attribution against exact patching: one note, ten chapter decisions, all 1,600 heads patched exactly at each. *Detached*: the hook of the attribution pass, which kept only each head's direct path; *intact*: the corrected hook, used at the locus. Integrated gradients: 5 steps. [R-C2]

	First-order		Integrated gradients	Direct-logit
	detached	intact	detached	attribution
Pearson with the exact effect, all heads	0.36	0.84	0.22	≈ 0 (−0.11 to +0.07)
Pearson, heads above the noise floor	0.47	0.85	0.25	—
Exact top-20 effect recovered by the top-20	48%	92%	29%	—

Exact patching of the nominated heads. At every P1 event and at a seeded half of the NULL events, the attribution pass mean-ablates exactly the top-40 heads by $|A_h|$ together with the running global top-16 (the heads with the largest mean first-order importance over the chapter events of the notes already processed): 48 exact patches per chapter event on average, and 26 per NULL event averaged over all of them. A head's importance is its exact effect where measured and the first-order estimate elsewhere. P2 and P2C events received no exact patches in this pass; the faithfulness pass later patched the same 56 nominated heads exactly at every P2C event of its subset.

Noise floor. The forward pass is not deterministic on this hardware: six identical single-token forwards give a margin standard deviation of 0.046 nats (range 0.09). Every exact effect is a difference

of two such forwards, so single-head effects below about three standard deviations (0.14 nats) cannot be told apart from zero. In the self-test, between 71 and 397 heads per event exceed that floor: a decision is influenced by many heads, most of them weakly. The paper therefore reports ranked sets and means over notes, never single-head effects near the floor.

Knock-in (sufficiency by denoising). Ablation shows that a set is necessary; sufficiency needs a context in which the model makes a different decision (Wang et al., 2023a; Heimersheim & Nanda, 2024). For each note, the corrupted run is the next note in corpus order (cyclically), a different patient, at a matched event: the same kind of event with a different decision (a different chapter letter at P1, a different character at P2 and P2C), preferring the same input token at q, then the same character depth, then the later position. The donor's input token at the matched position is replaced by the clean note's token at q, the clean run's per-head outputs at q are written in for the heads of the set, and the clean decision's margin is read. The donor event shared the input token at q for 55%, 75% and 95% of P1, P2 and P2C events, and the character depth for 100%, 87% and 85%; since the clean token is always the one run, the ceiling below is exact either way. Writing in all 1,600 heads reproduces the clean residual stream at q, so the MLPs see exactly their clean inputs and the all-heads knock-in must equal the clean margin. It does, to the noise floor, on every event of every kind (median difference 0.000 nats; 10th–90th percentile 0.00 to 0.00). This checks the wiring and sets the scale,

$$\text{recovery}(S) = \frac{\text{margin}_S - \text{margin}_{\text{corr}}}{\text{margin}_{\text{all}} - \text{margin}_{\text{corr}}},$$

which is 0 when the other patient's decision stands and 1 when the set alone recovers the clean one. Events whose range $\text{margin}_{\text{all}} - \text{margin}_{\text{corr}}$ is under 1 nat are excluded, because the corruption produced no different decision there: 0.5% of chapter, 9% of character and 29% of contrastive events. The corrupted context favours another chapter by a median of 8.7 nats (it is against the clean chapter on 99% of events) and another character by 3.6 nats (87%). At P2C events it is close to indifferent (0.4 nats; against the clean character on 54%), so contrastive knock-in measures how far a set moves an undecided context, over a median range of 3.1 nats.

NULL events and calibration. A ranking of 1,600 heads always has a top; the question is whether the top reflects the decision or the procedure. NULL events run the identical procedure where nothing is decided. Raw importance fails this check: Benjamini–Hochberg at FDR 0.05 declares 309 heads positive and 282 negative at NULL events, and the null top-16 effects average 0.16 nats against 0.26 at chapter events. Some heads have a generic effect on the letter logits, sharpening or flattening the distribution, and a generic effect moves any letter margin at any position, consistently. Decision-specific importance $I_c = I - I_{NULL}$ subtracts it per note: a head's mean importance over the note's events of a kind minus its mean over the note's NULL events, then averaged over notes. The same check on I_c, one half of the NULL events against the other, gives 0 discoveries in either direction. The generic heads and the decision heads are different heads: the raw and null rankings are unrelated (Spearman -0.24; they share 0 of their top-16 and 1 of their top-40), and the raw top-10 keep their ranks under I_c. For the chapter decision, 429 heads have $I_c > 0$ and 546 have $I_c < 0$ at FDR 0.05: consistent effects in both directions, not a sparse set; sparsity is a question of magnitude (Appendix D). A phase's ranking by I_c is its global set, and its pooled weights are $w_h = \text{softmax}(I_c/\tau)$ over the top-k, with τ the standard deviation of the top-k importances.

Statistics. The unit of every test is the note, because events within a note are not independent. Per note, a head's importance for a kind is its mean over the note's events of that kind; its global importance is the mean over notes.

- *Bootstrap intervals* resample notes 1,000 times and report percentile 95% intervals (Efron, 1979); the intervals of the I_c values tabulated in Appendices D and F exclude zero.
- A *one-sample t-test* across the 100 per-note means gives each head's p-value. Bootstrap percentile p-values were too coarse to feed a multiple-comparison correction: in a synthetic check on pure noise they gave 13 discoveries.
- *Benjamini–Hochberg* (Benjamini & Hochberg, 1995) controls the expected share of false discoveries among the heads it declares; under a complete null it should declare essentially none, which is the calibration check above.

- *Sign tests* across notes compare paired sets (global against random, global against the note's own ranking). They count wins and assume nothing about effect sizes; every "notes" count in the paper is such a count.
- *Recurrence* uses a binomial test per head with Bonferroni correction over the 1,600 heads (Appendix D).
- *Medians over notes of per-note means* are reported for necessity, a ratio to the clean margin that is unstable where that margin is near zero. Recovery is normalised per event by its ceiling and has no such problem.

Compute. Generating the 100 traces took 6.4 GPU-hours on H200s. A note's key–value cache is about 17 GiB at 110k tokens, and a single-token forward against it takes 72 ms on an H200 and about 100 ms on an RTX Pro 6000. The five passes over the corpus took: attribution, 6.3 GPU-hours (seven RTX Pro 6000 lanes, a median of 176 s per note); faithfulness, about 12 GPU-hours on B200 and RTX Pro 6000 GPUs (about 7,000 single-token forwards per note); read side, about 1 GPU-hour (one prefill with query and key capture per note); prose locus, about 9.3 GPU-hours on B200s only, because the multi-position backward keeps every layer's expanded keys and values (130 GiB at 111k tokens); and the fixed prose-time set, about 3 GPU-hours. Each pass ran in under five hours of wall time on two to seven lanes; HD-002 repeated all five on the shortlist prompts (Appendix G).

D FULL HEAD TABLES

Supports: Sec. 3.1, Sec. 3.2; [R-C3], [R-C4]

The global sets at every k. Table 7 is Table 1 at every set size tested, $k \in \{3, 8, 16, 40, 100\}$ (the main text prints four of the five).

Table 7: Global per-phase sets at every k (Table 1 with the $k = 16$ rows): necessity (share of the clean margin destroyed by mean-ablation; > 1: flipped) and knock-in recovery in another patient's context, against count-matched random and bottom-k sets. Medians over notes of per-note means; "notes": notes where the global set beats random (of 100; P2C: of 92). [R-C3]

	P1 (chapter)				P2 (character)				P2C (contrastive)			
k	global	random	bottom	notes	global	random	bottom	notes	global	random	bottom	notes
Necessity: share of the clean margin destroyed by mean-ablating the set (> 1: decision flipped)												
3	0.24	0.00	−0.15	92	0.41	0.01	−0.04	95	0.26	0.00	−0.06	90
8	0.62	0.00	−0.36	95	0.78	0.00	−0.18	98	0.63	0.00	−0.06	90
16	0.88	0.02	−0.05	97	1.10	0.06	0.11	97	0.72	−0.01	−0.12	90
40	1.24	0.04	0.89	100	1.61	0.06	0.76	97	0.86	0.04	−0.50	90
100	1.60	0.09	1.57	100	2.21	0.26	1.41	99	0.89	0.07	−0.16	90
Sufficiency by knock-in: share of the clean decision restored in another patient's context												
3	0.16	0.00	−0.16	100	0.24	0.00	−0.45	100	0.05	0.00	−0.33	80
8	0.67	0.00	−0.28	100	0.48	0.01	−0.54	100	0.69	0.00	−0.54	92
16	0.76	0.02	−0.48	100	0.64	0.01	−0.63	100	0.73	0.00	−0.56	92
40	0.86	0.04	−0.30	100	0.79	0.02	−0.81	100	0.85	0.02	−0.63	92
100	0.96	0.09	−0.33	100	0.89	0.05	−0.80	100	0.94	0.05	−0.78	92

Per-phase importances. Table 8 gives decision-specific importance for representative heads, and Table 9 completes it to every head in the chapter top-13, the character top-10 or the contrastive top-9. Importances are means over notes of per-note I_c in nats, with ranks among 1,600 heads; every listed I_c has a 95% bootstrap interval that excludes zero. Chapter importances are exact for the nominated heads, and character and contrastive importances are first-order (Appendix C). Profiles describe the write side: *shared* heads rank high in all three phases, while *chapter-*, *character-* and *contrastive-specific* heads rank high in one. Over all 1,600 heads, I_c correlates 0.75 between chapter and character decisions, 0.63 between chapter and contrastive decisions and 0.80 between character and contrastive decisions, and the top-40s share 25, 24 and 26 heads. The structure is a shared core that writes the code string (`L39H38`) with phase-specific heads around it: heads in layers 26–28

for the chapter, heads in layers 36–39 for the characters, and L30H11, ranked 1,500 on the chapter decision but 2 on the contrastive one, where the decision is against a previously named near-miss. Under Benjamini–Hochberg at FDR 0.05, 429, 543 and 340 heads have $I_c > 0$ for chapter, character and contrastive decisions: consistent small effects are widespread, and what makes a set is magnitude (see Sparsity below).

Table 8: Decision-specific importance I_c (nats; mean over notes; rank of 1,600) of representative heads, with write-side profiles and the layers of the chapter top-40. P1 importances are exact for the nominated heads, P2 and P2C importances first-order; the last column is the exact P2C refinement in raw importance I, not comparable with I_c. [R-C4]

	I_c (nats; rank of 1,600)			raw I	
Head	P1	P2	P2C	P2C exact	profile
L39H38	0.91 (1)	0.97 (1)	0.62 (1)	0.22 (4)	shared (emission)
L26H36	0.43 (3)	0.08 (29)	0.04 (28)	0.03 (14)	chapter
L26H37	0.29 (6)	0.10 (21)	0.05 (22)	0.04 (9)	chapter
L27H11	0.24 (7)	0.06 (44)	0.04 (32)	0.03 (17)	chapter
L27H13	0.21 (8)	0.08 (34)	0.03 (37)	0.01 (42)	chapter
L39H10	0.18 (13)	0.42 (2)	0.24 (7)	0.36 (2)	shared, character-heavy
L39H37	0.05 (72)	0.29 (4)	0.07 (18)	0.02 (29)	character
L39H12	−0.07 (1,580)	0.27 (8)	0.24 (6)	0.64 (1)	character
L38H22	−0.06 (1,572)	0.23 (10)	0.22 (8)	0.00 (1,517)	character
L36H10	0.06 (59)	0.24 (9)	0.02 (64)	0.02 (35)	character
L30H11	−0.02 (1,500)	−0.08 (1,588)	0.51 (2)	0.28 (3)	contrastive
L38H39	−0.02 (1,485)	−0.04 (1,565)	0.19 (9)	0.12 (7)	contrastive
P1 top-40 by layer: 0–11: 0 (0%)	12–23: 2 (2%)	24–31: 24 (53%)	32–39: 14 (45%)		

Table 9: Heads in the chapter top-13, the character top-10 or the contrastive top-9 that are not in Table 8: decision-specific importance I_c (nats; mean over notes; rank of 1,600) and write-side profile. [R-C4]

Head	P1	P2	P2C	profile
L39H13	0.47 (2)	0.12 (17)	0.02 (68)	chapter, some character
L38H24	0.39 (4)	0.27 (7)	0.27 (5)	shared
L28H21	0.36 (5)	0.27 (6)	0.11 (14)	shared
L35H18	0.21 (9)	0.13 (16)	0.14 (11)	shared
L39H29	0.19 (10)	0.31 (3)	0.31 (3)	shared, character-heavy
L37H6	0.19 (11)	0.28 (5)	0.28 (4)	shared
L26H18	0.18 (12)	0.06 (47)	0.02 (74)	chapter

The exact contrastive refinement. The contrastive global set, the one tested in Table 1 (and, measured on the shortlist, the one pooled for distillation), is the first-order ranking. Exact mean-ablation of the 56 heads nominated at every P2C event of the faithfulness pass (the event's top-40 by $|A_h|$ and the global top-16) gives the exact-merged ranking, exact where measured and first-order elsewhere (Table 10). It is in raw importance I, not NULL-calibrated, so it is not comparable with I_c. First-order attribution had under-estimated L39H12 and over-estimated L30H11. The exact-merged and first-order top-40s share 33 heads, every head of the exact top-8 is in both, and 313 heads have a positive exact-merged importance under Benjamini–Hochberg at FDR 0.05. The contrastive reading stands: L30H11 and L38H39 remain contrastive-specific, and two final-layer character heads, L39H12 and L39H10, carry the largest exact effects at the divergence character. This is why the first-order contrastive top-3 recovers only 0.05 on knock-in while the top-8 recovers 0.69: the top of the contrastive set is broader than the three heads that first-order attribution put first.

Layers. At the code token the chapter top-40 has 0 heads in layers 0–11 and 2 in layers 12–23; it sits in the middle and late layers, with 24 heads in layers 24–31 carrying 53% of its importance and 14 in layers 32–39 carrying 45% (Table 8, last row). At the prose locus the set that decides the chapter sits one block earlier (Appendix F).

Table 10: The contrastive ranking after exact refinement, in exact-merged order: raw importance I (nats; mean over the 92 notes with contrastive events; exact where measured) with its 95% bootstrap interval, the first-order raw importance, and the head's rank in the first-order I_c ranking that defines the contrastive global set. Raw I is not NULL-calibrated and not comparable with I_c. [R-C4]

Head	I exact-merged	95% interval	I first-order	rank by first-order I_c
L39H12	0.64	[0.54, 0.75]	0.19	6
L39H10	0.36	[0.29, 0.42]	0.23	7
L30H11	0.28	[0.25, 0.32]	0.53	2
L39H38	0.22	[0.20, 0.25]	0.25	1
L39H29	0.16	[0.14, 0.19]	0.16	3
L39H25	0.14	[0.13, 0.16]	0.14	10
L38H39	0.12	[0.11, 0.14]	0.21	9
L28H21	0.05	[0.04, 0.06]	0.06	14

Pooled weights at the operating point. The pooled weights $w_h = \mathrm{softmax}(I_c/\tau)$ over a phase's top-k concentrate on the leader. At $k = 3$ the leader gets 0.81 of the weight for chapter decisions and 0.66 for contrastive ones, and at the top-40 operating point L39H38 still holds 0.742, 0.813 and 0.444 of it (P1, P2, P2C): HD-001's pooled top-40 is close to one head. The distillation half pools the shortlist's sets (HD-002, Appendix G), whose weights are spread; their leaders are L27H11 (0.254), L39H10 (0.303) and L39H29 (0.215).

Necessity and knock-in in full: the global set against the note's own ranking. Table 1 tests the global set against random and bottom-k controls; Table 11 adds each note's own ranking, fitted to that note's own events. For the chapter decision the two are level on necessity (0.62 against 0.63 at $k = 8$; the global set destroys at least as much as the note's own on between 36 and 52 of 100 notes for $k \leq 40$, and on 72 at $k = 100$), and the global set is the better carrier on knock-in at every $k \geq 8$ (0.67 against 0.45 at $k = 8$). A ranking fitted to a note's own chapter events (a median of 19) plausibly picks up heads that were necessary on those events by happenstance, whereas the cross-note ranking picks the heads that carry the decision as such. For characters the note's own ranking is the better carrier at every k and removes more up to $k = 40$ (at $k = 8$, 0.59 against 0.48 on knock-in and 1.10 against 0.78 on necessity), as expected of a ranking fitted to about 70 events and tested on them; for contrastive events it is better on necessity at every k and on knock-in at every k except 8. At the top-40 operating point the global set's knock-in is within 0.04 of the note's own for characters and for contrastive events.

Table 11: The global set against each note's own ranking (fitted to that note's events), at every k: necessity (share of the clean margin destroyed; > 1: flipped) and knock-in recovery in another patient's context. Medians over notes of per-note means (P2C: 92 notes). Random and bottom-k controls are in Table 1. [R-C3]

	Necessity						Knock-in recovery					
	P1		P2		P2C		P1		P2		P2C	
k	global	own	global	own	global	own	global	own	global	own	global	own
3	0.24	0.30	0.41	0.55	0.26	0.45	0.16	0.18	0.24	0.33	0.05	0.35
8	0.62	0.63	0.78	1.10	0.63	0.75	0.67	0.45	0.48	0.59	0.69	0.63
16	0.88	0.91	1.10	1.38	0.72	0.84	0.76	0.61	0.64	0.69	0.73	0.76
40	1.24	1.26	1.61	1.81	0.86	0.97	0.86	0.78	0.79	0.82	0.85	0.87
100	1.60	1.45	2.21	2.14	0.89	1.00	0.96	0.89	0.89	0.91	0.94	0.98

Spread, and the jump from three to eight heads. Across notes the chapter top-8 destroys between 0.30 (10th percentile) and 1.79 (90th) of the margin. The ratio is unstable on the few notes whose clean margins sit near zero, which is why medians are reported, and the tail is heavier for characters, whose margins are more often near or below zero (Table 5). Knock-in, normalised by each event's ceiling, is tighter: across notes the chapter top-8 recovers between 0.56 and 0.74 (10th to 90th percentile), and more than half on 99 of 100 notes. Recovery jumps between three and eight heads,

from 0.16 to 0.67, and rises to 0.86 at 40 and 0.96 at 100: heads four to eight of the chapter ranking (L38H24, L28H21, L26H37, L27H11, L27H13) act jointly with the top three. For characters and contrastive decisions the top-8 recovers 0.48 and 0.69.

The bottom-k oppose the decision. Removing the eight most negatively ranked chapter heads raises the chapter margin (necessity −0.36), and writing the bottom sets' clean outputs into the corrupted run moves it further from the clean decision (recovery between −0.16 and −0.81; Table 1). These heads argue for the alternative, which is why the negative discoveries of the calibration (Appendix C) are effects rather than noise. At $k = 40$ and 100 the chapter bottom set is large enough to include heads with large generic effects, and removing it destroys the margin too (0.89 and 1.57).

Isolation is not a sufficiency test. The attribution pass first tested sufficiency by isolation: keep only the top-k of each note's own ranking and mean-ablate every other head at q. The margin that remains is about −0.6 of the clean margin at every k, so the decision flips, while the same pass's necessity bins read 0.29, 0.62, 0.88, 1.22 and 1.42 at $k = 3$ to 100, with the top-k above random on 96 to 100 of 100 notes, the same picture as Table 11. Mean-ablating some 1,560 heads removes the residual stream's base, not only the heads that argue for the competitors, so isolation asks whether the set works without the rest of the model, which is not the question. The set is necessary on top of a broad base. Knock-in, which holds everything outside the set at the corrupted run's values, is the sufficiency test the paper uses.

Recurrence. If each note's top-20 were a random draw from 1,600 heads, a given head would appear in it with probability 20/1,600, so the number of notes in which it appears would be binomial over 100 notes; a head is called recurrent if its count survives Bonferroni correction over all 1,600 heads. Using each note's top-20 by raw importance I (not NULL-calibrated), 41 heads are recurrent, against 1.25 notes expected per head by chance. L39H38 and L39H13 are in the top-20 of every note (100 and 100 of 100), followed by L26H36 (95), L26H37 (94), L28H21 (84), L26H18 (78), L27H13 (75), L27H11 (74), L24H1 (71) and L39H10 (68).

Sparsity. Counting chapter heads above the calibration noise floor (the 95th percentile of the absolute split-half NULL difference, 0.004 nats), 289 heads have I_c above the floor and 92 of them carry 80% of the decision-specific importance. Per note, counting raw I above the note's own NULL floor, a median of 64 heads carry 80%. The set is neither three heads nor diffuse. Importance is not additive under ablation, so forty heads that flip the decision when removed, forty heads that recover 0.86 of it when knocked in, and 92 heads that carry 80% of the summed importance are one picture seen from three sides.

E READ SIDE

Supports: Sec. 3.3; [R-C5]

Panel, events and position categories. The read side reads the query–key half of the same heads whose outputs the write side measured (Elhage et al., 2021). A fixed panel of 68 heads (the global top-16 of each phase, the global bottom-16 of the chapter and contrastive rankings, and 16 random heads drawn with a fixed seed) is read at every P1 event and at the P2 and P2C events of the faithfulness pass. One prefill per note with query and key capture gives each panel head's causal attention row at q, which is split over 15 categories of key position: the attention sink (the first 8 tokens); the chat template; in the inventory, the emitted code's own line, the other lines of its three-character category, the other lines of its chapter, other chapters' lines, the near-miss's line, and inventory tokens outside any line; the note; the instruction; and in the trace, earlier mentions of the code (S), earlier mentions of the near-miss (N), mentions of other codes, the characters of the current mention written so far (the code prefix), and the remaining prose.

Definitions. The region is the emitted chapter's inventory lines. *On region* is the share of a head's total attention that falls on the region. *Region lift* divides the share of the head's inventory attention that falls on the region by the region's share of inventory tokens, so 1 means that the head reads the inventory indifferently. n_{eff} is the effective number of inventory tokens the head attends, the inverse

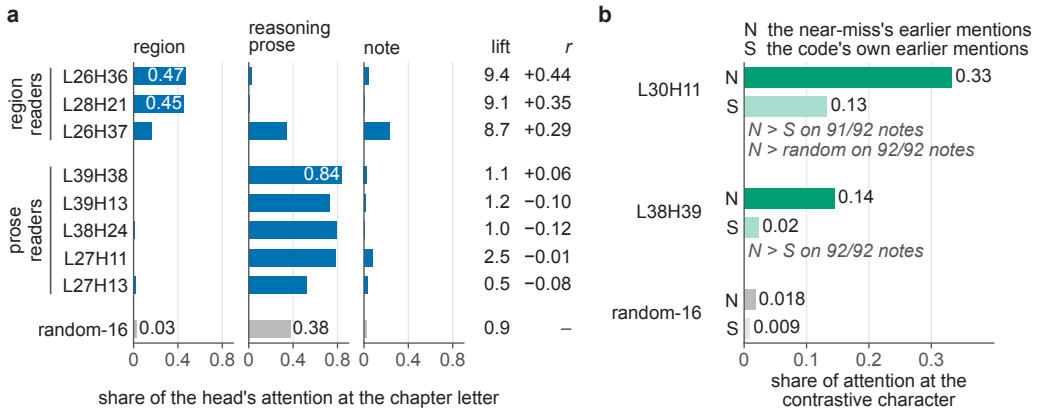


Figure 4: What the causal heads read (`Qwen3-14B`; attention at the event, medians over notes). (a) At chapter-letter events, the share of attention the global P1 top-8 put on the emitted chapter's inventory lines (region), on the reasoning so far (prose) and on the note, with 16 random heads as reference; right, each head's region lift and r, the correlation across events between its region share and its importance. Region readers read the region, prose readers the reasoning. (b) At contrastive events, the share on the earlier mentions of the near-miss (N) and of the code itself (S) for the two contrastive-specific heads and 16 random heads. Descriptive: attention was not intervened on. [R-C5]

participation ratio $1/\sum_j p_j^2$ of its attention renormalised over inventory tokens; an inventory line is about 18 tokens, so an n_{eff} of a few dozen means a few lines. r is the Pearson correlation, across a head's events, between its share on the region (on N at P2C) and its measured importance at that event. Values are medians over notes of per-note means, and group comparisons are sign tests across notes.

Group level. The global top-16 of every phase reads the emitted chapter's lines selectively, far above the random panel, on every note (Table 12). The bottom-16 read them almost as selectively: region selectivity is a property of heads that matter, with either sign, so what a head reads does not say which way it writes; the write side (Appendix D) does. Across the chapter top-16, Spearman's correlation between lift and I_c is 0.00, because the most important head, `L39H38`, is a prose reader.

Table 12: Read side by phase: region lift of the global top-16, 16 random heads and the global bottom-16 (P1 top-16 with its 10th–90th percentile over notes); notes on which the top-16's lift beats the random panel's; and the top-16's attention on the sink, the inventory's code lines, the note, the trace prose and the code prefix (random panel in brackets). Medians over notes. [R-C5]

	Region lift				Top-16 attention on				
Kind	top-16	random	bottom	top > random	sink	inventory lines	note	prose	code prefix
P1	3.54 (2.7–4.9)	0.87	3.11	100/100	0.10	0.28	0.07	0.38	—
P2	4.08	1.32	—	100/100	0.11	0.37	0.02	0.16	0.251 (0.135)
P2C	3.94	1.89	3.61	92/92	0.07	0.32	0.02	0.14	0.254 (0.151)

Chapter decisions: two families. Table 13 gives the chapter top-16 head by head; the random panel puts 0.03 of its attention on the region. Five heads in layers 24–28 are *region readers* (`L26H36`, `L28H21`, `L26H37`, `L26H18`, `L24H1`). `L26H36` and `L28H21` put 0.47 and 0.45 of their attention on the emitted chapter's lines (lift 9.4 and 9.1) over an effective few lines, and `L26H37` and `L26H18` split theirs between the region and the note. For the region readers, reading the region is what makes them matter: the more a region reader reads the region at an event, the larger its effect on that decision ($r = +0.44$ for `L26H36`, +0.35 for `L28H21`, +0.29 for `L26H37` and +0.22 for `L26H18`). The *prose readers* (the late-layer `L39H38`, `L39H13` and `L38H24`, and `L27H11`, which is chapter-specific on the write side) put 0.73 to 0.84 of their attention on the trace prose, and their effect does not depend on where in the prose they look ($r \approx 0$). For the mentions decided in the reasoning, they carry into the

output a chapter that is already elicitable from the prose (Appendix F). The top three heads already contain both families (the prose readers L39H38 and L39H13, the region reader L26H36); heads four to eight, which make the jump in knock-in recovery from three to eight heads (Appendix D), add two region readers (L28H21, L26H37) and three prose readers (L38H24, L27H11, L27H13). The boundary between the families and the "mixed" heads is a judgement call; the table gives the numbers.

Table 13: Chapter decisions: the global P1 top-16 at P1 events. I_c (nats); share of attention on the region (the emitted chapter's inventory lines), on the trace prose and on the sink; region lift; n_{eff} over inventory tokens; r between region share and importance across the head's events; family. Medians over notes. Descriptive: attention was not intervened on. [R-C5]

Head	I_c	region	prose	sink	lift	n_{eff}	r	family
L26H36	0.43	0.47	0.03	0.00	9.4	23	+0.44	region reader
L28H21	0.36	0.45	0.01	0.00	9.1	44	+0.35	region reader
L26H37	0.29	0.17	0.34	0.00	8.7	56	+0.29	region reader; note 0.23
L26H18	0.18	0.06	0.48	0.00	6.1	60	+0.22	region reader; note 0.25
L24H1	0.17	0.16	0.30	0.00	5.0	38	+0.11	region reader
L25H28	0.17	0.09	0.29	0.05	3.2	70	+0.32	mixed
L39H38	0.91	0.00	0.84	0.00	1.1	727	+0.06	prose reader
L39H13	0.47	0.00	0.73	0.17	1.2	1,949	−0.10	prose reader
L38H24	0.39	0.01	0.79	0.02	1.0	32	−0.12	prose reader
L27H11	0.24	0.00	0.78	0.00	2.5	54	−0.01	prose reader
L27H13	0.21	0.02	0.52	0.03	0.5	24	−0.08	prose reader; other chapters 0.32
L39H29	0.19	0.01	0.52	0.35	0.9	355	−0.03	prose reader
L35H18	0.21	0.08	0.05	0.17	2.2	33	+0.11	inventory, diffuse; other chapters 0.56
L37H6	0.19	0.07	0.05	0.14	1.9	107	+0.05	inventory, diffuse
L39H10	0.18	0.01	0.20	0.60	1.1	1,468	−0.09	sink
L26H33	0.17	0.01	0.12	0.00	1.6	227	−0.09	—

Character decisions. The character-specific heads read the code being written (Table 14): seven heads of the character top-16 put between 0.28 and 0.74 of their attention on the current mention's already-written characters, against 0.135 for the random panel as a group, a copy-and-continue mechanism over the code string. A second group reads the category's lines in the inventory, putting 0.10 to 0.24 of its attention on the emitted code's category siblings at lifts of 9 to 14: refinement within the category, on the read side.

Table 14: Character decisions: heads of the character global top-16 at P2 events. Left, share of attention on the current mention's already-written characters (the code prefix), with the random and top-16 panels as groups; right, share on the emitted code's category siblings in the inventory, and region lift. Medians over notes. [R-C5]

Code-prefix readers		Category readers		
Head	on prefix	Head	on category	lift
L39H25	0.74	L28H21	0.24	10
L39H10	0.61	L38H35	0.20	14
L39H29	0.60	L30H13	0.17	13
L36H10	0.55	L38H22	0.10	9
L39H36	0.39			
L39H37	0.35			
L39H12	0.28			
top-16 (group)	0.251			
random-16 (group)	0.135			

Contrastive decisions: the near-miss loop-back. At the character where a code diverges from the near-miss, the contrastive-specific head L30H11 puts 0.33 of its attention on the earlier mentions of the near-miss (N; 10th–90th percentile over notes 0.20 to 0.45), against 0.02 for the random panel

(Table 15). It is above the random panel on 92 of 92 notes and above its own share on the code's earlier mentions (S) on 91 of 92, and its measured effect grows with its share on N ($r = +0.30$). `L38H39`, the other contrastive-specific head, also reads N more than S, on 92 of 92 notes. The head re-reads the sibling named earlier. As a group, the contrastive top-16 put 0.074 of their attention on N and 0.254 on the code prefix (random panel 0.151). A query–key score, the similarity of a head's query at the event to its keys at S minus that to its keys at N, computed at the 2,208 contrastive events with earlier mentions of the code, does not separate heads: it tops out at 0.02, with a flat block of layer-34 and layer-35 heads at near-identical values, and since `L30H11` attends to N more than to S, such a score would rank the loop-back head negatively. The near-miss loop-back is therefore measured as attention mass on N, as M3 is in the distillation half.

Table 15: Contrastive decisions: attention on the earlier mentions of the near-miss (N) and of the code itself (S) at P2C events (medians over the 92 notes; random and bottom panels: means over their 16 heads), and the notes on which a head puts more on N than on S. I_c is the head's contrastive importance (nats). [R-C5]

Head	I_c (P2C)	on N	on S	$N > S$ (notes of 92)
`L30H11`	0.51	0.33	0.13	91
`L38H22`	0.22	0.21	0.15	72
`L38H39`	0.19	0.14	0.02	92
`L38H35`	0.12	0.11	0.10	49
random-16	—	0.02	0.01	—
bottom-16	—	0.05	0.04	—

Caveat. The read side is descriptive. It is linked to the write side by correlation across events, and the attention patterns were not intervened on; attention-pattern patching, the direct test, was not run. The families are read off medians over notes and hold across notes: the region readers' lift exceeds 4 on every note.

F PROSE LOCUS

Supports: Sec. 3.4; [R-C6]

The events of Appendix C measure heads at the code token, where the chapter margin is usually already decisive (Table 5). Two additions locate where it can be elicited and by which heads.

Forcing probe. Crop the cache to an earlier position p, append a short suffix that makes the model name a code, and read the chapter margin (target: the letter eventually emitted; competitors: the other inventory letters) at the suffix's last token. This asks which region the residual stream at p supports. The suffix is "`\nSo the code here is`" (suffix B). Four wordings were compared at the code token on a two-note pilot, and B was best there. On the full corpus the four wordings make the emitted letter decisive at the code token (margin at least 2 nats) for between 64% and 73% of mentions, the chosen one for 68%; the best is suffix D ("` The corresponding ICD-10-CM code is`"). We record this and did not rerun; we expect suffix D to recover a few more mentions rather than change the picture. A suffix asks for *the* code, and the model answers with its best-supported one, so a mention that the model lists without endorsing it (a differential, a rejected candidate) does not register. This is the probe's known limit.

Locus. Walk back from the code token in steps of 8 tokens, up to 512 tokens or the start of the trace, until the forced margin first drops below $\theta = 2$ nats; record four more steps for the shape of the curve; then refine the last crossing to the token. The locus is the earliest position from which the forced margin stays at least θ all the way to the code token. Each mention takes 4–9 probes, each a short forward against the cropped cache (about 0.1 s); the cache is never copied but cropped and, when a later probe needs more of it, regrown from the token ids. Every chapter mention receives one of five outcomes: *at the code token* (decisive only there, $d = 0$); *in the reasoning window* (the locus lies $d \geq 1$ tokens before the letter and after the end of the previous mention's code); *inherited* (the locus falls at or before the end of the previous mention's code, so the region was settled with an

earlier code; such mentions are counted, not attributed); *probe failure* (not decisive even with the cache cut one token before the letter); and *censored* (still decisive 512 tokens back). Table 16 gives the outcomes for all 2,064 chapter mentions (18,400 probes) by kind of mention; 0.3% of mentions are censored.

Table 16: Where the chapter can be elicited: the forcing probe's outcome for every chapter mention, as a share of each column's mentions. "Previous code" is the code mentioned just before, in the same or in another chapter. [R-C6]

	all	gold codes	wrong codes	previous code		first code
				same chapter	other chapter	of the note
mentions	2,064	226	1,838	906	1,058	100
at the code token	2%	2%	2%	1%	4%	1%
in the reasoning window	36%	59%	34%	12%	54%	71%
inherited from an earlier code	29%	23%	30%	67%	0%	0%
probe failure	32%	15%	34%	20%	42%	27%

Distance, step size and robustness. For the 750 mentions decided in the reasoning window, the locus lies a median of 7 tokens before the letter (10th–90th percentile 1 to 55); the first code of a note is decided further back (median 23). Per note, the window share is 37% (10th–90th percentile 21% to 55%) and the inherited share 26% (0% to 42%). Across the 8-token step in which the forced margin crosses θ it jumps by a median of 8.3 nats. The curve returns above θ at some earlier grid point for 21% of mentions, which is why the locus is the earliest position from which the margin *stays* decisive; the 8-token grid is fine enough. Lowering θ to 1 or raising it to 3 moves the share decisive at the code token from 68% to 77% or 60%. The inherited mass is inheritance within a chapter: 67% of the mentions whose previous code shares their chapter, and 63% of the mentions of a sibling of an earlier code, are decided with the earlier code and never re-decided. It is not evidence of narrowing far upstream.

The note alone. With the cache cropped at the end of the prompt, before anything is written, the forced margin is decisive for the emitted letter for 3% of mentions, 12% of first codes and 13% of gold first codes (median −3.6 nats). The region is not read off the note; it is written down in the reasoning.

What the locus token is. Trace tokens were classed by the word they belong to: a word of the emitted code's own inventory description (at least four letters, with structural words such as "unspecified" removed), a word of another description in the same chapter, a word of a description in another chapter, part of a code string, punctuation, or other prose (Table 17). 32% of the loci are a word of the emitted code's own description, 14 times its base rate, and together with the words of other descriptions in the emitted chapter they make up more than half of the loci.

Table 17: The locus token, for the 750 mentions decided in the reasoning window: share of each class among all attribution-window tokens (base rate), share among loci, and lift. [R-C6]

Class of token	base rate	at the locus	lift
word of the emitted code's own description	2.3%	32%	14
word of another description in the same chapter	5.4%	22.4%	4.2
other prose	48.0%	34.7%	0.7
word of a description in another chapter	15.8%	4.8%	0.3
punctuation	13.9%	3.7%	0.3
inside a code string	14.5%	2.4%	0.2

Window attribution. One forward of the 64 tokens that end at the locus, plus the suffix, against the cropped cache, with every head's output at every one of those positions exposed as a gradient leaf (graph intact), and one backward from the forced margin give $A[p,h] = (\mu_h - z_h(p)) \cdot \partial\,\text{margin}/\partial z_h(p)$: the first-order effect of mean-ablating head h at position p on the decision made there. Heads at the window positions are the prose-time computation; heads at

the suffix positions are its transport into an emission. NULL events (the same 20 prose positions per trace, with a random letter as target) get the same window and set the floor: prose-time importance is $I_c = I - I_{NULL}$ summed over the window positions, and transport importance the same over the suffix positions. The pass ran on the 804 attributed mentions (750 decided in the window, 48 at the code token, 6 censored) and on all 2,000 NULL positions. About half of a chapter decision's positive importance is at the suffix, the transport; the rest is in the prose, concentrated at the locus token and the three tokens before it, more so than at NULL positions, where nothing was decided (Table 18). Importance falls off with distance rather than piling up at the window's far edge, so 64 tokens are enough.

Table 18: Where the importance sits: share of each event's positive importance (window plus suffix, normalised to one) at the locus token, 1–3 tokens before it, 4–15 before, 16–63 before, and at the suffix. [R-C6]

Events	locus token	1–3 before	4–15 before	16–63 before	suffix
chapter, decided in the window (750)	8%	9%	13%	21%	48%
chapter, decided at the code token (48)	9%	13%	21%	26%	32%
NULL, random letter (2,000)	3%	5%	14%	31%	47%

The prose-time ranking. Table 19 gives the prose-time set's top-16; every interval excludes zero. The set shares 0 of its top-8, 1 of its top-16 and 9 of its top-40 heads with the emission set (the chapter global set), and the two importances correlate at Spearman +0.12 over all 1,600 heads. Its top-40 has 0 heads in layers 0–11, 31 in layers 12–23 (75% of its importance), 9 in layers 24–31 and 0 in layers 32–39: one block earlier than the emission set (Table 8). The one emission top-16 head in the prose-time top-16 is the region reader `L24H1`, third; the region readers `L26H37` and `L26H36` rank 18 and 29; and the emission set's final-layer prose readers (`L39H38`, `L39H13`, `L38H24`) are not in the prose-time top-40. Ranked on the 750 window-decided mentions alone, the top-16 has the same heads with small reorderings. `L24H19`, `L24H1` and `L22H16` split their importance between the prose and the transport (52%, 55% and 51% of it at the suffix), whereas `L18H4`, `L22H15` and `L18H3` act mostly in the prose, at and before the locus token (at most 27% at the suffix).

Table 19: The prose-time ranking: decision-specific importance summed over the attribution-window positions (nats; mean over notes) with its 95% bootstrap interval, and the head's rank and I_c in the emission-time chapter ranking. [R-C6]

Head	I_c (prose-time)	95% interval	emission rank	emission I_c
`L24H19`	0.265	[0.244, 0.288]	36	0.10
`L22H16`	0.249	[0.221, 0.276]	49	0.07
`L24H1`	0.228	[0.206, 0.251]	14	0.17
`L18H4`	0.184	[0.111, 0.251]	127	0.02
`L22H15`	0.177	[0.141, 0.213]	43	0.08
`L21H11`	0.160	[0.122, 0.198]	67	0.05
`L18H3`	0.158	[0.104, 0.216]	332	0.00
`L22H17`	0.155	[0.123, 0.186]	41	0.08
`L22H18`	0.151	[0.127, 0.175]	169	0.01
`L20H19`	0.143	[0.117, 0.170]	1,557	−0.04
`L24H16`	0.128	[0.108, 0.149]	37	0.09
`L19H38`	0.115	[0.075, 0.156]	123	0.02
`L21H1`	0.113	[0.087, 0.141]	40	0.09
`L19H37`	0.109	[0.081, 0.139]	1,362	−0.01
`L16H38`	0.109	[0.081, 0.135]	1,259	0.00
`L17H18`	0.106	[0.091, 0.121]	1,224	0.00

The transport ranking. Summed over the suffix positions instead, I_c ranks `L21H25` first (0.92 nats), followed by `L26H37`, `L27H13`, `L25H23`, `L26H36`, `L20H20`, `L26H18`, `L27H11` and `L24H1` (0.77 down to 0.50); the prose-time heads `L22H16` and `L24H19` rank 10 and 12, and `L39H38` and `L39H13` rank 14 and 15. The transport ranking overlaps the emission set (4 of the top-8, 9 of the top-16 and 21 of the top-40; Spearman +0.34 over all heads and +0.45 over the

emission top-100) and also contains the layer-22 and layer-24 prose-time heads. Its top-16 has 5 heads in layers 12–23 (32% of its importance), 9 in layers 24–31 (58%) and 2 in layers 32–39. One suffix and one decision give two rankings: what was decided in the prose was decided mostly by heads one block earlier than the ones that carry it out, and the dissociation is by position (window-only against suffix-only ablation, below), not by disjoint heads.

Exact faithfulness at the locus (pass A). Over the window and suffix positions, the locus pass also mean-ablated exactly the emission set's top-k, the event's own top-k by window attribution, random-k and the emission bottom-k, and, at $k = 16$, the two sets restricted to the window positions only and to the suffix positions only (Table 20). The emission set flips the forced decision, but through the suffix: restricted to the prose positions it removes 0.16 of the forced margin, while the heads ranked at the prose positions remove all of it from there (1.08). A per-event set is fitted to the event it is tested on; the fair test is a fixed set (pass B).

Table 20: Pass A: share of the forced chapter margin destroyed by mean-ablating a set at the locus (804 mentions, 100 notes; medians over notes of per-note means, emission set with its 10th–90th percentile). Top: over window and suffix positions; bottom: at $k = 16$, restricted by position. [R-C6]

k	emission top-k	event's own top-k	random-k	emission bottom-k	emission > random
8	0.74 (0.55–0.98)	1.62	0.00	−0.58	100/100
16	1.08 (0.80–1.34)	1.95	0.09	−0.14	100/100
40	1.58 (1.35–1.84)	2.16	0.20	1.38	100/100
Set, $k = 16$	window positions only	suffix positions only	both		
emission top-16	0.16	1.01	1.08		
event's own top-16	1.08	1.57	1.95		

The fixed prose-time set (pass B). The global prose-time set, the same top-k on every note, was mean-ablated at the locus against count-matched random heads and the bottom of the prose-time ranking (Table 21). Eight heads removed only at the prose positions remove the whole forced decision, where the emission set's top-16 removed 0.16 from the same positions; the same eight heads at the suffix remove 0.45, the mirror image of the emission set (1.01 at the suffix). On the 750 window-decided mentions alone the numbers agree within 0.04. The chapter is narrowed in the prose by one set of heads and executed at the code token by another.

Table 21: Pass B: share of the forced chapter margin destroyed by mean-ablating the fixed global prose-time top-k at the window positions only (with its 10th–90th percentile over notes), at the suffix positions only, and at both, against random-k and the prose-time bottom-k at the window positions (804 mentions, 100 notes; medians over notes of per-note means), and the notes on which the prose-time set beats random. [R-C6]

	Prose-time top-k			random-k	bottom-k	prose-time > random
k	window only	suffix only	both	window	window	(notes of 100)
8	1.00 (0.58–1.38)	0.45	1.26	0.02	−0.11	100
16	1.32 (0.88–1.76)	0.48	1.48	0.01	−0.01	100
40	1.69 (1.18–2.14)	0.77	1.71	0.07	0.09	99

What the prose-time heads read. Attention rows were read at the locus token in the natural context (no suffix) for four 16-head groups: the prose-time top-16, the emission top-16, the prose-time bottom-16 and 16 random heads (Table 22). The prose-time set's share on the emitted chapter's lines beats the random panel's on 84 of 100 notes (71 at the three window positions where the set's importance is largest), but its distinctive reading is the note: 0.25 of its attention, against 0.08 for the random heads and 0.07 for the emission top-16 at the code token (Table 12). Per head there are two families, different from those at the code token. *Note readers* in layers 18–24 (`L24H19`, `L24H1`, `L22H17`, `L21H1`, `L21H11`, `L18H4`, `L18H3`, `L24H16`) put 30% to 53% of their attention on the discharge summary, with a sharp 7% to 16% on the emitted chapter's lines (lifts 3 to 9, over 27 to 64 effective inventory tokens). *Prose readers* in layers 17–22 (`L22H16`, `L22H18`, `L19H38`, `L19H37`,

L17H18, L20H19) put 60% to 79% on the reasoning so far. One head, L22H15, reads the region at the locus (0.31, lift 6). The link between reading the region and mattering is weak here (r at most 0.21): at the locus the prose-time heads' effect does not depend on how much of the inventory they read. This is consistent with a narrowing driven by the note's evidence and the reasoning, which the emission set's region readers then match against the inventory lines at the code token (Appendix E).

Table 22: What the prose-time heads read: attention at the locus token in the natural context (no suffix), mean over the 804 loci, for four 16-head groups: share on the region (the emitted chapter's lines), on other chapters' lines, on the note, on the reasoning prose and on the sink; region lift; n_{eff} over inventory tokens. Descriptive: attention was not intervened on. [R-C6]

Group	region	other chapters	note	prose	sink	lift	n_{eff}
prose-time top-16	0.08	0.12	0.25	0.42	0.04	4.5	328
emission top-16	0.06	0.21	0.20	0.36	0.11	3.1	174
prose-time bottom-16	0.05	0.14	0.17	0.56	0.01	2.9	287
random-16	0.05	0.36	0.08	0.27	0.21	1.9	148

Scope. The locus results concern endorsed mentions decided in the reasoning window (36% of chapter mentions). Another 29% inherit the chapter from an earlier code and are counted, not attributed; for 32% (34% of wrong codes and 15% of gold ones) the probe never becomes decisive, and they are excluded. We do not claim a locus for every mention. The prose-time ranking is first-order (fidelity 0.84 under the intact graph, Appendix C); its standing as a set comes from the exact ablations of passes A and B, and attention was not intervened on.

G Replication on the retrieved shortlist

Supports: Sec. 3.5; [R-C7]

HD-002 ran the five passes of Appendices C–F, with unchanged scripts, on new teacher traces for the same 100 characterization notes (TR-007; decoding as in Appendix B), with the full inventory replaced by the retrieved shortlist of the distillation half (Appendix A: K = 1,000 plus the siblings of the top 200, about 1,550 lines). On these prompts the teacher scores micro-F1 0.199 (over the 99 of 100 traces that are valid), against 0.168 with the full inventory (all 100 valid). Table 23 compares the two runs.

What is preserved. The I_c rankings agree at Spearman 0.55, 0.63 and 0.55 (P1, P2, P2C) over all 1,600 heads, with 31, 31 and 32 of the top-40 and 5, 6 and 7 of the top-8 shared. The heads persist: the chapter top-8 keeps L26H36, L27H11, L27H13 and L28H21, the character top-8 keeps L39H10, L39H29, L39H12 and L37H6, and the contrastive top-8 keeps L30H11; the anchor-cosine heads of Appendix B stay non-causal. The global sets keep their causal standing: top-40 knock-in recovers 0.92, 0.82 and 0.85 against 0.04, 0.03 and 0.02 for random sets, and top-40 necessity flips the chapter and character decisions in both runs. With fewer lines to read, the chapter top-16 put more of their attention on the emitted chapter's lines (0.14 against 0.10) at a lower lift (3.04 against 3.54). The chapter can be elicited a median of 9 tokens before the letter (7 on the full inventory) from the same prose-time set (Spearman 0.79, 31 of the top-40 shared; top three L22H16, L24H19 and L22H17), and the dissociation reproduces: window-only ablation of the event's own locus top-16 removes 1.04 of the forced margin, that of the emission top-16 0.19.

What shifts. L39H38 leaves the chapter top-8 on the shortlist, and the chapter top-8's knock-in is lower (0.47 against 0.67), although the top-40's is higher. Fewer chapter mentions are decided in the reasoning window (0.27 against 0.36), more inherit the chapter (0.35 against 0.29), and more never become decisive under the probe (0.37 against 0.32). The fixed prose-time top-8 removes less of the forced margin from the window positions (0.73 against 1.00), and L30H11's share on the near-miss's earlier mentions is lower (0.263 against 0.33). The pooled weights spread out: the shortlist's top-40 is not dominated by one head (Appendix D).

Table 23: Full inventory (HD-001) vs retrieved shortlist (HD-002) on the same 100 notes. [R-C7]

Quantity	Full inventory (HD-001)	Shortlist (HD-002)
Contrast rankings over 1,600 heads (agreement between the two; P1 / P2 / P2C)		
Spearman	0.55 / 0.63 / 0.55	
top-40 shared (of 40)	31 / 31 / 32	
top-8 shared (of 8)	5 / 6 / 7	
Global per-phase sets (median over notes; P1 / P2 / P2C)		
knock-in, $k = 8$	0.67 / 0.48 / 0.69	0.47 / 0.46 / 0.71
knock-in, $k = 40$	0.86 / 0.79 / 0.85	0.92 / 0.82 / 0.85
knock-in, random-40	0.04 / 0.02 / 0.02	0.04 / 0.03 / 0.02
necessity, $k = 40$	1.24 / 1.61 / 0.86	1.35 / 1.57 / 0.93
Read side (top-16 heads)		
P1 share on the emitted chapter's lines	0.10	0.14
P1 region lift	3.54	3.04
P2C share on the near-miss's mentions	0.074	0.054
Prose locus (chapter mentions)		
median distance (tokens)	7	9
decided in the window	0.36	0.27
inherited from an earlier code	0.29	0.35
never decisive under the probe	0.32	0.37
prose-time ranking: Spearman / top-40 shared	0.79 / 31	
window-only ablation, locus top-16 (pass A)	1.08	1.04
window-only ablation, emission top-16 (pass A)	0.16	0.19
window-only ablation, global prose-time top-8 (pass B)	1.00	0.73

What the distillation half takes. The distillation half took its targets from HD-002: the global sets and their weights for w_h^T, and the prose-time set for the teacher's pool on the 8-token supervision window before each chapter letter (Sec. 4.1). Its teacher reference values are therefore HD-002's: the knock-in rows of Table 23 for M4, the locus distance for M5, and the values in Table 24. HD-001 remains the paper's primary characterization.

Table 24: Further teacher reference values used by the distillation half, on the full inventory (HD-001) and on the shortlist (HD-002, the values used), with the readout that uses each. Random-16 panel in brackets. [R-C7]

Quantity	HD-001	HD-002	used by
`L30H11` share on the near-miss's earlier mentions	0.33	0.263	M3 reference
P2C top-16 share on the code prefix	0.254 (0.151)	0.260 (0.087)	prefix-share readout
P1 top-16 region lift	3.54 (0.87)	3.04 (1.25)	M2 at P1
P2C top-16 region lift	3.94 (1.89)	3.38 (1.95)	M2 at P2C
prose-time top three	`L24H19`, `L22H16`, `L24H1`	`L22H16`, `L24H19`, `L22H17`	teacher's window pool

H MISTILL OBJECTIVE DETAILS

Supports: Sec. 4.1; [R-D10], [R-D9]

This appendix states the objective as it was trained, the pre-registered term that was measured and removed, how λ was selected, and what the objective costs to train. Everything in Sections H.2 and H.3 is the selection stage: `Qwen3-4B` trained on 100 notes for two epochs and evaluated on the 100 validation notes (the characterization notes, regenerated on the shortlist, which never enter training).

H.1 THE LOSS AS RUN

MISTILL minimises Eq. (1) with $\lambda_{P1} = \lambda_{P2} = 0.1$, teacher and student teacher-forced on the teacher's trace. $\mathcal{L}_{CoT}$ is the Vanilla CoT-KD loss unchanged (Appendix I): the mean cross-entropy over

the trace tokens plus, with weight 1, the KL divergence $\mathrm{KL}(p^T \| p^S)$ between the teacher's stored distribution over its top-32 tokens and a tail bin and the student's distribution over the same bins; for `SmolLM3-3B`, whose vocabulary differs, universal logit distillation (ULD; Boizard et al., 2025) over the same number of bins replaces the KL. Each phase adds two terms. With R_P the supervised rows of phase P and $E_P \subseteq R_P$ its event rows,

$$\mathcal{L}^P = \frac{1}{|R_P|} \sum_{q \in R_P} \mathrm{KL}\big(\bar{\alpha}_q^T \,\big\|\, \bar{\alpha}_q^S\big) + \frac{1}{|E_P|} \sum_{e \in E_P} \big(m_e^S c_e^S - m_e^T c_e^T\big)^2, \tag{4}$$

which is Eq. (3) with each term averaged over the rows where it is defined: the attention term on every supervised row, the interaction term only where a decision, and so a margin, exists. R_{P1} holds every P1 event and the W=8 positions before it, and E_{P1} the events alone; $R_{\mathrm{P2}} = E_{\mathrm{P2}}$ holds the P2 and P2C events. Each row is read through one head pool per model: the P1 set at a P1 event, the P2 set at a P2 event, the P2C set at a P2C event, and, on the window, the teacher's prose-time set and the student's P1 set (below).

Pooled attention, region mass and contribution. $\bar{\alpha}_q^M$ is Eq. (2), renormalised over buckets. At an event, m_e is the pooled mass on the decision's region: every shortlist line whose code starts with the emitted chapter letter at a P1 event; the code's own line, its earlier mentions (S) and its emitted prefix at a P2 or P2C event. The pooled heads' direct contribution to the event's margin is

$$c_e = \frac{\sum_h w_h \left(W_O^h z_{h,e}\right)^\top u_e}{\max\left(|\,\mathrm{margin}_e\,|,\, 1\right)}, \qquad u_e = U_t - \sum_{j \in C} \pi_j U_j, \quad \pi = \mathrm{softmax}\left(\mathrm{logit}_C\right), \tag{5}$$

where $z_{h,e}$ is head h's output at the event (its share of the input to the layer's output projection), W_O^h that head's slice of the projection, U the unembedding, and t and C the event's target and competitors in the model's own vocabulary. u_e is the gradient of margin_e with respect to the final residual stream, with the scale of the final normalisation held fixed. The denominator is the model's own margin, detached and floored at one nat so that a near-zero margin cannot inflate the term; c is a scalar per event, free of width, head count and tokenizer.

Span buckets. Every span is a character range in the untemplated text [prompt; trace], so one bucket list serves both models; chat-template tokens and the first-position attention sink fall into a separate bucket. The buckets are the shortlist header, one bucket per shortlist line (code and description), the note in windows of 32 words (the notes lack reliable sentence punctuation, so fixed windows stand in for sentences), the instruction, and the trace split at sentence boundaries, with pieces capped at 400 characters. At a P2 or P2C event five buckets are carved out of these: the code's own line, its category siblings' lines, the code's earlier mentions (S), the near-miss's earlier mentions (N) and the code's emitted prefix. Each token's attention goes to the bucket that contains its span, so the two models' pooled rows are distributions over one axis.

Head pools. Each model pools a causally ranked top-40 per phase with $w_h = \mathrm{softmax}(I_c/\tau)$ over the 40 heads, τ the standard deviation of their importances. The teacher's pools are HD-002's global sets on the shortlist (Appendix G) and, on the P1 window, its prose-time set, weighted the same way. The student's pools come from the same attribution pass (the calibrated ranking of Sec. 2.3) run once on the untrained student, teacher-forced on the 100 validation traces. On `Qwen3-4B` one late head, `L35H25`, tops all three phases and carries most of each pool's weight (0.742, 0.569 and 0.512 at P1, P2 and P2C); `SmolLM3-3B`'s pools are flatter and mid-layer, led at P1 by `L22H4`, `L19H0` and `L29H11`. The pools are fixed for the run, so the objective is stationary; re-ranking the student's pool during training was planned as an ablation and not run.

The window, and its asymmetry. The W=8 positions before each P1 event lie between the characterization's median locus distances, 7 tokens on the full inventory and 9 on the shortlist (Sec. 3.4; Appendix G). On them the teacher is pooled over its prose-time set and the student over its own P1 set, because no prose-time set had been measured for either student when training began: on the window, the heads of the student's P1 set are asked to reproduce the attention of the teacher's prose-time heads. A prose-time set for the untrained `Qwen3-4B` was later produced as a by-product of its M5 locus pass, which makes a symmetric-window arm possible. Neither that arm nor the no-window arm, which would measure what the window contributes, was run (Appendix L).

Events across tokenizers. Each event carries an anchor, the first non-blank character its decision emits. A model's position for the event is the last token that ends at or before the anchor, and the event is kept only if that model's next token begins at the anchor (up to leading whitespace), that is, if the model makes the decision there. A tokenizer that merges the decided character into a longer token, as `SmolLM3-3B`'s does with digit runs, has no such position. The event is then dropped rather than moved to a neighbouring token, which would match the teacher's attention at one character against the student's at another. A cross-tokenizer student is therefore supervised on the subset of the teacher's rows at which its own tokenization decides, and its target and competitor ids, hence its margin and c, are its own. Window rows are paired by their distance in tokens before the event: the same text for `Qwen3-4B`, whose tokenizer is the teacher's, and an approximation for `SmolLM3-3B`.

H.2 The write term: pre-registered, measured, removed

The pre-registered objective had a third term per phase, $\frac{1}{|E_P|}\sum_e (c_e^S - c_e^T)^2$, which matches the pooled heads' normalised contribution directly. At pilot size it was harmful in both forms tried (Table 25). As written, a squared error against a target the student does not yet track is minimised by a constant. Trained at $\lambda = (1, 1)$, the P1 write term fell by 93% while the student–teacher correlation of c went from +0.11 to −0.06; the pooled heads stopped carrying the chapter margin (M1 at P1 fell from 0.815 to 0.300), and every cell of the pre-registered 3×3 grid scored below the untrained student. Standardising c over each note's events before the squared error removes the constant minimum, and the term then learns the teacher's pattern (correlation +0.38). The arm still scores below Vanilla CoT-KD trained on the same 100 notes at every λ tried, below the untrained student on completed answers even at $\lambda = 0.1$ (0.073 against 0.077), and its M1 falls in every phase. The term was removed. Both forms remain in the code (`--terms`, `--write-norm`), and c remains inside the interaction term, where it is the pooled heads' share of the margin that M1 reads (Appendix J). The removal is a post-hoc change to the pre-registered objective and is labelled as one wherever the objective is described.

Table 25: Selection stage on the 100 validation notes (`Qwen3-4B`, 100 training notes, two epochs, seed 0; $\lambda_{P1} = \lambda_{P2} = \lambda$). The pre-registered objective falls below the untrained student and the standardised write term below Vanilla CoT-KD trained on the same notes; without the write term the objective reaches Vanilla CoT-KD at $\lambda = 0.1$. Multi-valued cells list $\lambda = 1 \,/\, 0.3 \,/\, 0.1$. [R-D10]

Arm	validation micro-F1	valid	M1 P2C
untrained student	0.076	0.99	0.752
Vanilla CoT-KD (100 notes)	0.101	0.90	0.752
pre-registered (attention + write + interaction), best of 3×3	0.048	0.99	0.766
standardised write term, $\lambda = 1 \,/\, 0.3 \,/\, 0.1$	0.056 / 0.071 / 0.066	0.71 / 0.76 / 0.89	0.743 / 0.743 / 0.750
attention + interaction (as run), $\lambda = 1 \,/\, 0.3 \,/\, 0.1$	0.086 / 0.074 / 0.110	0.75 / 0.83 / 0.92	0.778 / 0.775 / 0.773

H.3 Selecting λ on the validation notes

The pre-registered rule was: $\lambda_{P1}, \lambda_{P2} \in \{0.3, 1, 3\}$, nine runs at 100 training notes, evaluated on the validation notes; select by validation micro-F1 with ties broken by M1 at P2C; use the chosen pair for every MISTILL and ablation run. On the pre-registered objective the grid is flat and below the untrained student (micro-F1 from 0.033 to 0.048, against a one-seed noise of 0.005). It cannot be ranked, and it is reported as the write term's negative rather than as a sensitivity to λ. At $\lambda = 1$ the mechanism terms also outweighed the task loss about 10-fold (about 3-fold at 0.3), and the arms over-generated. Every generation that reached the 16,384-token cap on the validation notes was on one of the notes from which the discharge diagnosis had been removed, which are most of them; at $\lambda = 1$, 27–29% of those notes' generations reached the cap, against 11% for Vanilla CoT-KD and 1% for the untrained student. The range was therefore extended below its pre-registered floor to 0.1, disclosed as an extension.

On the objective without the write term only $\lambda = 0.1$ reaches Vanilla CoT-KD trained on the same notes (Table 25), and the objective's effect on the decisions scales with λ (Table 26): the contrastive decision moves toward the teacher in proportion to the weight, the chapter and character decisions move slightly away, and the task holds only at the smallest weight. The selected pair, $\lambda_{P1} = \lambda_{P2} = 0.1$, was confirmed against Vanilla CoT-KD on three seeds each (Table 27). On the task MISTILL is at

least Vanilla CoT-KD, with an edge that the per-note paired test supports and a comparison of three seed means cannot. On the mechanism the two differ clearly: MISTILL's supervised heads read the emitted code's region more, and it carries the contrastive decision about a third of the way from the untrained student to the teacher, where Vanilla CoT-KD leaves it at the untrained student's level on every seed, at a small, significant cost at the chapter decision, the phase that is nearly saturated already (Appendix J). The second axis was exercised once, at tier-1 size, as $\lambda_{P1} = 0.03$ with $\lambda_{P2} = 0.1$ (Appendix K). No tuning budget was spent on the baselines' own hyper-parameters (Appendix I).

Table 26: Dose-response of the objective without the write term on the validation notes: M1 differences from Vanilla CoT-KD trained on the same 100 notes (seed 0), paired per note over the notes with events in every phase (mean per-note difference; notes higher / lower; sign-test p). [R-D10]

λ	Δ M1 P2C	Δ M1 P1	Δ M1 P2
0.1	+0.024 (69/11, $p < 10^{-4}$)	−0.007 (18/33, p 0.05)	−0.002 (p 0.25)
0.3	+0.034 (73/14, $p < 10^{-4}$)	−0.015 (15/46, p 10^{-4})	−0.011 (34/56, p 0.03)
1	+0.041 (67/16, $p < 10^{-4}$)	−0.020 (19/48, p 5×10^{-4})	−0.016 (26/64, p 10^{-4})

Table 27: The selected objective ($\lambda = 0.1$) against Vanilla CoT-KD trained on the same 100 notes, three seeds each, on the validation notes. Anchors in the first column: the untrained student and the teacher on its own validation traces. Paired: seed-matched and pooled over the three seed pairs (300 note-pairs; notes higher / lower, sign-test p, mean per-note difference). M2: region lift of the supervised heads, top-16 median, at P1 / P2 / P2C; its paired entry gives ranges over the nine seed pairings (notes higher of 100, median difference, largest p). [R-D10]

Readout (anchors)	Vanilla CoT-KD, seeds 0 / 1 / 2	MISTILL, seeds 0 / 1 / 2	paired
micro-F1	0.101 / 0.092 / 0.091 = 0.095 ±0.006	0.110 / 0.099 / 0.103 = 0.104 ±0.006	116/86/98 tied, p 0.04, mean +0.012
valid	0.90 / 0.95 / 0.91	0.92 / 0.91 / 0.97	—
M1 P2C (untrained 0.752, teacher 0.807)	0.752 / 0.751 / 0.752	0.773 / 0.772 / 0.772 (37% of the gap)	202/31, $p < 10^{-4}$, mean +0.024
M1 P1 (untrained 0.815, teacher 0.887)	0.839 / 0.837 / 0.840 (33% of the gap)	0.832 / 0.831 / 0.832 (24%)	42/100, $p < 10^{-4}$, mean −0.007
M1 P2 (untrained 0.647, teacher 0.750)	0.665 / 0.664 / 0.671	0.663 / 0.663 / 0.667	105/146, p 0.01, mean −0.003
M2 lift, supervised heads	2.01–2.09 / 3.29–3.31 / 2.61–2.68	2.28–2.33 / 3.56–3.58 / 2.84–2.89	78–99 of 100 higher, +0.20 to +0.32, $p < 10^{-4}$

H.4 IMPLEMENTATION AND TRAINING COST

The teacher's side is computed once: for each of the 493 training notes with a valid teacher trace, the pooled bucket distribution at every supervised row and m^T, c^T at every event. Per note the objective supervises a median of 562 rows (at most 2,903) at a median of 255 events, 40 heads per row, where Feature-KD matches every head of four layer pairs at 64 sampled positions and hidden states at every position. On the student side, attention is computed only at the supervised positions: the pooled layers' attention rows are recomputed from their checkpointed input hidden states with the model's own projections, rotary embeddings and query/key norms, inside checkpointed chunks, and the heads' outputs are rebuilt from the same rows. MISTILL therefore needs no online teacher and trains on one GPU. Unit tests check that every term is zero when the student's pooled quantities equal the teacher's, that the attention term with a uniform pool over one layer's heads equals Feature-KD's bucket-pooled KL, and that the stored teacher quantities are reproduced by an independent route (full attention outputs and per-head output-projection effects) on a small model; a smoke run checks that each ablation flag changes only what it names.

Cost ([R-D9]). On the corpus's longest note the `Qwen3-4B` step peaks at 37.5 GiB with bf16 weights, flat in note length, at about 27 s per note; the tier-1 runs logged peaks of 52 GiB. At 100 training notes Vanilla CoT-KD trains in 24–34 min and MISTILL in 76–78 min. A 493-note MISTILL run takes 6 h 50 min on an H200 (3.1–3.3 min per eight-note step, and the same on an RTX PRO 6000: the step is largely CPU-bound), against 3.0 h for a 493-note Vanilla CoT-KD run with fp32 master weights on different hardware, so the comparison is indicative only. With fp32 master weights MISTILL needs about 97 GiB on one GPU: seed 0 was split over two 80 GB GPUs, seeds 1 and 2 ran on one 180 GB GPU each.

I BASELINES AND TRAINING DETAILS

Supports: Sec. 4.2; [R-D0], [R-D2]–[R-D4]

The comparison uses one strong representative per distillation family, run from its official code where that code exists and at its published settings. Two rules fixed the set before any run: a third-party method enters only if its official code can be used directly, without re-implementing someone else's system, and the set covers the families generically, so that it cannot be read as a few hand-picked weak baselines.

I.1 Models and software

The teacher is `Qwen3-14B` (`Qwen/Qwen3-14B`) in thinking mode, in bf16 with static YaRN ×4; its traces (one per note: temperature 0.6, top-p 0.95, top-k 20, seed 0, a 16,384-token cap) are the targets of every arm. `Qwen3-4B` (`Qwen/Qwen3-4B`; 36 layers of 32 heads, 8 key-value heads) shares the teacher's tokenizer and is reasoning-trained; it runs in thinking mode under static YaRN ×2 (an 81,920-token window) in training, generation and every teacher-forced metric, the teacher's positional scaling with one factor for the whole run. `SmolLM3-3B` (Bakouch et al., 2025) (`HuggingFaceTB/SmolLM3-3B`; 36 layers of 16 heads, 4 key-value heads; a 65,536-token native window, no YaRN) comes from another family and runs in its native thinking mode; its tokenizer writes digit runs as single tokens, so some of its P2 decisions span several characters (Appendix H). The LAS simulator is `Qwen3-1.7B` without thinking (Appendix J); the retriever encoder is gte-large (Appendix A). Checkpoints are the public Hugging Face releases, loaded offline from a local cache so that a missing artifact fails instead of resolving to another revision; their revisions, the Hugging Face commit hashes in the download records of the local copies, are

`Qwen/Qwen3-14B`	`40c069824f4251a91eefaf281ebe4c544efd3e18`	teacher
`Qwen/Qwen3-4B`	`1cfa9a7208912126459214e8b04321603b3df60c`	student
`HuggingFaceTB/SmolLM3-3B`	`a07cc9a04f16550a088caea529712d1d335b0ac1`	student
`Qwen/Qwen3-1.7B`	`70d244cc86ccca08cf5af4e1e306ecf908b1ad5e`	LAS simulator
`thenlper/gte-large`	`4bef63f39fcc5e2d6b0aae83089f307af4970164`	retriever encoder

Software: PyTorch 2.11, Transformers 5.16, vLLM 0.29 for generation and TRL 1.13.0 for the on-policy arms.

I.2 The shared training recipe

The off-policy arms (Vanilla CoT-KD, Feature-KD, DistiLLM-2, and MISTILL with its controls and ablations) share one trainer and one recipe: the 493 training notes with a valid teacher trace (the seven notes whose teacher trace is invalid or degenerate are excluded from every arm), two epochs (124 optimiser steps), one trace per micro-step with gradient accumulation to a global batch of 8, AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, no weight decay) at learning rate 10^{-5} with a 5% linear warm-up and cosine decay, gradient-norm clipping at 1, gradient checkpointing, and the loss on the trace tokens only (the vocabulary logits are computed chunk by chunk and never held in full). The epoch order is a function of the seed and the epoch, and a resumed run restores the optimiser, scheduler and random-number states, so it continues the same data order and random draws. The teacher's per-position top-32 distributions are precomputed once. Each run writes its full configuration (every argument, the example ids and the PyTorch version) to `config.json`.

I.3 The arms

Vanilla CoT-KD (sequence-level; Kim & Rush, 2016; Hinton et al., 2015): $\mathcal{L}_{\mathrm{CoT}}$ alone, the cross-entropy on the trace tokens plus the top-32 KL with weight 1 (Appendix H). On `SmolLM3-3B` every arm replaces the KL by ULD (Boizard et al., 2025), which compares the teacher's and the student's sorted probabilities over the same number of bins and needs no shared vocabulary, so that the `SmolLM3-3B` arms differ only in their own terms.

DistiLLM-2 (Ko et al., 2025): its contrastive loss, vendored verbatim from the official repository (commit `fe4cf9b`, Apache-2.0) with a unit test, at its published settings, with the teacher's trace as the preferred response and the untrained student's own generation for the same prompt as the dispreferred one; the teacher runs online. The loss is defined only for a shared vocabulary and the official code has no cross-tokenizer path, so the arm runs on `Qwen3-4B` only; GOLD is the cross-family logit-distillation row.

OPD (on-policy distillation in the GKD lineage; Agarwal et al., 2024): TRL's `DistillationTrainer` (von Werra et al., 2020). The student samples its own completions, capped at 12,288 tokens (no rollout reached the cap at any logged step), and matches the teacher's next-token distributions on them under the generalised Jensen–Shannon divergence at $\beta = 1$ (the reverse KL, TRL's default), for two epochs at learning rate 10^{-5}, with the student loaded in bf16 and no fp32 master copy. Trainer and vLLM rollout engine share one 141 GB H200, because the cluster's idle-GPU policy rules out a separate rollout server.

GOLD (Patiño et al., 2025), the cross-tokenizer on-policy arm for `SmolLM3-3B`: TRL's experimental `GOLDTrainer` with the ULD loss and its text-span token alignment. One override computes the logits only over the completion window, because full-sequence logits of 30k-token prompts do not fit; the loss, the alignment and the on-policy machinery are upstream's. Like OPD, it trains bf16 weights without an fp32 master copy. The arm ran under deviations recorded before its final run: one epoch (a compute cut); rollouts capped at 6,144 tokens (16,384 and then 8,192 ran out of memory in the hybrid ULD loss on a 141 GB H200); a per-device batch of one; an on-policy fraction of 0.5, because a fully on-policy first run collapsed to degenerate output with no valid answer and was discarded; and the teacher's completion, which the off-policy half trains on, truncated to the same cap to fit in memory, which cut the end of 200 of the 500 training traces. The resulting null, and why it is reported, are in Appendix K.2.

Feature-KD (the MiniLM / TinyBERT lineage; Jiao et al., 2020; Wang et al., 2020; Zagoruyko & Komodakis, 2017; Romero et al., 2015): a KL divergence between the teacher's and the student's head-averaged, bucket-pooled attention (the span buckets of Appendix H) in four uniformly mapped layer pairs, at 64 trace positions sampled per note, plus a mean squared error between hidden states at those layers (the student's through a linear adapter to the teacher's width) at every trace position, with weights 1 and 1, no phases and no events; the teacher runs online. The hidden states are RMS-normalised per position before the error, because a raw error against the teacher's massive activations swamped the loss (that run was discarded). It is the generic attention and representation control, at matched data and schedule.

The adapter-seed caveat. When the four Feature-KD runs were trained, the trainer seeded its random-number generators after constructing the adapters, so the seed did not control their initialisation. The order was fixed afterwards, and two runs at one seed then agreed to every printed digit; every MISTILL run was trained after the fix. The Feature-KD runs are therefore not bit-reproducible, and their seed spread (0.153 ± 0.010 on `Qwen3-4B`) is an upper bound on the true seed variance. The effect the row is used for, Feature-KD below Vanilla CoT-KD by -0.035 [-0.049, -0.021], is larger than that spread. No other arm has adapters.

Methods not run. UniCoTT (Zhuang et al., 2025) distils chain, tree and graph explanation structure into small NLU students, and on chain traces its structural loss is degenerate; MoRSD (Yan et al., 2025) has an empty public repository; CWT / Skip-Thinking (Chen et al., 2025a) and Circuit Distillation (Wadhwa et al., 2025) have no public code; SemCoT (He et al., 2025) makes the student emit implicit tokens, so it has no rationale to score for fidelity or mechanism.

I.4 WHAT IS MATCHED, SEEDS AND PRECISION

Matched and not matched. Every arm sees the same shortlist prompts, the same 493 notes and teacher traces, the same teacher targets, global batch, learning rate and optimiser, and the same evaluation script, decoding and metrics, for two epochs (GOLD one). Not matched: (i) hyper-parameter tuning: MISTILL's λ was selected on the validation notes, with the range extended below its pre-registered floor (Appendix H.3), while the baselines run at their published defaults (Feature-KD's weights, DistiLLM-2's β, OPD's β, and GOLD's on-policy fraction as forced by its collapse); the matching budget for each baseline's own hyper-parameter was not spent; (ii) wall-clock time and hardware (Appendix H.4); (iii) the on-policy arms' rollout budgets; (iv) master-weight precision, below.

Seeds. Three seeds: `Qwen3-4B` Vanilla CoT-KD (fp32 and bf16), Feature-KD and MISTILL (bf16 and fp32). Two: `SmolLM3-3B` Vanilla CoT-KD (fp32) and MISTILL, and, among the controls and

ablations, `Qwen3-4B` RandHeads and `SmolLM3-3B` bottom-40. One: every other cell (DistiLLM-2, OPD, GOLD, `SmolLM3-3B` Feature-KD and Vanilla CoT-KD (bf16), every other control and ablation), marked in every table. A seed fixes the data order and the arm's own random draws (Feature-KD's sampled positions, RandHeads' heads, RandPos' positions); evaluation sampling is seeded per note, independently of the training seed.

Precision. MISTILL, its controls and ablations, the Vanilla CoT-KD runs of the learning curve and the precision controls train with bf16 weights and no fp32 master copy, updated by plain AdamW at learning rate 10^{-5}. The off-policy baselines (Vanilla CoT-KD, Feature-KD, DistiLLM-2) keep fp32 master weights under bf16 autocast; the on-policy baselines (OPD, GOLD), trained through TRL, load the student in bf16 and keep no fp32 master copy either. An update smaller than half of bf16's spacing at a weight's magnitude rounds away, so a bf16 run can learn less from the same data. A matched pair shows it: the fp32 Vanilla CoT-KD pilot and the bf16 Vanilla CoT-KD run of the selection stage share their 100 notes, recipe, seed and optimiser steps, and their training cross-entropy falls from 0.536 to 0.419 and from 0.532 to 0.478. The difference was found by comparing the dumped configurations of compared arms. It touches every comparison between MISTILL (or the on-policy arms) and the fp32 baselines, and none inside the MISTILL family or against the untrained student. It was controlled directly (Appendix K.3): Vanilla CoT-KD trained with bf16 weights, three `Qwen3-4B` seeds and one `SmolLM3-3B` seed, with the reading fixed before the result (at MISTILL's task level, the tier-1 task gap is the precision's; at fp32 Vanilla CoT-KD's, the objective's; between, both). The task answer with bf16 weights is the precision's. MISTILL trained with fp32 master weights (`Qwen3-4B`; seed 0 with the student split over two H100s, seeds 1 and 2 on one B200 each) scores 0.165 and trails fp32 Vanilla CoT-KD (-0.024 [-0.039, -0.009]): with fp32 master weights the objective does cost task accuracy (Appendix K.3). The `SmolLM3-3B` Vanilla CoT-KD configurations lack the thinking flag, but for this model the chat template's default renders the identical thinking-mode prompt, so precision is the one real difference between those runs and MISTILL's.

I.5 EVALUATION PROTOCOL AND HARDWARE

Every arm, and each untrained student, generates on the 150 test notes in thinking mode under the teacher's decoding: temperature 0.6, top-p 0.95, top-k 20, sampling seeded per note, a 16,384-token cap, and the student's YaRN setting, in vLLM. Greedy decoding is not used: Qwen3's documentation warns that it loops in thinking mode, and the first pilot under greedy decoding looped to the cap on every note without an answer; the cap is the teacher's, because the teacher's own shortlist traces are long (median about 6.1k tokens, 90th percentile about 9.7k). A mismatch between `SmolLM3-3B`'s training and evaluation templates cost one round of its evaluations and was fixed before any reported row. Answers are parsed and scored as in Appendix J; each evaluation writes its decoding settings to `metrics.json`.

Training and evaluation ran on two university GPU clusters (NVIDIA RTX PRO 6000 with 96 GB, H200 with 141 GB, B200) and on rented cloud GPUs (H100, GH200, H200, RTX PRO 6000). MISTILL-family runs train on one GPU (fp32 MISTILL on two). Feature-KD and DistiLLM-2 need the teacher online: on a 96 GB card the teacher sits on its own card (for Feature-KD, beside a student split over two), while on a 141 GB H200 everything fits one card. OPD and GOLD colocate trainer and rollout engine on one H200. Evaluations ran in vLLM on RTX PRO 6000 or H200 cards (the ablation exceptions are listed in Appendix L.3). Rented instances held the data only while a run lasted and were terminated once their outputs had been copied back and verified. The baseline campaign used about 245 cluster GPU-hours, plus rented GPUs, and the MISTILL campaign about 570 GPU-hours, rented ones included (training, evaluation and the own-heads passes).

J METRICS

Supports: Sec. 4.2, Sec. 4.4; [R-D1]–[R-D6], [R-D15]

The metrics were fixed before any run, in three tiers, with the headline claim placed on the external tier. As measured, the external tier shows no detected difference when both objectives train bf16 weights and a cost with fp32 master weights (Sec. 4.3), and the claims rest on causal recovery on the

students' own heads (Sec. 4.4), stated beside the task result. Every metric is computed on the 150 test notes for every arm by one evaluation script.

Scoring an answer. An arm's answer is the last JSON list after its closing `</think>`, restricted to codes of the inventory. A generation that reaches the 16,384-token cap, or opens a reasoning block and never closes it, is invalid and scores as an empty answer; codes are never scraped from unfinished reasoning. The valid rate is reported beside the task metrics (Table 2), and the share of generations that reached the cap in Table 30.

External task metrics. micro-F1 pools true and false positives over the notes, and macro-F1 averages per-label F1 over the labels that occur in the gold sets or the predictions. Head and tail macro-F1 split the labels present in the gold sets at the median training-split frequency among them (tail: at most the median). P@k and R@k ($k = 8, 15$) rank the emitted codes in emission order and compute precision over k, so short answers are penalised. PSP@k and PSnDCG@k (Jain et al., 2016) weight each correct label by its inverse propensity, estimated from `MIMIC-IV` training-split label frequencies with the usual $A = 0.55$ and $B = 1.5$, and are normalised by the best value achievable on the note. The gap to the teacher recovered is (arm − base)/(teacher − base), with the teacher's test micro-F1 0.227 on the same prompts. These follow the ICD-coding and extreme-classification conventions (Mullenbach et al., 2018; Edin et al., 2023; Bhatia et al., 2016).

Fidelity to the teacher. Teacher-trace PPL is the student's perplexity on the teacher's trace of each test note under teacher forcing, median over notes. BLEU-4 (Papineni et al., 2002) compares the arm's trace with the teacher's at sentence level (whitespace tokens, brevity penalty, no smoothing), mean over notes. LAS is the leakage-adjusted simulatability used by Wang et al. (2023b), adapted to multi-label coding. One simulator, `Qwen3-1.7B` without thinking, is fine-tuned once on the teacher's training traces to predict the teacher's final answer from the note with the rationale and, mixed in at equal weight, from the note alone. Per note, $\mathrm{LAS} = \mathrm{F1}\big(\mathrm{sim}(\text{note} \oplus r), a_T\big) - \mathrm{F1}\big(\mathrm{sim}(\text{note}), a_T\big)$, with r the text inside the arm's reasoning block and a_T the teacher's answer. The simulator does not see the shortlist, which would add 30k tokens, so its absolute F1 is low and only the difference is read; the note-only term is shared by every arm.

M1, the supervised readout. M1 (decision-margin agreement) is read at the teacher's P1, P2 and P2C events on the teacher's test traces under teacher forcing: the student's margin for the event's target against its competitors, with the events re-derived in the student's tokenizer (chunk-level decisions for `SmolLM3-3B`, whose tokenizer merges digit runs), reported as $P(\text{margin} > 0)$ and as the median margin per kind. NULL events are a control and stay near zero for every arm (0.011 to 0.019). Tables report the share over all events of a kind; paired tests use each note's share, because events cluster by note. Every arm's cross-entropy acts on these tokens, and MISTILL's interaction term matches the pooled heads' share c of exactly this margin (Eq. 5), so M1 is the readout the objective supervises.

M2–M5 and the prefix share, on each arm's own heads. The attribution pass of Sec. 2.3 is run on each trained checkpoint over the 100 validation traces and gives that arm's own calibrated ranking per phase; every read below is then made on the 150 test notes, teacher-forced on the teacher's traces.

- *M4, causal recovery.* The arm's own top-k ($k = 8, 40$) is knocked into a donor note's context at matched events, and the recovery of Sec. 2.3, $(\text{margin}_{\text{set}} - \text{margin}_{\text{corr}})/(\text{margin}_{\text{all}} - \text{margin}_{\text{corr}})$, is averaged over the note's events; tables report the median over notes, beside the same measurement for a count-matched random set. Necessity at $k = 40$ is the margin drop in nats when the set is mean-ablated; it scales with the clean margin, so it is reported, not interpreted, and never compared across hardware.
- *M2, region lift.* For the arm's own top-16 heads at a phase: the share of their attention on the shortlist lines that falls on the emitted code's chapter (its own line, its category and sibling lines, the rest of its chapter), divided by that region's share of the shortlist tokens, so that 1 is chance. Per head it is the median over the note's events, averaged over the 16 heads; tables report the median over notes, beside 16 random heads, at P1 and P2C.

- *M3, near-miss loop-back.* The share of the own top-16 P2C heads' attention on the earlier mentions of the near-miss, against their share on the code's own earlier mentions, for the group and its top head.
- *M5, locus distance.* The forcing probe of Sec. 3.4 run on the student: the median distance in tokens from the locus to the chapter letter, and the shares decided at the event, in the window and inherited. Pre-registered for three arms: the untrained `Qwen3-4B`, its fp32 Vanilla CoT-KD seed 0 and its MISTILL seed 0.
- *Prefix share* (not pre-registered; added while reading the controls). The share of the own top-16 P2C heads' attention on the code's already-emitted prefix in the trace.

These heads are found after training, by the trained model's own attribution, and the quantities are not the ones the loss matches; the loss names only the pools fixed before training. The two overlap: in every arm the trained student's own top-40 shares most of its heads with the untrained student's top-40 (Appendix L.2), which is MISTILL's supervised pool. What makes M4 non-circular is therefore the re-selection and the measurement — restoring a decision by writing the heads' outputs into another note's context, on notes the loss never saw — not disjoint heads.

Floors, medians and means. The random floors of M4 differ by student: at $k = 40$ the untrained `Qwen3-4B`'s are 0.043 (P1) and 0.030 (P2C), the untrained `SmolLM3-3B`'s 0.134 and 0.107. Absolute recovery is therefore not comparable across students, and every M4 value is printed beside its floor. The random sets are redrawn for every bin; every own-heads cell ran the bins $k = 8$ and 40, the teacher's and the 20-note cells of the baseline campaign five bins, so floors differ by draw while the reported cells do not. M4 values in the tables are medians over notes of per-note means. Paired comparisons use the per-note means and report the median of the per-note differences; where the text quotes a mean over notes (the `SmolLM3-3B` seed spreads, the concentration at $k = 8$), it says so. Task and fidelity values are means over seeds of the corpus-level metric, with the between-seed standard deviation.

Statistics. Task and fidelity differences are paired on the 150 notes, with the seed-averaged per-note value as the unit: a percentile bootstrap 95% interval of the mean difference (1,000 resamples) and a two-sided Wilcoxon signed-rank p (zero differences dropped). For micro-F1 the per-note value is the note's own F1, so these intervals are for the mean per-note F1 difference, a different estimand from the difference of the corpus-level micro-F1 values in Table 2; no note-resampled interval of the corpus-level difference is reported. 150 paired notes resolve a per-note micro-F1 difference of about 0.025 at 80% power. Mechanism differences (M4, M2, and M1 per note) use two-sided sign tests on the per-note values: within a seed pair, and pooled over seed-matched pairs (seed i of one arm against seed i of the other; 450 note-pairs for three pairs, 300 for two), with Clopper–Pearson intervals on the win rate where given. The pooled counts treat a note's seed pairs as separate trials and are not clustered by note, so they overstate precision relative to a note-clustered test, which is not reported; the within-pair tests quoted in the text are the stricter reading. On `Qwen3-4B` a single-seed comparison is uninterpretable: two fp32 Vanilla CoT-KD seeds separate with a highly significant sign test (seed 2 minus seed 1, +0.040 at P2C, higher on 112 of 150 notes, p 10^{-9}), hence the seed matching and pooling.

Readouts that do not discriminate ([R-D15]). Four readouts are reported for completeness and not used to rank arms. *M4 at P1* is at ceiling in every arm, the untrained students included (0.946 and 0.985, against the teacher's 0.915): chapter selection is not what distillation changes. *M3* is present before distillation and barely moves on `Qwen3-4B` (the group's near-miss share lies between 0.040 and 0.082 across the twenty-three arms read, and one head, `L24H21`, is the top loop-back head in 17 of them), while on `SmolLM3-3B` it falls with distillation (0.081 untrained, 0.058 for Vanilla CoT-KD seed 0, 0.025 for MISTILL) as micro-F1 rises from 0.003 to 0.148: a quantity that moves the wrong way across that improvement does not measure what the task depends on. *M5* places the chapter decision 7–8 tokens before the letter in all three students read (untrained 8, Vanilla CoT-KD 7, MISTILL 7), against the teacher's 9. *M2 on `SmolLM3-3B`* follows which heads are supervised rather than recovery: MISTILL and RandPos share a P2C lift (2.65 and 2.52) at recoveries of 0.789 and 0.304. Across the ablations neither M2 nor the prefix share follows recovery on either student (Appendix L.2), so the claims rest on M4 at P2C.

K EXTENDED DISTILLATION RESULTS

Supports: Secs. 4.3 and 4.4; [R-D1]–[R-D6], [R-D11], [R-D12], [R-D15]–[R-D17]

K.1 PAIRED COMPARISONS AND PER-SEED ROWS

Table 28 gives the paired differences behind Table 2 against both references, fp32 Vanilla CoT-KD and Vanilla CoT-KD trained with bf16 weights like MISTILL; Table 30 gives every seed with its generation behaviour; and Table 29 gives the own-heads readouts of Table 3 for every seed, with the seed-matched comparisons Section 4.4 cites. For calibration: supervised classifiers trained on the release's full 89,099-note training split reach micro-F1 0.585 on its combined diagnosis-and-procedure task (PLM-ICD, ten seeds; Edin et al., 2023) — a different regime (full supervision, classification heads, all 7,942 labels) from the zero-shot teacher and few-hundred-note distillation studied here, quoted so the absolute micro-F1 level can be placed. Against fp32 Vanilla CoT-KD, DistiLLM-2 (fp32) is indistinguishable on the task and Feature-KD (fp32) is lower; OPD, which trains bf16 weights, is also indistinguishable on the task, while MISTILL and bf16 Vanilla CoT-KD, both off-policy with bf16 weights, are lower by similar amounts, and MISTILL trained with fp32 master weights is lower too (−0.024 [−0.039, −0.009]). Against bf16 Vanilla CoT-KD, the precision-matched reference, MISTILL is level on micro-F1, PSP@15 and LAS on both students. Plain Vanilla CoT-KD recovers 0.68 of the gap between the untrained `Qwen3-4B` and the teacher from 493 notes, and turns a cross-family student that could not finish a trace into a working one. Its cost shows in validity: at fp32 a sixth of its `Qwen3-4B` generations run to the cap in a repetition loop (Table 30), where the untrained student almost never does; bf16 Vanilla CoT-KD runs to the cap less often, and MISTILL is at its level. The MISTILL deficit against fp32 Vanilla CoT-KD lies in the completed answers, at both precision (0.243 against 0.296) and recall (0.108 against 0.137), and bf16 Vanilla CoT-KD shows the same deficit (0.239, 0.102). On `SmolLM3-3B` the untrained student opens a reasoning block on every note but rarely closes it (valid 0.153); fp32 Vanilla CoT-KD teaches it to stop on about half the notes (0.547), while bf16 Vanilla CoT-KD (0.380), MISTILL (0.297) and Feature-KD (0.360) leave most generations at the cap.

K.2 TWO DISSOCIATIONS: OPD AND THE GOLD NULL

OPD matches fp32 Vanilla CoT-KD on the answer and separates from it wherever the teacher is the reference. Its micro-F1 is 0.179 (against fp32 Vanilla CoT-KD −0.005 [−0.025, +0.015], p 0.67), but its LAS is lower (0.229, −0.038 [−0.065, −0.011], p 0.006), its teacher-trace PPL higher (1.638 against 1.516) and its M1 lower at P1 and P2C (0.865 and 0.770, against 0.888 and 0.791). Each of these is measured against the teacher, and each says the student has moved off the teacher's trace distribution, which is what training on its own rollouts is for. Its BLEU-4 is the table's highest (0.200 against 0.173), while its LAS is significantly lower: the traces are more teacher-like token for token and carry less of the decision. A comparison on surface fidelity alone would rank OPD first; LAS, which asks whether a frozen simulator recovers the answer from the rationale, ranks it below fp32 Vanilla CoT-KD. The comparison mixes precisions, since OPD trains bf16 weights: against bf16 Vanilla CoT-KD OPD is higher on BLEU-4 and on LAS alike (+0.031 [+0.022, +0.040] and +0.046 [+0.021, +0.073]; Table 28), and its M1 is level with bf16 Vanilla CoT-KD's (0.863 and 0.773 at P1 and P2C), so the dissociation holds against the fp32 reference only. It is still why LAS is kept in the metric set: a trace can be closer to the teacher's token for token and carry less of its decision. The compensation is validity: OPD is the most often valid distilled arm (0.973, cap 0.027), because the loops that cost the off-policy arms a sixth to a quarter of their answers are, for an on-policy objective, training data. Its tail macro-F1 is nominally the table's best (0.063 against 0.059), at one seed; no weight is put on it.

GOLD scores zero, and the mechanism metrics locate why. GOLD (bf16 weights, like OPD) scores micro-F1 0.000: no note yields a single scored code. The policy is not degenerate: 0.613 of its generations close a well-formed answer block, the reasoning is fluent and on topic and names correct billable codes in prose, and its teacher-trace PPL (2.132, the table's highest) and BLEU-4 (0.116) are finite. Its final answers list three-character ICD-10-CM categories (N18, say, where a billable code such as N18.9 is required), and every such entry fails inventory validation. M1 places the failure at Phase 2. Against the untrained `SmolLM3-3B` (P1 0.731, P2 0.379, P2C 0.721), GOLD improves the

Table 28: Paired differences behind Table 2: arm minus reference on the same 150 test notes, seed-averaged per note (for micro-F1, the mean per-note F1 difference, not the difference of Table 2's corpus-level values), with percentile bootstrap 95% intervals (1,000 resamples) and Wilcoxon signed-rank p; references fp32 Vanilla CoT-KD and bf16 Vanilla CoT-KD; OPD and GOLD train bf16 weights. [R-D1][R-D2][R-D3][R-D4][R-D5][R-D17]

Arm	Δ per-note F1 [95% CI]	Δ PSP@15 [95% CI]	Δ LAS [95% CI]
`Qwen3-4B`			
against Vanilla CoT-KD			
Base	−0.081 [−0.101, −0.063] p 3×10^{-11}	−0.064 [−0.080, −0.050] p 5×10^{-12}	−0.174 [−0.206, −0.142] p 3×10^{-18}
Vanilla CoT-KD (bf16)	−0.041 [−0.056, −0.026] p 4×10^{-6}	−0.037 [−0.050, −0.025] p 8×10^{-7}	−0.084 [−0.105, −0.065] p 10^{-12}
Feature-KD	−0.035 [−0.049, −0.021] p 2×10^{-5}	−0.034 [−0.045, −0.022] p 3×10^{-7}	−0.054 [−0.074, −0.035] p 3×10^{-8}
DistiLLM-2	+0.003 [−0.016, +0.021] p 0.57	+0.001 [−0.015, +0.017] p 0.80	+0.004 [−0.027, +0.032] p 0.37
OPD	−0.005 [−0.025, +0.015] p 0.67	−0.007 [−0.023, +0.010] p 0.37	−0.038 [−0.065, −0.011] p 0.006
MISTILL	−0.034 [−0.050, −0.018] p 4×10^{-5}	−0.030 [−0.044, −0.016] p 3×10^{-5}	−0.075 [−0.095, −0.056] p 10^{-11}
MISTILL (fp32)	−0.024 [−0.039, −0.009] p 0.001	−0.023 [−0.035, −0.010] p 0.001	−0.039 [−0.058, −0.021] p 4×10^{-5}
against Vanilla CoT-KD (bf16)			
Base	−0.040 [−0.056, −0.023] p 5×10^{-6}	−0.027 [−0.039, −0.015] p 2×10^{-5}	−0.090 [−0.115, −0.062] p 5×10^{-11}
Vanilla CoT-KD	+0.041 [+0.026, +0.056] p 4×10^{-6}	+0.037 [+0.025, +0.050] p 8×10^{-7}	+0.084 [+0.065, +0.105] p 10^{-12}
Feature-KD	+0.006 [−0.007, +0.021] p 0.57	+0.004 [−0.008, +0.015] p 0.57	+0.030 [+0.012, +0.048] p 0.001
DistiLLM-2	+0.044 [+0.024, +0.063] p 4×10^{-5}	+0.038 [+0.021, +0.054] p 3×10^{-5}	+0.088 [+0.060, +0.116] p 10^{-8}
OPD	+0.037 [+0.018, +0.055] p 4×10^{-4}	+0.030 [+0.014, +0.046] p 4×10^{-4}	+0.046 [+0.021, +0.073] p 0.001
MISTILL	+0.007 [−0.005, +0.019] p 0.27	+0.007 [−0.004, +0.019] p 0.13	+0.009 [−0.007, +0.025] p 0.30
MISTILL (fp32)	+0.017 [+0.004, +0.030] p 0.02	+0.015 [+0.004, +0.027] p 0.01	+0.045 [+0.025, +0.066] p 6×10^{-5}
`SmolLM3-3B`			
against Vanilla CoT-KD			
Base	−0.112 [−0.130, −0.094] p 3×10^{-19}	−0.083 [−0.097, −0.069] p 3×10^{-19}	−0.244 [−0.273, −0.217] p 8×10^{-23}
Vanilla CoT-KD (bf16)	−0.069 [−0.092, −0.047] p 2×10^{-8}	−0.054 [−0.072, −0.038] p 4×10^{-9}	−0.145 [−0.179, −0.112] p 5×10^{-12}
Feature-KD	−0.060 [−0.080, −0.038] p 2×10^{-7}	−0.046 [−0.062, −0.028] p 3×10^{-7}	−0.109 [−0.142, −0.077] p 2×10^{-9}
GOLD (6,144 cap)	−0.115 [−0.133, −0.097] p 3×10^{-19}	−0.085 [−0.100, −0.072] p 3×10^{-19}	−0.277 [−0.305, −0.251] p 7×10^{-24}
MISTILL	−0.071 [−0.089, −0.053] p 10^{-12}	−0.055 [−0.069, −0.041] p 4×10^{-12}	−0.145 [−0.174, −0.117] p 10^{-16}
against Vanilla CoT-KD (bf16)			
Base	−0.043 [−0.058, −0.028] p 2×10^{-6}	−0.029 [−0.040, −0.019] p 6×10^{-8}	−0.100 [−0.128, −0.068] p 3×10^{-9}
Vanilla CoT-KD	+0.069 [+0.047, +0.092] p 2×10^{-8}	+0.054 [+0.038, +0.072] p 4×10^{-9}	+0.145 [+0.112, +0.179] p 5×10^{-12}
Feature-KD	+0.009 [−0.011, +0.029] p 0.37	+0.009 [−0.006, +0.023] p 0.22	+0.036 [+0.001, +0.071] p 0.05
GOLD (6,144 cap)	−0.046 [−0.060, −0.032] p 10^{-7}	−0.031 [−0.041, −0.021] p 10^{-11}	−0.132 [−0.163, −0.104] p 2×10^{-13}
MISTILL	−0.002 [−0.019, +0.016] p 0.94	−0.001 [−0.013, +0.012] p 0.82	−0.001 [−0.034, +0.032] p 0.68

Table 29: Per-seed rows and seed-matched comparisons behind Table 3 (150 test notes; medians over notes; random-40 floors in parentheses; task: that seed's micro-F1). Seed-matched lines: per-note differences pooled over the seed pairs (median; note-pairs higher / lower; sign test), M2 the same over per-note lifts, and task the paired mean difference with its 95% interval. [R-D6][R-D12][R-D15][R-D17]

	M4 knock-in, own top-40			read side (P2C)		task
Arm	P2C (random)	P2	P1	M2 lift (random)	prefix share	micro-F1
Teacher (`Qwen3-14B`) HD-002, 100 val. notes	0.847 (0.015)	0.821	0.915	3.38 (1.95)	0.260	—
`Qwen3-4B`						
Base	0.788 (0.030)	0.815	0.946	2.56 (1.17)	0.185	0.104
Vanilla CoT-KD s0	0.840 (0.035)	0.749	0.945	3.23 (1.46)	0.196	0.193
s1	0.848 (0.038)	0.763	0.948	3.30 (1.28)	0.193	0.189
s2	0.891 (0.035)	0.664	0.943	3.81 (1.26)	0.236	0.180
Vanilla CoT-KD (bf16) s0	0.834 (0.037)	0.751	0.952	3.36 (1.53)	0.187	0.136
s1	0.835 (0.036)	0.758	0.951	3.60 (1.60)	0.188	0.148
s2	0.847 (0.037)	0.764	0.952	3.47 (1.62)	0.188	0.147
Feature-KD	0.844 (0.048)	0.888	0.962	3.05 (1.57)	0.228	0.142
MISTILL s0	0.873 (0.025)	0.736	0.956	3.65 (1.56)	0.264	0.151
s1	0.870 (0.025)	0.736	0.955	3.67 (1.55)	0.265	0.147
s2	0.874 (0.026)	0.734	0.958	3.57 (1.70)	0.277	0.150
RandHeads	0.869 (0.032)	0.880	0.958	3.66 (2.21)	0.290	0.133
RandPos	0.631 (0.042)	0.771	0.907	2.93 (1.45)	0.076	0.129
MISTILL − Vanilla CoT-KD seed-matched, 3 pairs	+0.011 250/200 p 0.02	+0.011 254/196 p 0.007	+0.011 319/131 p 4×10^{-19}	+0.113 273/177 p 7×10^{-6}	—	−0.034 [−0.050, −0.018]
MISTILL − Vanilla CoT-KD (bf16), seed-matched, 3 pairs	+0.032 304/146 p 8×10^{-14}	−0.019 129/321 p 6×10^{-20}	+0.006 286/164 p 10^{-8}	+0.268 340/110 p 2×10^{-28}	—	+0.007 [−0.005, +0.019]
`SmolLM3-3B`						
Base	0.416 (0.107)	0.809	0.985	3.57 (1.51)	0.121	0.003
Vanilla CoT-KD s0	0.541 (0.105)	0.735	0.975	3.75 (1.64)	0.165	0.148
s1	0.523 (0.098)	0.714	0.978	3.87 (1.34)	0.180	0.113
Vanilla CoT-KD (bf16)	0.476 (0.098)	0.867	0.985	3.29 (1.39)	0.151	0.055
Feature-KD	0.628 (0.099)	0.933	0.988	2.74 (1.37)	0.097	0.062
MISTILL s0	0.789 (0.121)	0.866	1.004	2.65 (1.31)	0.439	0.060
s1	0.774 (0.127)	0.868	1.006	2.65 (1.37)	0.441	0.042
RandHeads	0.585 (0.116)	0.725	0.999	3.93 (2.01)	0.284	0.061
RandPos	0.304 (0.095)	0.891	0.993	2.52 (1.36)	0.050	0.063
MISTILL − Vanilla CoT-KD seed-matched, 2 pairs	+0.248 257/43 p 3×10^{-38}	+0.137 273/27 p 2×10^{-52}	+0.028 282/18 p 4×10^{-62}	−1.279 1/299 p 3×10^{-88}	—	−0.071 [−0.089, −0.053]
MISTILL − Vanilla CoT-KD (bf16), seed-matched, 1 pair	+0.317 130/20 p 6×10^{-21}	+0.009 86/64 p 0.09	+0.017 115/35 p 4×10^{-11}	−0.656 5/145 p 9×10^{-37}	—	−0.002 [−0.019, +0.016]

Table 30: Per-seed rows behind Table 2, with generation behaviour (150 test notes). Valid: share of generations with a parsed final answer; cap: share that reached the 16,384-token cap; tokens: mean generated tokens. (bf16) rows train bf16 weights without an fp32 master copy, as MISTILL does; so do OPD and GOLD (TRL). GOLD ran with a 6,144-token rollout cap. [R-D1][R-D2][R-D3][R-D4][R-D5][R-D17]

Arm	seed	micro-F1	micro-P	micro-R	valid	cap	tokens	LAS	M1 P1	M1 P2	M1 P2C
`Qwen3-4B`											
Base	—	0.104	0.180	0.073	0.987	0.013	4,329	0.093	0.813	0.660	0.759
Vanilla CoT-KD	0	0.193	0.303	0.141	0.847	0.153	7,771	0.271	0.888	0.746	0.794
	1	0.189	0.306	0.137	0.793	0.207	8,149	0.264	0.887	0.746	0.791
	2	0.180	0.280	0.132	0.833	0.167	7,634	0.266	0.888	0.745	0.790
Vanilla CoT-KD (bf16)	0	0.136	0.227	0.097	0.900	0.100	6,876	0.179	0.864	0.713	0.773
	1	0.148	0.251	0.105	0.887	0.113	7,383	0.177	0.862	0.713	0.774
	2	0.147	0.239	0.106	0.907	0.093	7,191	0.192	0.862	0.713	0.773
Feature-KD	0	0.142	0.266	0.097	0.753	0.247	9,017	0.194	0.873	0.727	0.772
	1	0.163	0.281	0.115	0.833	0.167	8,062	0.217	0.871	0.727	0.773
	2	0.152	0.281	0.104	0.753	0.247	8,424	0.227	0.874	0.724	0.768
DistiLLM-2	0	0.187	0.287	0.139	0.880	0.120	7,285	0.271	0.891	0.736	0.788
OPD	0	0.179	0.275	0.132	0.973	0.027	6,991	0.229	0.865	0.718	0.770
MISTILL	0	0.151	0.235	0.111	0.907	0.093	6,886	0.201	0.857	0.712	0.802
	1	0.147	0.247	0.105	0.887	0.113	7,121	0.191	0.857	0.712	0.804
	2	0.150	0.247	0.108	0.887	0.113	7,128	0.185	0.859	0.715	0.804
`SmolLM3-3B`											
Base	—	0.003	0.043	0.002	0.153	0.847	14,147	−0.019	0.731	0.379	0.721
Vanilla CoT-KD	0	0.148	0.306	0.097	0.593	0.407	10,232	0.237	0.880	0.592	0.789
	1	0.113	0.285	0.070	0.500	0.500	11,201	0.214	0.879	0.591	0.786
Vanilla CoT-KD (bf16)	0	0.055	0.222	0.031	0.380	0.620	12,312	0.081	0.844	0.535	0.756
Feature-KD	0	0.062	0.245	0.036	0.360	0.640	12,471	0.117	0.861	0.560	0.746
GOLD	0	0.000	0.000	0.000	0.613	0.387	10,794	−0.051	0.832	0.355	0.732
MISTILL	0	0.060	0.246	0.034	0.320	0.680	13,044	0.081	0.845	0.538	0.741
	1	0.042	0.213	0.023	0.273	0.727	13,434	0.080	0.845	0.537	0.741

chapter decision (0.832) and leaves the contrastive one near the untrained level (0.732), but its P2 agreement falls below the untrained student's, to 0.355 with a negative median margin (−1.04 nats), where Vanilla CoT-KD seed 0 on the same student reaches 0.592 (median +1.08 nats). P2 events are exactly the choices of a billable code within a category, so the pre-registered phase-2 readout reports the failure that the category-level answers make visible: task and mechanism fail together and at the same phase, the strongest internal evidence available that the null is real rather than a scoring artefact. The cause is the run's cap: the teacher's median trace is about 6.1k tokens with a 90th percentile near 9.7k, so truncating the teacher's completion to 6,144 tokens removed the end of the longer traces, where the answer block and its full-code emissions are, on 200 of the 500 training traces, and the arm learned the truncated behaviour. Its training losses stayed healthy to the last step, and the loss-based collapse monitor that had caught the fully on-policy run reported health throughout; a generation-level probe on a few notes during training is what would have caught it (it was added to the trainer afterwards and never fired on a MISTILL-family run). The cross-family on-policy cell is therefore unresolved rather than refuted: what is measured is GOLD under a 6,144-token rollout cap, and the row carries that cap wherever it is quoted. A clean rerun needs more memory than a 141 GB GPU offers at the teacher's trace lengths, or memory work in the hybrid loss; it was not attempted. The same-family on-policy cell (OPD, no rollout ever capped) carries the on-policy comparison.

K.3 THE PRECISION CONTROL

MISTILL, its controls and the on-policy baselines (OPD, GOLD) train bf16 weights without an fp32 master copy; Vanilla CoT-KD, Feature-KD and DistiLLM-2 keep fp32 master weights (Appendix I.4). The control trains Vanilla CoT-KD with bf16 weights on the fp32 Vanilla CoT-KD recipe (493 notes, two epochs, thinking mode), three `Qwen3-4B` seeds (top-32 KL, YaRN ×2) and one `SmolLM3-3B` seed (ULD), so that each control's dumped configuration differs from MISTILL's in the objective only. Each is evaluated on the 150 test notes and read on its own heads with the instruments of Appendix J. The reading was fixed before the result: bf16 Vanilla CoT-KD at MISTILL's task level means the tier-1 task gap to the fp32 baselines is the precision's; at fp32 Vanilla CoT-KD's level, the objective's; between, both.

Table 31: Precision control, task side (150 test notes; mean ± sd over seeds). Δ: paired micro-F1 difference, arm minus the reference's seed-averaged notes, with its 95% bootstrap interval. [R-D17]

Arm	weights	seeds	micro-F1	LAS	valid	M1 P2C	Δ vs MISTILL	Δ vs fp32 Vanilla CoT-KD
`Qwen3-4B`								
Vanilla CoT-KD	fp32	3	0.187 ± 0.007	0.267 ± 0.004	0.824	0.791	+0.034 [+0.018, +0.050]	—
Vanilla CoT-KD	bf16	3	0.143 ± 0.007	0.183 ± 0.008	0.898	0.773	−0.007 [−0.019, +0.005]	−0.041 [−0.056, −0.026]
MISTILL	bf16	3	0.149 ± 0.002	0.192 ± 0.008	0.893	0.803	—	−0.034 [−0.050, −0.018]
MISTILL	fp32	3	0.165 ± 0.004	0.228 ± 0.014	0.824	0.779	+0.010 [−0.005, +0.025]	−0.024 [−0.039, −0.009]
`SmolLM3-3B`								
Vanilla CoT-KD	fp32	2	0.130 ± 0.025	0.226 ± 0.016	0.547	0.787	+0.071 [+0.053, +0.089]	—
Vanilla CoT-KD	bf16	1	0.055	0.081	0.380	0.756	+0.002 [−0.016, +0.019]	−0.069 [−0.092, −0.047]
MISTILL	bf16	2	0.051 ± 0.013	0.080 ± 0.000	0.297	0.741	—	−0.071 [−0.089, −0.053]

Table 32: Training cross-entropy on the trace tokens at optimiser steps 5, 20, 60 and 120 of 124, per seed. bf16 Vanilla CoT-KD tracks MISTILL step for step, and fp32 Vanilla CoT-KD falls below both from about step 20. [R-D17]

Student	Run (seeds)	weights	step 5	step 20	step 60	step 120
`Qwen3-4B`	Vanilla CoT-KD (0 / 1 / 2)	fp32	0.553 / 0.543 / 0.548	0.461 / 0.469 / 0.466	0.439 / 0.431 / 0.430	0.376 / 0.384 / 0.387
	Vanilla CoT-KD (0 / 1 / 2)	bf16	0.560 / 0.550 / 0.554	0.477 / 0.487 / 0.483	0.469 / 0.463 / 0.460	0.446 / 0.454 / 0.456
	MISTILL (0)	bf16	0.560	0.484	0.475	0.453
`SmolLM3-3B`	Vanilla CoT-KD (0 / 1)	fp32	0.660 / 0.645	0.541 / 0.555	0.506 / 0.501	0.424 / 0.431
	Vanilla CoT-KD (0)	bf16	0.667	0.561	0.545	0.517
	MISTILL (0)	bf16	0.668	0.565	0.548	0.516

The task: the precision's, on both students. bf16 Vanilla CoT-KD lands at MISTILL's task level, not fp32 Vanilla CoT-KD's (Table 31): its three `Qwen3-4B` seeds (0.136, 0.148, 0.147) all sit far below fp32 Vanilla CoT-KD and around MISTILL. Paired per note and pooled over the three `Qwen3-4B` seed pairs, MISTILL minus bf16 Vanilla CoT-KD splits the task 205 higher and 184 lower (61 ties; p 0.31), while M1 at P2C favours MISTILL on 348 note-pairs against 29 ($p < 10^{-4}$, mean +0.035): the supervised readout's lead holds, and is larger, at matched precision. At P1 the pooled comparison leans the other way (65 higher, 131 lower, mean −0.006), and at P2 it is not significant. On `SmolLM3-3B`, one seed pair, the task splits 32 higher and 26 lower (p 0.51) and M1 at P2C 36 higher and 82 lower (mean −0.013): as against fp32 Vanilla CoT-KD, MISTILL's advantage on `SmolLM3-3B` is not in M1; on its own heads it holds, and grows (below). The training loss agrees (Table 32). What bf16 training costs against fp32 master weights is −0.041 micro-F1 on `Qwen3-4B` and −0.069 on `SmolLM3-3B` for Vanilla CoT-KD, about the size of MISTILL's gap to fp32 Vanilla CoT-KD. MISTILL does not share that gain: trained with fp32 master weights it moves +0.010 [−0.005, +0.025] from its bf16 runs and trails fp32 Vanilla CoT-KD by −0.024 [−0.039, −0.009] on the task and −0.039 on LAS, with M1 at P2C 0.779 against fp32 Vanilla CoT-KD's 0.791. With fp32 master weights the objective costs task accuracy; its λ was selected with bf16 weights and not re-selected.

The mechanism side. On the students' own heads (Table 3), MISTILL's gain over bf16 Vanilla CoT-KD at P2C is +0.032 on `Qwen3-4B` (pooled over three seed pairs: 304 higher, 146 lower), against +0.011 over fp32 Vanilla CoT-KD, and +0.317 on `SmolLM3-3B` (one pair), against +0.248: at matched precision the contrastive gain is larger. At P2 on `Qwen3-4B` it reverses (−0.019, lower on 321 of 450). bf16 Vanilla CoT-KD's three seeds recover 0.834, 0.835 and 0.847 at P2C (spread 0.013, against MISTILL's 0.004 and fp32 Vanilla CoT-KD's 0.052), and two of their three pairs separate per note where none of MISTILL's do.

K.4 EXTRA TIER-1 CELLS

A lighter Phase-1 weight. $\lambda_{P1} = 0.03$ with $\lambda_{P2} = 0.1$ (`Qwen3-4B`, one seed) changes nothing: micro-F1 0.152 (paired against MISTILL's seed-averaged notes +0.001 [−0.015, +0.017]) and M1 0.860 / 0.714 / 0.804 against seed 0's 0.857 / 0.712 / 0.802; the Phase-1 term has no measurable effect at this dose. The phase ablations of Appendix L locate the rest.

A same-seed replicate on other hardware. MISTILL seed 2 retrained with the same configuration on a different GPU type gives micro-F1 0.155 against 0.150 (paired 60 higher, 64 lower, p 0.79) and M1 at P2C 0.802 against 0.804. The one-seed noise at tier-1 size is therefore about 0.005 micro-F1, 0.003 M1 at P2C and 0.02 LAS. The replicate enters no mean.

`SmolLM3-3B` with the attention term only. The `SmolLM3-3B` MISTILL row is at bf16 Vanilla CoT-KD's task level (Appendix K.3); before the precision control, its low task score was suspected to come from the interaction term, whose target carries a cross-tokenizer offset (`SmolLM3-3B`'s P2 interaction loss sat near 1.5 throughout training). The arm was retrained with the attention term only, same epochs and λ: micro-F1 0.036 (valid 0.300, cap 0.700), no better than the both-terms seed 0 (0.060; paired 25 higher, 32 lower, p 0.43), with M1 0.844 / 0.535 / 0.736 against 0.845 / 0.538 / 0.741 and LAS 0.082 against 0.081, while its own P2 attention loss trained normally (from 0.99 to 0.24). The offset is real in the loss but not the cause of the row.

K.5 INSTRUMENT CHECKS

Knock-in without the attribution pass. On the test notes the own-heads knock-in runs without its first pass (the attribution over all heads, which the knock-in tables do not read), gated by rerunning the baseline campaign's 20-note Vanilla CoT-KD seed-0 cell that way. Every reported cell was reproduced, the largest difference being 0.014 (P2C at $k = 8$, the noisiest cell at 20 notes), with the random floors close. Raw per-event margins differed by a median of 0.045 nats, but so did the unpatched clean margins of the two runs, which ran on different GPU types: that is the forward pass's own cross-hardware floor, not the mode's. The gate's cell was not retained; its readout-level numbers are in the experiment logbook.

The instrument's cross-hardware floor. The `SmolLM3-3B` $k = 8$ own-heads cell was computed twice from one checkpoint, on an RTX PRO 6000 (the cell of record) and on an H200. The attribution ranking's top-40 is identical at every phase and its top-8 in the same order, and M4 differs per note by +0.002, +0.000 and +0.001 at P2C, P2 and P1 (sign-test p 0.57, 0.93 and 0.16); in median the two reads differ by 0.006. Every per-note difference read as an effect in Appendix L is several times larger.

K.6 READINGS WITHDRAWN

The baseline campaign read M4 on 20 test notes. At that size a sign test has about 25% power against a 0.65 win rate, and two readings did not survive the rerun on all 150. It read `SmolLM3-3B` Vanilla CoT-KD at 0.685 and Feature-KD at 0.677 at P2C (Vanilla CoT-KD at least Feature-KD); at 150 notes they are 0.541 and 0.628, and the order inverts (Feature-KD higher on 101 of 150 notes). And it read "the mechanism ordering matches the task ordering on `Qwen3-4B`" from Vanilla CoT-KD seed 0 above Feature-KD (0.881 against 0.814); at 150 notes the seed-matched pair does not separate (0.840 against 0.844; per note +0.001, 76/74, p 0.93), though Vanilla CoT-KD's seed 2 sits above Feature-KD (+0.044, 98/52, p 2×10^{-4}). The 20-note cells were labelled descriptive; the 150-note cells replace them, and every mechanism number in the paper is on 150 notes.

L ABLATION DETAILS

Supports: Secs. 4.3 (training-set size) and 4.5; [R-D7], [R-D8], [R-D13], [R-D14]

The controls of Sec. 4.4 left three findings, each with a caveat they could not close: the ranked heads need the teacher's positions (one control seed per student); where the untrained student's recovery is low, random heads at the teacher's events yield only part of MISTILL's gain (was that the heads' ranking, or their number?); and MISTILL's seeds agree where Vanilla CoT-KD's do not (a property of the heads MISTILL supervises, or of supervising positions at all?). The ablations were built to close these caveats and were pre-registered before any result (Table 33); a tier on `SmolLM3-3B` (A′) was added the same day, also before its results. Every arm is one flag away from MISTILL and trained at tier-1 size (493 notes, two epochs, $\lambda = 0.1$ unless the arm varies it, seed 0 unless stated) with bf16 weights, as MISTILL is, so no comparison in this appendix depends on the precision question

of Appendix K.3. Each arm is read with the instruments of Appendix J on the 150 test notes: M4 knock-in at the arm's own ranked heads, M2 and the prefix share at P2C, and the task.

The reading rule. Fixed in advance: an arm whose P2C effect lies inside MISTILL's seed band and whose task effect lies inside the tier-1 noise band is reported as "no effect", not re-run, and every P2C comparison is paired per note against each MISTILL seed. As measured, MISTILL's three `Qwen3-4B` seeds give 0.873, 0.870 and 0.874 in median, a band narrower than the instrument's own spread (one checkpoint read on two GPUs differs by 0.006 in median; Appendix K.5). The rule is therefore applied through the paired tests, uniformly, as an operational reading rather than a pre-registered one: a difference is read as an effect only when it holds against every MISTILL seed (sign test $p < 0.05$ against each).

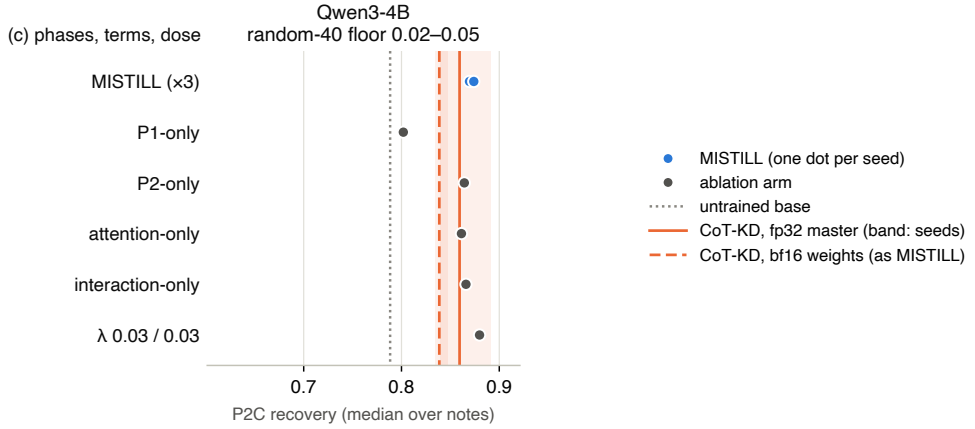


Figure 5: Figure 2, continued: (c) phases, terms and dose on `Qwen3-4B`. The P2C gain is the Phase-2 supervision's (P2-only keeps most of it, P1-only loses most of it); a third of the dose and either term alone leave it in place. All arms bf16, one seed unless marked. [R-D7][R-D14]

L.1 TIER BY TIER

Tier A: the positions, and whose heads. Both head-side expectations failed, and together they sharpen the reading of Sec. 4.4 rather than overturn it. RandBoth restores 0.863, with no paired difference from any MISTILL seed (Table 34) or from RandHeads (−0.006, p 0.46), and lies +0.221 above RandPos (higher on 140 of 150 notes). RandPos's collapse is therefore not "wrong positions" as such but the student's own ranked heads pointed at the wrong positions; random heads at random positions do no such harm. Bottom-40 falls below the untrained student: it restores 0.717 against 0.788, lower on 113 of 150 notes and below every MISTILL seed, so head identity has a floor even on the student where random heads can stand in for its top-ranked ones. The phases place the P2C gain in the Phase-2 supervision: P2-only restores 0.864 and P1-only 0.802 (P2-only minus P1-only +0.046, higher on 109 of 150). P2-only sits slightly below every MISTILL seed per note, outside MISTILL's seed band, so "keeps the gain" means about nine-tenths of it (P2-only minus the untrained student +0.073, against +0.075 to +0.083 for MISTILL's seeds). The task does not follow the P2C gain: P1-only scores 0.131 (−0.022 [−0.040, −0.001] against MISTILL) and P2-only 0.154 (+0.004 [−0.011, +0.020]).

Tier B: dose and terms. At a third of the weight neither the gain (0.880) nor the task (0.150, −0.002 [−0.021, +0.017] against MISTILL) moves: $\lambda = 0.1$ sits on a plateau, and a lighter dose does not buy task score back in this range. The term arms act on both phases, because the terms switch is global in the trainer; since the P1 term moves nothing at a third of its weight (Appendix K.4), this is close to switching the term at P2 alone. Attention-only restores 0.861 and interaction-only 0.866, no different from each other per note (p 0.37). Attention-only keeps MISTILL's gain over the untrained student (+0.078; below MISTILL only against seed 2) at a lower task score (0.130, −0.019 [−0.038, −0.002]); interaction-only sits below every MISTILL seed per note, keeping about four-fifths of the gain (+0.064 over the untrained student), at MISTILL's task score (0.157, +0.010 [−0.008, +0.027]).

Table 33: The ablations as pre-registered, before any result, with the outcome against each expectation (✓: as expected; ✗: not). Wording follows the pre-registration, except that bottom-40 is described by what its heads are (the 40 most negatively attributed) and "cost" means the tier-1 task gap to fp32 Vanilla CoT-KD, which Appendix K.3 attributes to bf16 training. [R-D7][R-D13][R-D14]

Tier	Arm (flag)	Question	Pre-registered expectation	Outcome
A: positions, not heads	RandBoth (`--heads random --positions random`)	Is RandPos's collapse the positions, or the combination?	≈ RandPos, below the untrained student	= MISTILL ✗
A	bottom-40 (`--heads bottom`)	The 40 most negatively attributed heads at the right events	≈ MISTILL / RandHeads; if it falls, head identity has a floor	below the untrained student: the floor branch
A	P1-only / P2-only ($\lambda_{P2} = 0$ / $\lambda_{P1} = 0$)	Which events carry the P2C gain?	P2-only keeps the gain and most of the cost; P1-only neither	P2-only ✓; P1-only loses the gain ✓ but keeps the cost ✗
B: the trade-off	λ 0.03 / 0.03	Does a lighter dose cost less task and keep the gain?	the gain shrinks with the cost, or not	not: neither moves
B	attention-only (`--terms attn`), interaction-only (`--terms inter`); both phases	Which term carries the gain, which the cost?	attention carries both; interaction adds the P2C margin	attention carries the cost; interaction alone keeps about four-fifths of the gain at no measurable cost ✗
C: reproducibility	RandHeads seed 1	Is the seed stability the positions' or MISTILL's heads'?	within 0.01 of RandHeads seed 0	fails (−0.033 at P2C, −0.089 at P2)
D: sensitivity	$k = 8$, $k = 100$	Pool size	within seed noise of MISTILL	✓ at P2C
A′: heads where the recovery must be built, on `SmolLM3-3B`	bottom-40 (two seeds); $k = 8$	Which heads, and how many, on the student where head identity matters	bottom-40 ≤ RandHeads (0.585); $k = 8$ near MISTILL if the top of the ranking carries it, between RandHeads and MISTILL if the pool size does	bottom-40 at or above MISTILL ✗; $k = 8$ between ✓

Table 34: Ablations on each arm's own heads (150 test notes; one seed unless marked). P2C and P2: M4 knock-in of the own top-40, median over notes, random-40 floor in parentheses. Δ P2C: per-note paired median difference against each MISTILL seed (`Qwen3-4B` seeds 0/1/2; `SmolLM3-3B` 0/1), sign-test p; vs base: against the untrained student, notes higher/lower. M2: P2C region lift (random). Prefix: share on the code's emitted prefix. micro-F1: with the paired difference from MISTILL's seed-averaged notes [95% interval]. All arms bf16. [R-D7][R-D13][R-D14]

Arm (tier)	P2C (random)	Δ P2C vs MISTILL, per seed (sign p)	vs base (higher/lower)	P2	M2 P2C lift (random)	prefix	micro-F1 (Δ vs MISTILL)
`Qwen3-4B`							
RandBoth (A)	0.863 (0.035)	−0.005 / −0.007 / −0.006 p 0.16 / 0.37 / 0.22	+0.071 119/31	0.738	2.69 (1.24)	0.139	0.138 −0.010 [−0.028, +0.007]
bottom-40 (A)	0.717 (0.049)	−0.173 / −0.175 / −0.180 p 2×10^{-24} / 3×10^{-26} / 2×10^{-24}	−0.088 37/113	0.851	3.68 (0.80)	0.146	0.114 −0.036 [−0.054, −0.020]
P1-only (A)	0.802 (0.033)	−0.057 / −0.055 / −0.062 p 7×10^{-8} / 10^{-10} / 3×10^{-9}	+0.028 89/61	0.794	3.40 (1.59)	0.189	0.131 −0.022 [−0.040, −0.001]
P2-only (A)	0.864 (0.026)	−0.006 / −0.006 / −0.008 p 8×10^{-4} / 0.02 / 2×10^{-4}	+0.073 121/29	0.732	3.53 (1.42)	0.283	0.154 +0.004 [−0.011, +0.020]
λ 0.03/0.03 (B)	0.880 (0.026)	+0.009 / +0.009 / +0.004 p 0.22 / 0.03 / 0.81	+0.091 130/20	0.748	3.75 (1.43)	0.233	0.150 −0.002 [−0.021, +0.017]
attention-only (B)	0.861 (0.039)	−0.008 / −0.009 / −0.012 p 0.29 / 0.09 / 0.02	+0.078 117/33	0.766	3.67 (1.71)	0.219	0.130 −0.019 [−0.038, −0.002]
interaction-only (B)	0.866 (0.023)	−0.010 / −0.010 / −0.015 p 10^{-5} / 8×10^{-4} / 5×10^{-5}	+0.064 119/31	0.712	3.24 (1.18)	0.285	0.157 +0.010 [−0.008, +0.027]
RandHeads s1 (C)	0.838 (0.040)	−0.031 / −0.027 / −0.028 p 0.001 / 10^{-5} / 10^{-6}	+0.051 110/40	0.791	4.10 (1.82)	0.244	$0.125^{(2\ \text{seed})}$ −0.023 [−0.037, −0.010]
$k = 8$ (D)	0.881 (0.029)	+0.007 / +0.008 / +0.007 p 0.12 / 0.01 / 0.29	+0.086 124/26	0.753	3.35 (1.65)	0.248	0.153 +0.005 [−0.013, +0.023]
$k = 100$ (D)	0.875 (0.030)	+0.006 / +0.006 / +0.001 p 0.37 / 0.09 / 0.81	+0.083 121/29	0.861	3.31 (1.76)	0.268	0.141 −0.008 [−0.025, +0.009]
`SmolLM3-3B`							
bottom-40 (A′)	0.841 (0.098)	+0.034 / +0.070 p 0.04 / 0.002	+0.409 141/9	0.857	2.45 (1.93)	0.279	$0.048^{(2\ \text{seed})}$ −0.004 [−0.018, +0.009]
bottom-40 s1 (A′)	0.835 (0.096)	+0.010 / +0.026 p 0.81 / 0.29	+0.407 137/13	0.862	2.46 (1.70)	0.277	$0.048^{(2\ \text{seed})}$ −0.004 [−0.018, +0.009]
$k = 8$ (A′)	0.671 (0.115)	−0.085 / −0.089 p 3×10^{-12} / 10^{-10}	+0.298 130/20	0.860	3.11 (1.29)	0.388	0.047 −0.004 [−0.020, +0.012]

Tier C: reproducibility. The pre-registered test fails. RandHeads' two seeds differ at P2C (seed 1 minus seed 0 −0.033, lower on 111 of 150, p 3×10^{-9}), at P2 (−0.089, lower on 147 of 150) and on the task (0.133 and 0.117), where MISTILL's seeds do not separate per note (seed 1 minus seed 0 +0.000, 75 higher and 75 lower). The seed stability tracks MISTILL's heads, the student's own ranked pool, rather than position supervision as such. Both RandHeads seeds train bf16 weights, as MISTILL's do, so bf16 training alone does not make seeds agree; it does narrow Vanilla CoT-KD's spread, from 0.052 to 0.013, but two of the three bf16 Vanilla CoT-KD seed pairs still separate per note (Appendix K.3).

Tier D: pool size (`Qwen3-4B`). Eight, forty or a hundred supervised heads give the same P2C recovery (0.881, 0.873 and 0.875; no paired difference from MISTILL holds against every seed). Only the widest pool lifts P2 (0.861, above MISTILL seed 0 on 150 of 150 notes), as RandHeads (0.880) and bottom-40 (0.851) also do; MISTILL's own P2 recovery (0.736) lies below the untrained student's (0.815).

Tier A′: which heads, and how many, where the recovery must be built. The untrained `SmolLM3-3B` restores 0.416 of the contrastive decision against a floor of 0.107, and on it both the identity and the number of the supervised heads matter. The base's 40 most negatively attributed heads yield at least as much as its top-40: 0.841 and 0.835 on two seeds against MISTILL's 0.789 and 0.774, seed 0 above both MISTILL seeds per note and seed 1 level with them (Table 34), both far above RandHeads (+0.278 and +0.262). The base's top-8 yield less (0.671: below both MISTILL seeds; +0.114 over RandHeads, higher on 104 of 150; −0.127 and −0.111 against bottom-40's seeds), and 40 random heads least (0.585). P2 does not separate the structured pools (bottom-40 0.857 and 0.862, $k = 8$ 0.860, MISTILL 0.866 and 0.868; $k = 8$ minus bottom-40 +0.008, p 0.57). Bottom-40's two seeds agree at P2 (+0.002, p 0.57), on the read side and in their carriers, but differ at P2C (−0.023, lower on 93 of 150, p 0.004), where MISTILL's do not (p 0.93): the pattern of tier C's RandHeads seeds on `Qwen3-4B`. Against the pre-registration, bottom-40 was expected at or below RandHeads; it is level with or above MISTILL, the highest cells on this student. $k = 8$ lands between RandHeads and MISTILL, the "pool size matters" branch, where on `Qwen3-4B` pool size did not matter. On the

task every A′ arm is at MISTILL's level (bottom-40 0.048 over two seeds, −0.004 [−0.018, +0.009]; $k = 8$ 0.047, −0.004 [−0.020, +0.012]), as is every attention-supervising arm on this student.

Task effects, read against the bf16 band. Every arm here trains bf16 weights, so task differences are read against bf16 Vanilla CoT-KD (0.143 ± 0.007 on `Qwen3-4B`, 0.055 on `SmolLM3-3B`), not as a cost against the fp32 baselines. The single-seed `Qwen3-4B` ablations span 0.114 (bottom-40) to 0.157 (interaction-only) around it, and the attention-supervising `SmolLM3-3B` arms, MISTILL's seeds included, span 0.042 to 0.063 around 0.055.

L.2 WHICH HEADS THE OBJECTIVE REACHES, AND WHERE THE MECHANISM LANDS

What bottom-40 is. `--heads bottom` gives the pool's weights to the 40 heads that the untrained student's calibrated ranking puts last, keeping the weight profile of the top-40 pool (as RandHeads does), so the arm differs from MISTILL in the ranking alone. The attribution is signed, so these are the most negatively attributed heads, whose contribution works against the decision, not heads without effect: at P2C on `SmolLM3-3B` their importances run from −0.035 to −0.334 (sum −3.66) against +0.048 to +0.319 (sum +4.49) for the top-40, with the median head at zero; on `Qwen3-4B` the sums are −2.66 against +3.71.

Where the gain lands. In every arm, on both students, the causal carriers after training are the untrained student's own top-ranked heads, whichever heads were supervised. Every one of the 29 arms read this way (10 on `SmolLM3-3B`, 19 on `Qwen3-4B`, controls and Vanilla CoT-KD at both precisions included) shares 27–38 of its own top-40 with the untrained student's top-40 at every phase. Bottom-40's own top-40 holds almost none of the 40 heads it supervised (0, 0 and 1 at P1, P2 and P2C on `SmolLM3-3B` seed 0, 0 at every phase on seed 1 and 0 on `Qwen3-4B`) and shares as many heads with the untrained student's top-40 as MISTILL's does, whose supervised heads are that top-40 (at P1 / P2 / P2C: on `SmolLM3-3B` 38 / 28 / 30 for seed 0 and 38 / 28 / 31 for seed 1, against MISTILL seed 0's 37 / 30 / 32; on `Qwen3-4B` 35 / 29 / 32 against MISTILL seed 0's 36 / 32 / 31); on `SmolLM3-3B` its own top-8 at P2C are the same eight heads on both seeds (the untrained student's ranks 0–4, 6, 9 and 18). `SmolLM3-3B` $k = 8$'s own top-40 holds all 8 of its supervised heads, which are the untrained student's top-8 (and 38 / 31 / 33 of the untrained student's top-40 at P1 / P2 / P2C). How strongly the recovery concentrates in those heads follows the size of the gain, on both students: knock-in at the arm's own top-8 restores a mean of 0.342 and 0.334 for `SmolLM3-3B` bottom-40's two seeds (recovery at $k = 40$ 0.841 and 0.835), against 0.232 and 0.228 for MISTILL and 0.170 for $k = 8$; on `Qwen3-4B`, bottom-40 (recovery 0.717) restores 0.165 against MISTILL's 0.605, lower on 150 of 150 notes. The supervised heads are an injection point, not the carriers.

The read side across the ablations. Across the arms of Sec. 4.4, M2 at P2C agreed with M4 on `Qwen3-4B` and the prefix share followed recovery on both students; the ablations break both. On `Qwen3-4B`, RandBoth restores 0.863 with an M2 lift of 2.69 (the untrained student 2.56, MISTILL 3.65) and a prefix share of 0.139, below the untrained student's 0.185, and bottom-40 restores 0.717 with MISTILL's lift (3.68). On `SmolLM3-3B`, M2 runs opposite to recovery across the three structured pools ($k = 8$ 3.11, MISTILL 2.65, bottom-40 2.45 at recoveries 0.671, 0.789, 0.841), and the prefix share puts bottom-40 (0.279) at RandHeads' level (0.284), below $k = 8$ (0.388) and MISTILL (0.439). M4 at P2C is the readout the conclusions rest on. M3 stays flat on `Qwen3-4B` and falls with distillation on `SmolLM3-3B` (bottom-40 0.042, $k = 8$ 0.037, MISTILL 0.025, Vanilla CoT-KD 0.058 and 0.061).

The instrument's floor. The same checkpoint read on two GPU types gives the identical top-40 and per-note differences of +0.002, +0.000 and +0.001 at P2C, P2 and P1 (Appendix K.5), about a third of the smallest per-note difference read as an effect here (P2-only against MISTILL seed 0, −0.006).

L.3 HOW THE RUNS WERE DONE, AND WHAT THEY DO NOT CONTROL FOR

The `Qwen3-4B` ablation arms trained on rented 4-GPU H100 nodes, one arm per GPU (the step is largely CPU-bound, so co-located arms slow each other), except $k = 8$ (a GH200) and $k = 100$ (a B200); their own-heads passes ran on the same GPUs or on a cluster. The `SmolLM3-3B` A′ arms

trained on rented RTX PRO 6000 and H200 instances or on a cluster RTX PRO 6000. Task evaluations ran on cluster RTX PRO 6000 cards, except P1-only, P2-only and attention-only (a rented evaluation instance) and `SmolLM3-3B` $k = 8$ (its H200, with a different vLLM build); generation is stochastic in any case, and the instrument's cross-hardware floor is above. Each run's `config.json` records its one differing flag.

P1-only and P2-only are λ-zero arms ($\lambda_{P2} = 0$ and $\lambda_{P1} = 0$): the trainer's literal single-phase path took two to three times the memory of the two-phase objective, for a cause not found, and a zero weight contributes exactly zero gradient. A literal P2-only run meant as a cross-check was stopped early, so the equivalence rests on that argument, not on a run. The term arms act on both phases. One seed per cell, except `Qwen3-4B` RandHeads and `SmolLM3-3B` bottom-40 (seeds 0 and 1). Three faults are disclosed rather than diagnosed: the attribution passes' host memory grew on a B200 node far beyond what the same code uses on an RTX PRO 6000 node (resumed per note, nothing lost); the single-phase path's memory growth; and one evaluation (`Qwen3-4B` bottom-40) that skipped LAS and M1 while exiting cleanly, resubmitted.

Not run. The temperature of the student pool ($\tau \times 0.5$ and $\times 2$), the no-window and symmetric-window cells, per-phase term arms, `SmolLM3-3B` P1-only and P2-only, a `SmolLM3-3B` attention-only cell at $\lambda = 0.03$, re-ranking the student's pool during training, and baseline tuning were planned and not run.

L.4 Training-set size

The learning curve (Figure 6) trains Vanilla CoT-KD and MISTILL on 100 and 250 notes (seed 0, both with bf16 weights) and uses the tier-1 rows at 493 notes (Table 35). With bf16 weights the two tie on the task at every size: at 100 notes MISTILL is higher on 58 notes and lower on 59 (p 1.00), at 250 notes 67 and 54 (p 0.28), and at 493 notes against bf16 Vanilla CoT-KD 205 and 184 note-pairs over three seed pairs (p 0.31). The one point where the arms separate, fp32 Vanilla CoT-KD at 493 notes (seed 0: 0.193 against MISTILL's 0.151), is the one that mixes precisions. MISTILL leads on M1 at P2C at every size, by a mean of +0.024, +0.032 and, at 493 notes, +0.035 against bf16 Vanilla CoT-KD (+0.012 against fp32 Vanilla CoT-KD, which reaches 0.791 on its own). The curve was run to decide whether to grow the corpus beyond 500 notes; it was not grown.

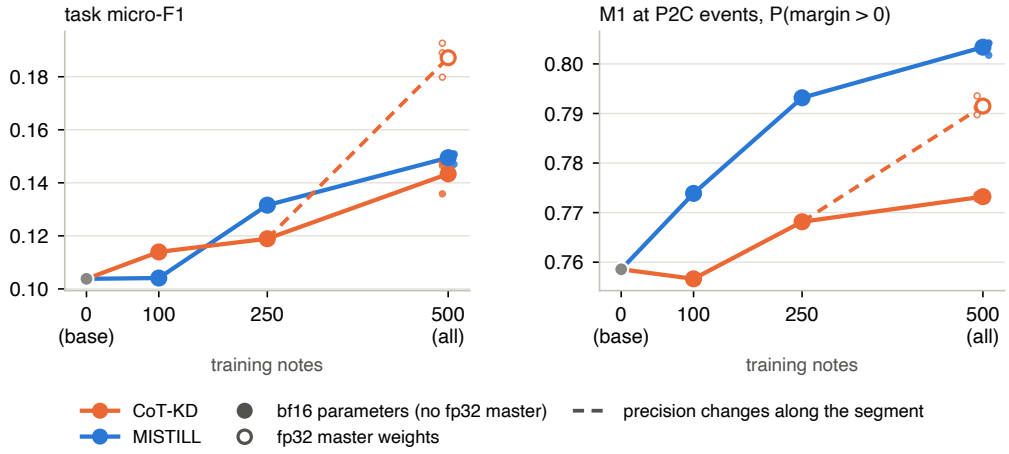


Figure 6: Learning curve on `Qwen3-4B` (seed 0 at 100 and 250 notes; seed means at all 493): task micro-F1 (left) and M1 at P2C events (right). Filled markers train bf16 weights, open markers fp32 master weights; the dashed segment is where the precision changes. With bf16 weights the two objectives tie on the task at every size, and MISTILL leads on its supervised readout. [R-D8][R-D17]

M Result index and reproduction entries

Supports: every [] in the paper

Table 35: Learning curve on `Qwen3-4B` (150 test notes; seed 0 at 100 and 250 notes, three seeds on the 500-note split, whose 493 notes with a valid teacher trace are the tier-1 training set, with the between-seed sd below). Paired: notes (or note-pairs, pooled over seed pairs) on which MISTILL is higher / lower, exact sign test. The second row at the full split is precision-matched. [R-D8][R-D17]

		micro-F1		M1 P2C		paired, per note	
training notes	precision (CoT-KD / MISTILL)	CoT-KD	MISTILL	CoT-KD	MISTILL	task	M1 P2C
0 (base)	—	0.104		0.759		—	—
100	bf16 / bf16	0.114	0.104	0.757	0.774	58/59	102/21
						p 1.00	$p < 10^{-4}$
250	bf16 / bf16	0.119	0.132	0.768	0.793	67/54	111/15
						p 0.28	$p < 10^{-4}$
500, 3 seeds	fp32 / bf16	0.187 ±.007	0.149 ±.002	0.791 ±.002	0.803 ±.001	167/234 p 0.001	241/137 $p < 10^{-4}$
500, 3 seeds	bf16 / bf16	0.143 ±.007	0.149 ±.002	0.773 ±.000	0.803 ±.001	205/184 p 0.31	348/29 $p < 10^{-4}$

One entry per result identifier; the code is at https://anonymous.4open.science/r/mistill-code-anon-3D07, whose `results/` folder lists every reported number with the statistic it is and the artifact and reader it comes from. Data require PhysioNet credentialed access to `MIMIC-IV` (Appendix A); no note or trace text is released. Reruns reproduce the reported statistics within the stated bands, not bit-exactly.

Layout. Code is in `src/mistill/` (events, buckets, losses, the objective, targets, the intervention hook), `scripts/` (training, evaluation, attribution and patching passes, readers) and `slurm/` (job files and the per-cluster site settings). Artifacts are per-experiment directories under `artifacts/` (paths below are relative to it); the experiment logbook `LOGBOOK.md` holds every run record with its job identifiers and every incident. Each training run writes `config.json`; each evaluation directory holds `metrics.json` (with the decoding settings), `predictions.jsonl`, `las.json`, `mechanism.json` (M1) and `generations/`; each own-heads cell `M6-<arm>/` holds `faithfulness2.json` (M4), `readside.json` (M2, M3, prefix share) and 150 per-note files, beside the arm's ranking in `HD-S-<arm>/head_sets_contrast.json`. Every number and generated table in the paper is written by `scripts/paper_numbers.py`, which wraps the readers named below and records its inputs in `paper_numbers.log`.

Data and pipeline. The distillation corpus (500 training and 150 test notes, one per patient) is built by `scripts/build_dt_corpus.py` and presented on the retrieved shortlist by `scripts/build_mimic_shortlist_set.py` (`data/mimic4/dtk`; the validation notes `data/mimic4/tr006k`). Teacher traces come from `scripts/pilot_teacher_mimic.py` (`TR-007/` validation, `TR-008/` corpus), teacher targets from `scripts/precompute_targets.py` (with `--mech` for MISTILL). Training: `slurm/dt_grid.sh` (baselines), `slurm/dt_opd.sbatch` (OPD, GOLD), `slurm/dt001_grid.sh` (MISTILL, pilot, tier 1, and `MODE=ablations` with the arm's one flag); evaluation: `slurm/dt_eval.sbatch` (`scripts/eval_student.py`, `scripts/las_simulator.py`, `scripts/eval_mechanism.py`); own heads: `slurm/dt001_s6_arm.sbatch` (attribution on the trained checkpoint, then `scripts/knockin_bins.py` with `--no-attr` at bins 8 and 40, then `scripts/read_side.py` with `--no-attr`, and `scripts/prose_locus.py` for the M5 arms).

Compute. One tier-1 training run takes about 3.0 h for fp32 Vanilla CoT-KD and 6 h 50 min for MISTILL on one GPU (Appendix H.4); an evaluation with M1 and LAS takes 1–2 h on an RTX PRO 6000; an own-heads chain takes about 3.5 h per `SmolLM3-3B` arm and 7 h per `Qwen3-4B` arm.

Nondeterminism classes, cited per entry. (G) Generation samples at temperature 0.6 with a per-note seed in vLLM; a fixed seed does not reproduce tokens across GPUs, library builds or batch composition, so task, fidelity and validity reproduce within the seed and replicate bands, not token for token. (K) Teacher-forced passes (margins, attribution, knock-in) vary slightly across GPUs:

the teacher's forward-pass noise is 0.046 nats (s.d.), one clean margin differs by a median of 0.045 nats between two GPU types, and one own-heads cell read on two GPUs differs per note by +0.002, +0.000 and +0.001 at P2C, P2 and P1 with an identical top-40. (T) Training: a same-seed rerun on other hardware differs by about 0.005 micro-F1, 0.003 M1 at P2C and 0.02 LAS; resumed runs restore every random-number state. (F) The Feature-KD runs are not bit-reproducible (Appendix I.3).

R-C1 Decisions are token-level events. Sec. 2.2, Appendix C, Table 4; Appendix C. *Artifacts:* `HD-001/<note>/events.jsonl`, `HD-001/READOUT.log` (audit block). *Reader:* `scripts/count_events.py`, `scripts/paper_numbers.py`. *Seeds:* one teacher trace per note. *Expected:* 2,064 P1, 7,688 P2, 3,855 P2C and 2,000 NULL events; median P1 margin +6.02 nats. Counts are exact given the stored traces; margins within (K).

R-C2 The attribution screen and the hook. Sec. 2.3; Appendices C and B. *Artifacts:* `HD-001/ANALYSIS.log`, per-note `attr.npz`; the detached-graph incident in the logbook. *Reader:* `scripts/attribute_heads.py` (self-test mode), `scripts/read_attribution.py`. *Seeds:* one teacher trace per note. *Expected:* correlation with exact effects 0.36 under the detached hook and 0.84 with the graph intact (0.47 and 0.85 above noise); noise floor 0.046 nats; all-heads knock-in equal to the clean margin (median difference 0.000 nats). *Nondeterminism:* K.

R-C3 Global sets are necessary and sufficient. Sec. 3.1, Table 1 (every k: Table 7); Appendix D. *Artifacts:* `HD-001/faithfulness2.json`, `HD-001/KNOCKIN.log`, per-note `patch2.jsonl`. *Reader:* `scripts/knockin_bins.py`, `scripts/read_knockin.py`. *Seeds:* one teacher trace per note; random sets drawn per bin. *Expected:* necessity of the top-8 0.62, 0.78 and 0.63 (P1, P2, P2C); knock-in of the top-40 0.86, 0.79 and 0.85 against random 0.04, 0.02 and 0.02; knock-in above random on 100, 100 and 92 notes. Medians over notes. *Nondeterminism:* K.

R-C4 Phase structure. Sec. 3.2, Table 8; Appendix D. *Artifacts:* `HD-001/head_sets_contrast.json`, `HD-001/PHASE_TABLE.log`, `HD-001/faithfulness2.json` (exact P2C refinement). *Reader:* `scripts/read_attribution.py` with `--contrast`, `scripts/read_knockin.py`. *Seeds:* one teacher trace per note. *Expected:* `L39H38` first in every phase (I_c 0.91, 0.97, 0.62); `L30H11` rank 2 at P2C and 1,500 at P1; 24 of the chapter top-40 in layers 24–31. P2 and P2C rankings are first-order. *Nondeterminism:* K.

R-C5 What the heads read (descriptive). Sec. 3.3, Figure 4; Appendix E. *Artifacts:* `HD-001/readside.json`, `HD-001/read_panel.json`, per-note `read.npz`. *Reader:* `scripts/read_side.py`, `scripts/read_readout.py`. *Seeds:* one teacher trace per note. *Expected:* `L26H36` region share 0.47, lift 9.4, correlation with importance +0.44; `L39H38` prose share 0.84; `L30H11` on the near-miss's mentions 0.33 (random 0.02). Not intervened on. *Nondeterminism:* K.

R-C6 The prose locus. Sec. 3.4, Figure 1; Appendix F. *Artifacts:* `HD-001/locus.json`, `HD-001/locus_sets.json`, `HD-001/head_sets_locus.json`, `HD-001/read_locus_panel.json`. *Reader:* `scripts/prose_locus.py` and `scripts/read_locus.py` (pass A), `scripts/locus_sets.py` and `scripts/read_locus_sets.py` (pass B). *Seeds:* one teacher trace per note. *Expected:* 36% of mentions decided in the window (29% inherited, 32% never decisive), median 7 tokens before the letter; prose-time top-8 window-only 1.00 (random 0.02) on 100 of 100 notes against the emission top-16's 0.16; prose-time heads on the note 0.25. *Nondeterminism:* K.

R-C7 Largely preserved on the shortlist. Sec. 3.5; Appendix G. *Artifacts:* `HD-002/compare.json` and the HD-002 readouts. *Reader:* the R-C3–R-C6 pipeline on `tr006k` / `TR-007`, then `scripts/compare_hd.py`. *Seeds:* one new teacher trace per note. *Expected:* Spearman 0.55, 0.63, 0.55; top-40 overlap 31, 31, 32; top-40 knock-in 0.92, 0.82, 0.85; locus 9 tokens (full inventory 7); prose-time Spearman 0.79; the chapter top-8 knock-in shifts (0.47 against 0.67). *Nondeterminism:* G (new traces), K.

R-D0 Setup. Sec. 4.2; Appendices A and I. *Artifacts:* `RT-003/bakeoff_*.json`, `DT-000/tokenizer_audit.json`, `TR-007/`, `TR-008/`. *Reader:* `scripts/build_mimic_shortlist_set.py` (recall sweep), `scripts/tokenizer_audit.py`. *Seeds:* teacher sampling seed 0. *Expected:* retriever recall 0.903 (validation), 0.936 (test split) and 0.939 (the 650-note corpus); teacher micro-F1 0.199 (validation) and 0.227 (test); 493 of 500 training traces valid. *Nondeterminism:* G (teacher traces).

R-D1 Untrained students. Table 2. *Artifacts:* `DT-000/eval/Qwen3-4B__base__s0/`, `DT-000/eval/SmolLM3-3B__base__s0/`, linked with every main-table row in `DT-001/eval_main/`. *Reader:* `scripts/read_dt.py` over `DT-001/eval_main` with `--reference vanilla --markdown`. *Seeds:* none (no training). *Expected:* micro-F1 0.104 (valid 0.987) and 0.003 (valid 0.153). *Nondeterminism:* G.

R-D2 Vanilla CoT-KD (fp32). Table 2; Appendix K.1. *Artifacts:* `DT-000/{runs,eval}/Qwen3-4B__vanilla__s{0,1,2}/`, `DT-000/{runs,eval}/SmolLM3-3B__vanilla__s0/`, `DT-001/{runs,eval}/SmolLM3-3B__vanilla__s1/`. *Reader:* as R-D1. *Seeds:* 3 and 2. *Expected:* `Qwen3-4B` 0.187 $\pm$0.007, against the untrained student +0.081 [+0.063, +0.101], gap recovered 0.68, LAS 0.267; `SmolLM3-3B` 0.130 $\pm$0.025 (0.148 and 0.113). *Nondeterminism:* G, T.

R-D3 Feature-KD below Vanilla CoT-KD. Table 2; Appendix I.3. *Artifacts:* `DT-000/{runs,eval}/Qwen3-4B__feature_kd__s{0,1,2}/`, `DT-000/{runs,eval}/SmolLM3-3B__feature_kd__s0/`. *Reader:* as R-D1. *Seeds:* 3 and 1. *Expected:* 0.153 $\pm$0.010, −0.035 [−0.049, −0.021] against fp32 Vanilla CoT-KD, LAS −0.054; `SmolLM3-3B` 0.062, −0.060 [−0.080, −0.038]. *Nondeterminism:* G, T, F.

R-D4 DistiLLM-2, OPD and GOLD. Table 2; Appendix K.2. *Artifacts:* `DT-000/{runs,eval}/Qwen3-4B__{distillm2,opd}__s0/`, `DT-000/eval/SmolLM3-3B__opd__s0/` (the GOLD arm; its checkpoint `DT-000/runs/SmolLM3-3B__opd__s0.podcopy/final/`). *Reader:* as R-D1. *Seeds:* 1 each; OPD and GOLD train bf16 weights. *Expected:* DistiLLM-2 0.187 (+0.003, p 0.57); OPD 0.179, LAS −0.038 (p 0.006), BLEU-4 0.200, valid 0.973; GOLD 0.000 with M1 at P2 0.355. *Nondeterminism:* G, T (on-policy rollouts are sampled).

R-D5 MISTILL at tier 1. Sec. 4.3, Table 2; Appendix K. *Artifacts:* `DT-001/{runs,eval}/Qwen3-4B__mech__s{0,1,2}/`, `DT-001/{runs,eval}/SmolLM3-3B__mech__s{0,1}/`. *Reader:* as R-D1, and `scripts/pilot_compare.py` (`pooled`) for the seed-matched M1 counts. *Seeds:* 3 and 2. *Expected:* `Qwen3-4B` 0.149 $\pm$0.002 against bf16 Vanilla CoT-KD's 0.143 $\pm$0.007 (per note −0.007 [−0.019, +0.005]); `SmolLM3-3B` 0.051 $\pm$0.013 against 0.055 (+0.002 [−0.016, +0.019]); M1 at P2C 0.803 against 0.773 (bf16) and 0.791 (fp32). *Nondeterminism:* G, T.

R-D6 Mechanism on each arm's own heads. Sec. 4.4, Table 3; Appendix J. *Artifacts:* `M6-<arm>/`, `HD-S-<arm>/`, `HD-002/` (teacher row). *Reader:* `scripts/read_m4.py` with `--prefix M6- --teacher artifacts/HD-002 --only`; `scripts/read_mech.py` with `--paired`; `scripts/m4_paired.py` with `--pooled`; `scripts/read_shares.py`. *Seeds:* as the arms' (Table 3); pooling pairs seed i with seed i. *Expected:* untrained `Qwen3-4B` 0.788 (teacher 0.847), untrained `SmolLM3-3B` 0.416 (floor 0.107); MISTILL minus fp32 Vanilla CoT-KD, seed-matched and pooled: `Qwen3-4B` +0.011 (250/200, p 0.02), M2 +0.113 at P2C and +0.170 (348/102, p 2×10^{-32}) at P1; `SmolLM3-3B` +0.248 (257 of 300). Against bf16 Vanilla CoT-KD: +0.032 at P2C (304/146), −0.019 at P2; `SmolLM3-3B` +0.317. *Nondeterminism:* K, T.

R-D7 Positions and heads: the controls. Secs. 4.4 and 4.5, Figure 2; Appendix L. *Artifacts:* `DT-001/{runs,eval}/{Qwen3-4B,SmolLM3-3B}__mech_{randheads,randpos}_t1__s0/`, `DT-001/{runs,eval}/Qwen3-4B__mech_randboth_t1__s0/`, and their `M6-` and `HD-S-` cells. *Reader:* as R-D6. *Seeds:* one per control. *Expected:* RandPos 0.631 and 0.304 (untrained 0.788 and 0.416); RandHeads 0.869 (`Qwen3-4B`) and 0.585 (`SmolLM3-3B`; MISTILL 0.789); RandBoth 0.863. *Nondeterminism:* K, T.

R-D8 Training-set size. Sec. 4.3, Figure 6; Appendix L.4. *Artifacts:* `DT-001/eval/Qwen3-4B__{mech,vanilla}_c100__s0/`, `DT-001/eval/Qwen3-4B__{mech,vanilla}_250__s0/`, the tier-1 rows and `DT-001/eval/Qwen3-4B__vanilla_bf16__s{0,1,2}/`. *Reader:* `scripts/pilot_compare.py` (`paired`, `pooled`); the figure from `scripts/fig_curve.py`. *Seeds:* 1 at 100 and 250 notes, 3 at 493. *Expected:* Table 35. *Nondeterminism:* G, T.

R-D9 Training cost. Sec. 6; Appendix H.4. *Artifacts:* each run's log and `metrics.jsonl` (logged peak memory and step times); the logbook's run records. *Seeds:* not applicable. *Expected:* peak 37.5 GiB and about 27 s per note on one GPU without an online teacher; at 100 notes 24–34 min for Vanilla CoT-KD against 76–78 min for MISTILL; 6 h 50 min for a 493-note MISTILL run. Hardware-dependent; indicative only.

R-D10 The objective's selection on the validation notes. Sec. 4.1; Appendix H. *Artifacts:* `DT-001/eval/<pilot-arm>/` for the grid (`Qwen3-4B__mech_l<a>_<b>__s0`), the write-term forms and the selected arm on three seeds, `vanilla_100__s{0,1,2}`, `Qwen3-4B__base__val` and `Qwen3-14B__teacher__val` (the M1 anchor); `DT-001/M2val-<arm>/readside.json`. *Reader:* `scripts/pilot_compare.py` (`breakdown`, `paired`, `pooled`, `m2`), `scripts/select_lambda.py`, `scripts/diag_mech_terms.py` (write-term diagnostics), `slurm/m2val.sh`. *Seeds:* 1, and 3 for the selected pair. *Expected:* Tables 25–27; one-seed noise 0.005 on validation micro-F1. *Nondeterminism:* G, T.

R-D11 Extra tier-1 cells. Appendix K.4. *Artifacts:* `DT-001/eval/Qwen3-4B__mech_l0.03_0.1__s0/`, `DT-001/eval/Qwen3-4B__mech__s2_gh200/` (the same-seed replicate), `DT-001/eval/SmolLM3-3B__mech_attn__s0/`. *Seeds:* 1 each. *Reader:* `scripts/read_dt.py`; `scripts/pilot_compare.py` (`paired`) for the replicate. *Expected:* 0.152; replicate 0.155 against 0.150; `SmolLM3-3B` attention only 0.036. The replicate's differences define (T).

R-D12 Reproducible recovery magnitude. Secs. 4.4 and 4.5, Table 3. *Artifacts:* `M6-<arm>/` of every seed. *Reader:* `scripts/m4_paired.py`. *Seeds:* 3 and 2 (MISTILL, Vanilla CoT-KD), 2 (RandHeads, bottom-40). *Expected:* MISTILL's `Qwen3-4B` seeds 0.873, 0.870, 0.874 against fp32 Vanilla CoT-KD's 0.840, 0.848, 0.891; `SmolLM3-3B` means over notes 0.786 and 0.788 (MISTILL) against 0.550 and 0.528 (Vanilla CoT-KD); RandHeads' seeds differ by −0.033, `SmolLM3-3B` bottom-40's by −0.023 (p 0.004). bf16 Vanilla CoT-KD's three `Qwen3-4B` seeds spread 0.013 at P2C (MISTILL's 0.004). *Nondeterminism:* K, T.

R-D13 Which heads, how many, and where the mechanism lands. Sec. 4.5, Figure 2; Appendix L.2. *Artifacts:* `DT-001/{runs,eval}/Qwen3-4B__mech_{bottom,k8,k100}_t1__s0/`, `DT-001/{runs,eval}/SmolLM3-3B__mech_{bottom_t1__s0,bottom_t1__s1,k8_t1__s0}/`, their `M6-` and `HD-S-` cells, and `HD-S-{Qwen3-4B,SmolLM3-3B}/` for the overlaps. *Reader:* `STUDENT=<student> bash scripts/read_ablations.sh` (wrapping `scripts/read_m4.py`, `scripts/read_mech.py`, `scripts/read_shares.py`, `scripts/m4_paired.py`, `scripts/read_dt.py` with `--compact`, and `scripts/head_overlap.py`). *Seeds:* 1 (`SmolLM3-3B` bottom-40: 2). *Expected:* `Qwen3-4B` bottom-40 0.717; `SmolLM3-3B` bottom-40 0.841 and 0.835; $k = 8$ 0.881 (`Qwen3-4B`) and 0.671 (`SmolLM3-3B`); $k = 100$ 0.875; carriers 27–38 of 40. *Nondeterminism:* K, T.

R-D14 Phases, terms and dose. Sec. 4.5, Figure 2; Appendix L.1. *Artifacts:* `DT-001/{runs,eval}/Qwen3-4B__mech_{p1only_lam0,p2only_lam0,attnonly,interonly}_t1__s0/`, `DT-001/{runs,eval}/Qwen3-4B__mech_l0.03_0.03__s0/`, `DT-001/{runs,eval}/Qwen3-4B__mech_randheads_t1__s1/`, their `M6-` and `HD-S-` cells. *Reader:* as R-D13. *Seeds:* 1 (RandHeads: 2). *Expected:* P1-only 0.802, P2-only 0.864, attention-only 0.861, interaction-only 0.866, λ 0.03 / 0.03 0.880; task in Table 34. *Nondeterminism:* K, T.

R-D15 Readouts that do not discriminate. Sec. 4.4; Appendix J. *Artifacts:* `M6-<arm>/readside.json`, `M6-<arm>/locus.json` (M5 arms). *Reader:* `scripts/read_mech.py`, `scripts/read_shares.py`. *Seeds:* as R-D6; M5 one seed per arm. *Expected:* P1 recovery 0.946 and 0.985 untrained; M5 8, 7 and 7 against the teacher's 9; M3 on `SmolLM3-3B` 0.081 to 0.025. *Nondeterminism:* K.

R-D16 Instrument checks. Appendix K.5. *Artifacts:* `_race_h200pod.HD-S-SmolLM3-3B__mech_k8_t1__s0/`, `_race_h200pod.M6-SmolLM3-3B__mech_k8_t1__s0/`; the gate's cell was not retained (its numbers are in the logbook). *Reader:* link the two copies side by side and run `scripts/read_m4.py` and `scripts/m4_paired.py` (pair `k8_h200:k8_rtx`); `scripts/gate_noattr.py` for the gate. *Seeds:* one cell each. *Expected:* largest gate difference 0.014; cross-hardware per-note differences +0.002, +0.000 and +0.001 at P2C, P2 and P1. *Nondeterminism:* K.

R-D17 Precision control. Secs. 4.2 and 4.3, Tables 2 and 3; Appendix K.3. *Artifacts:* `DT-001/{runs,eval}/Qwen3-4B__vanilla_bf16__s{0,1,2}/`, `DT-001/{runs,eval}/SmolLM3-3B__vanilla_bf16__s0/`, their `HD-S-` and `M6-` cells, `DT-001/{runs,eval}/Qwen3-4B__mech_fp32__s{0,1,2}/`. *Reader:* `scripts/read_precision.sh`; `scripts/pilot_compare.py` (`pooled`). *Seeds:* 3 and 1; fp32 MIST-

ILL 3. *Expected:* Table 31 (bf16 Vanilla CoT-KD minus fp32 −0.041 on `Qwen3-4B`, −0.069 on `SmolLM3-3B`); training cross-entropy as in Table 32; own heads as in Table 3; fp32 MISTILL 0.165 (−0.024 against fp32 Vanilla CoT-KD). *Nondeterminism:* G, T, K.

N EXTENDED RELATED WORK

Supports: Sec. 5

Chain-of-thought distillation. Training small students on a teacher's rationales transfers multi-step reasoning (Hsieh et al., 2023; Ho et al., 2023; Magister et al., 2023; Shridhar et al., 2023); later work makes rationales and answers consistent (Wang et al., 2023b; Chen et al., 2023), selects or chunks rationales (Yan et al., 2025; Chen et al., 2025a), distills agent trajectories (Kang et al., 2025), and asks which rationale properties matter (Wadhwa et al., 2024; Chen et al., 2025b). All transfer text, not the mechanism that uses it. MISTILL keeps the sequence-level loss (Kim & Rush, 2016) and adds a target for that mechanism.

Attention and representation distillation. Beyond output distributions (Hinton et al., 2015), students can match a teacher's hidden states (Romero et al., 2015), attention maps (Zagoruyko & Komodakis, 2017), layer-wise attention and hidden states (Jiao et al., 2020), or last-layer self-attention and value relations (Wang et al., 2020). Logit distillation for LLMs now includes contrastive (DistiLLM-2; Ko et al., 2025), on-policy (Agarwal et al., 2024) and cross-tokenizer objectives (ULD, DSKD, GOLD; Boizard et al., 2025; Zhang et al., 2024; Patiño et al., 2025). All act at every position, on the output distribution or on fixed layers. MISTILL chooses the heads (causally ranked) and positions (at and just before the teacher's decision events), and matches pooled attention plus its interaction with the heads' contribution to the decision margin. Circuit distillation aligns the representations of functionally corresponding circuit components in teacher and student (Wadhwa et al., 2025); MISTILL instead pools each model's own causally ranked heads, so it needs no correspondence between them.

Mechanistic interpretability. Causal mediation and activation patching localize behavior to components (Vig et al., 2020; Meng et al., 2022); circuit analyses find the components that implement a task (Elhage et al., 2021; Wang et al., 2023a; Conmy et al., 2023); attribution patching screens every component with a gradient approximation (Syed et al., 2024; Kramár et al., 2024); and metric and corruption choices matter (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Retrieval heads fetch information from long contexts (Wu et al., 2025b). New here are a full inventory of thousands of codes in context, decisions as token-level events scored by the model's own margin, and per-note necessity and knock-in sufficiency against a null calibration.

ICD coding and extreme classification. ICD coders are typically supervised multi-label classifiers, often with label-wise attention (Mullenbach et al., 2018); Edin et al. (2023) review and replicate them on MIMIC-III and MIMIC-IV (Johnson et al., 2023b). Large language models code poorly when asked for codes directly (Soroush et al., 2024), and recent agentic systems implement the official coding guidelines over the full ICD-10 system (Motzfeldt et al., 2025). Coding is extreme multi-label classification, whose benchmarks (Bhatia et al., 2016) report propensity-scored metrics that up-weight rare labels (Jain et al., 2016); recent work optimizes long-tail and macro-at-k metrics directly (Schultheis et al., 2023; 2024), trains dual encoders over the label space (Gupta et al., 2024), and scales training in large output spaces with dynamic sparsity and low precision (Ullah et al., 2024; Zhang et al., 2025; Ullah et al., 2026). LLMs tackle it by generating or retrieving candidate labels in context (Zhu & Zamani, 2024; D'Oosterlinck et al., 2024). Interpretability in coding has targeted trained classifiers (Mullenbach et al., 2018; Wu et al., 2025a; Edin et al., 2024); we ask how a reasoning-trained model selects codes from the full inventory in context, and whether that can be distilled.