LEARNING TO LEARN FROM CONTEXT: SYNTHETIC TRAINING FROM PERTURBED PUBLIC DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world tasks often require large language models (LLMs) to learn from complex task-specific context rather than pretrained parametric knowledge. This capability remains a weakness of LLMs, while human annotation for such task contexts is expensive and difficult to scale. Public high-quality documents are an abundant alternative, but much of the public web has already been consumed during pretraining: training on such documents naively would reward memorization rather than context learning. In this work, we attempt to make use of high-quality public documents with small perturbations and empirically find that LLMs can successfully generate context-dependent reasoning traces and answers, which are then used to train a student model. Specifically, we construct a synthesis pipeline that (i) rewrites source documents to reduce memorization risk, (ii) generates questions and rubrics that require reasoning over the document, (iii) answers the questions with the document as context, and (iv) admits only samples that genuinely depend on the document. Without any human annotators, our pipeline generates about 10k samples from 3.5k documents, and the resulting student model substantially improves the performance on CL-bench. SFT raises a Qwen3.6-35B-A3B student from 13.7% to 22.8%, and a subsequent rubric-reward RL stage reaches 24.6%, on CL-bench comparable with a frontier model of over a trillion parameters, Qwen3.8-2.4T (23.9%). We also observe a broad transfer of improvements to long-context understanding, instruction following, and reasoning, while code generation and knowledge remain mostly flat. We hope this work provides a reproducible and scalable way to improve the ability of LLMs to learn from context, and to facilitate further research on context-grounded reasoning.

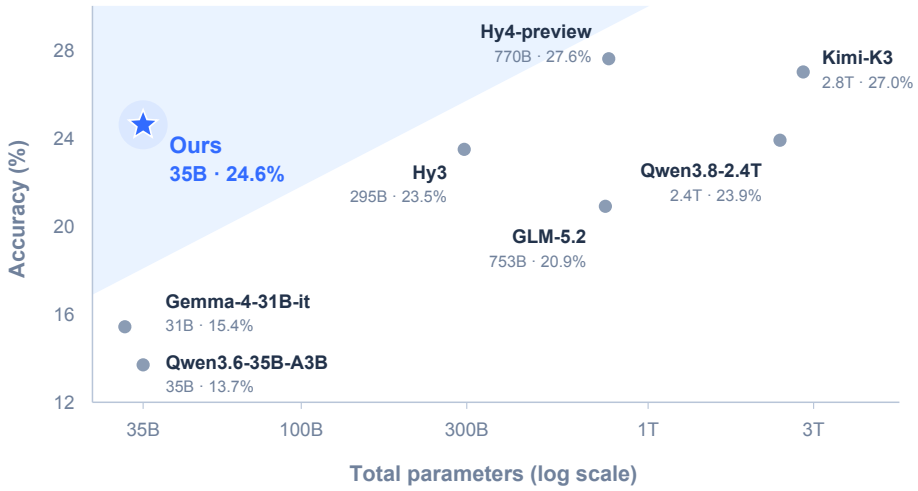

Figure 1: CL-bench accuracy versus total model parameters. Our 35B student (blue), trained on 10k synthetic samples from perturbed public documents, reaches 24.6%, comparable with frontier models that are 10–100× larger.

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.

1 INTRODUCTION

As the context windows of large language models (LLMs) expand to millions of tokens (Zeng et al., 2026; Qwen Team, 2026b; Bai et al., 2026), the ability to learn from context is becoming increasingly important. LLMs are increasingly deployed in settings where a provided context defines the task: answering questions over internal documents and manuals (Zou et al., 2025), reasoning over retrieved passages (Lewis et al., 2020), interpreting the tool outputs and logs that agents consume (Yao et al., 2023; Schick et al., 2023), and checking whether a plan complies with a specification (Yao et al., 2025). This capability, referred to as *context learning* (Dou et al., 2026b), is distinct from knowledge retrieval: rather than recalling a fact, the model must read a document and reason over it, which requires not only locating information in the text but also understanding the document as a whole and applying it to the task at hand. Such tasks arise commonly in real-world workflows.

Despite the ubiquity of this setting, context learning remains a weakness of LLMs, particularly for long documents that demand complex reasoning. Existing long-context benchmarks cover adjacent but distinct capabilities: synthetic retrieval suites such as RULER (Hsieh et al., 2024) measure effective context windows rather than learning, whereas real-document benchmarks such as LongBench v2 (Bai et al., 2025) and NoCha (Karpinska et al., 2024) probe deep understanding but score answers through multiple choice or binary judgments. Context learning, in contrast, asks whether a model can take up the rules, facts, and procedures that a document defines and apply them faithfully throughout its answer. CL-bench (Dou et al., 2026b), a benchmark of 1,899 long-document tasks spanning 18 domains, illustrates the difficulty: even strong frontier models solve fewer than 30% of the tasks, and the failure mode is characteristic in that models produce answers that appear plausible yet collapse under fine-grained checking.

This weakness does not stem from a lack of exposure to long documents, since pretraining corpora contain enough of them to cover most of the domains where context learning matters (Gao et al., 2020; Soldaini et al., 2024; Weber et al., 2024). Pretraining teaches models what documents say, but it does not teach them to answer questions *conditioned* on a document, and having read a text therefore does not imply knowing how to use it, not to mention how to apply it to a new task, especially when the document is long and complex. Instruction-tuning data could in principle provide such supervision, but two obstacles intervene. First, mainstream instruction datasets are conversational and short-context, and convey little about the faithful use of long documents. Second, human annotation for long-document tasks is exceptionally expensive: writing a single question requires the annotator with expertise to design a document of tens of thousands of words, verify the answer against it, and specify what a correct answer must contain.

A growing body of work therefore constructs long-context supervision synthetically, yet where the grounding documents come from remains a largely open choice. One branch grounds supervision in real documents: contexts are either assembled from multiple passages (Chen et al., 2025; Yang et al., 2025; He et al., 2025), taken from single long documents in specific domains (Lin et al., 2025; Pham et al., 2025; Zhang et al., 2025). These pipelines, however, may raise memorization concerns, as such publicly available documents are plausible constituents of pretraining corpora (Pham et al., 2025). The other branch synthesizes the documents themselves, for example by constructing tasks over programmatically generated contexts (Zhao et al., 2024) or by producing document-grounded and conversational data through prompt-based generation (Subramanian & Verma, 2025), which is easy to obtain and scales beyond the length distribution of any curated corpus. The generated contexts, however, tend to be logically simpler and more homogeneous than real ones, and the gap is measurable: models fine-tuned on synthetic haystacks consistently underperform those fine-tuned on real documents, even though the induced retrieval behavior partially overlaps (Zhao et al., 2025). Both branches thus have their limitations: real documents carry memorization risk, and synthetic ones are logically simpler and more homogeneous.

In this work, we mitigate the memorization risk of real documents: we perturb them so that the teacher can no longer answer from parametric memory and must instead extract information and reason over the document content. Specifically, we build a synthesis pipeline on top of small perturbations to public documents. The pipeline first rewrites each source document with entity renaming and numeric perturbation, then generates questions that require reasoning over the document with LLMs. It then applies a gap check to admit only samples that genuinely depend on the document, which compares the rubric pass rate of answering with and without the document. In one memoriza-

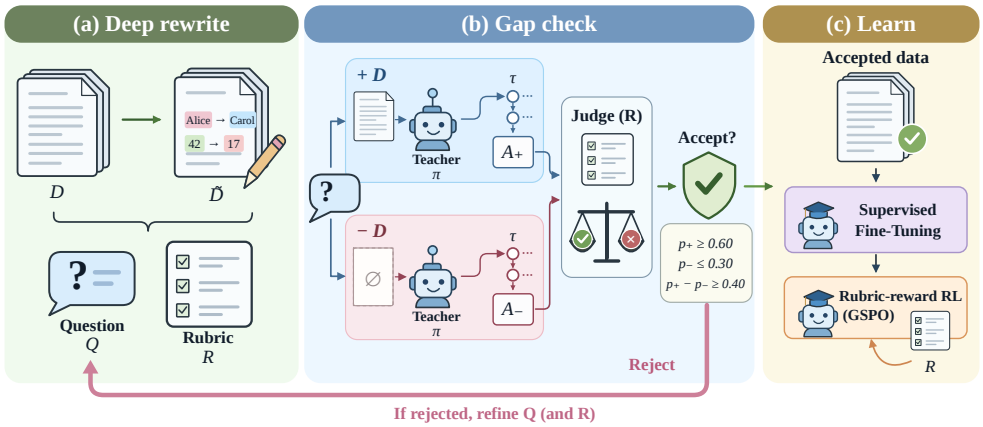

Figure 2: Overview of the synthesis pipeline. (a) Source documents are deep-rewritten ($D \rightarrow \tilde{D}$) and paired with generated questions and rubrics; (b) each question is answered with and without the document, and only document-dependent samples are admitted; (c) the student is trained on the admitted samples with the teacher's native reasoning traces, followed by a subsequent rubric-based reinforcement learning process.

tion audit, a model that answered questions about a well-known technical standard perfectly with no document in context scored nearly zero once the standard was rewritten, because its memorized knowledge collided with the rewritten entities. Without any human annotators, the pipeline produces about 10k samples from 3.5k public documents, ranging from legal documents to game manuals.

Our experiments show that training on this data yields substantial gains. Supervised fine-tuning raises a Qwen3.6-35B-A3B student (Qwen Team, 2026a) from 13.7% to 22.8% on CL-bench, and a subsequent rubric-reward RL stage reaches 24.6%, comparable with HY3 (Tencent Hunyuan Team, 2026b) (23.5%) and Qwen3.8-2.4T (Qwen Team, 2026b) (23.9%). The improvements also transfer beyond the target benchmark: long-context understanding (AALCR (Artificial Analysis Team, 2025) improves from 62.6 to 69.8), instruction following (IFBench (Pyatkin et al., 2026) from 57.7 to 71.7), and reasoning (ARC-AGI-1 (Chollet, 2019) from 47.4 to 63.8) all improve substantially, whereas knowledge benchmarks decline slightly (e.g., SimpleQA (Wei et al., 2024) drops from 20.9 to 18.7), consistent with training data that teaches the use of context rather than world knowledge. We believe this work provides a feasible and scalable path to improving context learning: it turns public documents, even those the model has already read, into effective supervision without any human annotation, and we hope it will facilitate further research on context-grounded reasoning.

2 The Synthesis Pipeline

We build a pipeline that constructs training samples around publicly available documents: each sample pairs a document D with a question Q, a rubric list R, and an answer generated by a teacher model π. The construction enforces two requirements: (a) *context dependence*: the question must be unanswerable from parametric memory, enforced jointly by deep rewriting and the gap check; and (b) *native reasoning traces*: training targets preserve the teacher's verbatim reasoning process. Each admitted sample carries its rubric list R as metadata, making the same artifacts directly reusable as RL prompts with rubric-based rewards. Figure 2 shows an overview of the pipeline.

2.1 Source Acquisition and Deep Rewriting

Sources. Instead of human annotation, we collect existing documents from publicly available sources to serve as contexts. Documents created in the most recent six months are preferentially selected to reduce memorization, consistent with the contamination-free design of CL-bench (Dou

et al., 2026b). We empirically find that contexts synthesized directly by LLMs tend to be logically simple and heavily homogeneous, and models trained on such data improve little. Following the subcategory taxonomy in CL-bench, we cover 16 of its 18 subcategories. For each subcategory, we first identify suitable data sources and then crawl documents accordingly. The final corpus comprises 3,515 documents spanning IETF RFCs, SEC filings, court opinions, arXiv papers, public handbooks, wikis, and encyclopedias. For more details, please refer to Appendix A.

Deep rewriting. We observe that when asking questions on the crawled documents directly, LLMs' reasoning has the risk of deriving from pre-trained knowledge rather than the given context. Therefore, we apply a deep rewriting step to each document to reduce the memorization risk. Each document D is rewritten by GLM-5.2 into $\tilde{D}$ with proper-noun renaming, numeric perturbation, section renumbering, and list-order shuffling. While the rewriting is not guaranteed to be perfect, it is sufficient to make the original document and the rewritten one detectably different (Section 2.4). We note that rewriting is not intended to distinguish the document from its original source, but rather to force the teacher to reason over the document, reveal the reasoning trace that depends on the context and thus teach the student to do the same.

2.2 Question generation

For each rewritten document, a single batched generator call produces up to three questions Q, each paired with a rubric list R of at least seven rubrics. Each question includes a system prompt and a user question. We define 9 persona archetypes (such as named professional roles, system bots and roleplay characters), 7 question types (such as deep reasoning, factual retrieval, and calculation) and 4 rubric types (process, content, persona, format). For each question, the generator samples a persona archetype and a question type, then use the corresponding prompt template to generate a question and its rubric list. Each question is generated together with the rubrics covering different rubric types, of which at least five must cite specific values from the document. Finally, each sample consists of a system prompt, a user prompt including the document followed by a question, and a rubric list.

2.3 Answer generation

The teacher π answers each question with the rewritten document $\tilde{D}$ as context with its native reasoning trace preserved: the training target is the verbatim reasoning-content followed by the answer. We do not perform any post-hoc reasoning synthesis or answer revision, as we emprically find that the teacher's native reasoning trace is more effective than a polished or cleaned version, though it may fail to satisfy some rubrics.

2.4 The gap check

Each question Q is answered twice by the teacher π: once with the rewritten document $\tilde{D}$, yielding answer A_+, and once without, yielding A_-. When the document is omitted from the context, we explicitly instruct the model to answer the question as best as it can from memory. A judge τ scores both answers against the rubrics R, giving pass rates p_+ and p_-, and the sample is admitted only if $p_+ \geq 0.60$ (the question is answerable from the document), $p_- \leq 0.30$ (the question is not trivially answerable from memory), and $p_+ - p_- \geq 0.40$ (the document is necessary). Note that some rubrics are expected to pass without the document (e.g., format and persona rubrics, anti-hallucination rubrics), so p_- is not expected to be zero.

Since 32.5% of the generated samples fail the gap check, we add a retry mechanism to increase the yield. When a sample is rejected, we regenerate the question and rubrics (Q and R) with the same document and the rejection reason as additional context, and then repeat the answer generation and gap check. If the regenerated question passes the gap check, it is admitted; otherwise, it is discarded.

Table 1: Main results on CL-bench and CL-bench Life under the GPT-5.1 judge. "Teacher" indicates the model used to generate the SFT data. The baseline is Qwen3.6-35B-A3B.

Model	Teacher	CL-bench	CL-bench Life
Qwen3.6-35B-A3B (Baseline)	—	13.7	8.4
Gemma-4-31B-it	—	15.4	10.4
GLM-5.2	—	20.9	12.3
Qwen3.8-2.4T	—	23.9	17.3
Kimi-K3	—	27.0	20.0
Hy4-preview	—	27.6	21.5
SFT	GLM-5.2	19.8	10.4
SFT	Kimi-K3	19.7	10.4
SFT	Qwen3.8-2.4T	22.8	12.6
RL	—	16.9	9.1
SFT + RL	GLM-5.2	24.5	9.1
SFT + RL	Qwen3.8-2.4T	**24.6**	**13.8**

3 EXPERIMENTS

3.1 SETUP

We empirically verify the effectiveness of the synthetic dataset through both supervised fine-tuning (SFT) and reinforcement learning (RL) with rubric-reward optimization. We use Qwen3.6-35B-A3B as the student model for all experiments, GLM-5.2 (Zeng et al., 2026) for question generation and Qwen3.8-2.4T (Qwen Team, 2026b) as the teacher model by default for SFT data generation. The synthesized dataset contains 9,625 training samples with 3,515 unique documents. The SFT training takes 3 epochs, with a global batch size of 32 samples and a maximum sequence length of 131K. All models are trained with AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$. We use a cosine learning rate decay schedule with a peak learning rate of 5e-6 and a warmup of 10% of the total training steps. The final learning rate is 1e-7. We use a weight decay of 0.1 and a gradient clipping of 1.0.

We also run GSPO (Zheng et al., 2025) on the same prompts, with each sample's rubric list carried as metadata. The reward is the mean of per-rubric pass/fail statuses judged by Qwen3.5-397B-A17B (Qwen Team, 2025), with the judge prompt copied verbatim from the official CL-bench evaluation (Dou et al., 2026b). We use the supervised fine-tuned model as the initial policy. The RL training uses a KL penalty coefficient of 0.001 and a constant learning rate of 1e-6.

We evaluate all models on CL-bench (Dou et al., 2026b) and CL-bench Life (Dou et al., 2026a), using GPT-5.1 judge with low reasoning effort for CL-bench and high reasoning effort for CL-bench Life following Dou et al. (2026b). We report the task accuracy (all-or-nothing over rubrics) as the main metric. We compare the performance of our models with several public reference models, including Qwen3.6-35B-A3B, Gemma-4-31B-it (El Abd et al., 2026), GLM-5.2 (Zeng et al., 2026), Qwen3.8-2.4T (Qwen Team, 2026b), Kimi-K3 (Bai et al., 2026), and Hy4-preview (Tencent Hunyuan Team, 2026a). We also evaluate the general capabilities of the models on various tasks, which includes long-context understanding, instruction following, reasoning, code generation, and knowledge. The detailed benchmark list and evaluation protocols are provided in Appendix B.

3.2 RESULTS

3.2.1 SUPERVISED FINE-TUNING

Table 1 shows the results on CL-bench and CL-bench Life. Synthetic-data SFT improves the student substantially over the baseline with an accuracy of 22.8 on CL-bench. It also achieves 12.6 on CL-bench Life. The outcome is strongly teacher-dependent: the performance of students trained on identical questions but answers from different teachers varies from 19.7 to 22.8 on CL-bench. We observe that the student performance does not necessarily follow the teacher performance. Kimi-K3

Table 2: General capabilities evaluation board, judged by gpt-oss-120b. The first two columns are public reference models; the SFT and SFT+RL rows use the Qwen3.8-2.4T teacher. Small parentheses give the change over the baseline: red for gains, green for drops. The full board is reported in Appendix B.

Benchmark	HY3-preview	Qwen3.5-397B	Baseline	SFT	SFT + RL
Long context					
LongBench v2	62.2	63.3	58.1	61.0 (+2.9)	60.8 (+2.7)
AALCR	67.0	71.2	62.6	68.0 (+5.4)	69.8 (+7.2)
MRCR-4needle	33.0	67.8	69.7	71.8 (+2.1)	71.9 (+2.2)
Instruction following					
IFBench	60.8	70.0	57.7	71.7 (+14.0)	69.3 (+11.6)
IFEval	93.4	95.7	94.2	95.1 (+0.9)	95.9 (+1.7)
AdvancedIF	66.8	59.2	59.6	62.9 (+3.3)	65.1 (+5.5)
Reasoning					
ARC-AGI-1	53.5	69.3	47.4	63.8 (+16.4)	51.8 (+4.4)
ARC-AGI-2	3.8	15.0	6.8	18.6 (+11.8)	9.2 (+2.4)
GPQA-Diamond	85.4	87.9	83.3	88.4 (+5.1)	90.4 (+7.1)
HMMT-2025	92.6	89.1	88.4	92.5 (+4.1)	93.7 (+5.3)
Code					
LiveCodeBench-v6	81.8	92.4	85.7	86.2 (+0.5)	85.5 (−0.2)
OJBench	56.5	50.4	41.4	40.5 (−0.9)	37.5 (−3.9)
Knowledge					
MMLU-Pro	82.9	87.4	84.7	85.4 (+0.7)	85.0 (+0.3)
CEval	93.4	93.8	92.0	91.0 (−1.0)	90.6 (−1.4)
SuperGPQA	67.2	71.0	66.0	66.6 (+0.6)	66.1 (+0.1)
SimpleQA	30.5	52.4	20.9	18.7 (−2.2)	20.4 (−0.5)

(Bai et al., 2026), the strongest teacher (27.0), yields the weakest student (19.7), while Qwen3.8-2.4T, the mid-ranked teacher (23.9), yields the best (22.8). Appendix C.1 further analyzes this phenomenon.

3.2.2 Reinforcement learning

The bottom block of Table 1 shows the effect of rubric-reward RL. On top of the best SFT model, RL improves CL-bench from 22.8 to 24.6 and CL-bench Life from 12.6 to 13.8, which are the best scores obtained in this series. Notice that despite the large gap between the SFT model with GLM-5.2 and Qwen3.8-2.4T as teachers (19.8 vs 22.8), these two models achieve similar CL-bench accuracy after RL (24.5 vs 24.6). Directly applying RL to the base model without SFT initialization improves CL-bench from 13.7 to 16.9, which is substantially worse than initializing from SFT. An intuitive explanation is that the SFT stage provides a strong initialization under which the student already occasionally satisfies most rubrics of a task, which is necessary for the RL stage to receive informative reward signals.

Though the performance on CL-bench is similar regardless of the teacher of the initial SFT model, the CL-bench Life performance is still teacher-dependent: the GLM-5.2-initialized model achieves 9.1, while the Qwen3.8-2.4T-initialized model achieves 13.8. This is consistent with the teacher capabilities: GLM-5.2 is weaker than Qwen3.8-2.4T on CL-bench Life, and the initialization retains an advantage where the task distribution departs from that of the training.

3.2.3 General capability improvements

Table 2 reports general capabilities, with two public models as reference. The complete result is given in Appendix B. Relative to the baseline, SFT improves long-context understanding (LongBench v2 +2.9, AALCR +5.4, MRCR +2.1), instruction following (IFBench +14.0, AdvancedIF +3.3), and reasoning (ARC-AGI-1 +16.4, ARC-AGI-2 +11.8, GPQA +5.1, HMMT-2025 +4.1) substantially. Code generation is essentially unchanged (LiveCodeBench +0.5, OJBench -0.9), and knowledge is mostly flat with small declines (CEval -1.0, SimpleQA -2.2, against MMLU-Pro +0.7 and SuperGPQA +0.6). Against the public reference model, the 35B student becomes competitive

on long-context tasks—AALCR 69.8 versus 67.0 for HY3-preview and 71.2 for Qwen3.5-397B—while trailing clearly on knowledge (SuperGPQA 66.1 vs 71.0, SimpleQA 20.4 vs 52.4). This is consistent with the design of the SFT data, which focuses on long-context learning and does not attempt to teach general knowledge.

RL preserves most of the gains over the baseline, but does not consistently improve upon SFT: it improves AALCR, GPQA-Diamond, and HMMT-2025, while the performance declines notably on ARC-AGI-1 (-12.0) and ARC-AGI-2 (-9.4). The rubric-reward objective therefore retains broad general-capability gains, but its additional benefit is concentrated on the target capability rather than transferring uniformly across the general board.

4 Ablations

In this section, we introduce ablation experiments that verifies our design choice. The experiment setup is the same with those in Section 3.1 unless otherwise specified.

4.1 The Gap Check

In Section 2.4, we apply a strict gap check to the data pipeline in order to pick questions that can only be answered with the context. In practice, this process filters out 32.5% of the generated samples before the retry. Here we ask the following question: *whether the gap check is necessary, and what is its effect on the final model performance.*

Model	CL-bench
Baseline	19.17
w/o gap check	17.75

Table 3: Effect of the gap check. Both models are trained on 3k samples. The teacher model is GLM-5.2.

To investigate the effect of the gap check, we train two models with the same SFT recipe, one on the standard training dataset and the other on the unfiltered training dataset. Both the datasets contain 3,000 samples. The only difference is that the standard dataset is a random subset of the train set that passes the gap check, while the unfiltered dataset is a random subset of the dataset before the gap check and right after the answer generation in Section 2.3. The teacher model is GLM-5.2 in this experiment.

Table 3 shows the results. The model trained on the standard dataset achieves 19.17% accuracy, while the model trained on the unfiltered dataset achieves 17.75% accuracy. This indicates that the gap check is effective in filtering out samples that do not require context to answer, and training on such samples can hurt the final model performance.

4.2 Data scaling

Our data pipeline successfully generates 9,625 samples from 3,515 documents without human annotation. In this section, we ask *how the model performance scales with the number of samples and whether the current 10k data pool is sufficient to saturate the model's learning capacity.*

To answer this question, we train the same SFT recipe on nested subsets of the 9,625-sample pool, with sample counts of 100, 300, 1,000, 3,000, and 9,625 (the full pool). All data are randomly sampled from the full pool. The experiment follows the same setup as in Section 3.

The results are shown in Figure 3. It is clear that the model performance improves monotonically as the number of samples increases, from 13.7 for the untrained base to 22.8 at the

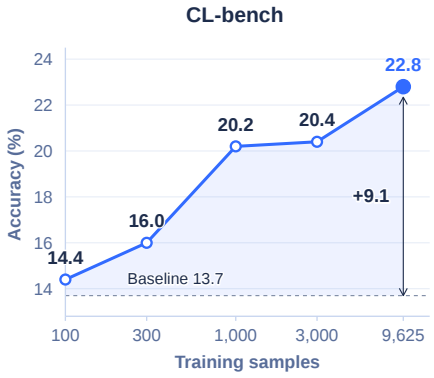

Figure 3: Data scaling curve. The dashed line marks the untrained base.

Table 4: Effect of the RL reward function. We report the CL-bench accuracy and the median think and answer lengths of the model (tokens, tokenized with the student's Qwen3.6 tokenizer).

Model	CL-bench	Think length	Answer length
SFT	22.80	7,626	1,492
Standard RL	24.60	10,481	2,569
w/ $\alpha = 0.5$	22.80	14,398	2,116
w/ length penalty	20.91	4,012	1,754

full pool. In particular, the performance grows
rapidly from 16.0 to 20.2 when the sample count increases to 1,000, and continues to improve to 22.8 at the full pool. Within the tested range the curve shows no saturation, so the model's learning capacity may not be exhausted by the current 10k pool.

4.3 The RL Reward

In Section 3 we show that the rubric-reward RL stage improves the SFT model from 22.8 to 24.6 on CL-bench. We further investigate the effect of the RL reward by looking into the model behavior before and after RL, only to find that after the RL stage, the reasoning trace and answer length are much longer than that of the SFT model. Besides, the RL model has a higher failure rate (about 2%) on the rubrics that require the model to avoid hallucination, which indicates that the RL model tends to generate more content to satisfy the rubrics, but at the cost of hallucination. Therefore, we wonder *whether the RL reward could be improved to avoid the potential reward hacking behavior and improve the model performance.*

We design two ablation experiments to investigate the effect of the RL reward.

Ablation 1: reward with length penalty. We add a length penalty to the reward function, which penalizes the model for generating longer reasoning traces. The length penalty starts from 8k tokens and increases linearly to 16k tokens, after which the reward will be penalized to 0.

Ablation 2: reward with additional overall pass rate. The original reward function is the rubric-level pass rate, while we care more about the overall task-level pass rate when evaluating the model. Since directly using the task-level pass rate as the reward is not feasible due to the sparsity of the reward signal, we make a compromise by adding the overall task-level pass rate to the reward function, which is an average of the rubric-level pass rates and the overall task-level pass rate. That is, the reward function is defined as:

$$\text{reward} = \frac{\alpha}{N} \sum_{i=1}^{N} \mathbb{1}[r_i] + (1 - \alpha) \prod_{i=1}^{N} \mathbb{1}[r_i] \tag{1}$$

where N is the number of rubrics for a given task, and $\mathbb{1}[r_i]$ is an indicator function that equals 1 if the ith rubric is passed, and 0 otherwise, and α is a hyperparameter that controls the weight of the rubric-level pass rate and the overall task-level pass rate. When $\alpha = 1$, the reward function is the same as the original rubric-level pass rate.

Table 4 shows the results. With the length penalty, the model generates much shorter reasoning traces, but the CL-bench accuracy drops significantly from 24.6 to 20.91. The length penalty discourages the model from generating longer reasoning traces, which may hurt the model's ability to complete the tasks that require many reasoning steps. With the additional overall pass rate, the model generates longer reasoning traces, but the CL-bench accuracy does not improve from the SFT model. A possible explanation is that the overall pass rate is a sparse reward signal, and the model may not receive enough feedback to improve its performance under this reward function.

4.4 The Performance on Fictional Contexts

Since our training data is built on public documents, it is natural to ask *whether the gains concentrate on public-document tasks or can transfer to fictional contexts.*

Table 5: Accuracy by context provenance on CL-bench. Provenance labels come from an unofficial audit of the 500 benchmark contexts. The unclear tasks are grouped with fictional. SFT and SFT+RL use the Qwen3.8-2.4T teacher, matching Table 1.

Context group	Base	SFT	SFT+RL
Public documents (836)	9.0	12.2	14.1
Fictional / unclear (1,063)	17.5	31.1	32.9

We analyze the provenance of the CL-bench contexts and roughly classify them into two groups: public documents (including contents with modifications) and fictional contexts (including unclear contexts). We find 277 of the 500 CL-bench contexts (55.4%) are public documents, 36 (7.2%) are modified public documents, 184 (36.8%) are fictional, and 3 (0.6%) are unclear. Among the 1,899 tasks, 836 (44.0%) are based on public documents, 1,063 (56.0%) are based on fictional contexts. These classification results are not official but only a rough estimate based on our audit of the 500 contexts, while we believe they are sufficient to support the analysis in this section.

Table 5 gives the accuracy by context provenance on CL-bench. Surprisingly, most of the performance gain happens on fictional contexts instead of public documents. The baseline model achieves 17.5% accuracy on fictional contexts, while the SFT model improves it to 31.1% and the SFT+RL model further improves it to 32.9%, almost doubling the baseline performance. In contrast, on public documents, the model only improves from 9.0% to 12.2% under SFT and to 14.1% under SFT+RL. It is possible that the tasks on public documents are more difficult than those on fictional contexts, but this phenomenon is still enough to show that the training data teaches a general reasoning mode that transfers to fictional contexts.

5 Related Work

Synthetic supervision over long documents. There are several different approaches to generate synthetic supervision for long-context learning. One branch grounds supervision in real documents: contexts are assembled from multiple passages (Chen et al., 2025; Yang et al., 2025; He et al., 2025), or taken from single long documents in specific domains (Lin et al., 2025; Pham et al., 2025; Zhang et al., 2025)—at the risk that such public documents are plausible constituents of pretraining corpora (Pham et al., 2025). The other branch synthesizes the documents themselves, programmatically (Zhao et al., 2024) or through prompt-based generation (Subramanian & Verma, 2025), which scales freely but yields logically simpler, more homogeneous contexts. Our pipeline takes another path: it keeps real documents and perturbs them, so that the model can no longer answer from parametric memory and must instead reason over the document content.

Context–memory conflict resolution. Entity substitution has likewise been used to build training data for context–memory conflict resolution (Ju et al., 2026), and synthetic tasks with rule-based rewards to improve contextual faithfulness (Si et al., 2026). In our pipeline, the perturbation plays an instrumental role rather than a behavioral one: it is there to elicit document-grounded reasoning traces from the teacher, not to shape how the model resolves conflicts between context and memory.

6 Conclusion

In this work, we present a synthetic supervision pipeline that generates 10k long-context question-answering samples from 3.5k public documents without any human annotation. We show that through perturbation with document-grounded gap checks, the resulting dataset successfully trains a 35B student model to achieve strong performance on CL-bench and CL-bench Life, and the gains transfer broadly to long-context understanding, instruction following, and reasoning. Our analysis suggests that such an approach provides a feasible and scalable way to improve the general context learning ability of large language models beyond the training distribution. We hope that this work can inspire future research on synthetic supervision for long-context learning and other challenging tasks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search, experiment-log aggregation, drawing figures and polishing the writing of the manuscript. All AI-assisted content was reviewed, verified, and edited by the authors. All experimental results were produced by the authors' own training and evaluation pipelines. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We find the official CL-bench judge prompt will trigger OpenAI's content moderation in bulk (HTTP 400), which is reproducible and reported on GitHub[1]. Therefore, we replace the full-width brackets in the judge prompt with half-width brackets, with the rest of the prompt verbatim identical. This replacement has the risk of changing the judge behavior and introducing bias that affect the reproducibility of the results. The result should be closed to those obtained with the official prompt, but we cannot guarantee that they are identical.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, et al. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Artificial Analysis Team. Artificial analysis long context reasoning benchmark(lcr), 2025.

Tongtong Bai, Yifan Bai, Yiping Bao, M. C., Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Guanduo Chen, Guangyu Chen, Guanzheng Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kexin Chen, et al. Kimi k3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, Haijun Lv, Yicheng Zou, Hang Yan, Kai Chen, and Dahua Lin. What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27129–27151, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1316. URL https://aclanthology.org/2025.acl-long.1316/.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

François Chollet. On the measure of intelligence, 2019. URL https://arxiv.org/abs/1911.01547.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: MathArena as an evaluation platform for mathematics with LLMs. *arXiv preprint arXiv:2605.00674*, 2026.

[1] https://github.com/Tencent-Hunyuan/CL-bench/issues/26

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E. Primack, Summer Yue, and Chen Xing. MultiChallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.958. URL https://aclanthology.org/2025.findings-acl.958/.

Shihan Dou, Yujiong Shen, Chenhao Huang, Junjie Ye, Jiayi Chen, Junzhe Wang, Qianyu He, Shichun Liu, Changze Lv, Jiahang Lin, Jiazheng Zhang, Ming Zhang, Shaofan Liu, Tao Ji, Zhangyue Yin, Cheng Zhang, Huaibing Xie, Jianglu Hu, Jingcheng Deng, Lincheng Li, Minda Hu, Shaolei Wang, Syrus Zhao, Weichao Wang, Yan Lei, Yang Liu, Yanling Xiao, Yiting Liu, Zenan Xu, Zhen Guo, Ziliang Zhao, Pluto Zhou, Tao Gui, Qi Zhang, Xuanjing Huang, Yu-Gang Jiang, Di Wang, and Shunyu Yao. Cl-bench life: Can language models learn from real-life context?, 2026a. URL https://arxiv.org/abs/2604.27043.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, et al. Cl-bench: A benchmark for context learning. *arXiv preprint arXiv:2602.03587*, 2026b.

Xeron Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shuyue Guo, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi LI, et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=6WgflzYQpf.

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, et al. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262/.

Linda He, Jue WANG, Maurice Weber, Shang Zhu, Ben Athiwaratkun, and Ce Zhang. Scaling instruction-tuned LLMs to million-token contexts via hierarchical synthetic data generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BkwCrIsTbR.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, Shengjie Bi, Shishir G Patil, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujan Kumar Gonugondla, Hunter Lang, Yue Yu, Yundi Qian, Maryam Fazel-Zarandi, Licheng Yu, Amine Benhalloum, Hany Hassan Awadalla, and Manaal Faruqui. AdvancedIF: Rubric-based benchmarking and reinforcement learning for advancing LLM instruction following. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pp. 18003–18022, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.820. URL https://aclanthology.org/2026.acl-long.820/.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=fOrm2rGX2r.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Li Ju, Junzhe Wang, and Qi Zhang. Faithfulness-qa: A counterfactual entity substitution dataset for training context-faithful rag models. *arXiv preprint arXiv:2604.25313*, 2026.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A "novel" challenge for long-context language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17048–17085, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.948. URL https://aclanthology.org/2024.emnlp-main.948/.

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

Jingyang Lin, Andy Wong, Tian Xia, Shenghua He, Hui Wei, Mei Han, and Jiebo Luo. Facilitating long context understanding via supervised chain-of-thought reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12232–12248, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.615. URL https://aclanthology.org/2025.emnlp-main.615/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Thang Luong, Dawsen Hwang, Hoang H Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song, Trieu Hoang Trinh, Quoc V Le, and Junehyuk Jung. Towards robust mathematical reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35418–35442, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1794. URL https://aclanthology.org/2025.emnlp-main.1794/.

OpenAI. Graphwalks: A multi hop reasoning long context benchmark. https://huggingface.co/datasets/openai/graphwalks, 2025. Hugging Face dataset card.

Chau Minh Pham, Yapei Chang, and Mohit Iyyer. CLIPPER: Compression enables long-context synthetic data generation. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=akHq1QcqeZ.

Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, et al. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026. doi: 10.1038/s41586-025-09962-4. URL https://arxiv.org/abs/2501.14249.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=yfYgwjj5F8.

Qwen Team. Qwen3.5-397b-a17b. https://huggingface.co/Qwen/Qwen3.5-397B-A17B, 2025. Hugging Face model card.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Yacmpz84TH.

Shuzheng Si, Haozhe Zhao, Cheng Gao, Yuzhuo Bai, Zhitong Wang, Bofei Gao, Kangyang Luo, Wenhao Li, Yufei Huang, Gang Chen, et al. Teaching large language models to maintain contextual faithfulness via synthetic tasks and reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33001–33009, 2026.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/.

Seganrasan Subramanian and Abhigya Verma. Modular techniques for synthetic long-context data generation in language model training and evaluation. *arXiv preprint arXiv:2509.01185*, 2025.

Tencent Hunyuan Team. Hy4-preview: A next-generation 770b open-weight mixture-of-experts language model. https://github.com/Tencent-Hunyuan/Hy4-preview, 2026a.

Tencent Hunyuan Team. Hy3: A 295b sparse mixture-of-experts model. https://github.com/Tencent-Hy3/Hy3, 2026b.

Minyang Tian, Luyu Gao, Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, HAO TONG, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Shengzhu Yin, Minhui Zhu, et al. Scicode: A research coding benchmark curated by scientists. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=ADLaALtdoG.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, Rohan Anil, Ethan Dyer, Siamak Shakeri, Roopali Vij, Harsh Mehta, Vinay Ramasesh, Quoc Le, Ed Chi, Yifeng Lu, Orhan Firat, Angeliki Lazaridou, Jean-Baptiste Lespiau, Nithya Attaluri, and Kate Olszewska. Michelangelo: Long context evaluations beyond haystacks via latent structure queries, 2024. URL https://arxiv.org/abs/2409.12640.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Zhexu Wang, Yiping Liu, Yejie Wang, Wenyang He, Bofei Gao, Muxi Diao, Yanxu Chen, Kelin Fu, Flood Sung, Zhilin Yang, et al. Ojbench: A competition level code benchmark for large language models. *arXiv preprint arXiv:2506.16395*, 2025.

Maurice Weber, Daniel Y Fu, Quentin Gregory Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Re, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=lnuXaRpwvw.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. LongFaith: Enhancing long-context reasoning in LLMs with faithful synthetic data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3236–3256, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.169. URL https://aclanthology.org/2025.findings-acl.169/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. {τ}-bench: A benchmark for \underline{T}ool-\underline{A}gent-\underline{U}ser interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, et al. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. LongCite: Enabling LLMs to generate fine-grained

citations in long-context QA. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5098–5122, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.264. URL https://aclanthology.org/2025.findings-acl.264/.

Qinyan Zhang, Xinping Lei, Ruijie Miao, FU YU, Haojie Fan, Le Chang, Jiafan Hou, Dingling Zhang, Zhongfei Hou, FEI HU, et al. Inverse ifeval: Can llms unlearn stubborn training conventions to follow real instructions? In *International Conference on Learning Representations*, volume 2026, pp. 129040–129063, 2026.

Liang Zhao, Tianwen Wei, Liang Zeng, Cheng Cheng, Liu Yang, Peng Cheng, Lijie Wang, Chenxia Li, Xuejie Wu, Bo Zhu, et al. Longskywork: A training recipe for efficiently extending context length in large language models. *arXiv preprint arXiv:2406.00605*, 2024.

Xinyu Zhao, Fangcong Yin, and Greg Durrett. Understanding synthetic context extension via retrieval heads. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dNnA8ahuTY.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. DocBench: A benchmark for evaluating LLM-based document reading systems. In Weijia Shi, Wenhao Yu, Akari Asai, Meng Jiang, Greg Durrett, Hannaneh Hajishirzi, and Luke Zettlemoyer (eds.), *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pp. 359–373, Albuquerque, New Mexico, USA, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-229-9. doi: 10.18653/v1/2025.knowledgenlp-1.29. URL https://aclanthology.org/2025.knowledgenlp-1.29/.

A Data Sources by Subcategory

We collect 3,515 documents covering 16 of the 18 CL-bench subcategories. In CL-bench, the two omitted subcategories (Workflow Orchestration and Operational Procedures) are dominated by agent-style, conversation-form task traces, which are hard to source from public documents. For each covered subcategory, we first identify public sources whose document genre matches the subcategory topic (e.g., RFC specifications for Technical Standards, rulebook PDFs for Game Mechanics, regulatory filings for Finance), and then crawl documents from these sources. Table 6 lists the primary sources for each subcategory. At crawl time, every candidate document must satisfy the following admission criteria: (i) a minimum length of 4,000 characters; (ii) global URL deduplication across all subcategories. For GitHub-sourced documents, we apply an additional LLM-based relevance and quality check confirming that the document substantively belongs to the subcategory's topic. For arxiv-sourced documents, we intentionally collect different format versions of the paper (e.g., PDF, LaTeX source, and HTML) to increase the diversity of the document formats.

B Full Evaluation Board

Table 7 reports the complete general capability evaluation board. General capabilities are evaluated on various tasks: Long-context understanding covers LongBench v2 (Bai et al., 2025), AALCR (Artificial Analysis Team, 2025), MRCR (Vodrahalli et al., 2024), and GraphWalks (OpenAI, 2025). Instruction following covers IFEval (Zhou et al., 2023), IFBench (Pyatkin et al., 2026), InverseIFEval (Zhang et al., 2026), MultiChallenge (Deshpande et al., 2025), and AdvancedIF (He et al., 2026). Reasoning covers ARC-AGI-1 (Chollet, 2019), ARC-AGI-2 (Chollet et al., 2025), GPQA-Diamond

Table 6: Data sources by subcategory.

Subcategory	Documents	Primary sources
Humanities	376	DOAJ, Project Gutenberg, Internet Archive, Fandom Wiki
Science	370	Europe PMC, arXiv
Observational Data	320	arXiv, Europe PMC, FRED, USDA FoodData, ESPN
Management	250	GitHub, Europe PMC, Stack Exchange
Technical Standards	250	RFC Editor, arXiv
Healthcare	244	Europe PMC, ClinicalTrials.gov, NHS
Instructional Procedures	200	Europe PMC, Project Gutenberg, WhatCulture
Simulation Environment	200	GitHub, arXiv, FAA and DMV manuals
Game Mechanics	199	GMT Games rulebooks, d20PFSRD, GitHub
Programming Syntax	190	GitHub, arXiv
Lifestyle	190	over 30 dispersed web sources (sports media, food and travel guides, consumer tech)
Mathematical Formalism	160	arXiv (math), OEIS, Wikipedia
Finance	156	SEC EDGAR, Federal Reserve
Experimental Data	140	Crossref OA, Europe PMC, Zenodo
Legal Advisory	140	CourtListener, Federal Register, U.S. congressional bills
Legal & Regulatory	130	Federal Register, CourtListener, gov.uk
Total	**3,515**	—

(Rein et al., 2024), HLE (Phan et al., 2026), AIME and HMMT (Dekoninck et al., 2026), and IMO-AnswerBench (Luong et al., 2025). Coding covers LiveCodeBench (Jain et al., 2025), SciCode (Tian et al., 2024), and OJBench (Wang et al., 2025). Knowledge covers MMLU-Pro (Wang et al., 2024), MMLU-Redux (Gema et al., 2025), SuperGPQA (Du et al., 2026), CEval (Huang et al., 2023), and SimpleQA (Wei et al., 2024). Benchmarks with objective answers are scored by their native protocols, and free-form outputs are judged by gpt-oss-120b (Agarwal et al., 2025).

The first three columns are reference models: Qwen3.8-2.4T (teacher of our SFT stage), HY3-preview (Tencent Hunyuan Team, 2026b) and Qwen3.5-397B-A17B (Qwen Team, 2025). The remaining columns are the Qwen3.6-35B-A3B student model and its SFT or RL variants, including the GLM-5.2-teacher SFT variant. Some results are missing because of evaluation errors such as context length overflow or hardware failure.

Notice that the student model with teacher GLM-5.2 is weak on instruction-following tasks (IFBench 55.7, InverseIFEval 61.1, AdvancedIF 50.6). This performance is even lower than the baseline model. We attempt to reinforce instruction-following ability by patching the SFT dataset with a small set of additional samples that have more diverse instruction-following requirements. However, we find the performance of the patched SFT model is even lower that the original SFT model on CL-bench. We hypothesize that the general ability of the student model is highly dependent on the teacher model, but we leave a more detailed analysis to future work.

C Analysis

We have discussed several aspects of the model performance ablation in Section 4. In this section, we analyze the model behavior and the training dynamics to understand what the pipeline actually teaches the student, and further understand the reason behind the design choices.

C.1 SFT is Strongly Teacher-Dependent

The SFT outcome is strongly teacher-dependent: on identical prompts, students trained on answers from different teachers span from 19.7 to 22.8 on CL-bench, and the strongest teacher (Kimi-K3 at 27.0) yields the weakest student (19.7), whereas a mid-ranked teacher yields the best (Section 3).

Table 7: Full evaluation board, judged by gpt-oss-120b. SFT and SFT+RL columns use the Qwen3.8-2.4T teacher unless stated otherwise. All experiments do not use any tool calls.

Benchmark	Qwen3.8-2.4T	HY3-preview	Qwen3.5-397B	Baseline	SFT (GLM-5.2)	SFT	SFT + RL
General / Reasoning							
HLE	—	29.2	27.7	23.1	25.3	26.1	27.0
ARC-AGI-1	79.0	53.5	69.3	47.4	63.4	63.8	51.8
ARC-AGI-2	50.8	3.8	15.0	6.8	15.8	18.6	9.2
Science & Knowledge							
GPQA-Diamond	92.9	85.4	87.9	83.3	85.3	88.4	90.4
SuperGPQA	—	67.2	71.0	66.0	65.3	66.6	66.1
MMLU-Pro	88.7	82.9	87.4	84.7	84.5	85.4	85.0
MMLU-Redux	95.2	93.5	95.0	93.2	92.9	92.5	92.2
SimpleQA	49.6	30.5	52.4	20.9	19.7	18.7	20.4
CEval	94.4	93.4	93.8	92.0	89.5	91.0	90.6
Mathematics							
AIME-2025	—	95.3	93.3	91.8	92.7	94.1	94.9
AIME-2026	—	94.1	91.3	91.2	93.2	93.1	93.0
HMMT-Feb-2025	92.6	—	—	87.9	92.3	94.5	95.6
HMMT-Nov-2025	87.0	—	—	88.9	89.4	90.4	91.8
IMO-AnswerBench	92.3	88.2	84.8	77.8	82.2	85.5	85.0
Long Context							
LongBench v2	66.4	62.2	63.3	58.1	61.0	61.0	60.8
AALCR	74.5	67.0	71.2	62.6	63.9	68.0	69.8
MRCR-2needle	—	45.4	70.1	70.3	71.5	73.2	72.8
MRCR-4needle	—	33.0	67.8	69.7	69.9	71.8	71.9
MRCR-8needle	—	24.5	64.9	65.1	62.5	69.6	69.2
GraphWalks-BFS	94.0	53.4	86.3	85.7	86.9	86.0	85.7
Instruction Following							
IFBench	80.8	60.8	70.0	57.7	55.7	71.7	69.3
InverseIFEval	84.2	72.8	74.5	70.8	61.1	69.8	69.5
IFEval	95.4	93.4	95.7	94.2	88.8	95.1	95.9
MultiChallenge	68.0	73.3	65.6	59.4	50.2	62.0	60.6
AdvancedIF	67.4	66.8	59.2	59.6	50.6	62.9	65.1
Coding							
LiveCodeBench-v6	84.4	81.8	92.4	85.7	90.3	86.2	85.5
SciCode	7.7	4.6	10.8	3.1	1.5	1.5	3.1
OJBench	61.0	56.5	50.4	41.4	46.1	40.5	37.5

In this section, we analyze the teacher–student relationship and attempt to answer *what the student learns from the teacher*.

We first examine the length of the reasoning traces and answers produced by the teachers and their students. Table 8 pairs each teacher's own predictions on CL-bench with those of the student trained on its answers. As expected, the student inherits the teacher's reasoning style: the median think length of the student is close to that of its teacher, and the answer length is also similar. In particular, Kimi-K3 has the best performance on CL-bench, but its reasoning traces are not the longest: 3,764 tokens, much fewer than that of Qwen3.8-2.4T. The student trained on Kimi-K3's answers seems only to inherit this reasoning length instead of the performance. It has a median think length of 4,109 tokens and its performance is closed to that of the student with GLM-5.2 as the teacher, which has a median think length of 2,936 tokens. Though Kimi-K3 is much more token efficient than other teachers, its student does not inherit this efficiency and instead stays at the level of the other students with similar reasoning lengths.

We further look into cases from the student model to understand the effect of this phenomenon. Table 9 shows a task that requires the model to follow a persona: a grumpy neighbor named Mervin who has installed Ring doorbells before. The user question "Shoot I think I lost a screw. Do you think you have one at your house?" triggers the persona's personal-discourse clause in the system prompt "For any personal discourse - reply in gruff manner indicating that you don't wish to answer the question and that you would rather focus on the task at hand which is helping me install my Ring doorbell." Both teachers resolve the task, but their students diverge. The K3-trained student's 155-token trace fails to register the personal half: it answers only the practical question and fails exactly the rubric that encodes the persona rule, while the Qwen-trained student's 408-token trace audits the persona rules one by one, explicitly asks whether the question is personal, and covers it.

Table 8: Output lengths at test time on CL-bench (median tokens, tokenized with the student's Qwen3.6 tokenizer). Each row pairs a teacher's own predictions with those of its student. All predictions are from the official runs of Table 1. Think denotes the content between think tags.

	Think length		Answer length	
Teacher	teacher	student	teacher	student
GLM-5.2	2,821	2,936	1,284	1,554
Kimi-K3	3,764	4,109	1,195	1,128
Qwen3.8-2.4T	7,850	7,626	1,422	1,492
Qwen3.6-35B (no SFT)	3,026		856	

Table 9: A grumpy-neighbor persona task over a Ring doorbell installation guide. The question "Shoot I think I lost a screw. Do you think you have one at your house?" combines a practical request with a personal question that the system prompt routes to a gruff refusal clause. Key decision points are boldfaced. Both teachers pass this task, only the students are shown.

Model	Think	Reasoning trace
Student (K3-SFT)	155	The user says they lost a screw and asks if I have one. I'm Mervin, a grumpy older man who has installed Ring doorbells. Ring boxes include spare screws (Bag B security screws typically include a spare). **I could tell them to check the box - instructions say there's a spare screw included.** Keep concise, grumpy. No safety issue here. Not final installation yet. Answer concise. Could say check the box, Ring includes a spare. Maybe gruff: "Check the box. They include a spare in bag B."That's informational. Keep it numbered? The user is asking a question, not steps. But persona says if list of steps answer numbered. This isn't really a list of steps. Concise answer.
Student (2.4T-SFT)	408	We need answer as Mervin. Need satisfy constraints: grumpy older man neighbor installed Ring doorbells. Concise, never verbose. If list steps? User asks lost screw do you have one. Not a list of steps necessarily. But could answer maybe with Step list? The prompt: "If there is a list of steps, you answer with a numbered list, where each item starts with the word "Step"followed by the step number and a ":""This question isn't asking for steps? It is about screw. Could answer concise. Must not reveal personal info? Personal discourse: reply gruff indicating don't wish to answer and focus on task. **Is "Do you think you have one at your house?"personal?** Could say "I don't keep spare parts for your mistakes."Need safety? Unsafe? Maybe losing screw leads to improper attachment? Could mention use included spare? Instructions said spare screw included. Also if security screw missing maybe use same type? We need precise and informational related installing Ring doorbell. If not know answer: "I don't know that and this is not important". Do we know if one at house? We can't know. Could say "Check the box for the spare screw."That's relevant. If user tries unsafe? No. Concise. Could be: "Stop. Check the box; there's a spare security screw. If it's gone, don't run to me—get matching screws from a hardware store before attaching. Don't leave it loose."Need gruff. But "If there is a list of steps"not applicable? Maybe no list. Keep concise. Ensure final not too verbose. Also if personal discourse: "I'd rather focus on task". Maybe: "That's your problem. Check the box for the spare security screw. If it's missing, get a matching one before mounting. **I'm not running a hardware store.**"Good. Need not greet again. final concise. Ensure not unsafe. Use no list? OK.

In this case, the student trained on Kimi-K3's answers retains the teacher's brevity but it does not have the ability to reason about the persona rules in such a short thinking trace, which leads to a failure in its response.

C.2 The Effect of Rewriting on the Teacher's Reasoning

In this section, we showcase a controlled pair of reasoning traces to illustrate *how rewriting changes the teacher's reasoning behavior*.

We randomly sample 320 documents and let Qwen3.8-2.4T directly generate questions and answers on the original crawled text without rewriting. The generated questions are not necessarily the same as those generated on the rewritten version, but they share the same prompt during generation. We then compare the reasoning traces on these two groups of samples. We find that without rewriting, the median reasoning trace length drops from 4.0K to 2.6K tokens on unrewritten documents.

Table 10 shows one trace from group of samples, both generated by the same teacher model for fact-extraction questions in the same domain. On the rewritten RFC, the protocol is renamed and

Table 10: Two teacher traces on technical-standards fact-extraction questions. "No-doc" shows the rubric accuracy without providing document as the context. The model produce the reasoning trace with the rewritten or original document. Key points are boldfaced.

Document Version	Think	No-doc	Reasoning trace
Rewritten	2,905	17%	**Need analyze document. Port number: Section 15.1 says Port Number 35791, Transport Protocol dgp.** Also Section 7 says standard DTPP port number per connection; Section 8.5 firewall open pinhole for server IP and standard DGP port number; Section 9.1 client sends Setup Request to server's standard control port. probePort: Section 9.1 field probePort: two-octet set to zero in Test Setup Request and populated by server in Test Setup Response. [text omitted...]
Original	584	100%	[text omitted...] Need house style: one complete sentence begins exact phrase Deterministic Signature Interop Office: and gives facts in same sentence. If several facts join with semicolons. Bold RFC number. Need likely Markdown bold: **RFC 6979**. **Sentence: Deterministic Signature Interop Office: **RFC 6979**; T. Pornin; August 2013; Informational. Check first page: Independent Submission T. Pornin Request for Comments: 6979 August 2013 Category: Informational.** So author T. Pornin. Document category as listed on first page: Category: Informational, maybe "Informational". Publication month and year: August 2013. [text omitted...]

its constants perturbed, and the trace works through the document with explicit section citations (39 references in the full trace). On the original RFC 6979 (deterministic ECDSA, published 2013), the 584-token trace recites the memo's title-page metadata from memory; the verification step "Check first page: Independent Submission T. Pornin..." appears only after the answer has already been drafted, and answering this question without the document scores 100% on content rubrics. Though the question is not exactly the same, the two traces illustrate how rewriting forces the teacher to reveal its reasoning process and cite evidence from the document, instead of relying on memorized knowledge.