

Audit Before You Distill: A Controlled Workflow for Reasoning Transfer into Compact Medical Language Models

Tiago Martinelli^{a,b,*}, João Paulo Papa^{a,c}, Adriano José Pereira^a

^a*Einstein Global Advanced Technologies for Equity (E-GATE), Hospital Israelita Albert Einstein, São Paulo, Brazil*

^b*Institute of Science and Technology Itau (ICTi), Itau Unibanco, São Paulo, SP, Brazil*

^c*School of Sciences, São Paulo State University, São Paulo, Brazil*

Abstract

Teacher rationales can help a smaller model answer health professions questions, but they can also reinforce errors. We examined when they help and how this evidence can guide training using Brazilian Portuguese questions from medical and multiprofessional residency examinations. First, we measured how a teacher rationale changed the student’s probability of selecting the correct answer and compared this effect with rationales drawn from other questions. The complete rationale increased this probability by 21 p.p. when the teacher was correct and decreased it by 14 p.p. when the teacher was wrong. By intervening on the reasoning content, we showed that shuffled rationales did not produce the same pattern. Guided by this result, we trained only with rationales attached to correct teacher answers and compared answer-only training, step-by-step distillation, and our proposed Distill-SFT method. Distill-SFT showed a modest advantage over step-by-step distillation on previously missed questions under reasoning, while using fewer training instances by combining the answer and rationale into a single target. We also investigated catastrophic forgetting using the out-of-domain ENEM dataset. Answer-only training reduced reasoning accuracy by 12 p.p. relative to baseline, compared with approximately 4 p.p. for either rationale-based training method. Together, these findings support a workflow that audits rationale content before distillation, verifies teacher answers before supervision, and trains compact models efficiently for health professions question answering.

Keywords: reasoning distillation, small language models, health informatics, multilingual question answering, causal intervention

1. Introduction

Frontier health assistants show how far language models can go, from medical question answering to diagnostic dialogue and health conversations grounded in personal records [1, 2, 3]. Bringing this capacity to compact models is attractive in educational and health-informatics settings, where model size and decoding costs constrain deployment. Rationale distillation offers one route: a large teacher generates a rationale for its answer, and that rationale becomes supervision for a smaller student [4, 5, 6].

Previous work has transferred rationales through separate label and explanation tasks [4], direct supervision of smaller models with teacher reasoning traces [5], multilingual or multi-teacher traces [6, 7], retrieved knowledge [8], student-oriented trace selection [9, 10], and compressed latent reasoning [11]. These approaches improve the way rationales are generated, selected, or encoded. Still, they do not establish by themselves whether transfer is caused by

question-relevant content rather than by rationale length, style, or template.

A rationale is not automatically a trustworthy training signal: it may be irrelevant to the question, support an incorrect answer, or constrain the student to a model-specific generation format. These risks are especially consequential outside of English, where compact open-weight students may be weaker, and reasoning behavior may be anchored in another language [12].

Three unresolved issues therefore limit confidence in rationale distillation. First, the benefit of rationale content is often not separated from the mere presence of a long, structured target. Second, teacher correctness may be treated as a filtering detail even though a rationale can transmit an incorrect answer. Third, comparisons based on individual runs can be unstable when accuracy varies substantially across data splits.

Our study asks how reasoning supervision can be diagnosed, selected, and transferred efficiently into a compact model, and why answer verification is essential.

*Corresponding author.

Email addresses: tiago.martinelli93@gmail.com (Tiago Martinelli), joao.papa@unesp.br (João Paulo Papa), adriano.pereira@einstein.br (Adriano José Pereira)

2. Material and Methods

To determine whether a rationale carried useful answer-directed content, we designed an exploratory *in silico* study with two linked stages. The work was conducted through a partnership between Hospital Israelita Albert Einstein, a not-for-profit quaternary hospital in São Paulo, Brazil, and São Paulo State University (UNESP), Bauru, Brazil. We distilled a Qwen3-32B teacher into a Qwen3-4B student using Brazilian Portuguese multiple-choice questions from medical and multiprofessional residency examinations. Stage 1 measured the frozen baseline student’s answer-choice probabilities while progressively revealing coherent or cross-question-shuffled teacher rationales. A complementary training ablation tested whether the same alignment effect persisted after supervised adaptation. Stage 2 applied the answer-verification rule and compared three training techniques over ten paired data splits: answer-only supervised fine-tuning, step-by-step multi-task distillation, and our proposed single-target distill-SFT. Questions answered correctly by the teacher but incorrectly by the baseline student formed the transfer subset, separating newly acquired answers from answers the student already knew. Statistical evaluation combined two-level bootstrap confidence intervals, paired split-level tests, and within-split McNemar diagnostics.

Data sources and transfer strata. Residence exams consist of multiple-choice questions in Brazilian Portuguese. We used multiple-choice examinations because they provide a reproducible test bed for health knowledge and controlled transfer, as in major medical-LLM evaluations. They do not substitute for assessing information gathering, guideline adherence, open-ended communication, or integration into clinical workflows [1, 13, 14, 2]. To probe forgetting of pretrained reasoning behavior outside the adaptation domain, we also used the public Exame Nacional do Ensino Médio (ENEM), Brazil’s national secondary-school examination. We defined Qwen3-4B before adaptation as the *baseline student*. Each item was assigned to one of four strata according to whether the teacher and baseline student answered it correctly: **A** (both correct), **B** (teacher correct, student wrong), **C** (student correct, teacher wrong), or **D** (both wrong). Subset B served as the transfer subset because it isolates questions answered correctly by the teacher but not by the student. Aggregate accuracy combines two outcomes—acquiring missing answers on B and preserving answers already known on A and C—that can move in opposite directions and conceal transfer. We therefore report stratum-specific results throughout.

Models and rationales. We used Qwen3-32B as the teacher and Qwen3-4B as the student [15]. On the analysis pool, the teacher answered 74.9% of items correctly versus 58.8% for the baseline student, leaving a 16.1-percentage-point

transfer gap that motivated distillation. Teacher rationales were generated once and normalized by removing the original thinking delimiters. When a rationale exceeded 250 whitespace-delimited words, preprocessing retained a tail-preserving window aligned to a sentence boundary, thereby preserving its concluding evidence and decision. Only rationales associated with a correct teacher answer were used as training targets. We refer to these as *answer-verified rationales*, because answer verification does not establish that every intermediate statement is valid. The student was adapted with low-rank adaptation (LoRA) [16]: rank 32, scaling 32, two epochs, learning rate 10^{-4} , effective batch size 16, and bfloat16. LoRA initialization was fixed and separated from the ten data-split seeds; the exact seeds are reported in [Appendix A](#).

Study workflow. Figure 1 shows how the rationale diagnostic in Stage 1 led to the training comparison in Stage 2. The complementary training ablation is summarized in the Results, and the delimiter control is reported in [Appendix A](#).

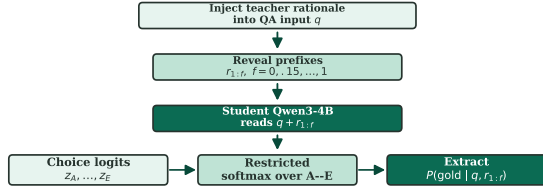
In Stage 1, the frozen baseline student read growing prefixes of coherent or cross-question-shuffled teacher rationales. The shuffle preserved length, style, and language while removing relevance to the question, and the trajectories were stratified by whether the teacher was correct. This diagnostic asked whether question-relevant rationale content changed answer-choice probabilities and whether the direction of that change followed the teacher’s answer.

We treated rationale assignment as a controlled intervention: the question, frozen reader, inference prompt, and evaluation rule were held fixed while question–rationale alignment was changed. In the potential-outcomes tradition, causal effects are defined by contrasts between outcomes under alternative interventions on the same treatment variable [17]. The coherent-versus-shuffled contrast therefore estimates the effect of question-relevant rationale content within this protocol; Fig. 2 shows the prefix construction and the resulting teacher-correct and teacher-wrong probability trajectories. This protocol-level contrast does not identify the model’s internal causal mechanism or establish transportability beyond the evaluated model family, item population, language, and training procedure.

Separately, the complementary training ablation trained matched students with either coherent rationales or cross-question-shuffled rationales while holding the optimization protocol, prompt, and evaluation rule fixed. The models were evaluated on a held-out subset B test set ($n = 164$ per split). This experiment asked whether question–rationale alignment also affected learned transfer; delimiter sensitivity was treated as a secondary control ([Appendix A](#)).

Stage 2 compared three techniques using training data from all four strata. Within each split, A–C were partitioned 85/15 into training and test items; D was similarly partitioned into training items and a disjoint held-out dif-

A Rationale influence diagnosis



B Answer-verified distillation

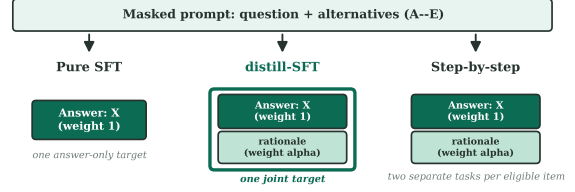


Figure 1: **Study workflow.** Panel A shows how the baseline student read increasing portions of a teacher rationale. After each portion, we measured the probability assigned to the correct answer among options A to E. We compared each rationale with one assigned from another question and separated cases in which the teacher was correct from those in which it was wrong. Panel B shows the training targets. Pure SFT uses the answer only. Distill-SFT places the rationale and answer in one target and gives rationale tokens weight α . Step-by-step distillation uses separate answer and rationale tasks, with weight α for the rationale task. Only rationales linked to correct teacher answers were used for training.

difficult test. The A–D training partitions were combined in the same model fit, and D was separated only for reporting so that performance on difficult questions remained visible. Train and test items never overlapped. *Pure SFT* supervised the answer only. Our *distill-SFT* used a single completion containing the answer-verified rationale followed by the answer, with rationale tokens weighted by α . *Step-by-step* distillation [4] created separate answer and rationale tasks. The distill-SFT target retained the Qwen3 thinking delimiters, and reasoning inference began with a Brazilian Portuguese thinking opener. Both rationale-trained techniques were evaluated under direct-answer and reasoning inference.

Rationale-conditioned choice probability. Before training, the baseline student read the question and growing prefixes $r_{i,f}$ of each stored teacher rationale, with the explicit final-answer span removed. At each checkpoint, we measured how much probability the student assigned to the correct option after reading that fraction of the rationale. We obtained this value by scoring the five answer-letter continuations and normalizing their logits over the restricted A–E set:

$$\tilde{p}_{i,f}(a) = \frac{\exp z_{i,f,a}}{\sum_{b \in \{A, \dots, E\}} \exp z_{i,f,b}}. \quad (1)$$

For teacher-correct and teacher-wrong groups g , the trajectory was the group mean probability of the gold answer,

$$\bar{p}_{g,c}(f) = \frac{1}{|g|} \sum_{i \in g} \tilde{p}_{i,f}(y_i), \quad c \in \{\text{coherent}, \text{shuffled}\}. \quad (2)$$

We reported the full-rationale change from baseline, the coherent-minus-shuffled contrast, and their teacher-correctness interaction. This was a restricted multiple-choice score, not a calibrated full-vocabulary probability.

Loss reweighting. We first express the balance between answer and rationale supervision as $\mathcal{L}_{\text{label}} + \alpha \mathcal{L}_{\text{rationale}}$: the label loss trains the final answer, the rationale loss trains

the explanation, and α controls their relative contribution. We evaluated $\alpha \in \{1.0, 0.3, 0.1\}$. Unit weight provided the reference; 0.3 reduced the contribution of the much longer rationale, and 0.1 was the lowest weight retained for multi-split evaluation. Values below 0.1 were not advanced beyond preliminary runs because the rationale objective no longer preserved a consistent reasoning signal. In the separate-task comparator, α scaled the entire rationale task. Our proposed distill-SFT applies this balance within a single completion. Since hundreds of rationale tokens can dominate the loss while the short answer carries the decision, we assigned weight α to rationale tokens and weight 1 to answer tokens:

$$\mathcal{L} = - \frac{\sum_t w_t \log p_\theta(y_t | y_{<t}, x)}{\sum_t w_t}, \quad (3)$$

$$w_t = \begin{cases} \alpha, & t \in \text{rationale}, \\ 1, & t \in \text{answer} \end{cases}.$$

The same α grid therefore varied token weights in our single-target loss and task weights in the comparator. All principal estimates were reconstructed from question-level predictions using a two-level nonparametric bootstrap (data splits resampled, then questions within splits; 10,000 replicates) [18], paired t tests across splits with Wilcoxon signed-rank tests as sensitivity analyses, and within-split exact McNemar tests [19]. We considered three principal comparison families: content relevance, technique by inference regime, and rationale loss weight. Delimiter sensitivity was a secondary control reported in Appendix A. Other pairwise comparisons were exploratory, and the minimum McNemar p across splits was treated as descriptive rather than as a global test. We also monitored generation truncation (maximum 0.45%), answer emission (minimum 95.7%), and native-English reasoning under the reasoning regime (0%).

3. Results

3.1. Rationales transmit the teacher’s answer direction

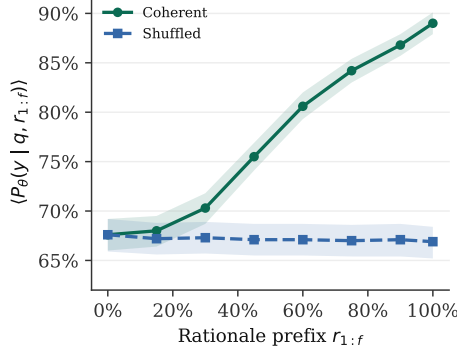
The first question was whether rationales carried the teacher’s answer into the smaller student, for better or

A Translated rationale prefix

A young woman with no psychiatric history;	15%
recent behavioral change includes	30%
persecutory thoughts, aggression, and insomnia.	45%
Headache and fever suggest an organic cause,	60%
such as a central nervous system infection,	75%
which may cause psychotic and neurologic symptoms.	90%
Sudden onset strengthens this suspicion.	100%

Final-answer span removed

B Teacher correct



C Teacher wrong

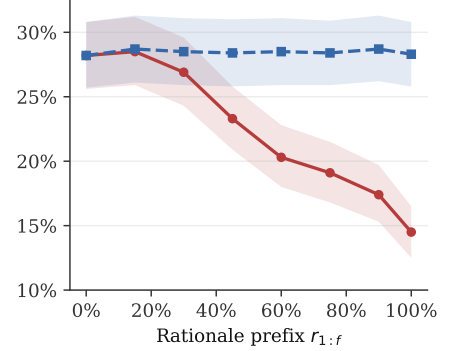


Figure 2: **Rationales move the student toward the teacher’s answer.** Panel A shows an English translation of one rationale originally written in Portuguese. The example illustrates how successive parts of the rationale were shown to the student. The curves in panels B and C were calculated using the original Portuguese rationales after removing the explicit final answer. They show the mean probability assigned to the correct answer among the five alternatives when the teacher was correct (B) or wrong (C). Solid lines use the original rationales. Dashed lines use rationales assigned from other questions. Shaded regions show 95% bootstrap intervals from resampling questions. With the complete rationale, the effects relative to rationales from other questions were +0.22 and −0.14, respectively, a difference of 0.36.

worse. With complete rationales, coherent content increased the mean restricted probability of the gold answer by 0.21 when the teacher was correct (Fig. 2B) and decreased it by 0.14 when the teacher was wrong (Fig. 2C). The coherent-versus-shuffled comparison confirmed this directional separation, with a correctness-by-coherence interaction of 0.36 (95% cluster-bootstrap CI 0.33–0.39). Coherent rationales thus carried the direction of the teacher’s answer: they transferred useful knowledge when the teacher was correct but made an incorrect answer more persuasive when the teacher was wrong. This result motivated the answer-verified supervision used in Stage 2.

The interaction was robust to overt option-wording cues. After jointly excluding rationales with exact normalized gold-option containment or an upper-decile gold-versus-distractor token-overlap margin, the effects retained their direction and the interaction interval remained above zero. This post-hoc sensitivity reused the stored probabilities and could not rule out semantic paraphrase; sample sizes, effect estimates, and confidence intervals are reported in Appendix Table B.3.

A complementary training ablation independently supported the importance of question–rationale alignment: coherent targets transferred better than cross-question-shuffled targets, while a secondary delimiter control found no detectable difference between the two coherent-target formats (Appendix A).

3.2. Efficient transfer, selective reasoning, and preserved capability

Stage 2 evaluated more than which technique produced the highest aggregate score: could answer-verified rationales recover answers the student lacked, preserve answers it already knew, and do so with a leaner training construction? This balance depended on the inference regime. Table 1 and Fig. 3 compare answer-only SFT, distill-SFT,

and multi-task step-by-step distillation. Neither rationale-trained technique was globally dominant. At $\alpha = 1.0$, no differences were detected in aggregate ABC accuracy or on difficult subset D under reasoning. On transfer subset B, however, distill-SFT gained 2.2 percentage points over step-by-step under reasoning (95% paired t CI 0.3–4.0), whereas step-by-step was stronger under direct inference. The technique-by-inference interaction remained positive at the lower rationale weights, supporting an advantage tied to how the student was used rather than uniform superiority. Appendix Table A.2 reports the paired contrasts, their uncertainty, and the sensitivity tests.

The step-by-step student also required reasoning decoding to reach this performance range. At $\alpha = 1.0$, reasoning improved its accuracy on both transfer and difficult items, with gains on subset B in all ten splits and on subset D in nine. Its direct mode remained competitive with pure SFT, with no difference detected on either subset, but did not reach the reasoning-enabled transfer regime (Appendix Table A.2).

Under matched reasoning inference, distill-SFT used 382 optimizer steps compared with 678 for step-by-step because it learned the rationale and answer in one target rather than duplicating rationale-bearing examples. Their reasoning outputs had similar lengths (303 vs. 297 tokens at $\alpha = 1.0$), and we detected no difference on difficult subset D or in the aggregate reasoning comparison. Distill-SFT therefore reached the same broad reasoning-performance range with 44% fewer optimizer updates. Also, no difference was detected between the two rationale-trained methods in the out-of-domain reasoning-retention analysis.

An exploratory frozen-reader audit showed that both techniques generated answer-directed rationales. After removing the explicit answer span, we compared the gold-choice probability induced by each rationale with that in-

Technique	α	Steps	B dir.	B reas.	D dir.	D reas.
Pure SFT	1.0	382	0.30	—	0.19	—
distill-SFT	1.0	382	0.27	0.43	0.15	0.24
distill-SFT	0.3	382	0.28	0.40	0.18	0.22
distill-SFT	0.1	382	0.29	0.41	0.18	0.24
Step-by-step	1.0	678	0.31	0.40	0.19	0.23
Step-by-step	0.3	678	0.33	0.39	0.21	0.23
Step-by-step	0.1	678	0.33	0.40	0.22	0.22

Table 1: **Stage 2 accuracy and optimizer steps.** Means across ten data splits. B is the transfer subset and D the held-out difficult subset; “dir.” and “reas.” denote direct and reasoning inference. A dash marks a regime not evaluated.

duced by a cross-question-shuffled rationale. No between-technique difference was detected in the mean probability shift on B or D. When both generators answered correctly, their rationales increased support for the gold answer; when both answered incorrectly, the effects reversed sign (Appendix Table B.4). These measures answer different questions: subset-B accuracy records whether the trained model crossed the final-answer decision boundary, whereas the frozen-reader contrast measures the mean probability shift induced by its rationale. The localized advantage in final decisions was therefore not accompanied by a detected difference in mean rationale influence. This dissociation reinforces the study’s central premise: end-task accuracy alone cannot determine which rationales are worth distilling, because their utility is conditional on the item and the receiving student.

Final-answer behavior was also sensitive to how rationale and answer losses were balanced. At full rationale weight, hundreds of rationale tokens dominated the short answer and reduced direct performance on held-out difficult items (subset D). Down-weighting the rationale improved distill-SFT direct accuracy at both $\alpha = 0.3$ and $\alpha = 0.1$, without a detected change under reasoning. Step-by-step showed the same pattern: the intermediate weight improved direct accuracy on B and the lowest weight improved it on D, without a detected reasoning improvement in either comparison (Table 1; paired contrasts in Appendix Table A.2). The transfer gain was not achieved by degrading the protected strata: distill-SFT was never significantly worse than step-by-step on A or C and was significantly better on both under reasoning at $\alpha = 0.3$ (Appendix Table B.5).

Reasoning selectively unlocks difficult-item gains. Reasoning did not improve every item, but its balance against the baseline student was favorable. Across the four test strata, the recovery of unknown and difficult items exceeded the erosion of already-known answers by a net 40.5 correct answers per split. This balance was measured against the baseline student and does not imply global superiority over pure SFT. The benefit was nevertheless selective: distill-SFT was most useful when reasoning-enabled trans-

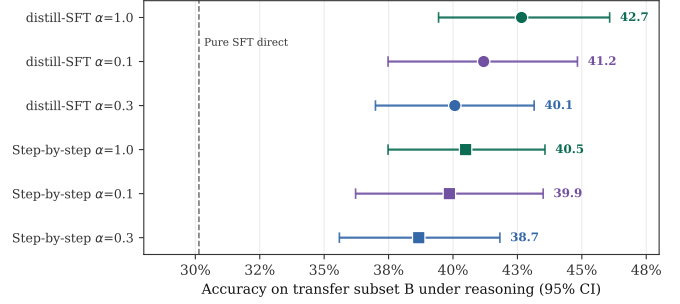


Figure 3: **A localized distill-SFT advantage under reasoning.** Points are ten-split means and lines are 95% two-level bootstrap intervals on transfer subset B under reasoning. Color encodes the rationale loss weight α and marker shape the technique. Distill-SFT is numerically ahead of step-by-step at every α , with a detected difference at $\alpha = 1.0$. The vertical dashed line marks pure-SFT direct answering. Table 1 shows the complementary direct-inference advantage of step-by-step.

fer was required, whereas step-by-step was stronger when the deployment objective was direct prediction. Because these experiments used gold-derived strata, they quantify the potential for selective reasoning rather than define a deployable router. Reasoning also increased decoding latency. A prospective gate should therefore predict the *benefit from reasoning*, rather than uncertainty alone [20, 21, 22, 23]. Appendix B reports the gain-loss decomposition and the routing and latency sensitivity analyses.

Rationale supervision mitigates forgetting of reasoning behavior. A student can retain an answer yet forget how to reason through it. We tested this risk on the out-of-domain ENEM examination. Direct-answer accuracy remained within one percentage point of the baseline across all adapted models, but the ability to externalize reasoning did not. The baseline model reached 0.85 with reasoning, whereas answer-only SFT fell to 0.75 and produced almost no reasoning text. Both rationale-trained students retained reasoning accuracy near 0.82; Fig. 4 reports the paired changes relative to the baseline. The rationale objectives therefore mitigated, but did not eliminate, forgetting of the pretrained reasoning behavior, consistent with rationale-replay findings in language-model adaptation [24].

The rationale objective also transferred the language of reasoning beyond the training domain. On ENEM, which was absent from the adaptation data, the baseline model reasoned predominantly in English, whereas the rationale-trained students externalized their reasoning in Brazilian Portuguese and did so more concisely. The change was therefore not limited to memorizing Portuguese explanations from the medical and multiprofessional residency examinations. It reflected an out-of-domain change in how the compact model expressed its reasoning, without establishing mechanistic faithfulness.

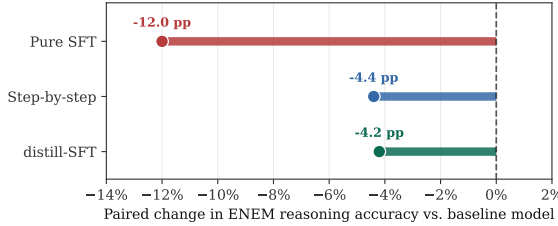


Figure 4: **Rationale supervision mitigates forgetting of pre-trained reasoning behavior.** Paired change in out-of-domain ENEM reasoning accuracy relative to the baseline model (ten data splits, $\alpha = 1.0$). Answer-only SFT shows the largest degradation; distill-SFT and step-by-step retain more of the pretrained reasoning behavior. The dashed line marks no change from baseline.

4. Discussion

The main result links the rationale diagnostic to the training experiment. A coherent rationale moved the student toward the teacher’s answer, whether that answer was right or wrong. A rationale from another question did not have the same effect. We therefore trained only with rationales attached to correct teacher answers. In the comparison across ten data splits, distill-SFT improved transfer on questions the baseline student had missed when the student reasoned. It also used 44% fewer optimizer steps than step-by-step distillation. Step-by-step remained stronger for direct answers. On ENEM, training with rationales reduced the loss of reasoning ability seen after pure SFT, addressing the risk of catastrophic forgetting. Together, these results connect the initial diagnosis to the choice of supervision, transfer, and retention.

Together, the Stage 1 diagnostic and the frozen-reader analysis sharpen this central argument. A localized final-answer gain without a detected difference in mean rationale influence shows why end-task accuracy alone cannot determine which traces are worth distilling. Rationale auditing is therefore not only a safeguard against misleading traces; it can become a selection mechanism that concentrates supervision on rationales that are question-relevant, answer-aligned, and informative for the target student. This student-aware view is consistent with recent rationale-selection methods in general and health-related reasoning distillation [10, 25, 26].

Distill-SFT placed the answer and rationale signals in one weighted completion, whereas the step-by-step comparator used the separate tasks introduced in prior work. In our generative setting, the step-by-step student required reasoning decoding to reach its best transfer accuracy. When both methods reasoned, their output lengths were similar, distill-SFT had a localized advantage on transfer questions, and single-target construction used substantially fewer optimization steps. Step-by-step was stronger under direct inference, and its direct mode remained competitive with pure SFT. The result is a reduction in training updates for the reasoning-enabled regime, not uniform compute superiority. Reasoning was also unnecessary for

every question: gains concentrated on transfer and difficult items, whereas direct answers remained preferable for knowledge already accessible to the student. These findings support a compact student with direct reasoning and behavioral behaviors and motivate a prospective gate that invokes reasoning only when it is likely to correct a direct response, consistent with selective-filtering approaches [23, 22].

The design had three strengths. We changed the rationale assigned to a question while keeping the student, prompt, and answer scoring fixed. We then tested the role of this alignment again during training. We also compared the methods on the same questions across ten data splits and reported transfer separately from answers the student already knew. Finally, ENEM let us assess whether the student retained its reasoning behavior outside the residency examination domain. These choices make the gains and losses easier to interpret without treating a difference that was not detected as proof that two methods are equivalent.

4.1. Limitations

The findings should be interpreted within the scope of the evaluated protocol. Answer verification applied only to the final option, so a correct answer could still rest on false, incomplete, or unsafe intermediate statements. The workflow is a first filter, not a factuality or clinical-safety assessment. The frozen-reader audit measured answer-directed influence rather than truth, completeness, or clinical safety; independent factual and clinical evaluation of generated rationales remains future work. Automatic answer extraction from unconstrained model outputs is a recognized source of evaluation error [27]. In our independent parsing audit, the estimated teacher accuracy changed by less than one percentage point, indicating that the main estimate was stable, although residual errors may affect a small fraction of teacher-correct and teacher-wrong assignments (details in Appendix B). The probability diagnostic was conditional on one reader, prompt, and restricted five-option scoring scheme. More broadly, the study covered one model family, one language, and multiple-choice examinations; selecting the keyed option does not establish open-ended clinical competence [28]. The ten data splits quantified sensitivity within the same item pool rather than across independent populations, and the gold-derived strata cannot serve as a deployed reasoning router. The study therefore establishes a controlled methodological foundation in medical-residency examinations, while broader prospective validation remains necessary before clinical deployment.

5. Conclusion

In Brazilian Portuguese residency examinations, question-relevant rationales moved a compact student toward the teacher’s answer, including when that answer

was wrong. This finding supports answer verification as a first filter for rationale supervision. Under paired evaluation, single-target distill-SFT used 44% fewer optimizer updates than step-by-step distillation and achieved higher accuracy on the target transfer subset under reasoning, whereas step-by-step was stronger under direct inference. Rationale supervision also mitigated forgetting of pretrained reasoning behavior outside the training domain. Together, these findings establish an auditable workflow for evaluating rationale supervision and its consequences when adapting compact models for health applications. Prospective evaluation on clinical tasks is needed to determine whether these benefits extend beyond examination performance.

Ethics statement

Ethics approval and informed consent were not required because the study involved no human participants, patient data, clinical records, or identifiable personal information. All experiments used publicly available examination questions and model-generated outputs.

Data availability

The study repository contains the question-level study data, model-generated teacher rationales, canonical seed definitions, aggregate analysis tables, and code needed to audit the reported analyses and construct comparable 85/15 seeded partitions. The original examining bodies retain all rights to the examination content; provenance and use restrictions are documented with the released data.

Code availability

Training, inference, masking self-tests, loss-weighting tests, and statistical analysis code are available at <https://github.com/martinellitiago/cot-distillation-medical-qa>. The versioned archive is available on Zenodo at <https://doi.org/10.5281/zenodo.22929248>.

CRediT authorship contribution statement

Tiago Martinelli: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing—original draft, Project administration. João Paulo Papa: Conceptualization, Validation, Supervision, Writing—review and editing. Adriano José Pereira: Conceptualization, Validation, Resources, Supervision, Funding acquisition, Writing—review and editing.

Declaration of AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used OpenAI Codex for language editing and Anthropic Claude Code to assist with the development and checking of analysis scripts. The authors reviewed and validated all text and code and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by capital expenditure (CAPEX) funding from Hospital Israelita Albert Einstein, a not-for-profit private hospital in São Paulo, Brazil, which coordinated the study. The funding provided a postdoctoral fellowship for TM and supported an academic visit by JPP as a visiting professor from São Paulo State University (UNESP), Bauru, Brazil.

Acknowledgements

The authors thank the Digital Transformation Office (DTO) and the leadership of Einstein Global Advanced Technologies for Equity for their support throughout this study. We also acknowledge the Institute of Science and Technology Itaú (ICTi) for providing the computational infrastructure used during the initial phase of this research. We also thank Cesar Gomes Miguel (Data Scientist Specialist), Gabriela Chiufta Tunes (Manager of Advanced Analytics and Multiomics), José Adenaldo Santos Bittencourt Júnior (Manager of High Performance Computing), and Edson Amaro Junior (Director of Global Data and Advanced Technologies) for their endorsement and assistance.

Appendix A. Training and analysis

Token-weighted distill-SFT loss. In distill-SFT, rationale tokens receive weight α and answer tokens receive weight 1 in Eq. 3. When no rationale is available, the target contains only the answer. We used $\alpha \in \{1.0, 0.3, 0.1\}$. For illustration, with 50 rationale tokens per answer token, these weights give the answer approximately 2%, 6%, and 17% of the total token weight, respectively. These percentages describe the weights assigned to tokens; the actual contribution to the loss also depends on each token’s prediction error. In step-by-step distillation, α weights the separate rationale task, while the answer task keeps weight 1. We evaluated both methods at all three weights and included pure SFT, giving seven configurations across ten data splits (70 runs).

Cross-question-shuffled control. Over the eligible teacher-correct items we draw, with a split-specific generator, a permutation π with no fixed point ($\pi(i) \neq i$ for all i), so each item receives another item’s rationale of comparable length, style, and language while its relevance to the question is destroyed. A hard assertion verifies zero self-assignments; across the ten canonical data splits none occurred.

Complementary training ablation and delimiter sensitivity. Matched students were trained with coherent or cross-question-shuffled rationale targets under the same optimization protocol and evaluated on a held-out subset B test set ($n = 164$ per split). Coherent rationales yielded mean accuracies of 0.390 with Qwen3 thinking delimiters and 0.391 as ordinary body text, whereas the corresponding shuffled-rationale accuracies were 0.237 and 0.285. No difference was detected between the two coherent-target formats (paired $p = 0.91$). Qwen3 supports thinking and non-thinking modes [15]; this secondary control does not establish equivalence between formats or imply that delimiters are generally irrelevant. It shows only that the delimiter provided no detectable benefit to coherent supervised transfer in this protocol.

Two-level bootstrap. We resample data splits with replacement and, within each chosen split, resample questions with replacement, over 10^4 replicates; the reported interval is the 2.5/97.5 percentile of the resulting distribution of the across-split mean.

Paired tests and effect reporting. The primary inferential comparison is a paired t test on the ten split-level accuracies, with the Wilcoxon signed-rank test as a sensitivity analysis. Within-split exact McNemar tests on discordant question pairs are item-level diagnostics; the minimum McNemar p across splits is descriptive and is never interpreted as a single global test. The manuscript focuses on three principal comparison families: content relevance, technique by inference regime, and rationale loss weight; delimiter sensitivity is treated as a secondary control. Other pairwise comparisons are treated as exploratory. Non-significant differences are reported as *not detected*, never as equivalence, because no equivalence margin was prespecified. Effects are given in percentage points and in questions gained or lost per split. Table A.2 consolidates the Stage 2 contrasts discussed in the main text. Its confidence intervals are paired t intervals computed from the ten split-level differences with nine degrees of freedom, distinct from the two-level bootstrap intervals for accuracy summaries. Tests and intervals are two-sided and unadjusted for multiplicity.

Confounder panel. We monitored three generation-related confounders: truncation at the generation budget (maximum 0.45%), answer emission (minimum 95.7%), and native-English reasoning under the reasoning regime (0%). None materially affected the principal comparisons.

Reproducibility. We used a single NVIDIA H100 80-GB GPU with bfloat16 precision. Inference used greedy decoding. The ten data-split seeds were 8, 12, 17, 23, 25, 31, 37, 44, 52, and 61; LoRA initialization used seed 3407. Latency measurements covered generation only, with GPU synchronization before and after each call and the first call after each restart excluded. Software versions and execution settings are documented at <https://github.com/martinellitiago/cot-distillation-medical-qa>.

Appendix B. Additional analyses

The teacher-answer check and the initial probability diagnostic used all 4,260 question records. These records contained 3,792 distinct normalized stems because some stems appeared more than once. All records were kept in the initial diagnostic; records sharing a stem were sampled together in the bootstrap. The lexical sensitivity checks then repeated the comparisons after excluding records with option-wording cues.

Appendix B.1. Answer extraction

We checked whether the rule used to extract the teacher’s final answer changed its measured accuracy. Extraction errors in free-text responses are a known source of evaluation error [27]. We applied a second parser that prioritized explicit final-answer phrases, such as **Respuesta correcta:** X, and the bold X) format. The two parsers returned different answers for 77 of 4,260 records (1.81%). Against the official answer key, only the original parser was correct in 44 cases, only the second parser was correct in 18, and both were wrong in 15. Estimated teacher accuracy changed from 74.9% to 74.3%. All training comparisons retained the original labels and item partitions. Neither parser was treated as a reference standard for ambiguous outputs.

Appendix B.2. Rationale probabilities

Answer scoring. For each question and rationale fraction, the reader produces logits for the five answer-letter continuations. Equation 1 applies a softmax only across these five candidates. The explicit final-answer span is removed from the rationale before scoring to prevent label leakage. Prefixes are evaluated at fractions 0, 0.15, 0.30, 0.45, 0.60, 0.75, 0.90, and 1.0. Coherent rationales are compared with a fixed cross-question-shuffled control, and all reported effects remain on the probability scale.

Probability contrasts. We made two comparisons, separately for questions the teacher answered correctly and incorrectly. Let g denote either group and $\bar{p}_{g,c}(f)$ the mean probability of the correct answer in that group, where c identifies the rationale condition and f is the fraction shown to the student. Thus $f = 0$ means no rationale and $f = 1$ means the complete rationale after removal of the explicit final answer.

Contrast	Set	Inference	Difference, p.p. (95% CI)	<i>t</i> - <i>p</i>	Wilcoxon <i>p</i>
<i>Technique and inference regime at $\alpha = 1.0$, unless specified</i>					
distill-SFT minus step-by-step	B	reasoning	+2.17 [+0.33, +4.00]	0.025	0.035
distill-SFT minus step-by-step	B	direct	−3.57 [−4.76, −2.38]	< 0.001	0.002
distill-SFT minus step-by-step	ABC	reasoning	+1.04 [−0.40, +2.48]	0.137	0.221
distill-SFT minus step-by-step	D	reasoning	+0.59 [−1.59, +2.78]	0.553	0.711
Technique-by-inference interaction	B	both	+5.73 [+3.00, +8.47]	0.001	0.004
Interaction at $\alpha = 0.3$	B	both	+6.22 [+2.21, +10.24]	0.007	0.014
Interaction at $\alpha = 0.1$	B	both	+4.62 [+0.46, +8.77]	0.033	0.025
Step-by-step: reasoning minus direct	B	both	+9.93 [+7.07, +12.79]	< 0.001	0.002
Step-by-step: reasoning minus direct	D	both	+3.71 [+1.86, +5.57]	0.001	0.006
Step-by-step minus pure SFT	B	direct	+0.42 [−2.13, +2.97]	0.719	0.691
Step-by-step minus pure SFT	D	direct	+0.30 [−1.93, +2.53]	0.770	0.641
<i>Rationale-weight contrasts: reduced weight minus $\alpha = 1.0$</i>					
distill-SFT, $\alpha = 0.3$	D	direct	+2.57 [+0.27, +4.88]	0.032	0.043
distill-SFT, $\alpha = 0.1$	D	direct	+2.82 [+0.75, +4.89]	0.013	0.012
distill-SFT, $\alpha = 0.3$	D	reasoning	−1.34 [−3.44, +0.77]	0.185	0.285
distill-SFT, $\alpha = 0.1$	D	reasoning	+0.35 [−1.71, +2.41]	0.712	0.900
Step-by-step, $\alpha = 0.3$	B	direct	+2.17 [+0.22, +4.12]	0.033	0.043
Step-by-step, $\alpha = 0.1$	D	direct	+2.33 [+0.59, +4.06]	0.014	0.010
Step-by-step, $\alpha = 0.3$	B	reasoning	−1.82 [−4.78, +1.15]	0.199	0.227
Step-by-step, $\alpha = 0.1$	D	reasoning	−1.09 [−2.76, +0.58]	0.174	0.219

Table A.2: **Paired training comparisons.** Differences are in p.p. across ten paired splits, with paired *t* intervals. The interaction is the reasoning contrast minus the direct contrast for distill-SFT versus step-by-step. Tests are two-sided and unadjusted; Wilcoxon tests provide a sensitivity check.

First, we measured the change after adding the question’s own rationale:

$$d_g = \bar{p}_{g,\text{coherent}}(1) - \bar{p}_{g,\text{coherent}}(0). \quad (\text{B.1})$$

Second, we compared that rationale with one assigned from another question:

$$C_g = \bar{p}_{g,\text{coherent}}(1) - \bar{p}_{g,\text{shuffled}}(1). \quad (\text{B.2})$$

Positive values indicate greater support for the correct answer; negative values indicate less support. The interaction compares this second effect between the two teacher groups:

$$I = C_{\text{correct}} - C_{\text{wrong}}. \quad (\text{B.3})$$

For example, $C_{\text{correct}} = +0.22$ and $C_{\text{wrong}} = -0.14$ give $I = 0.36$, or 36 p.p. This is a difference in probability shifts, not an accuracy gain.

Confidence intervals were calculated by resampling the stem groups described above.

Lexical-overlap sensitivity. We tested whether the probability shifts could be explained by rationales repeating the wording of the correct option. After normalizing case, accents, and punctuation, we flagged rationales containing the full correct option, provided it had at least two content words or 12 characters. We also measured the fraction of each option’s content words present in the rationale, excluding Portuguese stopwords and words from the question stem. When removing stem words left no content words, we used the option’s original content words. The overlap

margin was the coverage of the correct option minus the highest coverage among the incorrect options. We flagged values at or above its 90th percentile.

We repeated the comparisons after excluding either type of flagged record separately and then both together (Table B.3). This post-hoc analysis reused the saved probabilities and stem groups; the models were not run again. In each analysis, bootstrap samples included only the remaining records and their stem groups. After both exclusions, 2,920 records in 2,661 groups remained. The effects were +0.18 when the teacher was correct and −0.12 when it was wrong, with an interaction of 0.30 (95% CI 0.27 to 0.33). The same direction therefore persisted after removing these wording cues, although this check cannot rule out paraphrases of the correct option.

Generated rationales. We tested whether rationales produced by the trained models helped the baseline Qwen3-4B choose the correct answer. We used outputs from distill-SFT and step-by-step at $\alpha = 1.0$ on the same ten test splits of B and D. After removing the explicit final answer, we measured the baseline model’s answer probabilities with no rationale, with the question’s own rationale, and with one from another question. Its parameters remained unchanged throughout. We compared the last two conditions using the same paired tests and bootstrap procedure (Table B.4). We also summarized cases in which both trained models answered correctly or both answered incorrectly; these groups were defined after generation and are descriptive. This test measures answer influence within the Qwen family. It does not verify the factual content of

Sensitivity set	Records	Stem groups	C_{correct}	C_{wrong}	I (95% CI)
All records	4,260	3,792	+0.22	-0.14	+0.36 [+0.33, +0.39]
Exclude exact option wording	3,145	2,850	+0.18	-0.12	+0.30 [+0.27, +0.33]
Exclude high overlap margin	3,825	3,432	+0.21	-0.14	+0.35 [+0.32, +0.38]
Exclude both types of cue	2,920	2,661	+0.18	-0.12	+0.30 [+0.27, +0.33]

Table B.3: **Lexical sensitivity.** Records count question entries; stem groups collect repeated question stems. Effects follow Eqs. B.2 and B.3, with bootstrap intervals over the remaining groups.

Set	Generator outcomes	n	Distill	Step	Difference (95% CI)	t - p	Wilcoxon p
B	All	1,430	+0.21	+0.19	+0.02 [-0.02, +0.05]	0.183	0.232
B	Both correct	417	+0.59	+0.56	+0.03 [-0.02, +0.09]	0.120	0.193
B	Both wrong	658	-0.11	-0.09	-0.02 [-0.06, +0.02]	0.101	0.065
D	All	2,020	+0.08	+0.08	-0.01 [-0.04, +0.03]	0.630	0.770
D	Both correct	259	+0.64	+0.62	+0.02 [-0.07, +0.10]	0.553	0.322
D	Both wrong	1,333	-0.11	-0.10	-0.01 [-0.04, +0.01]	0.132	0.049

Table B.4: **Generated rationales.** Distill and Step show the change in correct-answer probability with the question’s own rationale versus another question’s rationale. Differences are distill-SFT minus step-by-step, with bootstrap intervals. n counts paired evaluations across splits. Outcome groups are descriptive; tests are unadjusted.

the explanation.

Appendix B.3. Transfer and retention

Table B.5 compares accuracy on A and C, the questions the baseline student already answered correctly. No lower accuracy was detected for distill-SFT than for step-by-step in these strata. At $\alpha = 0.3$, distill-SFT had higher accuracy on both under reasoning.

Sub.	Regime	α	Step	Distill	Δ (p.p.)	t - p
A	direct	1.0	93.1	93.9	+0.8	0.20
A	direct	0.3	93.4	93.4	+0.0	0.95
A	direct	0.1	92.7	93.1	+0.4	0.46
A	reasoning	1.0	82.0	82.7	+0.7	0.40
A	reasoning	0.3	80.0	82.2	+2.2	0.02
A	reasoning	0.1	81.9	82.5	+0.6	0.55
C	direct	1.0	79.6	80.7	+1.1	0.56
C	direct	0.3	78.0	80.7	+2.8	0.14
C	direct	0.1	77.8	80.9	+3.2	0.18
C	reasoning	1.0	61.1	61.1	+0.0	1.00
C	reasoning	0.3	60.7	67.2	+6.5	0.006
C	reasoning	0.1	63.2	62.6	-0.6	0.76

Table B.5: **Retention of previously correct answers.** Mean accuracy on subsets A and C in percent across ten splits. Differences are distill-SFT minus step-by-step in p.p., with paired t tests. Bold rows indicate $p < 0.05$.

Selective reasoning. With reasoning at $\alpha = 1.0$, distill-SFT gained a mean of 61.0 correct answers on B and 47.9 on D relative to the baseline student, but lost 47.4 on A and 21.0 on C. The net gain was 40.5 correct answers per split. Against pure-SFT direct answering, it gained 17.9 on B and 9.3 on D but lost 26.2 on A and 7.4 on C. Reasoning therefore helped on B and D without improving

overall accuracy over pure SFT. Table B.6 compares accuracy and latency when reasoning is used on all items or selected items. Selection by subset or by an oracle requires knowing which answers are correct, so these results describe potential gains rather than a tested deployment rule.

Set	Offline policy	Acc.	Latency (%)
ABC	answer all directly	0.72	1
ABC	route subset B to reasoning	0.77	31
ABC	reason on all items	0.68	100
ABC	per-question oracle	0.81	10
D	answer all directly	0.15	1
D	reason on all items	0.24	100
D	per-question oracle	0.31	17

Table B.6: **Selective reasoning.** Distill-SFT at $\alpha = 1.0$ across ten splits. Latency is relative to reasoning on every item. Subset and oracle policies use known answer outcomes and cannot be applied prospectively as tested.

References

- [1] K. Singhal, S. Azizi, T. Tu, et al., Large language models encode clinical knowledge, Nature 620 (2023) 172–180. doi:10.1038/s41586-023-06291-2.
- [2] T. Tu, M. Schaekermann, A. Palepu, et al., Towards conversational diagnostic artificial intelligence, Nature 642 (2025) 442–450. doi:10.1038/s41586-025-08866-7.
- [3] OpenAI, [Launching health in ChatGPT](#), accessed 27 July 2026 (Jul. 2026).

- [4] C.-Y. Hsieh, C.-L. Li, C.-k. Yeh, H. Nakhosht, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, T. 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, 2023, pp. 8003–8017. doi:10.18653/v1/2023.findings-acl.507.
- [5] K. Shridhar, A. Stolfo, M. Sachan, Distilling reasoning capabilities into smaller language models, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 7059–7073. doi:10.18653/v1/2023.findings-acl.441.
- [6] L. H. Li, J. Hessel, Y. Yu, X. Ren, K.-W. Chang, Y. Choi, Symbolic chain-of-thought distillation: Small models can also “think” step-by-step, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 2665–2679. doi:10.18653/v1/2023.acl-long.150.
- [7] Y. Tian, Y. Han, X. Chen, W. Wang, N. V. Chawla, Beyond answers: Transferring reasoning capabilities to smaller LLMs using multi-teacher knowledge distillation, in: Proceedings of the 18th ACM International Conference on Web Search and Data Mining (WSDM), Association for Computing Machinery, 2025, pp. 251–260. doi:10.1145/3701551.3703577.
- [8] M. Kang, S. Lee, J. Baek, K. Kawaguchi, S. J. Hwang, Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks, in: Advances in Neural Information Processing Systems, Vol. 36, 2023. doi:10.52202/075280-2109.
- [9] Y. Li, Y. Emad, K. Padthe, J. Lanchantin, W. Yuan, T. Nguyen, J. Weston, S.-W. Li, D. Wang, I. Kulikov, X. Li, [Naturalthoughts: Selecting and distilling reasoning traces for general reasoning tasks](#), arXiv preprint arXiv:2507.01921 (2025).
- [10] J. Yan, L. Liu, Y. Pan, S. Chen, Y. Xiang, B. Tang, Towards efficient CoT distillation: Self-guided rationale selector for better performance with fewer rationales, in: Findings of the Association for Computational Linguistics: EMNLP 2025, 2025, pp. 7818–7835. doi:10.18653/v1/2025.findings-emnlp.413.
- [11] Z. Shen, H. Yan, L. Zhang, Z. Hu, Y. Du, Y. He, CODI: Compressing chain-of-thought into continuous space via self-distillation, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 677–693. doi:10.18653/v1/2025.emnlp-main.36.
- [12] P. Payoungkhamdee, P. Limkonchotiawat, J. Baek, P. Manakul, C. Udomcharoenchaikit, E. Chuangsuwanich, S. Nutanong, An empirical study of multilingual reasoning distillation for question answering, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 7739–7751. doi:10.18653/v1/2024.emnlp-main.442.
- [13] P. Hager, F. Jungmann, R. Holland, et al., Evaluation and mitigation of the limitations of large language models in clinical decision-making, *Nature Medicine* 30 (2024) 2613–2622. doi:10.1038/s41591-024-03097-1.
- [14] R. K. Arora, J. Wei, R. S. Hicks, P. Bowman, J. Quiñonero-Candela, F. Tsimpouras, M. Sharman, M. Shah, A. Vallone, A. Beutel, J. Heidecke, K. Singhal, [HealthBench: Evaluating large language models towards improved human health](#), arXiv preprint arXiv:2505.08775 (2025).
- [15] A. Yang, A. Li, B. Yang, et al., [Qwen3 technical report](#), arXiv preprint arXiv:2505.09388 (2025).
- [16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, [LoRA: Low-rank adaptation of large language models](#), in: International Conference on Learning Representations, 2022.
- [17] P. W. Holland, Statistics and causal inference, *Journal of the American Statistical Association* 81 (396) (1986) 945–960. doi:10.1080/01621459.1986.10478354.
- [18] B. Efron, R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall/CRC, New York, 1993.
- [19] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* 12 (2) (1947) 153–157. doi:10.1007/BF02295996.
- [20] L. Chen, M. Zaharia, J. Zou, [FrugalGPT: How to use large language models while reducing cost and improving performance](#), Transactions on Machine Learning Research (2024).
- [21] I. Ong, A. Almahairi, V. Wu, W.-L. Chiang, T. Wu, J. E. Gonzalez, M. W. Kadous, I. Stoica, [RouteLLM: Learning to route LLMs from preference data](#), in: International Conference on Learning Representations, 2025.

- [22] Z. Sprague, F. Yin, J. D. Rodriguez, D. Jiang, M. Wadhwa, P. Singhal, X. Zhao, X. Ye, K. Mahowald, G. Durrett, [To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning](#), in: International Conference on Learning Representations (ICLR), 2025.
- [23] Y. Wu, Z. Zhang, H. Zhao, [Mitigating misleading chain-of-thought reasoning with selective filtering](#), in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 11325–11340.
- [24] J. Wu, Y. Zhao, Y. Xu, B. Liu, C. Zong, Boosting LLM translation skills without general ability loss via rationale distillation, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, 2025, pp. 12217–12236. [doi:10.18653/v1/2025.findings-acl.632](#).
- [25] H. Song, H. Lee, J. Shin, S. Cho, C. Ko, J. C. Park, Does rationale quality matter? enhancing mental disorder detection via selective reasoning distillation, in: Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 21738–21756. [doi:10.18653/v1/2025.findings-acl.1119](#).
- [26] Y. Yang, M. Lai, W. Zhao, X. Fan, Z. Xi, M. Wu, C. Huang, J. Zhao, H. Lv, J. Tong, Y. Zhou, Y. Zou, Q. Guo, T. Gui, Q. Zhang, X. Huang, Which reasoning trajectories teach students to reason better? a simple metric of informative alignment, in: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2026, pp. 42123–42150. [doi:10.18653/v1/2026.acl-long.1950](#).
- [27] Q. Yu, Z. Zheng, S. Song, Z. Li, F. Xiong, B. Tang, D. Chen, [xFinder: Large language models as automated evaluators for reliable evaluation](#), in: International Conference on Learning Representations (ICLR), 2025.
- [28] H. Wang, S. Zhao, Z. Qiang, N. Xi, B. Qin, T. Liu, [LLMs may perform MCQA by selecting the least incorrect option](#), in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 5852–5862.