LLMs meet Bloom's Taxonomy: A Cognitive View on Large Language Model Evaluations

Thomas Huber
University of St. Gallen, Switzerland
thomas.huber@unisg.ch

Christina Niklaus
University of St. Gallen, Switzerland
christina.niklaus@unisg.ch

Abstract

Current evaluation approaches for Large Language Models (LLMs) lack a structured approach that reflects the underlying cognitive abilities required for solving the tasks. This hinders a thorough understanding of the current level of LLM capabilities. For instance, it is widely accepted that LLMs perform well in terms of grammar, but it is unclear in what specific cognitive areas they excel or struggle in. This paper introduces a novel perspective on the evaluation of LLMs that leverages a hierarchical classification of tasks. Specifically, we explore the most widely used benchmarks for LLMs to systematically identify how well these existing evaluation methods cover the levels of Bloom's Taxonomy, a hierarchical framework for categorizing cognitive skills. This comprehensive analysis allows us to identify strengths and weaknesses in current LLM assessment strategies in terms of cognitive abilities and suggest directions for both future benchmark development as well as highlight potential avenues for LLM research. Our findings reveal that LLMs generally perform better on the lower end of Bloom's Taxonomy. Additionally, we find that there are significant gaps in the coverage of cognitive skills in the most commonly used benchmarks.

1 Introduction

Large Language Models (LLMs), such as GPT-4 (OpenAI et al., 2024), GPT-4o (OpenAI, 2024), Llama (Meta, 2024), Claude (Anthropic, 2024), Mistral (Jiang et al., 2023), Bloom (Workshop et al., 2023), or Gemma (Team et al., 2024a), have demonstrated impressive capabilities across a diverse range of tasks such as code generation (Zhong and Wang, 2024), logical reasoning (Wei et al., 2022), fact checking (Zhang and Gao, 2023) and many others (Yang et al., 2024). LLMs are published with evaluations of their performance on a variety of benchmarks. They serve as standardized tests and are used to give an overview of the capabilities of these models and to allow comparison between them. However, evaluations on these benchmarks provide a limited overview of the general capabilities of the models. A high performance on a reading comprehension benchmark for instance does not translate to a high performance on a task that tests other cognitive skills, such as arithmetic. Similarly, a low performance on a benchmark does not immediately highlight general weaknesses of the models. This makes it difficult to pinpoint potential avenues for research on how to improve model performance.

This problem is further exacerbated by the fact that benchmarks can consist of multiple subtasks that differ in the cognitive abilities required to solve them. Despite this, the scores on benchmarks with subtasks are often presented as an aggregate across all subtasks. Figure 1 shows examples of different subtasks found in benchmarks that test different cognitive and knowledge skills. For instance the subtask `BBH object_counting` is a simple reading comprehension task that does not require deep reasoning, while `BBH snarks` asks the models to detect sarcasm, which is a cognitively much more challenging task. Performance of models on individual subtasks can vary. For instance, we have evaluated GPT-4 on the individual BIG-Bench Hard subtasks and found that while it generally performs very well, achieving an aggregated average accuracy of ≈ 0.89, it struggled with the `salient_translation_error_detection`[1] task, achieving an accuracy of ≈ 0.48.[2][3] This discrepancy creates a risk of overestimating the validity of LLM-generated content, as a high overall score

[1]The task is to classify errors introduced by translation for pairs of original and corresponding translated sentences.

[2]We present the results of our full evaluation in Appendix A.

[3]Our code and results are available on https://github.com/ThHuberSG/coling-bloom.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5211–5246
January 19–24, 2025.

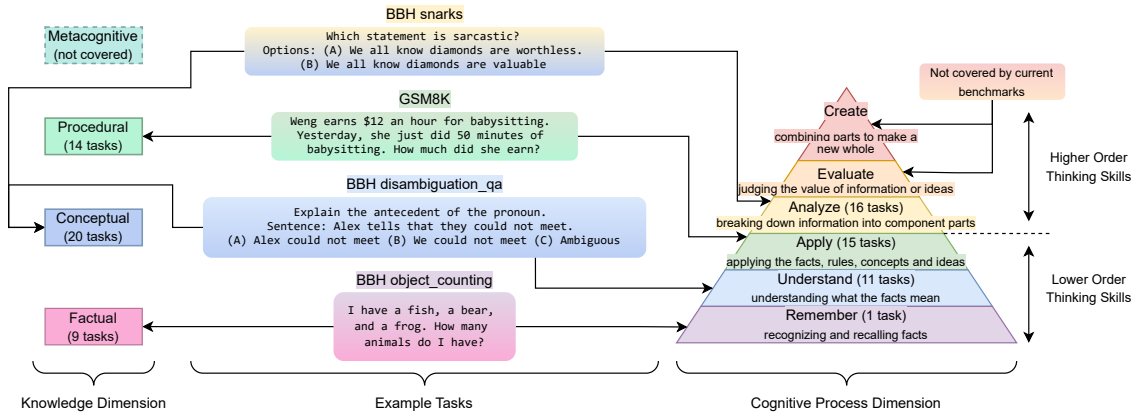


Figure 1: Distribution of currently used benchmarks for LLM evaluation when mapped to Bloom's Taxonomy. Metacognitive, Create and Evaluate are not covered by currently used benchmarks.

does not necessarily indicate a well-rounded performance across all cognitive levels. We argue that to gain a balanced overview of the models' performance it is necessary to change how we view benchmark scores and measure model performance. Instead of measuring performance on individual benchmarks, we propose to measure the models' performance in terms of cognitive abilities. Enhancing and measuring cognitive abilities typically belongs to the field of education. In this domain Bloom's Taxonomy (Bloom et al., 1956; Anderson and Krathwohl, 2001) is a widely used and established framework for classifying learning objectives in the cognitive dimension into six hierarchical levels, each representing increasing complexity of thought. An overview of the taxonomy is presented in Section 2.

We argue that analyzing LLM performance through the lens of Bloom's Taxonomy allows us to uncover the knowledge structures they possess and identify areas for improvement (Zhang et al., 2023). Firstly, it provides deeper insights into the cognitive abilities of LLMs, paving the way for more targeted development. Secondly, this approach can guide the selection of the most effective and appropriate use cases for LLMs. In this paper we map commonly used benchmarks into the taxonomy to assess how much of the taxonomy is covered by current LLM evaluation paradigms. We find that commonly used benchmarks do not sufficiently cover all levels of Bloom's Taxonomy, which suggests that the evaluations are not comprehensive enough and do not accurately represent a measurement of the LLMs' capabilities.

This work makes several key contributions to the field of LLM evaluation:

(i) We leverage Bloom's Taxonomy to establish a connection between commonly used benchmark tasks and the cognitive abilities they require. This provides an overview of the coverage that these popular benchmarks achieve on Bloom's Taxonomy.

(ii) Based on this mapping we measure the performance of LLMs on the levels of Bloom's Taxonomy. This reveals the types of learning and thinking LLMs excel at, along with areas where they struggle. This knowledge allows for a more targeted and effective development and application of LLMs.

(iii) Through Bloom's Taxonomy, we identify cognitive dimensions and knowledge types that are currently not well-represented by commonly used LLM benchmarks.

2 Background: Bloom's Taxonomy

Bloom's Taxonomy (Bloom et al., 1956; Anderson and Krathwohl, 2001)[4] serves as a foundational framework in education. It offers a two-dimensional framework for classifying learning objectives into six distinct levels of cognitive complexity, ranging from lower-order thinking skills that require less cognitive processing (*remember, understand, apply*) to higher-order thinking skills that require deeper learning and a greater degree of cognitive processing (*analyze, evaluate, create*). These levels represent a hierarchy of cognitive

[4]We employ the revised version of the taxonomy presented in Anderson and Krathwohl (2001).

skills, with each level building upon the foundation laid by the previous one (see the right part of Figure 1). The second dimension focuses on the type of knowledge students acquire when solving a task: factual, conceptual, procedural, and metacognitive.[5]

Bloom's Taxonomy serves two main purposes in the field of education: (i) It provides a structured framework for educators to categorize and organize learning goals into a hierarchy of increasing complexity. (ii) By understanding the different cognitive levels and knowledge types, educators can design activities, assessments, and teaching methods that effectively target these different areas. It enables them to accurately evaluate the different types of mental skills students develop, ensuring a well-rounded assessment of their learning.

This equips them with the tools to not only understand the different types of complex mental skills required for effective learning but also to effectively evaluate them in their students, fostering a more comprehensive learning journey for students.

This promotes a more well-rounded learning experience for students. In that way, it equips educators with the tools to not only understand the different types of complex mental skills required for effective learning but also to effectively evaluate them in their students.

In essence, Bloom's Taxonomy helps educators move beyond simply imparting information and instead focus on fostering critical thinking, analysis, creativity, and a deeper understanding in their students.

3 Methodology

We leverage Bloom's Taxonomy to explore the knowledge structures and cognitive abilities of LLMs. Specifically, we address the following research questions:

(i) RQ1: Can commonly used benchmarks used for LLM evaluation be mapped to cognitive capabilities they cover?

(ii) RQ2: What cognitive dimensions and knowledge types are underrepresented in current LLM benchmark tasks?

(iii) RQ3: What types of learning and thinking do LLMs excel at, and where do they struggle?

[5]For a detailed description of the knowledge and cognitive dimensions of Bloom's Taxonomy, the interested reader may refer to Section B in the appendix.

3.1 Mapping Benchmark Tasks to Bloom's Taxonomy

New LLMs are evaluated on a subset of benchmarks to provide a general overview of their capabilities. The chosen benchmarks are often similar across multiple models, but there is not one definitive selection. LLMs are used for a variety of tasks. It is widely accepted that they have very strong performance for language tasks such as grammar and spelling. This is reflected in the rankings of general language benchmarks such as GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) where LLMs rank highly. The cognitive skills required for these tasks however are different from tasks such as evaluating the quality of an argument in an essay. Benchmarks used in the evaluation of LLMs typically do not include an analysis of the cognitive skills required to solve them. We hypothesize that the frequently used benchmarks do not fully test all the capabilities of LLMs. We first investigate whether benchmarks used for LLM evaluation can be mapped to Bloom's Taxonomy, and thus can be considered to test cognitive abilities.

3.1.1 Benchmark Selection

To test this hypothesis, we have selected a set of commonly used benchmarks. We base our selection on their usage in the technical reports of recently published models, specifically GPT-4o (OpenAI, 2024), GPT-4 (OpenAI et al., 2024), Llama 3 (Meta, 2024) and Llama 3.1 (Dubey et al., 2024), Grok-2 (X.ai, 2024) and Claude 3 (Anthropic, 2024). We provide an overview in Appendix F in Table 16. Our choice of models is driven by general performance and popularity in the research community. We argue that the benchmarks these models are evaluated on present the current zeitgeist in the community on how LLMs should be evaluated to provide at least a basic overview of their capabilities. To assess whether or not these benchmarks cover cognitive abilities we first annotated the selected benchmarks by the levels they cover in Bloom's Taxonomy. For benchmarks that are made up of subtasks we considered each subtask to be a distinct task and annotated the subtasks separately. Benchmarks without a subtask categorization were considered as a whole and not split. Our selection approach left us with 43 distinct tasks in total. An overview of the used benchmarks and datasets is provided in Appendix C in Table 13.

We found that the results of benchmarks with subtasks, such as AGIEval (Zhong et al., 2023) and

BIG-Bench Hard (BBH) (Suzgun et al., 2023), are frequently not reported per subtask, but only as an aggregate over all subtasks. We disagree with this approach. The subtasks in each of these two benchmarks are very different and aggregating the scores across them leads to an arbitrary metric that does not accurately reflect the performance of the models. An aggregated score is a sensible approach when the individual tasks measure some kind of specific, narrow purpose, and the aggregate then provides a balanced view for this purpose. This is not the case for all benchmarks with subtasks. We argue that scores on benchmarks with subtasks should always be reported on a subtask level to allow for a more fine-grained analysis and comparison of the tested models.

We provide an overview of the selected benchmarks and model performance in Appendix A.

3.1.2 Labeling of Benchmark Tasks

Prior research in the field of Natural Language Processing (NLP) has not, to our knowledge, explored the classification of LLM benchmark tasks within Bloom's Taxonomy. Consequently, existing datasets and benchmarks that are commonly used for the evaluation of LLMs lack information regarding the knowledge types and cognitive dimensions that are targeted by these tasks.

To address this gap, we adopted a three-folded approach. First, we carried out a human annotation, where human annotators manually labeled the cognitive and knowledge dimensions present in the benchmark tasks. Second, we leveraged the capabilities of LLMs to analyze the tasks and identify these dimensions. Finally, we trained a machine learning classifier to solve this task. We aim to examine whether the tasks can be mapped to cognitive abilities by means of having a high inter-rater agreement as well as by being able to train a model to do so.

Annotator Agreement The boundaries between the levels of Bloom's Taxonomy are fuzzy, and we have found that in practice individual tasks fall somewhere in between two levels. For instance the `boolean_expressions` task from BIG-Bench Hard requires the evaluation of boolean expressions into True/False. This requires both recalling how to evaluate these expressions in general and simplify them, which falls into the *Understand* dimension, as well as carrying out the simplification across multiple steps, which can be considered to belong to the *Apply* dimension. These levels are consecutive in the taxonomy, and a disagreement between them is less severe than if the annotators assigned labels that are on opposite ends of the taxonomy hierarchy. For this reason we measure the inter-rater agreement using Cohen's weighted kappa (Cohen, 1968), which accounts for partial disagreements on nominal scales. The agreement scores can be found in Table 1. We discuss the scores in Section 4.1.

Manual Annotation In a first step, two annotators, with expertise in computer science but no formal background in pedagogy, manually labeled the cognitive and knowledge dimensions in the selected benchmarks independently of each other. For benchmarks that consist of multiple distinct subtasks, such as BIG-Bench Hard, the individual subtasks were annotated as opposed to the benchmark as a whole. After the annotation process they discussed the annotations where they disagreed and discussed the reasoning behind their choice until they reached a consensus. This follows the approach by Li et al. (2022), who performed a similar annotation. The annotators in our work had mainly the same issue as in the aforementioned publication. Multiple reading comprehension tasks in BBH for instance require some additional reasoning, which initially lead to different classifications. This was the main source of disagreement.

LLM Annotation To further strengthen the reliability of our human annotations, we used a simple prompt to instruct the current, high-performing LLMs GPT-4 and GPT-4o[6], Claude 3[7] and Llama 3[8] to annotate the same benchmark tasks with their corresponding cognitive and knowledge dimensions. The exact prompts can be found in Appendix D. We have used a subset of 20 samples of each task and assigned the majority label. We find that the models very rarely deviate in their label within a task and will instead almost always assign the same label, and therefore have opted to label only the subsets. Only the MMLU benchmark was an exception. We have found that, despite being split into subtasks, the individual task instances in each subtask can vary in terms of cognitive and knowledge skill required to solve them. We discuss this in Section 5. After annotation we measured the inter-rater agreement between the human annota-

[6] `gpt-4-0613` and `gpt-4o-2024-05-13`, through the OpenAI API

[7] `claude-3-haiku-20240307` through the Anthropic API

[8] Self-hosted `meta-llama/Meta-Llama-3-70B-Instruct`

	Cognitive Dimension						Knowledge Types					
	Human	claude3	gpt4	gpt4o	llama3	avg.	Human	claude3	gpt4	gpt4o	llama3	avg.
Human	-	0.51	0.69	0.66	0.67	0.63	-	0.34	0.17	0.38	0.33	0.31
claude3	0.51	-	0.59	0.59	0.7547	0.61	0.34	-	0.33	**0.66**	0.36	0.42
gpt4	**0.69**	0.59	-	0.77	0.7549	0.70	0.17	0.33	-	0.45	0.21	0.29
gpt4o	0.66	0.59	**0.77**	-	**0.78**	0.70	**0.38**	**0.66**	**0.45**	-	**0.34**	**0.46**
llama3	0.67	**0.7547**	0.7549	**0.78**	-	**0.74**	0.33	0.36	0.21	0.34	-	0.31

Table 1: Inter-model agreement for Cognitive Dimension and Knowledge Types

tions and the LLMs. Agreement scores be found in Table 1.

Training of a Classification Model To complement the human and LLM annotations we fine-tuned a classification model on the dataset presented by Li et al. (2022). This dataset contains 21,380 learning objectives, labelled with the corresponding cognitive dimension from Bloom's Taxonomy. The knowledge dimension is not included. We trained a small pre-trained RoBERTa model (Conneau et al., 2019), `xlm-roberta-base` from the HuggingFace repository[9] which has 279 million parameters. We used 10% of the full dataset for testing, 10% of the remaining data for validation and all the remaining data for training. Our fine-tuned model achieves an F1-Score of 0.92 on the test set. We used this classifier to predict the cognitive dimension labels for the benchmark tasks. The agreements with both humans and LLMs is very low despite the high performance of the classifier on the test set (highest agreement 0.04). This likely stems from the fact that the dataset used for training consists of abstract learning objectives and not specific tasks to be solved. As the performance of the classifier is very high, yet the agreement is low, we decided not to pursue this direction.

3.2 Performance of LLMs on Different Taxonomy Dimensions

Model Selection Due to the large number of available models[10] it is not feasible to evaluate them all. We have chosen GPT-4[11] (OpenAI et al., 2024), Llama 3[12] (Meta, 2024) and Claude 3[13] (Anthropic, 2024) for our analysis. We evaluated additional models on the BBH and AGIEval datasets: Phi-3 (Abdin et al., 2024), Gemma (Team et al., 2024b), Falcon (Almazrouei et al., 2023), Flan-T5 (Chung et al., 2022) and Bloomz (Muennighoff et al., 2023). We chose to perform our own evaluation on these two benchmarks because their subtasks are heterogenous and benchmark scores on all selected benchmarks are not publicly available for all models. The models were selected because they achieve high performance, are very recent and, we argue, are representative of the current paradigms in LLM evaluations. This is due to the fact that they share many of the same datasets in their evaluations. Our selection includes both open and closed models and range from 3.8b parameters (Phi-3) to 70b (Llama 3) and likely even higher in the case of GPT-4, as well as both an encoder-decoder model (Flan-T5) and decoder-only models. Similarly to how there is a large number of available models there exist many datasets and benchmarks for a variety of tasks in the broader NLP field[14]. The used LLMs place highly on many public leaderboards and are widely used and adopted by the NLP community.

Model Scoring Where they were available, we have used scores from the technical reports or publications. The exact scores and their source can be found in Appendix A. For the BIG-Bench Hard and AGIEval benchmarks we were unable to find scores reporting performance on a subtask level for all of the models. We have carried out our own evaluation on these benchmarks. For BIG-Bench Hard we have used the Chain-of-Thought prompts presented by Suzgun et al. (2023). We employed the same strategy for the AGIEval benchmark: Chain-of-Thought with zero-shot prompting. We then used Llama 3 to evaluate the output the models produced. The verification prompt is included in Figure 5 in Appendix D. We manually verified a small subset of the automated evaluations by comparing the verification output with both the correct

[9] `https://huggingface.co/FacebookAI/xlm-roberta-base`, accessed May 28, 2024

[10] `https://huggingface.co/models?pipeline_tag=text-generation` lists 135,873 models for text generation as of September 16, 2024

[11] `gpt-4-0613`

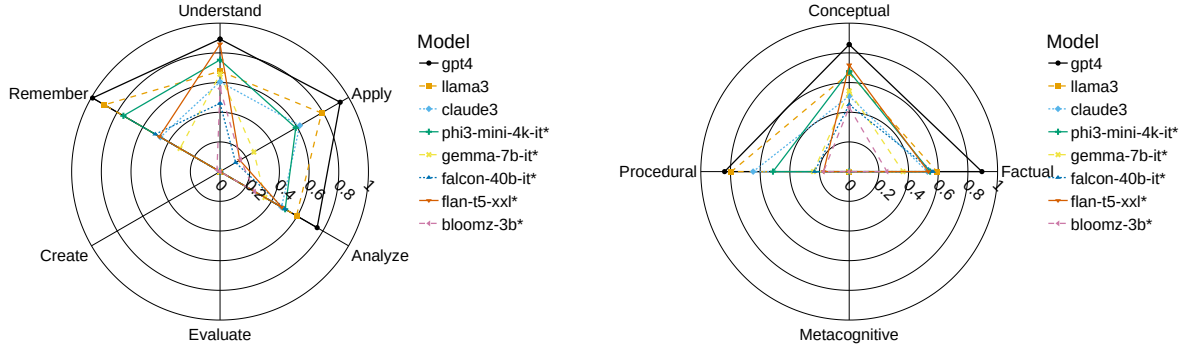
[12] `meta-llama/Meta-Llama-3-70B-Instruct`

[13] `claude-3-haiku-20240307`

[14] HuggingFace lists 16,287 NLP datasets as of September 16, 2024

Model	Cognitive Dimension						Knowledge Type			
	Remember	Understand	Apply	Analyze	Evaluate	Create	Factual	Conceptual	Procedural	Metacognitive
Llama3	0.90	0.68	0.79	0.59	N/A	N/A	0.59	0.66	0.79	N/A
GPT-4	0.99	0.89	0.93	0.76	N/A	N/A	0.89	0.86	0.84	N/A
Claude3	0.46	0.60	0.62	0.48	N/A	N/A	0.53	0.51	0.65	N/A
bloomz-3b*	0.02	0.56	0.16	0.27	N/A	N/A	0.25	0.43	0.17	N/A
bloomz-560m*	0.01	0.27	0.12	0.22	N/A	N/A	0.21	0.26	0.14	N/A
falcon-7b-instruct*	0.11	0.53	0.07	0.36	N/A	N/A	0.33	0.48	0.12	N/A
falcon-40b-instruct*	0.50	0.46	0.12	0.50	N/A	N/A	0.56	0.45	0.23	N/A
flan-t5-xxl*	0.47	0.86	0.15	0.48	N/A	N/A	0.53	0.72	0.17	N/A
gemma-7b-it*	0.31	0.66	0.26	0.35	N/A	N/A	0.36	0.54	0.24	N/A
phi3-mini-4k-instruct*	0.75	0.75	0.59	0.50	N/A	N/A	0.55	0.68	0.51	N/A

Table 2: Model performance on Bloom's Taxonomy's Cognitive Dimension and Knowledge Types. Model scores marked with * are based only on our own evaluation on BBH and AGIEval.



(a) Performance of models by cognitive process dimension

(b) Performance of models by knowledge type

Figure 2: Performance of models mapped to Bloom's Taxonomy. Model scores marked with * are based only on BBH and AGIEval. Unmarked scores are based on all benchmarks.

answer for the task as well as the generated answer and found no errors. Nevertheless we can not guarantee that these scores are perfectly accurate, but we do not aim to provide an exact evaluation of the model's capabilities but rather a comprehensive overview of their general performance mapped to the levels of Bloom's Taxonomy.

4 Results

For the experiments we have used the human-assigned labels as discussed in Section 3.1.2. For benchmarks that consist of multiple subtasks, such as BIG-Bench Hard, we consider each subtask as an individual, separate task. Benchmarks that do not consist of subtasks were considered as a whole. The labels assigned to the benchmarks by both humans and LLMs can be found in Appendix E.

4.1 RQ1: Mapping LLM Benchmark Tasks to Cognitive Capabilities

For the cognitive dimension, the scores between pairs of LLMs are high, and they align with the human annotations ($\kappa_{\text{weighted, avg.(humans/LLMs)}} \approx 0.63$). This indicates a high reliability of the cognitive labels assigned to the benchmarks, suggesting that the taxonomy can be applied to benchmark tasks in general.

The agreement scores for the knowledge types are more varied. Claude 3 and GPT-4o achieve a high agreement of $\kappa_{\text{weighted}} \approx 0.66$. The average agreement score between humans and LLMs is $\kappa_{\text{weighted, avg.(human/LLMs)}} \approx 0.30$. This low agreement could stem from low LLM performance for the task of assigning knowledge dimension labels. To the best of our knowledge no suitable dataset exists that could be used to evaluate the performance of LLMs on this task. The agreement score we report here gives a weak indication that LLMs have trouble with assigning the knowledge dimension label, but a suitable dataset to evaluate this is needed. An indication of weak performance is the fact that the GPT-4 family of models (both GPT-4 and GPT-4o) assigned the *Metacognitive* label, albeit rarely. This knowledge type requires self-reflection about

one's own learning and knowledge. The tasks in the benchmarks we considered are all straightforward and can be solved without reflection. They should not be considered to fall into this knowledge type for this reason. The human annotators never assigned this label.

The assigned labels can be found in Appendix E in Tables 14 and 15.

4.2 RQ2: Representation of Cognitive and Knowledge Dimensions in Benchmark Tasks

Cognitive Dimensions Of the 43 tasks we considered in our evaluation only one covers the *Remember* dimension, 11 the *Understand* dimension, 15 *Apply* and 16 *Analyze*. None of the tasks cover the *Evaluate* or *Create* dimensions. The assigned labels for the cognitive dimension show a focus towards the middle part of the taxonomy, and the very low and higher-order thinking skills *Remember* and *Evaluate*, *Create* are underrepresented.

Knowledge Types *Factual* knowledge is tested by 9 tasks and *Procedural* knowledge by 14. *Conceptual* knowledge is represented by 20 tasks. None of the benchmarks map to the *Metacognitive* level.

4.3 RQ3: Performance of LLMs according to Bloom's Taxonomy

As discussed in Section 3.2 we compare the performance of multiple models on the selected benchmarks. Detailed scores can be found in Appendix A. We include aggregated scores in Table 2. Figure 2 shows a visualization of the models' performance. We find that the models generally perform better on the lower end of the taxonomy. Models that score lowly on the *Remember* dimension still perform well on tasks belonging to *Understand*. We note that only one task of the benchmarks is mapped to *Remember*, `object_counting` of the BIG-Bench Hard benchmark. We manually analyzed the outputs of models on this task and found that they often misinterpret the question. For example for the task "I have a head of broccoli, four garlics, a yam, a stalk of celery, a cabbage, two potatoes, an onion, four lettuce heads, and a cauliflower. How many vegetables do I have?" Claude 3 counted hallucinated, non-existent fruits and gave that as the answer. The drop in performance on the taxonomy the higher the dimension indicates that the models may have a weakness in the higher-order thinking skills. This follows from the overall lower performance on the *Analyze* dimension, which is represented by 16 tasks.

For the knowledge dimension we observe that Llama 3's performance increases as the knowledge dimension increases, ranging from 0.59 at the *Factual* level, to 0.66 for *Conceptual* and 0.79 for tasks that require the application of *Procedural* knowledge. GPT-4 is consistent across the dimensions (0.89, 0.86, 0.84) while Claude 3 peaks at *Procedural* (0.53, 0.51, 0.65). The other models perform highly on *Conceptual* tasks and lowly on *Procedural* ones.

It is not clear where exactly the training process of LLMs falls in Bloom's Taxonomy. Many models are trained on next token prediction or masked language modeling, both of which can be considered to fall into the lower levels of the taxonomy. Additional fine-tuning and techniques such as reinforcement learning from human feedback make an exact placement difficult. Overall the entirety of the training process however does not reach into the higher levels of the taxonomy. As the training process has a large impact on the quality of the model this can explain the higher performance of the trained models on the lower levels of the taxonomy.

5 Discussion

In Section 4.2 we have analyzed the coverage of Bloom's Taxonomy by commonly used benchmarks. We show that none of the currently standard benchmarks to evaluate new LLMs cover the *Create* or *Evaluate* dimensions, and that *Remember* is underrepresented with only one task belonging to it. Tasks that cover the missing dimensions are easy to find or to construct, i.e. a simple task of "What is the capital of X?" can be one such task belonging to the *Remember* dimension. We do not claim that no suitable benchmarks exist, but rather that there is a gap in the current way that LLMs are evaluated because these benchmarks are not used consistently when presenting a new LLM. A report on the performance of models on a more diverse set of benchmarks that covers more capabilities can help give researchers a better view on the strengths and weaknesses of these models. Bloom's Taxonomy is a suitable framework for this goal as it is established as a way to categorize learning objectives and design evaluations that effectively target the different dimensions.

We noted a lack of tasks that require *Metacognitive* knowledge in the benchmarks. The GPT-4 family of models rarely suggested this label for a few tasks (refer to Table 14 for details). Tasks that test this kind of knowledge can be very beneficial as they require the models to attempt a more structured, deeper way of reasoning, which can give insight into their inner workings.

To enable a more comprehensive coverage of the taxonomy's levels both the *Create* and *Evaluate* cognitive dimensions as well as the *Metacognitive* knowledge type need to be included in model reports. Zheng et al. (2023) propose MT-Bench, a benchmark consisting of multiple categories, of which one is Writing. We argue that this task covers the *Create* dimension, but it has drawbacks: the scoring metrics do not have an upper bound, and ranking is relative to other models. This allows for a comparison between models but makes it difficult to measure absolute performance. Another suitable benchmark for the *Create* dimension is TuringAdvice (Zellers et al., 2021). Models are asked to generate advice for a given situation, and human annotators mark the advice as helpful or not helpful. Model performance can be measured in terms of their advice being preferred over human advice on the dataset, but the task itself remains open-ended. A model that always gives advice that is preferable over the human advice does not necessarily represent that the task is solved and no better advice can be achieved. These two benchmarks highlight a key difficulty when measuring the performance on the *Create* dimension. It is difficult to find a suitable metric, with clearly defined limits, for tasks of this category. Relative performance on the *Create* dimension can be measured but it remains difficult to measure the absolute performance. The *Evaluate* dimension can be covered by fact-checking tasks such as the recently proposed Factcheck-Bench by Wang et al. (2024). Comparing multiple texts and evaluating whether or not some claim is contained therein, thereby fact-checking it, covers this dimension. Unlike the *Create* dimension tasks, performance on such tasks can be measured in terms of metrics with upper bounds, which makes them suitable as a means to gain a balanced overview of LLM performance. Measuring *Metacognitive* performance can be difficult in the context of current LLMs. They are autoregressive and there is no real cognition to speak of. Nevertheless, one approach can be to ask a model to predict its own performance on a task that it will solve later. We are not aware of any such experiments being published, but argue that the models likely perform quite badly on such a task. This is due to their lack of cognition. Nevertheless, such an experiment is simple to set up and provides an overview of the metacognitive skills of LLMs. As it is out of scope of our work we leave it for future work.

A difference in performance on the various levels of the taxonomy can be observed. As discussed in Section 4.3 we observe a weakness in the higher-order thinking skills, the *Analyze* layer of the taxonomy, as well as a slight weakness in the *Understand* dimension. There is a lack of benchmarks that cover the full spectrum of the taxonomy, with the higher-order thinking skills underrepresented in current benchmarks. We theorize that higher-order thinking requires deeper reasoning skills than tasks that map to the lower levels of the taxonomy. Research on how to improve skills in these areas is still ongoing (Liu et al., 2022; Pan et al., 2023; Ling et al., 2023). To give a more comprehensive overview of current LLM capabilities we suggest that evaluations are extended with more benchmarks that cover the full range of cognitive abilities.

We note that during our analysis of the benchmarks we have found some that we do not agree are suitable as universal benchmarks. Tasks such as simple arithmetic or evaluating Boolean expressions test the capabilities of the models, but can be trivially solved with a rule-based or hard-coded approach. An ensemble model of LLM and a specialized tool that solves these tasks reliably could easily achieve a higher score than an LLM by itself. RAG models (Lewis et al., 2020) for instance can benefit from being evaluated on benchmarks that require external knowledge, but not every LLM is a RAG model. Benchmarks should be used to evaluate a specific skill or capability that the models can have as those benchmarks can reveal research areas where LLM improvements can have a tangible impact.

As discussed in Section 3.1.2 the MMLU benchmark (Hendrycks et al., 2020) consists of multiple subtasks, but individual task instances in each subtask nevertheless map to different levels in the taxonomy. For instance, one task in the `college_biology` subtask may require the recalling of simple facts (*Remember* dimension in the taxonomy), while another in the same subtask requires complex reasoning about specific scenarios (*Analyze / Evaluate* in the taxonomy). Using this

benchmark in an evaluation gives a distorted view on model performance for this reason. A model may perform highly on cognitively simple questions but fail on more complex ones. An evaluation purely on the entirety of a benchmark masks this, and, in rare cases such as MMLU, this is an issue even if the evaluation is done on a subtask level. This highlights the need for a shift in how we view model evaluations. The current approaches give only a distorted view which limits discussion around capabilities of current models and how to improve them.

Lastly a comparison to human performance on benchmarks can provide valuable insights. For the benchmarks that we have analyzed none of them readily had human results available. Comparing human performance, mapped to Bloom's Taxonomy, and comparing it to LLM performance could help identify gaps and strengths of these models compared to human capabilities. We hypothesize that certain tasks may be simple to solve for a human but LLMs may struggle on them and vice versa.

6 Related Work

Recent surveys provide an overview benchmarks and LLM evaluations (Chang et al., 2024; Zhao et al., 2023b). The recent Chatbot Arena Leaderboard (Chiang et al., 2024) uses crowdsourcing to compare different models. None of the popular benchmarks in the field of NLP include information about what cognitive skills they test, e.g. the coverage in Bloom's Taxonomy. A limited number of very recent studies have explored applying the taxonomy in the broader computer science domain. Shojaee et al. (2024) investigate the performance of LLMs in answering neurophysiology questions across two cognitive levels (lower-order and higher-order). They found no significant difference in LLM performance between cognitive levels. Herrmann-Werner et al. (2024) explore GPT-4's ability to answer psychosomatic medicine exam questions. A qualitative analysis using Bloom's Taxonomy revealed that errors predominantly occurred at the *Remember* and *Understand* cognitive levels. Fei et al. (2023) propose LawBench, a benchmark designed to evaluate the capabilities of LLMs across three cognitive levels: legal knowledge memorization, legal knowledge understanding, and legal knowledge application, which correspond to the three lower levels of Bloom's Taxonomy. However, all the approaches mentioned above are limited to specific domains.

To the best of our knowledge, the most comprehensive evaluation of LLMs based on Bloom's Taxonomy to date is presented by Zhang (2023) and Sun et al. (2024). With the objective of revealing the knowledge structures of LLMs, Zhang (2023) use the educational diagnostic assessment method (Bejar, 1984). To assess the cognitive capabilities of LLMs, they analyzed their performance and error patterns on MoocRadar (Yu et al., 2023), a student exercise dataset annotated with Bloom's Taxonomy. The results show that LLMs tend to struggle with tasks in the intermediate range the taxonomy, suggesting potential limitations in their reasoning and problem-solving abilities. Sun et al. (2024) present SciEval, a benchmark for scientific research ability evaluation of LLMs. Leveraging the cognitive levels of Bloom's Taxonomy, SciEval evaluates the capabilities of LLMs across four dimensions: basic knowledge, knowledge application, scientific calculation, and research ability. Similar to the findings by Zhang (2023), Sun et al. (2024) reveal that LLMs underperform in the scientific calculation domain, which aligns with the intermediate level of Bloom's Taxonomy, while demonstrating relatively superior performance in the other three domains. Our work differs from these previous works in that we map existing benchmarks to Bloom's Taxonomy and evaluate the coverage that current evaluation approaches and popular benchmarks achieve on the taxonomy.

7 Conclusion

We presented an analysis of the current LLM evaluation approaches by mapping commonly used benchmarks to Bloom's Taxonomy to identify the cognitive abilities that they cover. By doing so we identified gaps in the evaluation of LLMs. Currently used benchmarks do not sufficiently cover all levels of the taxonomy. Identifying the cognitive skills that the benchmarks cover allows for a deeper analysis of LLM strengths and weaknesses and can drive research to improve these models. The models we have considered show weaknesses towards the higher-order thinking skills in Bloom's Taxonomy. LLM evaluations should focus on a balanced selection of benchmarks that sufficiently covers the full range of cognitive skills to allow the research community to gain a focused view on the performance of the models.

8 Limitations

Bloom's Taxonomy was originally designed for classifying questions in an educational setting, such as exams. It is not clear whether it can be applied to LLM benchmarks without changes, but the high agreement scores between annotators suggests that this approach is feasible. The annotators do not have a background in pedagogy. The high agreements between annotators and LLMs suggest that the annotations are reliable but expert annotators might assign different labels on some benchmarks. One potential weakness with applying Bloom's Taxonomy we have identified is that tasks in the same benchmark, despite clearly mapping to the same cognitive and knowledge dimension levels, can have varying degrees of difficulty. Another factor is the length of the inputs. Longer inputs may not be a factor in a human-based evaluation setting but can potentially lead to the LLMs making mistakes. A recent work by Liu et al. (2024) investigates how LLMs use longer contexts. A more in-depth analysis is required as to whether these factors, length and difficulty, need to be incorporated into the taxonomy when applying it to the setting of LLM evaluation.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

AI@Meta. 2024. Llama 3 model card.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

L.W. Anderson and D.R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

Anthropic. 2024. Introducing the next generation of Claude. `https://www.anthropic.com/news/claude-3-family`. [Accessed 03-06-2024].

Isaac I Bejar. 1984. Educational diagnostic assessment. *Journal of educational measurement*, 21(2):175–189.

Benjamin S Bloom, Max D Engelhart, EJ Furst, Walker H Hill, and David R Krathwohl. 1956. Handbook i: cognitive domain. *New York: David McKay*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva,

Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *Preprint*, arXiv:2309.16289.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Anne Herrmann-Werner, Teresa Festl-Wietek, Friederike Holderried, Lea Herschbach, Jan Griewatz, Ken Masters, Stephan Zipfel, and Moritz Mahling. 2024. Assessing chatgpt's mastery of bloom's taxonomy using psychosomatic medicine exam questions: Mixed-methods study. *J Med Internet Res*, 26:e52113.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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

Yuheng Li, Mladen Rakovic, Boon Xin Poh, Dragan Gasevic, and Guanliang Chen. 2022. Automatic classification of learning objectives based on bloom's taxonomy. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 530–537, Durham, United Kingdom. International Educational Data Mining Society.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2365–2376, Dublin, Ireland. Association for Computational Linguistics.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date — ai.meta.com. `https://ai.meta.com/blog/meta-llama-3/`. [Accessed 03-06-2024].

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

OpenAI. 2024. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`. [Accessed 03-06-2024].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Hassan Shojaee, Reza Mohebbati, Mostafa Amiri, and Alireza Atarodi. 2024. Evaluating the strengths and weaknesses of large language models in answering neurophysiology questions. *Scientific Reports*, 14(1).

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19053–19061. AAAI Press.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tac-

chetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024a. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024b. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra,

Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

X.ai. 2024. Grok-2 Beta Release. `https://x.ai/blog/grok-2`. [Accessed 15-08-2024].

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Jifan Yu, Mengying Lu, Qingyang Zhong, Zijun Yao, Shangqing Tu, Zhengshan Liao, Xiaoya Li, Manli Li, Lei Hou, Hai-Tao Zheng, Juanzi Li, and Jie Tang. 2023. Moocradar: A fine-grained and multi-aspect knowledge repository for improving cognitive student modeling in moocs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2924–2934, New York, NY, USA. Association for Computing Machinery.

Rowan Zellers, Ari Holtzman, Elizabeth Clark, Lianhui Qin, Ali Farhadi, and Yejin Choi. 2021. TuringAdvice: A generative and dynamic evaluation of language use. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4856–4880, Online. Association for Computational Linguistics.

Jia Zhang. 2023. Exploring undergraduate translation students' perceptions towards machine translation: A qualitative questionnaire survey. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 1–10, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Xuan Zhang and Wei Gao. 2023. Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1011, Nusa Dua, Bali. Association for Computational Linguistics.

Zheyuan Zhang, Jifan Yu, Juanzi Li, and Lei Hou. 2023. Exploring the cognitive knowledge structure of large language models: An educational diagnostic assessment approach. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1643–1650, Singapore. Association for Computational Linguistics.

James Zhao, Yuxi Xie, Kenji Kawaguchi, Junxian He, and Michael Xie. 2023a. Automatic model selection with large language models for reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 758–783, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models. *CoRR*, abs/2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Li Zhong and Zilong Wang. 2024. Can LLM replace stack overflow? A study on robustness and reliability of large language model code generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 21841–21849. AAAI Press.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A Individual Task Results

We include the the scores used as the basis for our analysis in Table 3 (GPT-4), Table 4 (Llama 3), Table 5 (Claude 3), Table 6 (bloomz-3b), Table 7 (bloomz-560m), Table 8 (falcon-7b-instruct), Table 9 (falcon-40b-instruct), Table 10 (flan-t5-xxl), Table 11 (gemma-7b-it) and Table 12 (phi3-mini-4k-instruct).

All prompts used a chain-of-thought format and were zero-shot. Evaluations were performed automatically using a `meta-llama/Meta-Llama-3-70B-Instruct` instance and the prompt described in Figure 5. When we used scores from other publications we have added a reference to the publication to the table. AGIEval and BIG-Bench Hard results are available as an aggregated score but we were unable to find per-task results on all models we considered for this work. For this reason we have manually collected the data and publish the results below.

A.1 GPT-4

For the tasks where results were not available we used the OpenAI API and the model `gpt-4-0613` to calculate the scores.

A.2 Llama 3

We used a self-hosted `meta-llama/Meta-Llama-3-70B-Instruct` model from the HuggingFace repository for the tasks where results were not available.

A.3 Claude 3

We used the Anthropic API and the `claude-3-haiku-20240307` model for the tasks that we were unable to find results for.

A.4 bloomz-3b

We used the `bloomz-3b` model from HuggingFace to evaluate performance on AGIEval and BIG-Bench Hard subtasks. We used a simple CoT zero-shot prompting approach for both benchmarks.

A.5 bloomz-560m

We used the `bloomz-560m` model from HuggingFace to evaluate performance on AGIEval and BIG-Bench Hard subtasks. We used a simple CoT zero-shot prompting approach for both benchmarks.

A.6 falcon-7b-instruct

We used the `falcon-7b-instruct` model from HuggingFace to evaluate performance on AGIEval and BIG-Bench Hard subtasks. We used a simple CoT zero-shot prompting approach for both benchmarks.

A.7 falcon-40b-instruct

We used the `falcon-40b-instruct` model from HuggingFace to evaluate performance on AGIEval and BIG-Bench Hard subtasks. We used a simple CoT zero-shot prompting approach for both benchmarks.

A.8 flan-t5-xxl

We used the `flan-t5-xxl` model from HuggingFace to evaluate performance on AGIEval and BIG-Bench Hard subtasks. We used a simple CoT zero-shot prompting approach for both benchmarks.

A.9 gemma-7b-it

We used the `flan-t5-xxl` model from HuggingFace to evaluate performance on AGIEval and BIG-Bench Hard subtasks. We used a simple CoT zero-shot prompting approach for both benchmarks.

A.10 phi3-mini-4k-instruct

We used the `flan-t5-xxl` model from HuggingFace to evaluate performance on AGIEval and BIG-Bench Hard subtasks. We used a simple CoT zero-shot prompting approach for both benchmarks.

Model	Benchmark	Subtask	Score	Source
gpt4	AGIEval	aqua-rat	0.740000	https://github.com/microsoft/promptbase
gpt4	AGIEval	gaokao-english	0.931000	(Zhong et al., 2023)
gpt4	AGIEval	logiqa	0.627000	(Zhong et al., 2023)
gpt4	AGIEval	lsat-ar	0.344000	(Zhong et al., 2023)
gpt4	AGIEval	lsat-lr	0.845000	(Zhong et al., 2023)
gpt4	AGIEval	lsat-rc	0.877000	(Zhong et al., 2023)
gpt4	AGIEval	math	0.950000	(Zhong et al., 2023)
gpt4	AGIEval	sat-en	0.859000	(Zhong et al., 2023)
gpt4	AGIEval	sat-math	0.896000	(Zhong et al., 2023)
gpt4	ARC-Challenge	N/A	0.963000	(Zhong et al., 2023)
gpt4	BIG-Bench Hard	boolean_expressions	0.936000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	causal_judgement	0.716578	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	date_understanding	0.956000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	disambiguation_qa	0.876000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	dyck_languages	0.852000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	formal_fallacies	0.816000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	geometric_shapes	0.652000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	hyperbaton	0.972000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	logical_deduction_five_objects	0.804000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	logical_deduction_seven_objects	0.636000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	logical_deduction_three_objects	0.956000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	movie_recommendation	0.908000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	multistep_arithmetic_two	0.892000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	navigate	0.968000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	object_counting	0.992000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	penguins_in_a_table	0.986301	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	reasoning_about_colored_objects	0.968000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	ruin_names	0.904000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	salient_translation_error_detection	0.484000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	snarks	0.988764	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	sports_understanding	0.988000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	temporal_sequences	1.000000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	tracking_shuffled_objects_five_objects	1.000000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	1.000000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	tracking_shuffled_objects_three_objects	1.000000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	web_of_lies	1.000000	Own evaluation, CoT zero-shot prompting
gpt4	BIG-Bench Hard	word_sorting	0.996000	Own evaluation, CoT zero-shot prompting
gpt4	DROP	N/A	0.834000	https://github.com/openai/simple-evals
gpt4	GPQA	N/A	0.414000	https://github.com/openai/simple-evals
gpt4	GSM8K	N/A	0.956000	(Zhao et al., 2023a)
gpt4	HumanEval	N/A	0.882000	https://github.com/openai/simple-evals
gpt4	MATH	N/A	0.684000	https://github.com/microsoft/promptbase
gpt4	Winogrande	N/A	0.875000	(OpenAI et al., 2024)

Table 3: Benchmark scores for GPT-4

Model	Benchmark	Subtask	Score	Source
llama3	AGIEval	aqua-rat	0.767700	Own evaluation, CoT zero-shot prompting
llama3	AGIEval	gaokao-english	0.954200	Own evaluation, CoT zero-shot prompting
llama3	AGIEval	logiqa	0.735800	Own evaluation, CoT zero-shot prompting
llama3	AGIEval	lsat-ar	0.461000	Own evaluation, CoT zero-shot prompting
llama3	AGIEval	lsat-lr	0.892200	Own evaluation, CoT zero-shot prompting
llama3	AGIEval	lsat-rc	0.965400	Own evaluation, CoT zero-shot prompting
llama3	AGIEval	math	0.497000	Own evaluation, CoT zero-shot prompting
llama3	AGIEval	sat-en	0.966000	Own evaluation, CoT zero-shot prompting
llama3	AGIEval	sat-math	0.922700	Own evaluation, CoT zero-shot prompting
llama3	ARC-Challenge	N/A	0.930000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	boolean_expressions	0.764000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	causal_judgement	0.058824	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	date_understanding	0.044000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	disambiguation_qa	0.292000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	dyck_languages	0.348000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	formal_fallacies	0.068000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	geometric_shapes	0.392000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	hyperbaton	0.216000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	logical_deduction_five_objects	0.012000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	logical_deduction_seven_objects	0.660000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	logical_deduction_three_objects	0.980000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	movie_recommendation	0.808000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	multistep_arithmetic_two	0.852000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	navigate	0.956000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	object_counting	0.900000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	penguins_in_a_table	0.198630	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	reasoning_about_colored_objects	0.944000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	ruin_names	0.912000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	salient_translation_error_detection	0.800000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	snarks	0.219101	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	sports_understanding	0.960000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	temporal_sequences	1.000000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.992000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.984000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.996000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	web_of_lies	1.000000	Own evaluation, CoT zero-shot prompting
llama3	BIG-Bench Hard	word_sorting	0.988000	Own evaluation, CoT zero-shot prompting
llama3	DROP	N/A	0.797000	https://github.com/openai/simple-evals
llama3	GPQA	N/A	0.395000	https://github.com/openai/simple-evals
llama3	GSM8K	N/A	0.930000	(Meta, 2024)
llama3	HumanEval	N/A	0.817000	https://github.com/openai/simple-evals
llama3	MATH	N/A	0.504000	https://github.com/openai/simple-evals
llama3	Winogrande	N/A	0.831000	(AI@Meta, 2024)

Table 4: Benchmark scores for Llama 3

Model	Benchmark	Subtask	Score	Source
claude3	AGIEval	aqua-rat	0.633900	Own evaluation, CoT zero-shot prompting
claude3	AGIEval	gaokao-english	0.846400	Own evaluation, CoT zero-shot prompting
claude3	AGIEval	logiqa	0.411700	Own evaluation, CoT zero-shot prompting
claude3	AGIEval	lsat-ar	0.226100	Own evaluation, CoT zero-shot prompting
claude3	AGIEval	lsat-lr	0.482300	Own evaluation, CoT zero-shot prompting
claude3	AGIEval	lsat-rc	0.703100	Own evaluation, CoT zero-shot prompting
claude3	AGIEval	math	0.413000	Own evaluation, CoT zero-shot prompting
claude3	AGIEval	sat-en	0.873800	Own evaluation, CoT zero-shot prompting
claude3	AGIEval	sat-math	0.772700	Own evaluation, CoT zero-shot prompting
claude3	ARC-Challenge	N/A	0.892000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	boolean_expressions	0.324000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	causal_judgement	0.267380	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	date_understanding	0.712000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	disambiguation_qa	0.008000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	dyck_languages	0.232000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	formal_fallacies	0.596000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	geometric_shapes	0.560000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	hyperbaton	0.056000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	logical_deduction_five_objects	0.592000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	logical_deduction_seven_objects	0.492000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	logical_deduction_three_objects	0.896000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	movie_recommendation	0.384000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	multistep_arithmetic_two	0.820000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	navigate	0.960000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	object_counting	0.464000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	penguins_in_a_table	0.280822	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	reasoning_about_colored_objects	0.024000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	ruin_names	0.788000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	salient_translation_error_detection	0.780000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	snarks	0.089888	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	sports_understanding	0.012000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	temporal_sequences	0.680000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.712000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.708000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.732000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	web_of_lies	0.996000	Own evaluation, CoT zero-shot prompting
claude3	BIG-Bench Hard	word_sorting	0.904000	Own evaluation, CoT zero-shot prompting
claude3	DROP	N/A	0.784000	(Anthropic, 2024)
claude3	GPQA	N/A	0.333000	(Anthropic, 2024)
claude3	GSM8K	N/A	0.889000	(Anthropic, 2024)
claude3	HumanEval	N/A	0.759000	(Anthropic, 2024)
claude3	MATH	N/A	0.389000	(Anthropic, 2024)
claude3	Winogrande	N/A	NaN	Results not publicly available. We omitted this score.

Table 5: Benchmark scores for Claude 3

Model	Benchmark	Subtask	Score	Source
bloomz-3b	AGIEval	aqua-rat	0.192913	Own evaluation, zero-shot prompting
bloomz-3b	AGIEval	gaokao-english	0.663399	Own evaluation, zero-shot prompting
bloomz-3b	AGIEval	logiqa	0.284178	Own evaluation, zero-shot prompting
bloomz-3b	AGIEval	lsat-ar	0.213043	Own evaluation, zero-shot prompting
bloomz-3b	AGIEval	lsat-lr	0.254902	Own evaluation, zero-shot prompting
bloomz-3b	AGIEval	lsat-rc	0.327138	Own evaluation, zero-shot prompting
bloomz-3b	AGIEval	math	0.055000	Own evaluation, zero-shot prompting
bloomz-3b	AGIEval	sat-en	0.456311	Own evaluation, zero-shot prompting
bloomz-3b	AGIEval	sat-math	0.231818	Own evaluation, zero-shot prompting
bloomz-3b	ARC-Challenge	N/A	N/A	N/A
bloomz-3b	BIG-Bench Hard	boolean_expressions	0.412000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	causal_judgement	0.518717	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	date_understanding	0.184000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	disambiguation_qa	0.364000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	dyck_languages	0.032000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	formal_fallacies	0.488000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	geometric_shapes	0.484000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	hyperbaton	0.376000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	logical_deduction_five_objects	0.244000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	logical_deduction_seven_objects	0.172000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	logical_deduction_three_objects	0.348000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	movie_recommendation	0.160000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	multistep_arithmetic_two	0.004000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	navigate	0.492000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	object_counting	0.024000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	penguins_in_a_table	0.280822	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	reasoning_about_colored_objects	0.168000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	ruin_names	0.168000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	salient_translation_error_detection	0.296000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	snarks	0.314607	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	sports_understanding	0.524000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	temporal_sequences	0.308000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.152000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.136000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.304000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	web_of_lies	0.808000	Own evaluation, zero-shot prompting
bloomz-3b	BIG-Bench Hard	word_sorting	0.000000	Own evaluation, zero-shot prompting
bloomz-3b	DROP	N/A	N/A	N/A
bloomz-3b	GPQA	N/A	N/A	N/A
bloomz-3b	GSM8K	N/A	N/A	N/A
bloomz-3b	HumanEval	N/A	N/A	N/A
bloomz-3b	MATH	N/A	N/A	N/A
bloomz-3b	Winogrande	N/A	N/A	N/A

Table 6: Benchmark scores for bloomz-3b

Model	Benchmark	Subtask	Score	Source
bloomz-560m	AGIEval	aqua-rat	0.157480	Own evaluation, zero-shot prompting
bloomz-560m	AGIEval	gaokao-english	0.254902	Own evaluation, zero-shot prompting
bloomz-560m	AGIEval	logiqa	0.291859	Own evaluation, zero-shot prompting
bloomz-560m	AGIEval	lsat-ar	0.195652	Own evaluation, zero-shot prompting
bloomz-560m	AGIEval	lsat-lr	0.209804	Own evaluation, zero-shot prompting
bloomz-560m	AGIEval	lsat-rc	0.197026	Own evaluation, zero-shot prompting
bloomz-560m	AGIEval	math	0.035000	Own evaluation, zero-shot prompting
bloomz-560m	AGIEval	sat-en	0.276699	Own evaluation, zero-shot prompting
bloomz-560m	AGIEval	sat-math	0.177273	Own evaluation, zero-shot prompting
bloomz-560m	ARC-Challenge	N/A	N/A	N/A
bloomz-560m	BIG-Bench Hard	boolean_expressions	0.444000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	causal_judgement	0.534759	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	date_understanding	0.172000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	disambiguation_qa	0.144000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	dyck_languages	0.100000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	formal_fallacies	0.444000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	geometric_shapes	0.348000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	hyperbaton	0.236000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	logical_deduction_five_objects	0.184000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	logical_deduction_seven_objects	0.120000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	logical_deduction_three_objects	0.316000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	movie_recommendation	0.044000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	multistep_arithmetic_two	0.004000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	navigate	0.544000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	object_counting	0.008000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	penguins_in_a_table	0.227586	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	reasoning_about_colored_objects	0.137652	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	ruin_names	0.144000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	salient_translation_error_detection	0.823293	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	snarks	0.252809	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	sports_understanding	0.556000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	temporal_sequences	0.232000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.184000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.160000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.288000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	web_of_lies	0.484000	Own evaluation, zero-shot prompting
bloomz-560m	BIG-Bench Hard	word_sorting	0.000000	Own evaluation, zero-shot prompting
bloomz-560m	DROP	N/A	N/A	N/A
bloomz-560m	GPQA	N/A	N/A	N/A
bloomz-560m	GSM8K	N/A	N/A	N/A
bloomz-560m	HumanEval	N/A	N/A	N/A
bloomz-560m	MATH	N/A	N/A	N/A
bloomz-560m	Winogrande	N/A	N/A	N/A

Table 7: Benchmark scores for bloomz-560m

Model	Benchmark	Subtask	Score	Source
falcon-7b-instruct	AGIEval	aqua-rat	0.059055	Own evaluation, zero-shot prompting
falcon-7b-instruct	AGIEval	gaokao-english	0.432787	Own evaluation, zero-shot prompting
falcon-7b-instruct	AGIEval	logiqa	0.284178	Own evaluation, zero-shot prompting
falcon-7b-instruct	AGIEval	lsat-ar	0.256522	Own evaluation, zero-shot prompting
falcon-7b-instruct	AGIEval	lsat-lr	0.329412	Own evaluation, zero-shot prompting
falcon-7b-instruct	AGIEval	lsat-rc	0.567164	Own evaluation, zero-shot prompting
falcon-7b-instruct	AGIEval	math	0.020000	Own evaluation, zero-shot prompting
falcon-7b-instruct	AGIEval	sat-en	0.617647	Own evaluation, zero-shot prompting
falcon-7b-instruct	AGIEval	sat-math	0.142202	Own evaluation, zero-shot prompting
falcon-7b-instruct	ARC-Challenge	N/A	N/A	N/A
falcon-7b-instruct	BIG-Bench Hard	boolean_expressions	0.308000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	causal_judgement	0.208556	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	date_understanding	0.072000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	disambiguation_qa	0.224000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	dyck_languages	0.004000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	formal_fallacies	0.072000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	geometric_shapes	0.684000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	hyperbaton	0.300000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	logical_deduction_five_objects	0.064000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	logical_deduction_seven_objects	0.020000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	logical_deduction_three_objects	0.064000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	movie_recommendation	0.084000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	multistep_arithmetic_two	0.000000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	navigate	0.292000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	object_counting	0.108000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	penguins_in_a_table	0.000000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	reasoning_about_colored_objects	0.028000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	ruin_names	0.108000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	salient_translation_error_detection	0.020000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	snarks	0.426966	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	sports_understanding	0.540000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	temporal_sequences	0.084000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.016000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.056000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.044000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	web_of_lies	0.164000	Own evaluation, zero-shot prompting
falcon-7b-instruct	BIG-Bench Hard	word_sorting	0.004000	Own evaluation, zero-shot prompting
falcon-7b-instruct	DROP	N/A	N/A	N/A
falcon-7b-instruct	GPQA	N/A	N/A	N/A
falcon-7b-instruct	GSM8K	N/A	N/A	N/A
falcon-7b-instruct	HumanEval	N/A	N/A	N/A
falcon-7b-instruct	MATH	N/A	N/A	N/A
falcon-7b-instruct	Winogrande	N/A	N/A	N/A

Table 8: Benchmark scores for falcon-7b-instruct

Model	Benchmark	Subtask	Score	Source
falcon-40b-instruct	AGIEval	aqua-rat	0.137795	Own evaluation, zero-shot prompting
falcon-40b-instruct	AGIEval	gaokao-english	0.473684	Own evaluation, zero-shot prompting
falcon-40b-instruct	AGIEval	logiqa	0.405530	Own evaluation, zero-shot prompting
falcon-40b-instruct	AGIEval	lsat-ar	0.573913	Own evaluation, zero-shot prompting
falcon-40b-instruct	AGIEval	lsat-lr	0.564706	Own evaluation, zero-shot prompting
falcon-40b-instruct	AGIEval	lsat-rc	0.477612	Own evaluation, zero-shot prompting
falcon-40b-instruct	AGIEval	math	0.043043	Own evaluation, zero-shot prompting
falcon-40b-instruct	AGIEval	sat-en	0.446602	Own evaluation, zero-shot prompting
falcon-40b-instruct	AGIEval	sat-math	0.182648	Own evaluation, zero-shot prompting
falcon-40b-instruct	ARC-Challenge	N/A	N/A	N/A
falcon-40b-instruct	BIG-Bench Hard	boolean_expressions	0.516000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	causal_judgement	0.417112	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	date_understanding	0.308000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	disambiguation_qa	0.540000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	dyck_languages	0.060000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	formal_fallacies	0.456000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	geometric_shapes	0.684000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	hyperbaton	0.448000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	logical_deduction_five_objects	0.248000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	logical_deduction_seven_objects	0.172000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	logical_deduction_three_objects	0.332000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	movie_recommendation	0.292000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	multistep_arithmetic_two	0.092000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	navigate	0.524000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	object_counting	0.504000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	penguins_in_a_table	0.178082	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	reasoning_about_colored_objects	0.232000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	ruin_names	0.348000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	salient_translation_error_detection	0.176000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	snarks	0.747191	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	sports_understanding	0.844000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	temporal_sequences	0.120000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.088000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.072000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.268000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	web_of_lies	0.636000	Own evaluation, zero-shot prompting
falcon-40b-instruct	BIG-Bench Hard	word_sorting	0.052000	Own evaluation, zero-shot prompting
falcon-40b-instruct	DROP	N/A	N/A	N/A
falcon-40b-instruct	GPQA	N/A	N/A	N/A
falcon-40b-instruct	GSM8K	N/A	N/A	N/A
falcon-40b-instruct	HumanEval	N/A	N/A	N/A
falcon-40b-instruct	MATH	N/A	N/A	N/A
falcon-40b-instruct	Winogrande	N/A	N/A	N/A

Table 9: Benchmark scores for falcon-40b-instruct

Model	Benchmark	Subtask	Score	Source
flan-t5-xxl	AGIEval	aqua-rat	0.208661	Own evaluation, zero-shot prompting
flan-t5-xxl	AGIEval	gaokao-english	0.908497	Own evaluation, zero-shot prompting
flan-t5-xxl	AGIEval	logiqa	0.407066	Own evaluation, zero-shot prompting
flan-t5-xxl	AGIEval	lsat-ar	0.230435	Own evaluation, zero-shot prompting
flan-t5-xxl	AGIEval	lsat-lr	0.533333	Own evaluation, zero-shot prompting
flan-t5-xxl	AGIEval	lsat-rc	0.747212	Own evaluation, zero-shot prompting
flan-t5-xxl	AGIEval	math	0.043000	Own evaluation, zero-shot prompting
flan-t5-xxl	AGIEval	sat-en	0.810680	Own evaluation, zero-shot prompting
flan-t5-xxl	AGIEval	sat-math	0.209091	Own evaluation, zero-shot prompting
flan-t5-xxl	ARC-Challenge	N/A	N/A	N/A
flan-t5-xxl	BIG-Bench Hard	boolean_expressions	0.576000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	causal_judgement	0.582888	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	date_understanding	0.640000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	disambiguation_qa	0.624000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	dyck_languages	0.096000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	formal_fallacies	0.444000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	geometric_shapes	0.236000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	hyperbaton	0.676000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	logical_deduction_five_objects	0.528000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	logical_deduction_seven_objects	0.560000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	logical_deduction_three_objects	0.628000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	movie_recommendation	0.356000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	multistep_arithmetic_two	0.016000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	navigate	0.512000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	object_counting	0.472000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	penguins_in_a_table	0.404110	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	reasoning_about_colored_objects	0.544000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	ruin_names	0.228000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	salient_translation_error_detection	0.224000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	snarks	0.393258	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	sports_understanding	0.580000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	temporal_sequences	0.284000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.156000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.172000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.284000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	web_of_lies	0.576000	Own evaluation, zero-shot prompting
flan-t5-xxl	BIG-Bench Hard	word_sorting	0.096000	Own evaluation, zero-shot prompting
flan-t5-xxl	DROP	N/A	N/A	N/A
flan-t5-xxl	GPQA	N/A	N/A	N/A
flan-t5-xxl	GSM8K	N/A	N/A	N/A
flan-t5-xxl	HumanEval	N/A	N/A	N/A
flan-t5-xxl	MATH	N/A	N/A	N/A
flan-t5-xxl	Winogrande	N/A	N/A	N/A

Table 10: Benchmark scores for flan-t5-xxl

Model	Benchmark	Subtask	Score	Source
gemma-7b-it	AGIEval	aqua-rat	0.263780	Own evaluation, zero-shot prompting
gemma-7b-it	AGIEval	gaokao-english	0.686275	Own evaluation, zero-shot prompting
gemma-7b-it	AGIEval	logiqa	0.328725	Own evaluation, zero-shot prompting
gemma-7b-it	AGIEval	lsat-ar	0.173913	Own evaluation, zero-shot prompting
gemma-7b-it	AGIEval	lsat-lr	0.362745	Own evaluation, zero-shot prompting
gemma-7b-it	AGIEval	lsat-rc	0.531599	Own evaluation, zero-shot prompting
gemma-7b-it	AGIEval	math	0.149000	Own evaluation, zero-shot prompting
gemma-7b-it	AGIEval	sat-en	0.626214	Own evaluation, zero-shot prompting
gemma-7b-it	AGIEval	sat-math	0.377273	Own evaluation, zero-shot prompting
gemma-7b-it	ARC-Challenge	N/A	N/A	N/A
gemma-7b-it	BIG-Bench Hard	boolean_expressions	0.348000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	causal_judgement	0.347594	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	date_understanding	0.156000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	disambiguation_qa	0.392000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	dyck_languages	0.296000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	formal_fallacies	0.264000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	geometric_shapes	0.260000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	hyperbaton	0.440000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	logical_deduction_five_objects	0.172000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	logical_deduction_seven_objects	0.136000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	logical_deduction_three_objects	0.308000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	movie_recommendation	0.260000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	multistep_arithmetic_two	0.044000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	navigate	0.468000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	object_counting	0.312000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	penguins_in_a_table	0.321918	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	reasoning_about_colored_objects	0.192000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	ruin_names	0.092000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	salient_translation_error_detection	0.148000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	snarks	0.505618	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	sports_understanding	0.392000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	temporal_sequences	0.092000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.080000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.060000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.152000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	web_of_lies	0.304000	Own evaluation, zero-shot prompting
gemma-7b-it	BIG-Bench Hard	word_sorting	0.024000	Own evaluation, zero-shot prompting
gemma-7b-it	DROP	N/A	N/A	N/A
gemma-7b-it	GPQA	N/A	N/A	N/A
gemma-7b-it	GSM8K	N/A	N/A	N/A
gemma-7b-it	HumanEval	N/A	N/A	N/A
gemma-7b-it	MATH	N/A	N/A	N/A
gemma-7b-it	Winogrande	N/A	N/A	N/A

Table 11: Benchmark scores for gemma-7b-it

Model	Benchmark	Subtask	Score	Source
phi3-mini-4k-instruct	AGIEval	aqua-rat	0.645669	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	AGIEval	gaokao-english	0.787582	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	AGIEval	logiqa	0.525346	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	AGIEval	lsat-ar	0.286957	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	AGIEval	lsat-lr	0.545098	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	AGIEval	lsat-rc	0.669145	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	AGIEval	math	0.404000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	AGIEval	sat-en	0.718447	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	AGIEval	sat-math	0.718182	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	ARC-Challenge	N/A	N/A	N/A
phi3-mini-4k-instruct	BIG-Bench Hard	boolean_expressions	0.908000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	causal_judgement	0.636364	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	date_understanding	0.632000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	disambiguation_qa	0.752000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	dyck_languages	0.424000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	formal_fallacies	0.632000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	geometric_shapes	0.412000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	hyperbaton	0.840000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	logical_deduction_five_objects	0.552000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	logical_deduction_seven_objects	0.448000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	logical_deduction_three_objects	0.836000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	movie_recommendation	0.520000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	multistep_arithmetic_two	0.672000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	navigate	0.812000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	object_counting	0.752000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	penguins_in_a_table	0.869863	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	reasoning_about_colored_objects	0.828000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	ruin_names	0.596000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	salient_translation_error_detection	0.568000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	snarks	0.870787	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	sports_understanding	0.808000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	temporal_sequences	0.616000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	tracking_shuffled_objects_five_objects	0.928000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	tracking_shuffled_objects_seven_objects	0.876000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	tracking_shuffled_objects_three_objects	0.932000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	web_of_lies	0.976000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	BIG-Bench Hard	word_sorting	0.500000	Own evaluation, zero-shot prompting
phi3-mini-4k-instruct	DROP	N/A	N/A	N/A
phi3-mini-4k-instruct	GPQA	N/A	N/A	N/A
phi3-mini-4k-instruct	GSM8K	N/A	N/A	N/A
phi3-mini-4k-instruct	HumanEval	N/A	N/A	N/A
phi3-mini-4k-instruct	MATH	N/A	N/A	N/A
phi3-mini-4k-instruct	Winogrande	N/A	N/A	N/A

Table 12: Benchmark scores for phi3-mini-4k-instruct

B Revised Bloom's Taxonomy

Building on Bloom's original taxonomy (Bloom et al., 1956), Anderson and Krathwohl (2001) introduced a revised framework that incorporated two key dimensions: knowledge and cognitive processes. This transition from a singular focus on learning outcomes to a two-dimensional model allows for a more nuanced understanding of learning objectives. The knowledge dimension, displayed on the vertical axis, includes four types: factual, conceptual, procedural, and metacognitive knowledge (see Section B.2). The horizontal axis represents the cognitive process dimension, which comprises six levels: remembering, understanding, applying, analyzing, evaluating, and creating (see Section B.1). This intersection creates a matrix with 24 distinct cells, each representing a specific learning objective that integrates a type of knowledge with a cognitive process.[15]

B.1 Cognitive Dimensions

The cognitive dimension of the Revised Bloom's Taxonomy (Anderson and Krathwohl, 2001) classifies learning objectives into six hierarchical levels of cognitive complexity. These levels range from lower-order thinking skills, including *remembering*, *understanding*, and *applying*, to higher-order thinking skills that involve *analysing*, *evaluating*, and *creating*. This hierarchical structure emphasizes the progressive nature of learning, where each level builds upon the preceding one, requiring a deeper level of cognitive processing.

B.1.1 Remember

Within Bloom's Taxonomy's hierarchical framework, the lowest level, *remembering*, represents the ability to recognize and recall previously learned facts. This fundamental skill serves as a prerequisite for comprehension, as students cannot grasp a concept without first possessing the relevant factual knowledge.

Example 1 *List common food options available on campus (dining halls, vending machines, cafes).*[16] This task requires recalling factual information about the existing food system on campus.

Example 2 *Recall the environmental impact of food production (e.g., water usage, carbon footprint).* This task focuses on retrieving previously learned knowledge about the environmental consequences of food production.

B.1.2 Understand

The next level is *understanding*. It builds upon the foundation of knowledge retrieval. At this stage, students transcend simple recall by demonstrating understanding. This encompasses the ability to interpret the learned information, translate it into their own words, and effectively summarize key concepts.

Example 1 *Explain the concept of a "local food system" and its benefits for sustainability.* This task moves beyond simple recall, requiring an explanation of the concept and its connection to sustainability.

Example 2 *Describe the connection between food choices and personal health.* Here, the task involves understanding the relationship between diet and its impact on individual health.

B.1.3 Apply

The transition from comprehension to application is marked by students' ability to utilize acquired facts, ideas, and concepts in new contexts.

Example 1 *Identify features of food options on campus that promote or hinder sustainability (e.g., locally sourced ingredients, packaging).* This task requires applying your knowledge of sustainability to analyze existing food options and their environmental impact.

Example 2 *Apply your knowledge of sustainability to analyze your own eating habits on campus.* This application involves using your understanding of sustainability to assess your personal choices within the campus food system.

B.1.4 Analyze

Building upon the application of knowledge, the analysis level delves deeper by deconstructing concepts into their constituent components. *Analyzing* requires critical thinking skills to identify the relationships and interactions between these elements, enabling students to detect connections, draw inferences, and make attributions of cause and effect.

Example 1 *Analyze the strengths and weaknesses of different food options based on their environmental and health impact.* This task involves breaking

[15] In this work, we make use of the revised version of Bloom's Taxonomy by Anderson and Krathwohl (2001).

[16] The running example was created with the help of ChatGPT.

down different options, considering their environmental and health effects, and identifying both positive and negative aspects.

Example 2 *Compare meal plans offered by your campus and assess their sustainability practices.* Here, the analysis involves comparing different options (meal plans) and assessing them based on their commitment to sustainable practices.

B.1.5 Evaluate

The evaluation level transcends analysis by prompting students to make critical judgments about the presented concepts. This necessitates the application of established criteria and standards to assess the validity, usefulness, or effectiveness of the learned information, allowing students to defend or critique these concepts with justification.

Example *Evaluate the overall sustainability of your campus food system. Consider factors like waste generation, access to healthy options, and student preferences.* This task moves beyond analysis, requiring a judgment on the overall effectiveness of the campus food system in terms of sustainability. This evaluation involves weighing different factors (waste, health, preference) to form a well-rounded judgment about the campus food system.

B.1.6 Create

The *Create* layer represents the top tier of the Revised Bloom's Taxonomy. At this stage, students demonstrate their knowledge by applying the learned concepts to generate new ideas, products, or processes. Thus, this level exemplifies the transformation of learned concepts into meaningful and potentially original outcomes.

Example 1 *Develop a plan to personally adopt more sustainable food choices on campus.* This task moves beyond evaluation and requires the creation of a personalized plan to improve your own food choices on campus based on sustainability principles.

Example 2 *Create a campaign to raise awareness about sustainable food options and encourage positive change within the campus community.* Here, the highest level of Bloom's Taxonomy is reached. You're taking your understanding and creating a new initiative to promote sustainable food choices throughout the campus.

B.2 Knowledge Types

The revised version of Bloom's Taxonomy (Anderson and Krathwohl, 2001) identifies four levels of knowledge along a continuum of increasing abstraction, moving from concrete *factual* knowledge through *conceptual* and *procedural* knowledge to the highly abstract level of *metacognitive* knowledge.

B.2.1 Factual Knowledge

Factual knowledge is the most basic level of knowledge, which involves memorizing facts, details, and terminology.

Example 1 *List the types of agriculture used to produce commonly available food on campus (e.g., conventional, organic, local).* This task focuses on retrieving specific details and terminology related to food production methods. These are factual elements you should be able to recall.

Example 2 *Identify the different components of a food label and their meaning (e.g., ingredients, nutritional information).* Similar to the previous task, this requires recalling specific details and understanding their meaning within the context of food labels.

B.2.2 Conceptual Knowledge

Conceptual knowledge involves understanding the relationships between facts and ideas. This includes classifications, categories, principles, and generalizations.

Example 1 *Explain the concept of a "food system" and its key components (e.g., production, distribution, consumption, waste).* This task goes beyond factual recall. Here, you need to understand the concept of a food system and its various interconnected parts.

Example 2 *Describe the environmental impact of different food production methods (e.g., water usage, greenhouse gas emissions).* This requires understanding the relationship between different agricultural practices and their environmental consequences. You are explaining the concept, not just listing facts.

B.2.3 Procedural Knowledge

Procedural knowledge is knowing how to do things. It involves processes, methods, techniques, and algorithms.

Example 1 *Identify steps involved in life cycle assessment (LCA) of food products and explain its purpose.* This focuses on the "how" of sustainability. You need to know the specific steps involved in LCA, a process for evaluating the environmental impact of a product.

Example 2 *Research and outline the process of starting a campus garden or participating in a Community Supported Agriculture (CSA) program.* Here, you are not just recalling facts about these initiatives, but understanding the steps involved in setting them up, which is procedural knowledge.

B.2.4 Metacognitive Knowledge

Metacognitive knowledge is knowledge about one's own thinking. It includes strategic knowledge and self-awareness about learning processes.

Example *Compare and contrast different sources of information about sustainable food practices (e.g., scientific journals, news articles, advocacy websites). Evaluate the credibility and potential biases of information found about sustainable food choices.* This task requires reflecting on your knowledge and understanding different information sources. You need to evaluate their strengths and weaknesses, not just their content. You are not only consuming information, but actively analyzing it for credibility and potential biases, demonstrating metacognitive knowledge.

C Dataset Overview

We provide an overview over the used datasets and a description of the tasks contained therein in Table 13. Descriptions for the BIG-Bench Hard tasks are quoted directly from (Suzgun et al., 2023).

D Bloom Classify Prompts

We include the prompt we used to classify the cognitive dimension in Figure 3 and the one for the knowledge dimension in Figure 4. We used a self-hosted `meta-llama/Meta-Llama-3-70B-Instruct` model from the HuggingFace repository for evaluation. We promped it using the prompt shown in Figure 5.

E Bloom Taxonomy Labeling Results

We include the human- as well as the LLM-assigned labels for the benchmarks. The labels for the knowledge dimension can be found in Table 14 and the labels for the Cognitive process dimension in Table 15.

F Benchmark Usage in Technical Reports

Table 16 provides an overview of benchmark scores included in the technical reports of Llama 3 (Meta, 2024), Llama 3.1 (Dubey et al., 2024), GPT-4 (OpenAI et al., 2024), GPT-4o (OpenAI, 2024), Grok-2 (X.ai, 2024) and Claude 3 (Anthropic, 2024). We list only benchmarks used in more than one reports.

Cognitive Dimension Classification Prompt

Your task is to classify tasks into Bloom's Taxonomy. The classes and their description are provided below:
Remember: Recall facts and basic concepts.
[Examples]: define, duplicate, list, memorize, repeat, state
Understand: Explain ideas or concepts.
[Examples]: classify, describe, discuss, explain, identify, locate, recognize, report, select, translate
Apply: Use information in new situations.
[Examples]: execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch
Analyze: Draw connections among ideas.
[Examples]: differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test
Evaluate: Justify a stand or decision.
[Examples]: appraise, argue, defend, judge, select, support, value, critique, weigh
Create: Produce new or original work.
[Examples]: design, assemble, construct, conjecture, develop, formulate, author, investigate
Classify the following problem into the corresponding category of Bloom's Taxonomy: *{problem}*
Classification:

Figure 3: Prompt for classifying the cognitive dimension

Knowledge Dimension Classification Prompt

Your task is to classify tasks into the Knowledge dimension of Bloom's Taxonomy. The classes and their description are provided below:
Metacognitive: Knowledge of cognition in general as well as awareness and knowledge of one's own cognition.
Procedural: How to do something, methods of inquiry, and criteria for using skills, algorithms, techniques, and methods.
Conceptual: The interrelationships among the basic elements within a larger structure that enable them to function together.
Factual: The basic elements students must know to be acquainted with a discipline or solve problems in it.
Classify the following problem into the corresponding category of Bloom's Taxonomy: *{problem}*
Classification:

Figure 4: Prompt for classifying the knowledge dimension

Benchmark / Dataset	Subtask	Description
BIG-Bench Hard Suzgun et al. (2023)	boolean_expressions	Evaluate the truth value of a random Boolean expression consisting of Boolean constants (True, False) and basic Boolean operators (and, or, and not).
	causal_judgement	Given a short story (involving moral, intentional, or counterfactual analysis), determine how a typical person would answer a causal question about the story.
	date_understanding	Given a small set of sentences about a particular date, answer the provided question (e.g., "The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date yesterday in MM/DD/YYYY?").
	disambiguation_qa	Given a sentence with an "ambiguous" pronoun, either determine whether the sentence is inherently ambiguous or, if the pronoun can be implicitly deduced, state the antecedent of the pronoun.
	dyck_languages	Predict the sequence of the closing parentheses of a Dyck-4 word without its last few closing parentheses.
	formal_fallacies	Given a context involving a set of statements, determine whether an argument—presented informally—can be logically deduced from the provided context.
	geometric_shapes	Given a full SVG path element containing multiple commands, determine the geometric shape that would be generated if one were to execute the full path element.
	hyperbaton	Given two English-language sentences, determine the one with the correct adjective order.
	logical_deduction_five_objects	Deduce the order of a sequence of objects based on the clues and information about their spatial relationships and placements.
	logical_deduction_seven_objects	Deduce the order of a sequence of objects based on the clues and information about their spatial relationships and placements.
	logical_deduction_three_objects	Deduce the order of a sequence of objects based on the clues and information about their spatial relationships and placements.
	movie_recommendation	Given a list of movies a user might have watched and liked, recommend a new, relevant movie to the user out of four potential choices.
	multistep_arithmetic_two	Solve multi-step equations involving basic arithmetic operations (addition, subtraction, multiplication, and division).
	navigate	Given a series of navigation steps to an agent, determine whether the agent would end up back at its initial starting point.
	object_counting	Given a collection of possessions that a person has along with their quantities, determine the number of a certain object/item class.
	penguins_in_a_table	Given a unique table of penguins, answer a question about the attributes of the penguins.
	reasoning_about_colored_objects	Given a context, answer a simple question about the color of an object on a surface.
	ruin_names	Given an artist, band, or movie name, identify a one-character edit to the name that changes the meaning of the input and makes it humorous.
	salient_translation_error_detection	Given a source sentence written in German and its translation in English, determine the type of translation error that the translated sentence contains.
	snarks	Given two nearly-identical sentences, determine which one is sarcastic.
	sports_understanding	Determine whether a fictitious sentence related to sports is plausible.
	temporal_sequences	Given a series of events and activities a person has completed during the day, determine what time they might have been free to perform another activity.
	tracking_shuffled_objects_five_objects	Given the initial positions of a set of objects and a series of transformations (namely, pairwise swaps) applied to them, determine the final positions of the objects.
	tracking_shuffled_objects_seven_objects	Given the initial positions of a set of objects and a series of transformations (namely, pairwise swaps) applied to them, determine the final positions of the objects.
	tracking_shuffled_objects_three_objects	Given the initial positions of a set of objects and a series of transformations (namely, pairwise swaps) applied to them, determine the final positions of the objects.
	web_of_lies	Evaluate the truth value of a random Boolean function expressed as a natural-language word problem.
	word_sorting	Given a list of words, sort them lexicographically.
AGIEval Zhong et al. (2023)	aqua-rat	**A**lgebra **Qu**estion **A**nswering with **Rat**ionales. Contains algebraic problems described in textual form.
	gaokao-english	College entrace exam for Chinese students. The questions test the language capabilities in the English language.
	logiqa-en	Questions that evaluate logical reasoning and text comprehension skills.
	lsat-ar	Part of the law school admission test LSAT, focusing specifically on assessing analytical skills.
	lsat-lr	Logical reasoning questions that are part of the law school admission test LSAT. The tasks require text comprehension and analytical reasoning skills. Questions focus on arguments and how to analyze them.
	lsat-tc	Part of LSAT. Text comprehension tasks that are longer than in the other tasks. Task is to match one of the answer passages to the question passage, i.e. "What passage best describes what was said in the text?".
	math	Various mathematical problems. Less text-heavy than aqua-rat.
	sat-en	Questions similar to those used in the SAT. These questions test the English language skills.
	sat-math	Questions found similar to those used in the SAT. They test the mathematical capabilities.
Winogrande	N/A	Tests commonsense reasoning and text comprehension. Sentences with gaps need to be filled with the correct option from a selection of answers.
ARC-Challenge	N/A	Abstraction and Reasoning Corpus. Questions are comparatively complex and require reasoning and external knowledge.
DROP	N/A	The task is to answer questions regarding some paragraph.
GPQA	N/A	Graduate-level Google-Proof Q&A benchmark. Contains multiple choice questions from biology, physics and chemistry. These questions are designed to be difficult even for experts.
HumanEval	N/A	Tests code generation capabilities. Models are required to write a function based on the docstring of what the function should do.
GSM8K	N/A	Contains grade school math questions. The tasks are in textual form and require multi-step reasoning.
MATH	N/A	Challenging mathematical problems.

Table 13: Overview of the benchmarks and datasets used in our analysis.

<|begin_of_text|><|start_header_id|>system**<|end_header_id|>**
You are given both a task with its intended solution, as well as the solution by a student. Your task is to evaluate whether the student solved the task correctly. Respond with @Yes@ if it was solved correctly, and with @No@ if it was not.**<|eot_id|>**
<|start_header_id|>user**<|end_header_id|>**
[Task]
{task}
[/Task]
[Intended Solution]
{solution}
[/Intended Solution]
—
[Student Solution]
{student_solution}
[/Student Solution]
<|eot_id|><|start_header_id|>assistant**<|end_header_id|>**
@

Figure 5: Llama 3 evaluation prompt.

Benchmark / Dataset	Model Subtask	Human	claude3	gpt4	gpt4o	llama3
AGIEval	aqua-rat	Procedural	Conceptual	Procedural	Procedural	Procedural
	gaokao-english	Conceptual	Conceptual	Factual	Conceptual	Factual
	logiqa-en	Conceptual	Conceptual	Procedural	Conceptual	Procedural
	lsat-ar	Procedural	Conceptual	Procedural	Conceptual	Procedural
	lsat-lr	Factual	Conceptual	Metacognitive	Conceptual	Conceptual
	lsat-rc	Conceptual	Conceptual	Metacognitive	Conceptual	Factual
	math	Procedural	Procedural	Procedural	Procedural	Procedural
	sat-en	Conceptual	Conceptual	Conceptual	Conceptual	Factual
	sat-math	Procedural	Procedural	Procedural	Procedural	Procedural
ARC-Challenge	N/A	Conceptual	Conceptual	Conceptual	Factual	Factual
BIG-Bench Hard	boolean_expressions	Factual	Conceptual	Procedural	Procedural	Factual
	causal_judgement	Conceptual	Conceptual	Metacognitive	Conceptual	Conceptual
	date_understanding	Factual	Factual	Procedural	Factual	Factual
	disambiguation_qa	Conceptual	Procedural	Metacognitive	Conceptual	Factual
	dyck_languages	Factual	Procedural	Procedural	Procedural	Procedural
	formal_fallacies	Conceptual	Conceptual	Metacognitive	Conceptual	Procedural
	geometric_shapes	Factual	Procedural	Procedural	Procedural	Factual
	hyperbaton	Conceptual	Conceptual	Procedural	Factual	Factual
	logical_deduction_five_objects	Conceptual	Conceptual	Procedural	Conceptual	Factual
	logical_deduction_seven_objects	Conceptual	Conceptual	Procedural	Conceptual	Procedural
	logical_deduction_three_objects	Conceptual	Conceptual	Procedural	Conceptual	Factual
	movie_recommendation	Conceptual	Conceptual	Procedural	Conceptual	Factual
	multistep_arithmetic_two	Procedural	Procedural	Procedural	Procedural	Procedural
	navigate	Conceptual	Procedural	Procedural	Procedural	Procedural
	object_counting	Factual	Factual	Factual	Factual	Factual
	penguins_in_a_table	Factual	Factual	Factual	Factual	Factual
	reasoning_about_colored_objects	Conceptual	Factual	Procedural	Factual	Factual
	ruin_names	Conceptual	Factual	Factual	Factual	Factual
	salient_translation_error_detection	Conceptual	Conceptual	Procedural	Procedural	Factual
	snarks	Conceptual	Factual	Metacognitive	Metacognitive	Factual
	sports_understanding	Conceptual	Factual	Factual	Factual	Factual
	temporal_sequences	Procedural	Factual	Procedural	Conceptual	Procedural
	tracking_shuffled_objects_five_objects	Procedural	Procedural	Procedural	Procedural	Factual
	tracking_shuffled_objects_seven_objects	Procedural	Procedural	Procedural	Procedural	Factual
	tracking_shuffled_objects_three_objects	Procedural	Procedural	Procedural	Procedural	Factual
	web_of_lies	Procedural	Conceptual	Procedural	Conceptual	Procedural
	word_sorting	Factual	Procedural	Procedural	Procedural	Procedural
DROP	N/A	Factual	Factual	Factual	Factual	Factual
GPQA	N/A	Procedural	Conceptual	Procedural	Procedural	Factual
GSM8K	N/A	Procedural	Procedural	Procedural	Procedural	Procedural
HumanEval	N/A	Procedural	Procedural	Procedural	Procedural	Procedural
MATH (as a whole)	N/A	Procedural	Conceptual	Procedural	Procedural	Procedural
Winogrande	N/A	Conceptual	Factual	Factual	Factual	Factual

Table 14: Knowledge dimension labels assigned to the selected benchmarks.

Benchmark / Dataset	Model Subtask	Human	bloomberta	claude3	gpt4	gpt4o	llama3
AGIEval	aqua-rat	Apply	Analyze	Apply	Apply	Apply	Apply
	gaokao-english	Understand	Apply	Understand	Understand	Understand	Understand
	logiqa-en	Analyze	Analyze	Analyze	Analyze	Analyze	Analyze
	lsat-ar	Analyze	Apply	Analyze	Analyze	Analyze	Analyze
	lsat-lr	Analyze	Evaluate	Analyze	Evaluate	Analyze	Analyze
	lsat-rc	Analyze	Apply	Analyze	Evaluate	Evaluate	Analyze
	math	Apply	Apply	Analyze	Apply	Apply	Apply
	sat-en	Understand	Apply	Understand	Understand	Understand	Understand
	sat-math	Apply	Analyze	Apply	Apply	Apply	Apply
ARC-Challenge	N/A	Understand	Analyze	Analyze	Analyze	Analyze	Analyze
BIG-Bench Hard	boolean_expressions	Apply	Analyze	Analyze	Apply	Analyze	Evaluate
	causal_judgement	Analyze	Analyze	Analyze	Evaluate	Analyze	Analyze
	date_understanding	Apply	Analyze	Apply	Apply	Apply	Apply
	disambiguation_qa	Understand	Understand	Understand	Analyze	Understand	Understand
	dyck_languages	Apply	Apply	Apply	Apply	Apply	Apply
	formal_fallacies	Analyze	Evaluate	Analyze	Evaluate	Analyze	Evaluate
	geometric_shapes	Understand	Create	Understand	Understand	Understand	Understand
	hyperbaton	Understand	Analyze	Analyze	Understand	Understand	Understand
	logical_deduction_five_objects	Analyze	Understand	Analyze	Analyze	Analyze	Analyze
	logical_deduction_seven_objects	Analyze	Understand	Analyze	Analyze	Analyze	Analyze
	logical_deduction_three_objects	Analyze	Understand	Analyze	Analyze	Analyze	Analyze
	movie_recommendation	Analyze	Create	Apply	Analyze	Analyze	Understand
	multistep_arithmetic_two	Apply	Analyze	Apply	Apply	Apply	Apply
	navigate	Understand	Apply	Apply	Apply	Analyze	Apply
	object_counting	Remember	Understand	Remember	Remember	Remember	Remember
	penguins_in_a_table	Understand	Create	Remember	Remember	Remember	Remember
	reasoning_about_colored_objects	Apply	Analyze	Remember	Remember	Remember	Remember
	ruin_names	Understand	Analyze	Remember	Understand	Understand	Understand
	salient_translation_error_detection	Analyze	Remember	Analyze	Understand	Analyze	Analyze
	snarks	Analyze	Analyze	Analyze	Evaluate	Analyze	Analyze
	sports_understanding	Analyze	Create	Evaluate	Evaluate	Analyze	Evaluate
	temporal_sequences	Apply	Apply	Apply	Apply	Analyze	Analyze
	tracking_shuffled_objects_five_objects	Apply	Apply	Create	Apply	Apply	Analyze
	tracking_shuffled_objects_seven_objects	Apply	Apply	Create	Apply	Apply	Apply
	tracking_shuffled_objects_three_objects	Apply	Evaluate	Create	Apply	Apply	Apply
	web_of_lies	Analyze	Apply	Analyze	Analyze	Analyze	Analyze
	word_sorting	Apply	Analyze	Remember	Apply	Apply	Remember
DROP	N/A	Understand	Analyze	Remember	Remember	Remember	Remember
GPQA	N/A	Analyze	Analyze	Analyze	Apply	Apply	Analyze
GSM8K	N/A	Apply	Apply	Apply	Apply	Apply	Apply
HumanEval	N/A	Apply	Create	Apply	Apply	Create	Apply
MATH (as a whole)	N/A	Analyze	Apply	Apply	Apply	Apply	Apply
Winogrande	N/A	Understand	Analyze	Understand	Understand	Understand	Understand

Table 15: Cognitive Dimension labels assigned to the selected benchmarks.

Benchmark	Llama 3	Llama 3.1	GPT-4	GPT-4o	Grok-2	Claude 3
MMLU	✓	✓	✓	✓	✓	✓
HumanEval	✓	✓	✓	✓	✓	✓
GPQA	✓	✓		✓	✓	✓
MATH	✓	✓		✓	✓	✓
GSM8K	✓	✓	✓			✓
BIG-Bench Hard	✓	✓	✓			✓
ARC-Challenge	✓	✓	✓			✓
Winogrande	✓	✓	✓			
HellaSwag		✓	✓			✓
MGSM		✓		✓		✓
DROP	✓	✓		✓		
AGIEval	✓	✓				
CommonSenseQA	✓	✓				
TriviaQA	✓	✓				
SQuAD	✓	✓				
QuAC	✓	✓				
BoolQ	✓	✓				
MMLU-Pro		✓			✓	

Table 16: Benchmark usage reported in technical reports of LLMs.