

TO CoT OR NOT TO CoT? CHAIN-OF-THOUGHT HELPS MAINLY ON MATH AND SYMBOLIC REASONING

Anonymous authors

Paper under double-blind review

ABSTRACT

Chain-of-thought (CoT) via prompting is the de facto method for eliciting reasoning capabilities from large language models (LLMs). But for what kinds of tasks is this extra “thinking” really helpful? To analyze this, we conducted a quantitative meta-analysis covering over 100 papers using CoT and ran our own evaluations of 20 datasets across 14 models. Our results show that CoT gives strong performance benefits primarily on tasks involving math or logic, with much smaller gains on other types of tasks. On MMLU, directly generating the answer without CoT leads to almost identical accuracy as CoT *unless* the question or model’s response contains an equals sign, indicating symbolic operations and reasoning. Following this finding, we analyze the behavior of CoT on these problems by separating planning and execution and comparing against tool-augmented LLMs. Much of CoT’s gain comes from improving symbolic execution, but it underperforms relative to using a symbolic solver. Our results indicate that CoT can be applied selectively, maintaining performance while saving inference costs. Furthermore, they suggest a need to move beyond prompt-based CoT to new paradigms that better leverage intermediate computation across the whole range of LLM applications .

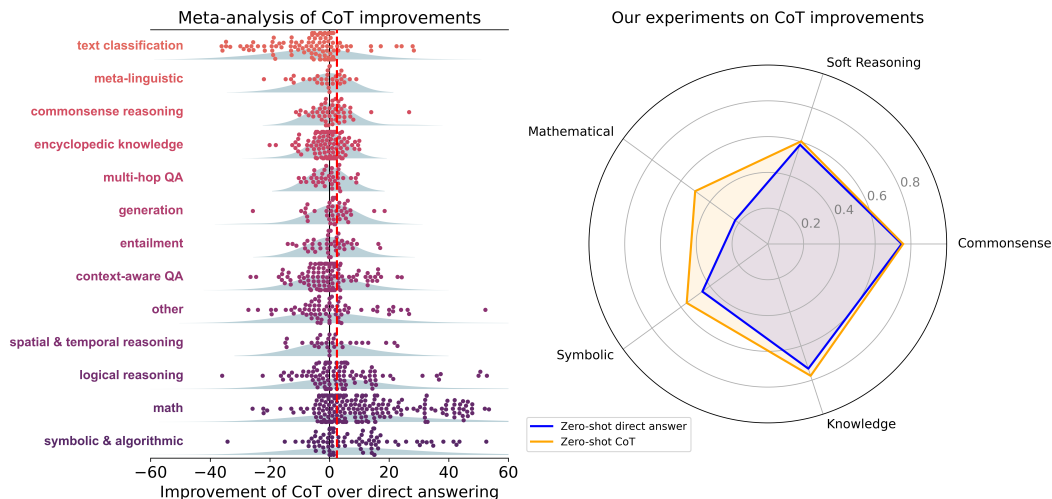


Figure 1: Left: meta-analysis of CoT literature; each point is a reported delta of CoT over direct answering for some (LLM, task) pair. Right: average performance of using zero-shot CoT v.s. direct answer prompts across five general reasoning categories, covering 20 datasets with 14 LLMs evaluated on each. In both sets of results, math and other kinds of symbolic reasoning are the domains that consistently see substantial improvements from CoT (red dotted line indicates the mean improvement from CoT across experiments).

1 INTRODUCTION

Chain-of-thought (CoT) (Nye et al., 2022; Wei et al., 2022) has become a widely used prompting technique for eliciting reasoning from language models. CoT can provide human-readable explanations of how problems are solved (Joshi et al., 2023; Lanham et al., 2023), but most frequently

it is invoked to improve an LLM’s ability to answer complex questions via intermediate computation (Madaan & Yazdanbakhsh, 2022; Wang et al., 2023a; Dziri et al., 2023). Current post-training schemes for LLMs heavily infuse CoT capabilities into models: systems like ChatGPT or Llama 3.1 default to CoT when given reasoning problems (OpenAI, 2023; Dubey et al., 2024).

CoT has seen widespread usage, but it is most heavily explored in the domain of mathematical reasoning (Zhou et al., 2023a; Fu et al., 2023; Chae et al., 2024; Xu et al., 2024b; Qi et al., 2024). In fact, many “reasoning” methods for LLMs are evaluated *only* in the math domain; for instance, Lightman et al. (2024) frame their paper as “complex multi-step reasoning” and Mixtral-Large2’s release ¹ cited effort “enhancing the model’s reasoning capabilities”, but performance is only reported on GSM8K and MATH. CoT is reported to be effective across a wide range of studies, but many of these studies focus on a narrow slice of the task space. In areas beyond math, results show that CoT is not as useful (Kambhampati et al., 2024a) or can even hurt performance (Wang et al., 2024).

In this work, we aim to evaluate where prompt-based CoT helps and why. We begin with a systematic meta-analysis of recent literature that reports performance of CoT versus direct answering (DA). We then augment this picture by conducting experiments on 20 datasets and 14 contemporary LLMs across zero-shot and few-shot prompt settings. **Finding 1: CoT only helps substantially on problems requiring mathematical, logical, or algorithmic reasoning.** Figure 1 shows this holds both across the literature and our own experiments. We find only a few cases of large gain in other kinds of tasks, and many of these outliers feature some component of symbolic reasoning. For instance, on MMLU (Hendrycks et al., 2021a) and MMLU Pro (Wang et al., 2024), we analyze the improvements from CoT and find that CoT *only* gives benefit on math slices of the dataset. **As much as 95% of the total performance gain from CoT on MMLU is attributed to questions containing “=” in the question or generated output.** For non-math questions, we find no features to indicate when CoT will help.

How can we better understand *why* CoT improves on these questions and only these questions? The math and formal logical reasoning datasets we consider can be broken down into two stages of processing: a planning step (e.g., parsing a problem into equations) and an execution step (building intermediate outputs and working towards a solution) (Ye et al., 2023; Wang et al., 2023b; Sun et al., 2024). **Finding 2: CoT primarily helps with the execution step that performs computation and symbolic manipulation, but falls short of what LLMs with tool augmentation can do.** We find that LMs prompted with CoT can generate executable formal solution plans and execute those plans better than direct answering. But using LMs to generate a solution plan and then using an external symbolic solver to solve the plan outperforms using CoT for both steps for these tasks.

These results paint a picture that CoT’s utility is often circumscribed by tool augmentation: on problems where CoT helps, we already have more powerful tools than CoT that we can employ, and on “soft reasoning” problems like commonsense where no tools exist, we see limited benefit from CoT. This characterization has two major implications. First, CoT is unnecessary for many problems where it is widely employed: there exist more efficient prompting strategies that yield similar performance for much lower inference cost. Second, we see a critical need to move beyond prompt-based CoT to more sophisticated approaches based on search, interacting agents, or models more heavily fine-tuned for CoT. Future work can explore how intermediate computation can be better used to solve challenging problems outside of the math and symbolic reasoning domains.

2 BACKGROUND: CHAIN-OF-THOUGHT

The tasks we consider in this work consist of a question $\mathbf{q} \in \Sigma^*$ for a vocabulary Σ and an answer $a \in \mathcal{L}(\mathbf{q})$ for a label set $\mathcal{L}(\mathbf{q})$. $\mathcal{L}(\mathbf{q})$ can consist of a data type like boolean or integer, classification labels, or problem-dependent labels like names of entities from \mathbf{q} . One exception that we still explore is BiGGen Bench (Kim et al., 2024), which instead relies on an LLM-as-a-judge (Dubois et al., 2023; Zheng et al., 2024b) to provide a label for generated long-form responses.

Prompting and chain-of-thought for reasoning A large language model places distributions over strings $p(\mathbf{y}) = \prod_{i=1}^n p_{\text{LM}}(y_i)$ where $\mathbf{y} \in \Sigma^*$. In practice, we can interpret these as conditional

¹<https://mistral.ai/news/mistral-large-2407/>

distributions $p(\mathbf{y} \mid \mathbf{x})$ where \mathbf{x} is a user’s prompt. Typical invocation of an LLM involves forming a prompt $\mathcal{I}(\mathbf{q})$ that wraps the question with additional instruction, then drawing a sample response $\tilde{\mathbf{y}} \sim p(\mathbf{y} \mid \mathcal{I}(\mathbf{q}))$, and finally returning $a = \text{extract}(\tilde{\mathbf{y}})$ using some kind of answer extractor.

For the tasks we consider in this work, the output $\tilde{\mathbf{y}}$ can take one of two forms. A **direct answer** only contains a string realization of a ; e.g., $\mathbf{y} = (_185, 4)$ which is detokenized as the answer $a = 1854$. A **chain of thought** is a longer sequence \mathbf{y} including other tokens beyond the answer, e.g., $\mathbf{y} = (_185, 6, _minus, _2, _equals, _185, 4)$. In both cases, the extract function must parse and detokenize the output; in CoT, there is some extra work to spot where the answer is placed.

Our prompts can explicitly encourage use of direct answer or chain of thought as strategies, which we denote as \mathcal{I}_{da} and \mathcal{I}_{cot} . For eliciting CoT, this includes strategies like telling a model to “*think step by step*” (Kojima et al., 2022). For directly answering a question, a prompt may say “*immediately generate the answer*”. We track the average location of the answer in the generated output for both CoT and direct prompts in Appendix F.3 to ensure that direct answer prompts give the answer early in the output. We also ensure that extract can parse answers from the generated output for each model, prompt, and dataset combination used in our experiments, tailoring the extract function as needed to ensure low unparseable rates for each model and task.² All prompts and outputs per dataset per model have been uploaded to Huggingface and we include examples of some of our prompts in the Appendix J. We also experiment with few-shot CoT prompts, which we find perform similarly to zero-shot prompts; details about these are given in Appendix E.

Symbolic reasoning Of key importance to this work is whether problems feature symbolic reasoning or not. We consider a problem to be **symbolic** if it can be grounded in a *natural, well agreed-upon* formal system. “ 12×4 ” is an example of a symbolic problem, which can be grounded in mathematics. Other systems include first-order logic (Saparov & He, 2023; Hua et al., 2024) or planning languages (Liu et al., 2023a; Valmeekam et al., 2023). Formally, for symbolic problems, we define a function f that acts as a map that produces some symbolic expression $\mathcal{S} = f(\mathbf{q})$ from the question. \mathcal{S} can be used as input for a solver to derive an answer, $\hat{a} = \text{solve}(\mathcal{S})$.

Conversely, a problem like *where on a river can you hold a cup upright to catch water on a sunny day?* from CommonsenseQA (Talmor et al., 2019) is **non-symbolic** by our definition. While this problem could be formalized with some kind of predicate logic (Zhou et al., 2022; Quan et al., 2024; Zhou et al., 2024) or grounded in some kind of physical simulation (Hao et al., 2023; Wong et al., 2023), there is not a natural nor well agreed-upon framework for solving it.

We view non-symbolic to symbolic reasoning as a spectrum. MuSR (Sprague et al., 2024) is a “semisymbolic” dataset in that it does contain an underlying formal system (e.g., for its murder mysteries portion, the notion that $\text{motive}(X) \wedge \text{means}(X) \wedge \text{opportunity}(X) \implies \text{murderer}(X)$), but also involves substantial commonsense reasoning that does not map onto a formal system. In these cases, we can still form $\mathcal{S} = f(\mathbf{q})$, but f must rely heavily on a language model and instantiate new information for \mathcal{S} that is not directly represented in \mathbf{q} .

Central claim Figure 1 shows that there are a large number of positive results on CoT reported in the literature. Informally, we believe many readers of the literature to hold the following view: \mathcal{I}_{cot} **will outperform \mathcal{I}_{da} on nearly all reasoning problems, whether those problems involve symbolic or non-symbolic reasoning.** Our evidence does *not* support this conjecture. We will show that this performance boost is strongest for symbolic and semi-symbolic tasks, while giving little to no improvement (or even hurting performance) on non-symbolic tasks.

3 RESULTS FROM THE LITERATURE

Criteria and Process We investigate all papers from ICLR 2024, a representative ML venue, and two representative NLP venues, EACL 2024 and NAACL 2024 (including Findings and Workshop papers). This resulted in 4,642 papers total that filtered using automatic and manual methods to papers including experiments comparing chain-of-thought, \mathcal{I}_{cot} , against direct answering prompts,

²We exclude a number of other “CoT-like” approaches in our analysis such as decomposed prompting (Khot et al., 2023; Zheng et al., 2024a) and multi-agent debate (Du et al., 2023; Chen et al., 2024). We focus on single prompt approaches. We deal with tool-augmented approaches in Section 5.

Table 1: A few categories for experimental comparisons. Full list in Appendix B.

Category	Description
Symbolic and algorithmic	Tasks involving symbol manipulation which can be solved by executing a program. This includes entity tracking datasets (e.g., SCONE, Coin Flip) and algorithmic tasks (e.g., BBH word sorting or finding shortest paths in a graph).
Math	Tasks requiring mathematical reasoning, from grade-school math to advanced mathematics, including physics questions.
Logical reasoning	Tasks designed to test for logical reasoning, whether deductive (Saparov & He, 2023, PrOntoQA), inductive (Bowen et al., 2024) or analogical (Ma et al., 2024) reasoning, including syllogisms and logical puzzles.
Encyclopedic knowledge	Tasks requiring expert-level in-depth knowledge beyond mere common-sense, usually in an open-book setting.
Mixed datasets	Datasets containing a variety of tasks, such as BIG-Bench Hard (BBH) or MMLU.
...	...

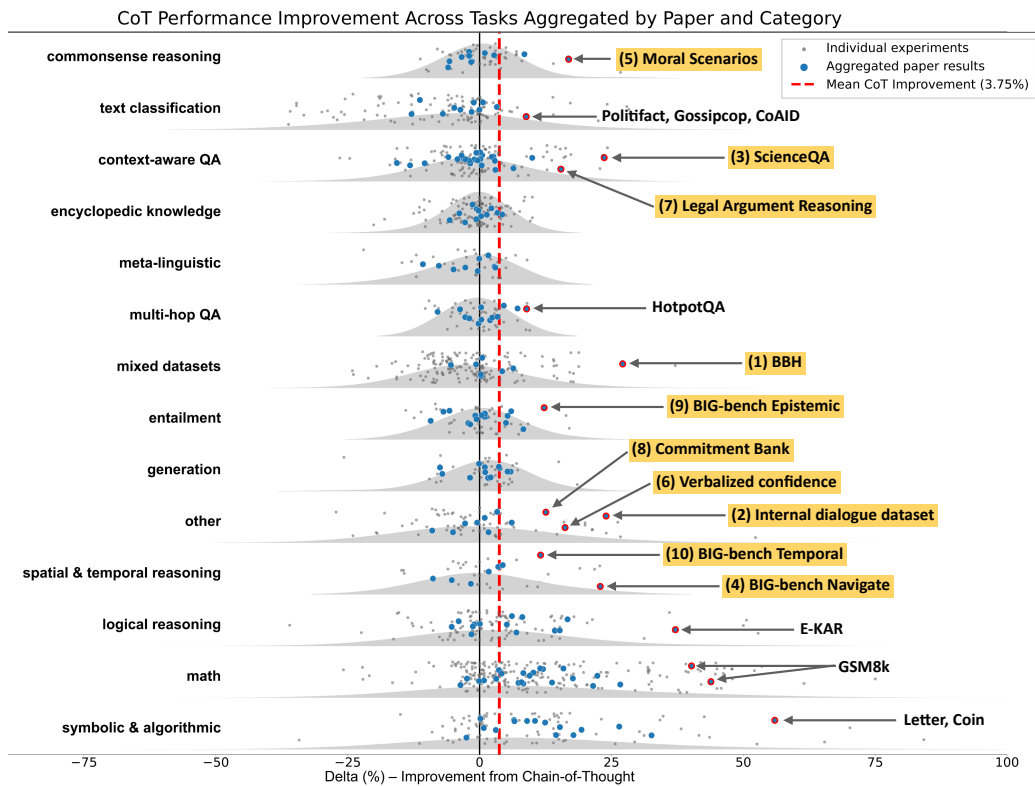


Figure 2: Results from our meta-analysis (grey dots) aggregated by paper and category (blue dots).

$\mathcal{I}_{\text{direct}}$. A total of 110 papers were found that matched our criteria with 1,218 experimental comparisons. We then grouped the comparisons by the types of tasks and datasets being evaluated. More details on our automatic and manual filtering, as well as our categorization, can be found in Appendix A and B.

Results Figure 2 shows the distribution of CoT deltas (CoT prompt minus the direct answer prompt performance) across our categorization of different task types found in the literature. Compared to Figure 1, we take the mean results per paper per category, indicated by blue dots, showing the trend across papers in the literature. The categories are ranked in order of ascending median

CoT delta. The three categories which benefited the most from CoT are symbolic reasoning, math, and logical reasoning, with average improvements of 14.2, 12.3, 6.9, respectively. Average performance on these top three tasks with CoT was 56.9, whereas performance without CoT was 45.5. For other categories, the average performance with CoT was 56.8, compared to 56.1 without CoT. We do not consider this small improvement a victory for CoT. CoT involves more computation than direct answering, and a truly fair comparison between the methods should match the compute of the two methods, e.g., ensembling across multiple prompts.

Do any non-math datasets benefit from CoT? On the right side of Figure 2, we show the top 10 outliers from our observed trend, namely papers with high CoT deltas averaged across experiments in tasks *other than* math, symbolic, or logical reasoning. Although not categorized as math or logic, several of these are related to logical, mathematical or symbolic reasoning in some way. From this list, the dataset which benefits the most most from CoT is BIG-bench Hard (BBH) (Suzgun et al., 2023), a benchmark consisting largely of problems requiring algorithmic, arithmetic or logical reasoning. For instance, BIG-bench Navigate is a spatial reasoning task, but relies heavily on a mathematical primitive of counting steps taken to derive a final conclusion. Similarly, while BIG-bench Temporal is a temporal reasoning task (answering questions about when certain events could have occurred), it requires deductive reasoning to solve. In addition, Legal Argument Reasoning (SemEval-2024 Task 5) (Bongard et al., 2022) was categorized as *context-aware QA*, but also requires substantial reasoning ability. Finally, MMLU-Moral Scenarios (Hendrycks et al., 2021a) requires answering two independent questions at once, which essentially involves a symbolic combination of two simpler questions.

There are a few outliers that less clearly follow the trend. ScienceQA (Lu et al., 2022) consists of multiple choice questions across a range of natural and social science disciplines, though it is hard to interpret gains without knowing breaking down performance by subject or question type. The dialogue evaluation dataset from Jia et al. (2024) sees large improvements with CoT, but this is a proprietary dataset, and we note that other essay scoring results in our meta-analysis (Li et al., 2024; Stahl et al., 2024) did not show improvements with CoT. Other non-math, symbolic or logical datasets that benefit from CoT are Commitment Bank (de Marneffe et al., 2019) and the task of eliciting verbalized confidence (Xiong et al., 2024). Nevertheless, these are exceptions to the rule. The majority of the reported benefits from using CoT in the NLP and ML literature comes from math or math-related tasks.

4 RESULTS FROM EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Dataset, Models, Prompts All datasets, models, and prompts we evaluate over can be found in detail in the tables 3, 4, and 5 of Appendix C. We restricted our experiments to English models commonly used and benchmarked on general reasoning datasets. Our datasets include those which are widely used in CoT and reasoning literature, including a mix of non-symbolic, semisymbolic, and symbolic reasoning. They span different formats, including multiple-choice, short-answer, and free-response; however, most of these datasets are multiple choice or short answer, as CoT is not typically used in long-form response settings. We also categorize each dataset into a larger category of reasoning required to solve it: Commonsense, Knowledge, Symbolic, Mathematical, and Soft Reasoning. We define Soft Reasoning as questions relying on commonsense and natural language but going beyond simple inferences about these statements. Finally, we explore several prompting strategies for eliciting reasoning from language models, as past work has emphasized the importance of the prompt (Yang et al., 2024). However, we generally found slight performance differences; see Appendix D for details. We therefore focus on prompts similar to Kojima et al. (2022) and Wei et al. (2022) for zero-shot and few-shot settings, respectively, with alterations to improve the model’s ability to produce desired behavior (i.e., formats that allow for easily parsed answers). We will upload all our prompts and outputs for each model for each prompting strategy on Huggingface.

Implementation Details We use a high-throughput inference package, vLLM (Kwon et al., 2023), for the model inference process. We use greedy decoding on all models. Our prompts are taken from the Llama 3.1 evaluations when available (Dubey et al., 2024), and minor adjustments are made to

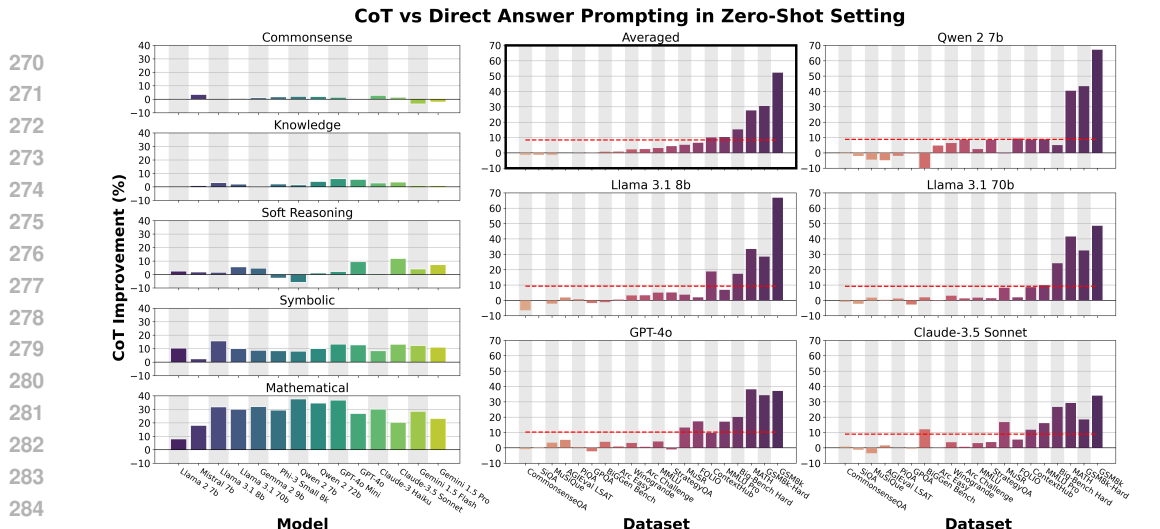


Figure 3: Left: Performance gain from using CoT for each reasoning category. Right: Performance gain from using CoT for each dataset, averaged across models and broken out across 5 representative models. Red lines indicate median improvement. In both plots we see a consistent trend: most improvements from using CoT are from math and symbolic reasoning.

unify prompting strategies. For other datasets, we either use the standard prompt for the dataset from the corresponding original paper or implement our own prompt. Our answer parser (extract) is tailored to each dataset and model. Specific details about each dataset, its prompts, and answer extractor can be found in Appendix C.

4.2 RESULTS

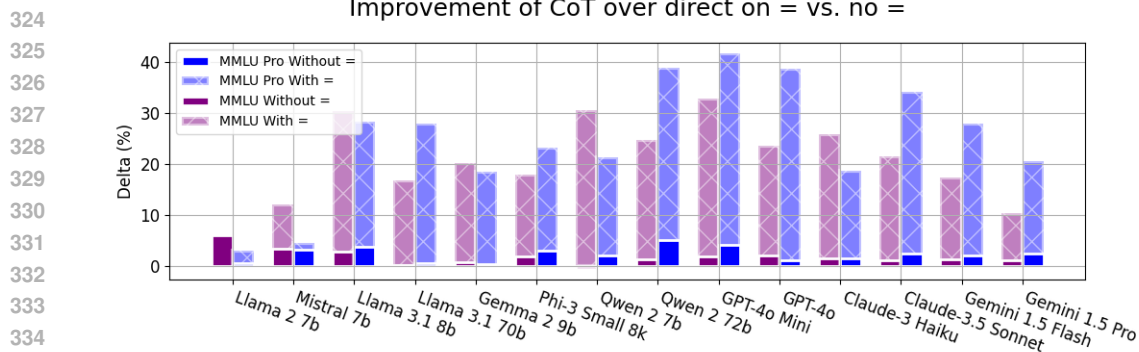
Where does zero-shot CoT improve over direct prompts? *On datasets that require math (MATH, GSM8K) or formal logic (ContextHub, MuSR to a lesser degree) to answer the problem.*

Figure 3 on the left shows the average CoT performance improvement for each reasoning category from Figure 1 (right); raw numbers can be found in Table 6 of the Appendix. On the right, Figure 3 shows the performance gain from using CoT for each dataset, averaged across all models and for a selection of individual models. On non-symbolic reasoning categories and datasets, specifically those that contain questions primarily involving commonsense (CSQA, PIQA, SiQA), language understanding (WinoGrande), and reading comprehension (AGI LSAT, ARC-Easy, ARC-Challenge), there is little to no separation between the performance of zero-shot CoT and zero-shot direct answer. Despite these datasets involving reasoning, CoT does not yield improvement.

By contrast, the mathematical and symbolic categories get larger boosts in improvements alongside symbolic and many semi-symbolic datasets. MATH and GSM8K show gains as large as 41.6% and 66.9%, respectively. The semi-symbolic datasets like ContextHub and MuSR Murder Mysteries show moderate gains. These datasets require the application of logical rules to reach the answer, e.g., first-order logic parsed from simple natural language (ContextHub) or more complex commonsense statements (MuSR Murder Mysteries). All results are shown in the Appendix F.1 as well as a full list of numeric results for both CoT and direct answer prompting in Table 7. We also explored the few-shot setting and found it had little impact on when CoT will help; see Appendix E.

Does the answer format impact where CoT will help? *Not much. Free response capabilities required for BigGen Bench may not benefit from pre-planning.*

Many of the commonly-used datasets for problems other than math are multiple choice. We highlight here that CoT has similar performance to direct answer across models for two datasets that are not multiple-choice and contain varying levels of non-symbolic reasoning. First, MuSiQue (Trivedi et al., 2022) is a short-form QA task requiring multi-hop reasoning. We consider this a semi-symbolic dataset as the questions have an explicit multi-hop structure. Because answer spans in MuSiQue can be paraphrased in many different ways, we use GPT-4o to judge if two answer spans are equivalent. Despite being semi-symbolic, we see no overall improvement from CoT.



336 Figure 4: CoT deltas between MMLU and MMLU Pro performance when a question or generated
337 response contains an “=” (With =) or not (Without =). We filter out any questions that do not result
338 in a final answer (degeneration, etc.). CoT primarily helps on the pairs of questions and generations
339 that contain an “=”, which indicates math-related questions.

340
341
342
343
344
345
346
347
348
349
350
351
352

Second, BiGGen Bench (Kim et al., 2024) uses free-form responses as the answer to a question, and an LLM-as-a-judge is used to evaluate these responses on a scale of 1 to 5. Because free-form responses blur the lines between CoT and direct answering, we create a new prompt that asks the language model to plan the free response before giving it. We then only pass the free response to the judge (GPT-4o-mini in our case) with the prompt from Kim et al. (2024). We also filter out any questions that explicitly state “Think step-by-step”. We plot the performance of BiGGen Bench as the number of times a response receives a score of 4 or better. **Despite including many reasoning questions (including several categories of math) and other categories, such as planning, we only see a mild improvement here. Because previous experiments show CoT helping on similar types of questions in the QA format, the lack of similar improvements here could imply that pre-planning is insufficient for unlocking reasoning capabilities in the LLM. Future work is needed to prove this.**

353
354
355
356
357
358
359
360
361
362
363
364
365
366

Are the gains in Knowledge, Soft Reasoning, and Commonsense significant? *Mostly no, except for MMLU, StrategyQA, and MuSR.*

We tested the significance of the improvements from CoT on the 13 datasets in the Knowledge, Soft Reasoning, and Commonsense reasoning categories using paired bootstrapping to assess whether CoT gives a significant improvement. To account for multiple comparisons, we applied a Bonferroni correction, setting the p-value to 0.00027 to account for the 14 models and 13 datasets. About 32% (59) of the datasets that show a benefit in these three reasoning categories were considered significant. Nearly half of these comparisons (26) are on MMLU and MMLU Pro. On these datasets, we find that CoT is mainly helping on math-related questions. StrategyQA and MuSR also received a consistent performance boost across 10 and 6 models respectively. StrategyQA is often used to benchmark reasoning methods and is built specifically to get a benefit from methods that decompose the question into steps, so a gain in performance is not unprecedented. MuSR, similarly, was built to have multiple steps of complex natural language reasoning, which may receive benefits from CoT. The remaining datasets that receive significant benefits are spread across the datasets and models.

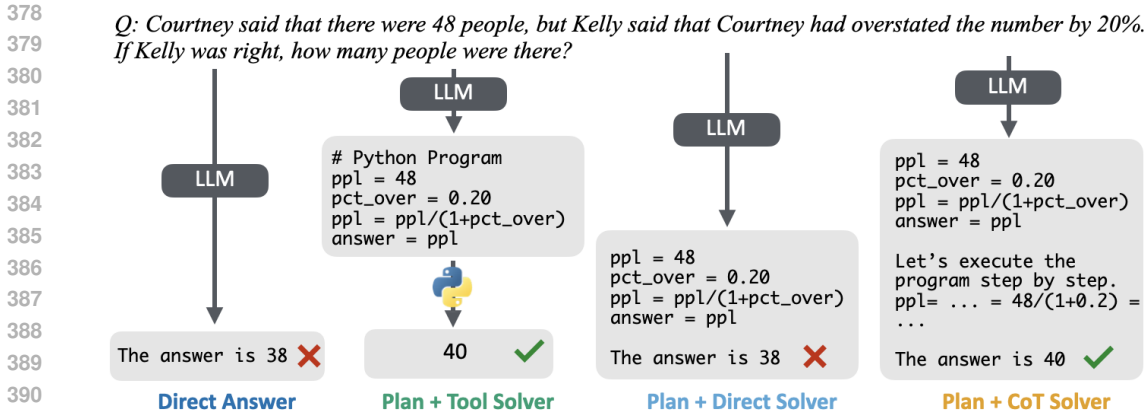
367
368
369
370
371
372
373

Why do MMLU and MMLU Pro get a boost? MMLU and MMLU Pro contain many different questions requiring different types of reasoning. We separated MMLU and MMLU Pro questions into two bins, those related to math and those not related to math, by checking if the questions text or generated response from the LLM includes an “=”. Figure 4 shows that a majority of the performance gain seen from MMLU and MMLU Pro is from the math slices of each dataset. See more details in Appendix G.

374
375
376
377

5 STRENGTHS AND WEAKNESSES OF CoT AT FORMAL REASONING

Previous sections establish that CoT primarily helps with symbolic reasoning tasks, but not why. Many symbolic and semi-symbolic tasks be broken down into two stages (Ye et al., 2023; Pan et al.,



392 Figure 5: Prompt variants that separate planning and execution for GSM8K. For all prompt variants
 393 besides direct answer and CoT (not shown), we few-shot prompt an LLM to first generate a Python
 394 program as a solution plan. For Plan + Direct Solver, the LLM is prompted to directly give an
 395 answer from the plan; for Plan + CoT Solver, the LLM is prompted to solve the plan step-by-step
 396 with CoT and give an answer; for Plan + Tool Solver, we feed the plan into a Python interpreter.

397
398
399 2023; Jiang et al., 2024): planning, either via a formal or informal specification via prompting (Sun
 400 et al., 2024; Wang et al., 2023b), and execution, using the same LM or external solvers. In this
 401 section, we attribute the performance gains from CoT on symbolic tasks to these two stages.

402 Given a question that requires symbolic reasoning, we define the **planning** stage as extracting all
 403 variables from the context into a formal specification and defining their relations. The **execution**
 404 stage uses a solver that takes as input a plan and can be run in an orderly fashion to derive the final
 405 answer. Using our notation from Section 2, let $f(\mathbf{q}) = \mathcal{I}_{\text{planning}}^m(\mathbf{q})$ be a mapping of the question
 406 \mathbf{q} to a symbolic plan $\mathcal{S}_{\text{plan}}$ that can be executed by the language model or by an external symbolic
 407 solver, $\hat{a} = \text{solve}(\mathcal{S}_{\text{plan}})$, where \hat{a} is the final answer for \mathbf{q} .

408 By separating planning and execution in this way, we can test how much a language model can gain
 409 from only having a plan, to having a plan and solving it with CoT, or to having a plan and then
 410 solving it with an external symbolic solver. Given a plan $\mathcal{S}_{\text{plan}} \sim \mathcal{I}_{\text{planning}}^m(\mathbf{q})$, we compare the
 411 performance of the settings below to evaluate at which stage LM is most effective and falls short.

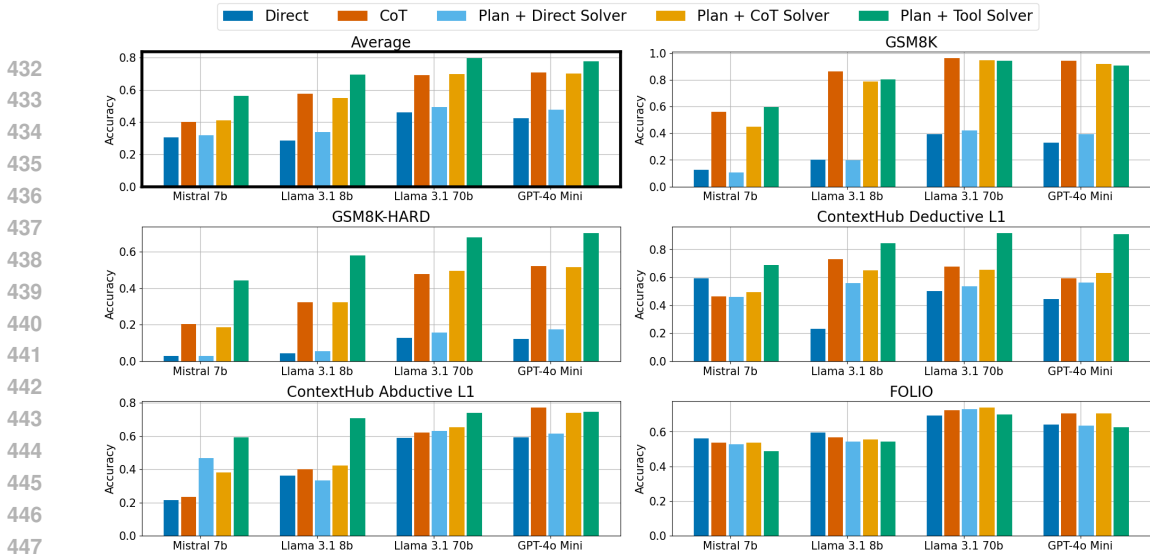
412
413
414 **5.1 SETTINGS EVALUATED**

415 **Settings 1 and 2: Few-shot direct answer and CoT:** We use the few-shot direct answer and CoT
 416 prompts from Section 4.1 as baselines. Figure 5 includes an example of each setting on GSM8K.

417 **Settings 3 and 4: Plan + Direct Solver and Plan + CoT Solver:** Here we use inspiration from Xu
 418 et al. (2024a) and generate a symbolic plan using the same strategy as Ye et al. (2023). Specifically,
 419 we use a few-shot prompt $\mathcal{I}_{\text{planning}}^m$ to generate a formal specification $\mathcal{S}_{\text{plan}}$ that should be executable
 420 by a symbolic solver. In the same prompt LMs are asked to solve their generated specification $\mathcal{S}_{\text{plan}}$
 421 and derive the final answer $\hat{y} \sim p(\mathbf{y} \mid \mathcal{I}_{\text{da}}(\mathcal{S}_{\text{plan}}))$, either directly giving the answer after generating
 422 the specification (**Plan + Direct Solver**) or providing step-by-step explanations and tracking of
 423 intermediate steps for the derivation (**Plan + CoT Solver**). Particularly, $\mathcal{S}_{\text{plan}}$ is a Python program
 424 for math datasets, and is a set of first-order logic specifications for logical reasoning datasets.

425 **Setting 5: Plan + Tool Solver:** We then evaluate how effective CoT can be at performing symbolic
 426 computations compared with external symbolic solvers. Following prior work on augmenting LMs
 427 with tools for math and logic questions (Ye et al., 2023; Pan et al., 2023; Gao et al., 2023; Chen
 428 et al., 2023), we generate $\mathcal{S}_{\text{plan}}$ the same way as in CoT Solver, but now feed in the plan into a
 429 symbolic solver (Python interpreter or a SMT Solver), such that $\hat{a} = \text{solve}(\mathcal{S}_{\text{plan}})$.

430 **Evaluation Setup:** We compare the performance of each setting on math (GSM8K) and logical
 431 reasoning (ContextHub and FOLIO) datasets. We follow Gao et al. (2023) to include GSM8K-Hard,



448 Figure 6: Performance of prompt variants that separate planning and execution for math and logical
 449 reasoning datasets. Despite outperforming direct answer for solving a formal plan and deriving the
 450 final answer, CoT is still limited in performing symbolic computations: there is a large performance
 451 boost from Plan + Tool Solver over CoT and Plan + CoT Solver on average across all models.
 452
 453
 454

455 a minimally modified version that replaces numbers of GSM8K with larger numbers, to account for
 456 the possibility of recent LLMs overfitting GSM8K by data contamination (Zhang et al., 2024).
 457

458 For Plan + Direct solver and Plan + CoT solver, we use the few-shot prompts from Ye et al. (2023).
 459 For Plan + Tool solver, we use state-of-the-art tool-augmented prompting methods. Particularly, for
 460 GSM8K, we use Program-aided Language Model (Gao et al., 2023, PAL) that executes the LM-
 461 generated plan with a Python interpreter. For logical reasoning datasets, we use Satisfiability-Aided
 462 Language Model (Ye et al., 2023, SatLM) that uses automated theorem prover Z3 (De Moura &
 463 Bjørner, 2008) to solve the generated specifications. If the generated plan cannot be parsed by the
 464 tool, we use random guessing when the question is multiple choice, and mark it incorrect otherwise.
 465
 466

467 5.2 EVALUATION RESULTS

468
 469 Figure 6 shows the results across a representative selection of models. Detailed numerical results,
 470 including the unparseable rates of model-generated plans, can be found in Appendix H.
 471

472 When comparing direct answer with Plan + Direct solver and Plan + CoT solver, we note that for
 473 many datasets and models, only having a plan does not account for most of the performance gain.
 474 **Compared with direct answer, CoT or Plan + CoT solver is needed for strong performance.**
 475 **Tracking the execution with one of these methods gives the strongest accuracy benefit, espe-**
 476 **cially for math-heavy datasets.**

477 Despite their strength over direct answer and Plan + Direct solver, **CoT and Plan + CoT solver are**
 478 **dominated by Plan + Tool solver in most settings.** LLMs are limited by their ability to execute
 479 and track steps compared with symbolic solvers.

480 We argue that these results provide an explanation of why CoT helps on symbolic tasks. While all
 481 tasks could feasibly benefit from a detailed description of how to solve each individual question (e.g.,
 482 a *plan* in the context of this section), CoT only outperforms direct answer when these steps require
 483 a substantial amount of tracing and computation. In these settings, we can see clear performance
 484 benefit from using symbolic solvers; **CoT appears to be a poor (but universal) approximation to**
 485 **such solvers.** When possible, LLMs should be paired with symbolic solvers at inference time when
 solving symbolic tasks to achieve consistently better performance over direct answer and CoT.

6 DISCUSSION AND RELATED WORK

Where is CoT helping and why? Our results showing CoT improvement for math and logic aligns well with early work on CoT for LLMs such as Scratchpads (Nye et al., 2022). As CoT gained popularity, its application has broadened to tasks that canonically do not require multiple steps. It can often yield small improvements over direct answering. We believe this led to the current prevailing sentiment that deliberation should improve performance on any task requiring some type of reasoning (our original claim from Section 2). However, our results show a clear separation between performance on non-symbolic and symbolic tasks. If, in theory, any question could benefit from deliberation, why is CoT only benefiting the questions that can be solved through symbolic manipulation? Our results from Section 5 suggest that the primary benefit of CoT comes in the ability to execute symbolic steps and track their output. Not all tasks have this feature: for example, questions from CommonsenseQA can hardly be translated into formally grounded and executable solution plans. Datasets like StrategyQA may feature multiple steps of reasoning, but executing those steps is not complex, so the benefits of CoT are small. It is unclear whether explicitly instilling models with particular modes of deliberation, like process of elimination for multiple choice questions, might make them more effective for non-symbolic tasks, or whether there’s a fundamental limitation imposed by their pre-training data. We leave this distinction for future work.

Can we improve CoT further? Our work treats chain-of-thought variants that explicitly don’t involve multiple inferences. There is evidence that using additional calls to LLMs can help (Du et al., 2023; Yao et al., 2023; Besta et al., 2023; Chen et al., 2024), but these methods use significantly increased computation, and careful benchmarking sometimes reveals that naive techniques are as good as iterative ones (Olausson et al., 2024). However, past theoretical results show that Transformers are augmented in a fundamental way by CoT (Liu et al., 2023b; Merrill & Sabharwal, 2024); we believe this indicates the potential for improving CoT beyond prompt-based CoT. On the other hand, recent methods showing benefit from “internalizing” CoT (Deng et al., 2024) may indicate that explicit generation of intermediate tokens is not used to its full potential.

Limitations One set of tasks we do not cover in our experiments (except for BiGGen Bench) is long-horizon planning. However, many works in the literature have already discussed the efficacy of planning with CoT. We also do not address the data contamination of some of these models on the datasets. We try to mitigate this by including multiple models, datasets (new and old), and our meta-analysis. For more discussion of planning and dataset contamination, see Appendix I.

7 CONCLUSION

In this work, we characterize the performance of prompt-based CoT through a meta-analysis of the literature and experiments across different models, datasets, and prompts. We find that CoT predominantly helps on math and formal logic, largely due to its ability to trace the intermediate steps of a problem. But CoT rarely outperforms tool-augmented approaches for these same problems. We believe that CoT remains a powerful technique, but to give improvement across a wider range of NLP tasks, research should move beyond prompt-based CoT to new paradigms like search, interacting agents, or better fine-tuned models.

REPRODUCIBILITY

For our experiments, we provide in-depth details of how we evaluated models on each dataset in Section 4.1 and Appendix C. Furthermore, we release all prompts for every dataset on Huggingface, including per model output and sampling parameters. For our meta-analysis of the literature, we describe our filtering criteria and process of annotating experiments into high-level categories in Section 3 and Appendix B. We also release the full list of papers in our meta-analysis together with extracted experimental comparisons and task category annotations.

REFERENCES

- 540
541
542 Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Has-
543 san Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Ben-
544 haim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C’esar Teodoro Mendes,
545 Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew
546 Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng
547 Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karam-
548 patziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi
549 Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi,
550 Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant,
551 Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil
552 Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji
553 Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Ji-
554 ahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng-Yuan Zhang, Cyril Zhang,
555 Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report:
556 A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219, 2024. URL
557 <https://api.semanticscholar.org/CorpusID:269293048>.
- 558 Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. a. URL [https://www-cdn.](https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf)
559 [anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.](https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf)
560 pdf.
- 561 Anthropic. Claude 3.5 Sonnet Model Card Addendum. b. URL [https://www-cdn.anthropic.](https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf)
562 [com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf](https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf).
- 563 Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Giani-
564 nazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoe-
565 fler. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI Con-*
566 *ference on Artificial Intelligence*, 2023. URL [https://api.semanticscholar.org/CorpusID:](https://api.semanticscholar.org/CorpusID:261030303)
567 261030303.
- 568 Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about
569 physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2019.
570 URL <https://api.semanticscholar.org/CorpusID:208290939>.
- 571 Leonard Bongard, Lena Held, and Ivan Habernal. The legal argument reasoning task in civil proce-
572 dure. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preotiuc-
573 Pietro (eds.), *Proceedings of the Natural Legal Language Processing Workshop 2022*, pp. 194–
574 207, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computa-
575 tional Linguistics. doi: 10.18653/v1/2022.nllp-1.17. URL [https://aclanthology.org/2022.](https://aclanthology.org/2022.nllp-1.17)
576 [nllp-1.17](https://aclanthology.org/2022.nllp-1.17).
- 577
578 Chen Bowen, Rune Sætre, and Yusuke Miyao. A comprehensive evaluation of inductive reasoning
579 capabilities and problem solving in large language models. In Yvette Graham and Matthew Purver
580 (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 323–339,
581 St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL [https://](https://aclanthology.org/2024.findings-eacl.22)
582 aclanthology.org/2024.findings-eacl.22.
- 583 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhari-
584 wal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agar-
585 wal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh,
586 Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Ma-
587 teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCand-
588 lish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot
589 learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Ad-*
590 *vances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Asso-
591 ciates, Inc., 2020. URL [https://proceedings.neurips.cc/paper_files/paper/2020/file/](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf)
592 [1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf).
- 593 Hyungjoo Chae, Yeonghyeon Kim, Seungone Kim, Kai Tzu-iunn Ong, Beong-woo Kwak,
Moohyeon Kim, Seonghwan Kim, Taeyoon Kwon, Jiwan Chung, Youngjae Yu, et al. Language

- 594 Models as Compilers: Simulating Pseudocode Execution Improves Algorithmic Reasoning in
595 Language Models. *arXiv preprint arXiv:2404.02575*, 2024.
- 596
- 597 Chih Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves
598 reasoning via consensus among diverse LLMs, 2024. URL [https://openreview.net/forum?](https://openreview.net/forum?id=Yo16nUVIJD)
599 [id=Yo16nUVIJD](https://openreview.net/forum?id=Yo16nUVIJD).
- 600
- 601 Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting:
602 Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Ma-*
603 *chine Learning Research*, 2023. ISSN 2835-8856. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=YfZ4ZPt8zd)
604 [YfZ4ZPt8zd](https://openreview.net/forum?id=YfZ4ZPt8zd).
- 605
- 606 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and
607 Oyvind Tafjord. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning
608 Challenge. *arXiv:1803.05457v1*, 2018.
- 609
- 610 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
611 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
612 Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL
613 <https://api.semanticscholar.org/CorpusID:239998651>.
- 614
- 615 Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: In-
616 vestigating projection in naturally occurring discourse. In *Proceedings of Sinn und Bedeutung 23*,
617 2019.
- 618
- 619 Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient SMT solver. In *Proceedings of the*
620 *Theory and Practice of Software, 14th International Conference on Tools and Algorithms for the*
621 *Construction and Analysis of Systems, TACAS’08/ETAPS’08*, pp. 337–340, Berlin, Heidelberg,
622 2008. Springer-Verlag. ISBN 3540787992.
- 623
- 624 Yuntian Deng, Yejin Choi, and Stuart Shieber. From Explicit CoT to Implicit CoT: Learning to
625 Internalize CoT Step by Step. *arXiv preprint arXiv:2405.14838*, 2024.
- 626
- 627 Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improv-
628 ing factuality and reasoning in language models through multiagent debate. *arXiv preprint*
629 *arXiv:2305.14325*, 2023.
- 630
- 631 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
632 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 Herd of
633 Models. *arXiv preprint arXiv:2407.21783*, 2024.
- 634
- 635 Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos
636 Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A Simulation Framework
637 for Methods that Learn from Human Feedback, 2023.
- 638
- 639 Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean
640 Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal,
641 Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of trans-
642 formers on compositionality. In *Thirty-seventh Conference on Neural Information Processing*
643 *Systems*, 2023. URL <https://openreview.net/forum?id=Fkckkr3ya8>.
- 644
- 645 Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting
646 for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*,
647 2023. URL <https://openreview.net/forum?id=yf1icZHC-19>.
- 648
- 649 Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and
650 Graham Neubig. Pal: program-aided language models. In *Proceedings of the 40th International*
651 *Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- 652
- 653 Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle
654 use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions*
655 *of the Association for Computational Linguistics*, 9:346–361, February 2021. ISSN 2307-387X.
656 doi: 10.1162/tacl.a.00370.

- 648 L. Guan, Yifan Zhou, Denis Liu, Yantian Zha, Heni Ben Amor, and Subbarao Kambhampati. "Task
649 Success" is not Enough: Investigating the Use of Video-Language Models as Behavior Critics
650 for Catching Undesirable Agent Behaviors. *ArXiv*, abs/2402.04210, 2024. URL [https://api.
651 semanticscholar.org/CorpusID:267500077](https://api.semanticscholar.org/CorpusID:267500077).
- 652 Atharva Gundawar, Mudit Verma, L. Guan, Karthik Valmeekam, Siddhant Bhambri, and Subbarao
653 Kambhampati. Robust Planning with LLM-Modulo Framework: Case Study in Travel Plan-
654 ning. *ArXiv*, abs/2405.20625, 2024. URL [https://api.semanticscholar.org/CorpusID:
655 270199944](https://api.semanticscholar.org/CorpusID:270199944).
- 656 Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy
657 Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian
658 Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty,
659 Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev.
660 FOLIO: Natural Language Reasoning with First-Order Logic. *arXiv preprint arXiv:2209.00840*,
661 2022. URL <https://arxiv.org/abs/2209.00840>.
- 662 Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reason-
663 ing with language model is planning with world model. In Houda Bouamor, Juan Pino, and
664 Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Lan-
665 guage Processing*, pp. 8154–8173, Singapore, December 2023. Association for Computational
666 Linguistics. doi: 10.18653/v1/2023.emnlp-main.507. URL [https://aclanthology.org/2023.
667 emnlp-main.507](https://aclanthology.org/2023.emnlp-main.507).
- 668 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob
669 Steinhardt. Measuring massive multitask language understanding. *Proceedings of the Interna-
670 tional Conference on Learning Representations (ICLR)*, 2021a.
- 671 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn
672 Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset.
673 *NeurIPS*, 2021b.
- 674 Hanxu Hu, Hongyuan Lu, Huajian Zhang, Wai Lam, and Yue Zhang. Chain-of-symbol prompting
675 elicits planning in large language models, 2023.
- 676 Wenyue Hua, Kaijie Zhu, Lingyao Li, Lizhou Fan, Shuhang Lin, Mingyu Jin, Haochen Xue, Zelong
677 Li, Jindong Wang, and Yongfeng Zhang. Disentangling Logic: The Role of Context in Large
678 Language Model Reasoning Capabilities. *ArXiv*, abs/2406.02787, 2024. URL [https://api.
679 semanticscholar.org/CorpusID:270258104](https://api.semanticscholar.org/CorpusID:270258104).
- 680 Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as
681 zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv preprint
682 arXiv:2201.07207*, 2022.
- 683 Jinghan Jia, Abi Komma, Timothy Leffel, Xujun Peng, Ajay Nagesh, Tamer Soliman, Aram Gal-
684 styan, and Anoop Kumar. Leveraging LLMs for dialogue quality measurement. In Yi Yang,
685 Aida Davani, Avi Sil, and Anoop Kumar (eds.), *Proceedings of the 2024 Conference of the
686 North American Chapter of the Association for Computational Linguistics: Human Language
687 Technologies (Volume 6: Industry Track)*, pp. 359–367, Mexico City, Mexico, June 2024.
688 Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-industry.30. URL
689 <https://aclanthology.org/2024.naacl-industry.30>.
- 690 Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh
691 Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lu-
692 cile Saulnier, L’elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,
693 Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*,
694 abs/2310.06825, 2023. URL <https://api.semanticscholar.org/CorpusID:263830494>.
- 695 Dongwei Jiang, Marcio Fonseca, and Shay B. Cohen. Leanreasoner: Boosting complex logical
696 reasoning with lean. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Pro-
697 ceedings of the 2024 Conference of the North American Chapter of the Association for Compu-
698 tational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024,
699 700 701

- 702 *Mexico City, Mexico, June 16-21, 2024*, pp. 7497–7510. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.416. URL <https://doi.org/10.18653/v1/2024.naacl-long.416>.
- 703
704
705
706 Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang,
707 Yejin Choi, and Xiang Ren. Are Machine Rationales (Not) Useful to Humans? Measuring and
708 Improving Human Utility of Free-text Rationales. *ArXiv*, abs/2305.07095, 2023. URL <https://api.semanticscholar.org/CorpusID:258676376>.
- 709
710 Subbarao Kambhampati. Can large language models reason and plan? *Annals of the New*
711 *York Academy of Sciences*, 1534:15 – 18, 2024. URL <https://api.semanticscholar.org/CorpusID:268249961>.
- 712
713 Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant
714 Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: LLMs can’t plan, but can help planning
715 in LLM-modulo frameworks. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian
716 Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st*
717 *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22895–22907. PMLR, 21–27 Jul 2024a. URL <https://proceedings.mlr.press/v235/kambhampati24a.html>.
- 718
719
720 Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant
721 Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: LLMs can’t plan, but can help plan-
722 ning in LLM-modulo frameworks. In *Forty-first International Conference on Machine Learning*,
723 2024b. URL <https://openreview.net/forum?id=Th8JPEmH4z>.
- 724
725 Liwei Kang, Zirui Zhao, David Hsu, and Wee Sun Lee. On the empirical complexity of reasoning
726 and planning in llms. *arXiv preprint arXiv:2404.11041*, 2024.
- 727
728 Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and
729 one pairs: A “novel” challenge for long-context language models. *ArXiv*, abs/2406.16264, 2024.
730 URL <https://api.semanticscholar.org/CorpusID:270703648>.
- 731
732 Tushar Khot, H. Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish
733 Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The*
734 *International Conference on Learning Representations*, volume abs/2210.02406, 2023. URL
<https://api.semanticscholar.org/CorpusID:252715485>.
- 735
736 Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeun Kim, Dongkeun Yoon, Guijin
737 Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. The BiGGen Bench: A Principled Bench-
738 mark for Fine-grained Evaluation of Language Models with Language Models. *arXiv preprint*
arXiv:2406.05761, 2024.
- 739
740 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
741 language models are zero-shot reasoners. In *Proceedings of the 36th International Conference*
742 *on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc.
ISBN 9781713871088.
- 743
744 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E.
745 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
746 serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating*
747 *Systems Principles*, 2023.
- 748
749 Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional
750 skills of sequence-to-sequence recurrent networks. In Jennifer G. Dy and Andreas Krause (eds.),
751 *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stock-*
752 *holmssäsan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine*
753 *Learning Research*, pp. 2879–2888. PMLR, 2018. URL <http://proceedings.mlr.press/v80/lake18a.html>.
- 754
755 Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Her-
nandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness
in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

- 756 Fangyu Lei, Qian Liu, Yiming Huang, Shizhu He, Jun Zhao, and Kang Liu. S3Eval: A synthetic,
757 scalable, systematic evaluation suite for large language model. In Kevin Duh, Helena Gomez,
758 and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter
759 of the Association for Computational Linguistics: Human Language Technologies (Volume 1:
760 Long Papers)*, pp. 1259–1286, Mexico City, Mexico, June 2024. Association for Computational
761 Linguistics. doi: 10.18653/v1/2024.naacl-long.69. URL <https://aclanthology.org/2024.naacl-long.69>.
- 763 Tianwen Li, Zhexiong Liu, Lindsay Matsumura, Elaine Wang, Diane Litman, and Richard Cor-
764 renti. Using large language models to assess young students’ writing revisions. In Ekaterina
765 Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaís Tack,
766 Victoria Yaneva, and Zheng Yuan (eds.), *Proceedings of the 19th Workshop on Innovative Use
767 of NLP for Building Educational Applications (BEA 2024)*, pp. 365–380, Mexico City, Mex-
768 ico, June 2024. Association for Computational Linguistics. URL [https://aclanthology.org/
769 2024.bea-1.30](https://aclanthology.org/2024.bea-1.30).
- 770 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan
771 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth
772 International Conference on Learning Representations*, 2024. URL [https://openreview.net/
773 forum?id=v8L0pN6EOi](https://openreview.net/forum?id=v8L0pN6EOi).
- 774 B. Liu, Yuqian Jiang, Xiaohan Zhang, Qian Liu, Shiqi Zhang, Joydeep Biswas, and Peter
775 Stone. Llm+p: Empowering large language models with optimal planning proficiency. *ArXiv*,
776 [abs/2304.11477](https://api.semanticscholar.org/CorpusID:258298051), 2023a. URL <https://api.semanticscholar.org/CorpusID:258298051>.
- 778 Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers
779 learn shortcuts to automata. In *The Eleventh International Conference on Learning Representa-
780 tions*, 2023b. URL <https://openreview.net/forum?id=De4FYqjFueZ>.
- 782 Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord,
783 Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for
784 science question answering. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave,
785 K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Confer-
786 ence on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA,
787 November 28 - December 9, 2022*, 2022. URL [http://papers.nips.cc/paper_files/paper/
788 2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html).
- 789 Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At
790 Which Training Stage Does Code Data Help LLMs Reasoning? In *The Twelfth International
791 Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenRe-
792 view.net, 2024. URL <https://openreview.net/forum?id=KIPJKST4gw>.
- 794 Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it
795 takes two to tango. *ArXiv*, [abs/2209.07686](https://api.semanticscholar.org/CorpusID:252355328), 2022. URL [https://api.semanticscholar.org/
796 CorpusID:252355328](https://api.semanticscholar.org/CorpusID:252355328).
- 797 William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought.
798 In *The International Conference on Learning Representations*, volume [abs/2310.07923](https://api.semanticscholar.org/CorpusID:263909434), 2024.
799 URL <https://api.semanticscholar.org/CorpusID:263909434>.
- 800 Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David
801 Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Au-
802 gustus Odena. Show your work: Scratchpads for intermediate computation with language models,
803 2022. URL <https://openreview.net/forum?id=iedYJm92o0a>.
- 804 Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-
805 Lezama. Is self-repair a silver bullet for code generation? In *The Twelfth International Conference
806 on Learning Representations*, 2024. URL <https://openreview.net/forum?id=y0GJXRungR>.
- 808 OpenAI. GPT-4 Technical Report. *ArXiv*, [abs/2303.08774](https://api.semanticscholar.org/CorpusID:257532815), 2023. URL [https://api.
809 semanticscholar.org/CorpusID:257532815](https://api.semanticscholar.org/CorpusID:257532815).

- 810 Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering large
811 language models with symbolic solvers for faithful logical reasoning. In Houda Bouamor,
812 Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics:
813 EMNLP 2023*, pp. 3806–3824, Singapore, December 2023. Association for Computational Lin-
814 guistics. doi: 10.18653/v1/2023.findings-emnlp.248. URL [https://aclanthology.org/2023.
815 findings-emnlp.248](https://aclanthology.org/2023.findings-emnlp.248).
- 816 Xiangyu Peng, Siyan Li, Sarah Wiegrefe, and Mark Riedl. Inferring the reader: Guiding auto-
817 mated story generation with commonsense reasoning. In Yoav Goldberg, Zornitsa Kozareva, and
818 Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp.
819 7008–7029, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational
820 Linguistics. doi: 10.18653/v1/2022.findings-emnlp.520. URL [https://aclanthology.org/
821 2022.findings-emnlp.520](https://aclanthology.org/2022.findings-emnlp.520).
- 822 Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lina Zhang, Fan Yang, and Mao Yang. Mutual Rea-
823 soning Makes Smaller LLMs Stronger Problem-Solvers. *arXiv preprint arXiv:2408.06195*, 2024.
824
- 825 Xin Quan, Marco Valentino, Louise Dennis, and Andre Freitas. Enhancing ethical explanations of
826 large language models through iterative symbolic refinement. In Yvette Graham and Matthew
827 Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association
828 for Computational Linguistics (Volume 1: Long Papers)*, pp. 1–22, St. Julian’s, Malta, March
829 2024. Association for Computational Linguistics. URL [https://aclanthology.org/2024.
830 eacl-long.1](https://aclanthology.org/2024.eacl-long.1).
- 831 Machel Reid and et. al. Gemini 1.5: Unlocking multimodal understanding across millions of to-
832 kens of context. *ArXiv*, abs/2403.05530, 2024. URL [https://api.semanticscholar.org/
833 CorpusID:268297180](https://api.semanticscholar.org/CorpusID:268297180).
- 834 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien
835 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A
836 Benchmark, 2023.
837
- 838 Gemma Team Morgane Riviere and et. al. Gemma 2: Improving open language models at a practical
839 size. 2024. URL <https://api.semanticscholar.org/CorpusID:270843326>.
- 840 Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: an ad-
841 versarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, aug 2021. ISSN
842 0001-0782. doi: 10.1145/3474381. URL <https://doi.org/10.1145/3474381>.
- 843 Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Com-
844 monsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Em-
845 pirical Methods in Natural Language Processing and the 9th International Joint Conference on
846 Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, Novem-
847 ber 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL
848 <https://aclanthology.org/D19-1454>.
- 849 Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis
850 of chain-of-thought. In *The Eleventh International Conference on Learning Representations*,
851 2023. URL <https://openreview.net/forum?id=qFVVBzXxR2V>.
- 852 Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the
853 limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference
854 on Learning Representations*, 2024. URL <https://openreview.net/forum?id=jenyQzue1>.
- 855 Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam
856 Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the
857 imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint
858 arXiv:2206.04615*, 2022.
- 859 Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. Exploring LLM prompt-
860 ing strategies for joint essay scoring and feedback generation. In Ekaterina Kochmar, Marie Bex-
861 te, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng
862
863

- 864 Yuan (eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educa-*
865 *tional Applications (BEA 2024)*, pp. 283–298, Mexico City, Mexico, June 2024. Association for
866 Computational Linguistics. URL <https://aclanthology.org/2024.bea-1.23>.
867
- 868 Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations
869 of large language models on reasoning and planning tasks. *ArXiv*, abs/2402.08115, 2024a. URL
870 <https://api.semanticscholar.org/CorpusID:267637077>.
- 871 Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. Chain of thoughtlessness? an
872 analysis of cot in planning. 2024b. URL [https://api.semanticscholar.org/CorpusID:](https://api.semanticscholar.org/CorpusID:269626390)
873 [269626390](https://api.semanticscholar.org/CorpusID:269626390).
- 874 Simeng Sun, Yang Liu, Shuohang Wang, Dan Iter, Chenguang Zhu, and Mohit Iyyer. PEARL:
875 Prompting large language models to plan and execute actions over long documents. In Yvette
876 Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chap-*
877 *ter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 469–486,
878 St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL [https:](https://aclanthology.org/2024.eacl-long.29)
879 [//aclanthology.org/2024.eacl-long.29](https://aclanthology.org/2024.eacl-long.29).
- 880 Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung,
881 Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench
882 tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and
883 Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*,
884 pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi:
885 10.18653/v1/2023.findings-acl.824. URL [https://aclanthology.org/2023.findings-acl.](https://aclanthology.org/2023.findings-acl.824)
886 [824](https://aclanthology.org/2023.findings-acl.824).
- 887 Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A ques-
888 tion answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Con-*
889 *ference of the North American Chapter of the Association for Computational Linguistics: Human*
890 *Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Min-
891 nesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL
892 <https://aclanthology.org/N19-1421>.
893
- 894 Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei,
895 Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas
896 Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes,
897 Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S.
898 Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian
899 Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut
900 Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov,
901 Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta,
902 Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh
903 Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov,
904 Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert
905 Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat
906 models. *ArXiv*, abs/2307.09288, 2023. URL [https://api.semanticscholar.org/CorpusID:](https://api.semanticscholar.org/CorpusID:259950998)
907 [259950998](https://api.semanticscholar.org/CorpusID:259950998).
- 908 Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-
909 hop questions via single-hop question composition. *Transactions of the Association for Compu-*
910 *tational Linguistics*, 2022.
- 911 Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the
912 planning abilities of large language models - a critical investigation. In *Thirty-seventh Conference*
913 *on Neural Information Processing Systems*, 2023. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=X6dEqXISeW)
914 [X6dEqXISeW](https://openreview.net/forum?id=X6dEqXISeW).
- 915 Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kamb-
916 hampati. PlanBench: An extensible benchmark for evaluating large language models on planning
917 and reasoning about change. In *Proceedings of the 37th International Conference on Neural*
Information Processing Systems, NIPS ’23, Red Hook, NY, USA, 2024. Curran Associates Inc.

- 918 Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. Theory of mind abilities of
919 large language models in human-robot interaction: An illusion? *Companion of the 2024*
920 *ACM/IEEE International Conference on Human-Robot Interaction*, 2024. URL <https://api.semanticscholar.org/CorpusID:266902529>.
921
- 922 Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun.
923 Towards understanding chain-of-thought prompting: An empirical study of what matters. In
924 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual*
925 *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2717–
926 2739, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/
927 v1/2023.acl-long.153. URL <https://aclanthology.org/2023.acl-long.153>.
928
- 929 Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim.
930 Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language
931 models. In *Annual Meeting of the Association for Computational Linguistics*, 2023b. URL <https://api.semanticscholar.org/CorpusID:258558102>.
932
- 933 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha
934 Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language
935 models. In *The Eleventh International Conference on Learning Representations*, 2023c. URL
936 <https://openreview.net/forum?id=1PL1NIMMrw>.
937
- 938 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming
939 Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi
940 Fan, Xiang Yue, and Wenhui Chen. MMLU-Pro: A More Robust and Challenging Multi-Task
941 Language Understanding Benchmark, 2024.
- 942 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
943 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
944 *Neural Information Processing Systems*, 35:24824–24837, 2022.
945
- 946 Li Siang Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka,
947 Jacob Andreas, and Joshua B. Tenenbaum. From word models to world models: Translating from
948 natural language to the probabilistic language of thought. *ArXiv*, abs/2306.12672, 2023. URL
949 <https://api.semanticscholar.org/CorpusID:259224900>.
- 950 Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and
951 Yu Su. TravelPlanner: A Benchmark for Real-World Planning with Language Agents. *ArXiv*,
952 abs/2402.01622, 2024. URL <https://api.semanticscholar.org/CorpusID:267406800>.
953
- 954 Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs
955 Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The*
956 *Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May*
957 *7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=gjeQKfXfPz>.
- 958 Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical
959 reasoning via symbolic chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar
960 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*
961 *(Volume 1: Long Papers)*, pp. 13326–13365, Bangkok, Thailand, August 2024a. Association for
962 Computational Linguistics. URL <https://aclanthology.org/2024.acl-long.720>.
963
- 964 Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, and Jian-Guang
965 Lou. Re-reading improves reasoning in language models, 2024b. URL <https://openreview.net/forum?id=3jXCF5dNpC>.
966
- 967 Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun
968 Chen. Large language models as optimizers. In *The Twelfth International Conference on Learning*
969 *Representations*, 2024. URL <https://openreview.net/forum?id=Bb4VGOWELI>.
970
- 971 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik
Narasimhan. Tree of Thoughts: Deliberate problem solving with large language models, 2023.

- 972 Xi Ye, Qiaochu Chen, Isil Dillig, and Greg Durrett. Satisfiability-aided language models using
973 declarative prompting. In *Advances in Neural Information Processing Systems*, 2023.
974
- 975 Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav
976 Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue.
977 A careful examination of large language model performance on grade school arithmetic. *ArXiv*,
978 abs/2405.00332, 2024. URL <https://api.semanticscholar.org/CorpusID:269484687>.
- 979 Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le,
980 and Denny Zhou. Take a step back: Evoking reasoning via abstraction in large language models.
981 In *The Twelfth International Conference on Learning Representations*, 2024a. URL <https://openreview.net/forum?id=3bq3jsvcQ1>.
982
- 983 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
984 Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica.
985 Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th Inter-*
986 *national Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA,
987 2024b. Curran Associates Inc.
- 988
- 989 Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied,
990 Weizhu Chen, and Nan Duan. AGIEval: A Human-Centric Benchmark for Evaluating Foundation
991 Models, 2023.
- 992 Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. Learning to decompose: Hypotheti-
993 cal question decomposition based on comparable texts. In Yoav Goldberg, Zornitsa Kozareva,
994 and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natu-*
995 *ral Language Processing*, pp. 2223–2235, Abu Dhabi, United Arab Emirates, December 2022.
996 Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.142. URL
997 <https://aclanthology.org/2022.emnlp-main.142>.
- 998 Ben Zhou, Hongming Zhang, Sihao Chen, Dian Yu, Hongwei Wang, Baolin Peng, Dan Roth, and
999 Dong Yu. Conceptual and unbiased reasoning in language models. *ArXiv*, abs/2404.00205, 2024.
1000 URL <https://api.semanticscholar.org/CorpusID:268820105>.
1001
- 1002 Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuur-
1003 mans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables
1004 complex reasoning in large language models. In *The Eleventh International Conference on Learn-*
1005 *ing Representations*, 2023a. URL <https://openreview.net/forum?id=WZH7099tgfM>.
- 1006 Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and
1007 Jimmy Ba. Large Language Models are Human-Level Prompt Engineers. In *The Eleventh In-*
1008 *ternational Conference on Learning Representations*, 2023b. URL [https://openreview.net/](https://openreview.net/forum?id=92gvk82DE-)
1009 [forum?id=92gvk82DE-](https://openreview.net/forum?id=92gvk82DE-).
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

A META-ANALYSIS EXPANDED DETAILS ON CRITERIA AND PROCESS

Automatic Selection and Paper Filtering We investigate all papers from ICLR 2024, a representative ML venue, and two representative NLP venues, EACL 2024 and NAACL 2024 (including Findings and Workshop papers). We filtered all 4,642 papers (2,259 from ICLR 2024 and 2,382 from the two ACL-affiliated conferences) for those with at least two occurrences of “CoT”, “chain-of-thought”, or “chain of thought”, resulting in 516 papers. There are conceivably papers using CoT called by another name (e.g., Scratchpads), but we believe these 516 give a representative sample appropriate for systematic analysis.

Manual Paper Filtering and Results Extraction We then filter down to papers that perform a comparison of CoT prompting vs. direct prompting, whether or not this is core to the paper’s research question. We manually filtered the 516 papers in question and extracted the key results from those that remained. We excluded multimodal models, CoT-fine-tuned models, any experiments where the “CoT” method involves multiple forward passes (e.g., self-consistency (Wang et al., 2023c) and tree-of-thought (Yao et al., 2023)),³ and systems that augment LLMs with external tools (discussed more in Section 5).

For each paper passing through these criteria, we manually extracted the results from key tables comparing CoT and direct answer prompts. We only include results where the CoT and direct prompts are run on the same model and same dataset while being on a scale of 0 to 100 (excluding Likert scale evaluations, for example) for a more direct comparison. When papers include various CoT or direct answer prompts (including zero/few-shot variants), we always take the best-performing prompt for both. We focus on key test results where applicable, excluding dev sets if they are reported alongside test and also excluding numbers from ablations or nonstandard subsets of datasets.

This resulted in a total of 1,218 experimental comparisons across 110 papers (35 from ICLR and 75 from NAACL and EACL) covering 264 datasets. Details and more information will be available in our GitHub Repo.

Categorization Given the large number of tasks and datasets being compared, we grouped each task into a set of 14 categories. These categories were determined based on the description (and possibly examples) of the task, not taking into account system performance. These categories abstract over traditional NLP task classifications (e.g., NER, reading comprehension) and take into account both the task format and the kinds of reasoning involved. Definitions for several categories are shown in Table 1 and the full description is given in Appendix B.

B QUANTITATIVE META-ANALYSIS

See the full list of categories and their descriptions that we used for the meta-analysis in Table 2.

C EXPANDED EXPERIMENTAL DETAILS

A full list of the datasets can be found in Table 4. Each model can be seen in Table 5. We use one answer parser for all datasets of the same answer response format (one for multiple choice, short answer, etc.); however, some datasets require special handling and have edge cases that we handle separately from the rest of the datasets. Similarly, for each model, we use the exact same prompt across them, except when closed source models require different prompts because they do not allow for partial completions (i.e., when we cannot put “*let’s think step by step*” to warm-start the assistant’s response). All prompts are given in our Huggingface repo, including the model output and what our answer parser extracted as the answer.

Experiments were conducted either by invoking APIs or by running open-source models on our own hardware, mostly on a machine with 8 A40s or 4 Quadro RTX 8000s. All locally hosted models were

³These systems use more compute than direct answer, and there is not a clear comparison to be made here. Moreover, our anecdotal coverage of these methods shows that they are most used for math, coding, and logic settings, for which we already have high representation among reported CoT methods.

Table 2: Categories and their descriptions for the meta-analysis.

Category	Description
Symbolic and algorithmic	Tasks involving symbol manipulation which can be solved by executing a program. This includes entity tracking datasets (e.g., SCONE, Coin Flip) and algorithmic tasks (e.g., BBH word sorting or finding shortest paths in a graph).
Math	Tasks requiring mathematical reasoning, from grade-school math to advanced mathematics, including physics questions.
Logical reasoning	Tasks designed to test for logical reasoning, whether deductive (Saparov & He, 2023, PrOntoQA), inductive (Bowen et al., 2024) or analogical (Ma et al., 2024) reasoning, including syllogisms and logical puzzles.
Commonsense reasoning	Datasets designed to test for commonsense knowledge and reasoning, i.e., world knowledge that most people would have, rather than specialized expert-level knowledge in a discipline acquired after years of study.
Encyclopedic knowledge	Tasks requiring expert-level in-depth knowledge beyond mere commonsense, usually in an open-book setting.
Spatial and temporal reasoning	Datasets designed to test for an understanding of space and spatial relations (e.g., navigation) or reasoning involving time and sequences over time.
Multi-hop QA	Questions involving the composition of multiple steps of reasoning in order to arrive at an answer, such as “What is the capital of the country whose scientist discovered penicillin?”
Context-aware QA	Tasks such as closed-book QA and reading comprehension involving reasoning about a given text in context. The context is often a short passage, but could also take the form of a knowledge graph (KBQA) or a table. This category also includes information extraction tasks, such as NER or relation extraction.
Entailment	Tasks involving establishing the inferential relation between two texts, prototypically NLI, but also including fact verification.
Text classification	Tasks involving the classification of a text into a small set of categories, such a topic or sentiment classification, but also involving tasks such as hate speech detection and misinformation detection.
Generation	Tasks involving text generation, including machine translation, dialogue, question generation, as well as code generation. Tasks such as SQL execution (Lei et al., 2024) or systematic transformations of data (e.g., SCAN (Lake & Baroni, 2018)) are excluded because they can be solved by executing a program.
Meta-linguistic	Tasks probing for models’ knowledge of linguistics, such as identifying the main subject of a sentence or solving linguistic puzzles.
Mixed datasets	Datasets containing a variety of tasks, such as BIG-Bench Hard (BBH) or MMLU.
Other	Tasks which did not fit in any of the other categories, such as evaluating AI safety, eliciting models’ verbalized confidence, or melody retrieval.

hosted with vLLM. All parameters given to the vLLM API endpoint are given in the Huggingface repo as well.

D OTHER CoT PROMPT VARIANTS

D.1 TESTING PERFORMANCE VOLATILITY ACROSS PROMPTS

To test the impact of prompt choice on performance, we performed our zero-shot experiment on Llama 3.1 8B with 7 different datasets and 4 different zero-shot CoT prompting strategies common in the literature (Kojima et al., 2022; Wang et al., 2023b; Zhou et al., 2023b; Yang et al., 2024).

Table 3: Models, datasets, and prompting strategies used in our experiments. Models marked with † are run with a 4k context size window. Note that Gemma has a larger than 4k context size window, but VLLM only supports up to a 4k context size window for it. Models marked with * indicate closed-source models that cannot handle prefixed assistant messages. Datasets marked with Δ do not have a few-shot setting.

Models	Llama 2 7B Chat † (Touvron et al., 2023), Mistral 7B Instruct v0.3 (Jiang et al., 2023), Llama 3.1 8B Instruct (Dubey et al., 2024), Llama 3.1 70B Instruct, Gemma 2 9B It† (Riviere & et. al, 2024), Phi-3 Small 8k Instruct (Abdin et al., 2024), gpt-4o-mini-2024-07-18*, gpt-4o-2024-08-06*, Gemini 1.5 Flash* (Reid & et. al, 2024), Gemini 1.5 Pro* (Reid & et. al, 2024), claude-3-haiku-20240307* (Anthropic, a), claude-3-5-sonnet-20240620* (Anthropic, b)
Datasets	CommonsenseQA (Talmor et al., 2019), StrategyQA (Geva et al., 2021), SiQA Δ Sap et al. (2019), PiQA Δ (Bisk et al., 2019), Winogrande Δ (Sakaguchi et al., 2021), GPQA (Rein et al., 2023), MuSR (Sprague et al., 2024), ContextHub (Levels 1 and 2 only) (Hua et al., 2024), ARC Δ (Clark et al., 2018), AGIEval LSAT (Zhong et al., 2023), MMLU (Hendrycks et al., 2021a), MMLU Pro (Wang et al., 2024), MATH (Hendrycks et al., 2021b), GSM8K (Cobbe et al., 2021), GSM8K-hard (Gao et al., 2023), FOLIO (Han et al., 2022), MuSiQue Δ (Trivedi et al., 2022), Big-Bench Hard (Suzgun et al., 2023; Srivastava et al., 2022), BiGGen Bench (Kim et al., 2024)
Prompts	zero-shot direct answer, zero-shot CoT (Kojima et al., 2022), few-shot direct answer (Brown et al., 2020), few-shot CoT (Wei et al., 2022)

Table 4: List of datasets used in our experiments. We categorize each dataset into one of five categories based on the type of reasoning required: Commonsense, Knowledge, Soft Reasoning, Symbolic, or Mathematical. We also report answer formats. When we use few-shot prompts, we mark how many examples those prompts contain. BiGGen Bench has many categories of questions that explicitly ask for CoTs in the response; we ignore those categories for our evaluation.

Dataset	Type	Answer Format	m -Shots
CommonsenseQA	Commonsense	Multiple choice	7
StrategyQA	Commonsense	True or False	6
SIQA	Commonsense	Multiple choice	0
PIQA	Commonsense	Multiple choice	0
Winogrande	Commonsense	Multiple choice	0
Arc Easy	Knowledge	Multiple choice	0
Arc Challenge	Knowledge	Multiple choice	0
AGIEval LSAT	Soft Reasoning	Multiple choice	3
BiGGen-Bench	Soft Reasoning	Free response	0
MMLU	Knowledge	Multiple Choice	5
MMLU Pro	Knowledge	Multiple Choice	5
BigBench-Hard	Symbolic	Multiple Choice	0
MuSR	Soft Reasoning	Multiple Choice	1
GPQA	Mathematical	Multiple Choice	3
MuSiQue	Soft Reasoning	Short Answer	0
GSM8K	Mathematical	Short Answer	8
GSM8K-Hard	Mathematical	Short Answer	8
FOLIO	Symbolic	True, False, or Unknown	4
ContextHub	Symbolic	True, False, or Neither	3
MATH	Mathematical	Short Answer	4

Table 5: List of models for our experiments. We focus on contemporary instruction-tuned models; although pretrained and smaller language models could be used, they are not the focus of our study. Prompts and outputs used for each model are available on Huggingface. * Note that Gemma can accept more than 4k input tokens, but we are restricted to 4k by vLLM.

Model	Context Length	Is Open Source
Llama 2 7B Chat	4k	True
Mistral 7B Instruct v0.3	8k	True
Llama 3.1 8B Instruct	128k	True
Llama 3.1 70B Instruct	128k	True
Gemma 2 9B It	4k*	True
Qwen 7B Instruct	131k	True
Qwen 72B Instruct	131k	True
GPT4o-Mini	128k	False
GPT4o	128k	False
Gemini 1.5 Pro	128k	False
Gemini Flash	1m	False
Claude 3.5 Sonnet	200k	False
Claude 3 Haiku	200k	False

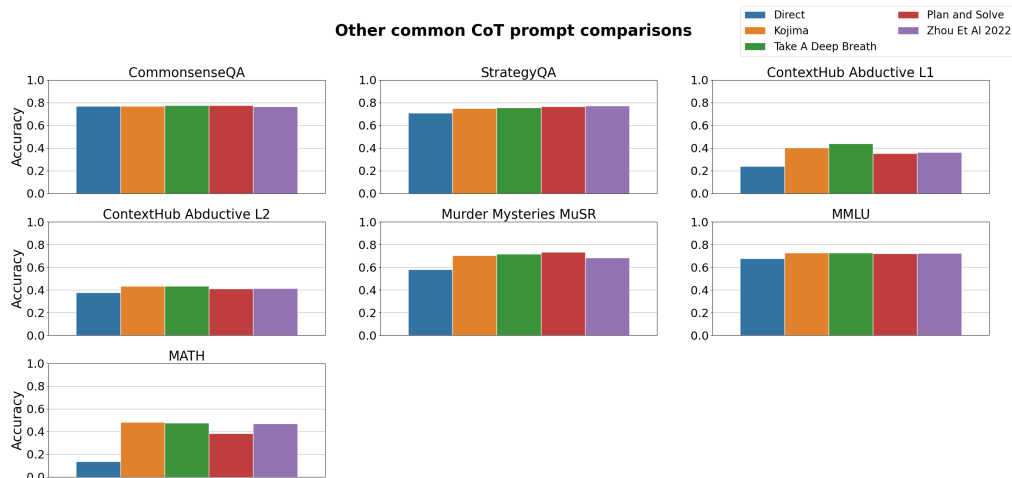


Figure 7: Performance of multiple prompts commonly used to elicit reasoning through CoT in the zero shot setting. Each prompt starts the assistant completion with a different phrase meant to elicit reasoning. All results are from using Llama 3.1 8B Instruct. For the Kojima variant, we explicitly place “Let’s think step by step.” in the assistant message. There is very little variation between the CoT prompts on average.

Figure 7 shows variation due to prompts is typically small and no prompt gives a consistent gain over the other. For our experiments, this suggests that different prompts have small effects on the overall outcome on average.

E FEW-SHOT EXPERIMENTS

Compared to a zero-shot prompt, a few-shot prompt additionally contains demonstrations of the relevant reasoning mode on different problem instances $\{(v(\mathbf{q}_i), \mathbf{y}_i^*)\}$. Few-shot prompts for direct answer simply encode the answer a_i as \mathbf{y}_i^* , whereas few-shot prompts for chain-of-thought include a reasoning trace ending in the correct answer. Now we can define the m -shot direct prompt as

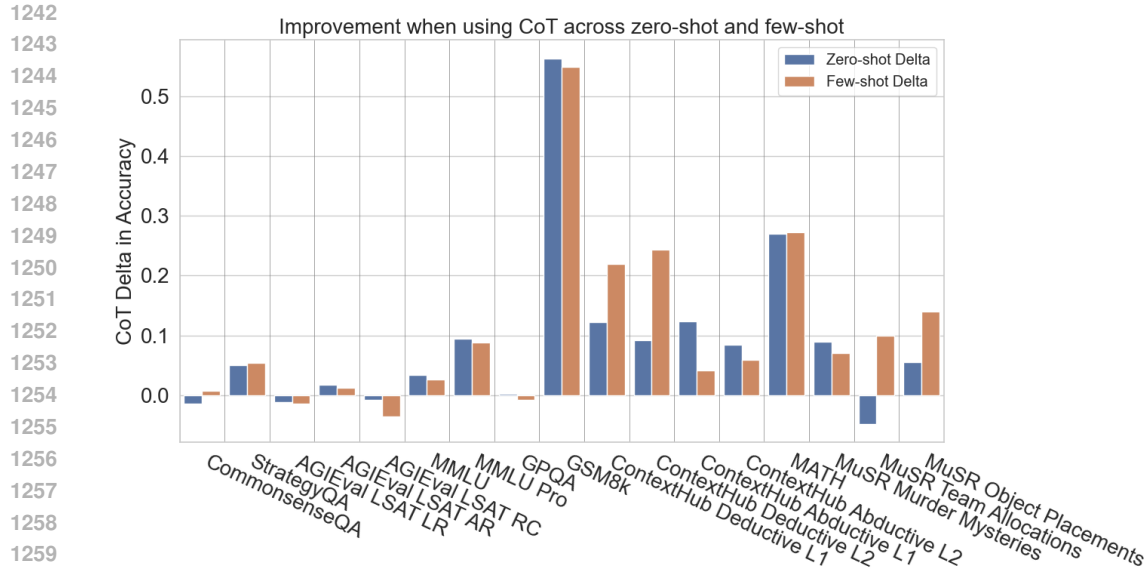


Figure 8: Average performance improvement from using CoT across different models in the zero-shot and few-shot settings. Each bar represents how much CoT improves the accuracy for that specific setting. In general, CoT in the few-shot setting does not change the qualitative performance of CoT versus zero-shot, though it can change the magnitude for symbolic datasets.

$$\mathcal{I}_{\text{da}}^m(\mathbf{q}) = v_{\text{da}}(\mathbf{q}_1)\mathbf{a}_1 v_{\text{da}}(\mathbf{q}_2)\mathbf{a}_2 \dots v_{\text{da}}(\mathbf{q}_m)\mathbf{a}_m v_{\text{da}}(\mathbf{q}) \text{ and the } m\text{-shot cot prompt as } \mathcal{I}_{\text{cot}}^m(\mathbf{q}) = v_{\text{cot}}(\mathbf{q}_1)\mathbf{y}_1^* v_{\text{cot}}(\mathbf{q}_2)\mathbf{y}_2^* \dots v_{\text{cot}}(\mathbf{q}_m)\mathbf{y}_m^* v_{\text{cot}}(\mathbf{q}).$$

Figure 8 shows the difference between few-shot prompting and the zero-shot setting discussed in the main text of the paper. We see that using CoT in the few-shot setting largely does not change the datasets that benefit from it. Only one dataset, MuSR Team Allocation, starts to improve with few-shot; however, we believe this to be an exception because the final step to derive the answer is complex in the prompt and clearer in the examples. The magnitude of improvement over direct answer prompting when using CoT is also similar to the zero-shot setting.

F EXPANDED CoT VS DIRECT EXPERIMENTAL RESULTS

F.1 FULL ZERO-SHOT RESULTS

Table 6: Direct answer and CoT accuracies for each reasoning category across models.

Model	Commonsense		Knowledge		Mathematical		Symbolic		Soft	
	DA %	CoT %	DA %	CoT %	DA %	CoT %	DA %	CoT %	DA %	CoT %
Claude-3 Haiku	74.3	77.2	73.0	76.1	18.1	48.2	38.6	48.7	55.9	56.6
Claude-3.5 Sonnet	84.3	85.8	83.8	88.8	38.7	59.0	53.2	67.1	67.6	75.7
GPT-4o Mini	81.8	83.2	73.6	83.1	22.9	59.7	48.1	60.9	61.1	63.5
Gemini 1.5 Flash	80.3	76.8	78.2	81.0	27.2	55.7	47.0	59.7	60.6	62.6
Gemini 1.5 Pro	80.4	78.3	80.9	83.8	35.4	58.5	52.9	62.6	64.1	67.8
Gemma 2 9b	75.0	76.1	74.9	76.9	18.5	50.5	46.7	55.8	58.2	60.5
Gpt-4o	87.3	87.7	82.9	88.6	36.5	63.3	55.7	68.3	65.9	74.0
Meta-Llama 2 7b	51.4	50.9	44.1	46.6	9.3	17.2	22.4	35.4	37.2	37.6
Meta-Llama 3.1 70b	84.2	84.7	82.4	85.6	24.9	54.9	49.0	60.0	65.7	69.5
Meta-Llama 3.1 8b	72.9	73.4	70.1	74.1	16.0	47.8	34.8	51.6	55.0	56.2
Mistral 7b	58.3	61.8	62.0	64.5	10.9	28.9	41.8	45.0	48.6	49.7
Phi-3 Small 8k	70.8	72.5	76.1	79.7	17.8	47.1	51.2	58.7	57.9	56.4
Qwen 2 72b	82.9	84.9	78.6	84.6	23.9	58.5	48.2	58.7	64.2	65.1
Qwen 2 7b	64.0	66.1	65.2	71.3	15.9	53.5	43.8	52.3	54.4	49.4
Average	74.8	75.7	73.3	77.5	22.6	50.2	45.2	56.1	58.3	60.3

Table 7: Zero-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	zero-shot CoT accuracy	zero-shot DA accuracy
MuSR Team Allocations	Soft Reasoning	Llama 2 7b	34.8	37.2
MuSR Team Allocations	Soft Reasoning	Mistral 7b	38.8	46.8
MuSR Team Allocations	Soft Reasoning	Llama 3.1 8b	44.0	48.0
MuSR Team Allocations	Soft Reasoning	Llama 3.1 70b	65.2	66.8
MuSR Team Allocations	Soft Reasoning	Gemma 2 9b	47.2	44.8
MuSR Team Allocations	Soft Reasoning	Phi-3 Small 8k	47.2	61.6
MuSR Team Allocations	Soft Reasoning	Qwen 2 7b	42.0	49.6
MuSR Team Allocations	Soft Reasoning	Qwen 2 72b	58.0	66.8
MuSR Team Allocations	Soft Reasoning	GPT-4o Mini	61.2	58.4
MuSR Team Allocations	Soft Reasoning	Gpt-4o	64.0	63.6
MuSR Team Allocations	Soft Reasoning	Claude-3 Haiku	56.8	59.2
MuSR Team Allocations	Soft Reasoning	Claude-3.5 Sonnet	80.4	63.2
MuSR Team Allocations	Soft Reasoning	Gemini 1.5 Flash	48.8	55.2
MuSR Team Allocations	Soft Reasoning	Gemini 1.5 Pro	58.4	62.4
SiQA	Commonsense	Llama 2 7b	53.4	55.9
SiQA	Commonsense	Mistral 7b	35.9	33.5
SiQA	Commonsense	Llama 3.1 8b	73.5	73.5
SiQA	Commonsense	Llama 3.1 70b	78.7	80.9
SiQA	Commonsense	Gemma 2 9b	74.9	76.3
SiQA	Commonsense	Phi-3 Small 8k	38.0	40.4
SiQA	Commonsense	Qwen 2 7b	37.3	39.3
SiQA	Commonsense	Qwen 2 72b	80.5	80.4
SiQA	Commonsense	GPT-4o Mini	79.0	80.0
SiQA	Commonsense	Gpt-4o	81.9	81.5
SiQA	Commonsense	Claude-3 Haiku	75.4	74.8
SiQA	Commonsense	Claude-3.5 Sonnet	79.7	81.0
SiQA	Commonsense	Gemini 1.5 Flash	74.5	79.1
SiQA	Commonsense	Gemini 1.5 Pro	73.9	78.2
MuSiQue	Soft Reasoning	Llama 2 7b	40.1	36.1
MuSiQue	Soft Reasoning	Mistral 7b	47.3	47.2
MuSiQue	Soft Reasoning	Llama 3.1 8b	62.6	64.7
MuSiQue	Soft Reasoning	Llama 3.1 70b	74.0	72.2
MuSiQue	Soft Reasoning	Gemma 2 9b	67.7	68.7
MuSiQue	Soft Reasoning	Phi-3 Small 8k	58.3	64.3
MuSiQue	Soft Reasoning	Qwen 2 7b	60.7	65.1
MuSiQue	Soft Reasoning	Qwen 2 72b	56.3	69.0
MuSiQue	Soft Reasoning	GPT-4o Mini	71.3	68.2
MuSiQue	Soft Reasoning	Gpt-4o	73.5	70.1
MuSiQue	Soft Reasoning	Claude-3 Haiku	54.8	56.0
MuSiQue	Soft Reasoning	Claude-3.5 Sonnet	66.9	70.4
MuSiQue	Soft Reasoning	Gemini 1.5 Flash	69.8	66.2
MuSiQue	Soft Reasoning	Gemini 1.5 Pro	69.8	71.3
AGIEval LSAT RC	Soft Reasoning	Llama 2 7b	31.2	36.4
AGIEval LSAT RC	Soft Reasoning	Mistral 7b	61.7	61.0
AGIEval LSAT RC	Soft Reasoning	Llama 3.1 8b	71.0	68.8
AGIEval LSAT RC	Soft Reasoning	Llama 3.1 70b	84.4	87.0
AGIEval LSAT RC	Soft Reasoning	Gemma 2 9b	75.1	78.1
AGIEval LSAT RC	Soft Reasoning	Phi-3 Small 8k	68.8	69.9
AGIEval LSAT RC	Soft Reasoning	Qwen 2 7b	61.0	66.5
AGIEval LSAT RC	Soft Reasoning	Qwen 2 72b	83.6	84.4

Table 7: Zero-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	zero-shot CoT accuracy	zero-shot DA accuracy
1350	AGIEval LSAT RC	Soft Reasoning GPT-4o Mini	77.3	74.3
1351	AGIEval LSAT RC	Soft Reasoning Gpt-4o	88.1	81.4
1352	AGIEval LSAT RC	Soft Reasoning Claude-3 Haiku	71.7	65.1
1353	AGIEval LSAT RC	Soft Reasoning Claude-3.5 Sonnet	90.0	89.6
1354	AGIEval LSAT RC	Soft Reasoning Gemini 1.5 Flash	78.1	81.0
1355	AGIEval LSAT RC	Soft Reasoning Gemini 1.5 Pro	82.2	85.9
1356	CommonsenseQA	Commonsense Llama 2 7b	49.4	54.6
1357	CommonsenseQA	Commonsense Mistral 7b	68.0	68.0
1358	CommonsenseQA	Commonsense Llama 3.1 8b	68.5	74.9
1359	CommonsenseQA	Commonsense Llama 3.1 70b	83.5	84.4
1360	CommonsenseQA	Commonsense Gemma 2 9b	79.2	80.1
1361	CommonsenseQA	Commonsense Phi-3 Small 8k	81.8	80.3
1362	CommonsenseQA	Commonsense Qwen 2 7b	78.5	79.0
1363	CommonsenseQA	Commonsense Qwen 2 72b	87.4	87.3
1364	CommonsenseQA	Commonsense GPT-4o Mini	82.5	83.9
1365	CommonsenseQA	Commonsense Gpt-4o	86.5	87.3
1366	CommonsenseQA	Commonsense Claude-3 Haiku	80.6	79.0
1367	CommonsenseQA	Commonsense Claude-3.5 Sonnet	85.1	84.3
1368	CommonsenseQA	Commonsense Gemini 1.5 Flash	79.7	82.6
1369	CommonsenseQA	Commonsense Gemini 1.5 Pro	79.9	82.9
1370	GPQA	Mathematical Llama 2 7b	28.3	24.3
1371	GPQA	Mathematical Mistral 7b	23.0	24.3
1372	GPQA	Mathematical Llama 3.1 8b	24.1	25.9
1373	GPQA	Mathematical Llama 3.1 70b	23.2	25.9
1374	GPQA	Mathematical Gemma 2 9b	26.3	21.2
1375	GPQA	Mathematical Phi-3 Small 8k	22.3	20.8
1376	GPQA	Mathematical Qwen 2 7b	24.1	24.6
1377	GPQA	Mathematical Qwen 2 72b	21.0	18.1
1378	GPQA	Mathematical GPT-4o Mini	21.0	24.0
1379	GPQA	Mathematical Gpt-4o	23.7	25.9
1380	GPQA	Mathematical Claude-3 Haiku	25.4	22.3
1381	GPQA	Mathematical Claude-3.5 Sonnet	25.4	25.9
1382	GPQA	Mathematical Gemini 1.5 Flash	22.3	22.8
1383	GPQA	Mathematical Gemini 1.5 Pro	21.0	23.7
1384	AGIEval LSAT LR	Soft Reasoning Llama 2 7b	29.4	33.5
1385	AGIEval LSAT LR	Soft Reasoning Mistral 7b	44.1	47.8
1386	AGIEval LSAT LR	Soft Reasoning Llama 3.1 8b	59.0	53.9
1387	AGIEval LSAT LR	Soft Reasoning Llama 3.1 70b	81.4	81.0
1388	AGIEval LSAT LR	Soft Reasoning Gemma 2 9b	64.9	67.6
1389	AGIEval LSAT LR	Soft Reasoning Phi-3 Small 8k	64.5	64.1
1390	AGIEval LSAT LR	Soft Reasoning Qwen 2 7b	50.6	58.4
1391	AGIEval LSAT LR	Soft Reasoning Qwen 2 72b	77.3	75.1
1392	AGIEval LSAT LR	Soft Reasoning GPT-4o Mini	65.3	68.2
1393	AGIEval LSAT LR	Soft Reasoning Gpt-4o	87.3	83.9
1394	AGIEval LSAT LR	Soft Reasoning Claude-3 Haiku	55.7	54.7
1395	AGIEval LSAT LR	Soft Reasoning Claude-3.5 Sonnet	83.7	82.7
1396	AGIEval LSAT LR	Soft Reasoning Gemini 1.5 Flash	70.0	71.2
1397	AGIEval LSAT LR	Soft Reasoning Gemini 1.5 Pro	79.4	80.4
1398	PiQA	Commonsense Llama 2 7b	62.1	64.7
1399	PiQA	Commonsense Mistral 7b	78.6	77.7
1400	PiQA	Commonsense Llama 3.1 8b	85.0	84.2
1401	PiQA	Commonsense Llama 3.1 70b	91.8	90.6
1402	PiQA	Commonsense Gemma 2 9b	84.0	84.8
1403	PiQA	Commonsense Phi-3 Small 8k	89.1	85.5
1404	PiQA	Commonsense Qwen 2 7b	84.3	86.2
1405	PiQA	Commonsense Qwen 2 72b	92.9	89.1
1406	PiQA	Commonsense GPT-4o Mini	93.1	88.6
1407	PiQA	Commonsense Gpt-4o	95.9	95.5
1408	PiQA	Commonsense Claude-3 Haiku	85.9	86.6
1409	PiQA	Commonsense Claude-3.5 Sonnet	94.6	94.5
1410	PiQA	Commonsense Gemini 1.5 Flash	84.6	89.8
1411	PiQA	Commonsense Gemini 1.5 Pro	88.1	91.3
1412	Arc Easy	Knowledge Llama 2 7b	71.1	69.8
1413	Arc Easy	Knowledge Mistral 7b	87.5	86.7
1414	Arc Easy	Knowledge Llama 3.1 8b	93.0	92.5
1415	Arc Easy	Knowledge Llama 3.1 70b	97.5	97.9
1416	Arc Easy	Knowledge Gemma 2 9b	94.9	95.8
1417	Arc Easy	Knowledge Phi-3 Small 8k	96.0	96.3
1418	Arc Easy	Knowledge Qwen 2 7b	89.5	84.7
1419	Arc Easy	Knowledge Qwen 2 72b	97.9	97.4
1420	Arc Easy	Knowledge GPT-4o Mini	96.8	94.6
1421	Arc Easy	Knowledge Gpt-4o	98.9	98.1
1422	Arc Easy	Knowledge Claude-3 Haiku	95.1	95.4
1423	Arc Easy	Knowledge Claude-3.5 Sonnet	98.6	98.4
1424	Arc Easy	Knowledge Gemini 1.5 Flash	96.8	97.2
1425	Arc Easy	Knowledge Gemini 1.5 Pro	97.2	94.6
1426	Arc Challenge	Knowledge Llama 2 7b	49.2	45.2

Table 7: Zero-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	zero-shot CoT accuracy	zero-shot DA accuracy	
1404					
1405					
1406	Arc Challenge	Knowledge	Mistral 7b	78.3	76.6
1407	Arc Challenge	Knowledge	Llama 3.1 8b	86.0	82.6
1408	Arc Challenge	Knowledge	Llama 3.1 70b	95.0	93.6
1409	Arc Challenge	Knowledge	Gemma 2 9b	91.0	89.6
1410	Arc Challenge	Knowledge	Phi-3 Small 8k	91.6	91.0
1411	Arc Challenge	Knowledge	Qwen 2 7b	83.9	75.3
1412	Arc Challenge	Knowledge	Qwen 2 72b	96.3	94.6
1413	Arc Challenge	Knowledge	GPT-4o Mini	93.3	82.6
1414	Arc Challenge	Knowledge	Gpt-4o	96.0	95.3
1415	Arc Challenge	Knowledge	Claude-3 Haiku	89.3	89.3
1416	Arc Challenge	Knowledge	Claude-3.5 Sonnet	96.0	95.3
1417	Arc Challenge	Knowledge	Gemini 1.5 Flash	92.3	93.6
1418	Arc Challenge	Knowledge	Gemini 1.5 Pro	91.6	90.6
1419	AGIEval LSAT AR	Soft Reasoning	Llama 2 7b	17.0	17.4
1420	AGIEval LSAT AR	Soft Reasoning	Mistral 7b	21.7	19.1
1421	AGIEval LSAT AR	Soft Reasoning	Llama 3.1 8b	20.4	26.1
1422	AGIEval LSAT AR	Soft Reasoning	Llama 3.1 70b	32.6	28.7
1423	AGIEval LSAT AR	Soft Reasoning	Gemma 2 9b	24.8	23.0
1424	AGIEval LSAT AR	Soft Reasoning	Phi-3 Small 8k	28.3	26.5
1425	AGIEval LSAT AR	Soft Reasoning	Qwen 2 7b	27.0	23.9
1426	AGIEval LSAT AR	Soft Reasoning	Qwen 2 72b	29.1	28.3
1427	AGIEval LSAT AR	Soft Reasoning	GPT-4o Mini	32.2	23.0
1428	AGIEval LSAT AR	Soft Reasoning	Gpt-4o	37.8	30.0
1429	AGIEval LSAT AR	Soft Reasoning	Claude-3 Haiku	24.8	23.5
1430	AGIEval LSAT AR	Soft Reasoning	Claude-3.5 Sonnet	38.3	33.9
1431	AGIEval LSAT AR	Soft Reasoning	Gemini 1.5 Flash	27.8	27.8
1432	AGIEval LSAT AR	Soft Reasoning	Gemini 1.5 Pro	30.0	31.7
1433	BiGGen Bench	Soft Reasoning	Llama 2 7b	61.6	56.8
1434	BiGGen Bench	Soft Reasoning	Mistral 7b	70.1	68.1
1435	BiGGen Bench	Soft Reasoning	Llama 3.1 8b	66.5	67.7
1436	BiGGen Bench	Soft Reasoning	Llama 3.1 70b	78.9	76.9
1437	BiGGen Bench	Soft Reasoning	Gemma 2 9b	64.7	64.5
1438	BiGGen Bench	Soft Reasoning	Phi-3 Small 8k	69.7	63.0
1439	BiGGen Bench	Soft Reasoning	Qwen 2 7b	46.2	69.9
1440	BiGGen Bench	Soft Reasoning	Qwen 2 72b	74.3	79.9
1441	BiGGen Bench	Soft Reasoning	GPT-4o Mini	70.3	77.7
1442	BiGGen Bench	Soft Reasoning	Gpt-4o	86.0	82.0
1443	BiGGen Bench	Soft Reasoning	Claude-3 Haiku	80.0	80.0
1444	BiGGen Bench	Soft Reasoning	Claude-3.5 Sonnet	91.4	79.3
1445	BiGGen Bench	Soft Reasoning	Gemini 1.5 Flash	73.9	68.5
1446	BiGGen Bench	Soft Reasoning	Gemini 1.5 Pro	78.7	67.1
1447	Winogrande	Commonsense	Llama 2 7b	49.9	50.4
1448	Winogrande	Commonsense	Mistral 7b	60.4	56.5
1449	Winogrande	Commonsense	Llama 3.1 8b	66.5	63.3
1450	Winogrande	Commonsense	Llama 3.1 70b	84.2	81.2
1451	Winogrande	Commonsense	Gemma 2 9b	68.7	67.7
1452	Winogrande	Commonsense	Phi-3 Small 8k	81.5	81.6
1453	Winogrande	Commonsense	Qwen 2 7b	67.1	60.7
1454	Winogrande	Commonsense	Qwen 2 72b	81.9	80.7
1455	Winogrande	Commonsense	GPT-4o Mini	79.2	71.9
1456	Winogrande	Commonsense	Gpt-4o	89.7	86.5
1457	Winogrande	Commonsense	Claude-3 Haiku	70.7	66.2
1458	Winogrande	Commonsense	Claude-3.5 Sonnet	89.4	85.7
1459	Winogrande	Commonsense	Gemini 1.5 Flash	72.5	74.8
1460	Winogrande	Commonsense	Gemini 1.5 Pro	75.5	78.3
1461	MMLU	Knowledge	Llama 2 7b	46.3	41.7
1462	MMLU	Knowledge	Mistral 7b	60.5	56.5
1463	MMLU	Knowledge	Llama 3.1 8b	72.6	67.5
1464	MMLU	Knowledge	Llama 3.1 70b	85.0	83.2
1465	MMLU	Knowledge	Gemma 2 9b	73.8	71.4
1466	MMLU	Knowledge	Phi-3 Small 8k	76.3	73.6
1467	MMLU	Knowledge	Qwen 2 7b	67.0	64.5
1468	MMLU	Knowledge	Qwen 2 72b	81.3	77.8
1469	MMLU	Knowledge	GPT-4o Mini	79.9	74.8
1470	MMLU	Knowledge	Gpt-4o	87.5	83.4
1471	MMLU	Knowledge	Claude-3 Haiku	72.2	68.4
1472	MMLU	Knowledge	Claude-3.5 Sonnet	87.2	84.0
1473	MMLU	Knowledge	Gemini 1.5 Flash	76.3	74.7
1474	MMLU	Knowledge	Gemini 1.5 Pro	81.3	81.1
1475	StrategyQA	Commonsense	Llama 2 7b	39.5	31.2
1476	StrategyQA	Commonsense	Mistral 7b	66.1	55.8
1477	StrategyQA	Commonsense	Llama 3.1 8b	73.7	68.6
1478	StrategyQA	Commonsense	Llama 3.1 70b	85.3	83.8
1479	StrategyQA	Commonsense	Gemma 2 9b	73.7	66.4
1480	StrategyQA	Commonsense	Phi-3 Small 8k	72.3	66.0
1481	StrategyQA	Commonsense	Qwen 2 7b	63.2	54.8
1482	StrategyQA	Commonsense	Qwen 2 72b	81.7	76.9

Table 7: Zero-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	zero-shot CoT accuracy	zero-shot DA accuracy	
1458	StrategyQA	Commonsense	GPT-4o Mini	82.2	84.5
1459	StrategyQA	Commonsense	Gpt-4o	84.5	85.5
1460	StrategyQA	Commonsense	Claude-3 Haiku	73.4	65.0
1461	StrategyQA	Commonsense	Claude-3.5 Sonnet	80.1	76.3
1462	StrategyQA	Commonsense	Gemini 1.5 Flash	72.5	75.2
1463	StrategyQA	Commonsense	Gemini 1.5 Pro	74.0	71.4
1464	MuSR Object Placements	Soft Reasoning	Llama 2 7b	36.3	30.5
1465	MuSR Object Placements	Soft Reasoning	Mistral 7b	50.8	43.4
1466	MuSR Object Placements	Soft Reasoning	Llama 3.1 8b	55.5	53.5
1467	MuSR Object Placements	Soft Reasoning	Llama 3.1 70b	65.6	43.8
1467	MuSR Object Placements	Soft Reasoning	Gemma 2 9b	63.3	57.0
1468	MuSR Object Placements	Soft Reasoning	Phi-3 Small 8k	53.1	55.1
1468	MuSR Object Placements	Soft Reasoning	Qwen 2 7b	48.8	48.4
1469	MuSR Object Placements	Soft Reasoning	Qwen 2 72b	61.7	45.7
1470	MuSR Object Placements	Soft Reasoning	GPT-4o Mini	59.0	55.0
1471	MuSR Object Placements	Soft Reasoning	Gpt-4o	67.6	45.3
1471	MuSR Object Placements	Soft Reasoning	Claude-3 Haiku	46.9	52.3
1472	MuSR Object Placements	Soft Reasoning	Claude-3.5 Sonnet	69.5	51.2
1472	MuSR Object Placements	Soft Reasoning	Gemini 1.5 Flash	61.7	56.2
1473	MuSR Object Placements	Soft Reasoning	Gemini 1.5 Pro	66.4	50.0
1474	FOLIO	Symbolic	Llama 2 7b	36.5	33.0
1475	FOLIO	Symbolic	Mistral 7b	50.7	41.9
1475	FOLIO	Symbolic	Llama 3.1 8b	58.6	56.7
1476	FOLIO	Symbolic	Llama 3.1 70b	70.9	69.0
1476	FOLIO	Symbolic	Gemma 2 9b	66.0	55.7
1477	FOLIO	Symbolic	Phi-3 Small 8k	68.0	59.6
1478	FOLIO	Symbolic	Qwen 2 7b	60.6	51.2
1478	FOLIO	Symbolic	Qwen 2 72b	65.0	65.0
1479	FOLIO	Symbolic	GPT-4o Mini	65.0	58.1
1480	FOLIO	Symbolic	Gpt-4o	79.8	62.6
1481	FOLIO	Symbolic	Claude-3 Haiku	61.6	48.8
1481	FOLIO	Symbolic	Claude-3.5 Sonnet	73.9	68.5
1482	FOLIO	Symbolic	Gemini 1.5 Flash	74.9	69.5
1482	FOLIO	Symbolic	Gemini 1.5 Pro	73.9	74.4
1483	ContextHub Deductive L2	Symbolic	Llama 2 7b	34.8	12.6
1484	ContextHub Deductive L2	Symbolic	Mistral 7b	48.8	55.1
1485	ContextHub Deductive L2	Symbolic	Llama 3.1 8b	52.8	21.5
1485	ContextHub Deductive L2	Symbolic	Llama 3.1 70b	50.0	41.1
1486	ContextHub Deductive L2	Symbolic	Gemma 2 9b	50.0	43.0
1487	ContextHub Deductive L2	Symbolic	Phi-3 Small 8k	52.4	49.1
1487	ContextHub Deductive L2	Symbolic	Qwen 2 7b	51.3	39.8
1488	ContextHub Deductive L2	Symbolic	Qwen 2 72b	52.8	44.0
1489	ContextHub Deductive L2	Symbolic	GPT-4o Mini	47.0	42.0
1489	ContextHub Deductive L2	Symbolic	Gpt-4o	54.5	45.6
1490	ContextHub Deductive L2	Symbolic	Claude-3 Haiku	45.2	41.8
1491	ContextHub Deductive L2	Symbolic	Claude-3.5 Sonnet	53.0	46.2
1491	ContextHub Deductive L2	Symbolic	Gemini 1.5 Flash	45.0	39.5
1492	ContextHub Deductive L2	Symbolic	Gemini 1.5 Pro	57.3	43.3
1493	ContextHub Abductive L2	Symbolic	Llama 2 7b	34.3	31.9
1493	ContextHub Abductive L2	Symbolic	Mistral 7b	34.0	25.7
1494	ContextHub Abductive L2	Symbolic	Llama 3.1 8b	41.3	37.3
1495	ContextHub Abductive L2	Symbolic	Llama 3.1 70b	51.0	44.4
1495	ContextHub Abductive L2	Symbolic	Gemma 2 9b	41.5	32.9
1496	ContextHub Abductive L2	Symbolic	Phi-3 Small 8k	44.3	32.8
1497	ContextHub Abductive L2	Symbolic	Qwen 2 7b	37.8	33.4
1497	ContextHub Abductive L2	Symbolic	Qwen 2 72b	45.5	32.2
1498	ContextHub Abductive L2	Symbolic	GPT-4o Mini	65.0	55.0
1499	ContextHub Abductive L2	Symbolic	Gpt-4o	57.5	46.8
1499	ContextHub Abductive L2	Symbolic	Claude-3 Haiku	37.0	31.4
1500	ContextHub Abductive L2	Symbolic	Claude-3.5 Sonnet	56.8	40.4
1501	ContextHub Abductive L2	Symbolic	Gemini 1.5 Flash	53.1	32.2
1501	ContextHub Abductive L2	Symbolic	Gemini 1.5 Pro	53.5	43.7
1502	MMLU Pro	Knowledge	Llama 2 7b	19.9	19.6
1503	MMLU Pro	Knowledge	Mistral 7b	31.6	28.4
1504	MMLU Pro	Knowledge	Llama 3.1 8b	44.8	38.0
1504	MMLU Pro	Knowledge	Llama 3.1 70b	64.9	55.0
1505	MMLU Pro	Knowledge	Gemma 2 9b	48.1	42.7
1505	MMLU Pro	Knowledge	Phi-3 Small 8k	54.8	43.7
1506	MMLU Pro	Knowledge	Qwen 2 7b	45.0	36.2
1507	MMLU Pro	Knowledge	Qwen 2 72b	62.8	44.3
1507	MMLU Pro	Knowledge	GPT-4o Mini	62.3	42.6
1508	MMLU Pro	Knowledge	Gpt-4o	72.1	55.0
1509	MMLU Pro	Knowledge	Claude-3 Haiku	47.6	39.0
1509	MMLU Pro	Knowledge	Claude-3.5 Sonnet	73.4	57.2
1510	MMLU Pro	Knowledge	Gemini 1.5 Flash	58.5	47.2
1511	MMLU Pro	Knowledge	Gemini 1.5 Pro	65.3	57.4
1511	MuSR Murder Mysteries	Soft Reasoning	Llama 2 7b	50.0	50.0

Table 7: Zero-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	zero-shot CoT accuracy	zero-shot DA accuracy	
1512					
1513					
1514	MuSR Murder Mysteries	Soft Reasoning	Mistral 7b	62.8	55.6
1515	MuSR Murder Mysteries	Soft Reasoning	Llama 3.1 8b	70.4	57.2
1516	MuSR Murder Mysteries	Soft Reasoning	Llama 3.1 70b	73.6	69.6
1517	MuSR Murder Mysteries	Soft Reasoning	Gemma 2 9b	76.8	61.6
1518	MuSR Murder Mysteries	Soft Reasoning	Phi-3 Small 8k	61.6	58.8
1519	MuSR Murder Mysteries	Soft Reasoning	Qwen 2 7b	59.2	53.2
1520	MuSR Murder Mysteries	Soft Reasoning	Qwen 2 72b	80.8	64.4
1521	MuSR Murder Mysteries	Soft Reasoning	GPT-4o Mini	71.2	63.6
1522	MuSR Murder Mysteries	Soft Reasoning	Gpt-4o	87.6	70.8
1523	MuSR Murder Mysteries	Soft Reasoning	Claude-3 Haiku	62.4	56.8
1524	MuSR Murder Mysteries	Soft Reasoning	Claude-3.5 Sonnet	85.2	70.4
1525	MuSR Murder Mysteries	Soft Reasoning	Gemini 1.5 Flash	70.8	58.4
1526	MuSR Murder Mysteries	Soft Reasoning	Gemini 1.5 Pro	77.6	64.0
1527	ContextHub Deductive L1	Symbolic	Llama 2 7b	47.7	8.3
1528	ContextHub Deductive L1	Symbolic	Mistral 7b	50.3	67.3
1529	ContextHub Deductive L1	Symbolic	Llama 3.1 8b	50.7	23.3
1530	ContextHub Deductive L1	Symbolic	Llama 3.1 70b	53.8	40.7
1531	ContextHub Deductive L1	Symbolic	Gemma 2 9b	56.3	39.2
1532	ContextHub Deductive L1	Symbolic	Phi-3 Small 8k	54.8	50.2
1533	ContextHub Deductive L1	Symbolic	Qwen 2 7b	59.3	43.3
1534	ContextHub Deductive L1	Symbolic	Qwen 2 72b	51.5	44.0
1535	ContextHub Deductive L1	Symbolic	GPT-4o Mini	49.3	41.5
1536	ContextHub Deductive L1	Symbolic	Gpt-4o	59.3	49.0
1537	ContextHub Deductive L1	Symbolic	Claude-3 Haiku	50.5	39.7
1538	ContextHub Deductive L1	Symbolic	Claude-3.5 Sonnet	54.5	47.0
1539	ContextHub Deductive L1	Symbolic	Gemini 1.5 Flash	47.3	38.5
1540	ContextHub Deductive L1	Symbolic	Gemini 1.5 Pro	57.3	46.0
1541	ContextHub Abductive L1	Symbolic	Llama 2 7b	29.4	16.4
1542	ContextHub Abductive L1	Symbolic	Mistral 7b	46.9	25.8
1543	ContextHub Abductive L1	Symbolic	Llama 3.1 8b	43.6	24.2
1544	ContextHub Abductive L1	Symbolic	Llama 3.1 70b	55.3	43.9
1545	ContextHub Abductive L1	Symbolic	Gemma 2 9b	61.9	58.9
1546	ContextHub Abductive L1	Symbolic	Phi-3 Small 8k	62.5	60.3
1547	ContextHub Abductive L1	Symbolic	Qwen 2 7b	52.2	47.5
1548	ContextHub Abductive L1	Symbolic	Qwen 2 72b	61.9	45.0
1549	ContextHub Abductive L1	Symbolic	GPT-4o Mini	61.1	42.2
1550	ContextHub Abductive L1	Symbolic	Gpt-4o	74.2	65.6
1551	ContextHub Abductive L1	Symbolic	Claude-3 Haiku	35.3	22.8
1552	ContextHub Abductive L1	Symbolic	Claude-3.5 Sonnet	80.8	60.3
1553	ContextHub Abductive L1	Symbolic	Gemini 1.5 Flash	66.4	47.2
1554	ContextHub Abductive L1	Symbolic	Gemini 1.5 Pro	62.2	60.0
1555	Big-Bench Hard	Symbolic	Llama 2 7b	29.8	31.9
1556	Big-Bench Hard	Symbolic	Mistral 7b	39.3	35.1
1557	Big-Bench Hard	Symbolic	Llama 3.1 8b	62.8	45.6
1558	Big-Bench Hard	Symbolic	Llama 3.1 70b	78.9	54.8
1559	Big-Bench Hard	Symbolic	Gemma 2 9b	58.7	50.8
1560	Big-Bench Hard	Symbolic	Phi-3 Small 8k	70.0	55.1
1561	Big-Bench Hard	Symbolic	Qwen 2 7b	52.6	47.6
1562	Big-Bench Hard	Symbolic	Qwen 2 72b	75.1	59.0
1563	Big-Bench Hard	Symbolic	GPT-4o Mini	77.7	49.7
1564	Big-Bench Hard	Symbolic	Gpt-4o	84.6	64.5
1565	Big-Bench Hard	Symbolic	Claude-3 Haiku	62.4	47.3
1566	Big-Bench Hard	Symbolic	Claude-3.5 Sonnet	83.6	56.9
1567	Big-Bench Hard	Symbolic	Gemini 1.5 Flash	71.3	55.4
1568	Big-Bench Hard	Symbolic	Gemini 1.5 Pro	71.6	50.3
1569	MATH	Mathematical	Llama 2 7b	4.2	4.0
1570	MATH	Mathematical	Mistral 7b	12.4	6.1
1571	MATH	Mathematical	Llama 3.1 8b	47.2	13.8
1572	MATH	Mathematical	Llama 3.1 70b	64.4	22.8
1573	MATH	Mathematical	Gemma 2 9b	45.6	19.1
1574	MATH	Mathematical	Phi-3 Small 8k	43.2	18.5
1575	MATH	Mathematical	Qwen 2 7b	53.7	13.3
1576	MATH	Mathematical	Qwen 2 72b	63.5	23.8
1577	MATH	Mathematical	GPT-4o Mini	69.6	24.3
1578	MATH	Mathematical	Gpt-4o	73.3	35.2
1579	MATH	Mathematical	Claude-3 Haiku	32.7	17.4
1580	MATH	Mathematical	Claude-3.5 Sonnet	63.8	34.6
1581	MATH	Mathematical	Gemini 1.5 Flash	54.5	31.3
1582	MATH	Mathematical	Gemini 1.5 Pro	62.1	39.4
1583	GSM8k-Hard	Mathematical	Llama 2 7b	6.7	1.8
1584	GSM8k-Hard	Mathematical	Mistral 7b	21.0	3.0
1585	GSM8k-Hard	Mathematical	Llama 3.1 8b	34.4	6.0
1586	GSM8k-Hard	Mathematical	Llama 3.1 70b	46.6	14.0
1587	GSM8k-Hard	Mathematical	Gemma 2 9b	40.9	8.8
1588	GSM8k-Hard	Mathematical	Phi-3 Small 8k	33.0	6.9
1589	GSM8k-Hard	Mathematical	Qwen 2 7b	48.4	5.0
1590	GSM8k-Hard	Mathematical	Qwen 2 72b	54.8	13.7

Table 7: Zero-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	zero-shot CoT accuracy	zero-shot DA accuracy
GSM8k-Hard	Mathematical	GPT-4o Mini	53.9	11.7
GSM8k-Hard	Mathematical	Gpt-4o	60.3	26.0
GSM8k-Hard	Mathematical	Claude-3 Haiku	45.3	9.6
GSM8k-Hard	Mathematical	Claude-3.5 Sonnet	50.8	32.3
GSM8k-Hard	Mathematical	Gemini 1.5 Flash	54.6	16.2
GSM8k-Hard	Mathematical	Gemini 1.5 Pro	58.2	26.2
GSM8k	Mathematical	Llama 2 7b	29.6	6.9
GSM8k	Mathematical	Mistral 7b	59.2	10.2
GSM8k	Mathematical	Llama 3.1 8b	85.4	18.5
GSM8k	Mathematical	Llama 3.1 70b	85.6	37.0
GSM8k	Mathematical	Gemma 2 9b	89.2	24.9
GSM8k	Mathematical	Phi-3 Small 8k	90.0	24.9
GSM8k	Mathematical	Qwen 2 7b	87.9	20.7
GSM8k	Mathematical	Qwen 2 72b	94.6	40.1
GSM8k	Mathematical	GPT-4o Mini	94.1	31.8
GSM8k	Mathematical	Gpt-4o	95.8	58.8
GSM8k	Mathematical	Claude-3 Haiku	89.4	22.9
GSM8k	Mathematical	Claude-3.5 Sonnet	96.1	62.2
GSM8k	Mathematical	Gemini 1.5 Flash	91.4	38.6
GSM8k	Mathematical	Gemini 1.5 Pro	92.7	52.4

F.2 FULL FEW-SHOT RESULTS

Table 8: Few-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	few-shot CoT accuracy	few-shot DA accuracy
AGIEval LSAT RC	Soft Reasoning	Llama 2 7b	33.1	38.7
AGIEval LSAT RC	Soft Reasoning	Mistral 7b	52.4	57.2
AGIEval LSAT RC	Soft Reasoning	Llama 3.1 8b	60.2	70.3
AGIEval LSAT RC	Soft Reasoning	Llama 3.1 70b	84.4	88.8
AGIEval LSAT RC	Soft Reasoning	Gemma 2 9b	74.3	79.2
AGIEval LSAT RC	Soft Reasoning	Phi-3 Small 8k	63.2	65.1
AGIEval LSAT RC	Soft Reasoning	Qwen 2 7b	61.7	68.8
AGIEval LSAT RC	Soft Reasoning	Qwen 2 72b	85.9	85.9
AGIEval LSAT RC	Soft Reasoning	GPT-4o Mini	77.3	71.4
AGIEval LSAT RC	Soft Reasoning	Gemini 1.5 Flash	79.2	81.8
AGIEval LSAT LR	Soft Reasoning	Llama 2 7b	33.7	34.7
AGIEval LSAT LR	Soft Reasoning	Mistral 7b	46.1	48.0
AGIEval LSAT LR	Soft Reasoning	Llama 3.1 8b	55.7	58.0
AGIEval LSAT LR	Soft Reasoning	Llama 3.1 70b	83.3	85.1
AGIEval LSAT LR	Soft Reasoning	Gemma 2 9b	65.7	68.2
AGIEval LSAT LR	Soft Reasoning	Phi-3 Small 8k	64.7	59.2
AGIEval LSAT LR	Soft Reasoning	Qwen 2 7b	54.1	61.2
AGIEval LSAT LR	Soft Reasoning	Qwen 2 72b	77.5	79.6
AGIEval LSAT LR	Soft Reasoning	GPT-4o Mini	68.4	64.5
AGIEval LSAT LR	Soft Reasoning	Gemini 1.5 Flash	68.6	72.9
GPQA	Mathematical	Mistral 7b	23.0	25.9
GPQA	Mathematical	Llama 3.1 8b	22.1	27.2
GPQA	Mathematical	Llama 3.1 70b	24.8	24.3
GPQA	Mathematical	Gemma 2 9b	19.9	22.3
GPQA	Mathematical	Phi-3 Small 8k	23.9	22.5
GPQA	Mathematical	Qwen 2 7b	23.4	21.2
GPQA	Mathematical	Qwen 2 72b	22.8	19.9
GPQA	Mathematical	GPT-4o Mini	20.0	20.0
GPQA	Mathematical	Gemini 1.5 Flash	21.9	24.6
CommonsenseQA	Commonsense	Llama 2 7b	18.2	19.2
CommonsenseQA	Commonsense	Mistral 7b	73.6	70.4
CommonsenseQA	Commonsense	Llama 3.1 8b	74.0	76.5
CommonsenseQA	Commonsense	Llama 3.1 70b	84.7	84.6
CommonsenseQA	Commonsense	Gemma 2 9b	81.8	80.8
CommonsenseQA	Commonsense	Phi-3 Small 8k	80.8	80.4
CommonsenseQA	Commonsense	Qwen 2 7b	80.3	72.9
CommonsenseQA	Commonsense	Qwen 2 72b	88.4	87.8
CommonsenseQA	Commonsense	GPT-4o Mini	84.7	84.7
CommonsenseQA	Commonsense	Gemini 1.5 Flash	81.7	83.3
AGIEval LSAT AR	Soft Reasoning	Llama 2 7b	19.6	18.7
AGIEval LSAT AR	Soft Reasoning	Mistral 7b	20.9	22.6
AGIEval LSAT AR	Soft Reasoning	Llama 3.1 8b	24.8	26.1
AGIEval LSAT AR	Soft Reasoning	Llama 3.1 70b	36.1	30.9
AGIEval LSAT AR	Soft Reasoning	Gemma 2 9b	22.2	28.7
AGIEval LSAT AR	Soft Reasoning	Phi-3 Small 8k	27.8	20.0
AGIEval LSAT AR	Soft Reasoning	Qwen 2 7b	24.3	23.0
AGIEval LSAT AR	Soft Reasoning	Qwen 2 72b	27.0	30.0

Table 8: Few-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	few-shot CoT accuracy	few-shot DA accuracy
1620	AGIEval LSAT AR	Soft Reasoning GPT-4o Mini	28.7	26.1
1621	AGIEval LSAT AR	Soft Reasoning Gemini 1.5 Flash	28.3	20.4
1622	MMLU	Knowledge Llama 2 7b	49.0	42.8
1623	MMLU	Knowledge Mistral 7b	63.0	57.0
1624	MMLU	Knowledge Llama 3.1 8b	71.7	69.3
1625	MMLU	Knowledge Llama 3.1 70b	84.3	83.7
1626	MMLU	Knowledge Gemma 2 9b	74.7	72.4
1627	MMLU	Knowledge Phi-3 Small 8k	77.3	75.2
1628	MMLU	Knowledge Qwen 2 7b	69.9	68.6
1629	MMLU	Knowledge Qwen 2 72b	82.7	81.8
1630	MMLU	Knowledge GPT-4o Mini	82.3	77.8
1631	MMLU	Knowledge Gemini 1.5 Flash	78.1	79.0
1632	StrategyQA	Commonsense Llama 2 7b	57.9	30.9
1633	StrategyQA	Commonsense Mistral 7b	70.7	72.0
1634	StrategyQA	Commonsense Llama 3.1 8b	74.4	65.8
1635	StrategyQA	Commonsense Llama 3.1 70b	87.1	84.2
1636	StrategyQA	Commonsense Gemma 2 9b	77.1	73.3
1637	StrategyQA	Commonsense Phi-3 Small 8k	75.0	71.1
1638	StrategyQA	Commonsense Qwen 2 7b	71.9	58.9
1639	StrategyQA	Commonsense Qwen 2 72b	83.2	80.1
1640	StrategyQA	Commonsense GPT-4o Mini	83.0	86.2
1641	StrategyQA	Commonsense Gemini 1.5 Flash	77.0	80.3
1642	ContextHub Abductive L2	Symbolic Llama 2 7b	36.2	35.0
1643	ContextHub Abductive L2	Symbolic Mistral 7b	33.8	30.0
1644	ContextHub Abductive L2	Symbolic Llama 3.1 8b	32.7	36.1
1645	ContextHub Abductive L2	Symbolic Llama 3.1 70b	54.6	51.2
1646	ContextHub Abductive L2	Symbolic Gemma 2 9b	44.8	33.2
1647	ContextHub Abductive L2	Symbolic Phi-3 Small 8k	49.8	34.2
1648	ContextHub Abductive L2	Symbolic Qwen 2 7b	39.6	35.0
1649	ContextHub Abductive L2	Symbolic Qwen 2 72b	54.7	34.9
1650	ContextHub Abductive L2	Symbolic GPT-4o Mini	62.0	60.0
1651	ContextHub Abductive L2	Symbolic Gemini 1.5 Flash	48.6	47.8
1652	ContextHub Abductive L1	Symbolic Llama 2 7b	21.4	16.7
1653	ContextHub Abductive L1	Symbolic Mistral 7b	23.6	21.7
1654	ContextHub Abductive L1	Symbolic Llama 3.1 8b	40.0	36.1
1655	ContextHub Abductive L1	Symbolic Llama 3.1 70b	62.2	58.9
1656	ContextHub Abductive L1	Symbolic Gemma 2 9b	48.9	59.4
1657	ContextHub Abductive L1	Symbolic Phi-3 Small 8k	59.2	56.4
1658	ContextHub Abductive L1	Symbolic Qwen 2 7b	48.6	38.9
1659	ContextHub Abductive L1	Symbolic Qwen 2 72b	53.3	56.1
1660	ContextHub Abductive L1	Symbolic GPT-4o Mini	77.2	59.2
1661	ContextHub Abductive L1	Symbolic Gemini 1.5 Flash	79.7	68.6
1662	MuSR Murder Mysteries	Soft Reasoning Mistral 7b	62.0	56.4
1663	MuSR Murder Mysteries	Soft Reasoning Llama 3.1 8b	61.6	61.2
1664	MuSR Murder Mysteries	Soft Reasoning Llama 3.1 70b	73.2	68.0
1665	MuSR Murder Mysteries	Soft Reasoning Gemma 2 9b	81.6	62.0
1666	MuSR Murder Mysteries	Soft Reasoning Phi-3 Small 8k	62.0	53.6
1667	MuSR Murder Mysteries	Soft Reasoning Qwen 2 7b	56.0	55.6
1668	MuSR Murder Mysteries	Soft Reasoning Qwen 2 72b	80.4	66.0
1669	MuSR Murder Mysteries	Soft Reasoning GPT-4o Mini	76.0	69.6
1670	MuSR Murder Mysteries	Soft Reasoning Gemini 1.5 Flash	70.0	66.4
1671	MuSR Team Allocations	Soft Reasoning Mistral 7b	42.8	43.2
1672	MuSR Team Allocations	Soft Reasoning Llama 3.1 8b	59.6	51.6
1673	MuSR Team Allocations	Soft Reasoning Llama 3.1 70b	89.2	63.6
1674	MuSR Team Allocations	Soft Reasoning Gemma 2 9b	48.4	45.6
1675	MuSR Team Allocations	Soft Reasoning Phi-3 Small 8k	66.0	46.4
1676	MuSR Team Allocations	Soft Reasoning Qwen 2 7b	34.0	40.8
1677	MuSR Team Allocations	Soft Reasoning Qwen 2 72b	56.0	66.4
1678	MuSR Team Allocations	Soft Reasoning GPT-4o Mini	75.6	60.0
1679	MuSR Team Allocations	Soft Reasoning Gemini 1.5 Flash	90.0	54.4
1680	MMLU Pro	Knowledge Llama 2 7b	21.5	20.4
1681	MMLU Pro	Knowledge Mistral 7b	34.8	26.7
1682	MMLU Pro	Knowledge Llama 3.1 8b	44.7	38.0
1683	MMLU Pro	Knowledge Llama 3.1 70b	64.4	55.1
1684	MMLU Pro	Knowledge Gemma 2 9b	48.5	42.4
1685	MMLU Pro	Knowledge Phi-3 Small 8k	54.8	43.2
1686	MMLU Pro	Knowledge Qwen 2 7b	46.6	39.0
1687	MMLU Pro	Knowledge Qwen 2 72b	62.5	51.6
1688	MMLU Pro	Knowledge GPT-4o Mini	63.0	45.0
1689	MMLU Pro	Knowledge Gemini 1.5 Flash	59.4	50.6
1690	MuSR Object Placements	Soft Reasoning Mistral 7b	55.5	41.0
1691	MuSR Object Placements	Soft Reasoning Llama 3.1 8b	66.8	50.4
1692	MuSR Object Placements	Soft Reasoning Llama 3.1 70b	67.2	57.4
1693	MuSR Object Placements	Soft Reasoning Gemma 2 9b	68.0	58.2
1694	MuSR Object Placements	Soft Reasoning Phi-3 Small 8k	62.1	51.6
1695	MuSR Object Placements	Soft Reasoning Qwen 2 7b	46.9	43.8
1696	MuSR Object Placements	Soft Reasoning Qwen 2 72b	66.4	43.0

Table 8: Few-shot accuracy for direct answering and CoT prompts on all datasets

Dataset	Type	Model	few-shot CoT accuracy	few-shot DA accuracy
MuSR Object Placements	Soft Reasoning	GPT-4o Mini	67.0	47.0
MuSR Object Placements	Soft Reasoning	Gemini 1.5 Flash	73.0	54.7
ContextHub Deductive L2	Symbolic	Llama 2 7b	34.7	15.0
ContextHub Deductive L2	Symbolic	Mistral 7b	63.8	51.4
ContextHub Deductive L2	Symbolic	Llama 3.1 8b	76.1	27.3
ContextHub Deductive L2	Symbolic	Llama 3.1 70b	82.6	53.6
ContextHub Deductive L2	Symbolic	Gemma 2 9b	61.9	47.6
ContextHub Deductive L2	Symbolic	Phi-3 Small 8k	61.5	54.0
ContextHub Deductive L2	Symbolic	Qwen 2 7b	55.3	36.4
ContextHub Deductive L2	Symbolic	Qwen 2 72b	80.2	54.0
ContextHub Deductive L2	Symbolic	GPT-4o Mini	59.0	41.0
ContextHub Deductive L2	Symbolic	Gemini 1.5 Flash	90.2	42.5
ContextHub Deductive L1	Symbolic	Llama 2 7b	34.7	16.0
ContextHub Deductive L1	Symbolic	Mistral 7b	46.2	59.2
ContextHub Deductive L1	Symbolic	Llama 3.1 8b	73.0	23.0
ContextHub Deductive L1	Symbolic	Llama 3.1 70b	67.5	50.0
ContextHub Deductive L1	Symbolic	Gemma 2 9b	66.0	45.7
ContextHub Deductive L1	Symbolic	Phi-3 Small 8k	74.8	51.8
ContextHub Deductive L1	Symbolic	Qwen 2 7b	58.8	37.5
ContextHub Deductive L1	Symbolic	Qwen 2 72b	70.7	42.8
ContextHub Deductive L1	Symbolic	GPT-4o Mini	59.2	44.3
ContextHub Deductive L1	Symbolic	Gemini 1.5 Flash	89.3	49.8
MATH	Mathematical	Llama 2 7b	4.7	3.9
MATH	Mathematical	Mistral 7b	13.7	7.1
MATH	Mathematical	Llama 3.1 8b	41.2	14.2
MATH	Mathematical	Llama 3.1 70b	61.9	24.2
MATH	Mathematical	Gemma 2 9b	47.5	19.8
MATH	Mathematical	Phi-3 Small 8k	42.4	18.9
MATH	Mathematical	Qwen 2 7b	55.0	15.0
MATH	Mathematical	Qwen 2 72b	65.3	26.2
MATH	Mathematical	GPT-4o Mini	71.7	24.6
MATH	Mathematical	Gemini 1.5 Flash	54.7	32.3
GSM8K	Mathematical	Llama 2 7b	29.0	7.7
GSM8K	Mathematical	Mistral 7b	56.2	12.5
GSM8K	Mathematical	Llama 3.1 8b	86.4	20.1
GSM8K	Mathematical	Llama 3.1 70b	96.1	39.1
GSM8K	Mathematical	Gemma 2 9b	89.2	24.9
GSM8K	Mathematical	Phi-3 Small 8k	90.4	24.5
GSM8K	Mathematical	Qwen 2 7b	87.6	21.4
GSM8K	Mathematical	Qwen 2 72b	93.2	40.6
GSM8K	Mathematical	GPT-4o Mini	94.2	32.8
GSM8K	Mathematical	Gemini 1.5 Flash	90.6	40.4

F.3 ANSWER EXTRACTOR AND AVERAGE ANSWER SPAN RESULTS

In this section, we report the number of generations from each model on each dataset that our answer parser could not extract. “-1” denotes that a model was not run on a certain dataset due to context length limitations in the few-shot setting. We see that these unparseable rates are generally low across the board. The weakest models struggle on some of the most challenging datasets, but unparseable rates are all at or below 15%.

We also report the average character index of the beginning of the answer span that the answer parser extracted. Of particular note is that the direct answer prompts all return an answer within the first 60 characters, indicating that the answers are returned almost immediately, as desired. CoT completions are much longer.

G ZOOM-IN: MMLU AND MMLU PRO

MMLU and MMLU Pro show gains from adding CoT, but because these datasets are so broad, they defy simple characterization. We explore the performance of CoT on each category of MMLU to understand divergences in CoT performance between these domains. We list the top three categories where CoT gives the largest error reduction for Llama 3.1 8B and 70B on MMLU and MMLU Pro in Table 17. Some of these categories are explicitly mathematical in nature, as we might expect from Figure 3. We can also see that CoT is helping on categories like “business”; upon closer inspection, we found that these categories frequently involve math as well (e.g., business questions may involve computations surrounding wealth). We need to more carefully characterize MMLU at the *instance*

1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

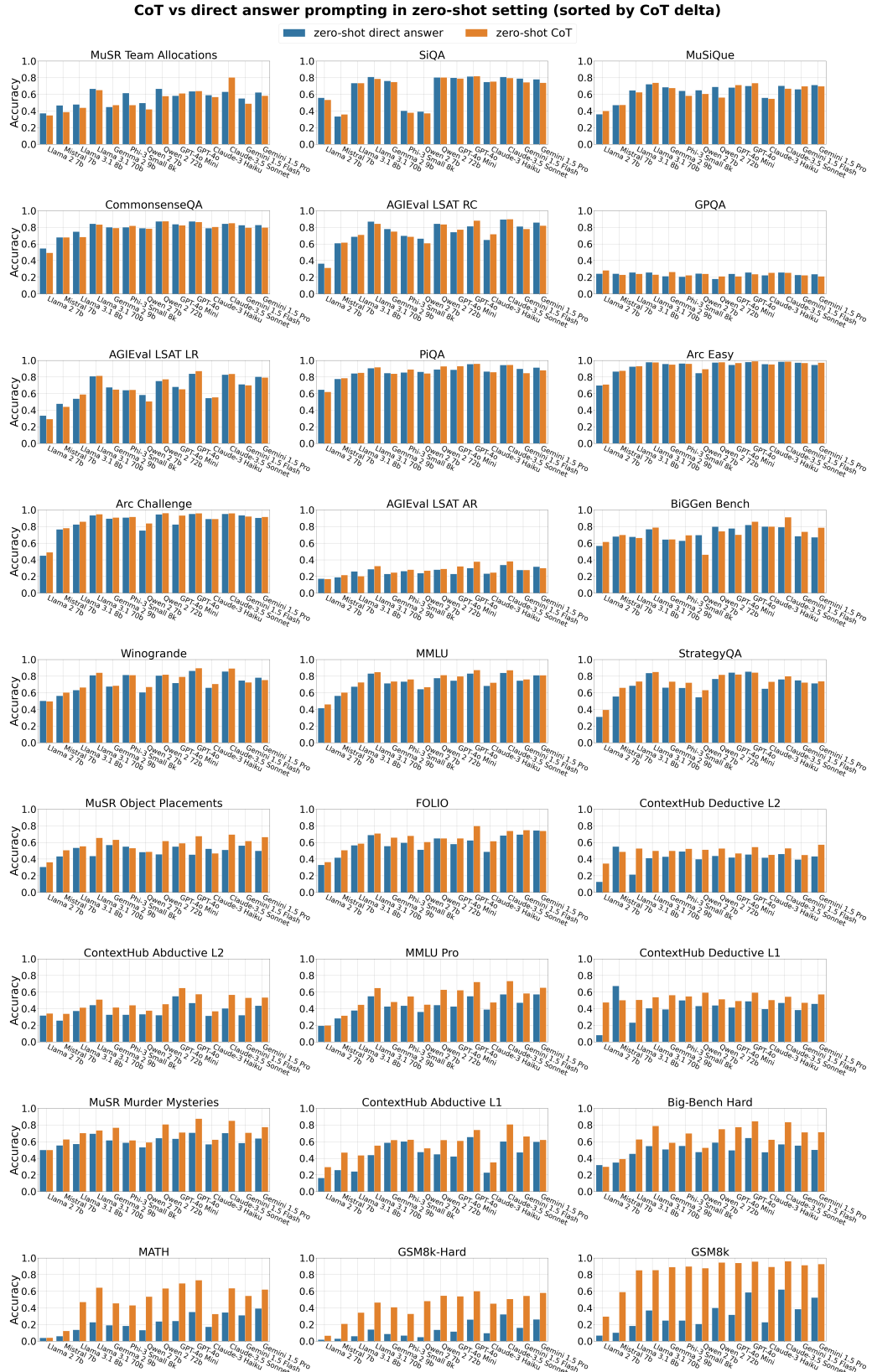


Figure 9: Performance of zero-shot direct (blue) and zero-shot CoT (orange) across datasets and models. Graphs are sorted in ascending order by median delta (CoT, direct). The datasets benefiting substantially are all symbolic or semi-symbolic in nature.

Table 9: Percentage of responses per dataset per model that our answer parser could not extract an answer for in the zero-shot direct answer setting. Prompt modifications were made to decrease these numbers. No model is above 15%.

Zero-shot Direct Answer Unparsable Answer Rate by Percentage														
dataset	Meta-Llama 2 7b	Mistral 7b	Meta-Llama 3.1 8b	Meta-Llama 3.1 70b	Gemma 2 9b	Phi-3 Small 8k	Qwen 2 7b	Qwen 2 72b	GPT-4o Mini	Gpt-4o	Claude-3 Haiku	Claude-3.5 Sonnet	Gemini 1.5 Flash	Gemini 1.5 Pro
CommonsenseQA	1.9	2.5	1.1	0.0	0.8	0.1	1.6	0.7	0.0	0.0	0.1	0.0	0.1	0.2
StrategyQA	0.0	1.9	0.1	0.0	11.7	0.5	4.9	2.7	0.0	0.0	0.0	0.0	0.0	0.2
SiQA	0.2	6.6	0.0	0.1	3.9	0.3	0.1	3.0	0.1	0.1	0.0	0.0	0.0	0.4
PiQA	0.4	6.0	0.0	0.1	3.3	2.1	0.0	5.5	0.2	0.0	0.1	0.0	0.1	0.9
Winogrande	0.0	3.0	0.1	0.0	2.1	0.2	5.1	0.4	0.0	0.0	0.0	0.0	0.0	3.6
Arc Easy	0.0	1.8	0.5	0.0	0.0	0.2	9.1	0.7	3.5	0.4	0.2	0.0	0.0	3.2
Arc Challenge	0.0	2.3	1.0	0.0	0.3	0.7	10.7	0.7	10.0	0.7	0.0	0.0	0.0	5.0
AGEval LSAT LR	0.4	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	2.5	0.0	0.0	0.2	0.2
AGEval LSAT AR	0.4	0.0	0.0	0.0	4.3	3.9	0.0	0.0	0.0	8.7	0.0	0.0	0.0	0.0
AGEval LSAT RC	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.7	0.0	0.0	0.4	0.4
ContextHub Deductive L1	0.0	0.0	0.0	0.0	1.2	0.0	2.3	0.0	0.0	0.0	0.2	0.0	0.0	0.2
ContextHub Deductive L2	0.0	0.0	0.0	0.0	0.0	0.0	2.2	1.0	0.0	0.0	2.8	0.0	0.0	0.0
ContextHub Abductive L1	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ContextHub Abductive L2	0.0	0.0	0.0	0.0	0.0	0.1	1.5	0.2	0.0	0.0	0.8	0.0	0.0	0.0
MuSR Murder Mysteries	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MuSR Team Allocations	0.0	0.0	0.0	0.0	3.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.4	0.4
MuSR Object Placements	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MMLU	0.1	0.0	0.0	0.0	0.1	0.2	3.6	1.2	0.6	0.0	1.3	0.3	0.2	0.7
MMLU Pro	0.7	1.3	1.0	0.3	1.0	3.7	6.8	12.2	0.4	0.3	0.4	0.6	0.8	0.8
GPQA	1.3	7.1	0.0	0.0	8.7	12.7	5.4	15.2	0.0	0.0	1.6	0.0	0.0	0.7
MATH	0.6	6.9	0.3	0.2	0.1	0.1	3.5	3.0	0.8	0.0	0.3	0.0	0.4	0.6
GSM8k	0.2	4.1	2.5	0.0	2.7	0.0	1.7	0.2	0.0	0.0	12.7	5.5	0.0	0.0
BigGen Bench	4.6	0.3	0.9	0.1	0.5	1.0	1.3	1.0	1.3	0.0	0.0	0.1	0.4	0.3
GSM8k-Hard	4.8	7.6	2.0	0.4	0.4	0.2	3.2	1.1	0.1	0.5	5.2	0.5	0.2	0.0
MuSiQue	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1
Folio	4.4	0.0	0.0	0.0	0.0	0.0	3.9	0.0	0.0	12.3	0.0	0.0	0.0	0.5
BigBench-Hard	0.0	0.0	0.0	7.4	0.0	0.2	0.0	0.0	0.0	0.3	0.0	4.5	0.2	12.8

Table 10: Percentage of responses per dataset per model that our answer parser could not extract an answer for in the zero-shot CoT setting. Prompt modifications were made to decrease these numbers. No model is above 15%.

Zero-shot CoT Unparsable Answer Rate by Percentage														
dataset	Meta-Llama 2 7b	Mistral 7b	Meta-Llama 3.1 8b	Meta-Llama 3.1 70b	Gemma 2 9b	Phi-3 Small 8k	Qwen 2 7b	Qwen 2 72b	GPT-4o Mini	Gpt-4o	Claude-3 Haiku	Claude-3.5 Sonnet	Gemini 1.5 Flash	Gemini 1.5 Pro
CommonsenseQA	2.9	1.3	8.6	0.0	0.6	0.1	0.0	0.0	1.6	0.0	0.2	0.3	2.4	2.6
StrategyQA	1.0	0.1	1.1	0.8	0.3	0.4	0.3	0.0	0.0	0.0	0.0	0.0	2.1	4.4
SiQA	0.8	1.8	0.3	0.1	1.6	0.0	0.1	0.1	0.0	0.0	0.3	0.1	3.5	4.0
PiQA	1.6	1.6	0.2	0.1	2.8	0.3	0.5	0.3	0.0	0.0	1.4	0.3	4.6	4.6
Winogrande	0.9	1.4	0.2	0.2	0.9	0.4	0.3	0.0	0.0	0.0	0.0	0.0	2.0	3.4
Arc Easy	0.2	0.4	0.2	0.0	0.5	1.6	1.6	0.0	0.5	0.0	0.0	0.0	0.4	0.5
Arc Challenge	0.0	0.7	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0	0.7	0.7
AGEval LSAT LR	3.3	2.2	0.0	0.0	1.2	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.8	0.2
AGEval LSAT AR	4.8	7.0	6.1	2.2	5.7	5.2	4.3	0.4	1.3	1.3	0.0	0.4	4.8	1.7
AGEval LSAT RC	7.1	1.1	0.0	0.0	0.7	3.0	6.7	0.0	0.0	0.0	0.0	0.0	0.0	0.4
ContextHub Deductive L1	0.7	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3
ContextHub Deductive L2	0.2	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.4
ContextHub Abductive L1	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ContextHub Abductive L2	0.0	0.2	0.1	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.4
MuSR Murder Mysteries	0.0	0.4	0.0	0.0	0.0	11.6	0.4	0.0	0.0	0.0	0.0	0.0	6.8	3.6
MuSR Team Allocations	5.2	3.2	0.8	0.0	0.8	0.4	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MuSR Object Placements	0.0	1.6	0.0	0.0	0.4	0.8	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.4
MMLU	1.9	0.6	1.0	0.2	1.5	1.0	0.4	0.2	0.0	0.1	0.0	0.1	3.1	3.2
MMLU Pro	4.4	5.4	13.1	3.3	12.5	3.6	5.4	2.0	2.4	1.9	0.4	0.4	5.0	4.4
GPQA	4.5	10.3	9.4	1.6	8.5	1.8	3.8	0.7	0.0	0.0	0.0	0.0	11.8	15.0
MATH	1.6	5.5	8.2	2.5	2.3	1.6	3.0	0.4	0.4	0.5	0.9	0.0	1.7	1.0
GSM8k	1.7	1.4	0.7	0.5	0.4	0.6	0.4	0.0	0.0	0.0	0.3	0.0	0.1	0.1
BigGen Bench	5.0	0.4	0.5	0.1	0.5	0.4	0.3	9.5	0.0	0.0	0.0	0.1	0.4	0.1
GSM8k-Hard	2.1	8.7	10.2	4.5	10.7	3.2	3.5	1.0	0.8	0.5	3.0	1.8	0.4	2.7
MuSiQue	1.4	0.0	8.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	3.1
Folio	0.0	0.0	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1.5
BigBench-Hard	3.8	5.4	1.8	0.4	1.3	0.1	0.4	0.3	0.0	0.0	0.0	0.0	1.2	0.9

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

Table 11: Percentage of responses per dataset per model that our answer parser could not extract an answer for in the few-shot direct answer setting. Prompt modifications were made to decrease these numbers. No model is above 15%.

Few-shot Direct Answer Unparseable Answer Rate by Percentage										
dataset	Meta-Llama 2 7b	Mistral 7b	Meta-Llama 3.1 8b	Meta-Llama 3.1 70b	Gemma 2 9b	Phi-3 Small 8k	Qwen 2 7b	Qwen 2 72b	GPT-4o Mini	Gemini 1.5 Flash
CommonsenseQA	0.0	0.1	0.2	0.0	1.3	0.9	9.9	1.3	0.0	0.6
AGIEval LSAT LR	6.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2
AGIEval LSAT AR	2.6	0.0	0.0	0.0	3.5	5.2	0.0	0.0	0.0	0.0
AGIEval LSAT RC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ContextHub Deductive L1	0.0	2.8	0.0	0.0	0.0	10.7	0.3	0.0	0.0	0.0
ContextHub Deductive L2	0.0	0.1	0.0	0.0	0.0	0.3	0.2	0.0	0.0	0.0
ContextHub Abductive L1	0.0	2.8	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0
ContextHub Abductive L2	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MuSR Murder Mysteries	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4
MuSR Team Allocations	-1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MuSR Object Placements	-1.0	0.0	0.0	0.0	0.4	1.2	0.0	0.0	0.0	0.0
MMLU	4.2	0.2	0.0	0.0	0.1	0.0	0.4	0.1	0.0	0.2
MMLU Pro	5.1	1.2	2.4	0.3	1.0	9.1	0.5	2.6	0.4	0.5
GPQA	-1.0	1.3	0.0	0.0	3.6	7.4	13.4	1.1	0.0	0.0
MATH	0.3	5.9	0.3	0.2	0.1	0.1	1.6	2.2	0.0	0.3
GSM8k	0.1	0.1	0.5	0.0	0.1	2.2	0.0	0.2	0.0	0.0

Table 12: Percentage of responses per dataset per model that our answer parser could not extract an answer for in the few-shot CoT setting. Prompt modifications were made to decrease these numbers. No model is above 15%.

Few-shot CoT Unparseable Answer Rate by Percentage										
dataset	Meta-Llama 2 7b	Mistral 7b	Meta-Llama 3.1 8b	Meta-Llama 3.1 70b	Gemma 2 9b	Phi-3 Small 8k	Qwen 2 7b	Qwen 2 72b	GPT-4o Mini	Gemini 1.5 Flash
CommonsenseQA	0.7	0.9	1.8	0.1	0.2	0.1	0.0	0.0	0.0	3.4
AGIEval LSAT LR	0.6	0.8	0.4	0.0	1.4	3.1	0.8	0.0	0.0	0.6
AGIEval LSAT AR	2.2	9.1	3.9	0.9	11.7	3.0	3.5	1.7	0.0	1.3
AGIEval LSAT RC	7.8	5.9	0.0	0.0	1.9	9.3	2.6	0.0	0.0	2.2
ContextHub Deductive L1	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.3
ContextHub Deductive L2	0.9	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.3
ContextHub Abductive L1	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ContextHub Abductive L2	3.1	0.0	5.3	0.1	0.0	0.2	0.0	0.0	0.0	0.7
MuSR Murder Mysteries	-1.0	1.2	0.0	0.0	0.4	0.8	0.0	0.0	0.0	14.0
MuSR Team Allocations	-1.0	2.4	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.4
MuSR Object Placements	-1.0	0.4	0.0	0.0	1.2	0.4	0.0	0.0	0.0	0.0
MMLU	0.6	0.8	1.1	0.2	1.5	0.7	0.3	0.2	0.2	2.5
MMLU Pro	0.6	1.9	8.5	2.1	14.1	1.8	1.9	0.8	1.1	3.9
GPQA	-1.0	12.1	10.3	0.9	12.9	6.0	5.6	3.3	0.0	13.6
MATH	1.5	6.8	8.2	2.4	11.1	2.6	2.9	1.1	0.5	1.8
GSM8k	0.8	1.3	1.0	0.1	0.5	0.5	0.1	0.0	0.1	0.1

1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943

Table 13: Average character index of where the answer span begins in a generated response for each dataset and model pair for the zero-shot direct answer setting. We use these numbers as a proxy for the model following instructions (i.e. generating reasoning before an answer). Prompt modifications were made to ensure CoT prompts resulted in longer generations and direct answer prompts led to short generations.

Zero-shot Direct Answer Span Location By Character Index														
dataset	Meta-Llama 2.7b	Mistral 7b	Meta-Llama 3.1 8b	Meta-Llama 3.1 70b	Gemma 2 9b	Pko-3.5 Small 8k	Qwen 2.7b	Qwen 2.72b	GPT-4o Mini	Gpt-4o	Claude-3 Haiku	Claude-3.5 Sonnet	Gemini 1.5 Flash	Gemini 1.5 Pro
CommonsenseQA	9	8	27	8	8	8	10	8	8	10	7	7	8	8
StrategyQA	44	45	27	44	44	44	46	44	8	8	42	41	8	7
SiQA	8	8	8	8	8	8	29	8	8	8	6	6	8	8
PiQA	7	8	8	8	8	8	25	8	8	8	4	5	8	8
Winogrande	8	9	8	8	8	8	9	8	8	9	5	4	8	8
Arc Easy	9	8	8	8	8	8	9	8	8	8	7	7	8	8
Arc Challenge	8	8	8	8	8	8	9	8	8	8	7	7	8	8
AGIEval LSAT LR	25	24	24	24	24	24	25	24	43	21	25	25	26	26
AGIEval LSAT AR	25	24	24	24	24	24	26	24	48	23	25	25	26	27
AGIEval LSAT RC	25	24	24	24	24	24	25	24	31	18	25	25	26	25
ContextHub Deductive L1	19	19	19	19	20	19	19	19	19	19	20	20	19	20
ContextHub Deductive L2	19	19	19	19	19	19	19	19	19	19	20	20	19	19
ContextHub Abductive L1	19	19	19	19	20	19	19	19	19	19	20	20	19	19
ContextHub Abductive L2	19	19	19	19	20	19	19	19	19	19	20	20	19	19
MuSR Murder Mysteries	8	8	27	8	8	8	8	8	8	8	6	4	8	8
MuSR Team Allocations	27	22	19	19	27	23	26	22	8	8	30	20	8	8
MuSR Object Placements	8	8	27	8	8	8	8	8	8	8	7	6	8	8
MMLU	19	18	19	19	20	18	18	18	19	19	19	19	19	20
MMLU Pro	20	19	38	19	21	19	19	20	19	19	20	20	19	19
GPQA	19	19	19	19	21	19	19	19	19	19	20	20	19	20
MATH	30	31	28	28	28	30	30	33	28	28	31	29	28	28
GSM8k	22	29	30	28	28	37	24	28	28	28	29	28	28	28
GSM8k-Hard	9	57	11	9	13	40	20	7	8	8	8	8	8	8
Folio	39	8	8	8	8	8	31	13	8	16	5	6	8	70
BigBench-Hard	39	22	25	21	26	32	29	26	28	19	28	28	10	16

Table 14: Average character index of where the answer span begins in a generated response for each dataset and model pair for the zero-shot CoT setting. We use these numbers as a proxy for the model following instructions (i.e. generating reasoning before an answer). Prompt modifications were made to ensure CoT prompts resulted in longer generations and direct answer prompts led to short generations.

Zero-shot CoT Answer Span Location By Character Index														
dataset	Meta-Llama 2.7b	Mistral 7b	Meta-Llama 3.1 8b	Meta-Llama 3.1 70b	Gemma 2 9b	Pko-3.5 Small 8k	Qwen 2.7b	Qwen 2.72b	GPT-4o Mini	Gpt-4o	Claude-3 Haiku	Claude-3.5 Sonnet	Gemini 1.5 Flash	Gemini 1.5 Pro
CommonsenseQA	441	564	845	1237	236	466	577	341	899	1086	626	1103	214	165
StrategyQA	726	434	996	1131	267	460	363	358	692	1033	754	1158	256	195
SiQA	569	423	841	965	235	528	472	420	847	1094	602	1016	196	169
PiQA	699	455	869	914	207	532	447	364	683	935	578	1092	200	150
Winogrande	377	324	645	694	187	326	391	298	634	750	408	889	200	173
Arc Easy	684	581	1154	1319	367	610	534	355	990	1239	789	1222	340	231
Arc Challenge	763	644	1178	1316	422	596	571	387	1020	1269	828	1240	372	267
AGIEval LSAT LR	2053	1324	1163	1675	524	689	1560	768	949	998	1561	728	906	886
AGIEval LSAT AR	1377	1791	1422	2182	712	1027	1819	1264	1230	1151	1202	849	817	871
AGIEval LSAT RC	1977	1032	1103	1575	779	590	1170	660	973	1079	1628	786	703	709
ContextHub Deductive L1	694	368	759	711	383	327	539	402	540	580	542	556	320	254
ContextHub Deductive L2	842	472	1095	990	614	442	789	585	840	758	777	655	515	503
ContextHub Abductive L1	577	461	747	879	464	440	754	638	788	879	683	594	368	325
ContextHub Abductive L2	861	600	1270	1229	686	571	976	856	1115	1113	894	894	601	551
MuSR Murder Mysteries	495	1592	1958	1847	1210	1246	1241	1718	1961	1965	1671	1759	1349	1213
MuSR Team Allocations	1212	1845	2294	2310	1513	1433	2021	2213	2562	2698	1479	1856	1596	1607
MuSR Object Placements	917	625	1354	1266	695	641	904	819	1593	1536	1210	1455	616	429
MMLU	834	512	663	622	503	277	497	407	400	461	447	409	630	413
MMLU Pro	1371	513	788	716	640	518	954	699	926	940	590	653	660	774
GPQA	1034	778	917	901	806	500	1018	628	541	666	486	472	981	735
MATH	742	1118	1222	1179	748	670	1189	1145	1125	1153	677	675	679	698
GSM8k	572	637	834	719	453	521	709	645	1048	1035	708	680	541	437
GSM8k-Hard	916	939	1027	1069	555	766	1083	1053	1350	1266	594	815	605	512
Folio	724	765	1479	1379	733	668	919	488	1285	1583	907	1194	934	492
BigBench-Hard	596	230	876	861	429	349	315	443	877	973	545	863	455	346

1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

Table 15: Average character index of where the answer span begins in a generated response for each dataset and model pair for the few-shot direct answer setting. We use these numbers as a proxy for the model following instructions (i.e. generating reasoning before an answer). Prompt modifications were made to ensure CoT prompts resulted in longer generations and direct answer prompts led to short generations.

Few-shot Direct Answer Span Location By Character Index										
dataset	Meta-Llama 2.7b	Mistral 7b	Meta-Llama 3.1 8b	Meta-Llama 3.1 70b	Gemma 2.9b	Phi-3 Small 8k	Qwen 2.7b	Qwen 2.72b	GPT-4o Mini	Gemini 1.5 Flash
CommonsenseQA	87	8	27	8	8	8	10	8	8	8
AGIEval LSAT LR	25	24	24	24	24	24	24	24	31	24
AGIEval LSAT AR	25	24	24	24	24	24	24	24	27	24
AGIEval LSAT RC	25	24	24	24	24	24	24	24	25	24
ContextHub Deductive L1	19	19	19	19	19	19	19	19	19	19
ContextHub Deductive L2	19	19	19	19	19	19	19	19	19	19
ContextHub Abductive L1	19	19	19	19	19	19	19	19	19	19
ContextHub Abductive L2	19	19	19	19	19	19	19	19	19	19
MuSR Murder Mysteries	-1	8	27	8	8	8	8	8	8	8
MuSR Team Allocations	-1	21	19	19	27	21	27	23	8	8
MuSR Object Placements	-1	8	27	8	8	8	8	8	8	8
MMLU	19	18	19	19	19	18	18	18	19	19
MMLU Pro	19	19	38	19	20	20	19	19	19	19
GPQA	-1	19	19	19	19	19	19	19	19	19
MATH	29	36	29	29	28	30	30	41	28	28
GSM8k	22	23	23	22	22	23	22	24	27	28

Table 16: Average character index of where the answer span begins in a generated response for each dataset and model pair for the few-shot CoT setting. We use these numbers as a proxy for the model following instructions (i.e. generating reasoning before an answer). Prompt modifications were made to ensure CoT prompts resulted in longer generations and direct answer prompts led to short generations.

Few-shot CoT Answer Span Location By Character Index										
dataset	Meta-Llama 2.7b	Mistral 7b	Meta-Llama 3.1 8b	Meta-Llama 3.1 70b	Gemma 2.9b	Phi-3 Small 8k	Qwen 2.7b	Qwen 2.72b	GPT-4o Mini	Gemini 1.5 Flash
CommonsenseQA	301	195	470	921	145	192	280	174	219	158
AGIEval LSAT LR	1037	510	464	539	437	359	530	599	894	523
AGIEval LSAT AR	1024	1247	886	768	573	1025	750	835	1033	670
AGIEval LSAT RC	799	378	131	206	164	111	241	205	1086	266
ContextHub Deductive L1	383	386	406	376	359	376	388	364	416	366
ContextHub Deductive L2	736	767	829	822	823	855	612	807	884	809
ContextHub Abductive L1	301	386	428	450	431	413	541	447	575	379
ContextHub Abductive L2	709	586	967	754	804	784	829	821	905	815
MuSR Murder Mysteries	-1	1280	1693	1702	1225	1338	1246	1719	1974	1419
MuSR Team Allocations	-1	2195	2087	2160	1628	1755	2181	2156	2632	1841
MuSR Object Placements	-1	907	1104	1213	706	919	676	963	1351	853
MMLU	282	266	333	245	265	260	267	243	392	218
MMLU Pro	429	397	424	411	516	425	541	325	681	396
GPQA	-1	848	782	774	615	711	662	703	670	594
MATH	630	705	584	640	747	529	1074	848	1261	553
GSM8k	374	332	352	352	398	372	415	341	651	314

Table 17: The top 3 slices benefiting the most from CoT across MMLU and MMLU Pro for Llama 3.1 8b and 70b. 6 out of 12 of these top slices directly contain “math” or “mathematics.” We dive deeper into each category subsequently and observe that the questions leading to improvements in the other categories are mathematical in nature as well.

Model	Subject	MMLU				MMLU Pro				
		Direct (%)	CoT (%)	Err. Red. (%)	N	Subject	Direct (%)	CoT (%)	Err. Red. (%)	N
Llama 3.1 8b	elementary_mathematics	46.8	88.4	78.1	378	math	23.6	44.8	27.8	1350
Llama 3.1 8b	high_school_mathematics	39.6	71.5	52.8	270	business	29.4	45.6	23.0	789
Llama 3.1 8b	miscellaneous	83.9	89.9	37.3	783	physics	27.9	41.4	18.8	1299
Llama 3.1 70b	elementary_mathematics	82.3	94.7	70.1	378	math	44.5	68.3	42.9	1351
Llama 3.1 70b	medical_genetics	93.0	97.0	57.1	100	business	44.0	67.8	42.5	789
Llama 3.1 70b	high_school_mathematics	61.5	82.2	53.8	270	chemistry	40.5	64.0	39.6	1132

level. In doing so, we can test our hypotheses with much finer granularity than possible by relying on subjective groupings into tasks and categories.

Breakdown by the presence of equations We aim to design an instance-level classifier to determine if CoT is expected to help on a question or not. That is, we want a function $g : \mathbf{q} \rightarrow \{0, 1\}$ where $g(\mathbf{q})$ returns 1 if $\text{extract}(\tilde{\mathbf{y}}_{cot}) = \mathbf{y}^*$ and $\text{extract}(\tilde{\mathbf{y}}_{da}) \neq \mathbf{y}^*$ where \mathbf{y}^* is the gold answer to \mathbf{q} . We explored different forms of g ; however, we ultimately found it most effective to use a classifier $g : (\mathbf{q}, \tilde{\mathbf{y}}_{cot}) \rightarrow \{0, 1\}$ which also consults the chain-of-thought produced by the model. This allows us to featurize how the LM solves the problem, particularly whether it uses symbolic reasoning or not.

We find that g can be implemented with a **single feature**: does \mathbf{q} or $\tilde{\mathbf{y}}_{cot}$ contain a “=”? The “=” token very strongly indicates the presence of equations in the problem or its solution, which turn out to be a strong hallmark of symbolic reasoning.⁴

We plot the overall CoT delta (performance of CoT minus the performance of direct answer) for both MMLU and MMLU Pro across multiple models between two bins according to this classifier g , labeled as “With =” and “Without =”, in Figure 4. We also report the amount of performance gain explained by questions having an “=” vs. not in Appendix G.1. We find that the majority of the performance gain from CoT on MMLU and MMLU Pro comes from questions that have an “=” in the question or generated responses. Because “=” are usually found in math problems, we equate this to CoT primarily benefiting MMLU and MMLU Pro on the math-related questions with very little to no gain (depending on the model) for non-math questions.

G.1 PERFORMANCE IMPACTS OF “=” ON MMLU AND MMLU PRO

Tables 18 and 19 show the amount of total improvement from using CoT over direct prompting that can be explained by the presence of “=” on MMLU and MMLU Pro over multiple models.

H FULL RESULTS OF EVALUATIONS ON FORMAL REASONING DATASETS

As discussed in Section 5, we include detailed evaluation results of few-shot direct answer, few-shot CoT, direct answer solver, CoT solver, and tool-augmented prompting in Table 20. The unparseable rate stands for the rate of unparseable model responses that either fail to pass our answer extraction parser (for all methods except tool-augmented prompting) or fail to be executed by symbolic solvers. For FOLIO and ContextHub, we compute the accuracy by making a random guess for the unparseable responses; for GSM8K and GSM8K-Hard, we consider the unparseable responses as incorrect.

We note that all models have a low unparseable rate ($< 10\%$) for all methods except tool-augmented prompting. By manually inspecting the outputs, we observe that the high unparseable rate for some models with tool-augmented prompting is caused by these models generating Python programs or

⁴We explored implementing g with a logistic regression classifier with tf-idf features over the $(\mathbf{q}, \tilde{\mathbf{y}}_{cot})$ pairs, trained over a subset of the data from MMLU and MMLU Pro. This classifier actually allowed us to discover the “=” feature, but its accuracy did not exceed the accuracy of that single feature.

2052
 2053
 2054
 2055
 2056
 2057
 2058
 2059
 2060
 2061
 2062
 2063
 2064
 2065
 2066
 2067
 2068
 2069
 2070
 2071
 2072
 2073
 2074
 2075
 2076
 2077
 2078
 2079
 2080
 2081
 2082
 2083
 2084
 2085
 2086
 2087
 2088
 2089
 2090
 2091
 2092
 2093
 2094
 2095
 2096
 2097
 2098
 2099
 2100
 2101
 2102
 2103
 2104
 2105

Table 18: Total CoT deltas on MMLU broken down by the total gain from questions and responses with an “=” vs. without an “=”.

Model	Total CoT Delta	CoT delta w/ =	CoT delta w/o =	Perf. Gain w/ =	Fraction of N w/ =
Llama 2 7b	6.0	0.6	5.4	9.8%	10.9%
Mistral 7b	4.1	1.2	2.9	28.6%	9.8%
Llama 3.1 8b	5.5	2.9	2.6	52.9%	9.6%
Llama 3.1 70b	1.9	1.8	0.1	94.0%	10.6%
Gemma 2 9b	2.6	2.0	0.6	78.5%	10.0%
Phi-3 Small 8k	3.1	1.5	1.7	47.4%	8.3%
Qwen 2 7b	2.5	3.0	-0.5	100.0%	9.8%
Qwen 2 72b	3.5	2.4	1.1	67.8%	9.6%
GPT-4o Mini	5.2	3.5	1.7	66.9%	10.5%
GPT-4o	4.2	2.4	1.8	57.6%	10.3%
Claude-3 Haiku	3.7	2.4	1.3	64.4%	9.3%
Claude-3.5 Sonnet	3.2	2.3	0.9	72.1%	10.7%
Gemini 1.5 Flash	3.0	1.7	1.2	59.0%	10.1%
Gemini 1.5 Pro	1.9	1.0	0.9	51.9%	9.6%

Table 19: Total CoT deltas on MMLU Pro broken down by the total gain from questions and responses with an “=” vs. without an “=”.

Model	Total CoT Delta	CoT delta w/ =	CoT delta w/o =	Perf. Gain w/ =	Fraction of N w/ =
Llama 2 7b	1.6	1.3	0.3	79.6%	43.6%
Mistral 7b	3.8	1.9	1.9	50.7%	41.8%
Llama 3.1 8b	12.4	10.0	2.4	80.8%	35.2%
Llama 3.1 70b	11.4	11.1	0.3	97.6%	39.6%
Gemma 2 9b	7.6	7.4	0.2	97.9%	40.2%
Phi-3 Small 8k	11.6	9.9	1.7	85.7%	42.7%
Qwen 2 7b	10.0	8.9	1.1	88.6%	41.6%
Qwen 2 72b	19.0	16.1	2.9	84.7%	41.4%
GPT-4o Mini	20.6	18.4	2.3	89.0%	44.0%
GPT-4o	17.7	17.1	0.6	96.7%	44.1%
Claude-3 Haiku	8.7	7.8	0.9	90.1%	42.0%
Claude-3.5 Sonnet	16.2	14.8	1.3	91.9%	43.4%
Gemini 1.5 Flash	12.9	11.8	1.1	91.3%	42.3%
Gemini 1.5 Pro	10.0	8.6	1.4	85.7%	41.8%

Table 20: Performance and unparseable rates for few-shot direct answer, few-shot CoT, Plan + Direct Solver, Plan + CoT Solver, and Plan + Tool Solver Solver. “Acc.” stands for accuracy and “% Unp.” stands for the rate of unparseable model responses that either fail to pass our answer extraction parser (for all methods except Plan + Tool Solver prompting) or fail to be executed by symbolic solvers. For FOLIO and ContextHub, we compute the accuracy by making a random guess for the unparseable responses; for GSM8K and GSM8K-Hard, we consider the unparseable responses as incorrect.

Dataset	Method	Mistral 7b		Llama 3.1 8b		Llama 3.1 70b		GPT-4o Mini	
		Acc.	% Unp.	Acc.	% Unp.	Acc.	% Unp.	Acc.	% Unp.
GSM8K	Direct Answer	12.5	0.1	20.1	0.5	39.1	0.0	32.8	0.0
GSM8K	CoT	56.2	1.4	86.4	1.0	96.1	0.1	94.2	0.1
GSM8K	Plan + CoT Solver	45.0	1.0	78.7	0.4	94.7	0.0	92.0	0.1
GSM8K	Plan + Direct Solver	10.6	0.1	19.6	0.1	42.2	0.0	39.3	0.0
GSM8K	Plan + Tool Solver	59.8	8.6	80.3	1.3	94.4	0.4	90.5	1.5
GSM8K-Hard	Direct Answer	2.9	0.7	4.4	0.6	12.8	0.7	12.3	7.6
GSM8K-Hard	CoT	20.3	5.0	32.4	9.6	47.8	4.4	52.2	0.5
GSM8K-Hard	Plan + CoT Solver	18.7	2.6	32.4	1.3	49.7	0.6	51.5	0.3
GSM8K-Hard	Plan + Direct Solver	3.0	0.5	5.5	0.8	15.8	0.1	17.4	0.3
GSM8K-Hard	Plan + Tool Solver	44.2	8.9	57.9	1.2	68.0	0.5	70.4	1.4
ContextHub Deductive L1	Direct Answer	59.2	2.8	23.0	0.0	50.0	0.0	44.3	0.0
ContextHub Deductive L1	CoT	46.2	0.2	73.0	0.2	67.5	0.0	59.2	0.0
ContextHub Deductive L1	Plan + CoT Solver	49.5	0.0	64.8	0.0	65.5	0.0	63.2	0.0
ContextHub Deductive L1	Plan + Direct Solver	45.8	3.0	55.8	0.0	53.5	0.0	56.2	0.0
ContextHub Deductive L1	Plan + Tool Solver	68.8	27.8	84.2	11.8	91.7	9.8	90.7	7.8
ContextHub Abductive L1	Direct Answer	21.7	2.8	36.1	0.0	58.9	0.0	59.2	0.0
ContextHub Abductive L1	CoT	23.9	0.0	40.0	0.0	62.2	0.0	76.9	0.0
ContextHub Abductive L1	Plan + CoT Solver	38.3	0.0	42.5	0.0	65.6	0.0	74.2	0.0
ContextHub Abductive L1	Plan + Direct Solver	46.9	3.9	33.3	0.3	63.1	0.0	61.7	0.0
ContextHub Abductive L1	Plan + Tool Solver	59.2	35.8	70.8	9.7	73.9	4.2	74.7	10.3
FOLIO	Direct Answer	56.2	12.3	59.6	0.0	69.5	0.0	64.0	0.0
FOLIO	CoT	53.7	1.5	56.7	2.5	72.4	2.0	70.4	0.0
FOLIO	Plan + CoT Solver	53.7	0.0	55.7	0.0	73.9	0.5	70.4	0.0
FOLIO	Plan + Direct Solver	52.7	0.0	54.2	0.0	72.9	0.0	63.5	0.0
FOLIO	Plan + Tool Solver	48.8	46.8	54.2	28.6	70.0	16.7	62.6	25.1

formal specifications that fail to follow the format of the formal language (Python or z3) and that lead to execution errors. Such an issue is particularly severe for the smaller models. However, we note that despite the high unparseable rate, the overall accuracy of these models with tool augmentation is still on par with or outperforms other methods.

I DISCUSSION OF LIMITATIONS

I.1 LONG HORIZON PLANNING

One set of tasks where symbolic reasoning helps substantially that our experiments haven’t covered as thoroughly (with the exception of BiGGen-Bench) is long-horizon planning (Valmeekam et al., 2023; Xie et al., 2024; Gundawar et al., 2024; Valmeekam et al., 2024). There are two reasons we don’t treat it here. First, we are primarily interested in tasks that are conveyed in language, and we see less complex planning in language-only tasks. Second, there has already been a large debate on the effectiveness of CoT, both pro (Huang et al., 2022; Hu et al., 2023) and against (Valmeekam et al., 2023; Kambhampati, 2024; Kambhampati et al., 2024b; Stechly et al., 2024a; Guan et al., 2024; Verma et al., 2024; Gundawar et al., 2024; Stechly et al., 2024b) using CoT and its derivatives like tree-of-thought (Yao et al., 2023; Kang et al., 2024), that has resulted in complex systems to help solve planning problems better. While story generation and interpretation involve elements of planning with natural language (Peng et al., 2022; Karpinska et al., 2024), such tasks are not conventionally formalized and benchmarked as planning and reasoning.

I.2 DATASET CONTAMINATION

One limitation of our study is the presence of possible data contamination: it is unknown which benchmarks may have been explicitly pre-trained on by language models. If a model had memorized answers to benchmark questions, we would expect direct answering to close some of the gap with CoT, as the model can just reproduce a known answer rather than deriving it from scratch. We argue

2160 there are four reasons that our general conclusions are still trustworthy. First, we use a range of
2161 language model scales, including small models that have less capacity to memorize. Second, datasets
2162 with poor direct answering performance like GSM8K-Hard are unlikely to have been substantially
2163 memorized. Third, the inclusion of recent datasets such as MuSR (Sprague et al., 2024) and BiGGen
2164 Bench (Kim et al., 2024) helps to defray this risk. Fourth, our survey of the literature includes papers
2165 that were submitted to conferences in 2023, representing a range of older LLMs trained at various
2166 times.

2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213

2214 J EXAMPLE PROMPTS

2215

2216

2217

2218

2219

2220

2221

2222

2223

2224 We will release all prompts and model responses on our Huggingface repo. We list a few prompt
2225 response pairs here in this section.

2226

2227

2228

2229

2230

2231

2232

2233

2234

2235

2236

2237

2238

AGIEval LSAT AR zero-shot CoT prompt for Llama 3.1 70B

2239

2240

```
<|start_header_id|>user<|end_header_id|>
```

2241

2242

2242 Explain your reasoning step-by-step for each question before answering.
2243 Give your final answer in the format \"The answer is therefore <A, B, C,
2244 D, E>\". Failure to comply with the answer formatting will result in no
2245 credit.

2246

2247

2248

2249

2250

2251

2252

2253

2254

2255

2256

2257

2258

2259

2260

2261

2262

2263

2264

2264 Let's think step by
2265 step.<|eot_id|><|start_header_id|>assistant<|end_header_id|>

2266

2267

2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321

AGIEval LSAT AR zero-shot CoT response for Llama 3.1 70B

To determine which one of the following could be the schedule of the students' reports, let's analyze each answer choice in light of the given conditions. First, let's recall the conditions: 1. Exactly six students will give reports over three days. 2. Tuesday is the only day George can give a report. 3. Neither Olivia nor Robert can give an afternoon report. 4. If Nina gives a report, then on the next day Helen and Irving must both give reports, unless Nina's report is given on Wednesday. Let's evaluate each answer choice: (A) Mon. morning: Helen; Mon. afternoon: Robert Tues. morning: Olivia; Tues. afternoon: Irving Wed. morning: Lenore; Wed. afternoon: Kyle - This choice does not violate any conditions directly. However, we need to check if it aligns with all the rules, especially considering the Nina rule, which isn't directly applicable here since Nina isn't scheduled. George's requirement is also not met since George isn't scheduled on Tuesday. (B) Mon. morning: Irving; Mon. afternoon: Olivia Tues. morning: Helen; Tues. afternoon: Kyle Wed. morning: Nina; Wed. afternoon: Lenore - This choice violates the condition that if Nina gives a report, then on the next day Helen and Irving must both give reports, unless Nina's report is given on Wednesday. Since Nina's report is on Wednesday, this condition is technically not violated, but we need to ensure all other conditions are met. George is not scheduled on Tuesday, which is a requirement. (C) Mon. morning: Lenore; Mon. afternoon: Helen Tues. morning: George; Tues. afternoon: Kyle Wed. morning: Robert; Wed. afternoon: Irving - This choice meets the condition that George gives a report on Tuesday. It also does not violate any other conditions directly. However, we need to ensure that the Nina rule isn't applicable or is met, which in this case, it's not applicable since Nina isn't scheduled. (D) Mon. morning: Nina; Mon. afternoon: Helen Tues. morning: Robert; Tues. afternoon: Irving Wed. morning: Olivia; Wed. afternoon: Lenore - This choice violates the condition that if Nina gives a report, then on the next day Helen and Irving must both give reports. Here, Helen is already scheduled on the same day as Nina, and while Irving is scheduled the next day, the condition specifies both Helen and Irving must give reports the next day, which is not the case since Helen's report is on the same day as Nina's. (E) Mon. morning: Olivia; Mon. afternoon: Nina Tues. morning: Irving; Tues. afternoon: Helen Wed. morning: Kyle; Wed. afternoon: George - This choice violates the condition that Tuesday is the only day George can give a report, as George is scheduled on Wednesday. Given the analysis, the only choice that does not directly violate any of the given conditions and meets the specific requirements (especially considering the placement of George and the handling of the Nina rule where applicable) is: The answer is therefore (C).

2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375

MATH zero-shot CoT prompt for Llama 3.1 70B

<|begin_of_text|><|start_header_id|>user<|end_header_id|>

Solve the following math problem efficiently and clearly:

- For simple problems (2 steps or fewer):\nProvide a concise solution with minimal explanation.

- For complex problems (3 steps or more):

Use this step-by-step format:\n\n## Step 1: [Concise description]\n[Brief explanation and calculations]

Step 2: [Concise description]

[Brief explanation and calculations]

...

Regardless of the approach, always conclude with:

Therefore, the final answer is: $\boxed{\text{answer}}$. I hope it is correct.

Where [answer] is just the final number or expression that solves the problem.

Problem: Chris labels every lattice point in the coordinate plane with the square of the distance from the point to the origin (a lattice point is a point such that both of its coordinates are integers). How many times does he label a point with the number

25?<|eot_id|><|start_header_id|>assistant<|end_header_id|>

MATH zero-shot CoT response for Llama 3.1 70B

Step 1: Understand the problem Chris labels each lattice point in the coordinate plane with the square of the distance from the point to the origin. We need to find how many times he labels a point with the number 25. ## Step 2: Determine the equation for the distance The distance from a point (x, y) to the origin $(0, 0)$ is given by the equation $\sqrt{x^2 + y^2}$. Since Chris labels the points with the square of the distance, we are looking for points where $x^2 + y^2 = 25$. ## Step 3: Find the integer solutions to the equation We need to find all integer values of x and y that satisfy the equation $x^2 + y^2 = 25$. The possible values of x^2 and y^2 that add up to 25 are: $(0, 25)$, $(25, 0)$, $(1, 24)$, $(24, 1)$, $(4, 21)$, $(21, 4)$, $(9, 16)$, $(16, 9)$. ## Step 4: Count the number of solutions However, we need to consider that some of these pairs have negative counterparts, and we should also count those. For example, $(-0, 5)$ and $(0, -5)$ are also solutions, as well as $(-5, 0)$ and $(5, 0)$. Similarly, $(-3, 4)$, $(3, -4)$, $(-4, 3)$, and $(4, -3)$ are also solutions. So, we have a total of 12 solutions. The final answer is: $\boxed{12}$.

2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429

```
CommonsenseQA zero-shot direct answer prompt for Llama 3.1 70B

<|start_header_id|>system<|end_header_id|>

You are a helpful AI assistant that will answer reasoning questions. You
will always say at the end \"Answer: <Your Answer Letter Choice>\". You
must only pick one answer and you must end your response with \"Answer:
<Your Answer Letter Choice>\"
everytime!<|eot_id|><|start_header_id|>user<|end_header_id|>

Question: Where would you find magazines along side many other printed
works?

Answer Choices:
( A ) doctor
( B ) bookstore
( C ) market
( D ) train station
( E ) mortuary

Only write the answer. Write the answer in the following format: \"Answer:
<your answer>\". You must always give an answer. You may only pick one
answer choice, if you think multiple are correct only pick the one you
think is best.<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

```
CommonsenseQA zero-shot direct answer response for Llama 3.1 70B

B
```