AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models

Anonymous ACL submission

Abstract

Assessing foundation models' abilities for human-level tasks is crucial for Artificial General Intelligence (AGI) development. Traditional benchmarks, which rely on artificial datasets, may not accurately represent these capabilities. In this paper, we introduce AGIEval, a novel bilingual benchmark designed to assess foundation models in the context of human-centric standardized exams, such as college entrance exams, law school admission tests, math competitions, and lawyer qualification tests. We evaluate several state-of-the-art foundation models on our benchmark. Impressively, we show that GPT-4 exceeds the average human performance in SAT, LSAT, and math contests, with 95% accuracy on SAT Math and 92.5% on the Chinese college entrance English exam. This demonstrates the exceptional performance of contemporary foundation models. In contrast, we also find that GPT-4 is less proficient in tasks requiring complex reasoning or specific domain knowledge. Our comprehensive analyses of model capabilities (understanding, knowledge, reasoning, and calculation) reveal their strengths and limitations, providing valuable insights into future directions for enhancing general capabilities. By concentrating on tasks pertinent to human cognition and decision-making, our benchmark delivers a meaningful and robust evaluation of foundation models' performance in real-world scenarios[1].

1 Introduction

Recently, large foundation models, such as the large language models (LLMs) ChatGPT(OpenAI, 2022) and GPT-4 (OpenAI, 2023), exhibited remarkable versatility and adaptability, with plethora of applications spanning various domains as a decision-making assistant, from processing daily events to assisting in specialized fields such as law and finance. With these advancements, AI systems are inching closer to achieving Artificial General Intelligence (AGI). As these AI systems continue to evolve and become more integrated into our daily lives, it is essential to effectively assess their general abilities in handling human-centric tasks, identify potential shortcomings, and ensure that they can handle complex, human-centric tasks effectively. Moreover, evaluating their reasoning abilities is also crucial to ensure their reliability and trustworthiness across diverse settings.

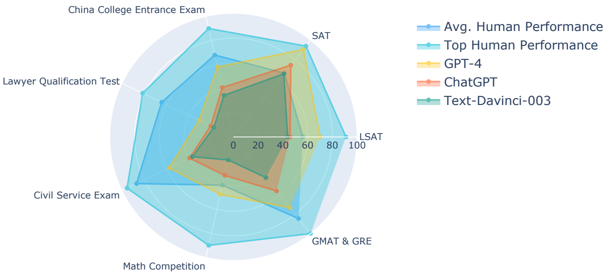


Figure 1: The performance of LLMs (text-davinci-003, ChatGPT, and GPT-4) was evaluated on several human-centric exams under zero-shot learning with a Chain-of-Thought (CoT) prompting setting. Human performance (avg.) refers to the average performance of all test takers, while human performance (top) refers to the performance of the top 1% of test takers. Compared to the averaged human performance, GPT-4 achieves better scores on the SAT, LSAT, and math competitions.

Traditional benchmarks for evaluating foundation models often fall short in providing an accurate assessment of their general abilities in handling human-level tasks. This is primarily due to the use of artificial datasets and a lack of emphasis on real-world tasks that require human-like cognitive capabilities. Moreover, these benchmarks often focus on tasks that do not truly represent the complexities and nuances of real-world human cognition and decision-making, leading to a skewed evaluation of models' capabilities and limiting their ability to provide meaningful insights into the models' real-world applicability. Consequently, there is a growing need for a more human-centric benchmark that allows for a robust evaluation of foundation

[1] The data and code are released in the supplementary

model in the context of tasks that are relevant to human reasoning and problem-solving.

We introduce a human-centric benchmark, AGIEval, specifically designed to evaluate the general abilities of foundation models in tasks pertinent to human-level problem-solving. This benchmark is derived from official, public, and high-standard admission and qualification exams intended for general human test-takers, such as general college admission tests (e.g., Chinese College Entrance Exam (Gaokao) and American SAT), law school admission tests, math competitions, lawyer qualification tests, and national civil service exams. These exams are taken by a diverse range of individuals seeking entry into higher education institutions or new career paths, with millions participating annually (e.g., 12 million for the Chinese Gaokao and 1.7 million for the American SAT). As a result, these exams establish officially recognized standards for assessing human-level capabilities. Additionally, the benchmark covers bilingual tasks in both Chinese and English, allowing for a more comprehensive evaluation. By concentrating on these tasks, our benchmark provides a more meaningful and comprehensive evaluation of large language model performance in scenarios directly relevant to human decision-making.

We employ 20 human-centric tasks across a wide variety of subjects in our benchmark to assess the performance of cutting-edge foundation models, encompassing close-source models, i.e., text-davinci-003, ChatGPT and GPT-4, and an open-source model, Vicuna (Chiang et al., 2023). Our experiments explore their performance under various settings, including few-shot learning, zero-shot learning, and chain-of-thought prompting techniques. We compare the performance of these models with human performance, as illustrated in Fig. 1. Remarkably, the results reveal that GPT-4 outperforms the average human performance on LSAT, SAT, and math competitions under the zero-shot chain-of-thought (CoT) setting, demonstrating its capability on human-centric tasks. However, there remains a gap between GPT-4 and the top human performance, indicating opportunities for future improvement. We also discover that these models struggle with tasks requiring complex reasoning (e.g., LSAT-analytical reasoning and physics) or specific domain knowledge, such as law and chemistry. Moreover, our comprehensive qualitative analyses of the four dimensions of model capabilities (i.e., *understanding, knowledge, reasoning, and calculation*) delve into their respective strengths and limitations, providing valuable insights into their general capabilities. This multi-faceted approach enables us to examine the models' single-task behavior and identify general patterns, ultimately contributing to a more robust understanding of these state-of-the-art models and their potential applications in tackling human-level tasks.

2 Background and Related Work

Large Foundation Model: Recently, large foundation models, like LLMs (e.g., GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), OPT (Zhang et al., 2022a) and FLAN-T5 (Chung et al., 2022)) have successfully demonstrated unprecedented performance in a wide range of natural language tasks. The success of these models can be attributed to advances in deep learning techniques, architectural improvements, and the availability of massive amounts of data for training. The most recent cutting-edge language models, such as ChatGPT(OpenAI, 2022) and GPT-4 (OpenAI, 2023), have continued to demonstrate substantial adaptability to a diverse array of tasks and domains and have served as a daily decision-making assistant for human beings. However, despite their impressive performance on various benchmarks, concerns have been raised about the reasoning abilities, trustfulness and real-world applicability of these models (Marcus and Davis, 2019).

Evaluation of Language Models: Constructing benchmarks is a reliable way to establish evaluation standards and monitor model performance. Numerous benchmarks (Thorne et al., 2018; Rajpurkar et al., 2016) have been proposed and widely adopted for evaluating single-task performance, such as SQuAD (Rajpurkar et al., 2016) for assessing answer extraction ability and SNLI (Bowman et al., 2015) for evaluating natural language inference capability. The emergence of general language models (LMs) like BERT (Devlin et al., 2019) has made it increasingly essential to develop more comprehensive benchmarks to assess the general capabilities of these LMs. GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) are popular benchmarks that evaluate language model performance across diverse NLP tasks. GLUE series benchmarks have significantly influenced language model development, encouraging researchers to enhance their models' generalization capabilities. The LAMBADA language modeling task (Paperno et al., 2016) assesses language models' ability to capture long-range dependencies in text. SentEval (Conneau and Kiela, 2018) and DecaNLP (McCann et al., 2018) also set benchmarks for evaluating models' general capabilities. ToxiGen (Hartvigsen et al., 2022) and BOLD (Dhamala et al., 2021) evaluate the bias in language models. Despite their broad applicability, these benchmarks mainly consist of artificially curated datasets designed to evaluate specific machine skills, rather than real-world problems aimed at assessing human behaviors. Consequently, these benchmarks primarily focus on simpler textual understanding rather than complex reasoning abilities aligned with real-world applicability. MMLU (Hendrycks et al., 2020) addresses this issue by collecting questions from online sources covering a diverse set of subjects (e.g., history, humanities) that humans learn, pushing towards human-centric evaluation. Our work differs from MMLU in two main ways: (1) We derive our benchmark from high-standard human-centric exams like college admissions tests, ensuring a robust, standardized evaluation, unlike MMLU which lacks explicit sourcing details. (2) AGIEval is bilingual (English and Chinese), broadening the assessment scope across languages and cultures, whereas MMLU is solely English-based. The official technical report of GPT-4 (OpenAI, 2023) also underscored the importance of evaluating models' behaviors on human exams and analyzed GPT-4's performance on several such exams. However, the relevant benchmarks in these reports and the corresponding model outputs are not publicly available, and the evaluation metric is also not transparent. These factors limit further research to follow up their evaluation.

3 Human-Centric Benchmark

3.1 Design Principles

Emphasis on human-level cognitive tasks: Our human-centric benchmark is designed to mimic human cognition and problem-solving, aiming for a comprehensive evaluation of foundation models. We use a diverse set of public, official exams, such as college admission tests, law tests, and national civil service exams. These exams, taken by millions seeking further education or careers, provide standards for assessing human-level capabilities, making our benchmark directly relevant to human cognition and decision-making.

Relevance to real-world scenarios: The second design principle is emphasizing tasks relevant to real-world situations. By utilizing high-standard admission and qualification exams, we capture the complexity and practicality of challenges in various fields. This not only measures model performance against human cognition, but also their applicability in real-life scenarios, fostering AI development that is reliable, practical, and capable of solving diverse real-world problems.

3.2 Exam Selection

Our human-centric benchmark features various standardized exams, each serving unique assessment roles. Some exams are participated by millions of human test-takers annually. For example, 12 millions of students participate in Gaokao. Statistics of annual human participants are reported in Table 5. **Dataset collection is introduced in Appendix B**. The following categories of human-centric exams are included in our benchmark:

General College Entrance Exams: Including the GRE, SAT, and Gaokao, these exams assess critical thinking, problem-solving, and analytical skills for entry into higher education. We selected tasks from eight subjects in the Gaokao and mathematical questions from the GRE and SAT. These exams are designed to assess the general aptitude and subject-specific knowledge of humans.

Law School Admission Test: LSAT measures reasoning and analytical skills of prospective law students. These tests include sections on logical reasoning, reading comprehension, and analytical reasoning, aiding us in evaluating language models' legal reasoning abilities and ability to analyze complex information and draw accurate conclusions.

Lawyer Qualification Test: Including the bar exam, these tests assess legal knowledge, analytical skills, and ethical understanding. Questions from Chinese lawyer qualification tests are included. By incorporating lawyer qualification tests in our benchmark, we can evaluate language models' performance in the context of professional legal expertise and ethical judgment.

Graduate Management Admission Test (GMAT): The GMAT is a standardized exam designed to assess the analytical, quantitative, verbal, and integrated reasoning skills of prospective graduate business school students. It assess LLMs' potential to assist in decision-making and problem-solving in management scenarios.

Exams	#Participants	Language	Tasks	Subject	# Instance	#Avg. Token
Gaokao	12M	Chinese	GK-geography	Geography	199	144
			GK-biology	Biology	210	141
			GK-history	History	243	116
			GK-chemistry	Chemistry	207	113
			GK-physics	Physics	200	124
			GK-En	English	306	356
			GK-Ch	Chinese	246	935
			GK-Math-QA	Math	351	68
			GK-Math-Cloze	Math	118	60
SAT	1.7M	English	SAT-En.	English	206	656
			SAT-Math	Math	220	54
Lawyer Qualification Test	820K	Chinese	JEC-QA-KD	Law	1000	146
			JEC-QA-CA	Law	1000	213
Law School Admission Test (LSAT)	170K	English	LSAT-AR	Law-Analytics	230	154
			LSAT-LR	Law-Logic	510	178
			LSAT-RC	Law-Reading	260	581
Civil Service Examination	2M	English	LogiQA-en	Logic	651	144
	2M	Chinese	LogiQA-ch	Logic	651	242
GRE	340K	English	AQuA-RAT	Math	254	77
GMAT	150K	English				
AMC	300K	English	MATH	Math	1000	40
AIME	3000	English				

Table 1: Exams included in AGIEval. We highlight the number of human participants taking these exams annually (column “# Participants"). We also report the number of instances and average token number in AGIEval.

High School Math Competitions: Math competitions like **American Mathematics Competitions (AMC)** and the **American Invitational Mathematics Examination (AIME)** test mathematical abilities, creativity, and problem-solving skills, helping to evaluate models' proficiency in tackling complex mathematical problems.

Chinese Civil Service Examination: This exam assesses a range of competencies for prospective civil servants. These exams evaluate a range of competencies, such as general knowledge, reasoning abilities, language skills, and subject-specific expertise, allowing us to gauge models' performance in public administration contexts.

4 Evaluation of Foundation Models

4.1 Model Selection

In this section, we evaluate the performance of various state-of-the-art language models on our benchmark dataset. **(1) GPT-4:** The fourth iteration of the GPT series, GPT-4 is a large-scale, generative pre-trained transformer with enhanced performance and a broad knowledge base. It exhibits human-level performance in numerous scenarios, including factuality, steerability, and adherence to guardrails. **(2) ChatGPT:** An OpenAI-developed conversational model, ChatGPT is trained on extensive instruction data and fine-tuned using reinforcement learning with human feedback, enabling contextually relevant responses. **(3) text-davinci-003:** As an intermediate version between GPT-3 and GPT-4, GPT-3.5 offers improved performance, providing a comparative perspective. We specifically evaluate the text-davinci-003 variant. **(4) Vicuna-13B (Chiang et al., 2023):** It is an open-source LLM, trained on user-shared conversations from ShareGPT by fine-tuning LLaMA. It achieves over 90% of the quality of OpenAI's ChatGPT.

4.2 Experimental Setup

To gauge the adaptability of LLMs, we conduct two types of evaluations: zero-shot and few-shot. We further implement a “Chain-of-Thought (CoT)” reasoning evaluation. Fig. 2 describes the concrete prompting examples for zero-shot testing, few-shot testing and chain-of-thought prompting.

4.2.1 Zero-shot and Few-shot Evaluation

In the zero-shot setting, models were evaluated on the questions without being provided examples of the specific tasks. This scenario tests the models' innate ability to reason and solve problems without explicit training. In the few-shot setting, models were given a small number of examples (e.g., 5) from the same task before being evaluated on the test samples. This evaluation setup tests the models' ability to quickly adapt from limited examples.

Task/Model	Human		Zero-Shot			Zero-Shot CoT			Few-Shot			Few-Shot CoT		
	Avg.	Top	TD	CG	G4	TD	CG	G4	TD	CG	G4	TD	CG	G4
AQuA-RAT	85	100	29.9	31.9	40.6	42.1	55.9	73.2	30.3	31.1	50.8	47.2	60.6	74.0
MATH	40	90	11.9	26.4	35.7	19.1	31.9	47.7	10.3	14.8	15.1	15.1	30.1	25.3
LogiQA (English)	86	95	22.7	35.0	49.3	36.9	39.9	57.8	43.5	43.5	63.9	37.5	38.9	62.7
LogiQA (Chinese)	88	96	40.3	41.0	58.8	36.7	38.9	57.5	43.2	46.2	65.0	40.0	38.6	61.9
JEC-QA-KD	71	78	21.9	21.1	33.4	18.4	21.2	31.9	22.4	27.6	41.3	23.6	23.4	40.4
JEC-QA-CA	58	85	21.0	22.0	31.1	16.7	19.6	29.8	22.2	25.1	37.4	16.1	20.0	34.7
LSAT-AR	56	91	21.7	24.4	35.2	23.9	22.6	34.4	22.6	25.7	33.9	22.6	25.2	31.7
LSAT-LR	56	91	47.5	52.6	80.6	50.0	52.6	80.6	60.4	59.2	85.9	51.2	52.2	84.5
LSAT-RC	56	91	64.7	65.4	85.9	57.6	62.1	85.1	70.6	67.7	87.7	64.3	57.6	87.7
SAT-Math	66	94	35.5	42.7	64.6	54.6	70.9	95.0	44.6	40.9	71.8	55.5	65.0	89.6
SAT-English	66	94	74.8	81.1	88.8	75.7	77.7	85.9	84.0	81.1	88.8	76.7	78.2	85.9
GK-Cn	65	85	43.9	39.0	53.3	35.4	33.7	44.7	25.6	41.5	61.4	29.3	37.8	51.6
GK-En	69	91	81.4	84.9	91.9	83.0	84.3	92.5	86.9	86.3	93.8	80.7	84.6	93.1
GK-geography	65	85	53.3	59.8	76.9	48.7	55.8	72.4	59.8	63.8	75.9	52.3	61.8	76.4
GK-history	64	85	47.3	59.7	77.4	37.0	50.2	76.5	49.0	57.6	77.8	51.9	58.4	78.2
GK-biology	68	89	40.5	52.9	75.7	30.0	42.4	71.9	44.3	52.4	80.0	32.9	50.0	72.9
GK-chemistry	66	86	27.1	38.7	51.7	24.6	33.8	52.2	32.4	44.0	54.6	35.8	33.8	54.1
GK-physics	71	94	22.0	33.0	39.0	18.5	29.5	45.5	31.0	33.5	43.5	27.5	36.5	54.5
GK-Math-QA	73	96	28.2	36.5	47.0	28.8	33.3	50.7	27.6	31.3	39.9	33.1	31.6	49.0
GK-Math-Cloze	73	96	17.0	7.6	16.1	4.2	5.1	15.3	5.9	5.9	11.0	5.93	8.5	16.1
Average	67	91	38.1	42.9	56.4	37.4	43.2	58.4	41.2	44.4	59.2	40.4	45	61.3

Table 2: Performance of close-source LLMs on 20 tasks under **zero-shot**, **zero-shot CoT**, **few-shot** and **few-shot CoT** settings. We also report human performance on each task. For LSAT, Gaokao and SAT, we report average (50%) and top (1%) human performance. The Text-Davinci-003 is abbreviated as TD, ChatGPT is abbreviated as CG, and GPT-4 is abbreviated as G4.

4.2.2 Chain-of-Thought (CoT) Reasoning

We employ the Chain-of-Thought (CoT) prompting method (Wei et al., 2022) to assess models' reasoning capabilities. CoT enables large language models to break down a complex question to a series of decomposed reasoning steps. As shown in Fig. 2, CoT involves two steps: Firstly, with prompt "*[question] Let's think step by step:* "(Zhang et al., 2022b), the model generates an explanation for a given question, which evaluates its comprehension and problem-solving strategy identification. Secondly, the model provides an answer based on its explanation, testing its ability to generate a solution using its self-derived reasoning, mirroring human problem-solving processes. In the few-shot CoT setting, the explanation and answer are generated simultaneously.

4.2.3 Evaluation Metrics

We use both quantitative and qualitative evaluation metrics. Quantitative metrics included accuracy for multi-choice questions and use Exact Match (EM) for fill-in-blank questions. We also perform qualitative evaluations, which involved human evaluators assessing the models' responses in terms of semantic understanding capability, knowledge utilization, and reasoning and calculation.

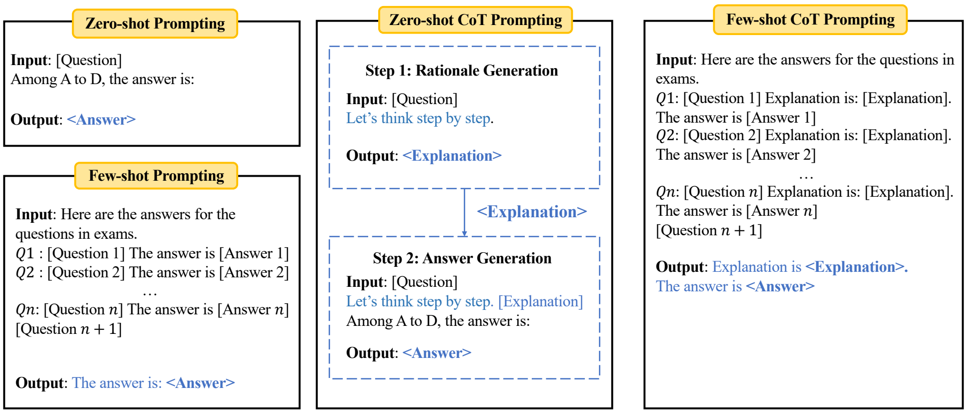
4.3 Main Results

The results of closed-source models are reported in Table 2, while the results of the open-source model are reported in Table 3. We also report average and top human performance on each task. From the results, we highlight the following findings.

(1) Superior Performance of GPT-4: On average, GPT-4 significantly outperforms its counterparts (e.g., ChatGPT) across all settings. Impressively, GPT-4 achieves 93.8% accuracy on Gaokao-English and 95% accuracy on SAT-MATH, demonstrating its superior capabilities.

(2) ChatGPT v.s. TD-003: ChatGPT excels over text-davinci-003 in tasks requiring extensive knowledge like geography, biology, chemistry, physics, and mathematics, implying a stronger knowledge base of ChatGPT. In tasks emphasizing simple comprehension and logical reasoning, like English and LSAT tasks, both models perform comparably, indicating their proficiency in language understanding and logical reasoning.

(3) Challenge of Complex Tasks: All models face difficulties with complex tasks, such as those in MATH or LSAT-AR, revealing limitations in handling advanced reasoning. This presents future research opportunities to bolster models' reasoning abilities.



Zero-shot Prompting

Input: [Question]
Among A to D, the answer is:

Output: <Answer>

Few-shot Prompting

Input: Here are the answers for the questions in exams.
$Q1$: [Question 1] The answer is [Answer 1]
$Q2$: [Question 2] The answer is [Answer 2]
…
Qn: [Question n] The answer is [Answer n]
[Question $n+1$]

Output: The answer is: <Answer>

Zero-shot CoT Prompting

Step 1: Rationale Generation
Input: [Question]
Let's think step by step.

Output: <Explanation>

<Explanation>

Step 2: Answer Generation
Input: [Question]
Let's think step by step. [Explanation]
Among A to D, the answer is:

Output: <Answer>

Few-shot CoT Prompting

Input: Here are the answers for the questions in exams.
$Q1$: [Question 1] Explanation is: [Explanation].
The answer is [Answer 1]
$Q2$: [Question 2] Explanation is: [Explanation].
The answer is [Answer 2]
…
Qn: [Question n] Explanation is: [Explanation].
The answer is [Answer n]
[Question $n+1$]

Output: Explanation is **<Explanation>**.
The answer is **<Answer>**

Figure 2: Prompting examples of different settings.

Task/Model	Computation		LogiQA		JEC-QA		LSAT			SAT		GK								
	AQaA	MATH	En.	Cn.	KD	CA	AR	LR	RC	Math	En.	Cn	En	Geo.	His.	Bio.	Che.	Phy.	M.-QA	M.-Cloze
Vicuna (ZS)	26.4	6.8	18.4	23.5	14.3	12.4	22.2	25.5	30.5	24.6	50.5	25.6	50.7	24.6	28.9	20.5	26.6	15	22.5	2.5
Vicuna (ZS-CoT)	22.1	6.6	30.3	27.1	14.9	15.2	20.9	36.1	44.2	35.5	57.8	23.6	67	28.6	34.9	24.3	23.2	17	21.7	1.7

Table 3: Performance of Vicuna-13B under zero-shot and zero-shot CoT setting. Task names are abbreviated.

(4) Few-shot Learning vs. Zero-shot Learning: Few-shot learning marginally outperforms zero-shot learning, suggesting that LLMs' zero-shot capabilities are nearing their few-shot performance. This development, a marked improvement from the original GPT-3 (few-shot performance of GPT-3 is significantly better), may stem from enhanced human-alignment and instruction tuning in recent models. This progress demonstrates the effectiveness of recent advancements in LLM tuning, which allows them to better understand the meaning and context of tasks even in zero-shot settings. As shown in Fig. 3, Vicuna, despite excelling on OpenLLM leaderboard (Beeching et al., 2023) and its claimed comparable ability with ChatGPT, falls short on AGIEval, highlighting the valuable challenges AGIEval presents to open-source models.

4.4 Analyses of Chain-of-thought Prompting

As reported in Table 2, the CoT prompting demonstrates its potential by improving performance. However, the performance gains from CoT are not consistently observed across all tasks. Our analysis leads to the following findings:

(1) Performance Variability: CoT mainly enhances performance in English math and logic reasoning tasks but degrades performance in others, implying inconsistent effects on different tasks, which may be a consequence of the generated misleading reasoning processes. It's vital to understand what drives these variations to uniformly optimize CoT for diverse tasks.

(2) Backbone Dependency: CoT's efficacy is linked to the base model. GPT-4, for instance, generates more illustrative reasoning processes, improving CoT performance. This underscores the importance of model compatibility with CoT.

(3) Language Sensitivity: CoT performance varies with language. For LogiQA, CoT improves English tests but decreases Chinese ones. Similar findings are observed in mathematical tests, where performance increase on English math tests (MATH, AQuA) but decrease on Chinese math exam in Gaokao. This suggests CoT's sensitivity to language differences, necessitating further optimization across languages to ensure its consistent and generalizable reasoning capabilities.

In conclusion, CoT's effectiveness is relevant to task, model capability, and language. These factors need careful consideration when employing CoT or developing future models.

4.5 Qualitative Analyses of Model Capabilities

We conduct a qualitative analysis of ChatGPT's outputs under a zero-shot CoT setting, with 100 erroneously answered instances for each task, to assess its alignment with human capabilities. We enlist human annotators with expert knowledge, such as Ph.D. students and professional researchers, to evaluate the model outputs (i.e., explanations and answers) along the following four dimensions and report average scores for tasks. **(1) Understanding**: Assessing whether the model comprehends the con-

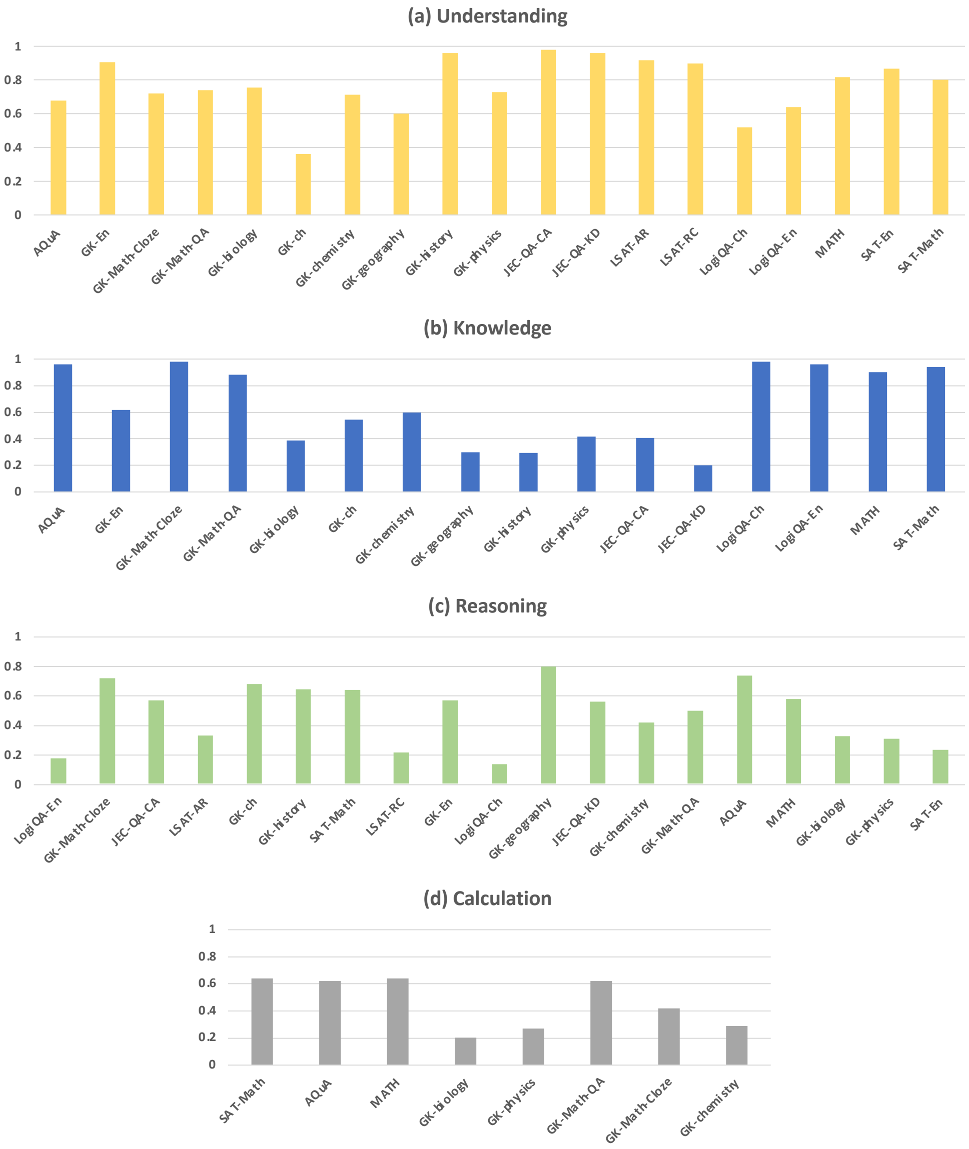


Figure 3: Qualitative assessment of inaccurately answered questions by the model focuses on four dimensions of capabilities: understanding, knowledge acquisition, reasoning and calculation.

text and questions. **(2) Knowledge**: Evaluating the model's ability to recall relevant knowledge or formula for problem-solving. **(3) Reasoning**: Determining the model's ability to reason accurately. **(4) Calculation**: Evaluating the model's correctness in mathematical calculations.

Each instance is scored 1 for correct skill application and 0 otherwise. Certain tasks like LSAT and English reading tasks, primarily emphasize understanding not requiring external knowledge or calculations, were excluded from respective skill analyses. This detailed evaluation provides insights into the models' strengths and weaknesses, guiding future improvements of LLMs. Annotators also provided insights into the models' behavior patterns. We summarize the overall trend in the paper and give **detailed analyses about strength and weaknesses in Appendix D**.

4.5.1 Overall Trend of Model Capabilities

The average scores on tasks for the four dimensions of capabilities are shown in Fig. 3. As shown From the qualitative analysis, we summarize the following observations:

Understanding: The model generally performs well in understanding. For most tasks, it can accurately interpret the meaning of questions, demonstrating its ability to comprehend context.

Knowledge: In the knowledge dimension, the model demonstrates proficiency in identifying correct knowledge or formulas for tasks. However, it encounters difficulties in recalling specific domain knowledge, such as law, biology, and physics. This observation emphasizes the significance of integrating more domain-specific knowledge into the model, potentially through the utilization of specialized domain-specific knowledge bases or knowledge-enhanced pre-training techniques.

Reasoning: Among the four dimensions, the model's reasoning capability appears to be relatively worse. For tasks necessitating complex, multi-step reasoning (e.g., LSAT-AR, LogiQA, and GK-Physics), the model struggles to accurately execute multi-step reasoning process. This underlines the importance of research concentrating on augmenting the model's reasoning capabilities, potentially through the exploration of prompting methods or training strategies that encourage complex reasoning and problem-solving skills.

Calculation: The model's calculation ability is weaker than their understanding capacity and displays variability across different subjects. They perform better in math exams, but face challenges in chemistry and biology exams, which often require variable substitution involving chemical elements. This suggests that enhancing the calculation and combinatorial abstraction and calculation ability of the model, particularly in subject areas with specialized notations or customized symbol substitutions, is a crucial challenge for further improvement.

4.6 Data Contamination Issue

The issues surrounding data contamination and future web scrapes on training data for LLMs are noteworthy. Most of current benchmarks and datasets up to date suffer from these vulnerabilities. To exam the situation of contamination, we provided timestamp for the 4 new Gaokao datasets and we can evaluate on the latest tests (later than 2022) released later than the training data timestamp of ChatGPT and GPT-4. Hereinafter, from AGIEval, we provide results comparing the GPT-4 zero-shot performance on six Gaokao subjects with and without risk of data contamination (Chinese, English, and History have not been included in this analysis due to the constrained size of the exams for these subjects). The uncontaminated dataset comprises entries released in 2022, which postdates the GPT-4 training data's timestamp (September 2021). The results are reported on Table 4. Evidently, we observe that barring the Mathematics subjects, the performance experiences a minor drop in the absence of contamination, yet remains proximate to the performances on the complete datasets. This finding substantiates that while AGIEval still retains its value as a useful and effective human-centric benchmark for evaluating the abilities of foundation models against complex human-oriented tasks.

	#test	Full acc.	Un. acc.
Gaokao-geo.	37	76.9%	73%
Gaokao-bio.	58	75.7%	77.6%
Gaokao-chem.	64	51.7%	42.2%
Gaokao-phy.	20	40%	40%

Table 4: Analysis on data contamination risk on AGIEval. The uncontaminated set (performance on the last column) includes examples released later than the time stamp of training data of ChatGPT and GPT-4.

5 Conclusion

We introduce AGIEval, a novel benchmark specifically designed to assess the general capabilities of large foundation models with respect to human-level cognition. The benchmark comprises high-quality official admission tests, qualification exams, and advanced competitions tailored for human participants, including law school admission tests and college entrance examinations. These assessments establish officially recognized standards for gauging human capabilities, making them well-suited for evaluating foundation models in the context of human-centric tasks. Additionally, AGIEval incorporates bilingual tasks in both Chinese and English, offering a more comprehensive assessment of model behavior. We have carried out an extensive evaluation of three cutting-edge large foundation models: text-davinci-003, ChatGPT, and GPT-4, using AGIEval. Remarkably, GPT-4 surpasses average human performance on LSAT, SAT, and math competition, attaining a 95% accuracy rate on the SAT Math test and a 92.5% accuracy on the Gaokao English test, demonstrating the impressive performance of contemporary foundation models. Despite their significant achievements, our in-depth manual analyses also reveal the limitations of these large language models in terms of understanding, knowledge utilization, reasoning and calculation. Guided by these findings, we explore potential future research avenues in this domain. By assessing these foundation models on human-centric tasks and probing their capabilities more deeply, we strive to foster the development of models that are more closely aligned with human cognition.

6 Limitation

Until the time we finished this work, state-of-the-art foundation models, such as text-davinci-003, ChatGPT, and GPT-4, only have publicly available APIs for language-only tasks. Therefore, we release the language-only version of AGIEval and focus on evaluating a wider range of large language models in the present paper. In the future, we will study on the multi-modal test set.

References

Edward Beeching, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

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*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *Sort*, 2(4):0–6.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.

Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

OpenAI. 2022. Chatgpt. https://chat.openai.com/chat.

OpenAI. 2023. Gpt-4 technical report.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2022. From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2201–2216.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: A legal-domain question answering dataset. In *Proceedings of AAAI*.

Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Yining Chen, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2022. Analytical reasoning of text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2306–2319, Seattle, United States. Association for Computational Linguistics.

A Discussion about Future Directions

In light of the findings and limitations identified in our analysis, we point out several potential future directions for the development of large foundation models. These directions aim to address the weaknesses observed and further improve the models' capabilities in various human-centric tasks.

Inclusion of External Knowledge and Formulas: Enriching the models with external knowledge sources, like formulas and domain-specific knowledge can help enhance their performance in mathematical and knowledge-intensive tasks. Specifically, developing models that can effectively handle domain-specific tasks, such as those in law, biology, or physics, requires the integration of specialized knowledge bases and expertise into the model, and enables the model to adapt to different verticals more effectively. This could involve integrating structured knowledge repositories, mathematical and scientific concepts into the models with pretraining or knowledge-enhanced prompting methods, allowing them to access and apply relevant information more efficiently.

Strict Complex Logical Reasoning: Improving the models' capacity for strict complex logical reasoning is crucial for their performance in a wide range of human-centric tasks. This could involve the creation of new datasets that emphasize complex reasoning, as well as incorporating APIs and external symbolic compilers that can execute strict logical or mathematical deduction, and use the execution results to further facilitate logical analysis and reasoning verification.

Multi-lingual Reasoning Capabilities Generalization: As mentioned in Sec. 4.4, the reasoning capabilities of models are variant across different language, where the reasoning ability is relatively better for rich-resourced language like English. Enhancing the models' multi-lingual reasoning capabilities is essential for their applicability in a diverse range of real-world scenarios. Therefore, future directions can put more focus on enhancing the multilingual generalization of the reasoning capability of foundation models.

Multi-modal Evaluation: Expanding the evaluation framework to include multi-modal tasks can provide a more comprehensive assessment of the models' capabilities. This could involve incorporating visual, auditory, or interactive tasks that require the models to process and reason with multiple types of input simultaneously and generate multi-modal outputs for comprehensive real-world applications. In future work, we will focus on the multi-modal version of AGIEval.

Better Automatic Evaluation Metrics for Human-centric Tasks: Developing more robust and meaningful automatic evaluation metrics is crucial for the objective assessment of large language models' performance. Future research should focus on devising metrics that can accurately capture the models' understanding, knowledge, and reasoning abilities while taking into account the nuances and complexities of real-world tasks.

Robustness of Reasoning Capability: Improving the robustness of the models' reasoning capabilities is essential for ensuring their consistency and reliability across various contexts. This can be achieved by exploring techniques that enhance the models' ability to maintain consistent reasoning performance, even when faced with changes in the surrounding context or variations in the input data.

By addressing these future directions, foundation models can be further developed and refined to exhibit more advanced capabilities that align closely with human cognition, ultimately enabling them to tackle a broader range of complex, human-centric tasks with greater accuracy and reliability.

B Dataset Collection

As previously mentioned, our human-centric benchmark comprises questions from a diverse range of official and high-quality exams, originally designed for human test-takers. These exams include general college admission tests (GRE, Gaokao, SAT), entrance exams for specific majors (such as LSAT and GMAT), high school math competitions (AMC and AIME), as well as the national civil service examination and lawyer qualification test in China.

Since evaluating model performance on subjective questions is challenging without human expert scoring, we believe such questions are unsuitable for inclusion in this benchmark for consistent assessment. To ensure a robust and standardized evaluation metric, we have removed all subjective questions, retaining only objective ones, such as multiple-choice and fill-in-the-blank questions.

With regard to data collection, we gather Gaokao[2] and SAT questions[3] from publicly available online sources, along with their corresponding solutions or explanations. Throughout our data collection phase, we encountered various challenges. Consider the instance of Gaokao: our approach encompassed not only discerning reliable sources while respecting copyright regulations, but also the annotation and removal of examples with multi-modal components, elimination of duplications, identification of items unsuitable for the QA format, as well as reformatting and connecting passages and questions. Furthermore, we invite professional human experts to manually check the correctness of latex formula in each question and answer, to ensure the correctness and robustness of QA pairs.

For the LSAT, we utilize data from Wang et al. (2022) and Zhong et al. (2022), which encompasses three tasks (logical reasoning, reading comprehension, and analytical reasoning) from the LSAT administered between 1991 and 2016. For Chinese civil service examinations, we repurpose data from LogiQA (Liu et al., 2021), a dataset built on various types of logical reasoning questions collected from the National Civil Servants Examination of China. It is worth noting that LogiQA consists of bilingual questions (English and Chinese), where the English version is a translated version of the original Chinese version.

For high school math competitions, we employ data from the MATH dataset (Hendrycks et al.), comprising questions from AMC and AIME. Furthermore, we incorporate GRE and GMAT questions from AQaA-RAT (Ling et al., 2017), which emphasizes algebraic word problems. In the case of the Chinese Civil Service Examination, we reuse instances from JEC-QA (Zhong et al., 2020), a large-scale dataset derived from the National Judicial Examination of China. We down-sample the two types of JEC-QA and MATH to 1,000 instances each.

As a result, we construct a benchmark consisting of 8,062 questions for evaluation. Detailed data statistics are presented in Table 5. It is worth noting that our benchmark is bilingual, encompassing both **English and Chinese tests**. This design enables the evaluation of a broader scope of model capabilities, reflecting their performance and adaptability across different languages. A few data examples in Gaokao are shown in Fig. 4, and an example in SAT and the corresponding Chain-of-Thought reasoning process generated by GPT-4 is shown in

[2] Gaokao questions are collected from officially announced exam questions and answers like http://www.hbccks.cn/html/gkgzzt/ggsjjda/.

[3] https://satsuite.collegeboard.org/sat/practice-preparation/practice-tests/paper

Exams	#Participants	Language	Tasks	Subject	# Instance	#Avg. Token
Gaokao	12M	Chinese	GK-geography	Geography	199	144
			GK-biology	Biology	210	141
			GK-history	History	243	116
			GK-chemistry	Chemistry	207	113
			GK-physics	Physics	200	124
			GK-En	English	306	356
			GK-Ch	Chinese	246	935
			GK-Math-QA	Math	351	68
			GK-Math-Cloze	Math	118	60
SAT	1.7M	English	SAT-En.	English	206	656
			SAT-Math	Math	220	54
Lawyer Qualification Test	820K	Chinese	JEC-QA-KD	Law	1000	146
			JEC-QA-CA	Law	1000	213
Law School Admission Test (LSAT)	170K	English	LSAT-AR	Law-Analytics	230	154
			LSAT-LR	Law-Logic	510	178
			LSAT-RC	Law-Reading	260	581
Civil Service Examination	2M	English	LogiQA-en	Logic	651	144
	2M	Chinese	LogiQA-ch	Logic	651	242
GRE	340K	English	AQuA-RAT	Math	254	77
GMAT	150K	English				
AMC	300K	English	MATH	Math	1000	40
AIME	3000	English				

Table 5: Exams included in AGIEval. We highlight the number of human participants taking these exams annually (column “# Participants"). We also report the number of instances and average token number in AGIEval.

Fig. 5.

C Implementation Details

C.1 API Details

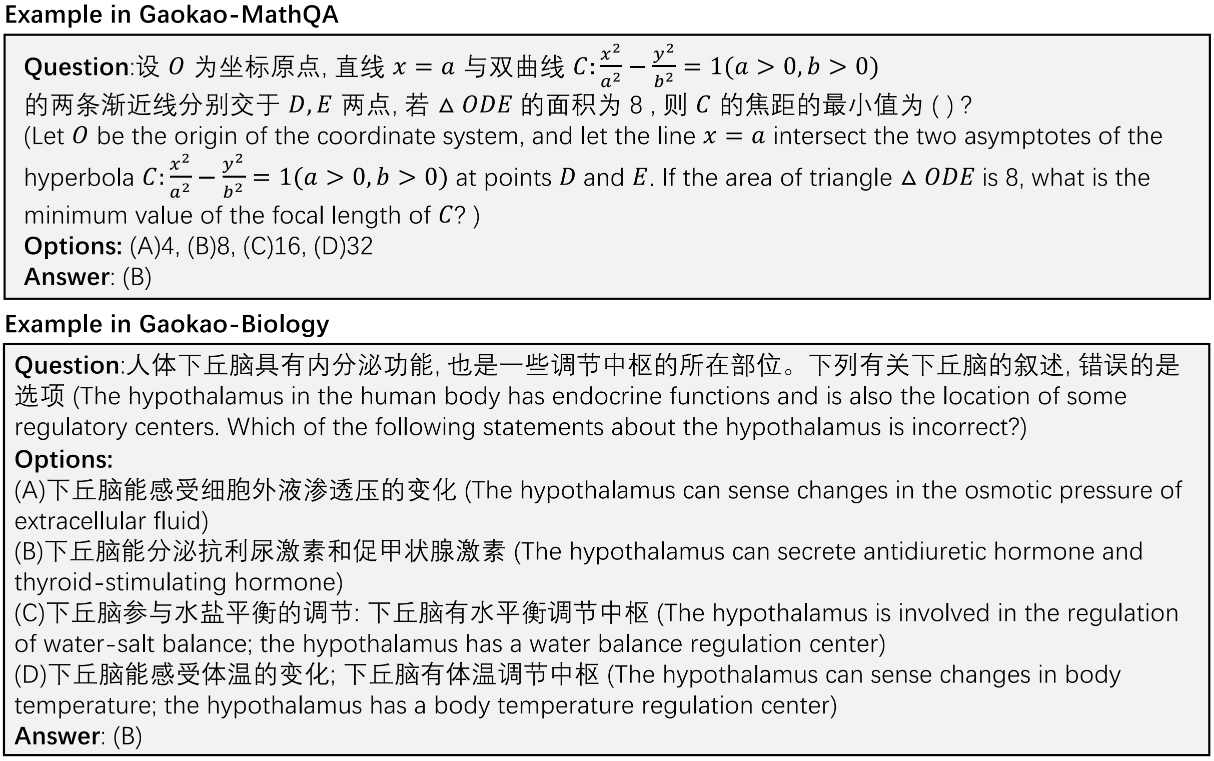
All experiments were conducted using the respective language models' API provided by Azure OpenAI Service[4]. The Azure OpenAI services offer two types of APIs: completion and chat completion. The completion API generates text based on prompts, while the chat completion API generates the next AI response based on the conversation history and new human input. For text-davinci-003 and few-shot ChatGPT, we use the completion API, and for zero-shot ChatGPT and GPT-4, we use the chat completion API. Notably, only the chat completion API is available for GPT-4 at present. We use a temperature of zero to generate output using greedy search and set the maximum number of tokens for generation to 2048. Additionally, we set the frequency penalty to zero and top p to 1, which are the default values for these APIs.

The Chat Completion API exhibits distinct properties in comparison to the Completion API. In a zero-shot context, the Chat Completion API has the potential to autonomously generate reasoning steps, eliminating the necessity for prompt engineering and potentially enhancing performance. For few-shot scenarios, it is imperative to adapt the few-shot examples into conversational history, as recommended in the Azure guidelines. The inquiry is transformed into a user input, while the AI's response is composed of a chain-of-thought explanation and answer. However, we have observed that the models, particularly ChatGPT, encounter difficulties in adhering to the pattern using the Chat Completion API. Consequently, we employ the Completion API to conduct few-shot experiments with ChatGPT, which is analogous to text-davinci-003, in order to gain a deeper understanding of the disparities between text-davinci-003 and ChatGPT. If a completion API for GPT-4 become accessible in the future, we will revise and update the few-shot outcomes accordingly.

C.2 Few-shot Examples Construction:

For AQuA-RAT, LogiQA and LSAT, we randomly sample five examples of medium sentence length of the test set from the provided training set. Similarly, for Gaokao and SAT, we randomly select five examples of medium sentence length from the dataset that was initially collected and exclude them from the test set. For JEC-QA, given that the test set is not publicly available, we take the first 1,000

[4] https://azure.microsoft.com/en-us/products/cognitive-services/openai-service



Example in Gaokao-MathQA

Question:设 O 为坐标原点, 直线 $x=a$ 与双曲线 $C:\frac{x^2}{a^2}-\frac{y^2}{b^2}=1(a>0,b>0)$ 的两条渐近线分别交于 D,E 两点, 若 $\Delta\ ODE$ 的面积为 8 , 则 C 的焦距的最小值为 ( ) ? (Let O be the origin of the coordinate system, and let the line $x=a$ intersect the two asymptotes of the hyperbola $C:\frac{x^2}{a^2}-\frac{y^2}{b^2}=1(a>0,b>0)$ at points D and E. If the area of triangle $\Delta\ ODE$ is 8, what is the minimum value of the focal length of C?)

Options: (A)4, (B)8, (C)16, (D)32

Answer: (B)

Example in Gaokao-Biology

Question:人体下丘脑具有内分泌功能, 也是一些调节中枢的所在部位。下列有关下丘脑的叙述, 错误的是选项 (The hypothalamus in the human body has endocrine functions and is also the location of some regulatory centers. Which of the following statements about the hypothalamus is incorrect?)

Options:

(A)下丘脑能感受细胞外液渗透压的变化 (The hypothalamus can sense changes in the osmotic pressure of extracellular fluid)

(B)下丘脑能分泌抗利尿激素和促甲状腺激素 (The hypothalamus can secrete antidiuretic hormone and thyroid-stimulating hormone)

(C)下丘脑参与水盐平衡的调节: 下丘脑有水平衡调节中枢 (The hypothalamus is involved in the regulation of water-salt balance; the hypothalamus has a water balance regulation center)

(D)下丘脑能感受体温的变化; 下丘脑有体温调节中枢 (The hypothalamus can sense changes in body temperature; the hypothalamus has a body temperature regulation center)

Answer: (B)

Figure 4: Data examples in Gaokao.

examples from the training set as the test set and again sample five examples of medium sentence length from the rest. For MATH, we use the same instances as in the appendices of Lewkowycz et al. (2022).

To generate explanations for few-shot CoT experiments, for AQuA-RAT and MATH, we use the existing rationales from these datasets. For Gaokao and SAT, we collected expert annotations. For LogiQA, JEC-QA and LSAT, we use ChatGPT to generate explanations given the questions and the answers. We release all CoT demonstrations in the Github repository.

D Qualitative Analysis Details

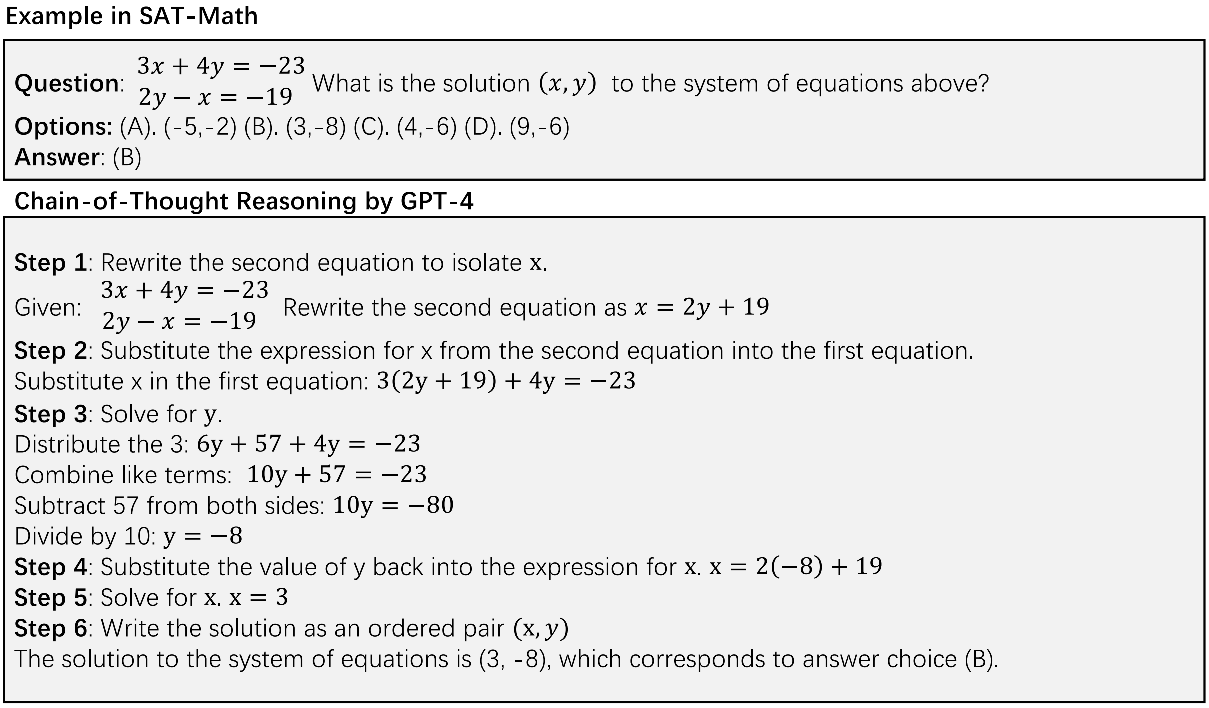
By closely examining the models' output explanations and analyzing their behavior patterns, we identify several strengths that highlight the capabilities of these models in handling various aspects of problem-solving. The models demonstrate remarkable performance in the following areas:

Good Understanding: The models excel in accurately comprehending the semantic meaning of context and questions. They effectively discern nuances, interpret complex questions, and parse intricate sentences, showcasing their strong natural language understanding skills. This capability enables them to grasp the core concepts of a problem and lays the foundation for subsequent reasoning and problem-solving steps.

Proficiency in Simple Reasoning and Deduction: The models are adept at handling tasks that require simple reasoning and deduction. They can draw straightforward conclusions, identify logical connections, and perform basic inference, which is crucial for addressing a wide variety of problems. Their ability to effectively perform simple reasoning tasks is an essential component of their overall problem-solving skillset.

For example, the model can comprehend "could be true except" is equals to "cannot be true". Also, taking a question in the LSAT-AR task as an example, it requires the model to place 8 books to a bookcase with three shelves following conditions: "*each shelf should have at least 2 books and more books should be placed on the bottom shelf than on the top shelf*". The model successfully deduced that "*there are at least 3 books on the bottom shelf and at most 2 books on the top shelf.*"

Grasping General Reasoning Process: The models demonstrate an ability to understand and generate the general idea of reasoning processes. They can identify the main components of a problem, recognize the structure of a solution, and out-



Example in SAT-Math

Question: $3x + 4y = -23$ $2y - x = -19$ What is the solution (x, y) to the system of equations above?

Options: (A). (-5,-2) (B). (3,-8) (C). (4,-6) (D). (9,-6)

Answer: (B)

Chain-of-Thought Reasoning by GPT-4

Step 1: Rewrite the second equation to isolate x.

Given: $3x + 4y = -23$ $2y - x = -19$ Rewrite the second equation as $x = 2y + 19$

Step 2: Substitute the expression for x from the second equation into the first equation.

Substitute x in the first equation: $3(2y + 19) + 4y = -23$

Step 3: Solve for y.

Distribute the 3: $6y + 57 + 4y = -23$

Combine like terms: $10y + 57 = -23$

Subtract 57 from both sides: $10y = -80$

Divide by 10: $y = -8$

Step 4: Substitute the value of y back into the expression for x. $x = 2(-8) + 19$

Step 5: Solve for x. $x = 3$

Step 6: Write the solution as an ordered pair (x, y)

The solution to the system of equations is (3, -8), which corresponds to answer choice (B).

Figure 5: Data example in SAT and corresponding CoT reasoning process.

line a high-level reasoning strategy. This capability allows them to generate meaningful explanations and provides a starting point for more detailed reasoning and problem-solving tasks.

These strengths indicate that the models have made significant progress in aligning with human problem-solving capabilities. However, there is still room for improvement, especially in complex reasoning tasks and domain-specific knowledge, as discussed in the subsequent section on weaknesses.

D.1 Weaknesses

Despite the significant strengths displayed by the models, there are certain limitations that need to be addressed to improve their overall performance. We outline these weaknesses based on the analysis of the models' output explanations:

Understanding:

- *Difficulty with Variable Substitution:* The models struggle to understand questions that require variable substitution, often failing to recognize the need for this operation and how it should be applied to solve the problem. This limitation can hinder their ability to tackle a wide range of mathematical and logical tasks. For instance, the model frequently struggles to answer chemistry questions that involve substituting a variable in a chemical equation with a chemical element and analyzing its properties.

- *Challenges with Complex Math Concepts and Symbols:* The models find it difficult to comprehend complex mathematical concepts and interpret the meaning of symbols, particularly when multiple symbols are involved. This weakness limits their ability to effectively address advanced mathematical problems.

- *Confusion with Similar Concepts:* The models can easily be confused by similar concepts or terms, sometimes leading to incorrect or misleading reasoning. For example, in the physics exam, the model is confused by the difference between vertical speed and horizontal speed of moving object. This issue underscores the need for better disambiguation and concept understanding techniques in future model iterations.

- *Difficulty in Handling Long Contexts:* The models are prone to being disrupted by long contexts, leading to a decline in their comprehension and reasoning abilities. Improving the models' capacity to maintain focus and process extensive information is essential for enhancing their performance in real-world scenarios.

Knowledge:

- *Insufficiency in Commonsense and Domain-Specific Knowledge:* The models occasionally demonstrate a lack of commonsense or domain-specific knowledge, which hinders their ability to generate plausible explanations and provide accurate answers. This limitation underscores the importance of incorporating diverse knowledge sources into the training data and exploring techniques that can more effectively integrate and access this information within the models. Moreover, it emphasizes the necessity to broaden the models' exposure to a wider array of subjects and fields, ensuring a more comprehensive understanding of various domains.

  For instance, given the conditions "*if Julio and Kevin both lead morning sessions, we know that Kevin and Rebecca must lead sessions that meet on the same day*," the model incorrectly deduces that "*Therefore, Rebecca must also lead a morning session.*" This indicates a lack of commonsense knowledge about the relationship between *morning* and *day*, leading to an erroneous explanation. Additionally, the model generally performs poorly on tasks requiring specific domain knowledge, such as law and chemistry.

- *Difficulty Identifying Correct Formulas:* The models occasionally struggle to recall and apply the appropriate formulas necessary to solve particular problems, especially in tasks that demand specialized knowledge or expertise. This shortcoming suggests that there is potential for improvement in the models' knowledge retrieval mechanisms and their ability to recognize the relevance of specific formulas to a given problem. Developing strategies to enhance the models' proficiency in identifying and applying correct formulas will be essential for improving their performance in tasks requiring a deep understanding of domain-specific concepts and techniques.

Addressing these weaknesses in knowledge will contribute to the development of more robust and versatile large language models, better equipped to tackle a broader range of human-centric tasks and exhibit a more comprehensive understanding of various domains.

Reasoning:

- *Challenges in Strict Logical Deduction:* The models frequently encounter difficulties when attempting to perform strict logical deduction accurately. Common issues include ignoring premise conditions, misconstruing sufficient and necessary conditions, or making errors in logical chaining. These types of errors are commonly observed in manual analyses.

  For instance, given a condition, "*If Myers is on the team, neither Ortega nor Paine can be*", and a solution, "*Ortega, Paine, Thomson, and Zayre are on the team*", the model incorrectly states that this solution is wrong because "*Paine and Ortega are on the team*", neglecting to first satisfy the premise condition "*If Myers is on the team*". Furthermore, the model demonstrates a misunderstanding of the difference between sufficient and necessary conditions in its explanation of another question and states: "*If Kayne is assigned to an ambassadorship, then so is Jaramillo. This constraint is essentially the same as the given constraint that if Jaramillo is assigned to one of the ambassadorships, then so is Kayne*".

  To address these limitations, it is essential to improve the models' abilities to recognize and apply logical rules and refine their understanding of logical structures.

- *Difficulty with Counterfactual Reasoning:* The models consistently struggle with counterfactual reasoning tasks. They have difficulty generating alternative scenarios, evaluating hypothetical outcomes, or exploring potential consequences based on varying assumptions. For instance, the models frequently make incorrect judgments for counterfactual questions in the LSAT-AR task: "*Which one of the following, if substituted for the constraint that [Constraint A], would have the same effect in determining the assignment?*" Enhancing the models' capabilities in handling counterfactual reasoning tasks is vital for developing a more comprehensive problem-solving skillset.

- *Struggles in Multi-hop Complex Reasoning:* The models have difficulty accurately executing multi-hop complex reasoning tasks, often displaying inconsistent logic, omitting inference steps, or producing flawed reasoning chains. To address a broader range of complex problems, it is crucial to improve the models' abilities to systematically navigate and process multi-step reasoning tasks.

- *Establishing Incorrect Conclusions and Contradictory Reasoning:* The models occasionally set an incorrect conclusion first and then generate contradictory reasoning based on that faulty foundation. This behavior emphasizes the need for improved reasoning verification and error correction techniques in the models' problem-solving processes.
- *Concealed Substitution of Concepts:* The models sometimes covertly substitute one concept with another similar one, leading to inaccurate or misleading reasoning. For example, in a biology exam, the model replaces the concept of "*isotopically labeled amino acids*" with "*isotopically labeled tRNA (a tool for transporting amino acids)*", resulting in erroneous reasoning. This issue underscores the importance of better concept disambiguation and reasoning coherence in future model iterations.
- *Difficulty in Identifying Solutions:* The models occasionally struggle to discover feasible solutions for specific problems, possibly due to limitations in their knowledge, reasoning capabilities, or problem-solving strategies. Addressing this shortcoming involves refining the models' ability to explore, evaluate, and select appropriate solutions based on the given problem context.
- *Vulnerability to Contextual Disturbance:* The reasoning ability of large language models is often easily disrupted by changes in the surrounding context. When the context is modified, the models may produce different deductions for the same condition, suggesting that the robustness of their reasoning ability is not yet sufficient. This observation emphasizes the need to develop models that can maintain consistent reasoning performance, even in the presence of varying contextual information, ensuring more reliable and stable problem-solving capabilities.

Calculation: The model is prone to making calculation errors, particularly when dealing with complex variable substitutions. This may be attributed to the inherent limitations of the model's computation process in handling mathematical operations, as well as its difficulty in parsing intricate relationships between variables. Consequently, the model may struggle to maintain accuracy and precision when attempting to solve problems involving advanced algebraic manipulations or multi-step calculations. To address this issue, future iterations of the model should focus on enhancing its mathematical reasoning capabilities and improving its ability to recognize and apply relevant mathematical rules. This could involve incorporating specialized modules or mechanisms specifically designed to handle complex calculations, variable substitutions, and numerical problem-solving tasks. By refining the model's ability to accurately process and solve intricate mathematical problems, we can expand its applicability across a broader range of disciplines and domains, ensuring a more comprehensive and robust problem-solving skillset.

By addressing these reasoning weaknesses, future large language models can be developed with more robust problem-solving capabilities, enabling them to effectively tackle a broader range of human-centric tasks and exhibit more sophisticated reasoning skills that align closely with human cognition.