Black-Box Knowledge Distillation for Reasoning in Small Language Models

Varun Saravanan
Computer Science Student
The University of Texas at Austin
varunvsaran@utexas.edu

Abstract

This project studies whether black-box knowledge distillation can transfer reasoning behavior from a larger teacher language model to a smaller student language model using only teacher-generated text outputs. The central question is whether the format of the teacher supervision signal matters: does a student learn more effectively from final answers only, reasoning traces only, or full chain-of-thought traces followed by final answers? To test this, I use Claude Haiku 4.5 as a black-box teacher to generate filtered reasoning traces for GSM8K and fine-tune Qwen2.5-1.5B-Instruct under three supervision conditions using QLoRA. The resulting models are evaluated on GSM8K, six BIG-Bench Hard (BBH) tasks, and an LLM-as-judge qualitative evaluation. Full chain-of-thought plus answer supervision improves GSM8K accuracy from 30.02% to 46.02%, while the teacher upper bound reaches 64.75%. The BBH results are more difficult to interpret because some tasks appear sensitive to the prompting and answer-extraction protocol, but the per-task results still show that different supervision formats produce different downstream behavior. Qualitative judge results suggest that accuracy gains do not automatically imply better displayed reasoning traces, although this conclusion is limited by the use of a single LLM judge and a 100-sample evaluation. Overall, black-box distillation improves task-specific answer accuracy, but the supervision format, answer anchor, and evaluation protocol strongly shape the observed behavior.

1 Introduction

Large language models often show stronger mathematical and multi-step reasoning ability than smaller models, but their size makes them harder to deploy, fine-tune, and run under limited compute. Knowledge distillation offers a possible solution: a smaller student model can be trained to imitate outputs from a stronger teacher model. In many realistic settings, however, the teacher is only available as a black box through an API. This means the student cannot use teacher logits, weights, hidden states, or internal representations. It can only learn from the teacher's text.

This project investigates black-box knowledge distillation for reasoning in small language models. Rather than asking only whether distillation improves performance, the project asks what kind of teacher output is most useful. In particular, I compare three supervision formats: answer-only supervision, reasoning-only supervision, and full chain-of-thought supervision with a final answer. This setup tests whether reasoning traces themselves help the student learn better reasoning behavior, whether final answer formatting is necessary, and whether the benefits of arithmetic reasoning distillation transfer beyond the training domain.

The main research question is:

> **Does the format of black-box teacher supervision affect reasoning performance in a small language model?**

A secondary research question is:

> **Do reasoning gains from GSM8K distillation generalize to harder symbolic and multi-step tasks in BBH, or are they mostly task-specific?**

I hypothesized that the full chain-of-thought plus answer condition would perform best on GSM8K because it gives the student both a reasoning scaffold and a clear answer target. I also expected transfer to BBH to be limited because GSM8K arithmetic reasoning does not cover all forms of symbolic, temporal, or logical reasoning.

2 Task and Dataset

2.1 Task Definition

The task is open-ended reasoning generation followed by exact-answer extraction. Given a natural language problem, the model is prompted to solve it step by step and produce a final answer in a standardized format. The evaluation then extracts the predicted answer and compares it to the ground truth.

2.2 GSM8K Distillation Dataset

The training data comes from the GSM8K training split, which contains grade-school math word problems. I sampled 3,000 examples from the approximately 7,400-example training split for teacher querying. This sample size was chosen to keep teacher querying and student fine-tuning feasible under the course compute constraints, but it also means the experiment does not test whether the same trends would hold with the full training set. Each problem was sent to Claude Haiku 4.5 using the same chain-of-thought prompt:

```
Solve the following problem step
by step.  Show your reasoning
clearly, then state your final
answer on a new line as:  Answer:
value.
```

I kept a teacher-generated response only if the extracted teacher answer matched the ground truth and the reasoning trace contained at least 30 words. This filtering step removed incorrect teacher outputs and degenerate responses with little or no reasoning. However, it also creates selection bias because the student trains only on examples the teacher already solved correctly. As a result, the distilled data may be easier or cleaner than the full GSM8K distribution.

Each retained record contains the original question, the teacher reasoning trace, the teacher extracted answer, and the ground-truth answer. The three experimental conditions all use this same underlying filtered dataset; only the formatting of the student training target changes.

2.3 Evaluation Datasets

The main evaluation benchmark is the GSM8K test split, which contains 1,319 examples. This tests whether the student improves on the same type of arithmetic reasoning used for distillation.

To test cross-task generalization, I also evaluate on six BBH tasks:

- `multistep_arithmetic_two`
- `logical_deduction_five_objects`
- `date_understanding`
- `tracking_shuffled_objects_three_objects`
- `word_sorting`
- `causal_judgement`

These tasks cover arithmetic, symbolic manipulation, logical deduction, temporal reasoning, ordering, and commonsense judgment. I report the six-task macro-average, but I also emphasize per-task results because the teacher and students behave very differently across tasks. In particular, the teacher's near-zero score on date understanding suggests that the BBH protocol may be poorly matched to at least some tasks.

3 Models and Methods

3.1 Teacher Model

The teacher model is Claude Haiku 4.5, specifically `claude-haiku-4-5`, accessed through the Anthropic API. It is treated as a pure black box. I do not use teacher weights, logits, gradients, attention maps, hidden states, or any other internal information. The teacher only provides text outputs in response to prompts.

The teacher has two roles. First, it generates the distilled GSM8K training data. Second, it serves as an upper-bound reference during evaluation. Under the same evaluation protocol, Claude Haiku 4.5 achieves 64.75% accuracy on GSM8K and 47.90% macro-average accuracy on BBH.

3.2 Student Model

The student model is Qwen2.5-1.5B-Instruct, a 1.5-billion-parameter instruction-tuned language model. This model was chosen because it is small enough to fine-tune under constrained compute while still capable of instruction-following and reasoning-style generation.

Fine-tuning uses QLoRA. The base model is loaded in 4-bit NF4 quantization with bitsandbytes, and a LoRA adapter is trained on top. The LoRA configuration uses rank 16, alpha 32, and dropout 0.05.

The shared training setup across all three fine-tuned conditions is shown in Table 1.

Hyperparameter	Value
Training method	TRL SFTTrainer
Epochs	3
Per-device batch size	2
Gradient accumulation	8
Effective batch size	16
Learning rate	0.0002
Scheduler	Cosine
Warmup ratio	0.05
Max sequence length	512
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
Random seed	42

Table 1: Shared training setup for all fine-tuned student conditions.

Using the same base model, data source, and hyperparameters across all conditions makes the comparison focused on the effect of supervision format rather than training differences. However, the single-seed setup means the results should be interpreted as one controlled comparison rather than as a fully variance-estimated study.

3.3 Experimental Conditions

I compare one teacher upper bound, one untuned baseline, and three fine-tuned student conditions.

Teacher: Claude Haiku 4.5. The teacher model is `claude-haiku-4-5`. It is evaluated directly on the same benchmarks and serves as the upper-bound reference for both answer accuracy and qualitative reasoning scores.

Baseline: Untuned Qwen2.5-1.5B-Instruct. The baseline is the original instruction-tuned student model with no additional fine-tuning. It establishes the student's starting reasoning ability.

Condition A: Answer Only. The student is trained only on the teacher's final answer. Each example contains the question followed by a final `Answer: value` line. The teacher's reasoning trace is removed. This condition tests whether answer imitation alone changes downstream performance.

Condition B: Reasoning Only. The student is trained on the teacher's reasoning trace without the final answer line. This condition tests whether exposure to reasoning structure alone helps the student, even without explicit answer supervision.

Condition C: Full Chain-of-Thought + Answer. The student is trained on the complete teacher output: the full reasoning trace followed by the final answer in the standardized `Answer: value` format. This is the most direct form of black-box reasoning distillation.

4 Evaluation Protocol

4.1 GSM8K Evaluation

For GSM8K, all models are prompted with the same chain-of-thought prompt used during teacher data collection. Models generate up to 256 new tokens using greedy decoding. The predicted answer is extracted by first searching for the `Answer:` tag. If no tag is found, the evaluation script falls back to the last number appearing in the model output.

Accuracy is exact-match after normalizing for commas and case. This protocol is applied consistently across the teacher, baseline, and all fine-tuned conditions.

4.2 BBH Evaluation

For BBH, I evaluate the models on six selected tasks and compute macro-average accuracy across tasks. Because the teacher performs near zero on date understanding and poorly on tracking shuffled objects, I treat the six-task macro-average as descriptive rather than definitive. I therefore report both the full six-task macro-average and a four-task macro-average that excludes date understanding and tracking shuffled objects as a robustness check.

The same prompt template and answer-extraction procedure are used across models. I do not run prompt sensitivity or prompt ablation experiments, so the results may depend on the exact prompt and extraction format.

4.3 Qualitative Judge Evaluation

A secondary qualitative evaluation uses an LLM-as-judge pipeline. Claude Haiku 4.5 judges a random sample of 100 GSM8K outputs per condition on three dimensions:

- **reasoning_present:** whether the output shows step-by-step reasoning, rated as 0 or 1
- **reasoning_quality:** logical coherence and correctness, rated from 1 to 5

Model / Condition	GSM8K Accuracy
Teacher: Claude Haiku 4.5	**64.75%**
Baseline	30.02%
Condition A: Answer Only	8.87%
Condition B: Reasoning Only	27.07%
Condition C: Full CoT + Answer	46.02%

Table 2: GSM8K exact-match accuracy for the teacher upper bound, baseline, and three distillation conditions.

- **faithfulness:** whether the output follows a structured reasoning style similar to the teacher, rated from 1 to 5

This judge evaluation is designed to capture qualitative differences that exact-match accuracy may miss. However, the same model family is used as both teacher and judge, so the judge may prefer outputs that resemble its own style. The evaluation also uses only 100 samples per condition and does not include confidence intervals, human ratings, a second judge, or inter-rater reliability. I therefore use the judge results as suggestive qualitative evidence rather than as a definitive measurement of reasoning quality.

5 Results

5.1 GSM8K Accuracy

Condition C is the best student model on GSM8K, improving over the baseline by approximately 16 percentage points. This supports the hypothesis that full chain-of-thought plus answer distillation can improve arithmetic answer accuracy in a small language model. However, Claude Haiku 4.5 reaches 64.75%, leaving an 18.73-point gap between the teacher and the best student. This shows that substantial teacher ability is not recovered through black-box distillation alone.

Condition B degrades relative to the baseline, falling from 30.02% to 27.07%. Although the drop is not enormous, it is still a decrease after training on reasoning traces, so it should not be described as an improvement. This result suggests that reasoning text alone is not sufficient to teach the model how to complete the task.

Condition A performs much worse than expected on GSM8K, dropping from 30.02% to 8.87%. Because this result is anomalous and has not been replicated across seeds, I do not use it as strong evidence for a broad theory of answer-only training. Its BBH behavior is still reported because it reveals task-specific effects, but the GSM8K result should be treated cautiously.

Model / Condition	6-Task Avg.	4-Task Avg.
Teacher: Claude Haiku 4.5	**47.90%**	**69.85%**
Baseline	8.32%	12.08%
Condition A: Answer Only	19.98%	14.46%
Condition B: Reasoning Only	8.20%	11.50%
Condition C: Full CoT + Answer	18.59%	23.19%

Table 3: BBH macro-average accuracy. The four-task average excludes date understanding and tracking shuffled objects because the teacher performs poorly on those tasks under this protocol.

5.2 BBH Macro-Average Accuracy

The BBH results show a more complicated pattern than GSM8K. On the full six-task average, Condition A has the highest student macro-average. However, this average is driven by large task-specific gains on date understanding, tracking shuffled objects, and causal judgement rather than broad gains across all tasks. When the two most protocol-sensitive tasks are excluded, Condition C becomes the strongest student model. Therefore, the BBH results should not be summarized with the six-task macro-average alone.

The teacher reaches 47.90% on the six-task macro-average and 69.85% on the four-task macro-average. The size of these gaps shows that the student recovers only part of the teacher's broader reasoning ability.

5.3 BBH Per-Task Accuracy

Condition C performs best among student models on multistep arithmetic, which is the task most similar to GSM8K. This supports the idea that GSM8K chain-of-thought distillation transfers most strongly to similar arithmetic reasoning tasks.

Condition A performs surprisingly well on date understanding, tracking shuffled objects, and causal judgement, even though it performs badly on GSM8K and on multistep arithmetic. It is the best student condition on three of the six BBH tasks. However, two of those tasks are also tasks where the teacher itself performs poorly under the same evaluation protocol. This makes Condition A's BBH behavior interesting, but not strong evidence that answer-only training improves general reasoning. Its high six-task macro-average is driven by sharp task-specific gains rather than consistent transfer.

The teacher per-task results clarify several student failures. Claude Haiku 4.5 scores 100.00% on multistep arithmetic and 94.00% on word sorting, but only 0.40% on date understanding and 7.60% on tracking shuffled

Condition	Multistep Arith.	Logical Deduct.	Date Underst.	Tracking Objects	Word Sorting	Causal Judge.
Teacher	**100.00**	**80.00**	0.40	7.60	**94.00**	5.41
Baseline	20.80	0.00	0.40	1.20	4.00	23.53
Condition A	2.00	11.60	**32.80**	**29.20**	0.40	**43.85**
Condition B	38.80	1.60	0.00	3.20	2.40	3.21
Condition C	46.00	2.00	3.60	15.20	6.80	37.97

Table 4: Per-task accuracy on six BBH tasks. Values are percentages.

objects. This means the weak student results on date understanding are not simply a distillation failure; the teacher itself nearly fails the task under the same prompting and answer-extraction protocol. This suggests that the evaluation format may be poorly matched to that task.

Condition B performs well on multistep arithmetic but poorly almost everywhere else. This suggests that reasoning-only supervision may encourage mathematical elaboration but does not reliably teach the model to produce final task answers.

5.4 Combined Accuracy and Judge Results

Table 5 combines the main accuracy results with the LLM-as-judge qualitative evaluation. The teacher serves as both the performance and qualitative upper bound, reaching 64.75% GSM8K accuracy, 47.90% BBH macro-average accuracy, 4.92 reasoning quality, and 4.84 faithfulness. Among students, Condition C gives the strongest GSM8K accuracy, while the untuned baseline receives the strongest qualitative reasoning scores. Because the judge is also the teacher model and no variance estimates are available, these qualitative scores should be read as suggestive rather than definitive.

6 Analysis and Discussion

6.1 Full CoT Distillation Improves In-Domain Arithmetic Accuracy

The clearest accuracy result is that Condition C substantially improves GSM8K performance. The model receives both the teacher's reasoning process and a clear final answer target, and this combination appears to help the student learn a more effective generation strategy for arithmetic word problems.

This result supports the project hypothesis for GSM8K answer accuracy. Full chain-of-thought distillation does more than teach surface answer patterns: it gives the model an output structure for working through arithmetic problems and then committing to a standardized answer. However, because the experiment uses one training run and one data sample, the exact magnitude of the improvement should be treated as an estimate rather than a statistically established effect size.

6.2 Accuracy Gains Do Not Necessarily Imply Better Displayed Reasoning

The judge evaluation suggests a dissociation between answer accuracy and perceived reasoning quality. Condition C achieves the best student GSM8K accuracy at 46.02%, but its reasoning quality score is 3.87 and its faithfulness score is 3.62. The untuned baseline, despite lower GSM8K accuracy at 30.02%, receives higher judge scores: 4.12 for reasoning quality and 3.97 for faithfulness.

This finding should be interpreted cautiously. The judge evaluation uses 100 examples per condition, no confidence intervals, and Claude Haiku 4.5 as both teacher and judge. Still, the pattern suggests that fine-tuning can improve final-answer accuracy without clearly improving the displayed reasoning trace. One possible explanation is that the base Qwen2.5-1.5B-Instruct model already has strong instruction-following behavior, allowing the base model to produce coherent-looking reasoning. Supervised fine-tuning on a narrower teacher-generated dataset may improve answer accuracy while making the reasoning style more formulaic or less coherent.

6.3 Answer-Only Supervision Produces Narrow but Uneven Effects

Condition A shows that the supervision format directly affects the model's output behavior. It has reasoning_present of 0.00, reasoning quality of 1.00, and faithfulness of 0.95. In other words, answer-only training teaches the model to produce bare answers without step-by-step working.

Condition	GSM8K Acc.	BBH Avg.	Reasoning Present	Reasoning Quality	Faithfulness
Teacher	**64.75%**	**47.90%**	1.00	**4.92**	**4.84**
Baseline	30.02%	8.32%	1.00	4.12	3.97
Condition A	8.87%	19.98%	0.00	1.00	0.95
Condition B	27.07%	8.20%	1.00	3.80	3.52
Condition C	46.02%	18.59%	1.00	3.87	3.62

Table 5: Combined accuracy and qualitative judge results. Judge scores are averages over 100 randomly sampled GSM8K outputs per condition. No confidence intervals are available, so small score differences should be interpreted cautiously.

At the same time, Condition A's BBH behavior shows that answer-only training did not simply make the model worse at every task. It achieves the best student scores on date understanding, tracking shuffled objects, and causal judgement. These gains are unusual because those tasks are not the ones most improved by full CoT distillation or reasoning-only distillation. A plausible interpretation is that some BBH tasks reward short categorical answer behavior more than extended reasoning traces. Because Condition A is trained to provide direct final answers, it may be better aligned with the answer format of these tasks even though it loses the ability to show useful reasoning.

This interpretation is limited by two important caveats. First, Condition A's GSM8K result is anomalous and unreplicated. Second, two of the tasks driving its BBH macro-average are also tasks where the teacher performs poorly, suggesting that the evaluation protocol may be unstable for those tasks. For that reason, Condition A's BBH macro-average should not anchor a broad theoretical claim about answer-only distillation.

6.4 Reasoning Without Answers Degrades GSM8K Accuracy

Condition B is informative because it isolates reasoning traces from final-answer supervision. On GSM8K, it drops from the 30.02% baseline to 27.07%. On BBH, its macro-average accuracy is almost identical to baseline.

This suggests that reasoning traces alone may not teach the model how to finish the task. The student may learn to produce reasoning-like text without learning to commit to the correct final answer. This is especially plausible because the evaluation depends on answer extraction, and Condition B never sees the explicit `Answer:` target during training.

The answer-anchor explanation is plausible, but it is not directly proven by this experiment. A cleaner test would add an additional condition such as reasoning plus a dummy answer marker or reasoning plus a differently formatted final answer. In this paper, I therefore treat the answer-anchor idea as a hypothesis that explains the observed gap between Condition B and Condition C, not as an established causal mechanism.

6.5 Cross-Task Transfer Is Uneven and Protocol-Sensitive

The BBH results show that GSM8K distillation does not produce uniform improvement across reasoning tasks. Condition C improves the six-task BBH macro-average from 8.32% to 18.59%, but the teacher remains far ahead at 47.90%. Most student gains come from specific tasks rather than broad improvement across all reasoning categories.

At the same time, the BBH protocol is clearly imperfect. Claude Haiku 4.5 performs extremely well on multistep arithmetic and word sorting, but nearly fails date understanding and tracking shuffled objects under the same protocol. Because of this, I avoid treating the six-task BBH macro-average as a definitive measure of general reasoning. The four-task average excluding these two protocol-sensitive tasks gives a different student ranking, with Condition C ahead of Condition A.

Overall, the BBH results suggest limited and uneven cross-task transfer rather than complete general reasoning transfer. GSM8K examples appear most useful for related arithmetic reasoning, and less reliable for tasks requiring different forms of symbolic, temporal, or categorical reasoning.

6.6 Macro-Average Accuracy Can Hide Different Failure Modes

Condition A and Condition C have similar six-task BBH macro-average scores, but their per-task profiles are very different. Condition C is strongest among students on multistep arithmetic, while Condition A is strongest on

date understanding, tracking shuffled objects, and causal judgement.

This means the macro-average alone is not enough to interpret the experiment. The same average score can come from different abilities and different failures. In this case, Condition C's BBH score is more consistent with arithmetic transfer from GSM8K, while Condition A's score is driven by unexpected gains on tasks where direct answer behavior appears to help. The per-task breakdown is therefore essential for interpreting the supervision-format comparison.

7 Code Architecture

The project code is organized as a Python pipeline with separate scripts for data collection, training, evaluation, and judging.

The full pipeline is:

1. `collect.py`: queries Claude Haiku 4.5 on GSM8K and writes teacher-distilled JSONL records.
2. `train.py`: fine-tunes the Qwen student model using QLoRA.
3. `evaluate.py`: evaluates the teacher, baseline, and fine-tuned models on GSM8K and BBH.
4. `judge.py`: runs the LLM-as-judge qualitative evaluation on sampled GSM8K outputs.

Training uses the `transformers`, `peft`, `trl`, `bitsandbytes`, and `accelerate` libraries. Teacher data collection and judge evaluation use the Anthropic Python SDK. Configurations are YAML-based, with one config file per training condition. Each config specifies the model ID, LoRA hyperparameters, data path, output directory, and training hyperparameters.

All random seeds are fixed to 42. The training setup is designed to fit within the course compute constraints by using a 1.5B model, 4-bit quantization, LoRA adapters, maximum sequence length 512, and a 3,000-example teacher-query subset.

8 Limitations and Future Work

This project has several limitations. First, the teacher-generated dataset is filtered for correctness, which improves supervision quality but may bias the dataset toward examples Claude Haiku 4.5 already solves cleanly. Second, the experiment uses 3,000 GSM8K examples as the only distillation training source, so the student may learn arithmetic-specific reasoning rather than general reasoning. Third, each condition is trained once with seed 42, so there is no variance estimate across seeds or training runs. Fourth, the Condition A GSM8K result is anomalous and should be replicated before drawing strong conclusions about answer-only training. Fifth, the BBH results are sensitive to prompting and answer extraction, as shown by the teacher's near-zero score on date understanding. Sixth, the LLM-as-judge evaluation uses the teacher model as the judge, only 100 samples per condition, and no inter-rater reliability.

There are several useful extensions. Rerunning all conditions across multiple random seeds would provide variance estimates and help determine whether the observed differences are robust. Training with more GSM8K examples would test whether the experiment is data-limited rather than format-limited. A prompt ablation study would clarify whether BBH results are sensitive to prompt wording. Running a 3B or 7B untuned baseline would show whether Condition C's gains are more useful than simply using a larger off-the-shelf model. Finally, distilling from a more diverse reasoning dataset could test whether broader teacher supervision improves transfer across BBH task categories.

9 Conclusion

This project shows that black-box knowledge distillation can improve answer accuracy in a small language model when the supervision format provides both reasoning traces and final answers. Full chain-of-thought plus answer distillation raises GSM8K accuracy from 30.02% to 46.02%, a substantial improvement over the untuned baseline. However, Claude Haiku 4.5 still reaches 64.75%, leaving a large gap between teacher ability and what the student recovers through text-only distillation.

The qualitative results complicate the picture. The untuned baseline receives higher reasoning quality and faithfulness scores than any fine-tuned student condition, even though Condition C is much more accurate. Because this result comes from a limited LLM-as-judge setup, it should be treated cautiously, but it suggests that improving final-answer accuracy does not automatically improve the quality of the displayed reasoning trace.

The transfer results are also mixed. On BBH, fine-tuned models improve on some tasks but do not show uniform reasoning transfer, and some BBH tasks appear poorly matched to the evaluation protocol. Condition A's

high six-task BBH macro-average is especially important to interpret carefully: it comes from strong gains on date understanding, tracking shuffled objects, and causal judgement, not from general improvement across all reasoning tasks. Overall, the results suggest that black-box distillation can teach useful task-specific answer behavior, but those behaviors are shaped heavily by the teacher output format, the answer anchor, the training domain, and the evaluation protocol.

Acknowledgments

I would like to thank the CS371N course staff for a great semester. The concepts, assignments, and discussions throughout the course helped shape how I think about NLP research, evaluation, and model behavior, and I hope to apply this knowledge in future projects and work.

References

[1] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

[2] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[4] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*.