FAC²E: Better Understanding Large Language Model Capabilities by Dissociating Language and Cognition

Xiaoqiang Wang[1,2], **Lingfei Wu**[3], **Tengfei Ma**[4] **and Bang Liu**[1,2†]
[1]DIRO & Institut Courtois, Université de Montréal
[2]Mila - Quebec AI Institute; [3]Anytime.AI; [4]Stony Brook University
{xiaoqiang.wang, bang.liu}@umontreal.ca
lwu@anytime-ai.com, tengfei.ma@stonybrook.edu

Abstract

Large language models (LLMs) are primarily evaluated by overall performance on various text understanding and generation *tasks*. However, such a paradigm fails to comprehensively differentiate the fine-grained language and cognitive *skills*, rendering the lack of sufficient interpretation to LLMs' *capabilities*. In this paper, we present **FAC²E**, a framework for **F**ine-gr**A**ined and **C**ognition-grounded LLMs' **C**apability **E**valuation. Specifically, we formulate LLMs' evaluation in a multi-dimensional and explainable manner by dissociating the language-related capabilities and the cognition-related ones. Besides, through extracting the intermediate reasoning from LLMs, we further break down the process of applying a specific capability into three sub-steps: recalling relevant knowledge, utilizing knowledge, and solving problems. Finally, FAC²E evaluates each sub-step of each fine-grained capability, providing a **two**-faceted diagnosis for LLMs. Utilizing FAC²E, we identify a common shortfall in knowledge utilization among models and propose a straightforward, knowledge-enhanced method to mitigate this issue. Our results not only showcase promising performance enhancements but also highlight a direction for future LLM advancements.

1 Introduction

Large language models (LLMs) (Brown et al., 2020), especially instruction-tuned LLMs (Ouyang et al., 2022; Bai et al., 2022; Touvron et al., 2023b; Chiang et al., 2023) revolutionized natural language processing and have surpassed human performance on tasks that require nontrivial reasoning (Guo et al., 2023; Malinka et al., 2023), while showing great potential in applications from conversational assistants (OpenAI, 2022; Achiam et al., 2023) to expertise problem-solving (Nori et al., 2023; Zhou et al., 2023; Suzgun and Kalai, 2024). However, despite the impressive performance, LLMs also show poor robustness on complex tasks (Ullman, 2023) and significantly inconsistent evaluation results under different settings, such as binary preference (Xu et al., 2023) and automatic metrics (Gudibande et al., 2023). Therefore, it is crucial to attain an overarching understanding of the capabilities and limitations of LLMs.

To address this challenge, some studies have assessed the performance of LLMs on different tasks based on independent benchmarks from various dimensions (Liang et al., 2022; Srivastava et al., 2023; Gao et al., 2023), such as commensense (Zellers et al., 2019), knowledge (Hendrycks et al., 2020; Yu et al., 2023), instruction-following (Gu et al., 2024), and trustworthy (Sun et al., 2024). Besides, motivated by building LLM-based AI assistants, other studies propose highly curated benchmarks with instance-level fine-grained annotations, such as difficulty and reasoning skills (Mialon et al., 2023; Ye et al., 2024), for holistic evaluation of LLMs.

However, the existing studies, assessing effectiveness across various tasks, provided limited insight into the models' true capabilities, as it only indicates their overall performance on specific datasets, without revealing the fine-grained capabilities acquired or their proficiency levels within the multiple capabilities involved. For instance, in the context of generative question answering, a model adept at extracting information but struggling to form a coherent understanding may exhibit similar overall performance to another model with profound understanding insights but difficulties in articulating accurate responses.

We argue that a fine-grained understanding of LLMs' capabilities can not only accurately unveil their inherent limitations, but also help us to better identify why one model outperforms the other

†Corresponding author. Canada CIFAR AI Chair.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13228–13243
November 12-16, 2024

Capability	Description	Skill Example	
LINGUISTIC KNOWLEDGE	Encoding grammatical concepts support linguistic operations regarding word meanings and their combinatorial processing.	Grammaticality: Semantics:	agreements, licensing, long-distance dependencies, and garden-path effects. synonymy, antonymy, and hypernymy.
FORMAL KNOWLEDGE	Conducting word-based formal reasoning through understanding lexical semantics.	Mechanism: Skill:	deductive, inductive, and analogical. numeric, logic, and manipulation.
WORLD MODELING	Understanding text based on given context and associating it with world knowledge.	Remember: Understand:	factual knowledge, context, and commensense. narrative structure and discourse comprehension.
SOCIAL MODELING	Infering mental state behind text and intended meaning beyond literal content.	Pragmatics: Theory-of-mind	polite deceits, irony, maxims of conversation, metaphor, indirect speech, and humor. unexpected content and unexpected transfer tasks.

Table 1: Formulation of cognition-grounded LLMs' capabilities. See Section 2.1 for details.

and how the different capabilities correlate. Additionally, such insights allow us to provide tailored guidance to improve training efficiency or facilitate more advanced model development.

In this paper, we propose $\texttt{FAC}^2\texttt{E}$, a fine-grained capability evaluation framework for LLMs. Specifically, $\texttt{FAC}^2\texttt{E}$ dissociates the language-related and cognition-related capabilities of LLMs and organizing them into four distinct axes: LINGUISTIC KNOWLEDGE, FORMAL KNOWLEDGE, WORLD MODELING, and SOCIAL MODELING. This categorization is grounded in neuroscience evidence manifesting that language processing and cognitive processes, like memory and reasoning, operate differently in the brain. Drawing from this insight, we adapt a range of existing benchmarks into a unified question-answering format. We then develop specific instructions for each capability, allowing $\texttt{FAC}^2\texttt{E}$ to evaluate LLMs through a method known as few-shot instruction-following.

Furthermore, we break down the application of a specific capability into three sub-steps: knowledge recall, knowledge utilization, and problem-solving, by iteratively drawing out the model's intermediate reasoning. After evaluating each sub-step, $\texttt{FAC}^2\texttt{E}$ can reveal the quality of knowledge encoded in the model, and effectiveness in applying relevant knowledge to solve practical problems, offering a more comprehensive evaluation than a single performance metric could.

Our findings reveal a notable gap in capabilities between open-source and proprietary models, especially for cognition-related capabilities. Additionally, we found that many models have difficulties in applying knowledge effectively. To address this, we suggest a knowledge-enhance remedy by incorporating relevant knowledge text as additional input. Experimental results show that it can help the backbone model (*e.g.* LLaMA 2) achieve approximately 90% of the performance of its instruction-tuned counterpart (*e.g.* LLaMA 2-Chat).

2 Methodology

In this section, we introduce $\texttt{FAC}^2\texttt{E}$ framework, designed for fine-grained and cognition-grounded LLMs' capability evaluation. Specifically, we first define the taxonomy for LLMs' capabilities based on the distinction between language and cognition, which is drawn upon insights from neuroscience (Fedorenko and Varley, 2016; Mahowald et al., 2023). Based on this, we transform a variety of existing benchmarks into the unified question-answering format, design capability-specific instruction, and frame $\texttt{FAC}^2\texttt{E}$ via few-shot instruction-following. Furthermore, we break down the evaluation process for each capability into a three-step reasoning approach. This involves identifying the knowledge pertinent to the input, examining how the model applies this knowledge in practical contexts, and assessing the effectiveness of its problem-solving. By evaluating each of these steps, $\texttt{FAC}^2\texttt{E}$ provides a comprehensive overview of the model's performance, offering a more nuanced understanding of LLMs' intrinsic capabilities.

2.1 Formulation of LLMs' Capabilities

Human language processing, long studied in cognitive science and neuroscience, robustly attributes language and cognition to different brain areas, namely "language network" and "multi-demand network" (Duncan, 2010; Scott et al., 2017). The former is sensitive to linguistic regularities and formal operations, with damage leading to linguistic deficits, while the latter responds actively to various cognitively demanding processes, such as reasoning and memory. Similarly, prior analyses have identified core linguistic regions (Zhang et al., 2024b) and language-independent knowledge neurons (Chen et al., 2023) in LLMs, represented by different parameter sets and subnetworks, each contributing distinctly to language and reasoning tasks.

Motivated by this separated relationship, we define the LLMs' capabilities as a 4-dimensional schema as shown in Table 1. Compared to the fine-grained definitions of high-level knowledge (Hendrycks et al., 2020; Yu et al., 2023) and reasoning skills (Ye et al., 2024) in related works, this formulation aims to dissociate language-related and cognition-related capabilities to define a broader range of both high-level and more fundamental low-level functionality. Additionally, it seeks to minimize the coupling between these capabilities to facilitate nuanced analysis (Sec. 3.1) and targeted improvement of the model (Sec. 3.2).

Linguistic Knowledge. To effectively generate language, LLMs first understand text at a basic linguistic level, including both grammar and semantics. Grammaticality encompasses the rules that govern language structure, spanning from the sounds (phonological) and words (lexical) to the arrangement of words in phrases and sentences. To capture this grammatical structure as comprehensively as possible, especially given the challenges conventional models face in benchmarks like BLiMP (Warstadt et al., 2020), we focus on four key skills: agreement (anaphora and subject-verb relationships), licensing (negative polarity items and reflexive pronouns), managing long-distance dependencies (filler-gap constructions and cleft sentences), and navigating garden-path sentences, which contain temporary ambiguities that must be resolved for correct understanding. For example, "*the horse raced past the barn fell,*" the initial interpretation is that the horse is racing, but upon reaching "*fell,*" it becomes clear that the horse is being raced by another (unnamed) entity.

Semantics, on the other hand, while more closely related to high-level cognitive understanding, within the context of Linguistic Knowledge, pertains to the meanings of individual words or lexical semantics (Geeraerts, 2009). This aspect is distinct from conceptual knowledge, which falls under other dimensions of LLM capabilities, highlighting the meaning-related understanding, such as synonymy, antonymy, and hypernymy.

Formal Knowledge. Beyond encoding linguistic structures and word meanings, an essential aspect of language capability involves understanding formal operations among words, or word-based reasoning. This means LLMs should be capable of recognizing relationships between words and deducing missing elements in a given pattern, such as completing analogies (*e.g.* "*man:woman :: king:_*"). FAC^{2}E includes three types of reasoning mechanisms—deductive, inductive, and analogical reasoning—between words (Bang et al., 2023), and includes three symbol-based formal skills: numeric (dealing with numbers), logic (applying logical operations), and manipulation (altering the inputs in a rule-based manner) (Wei et al., 2022b). An example task is concatenating the last letters of a word list ("*think, machine, learning*" → "*keg*").

World Modeling. To step towards cognitive capabilities, well-grounded comprehension of factual and commonsense knowledge is required. Precisely, we decompose this capability into two primary mechanisms: *remember* and *understand*, respectively modeling the retrieval-based and comprehension-based capability (Sugawara et al., 2020). Considering the versatility of knowledge sources, we instantiate the *remember* sub-capability as recalling factual knowledge (open-ended facts), reading comprehension (facts in context), and applying commonsense reasoning. Based on the multiple granularities of text comprehension and hierarchy of input text, we characterize the *understand* sub-capability as two skills: understanding narrative or event structure (paragraph-level), and discourse comprehension (document-level).

Social Modeling. The utility of human language lies in not only the understanding of the text itself but also the social context and mental states underlying communication (Adolphs, 2009), *i.e.* serving as a medium for information exchange between individuals. Specifically, there are a lot of phenomena about non-literal language comprehension in daily life, such as jokes, sarcasm, and indirect speech, successful LLMs should be capable of applying social inference skills to attain the intended meaning beyond the literal content. In this paper, we incorporate two kinds of social modeling into FAC^{2}E, encompassing pragmatics and theory of mind (ToM) reasoning. Pragmatics is evaluated by six kinds of dialogue, including polite deceits, irony, maxims of conversation, metaphor, indirect speech, and humor, while ToM is based on the "unexpected tasks" devised by Kosinski (2023).

2.2 FAC^{2}E

Based on the formulation of LLMs' capabilities, we collect input and output pairs from various benchmarks and modify collected instances, yielding a unified question-answering (QA) format. After that, following the widely adopted few-shot in-context learning (ICL) (Brown et al., 2020), we devise capability-specific instruction and frame FAC^{2}E via instruction following (Ouyang

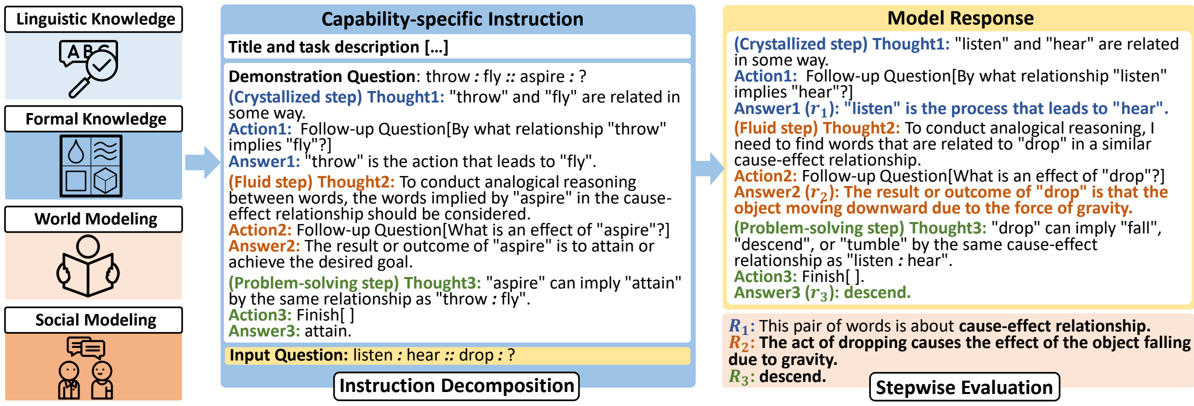

Figure 1: Illustration of FAC^{2}E pipeline. The input question is decomposed into two intermediate follow-up questions, which are used to help the model talk with itself to elicit reasoning sub-steps. FAC^{2}E evaluates each sub-step to reveal crystallized performance, fluid performance, and corresponding problem-solving performance. The content in the round parentheses is purely illustrative and is not part of the model input. The instruction has been omitted here for clarity. Please refer to Appendix B for full version example.

et al., 2022). We further leverage chain-of-thought (CoT) (Wei et al., 2022a,b) style prompting to elicit two intermediate reasoning steps from the model, namely *crystallized* step and *fluid* step. The terms "*crystallized*" and "*fluid*" are borrowed from Cattel's theory (Cattell, 1963), a foundational building block of cognitive science about the source of intelligence. Cattel's theory delineates that crystallized intelligence is semantic knowledge from past experiences, fluid intelligence is the ability to navigate novel situations, and problem-solving uses both. Therefore, we add two intermediate steps to operationalize the two kinds of mechanisms, and aim to measure how well the model recalls and applies knowledge. Last, we compare the intermediate results with the reference answers to score the reasoning sub-steps, hence providing an assessment of the crystallized performance and fluid performance as well as problem-solving performance.

As depicted in Figure 1, the pipeline FAC^{2}E can be divided into three steps, including capability-specific instruction design, instruction decomposition, and stepwise evaluation. Specifically, we devise natural instruction for each capability-related task. Borrowing the widely used template schema of instruction-following (Wang et al., 2022b; Mishra et al., 2022), the capability-specific instruction $\mathcal{I}_c$ is comprised of three parts: title, task description and few-shot demonstrations. Precisely, the title defines a given QA task in high-level natural language and highlights the associated skills, while the task description not only presents a complete clarification of how an input text is expected to be mapped to final output, but also define the output of reasoning sub-steps through instruction decomposition. After that, following the given task description, a few in-context demonstrations are provided to better steer the response generation. At last, we collect the response results for each reasoning sub-steps, denoted as $\{r_i\}_{i=1}^3$, and respectively evaluate them with the reference answer, denoted as $\{R_i\}_{i=1}^3$, which are directly extracted from corresponding benchmarks. Formally, the procedure can be represented as:

$$\{r_i\}_{i=1}^3 = \mathcal{M}(\mathcal{I}_c) \tag{1}$$

$$s_i = \text{Criterion}_i(r_i, R_i) \tag{2}$$

where $\mathcal{M}$ denotes the examined LLM, while Criterion_i and s_i represent the employed automatic metric and corresponding score, respectively.

Instruction decomposition. We leverage CoT-like iterative prompting strategy to elicit the intermediate reasoning from the model to frame crystallized and fluid steps. Differing from the standard CoT that outputs a continuous rationale before the final answer, we first decompose the given question as follow-up sub-questions. After that, these sub-questions are used to help the model talk with itself to respectively discover (i) what knowledge this question is about, (ii) how to apply relevant knowledge to the given instance, and (iii) the final answer. In other words, FAC^{2}E convert the CoT continuous rationale into easily parseable multi-step rationales, which externalizes reasoning of the model (Shwartz et al., 2020; Zhou et al., 2022b) and enables the evaluation of crystallized performance and fluid performance. Formally, as depicted in Figure 1, we expect that the model outputs as: `[Thought]`, `[Action]`, `[Answer]` , where `[Thought]` can reason about the current situation, `[Action]`

Capability	Benchmark	#Samples	Length	QA
Agreements	BLiMP (Warstadt et al., 2020)	400	33	M
Licensing	Marvin and Linzen (2018)	1,500	36	M
Long-distance dependency	Wilcox et al. (2019)	660	48	M
Garden-path effects	Futrell et al. (2018)	450	40	M
Lexical semantics	Petersen and Potts (2023)	1,000	8	M
Deductive	Bang et al. (2023)	1,600	18	M
Inductive	Bang et al. (2023)	1,500	16	M
Analogical	Webb et al. (2023)	800	4	G
Numeric	MAWPS (Koncel-Kedziorski et al., 2016)	400	33	G
Logic	Tian et al. (2021)	1,000	87	G
Manipulatation	Wei et al. (2022b)	250	31	G
Factual Knowledge	LAMA (Petroni et al., 2019)	500	48	G
Reading Comprehension	Dua et al. (2019)	1,000	195	M
Commonsense	Talmor et al. (2019)	800	13	M
Discourse	Wang et al. (2023c)	400	439	G
Narrative	Xu et al. (2022)	200	395	G
Pragmatics	Hu et al. (2023)	150	288	M
Theory of mind	Ullman (2023)	80	152	G

Table 2: Breakdown statistics on source benchmarks employed by FAC^{2}E and re-formulation types (**G**enerative or **M**ultiple-choice QA), where Length refers to the average input length of examples.

can be either (1) [Follow-up Question], which returns a sub-question, or (2) [Finish], and [Answer] is extracted as the reasoning result of a sub-step.

Stepwise evaluation. Given the reasoning results of three sub-steps, *i.e.* $\{r_i\}_{i=1}^3$, we engage automatic metrics as the criterion to evaluate them. Specifically, r_1 and r_2 are free-form rationales for intermediate reasoning steps. Considering the diversity of rationale generation, we resort to BARTScore-Recall (Yuan et al., 2021), one of the most superior metrics for natural language generation to evaluate the quality of generated rationale automatically. BARTScore-Recall gauges how many semantic content units from reference texts are covered by the generated candidates, and will not penalize the redundant and instance-specific information in the model response. For the last response r_3, since it is expected to be the final answer for the given question, it is evaluated by the BARTScore-Recall (Yuan et al., 2021) or accuracy for generative QA re-formulation and multiple choice QA re-formulation, respectively.

3 Experiments

Evaluation data construction. In Table 2, we present a collection of 17 widely adopted English benchmarks and modify the corresponding input-output into a unified QA format, *i.e.* generative QA or multiple-choice QA. The choice to utilize these benchmarks is rooted in considerations of data quality, availability, and specificity of focus; hence, some widely recognized benchmarks may not be included for these reasons. For example, PIQA (Bisk et al., 2020) focuses on physical commonsense, which, while valuable, represents only a single facet of commonsense reasoning. In contrast, MMLU (Hendrycks et al., 2020) encompasses a broad spectrum of subjects, but requires both commonsense and contextual understanding, which might not align with our goal of ensuring a broad range of capabilities while minimizing the coupling between abilities during evaluation.

Model	Model size	Pre-training	Fine-tuning
T5	11B	1.0T tokens	✗
Flan-T5	11B	as above	IT
Flan-Alpaca	11B	as above	IT
LLaMA	7B	1.4T tokens	✗
Alpaca	7B	as above	IT
Vicuna	7B	as above	IT
TÜLU 1	7B,13B,30B,65B	as above	IT
LLaMA 2	7B	2.0T tokens	✗
LLaMA 2-Chat	7B	as above	IT+RLHF
LLaMA 3-Instruct	8B	15.0T	IT+RLHF
LLaMA 3.1-Instruct	8B	16.4T	IT+RLHF
GPT-3.5	175B	-	IT+RLHF
InstructGPT	175B	-	IT+RLHF
GPT-4	-	-	IT+RLHF
Bard	137B	-	IT+RLHF

Table 3: Statistics of examined LLMs, where fine-tuning techniques indicating whether the model is built with instruction tuning (IT) and reinforcement learning with human feedback (RLHF) or not.

The reference answers of the benchmarks are directly used as the final answer R_3, while the reference rationales (R_1 and R_2) for the intermediate reasoning steps are constructed automatically. Specifically, on the one hand, R_1, *i.e.* the reference rationale for the first reasoning step, is based on the rationale templates and the gold labels of the employed benchmarks. For example, when evaluating the grammaticality regarding negative polarity item (NPI) licensing, the rationale template for R_1 could be "*The word* [ ] *is a negative polarity item: it can only be used in the scope of negation.*". The blank is then filled with gold labels (licensing contexts or trigger words), such as "*any*", "*ever*", and "*even*", to build the final R_1 for corresponding NPI licensing samples. On the other hand, R_2, *i.e.* the reference rationale for the second reasoning step is built on the instance-wise annotations of human evaluation publicly released by the authors of corresponding benchmarks, which annotates necessary explanations as well as final answer R_3 for a given question. Although this will leave few benchmarks available and lead to a limited number of evaluation data, it provides relatively reliable references and especially enables reproducible evaluation.

Examined models. As summarized in Table 3, the examined LLMs can be categorized into publicly available open-source models and proprietary ones whose responses are provided through private APIs. Open-source models include three backbone

Model	Linguistic Knowledge			Formal Knowledge			World Modeling			Social Modeling		
	s_1	s_2	s_3	s_1	s_2	s_3	s_1	s_2	s_3	s_1	s_2	s_3
T5	83.99	26.39	47.39	77.97	28.10	33.26	74.61	24.74	26.53	66.79	18.31	19.16
Flan-T5	84.96	42.50	64.12	80.10	35.22	42.58	74.82	36.13	34.64	67.73	27.23	21.27
Flan-Alpaca	85.25	39.41	60.15	79.88	34.99	41.72	75.54	37.42	36.57	68.38	28.74	23.82
LLaMA	85.34	31.36	53.77	80.02	31.18	40.04	75.57	27.46	30.47	67.54	20.14	20.75
Alpaca	86.02	45.78	68.39	82.03	38.10	53.85	77.30	41.91	49.42	69.93	29.61	32.96
Vicuna	85.23	47.45	72.66	84.35	40.10	57.07	75.33	43.38	44.37	65.68	26.87	30.63
TÜLU 1	84.14	45.84	70.72	82.32	39.29	51.21	75.90	43.30	40.80	69.73	26.77	27.02
LLaMA 2	83.19	34.56	57.89	82.19	34.18	46.15	77.48	34.84	40.92	68.22	24.74	24.20
LLaMA 2-Chat	87.04	48.95	74.46	84.05	43.21	57.13	78.43	46.09	44.46	71.06	28.89	29.59
LLaMA 3-Instruct	87.78	50.20	78.57	85.81	43.88	60.79	80.18	46.74	47.72	76.99	39.10	30.52
LLaMA 3.1-Instruct	88.21	51.47	82.71	87.33	44.94	65.01	81.54	47.11	50.39	77.78	30.51	31.75
GPT-3.5	87.91	53.91	82.72	85.93	45.20	70.47	81.53	53.18	67.68	77.23	36.34	40.56
InstructGPT	88.52	55.50	85.19	85.12	44.18	67.48	80.34	51.78	65.16	74.17	39.90	45.95
GPT-4	89.32	58.98	89.62	87.64	47.99	75.97	81.86	54.78	69.43	81.24	40.99	45.71
Bard	87.74	52.37	86.16	86.97	46.08	71.62	79.30	49.09	61.31	78.64	38.27	42.53

Table 4: Quantitative results in terms of four capability dimensions. As stated in Section 2.2, s_1, s_2, and s_3 refer to crystallized performance, fluid performance, and problem-solving performance, respectively. The color of the text indicates the model type: blue for open-source and red for proprietary models. The shade represents the ranking, where the darker shade represents the highest score, and the lighter shade represents the second highest score.

models, *i.e.* T5 (Raffel et al., 2020), LLaMA (Touvron et al., 2023a) and LLaMA 2 (Touvron et al., 2023b), which are pre-trained on large scale corpus and not applied to any fine-tuning. Initialized with T5, Flan-T5 (Longpre et al., 2023) and Flan-Alpaca (Chia et al., 2023) are instruction-tuned on Flan V2 (Longpre et al., 2023) and Alpaca (Taori et al., 2023), respectively. Built on LLaMA, Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023) are instruction-tuned with responses generated by GPT-3.5, while TÜLU 1 (Wang et al., 2023d) are instruction-tuned with a mixture of both manually curated and distilled dataset. Based on LLaMA 2, LLaMA 2-Chat (Touvron et al., 2023b) is firstly instruction-tuned with high-quality collected annotations, and then aligned with human preferences for the chat use case. LLaMA 3-Instruct and LLaMA 3.1-Instruct respectively fine-tune LLaMA 3 and LLaMA 3.1 to better understand and follow human instructions. Besides, to perform a fair comparison w.r.t instruction-tuning dataset, we also evaluate LLaMA checkpoints fine-tuned on other datasets, such as Flan V2 (Longpre et al., 2023) (human-written), Alpaca (model-generated) (Taori et al., 2023), ShareGPT (user prompt with model response).

Proprietary models consist of OpenAI's GPT-3.5 (`gpt-3.5-turbo`) (OpenAI, 2022), InstructGPT (`gpt-3.5-turbo-instruct`) (Ouyang et al., 2022), GPT-4 (`gpt-4-turbo`) (Achiam et al., 2023), and Google's Bard (Google, 2023) (also known as Gemini (Team et al., 2023)).

3.1 Main results

The difference in problem-solving performance is significantly greater than the difference in crystallized performance. On the one hand, in terms of problem-solving performance (s_3), as shown in Table 4, open-source models usually underperform the proprietary ones across various capabilities, especially cognition-related ones, such as world modeling and social modeling. For example, the most competitive open-source model LLaMA 3.1-Instruct achieves a 50.39 accuracy in the world modeling dimension, while GPT-4 produces a performance of 69.43, excelling Alpaca by a substantial margin (around 40%). A similar conclusion can also be drawn from the other dimensions, such as the best problem-solving performance of proprietary models exceeds that of open-source models by about 20%, 32%, and 40% in linguistic knowledge, formal knowledge, and social modeling, respectively. On the other hand, in terms of crystallized performance (s_1), there is a rather smaller gap between open-source and proprietary models compared to problem-solving accuracy. For example, the maximal difference of s_1 in the world modeling among all the examined models is about 9%, *i.e.* GPT-4's 81.86 vs. T5's 74.61, while their difference of s_3 is GPT-4's 69.43 vs. T5's 26.53. This inconsistency between s_1 and s_3 can also be observed in other capability dimensions,

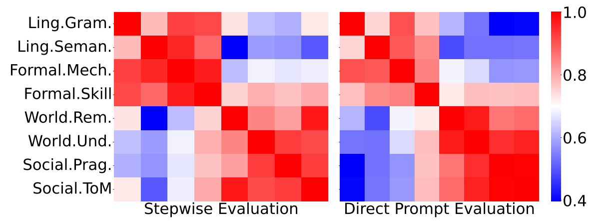

Figure 2: Pairwise correlation of problem-solving performance (s_3) among different capabilities. Please refer to Table 1 for full label names.

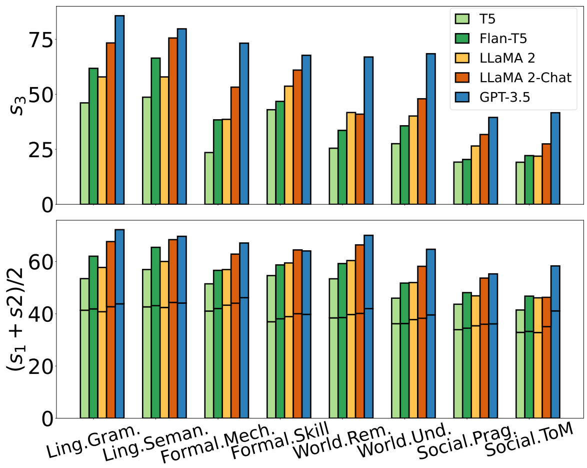

Figure 3: Bar diagram illustrating the relationship between problem-solving performance (s_3) and intermediate performance ($(s_1+s_2)/2$). Each bar of intermediate performance is divided into two stacked segments, the *lower* one denotes s_1, while the *upper* one denotes s_2.

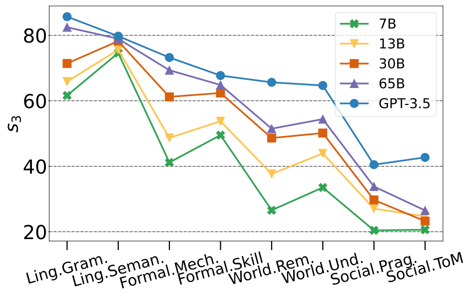

Figure 4: Problem-solving performance of instruction-tuned LLaMA with different model sizes.

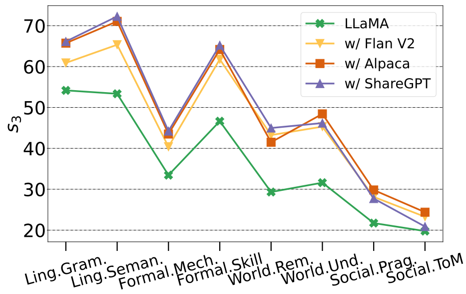

Figure 5: Problem-solving performance of LLaMA on different instruction-tuning datasets.

potentially showing that either pre-training or fine-tuning of LLMs can encode sufficient knowledge into the model, but the final task performance does not just depend on the amount or quality of knowledge.

Notably, the crystallized performance (s1) of LLaMA 3-Instruct and LLaMA 3.1-Instruct has significantly improved compared to LLaMA 2-Chat. For example, in social modeling, LLaMA 3-Instruct scored 76.99, and LLaMA 3.1-Instruct scored 77.78, compared to LLaMA 2-Chat's 71.06. This suggests that these models better encode knowledge relevant to higher-level cognitive tasks, likely due to training and fine-tuning LLaMA 3 on higher-quality and larger-scale data compared to the LLaMA 2 model.

Linguistic capabilities show a relatively weak correlation with cognitive capabilities. Figure 2 presents the correlation results between different capabilities. Both the language-related and cognition-related capabilities exhibit stronger (Pearson's r > 0.7 (Krippendorff, 2004)) intra-dimension correlation (*e.g.* world modeling vs. social modeling) when compared to inter-dimension correlation (*e.g.* world modeling vs. linguistic knowledge). This indicates that excellence in language processing does not necessarily equate to a similar level of cognitive capability. These results can also be observed from direct prompt evaluation, which assesses problem-solving performance (s_3) without prompting intermediate reasoning steps, further demonstrating the rationality of dissociating language and cognition.

A possible reason behind it could be that dedicated structures of the model or subsets of parameters are highly correlated with language (Zhang et al., 2024b; Tang et al., 2024), whereas others serve as cognition (Chen et al., 2023), and they are optimized at different training stages and function as different mechanisms during inference, which has been verified by recent studies on knowledge locating and editing of LLMs Dai et al. (2022); Meng et al. (2022); Zhang et al. (2024a).

The crystallized step impacts problem-solving more than the fluid step. Figure 3 illustrates the relationship between problem-solving performance (s_3) and sum of crystallized (s_1) and fluid (s_2) performance. Both s_1 and s_2 make a difference to the final s_3, showing that problem-solving not only depends on the amount or quality of stored knowledge but also is reflective of the effectiveness of knowledge utilization. For example, LLaMA 2 underperforms LLaMA 2-Chat in terms of s_3 across various capabilities. When taking a closer look at the intermediate results, we can observe that both the models do well in crystallized step, but LLaMA 2 shows an worse result in fluid step, leading the worse problem-solving performance than LLaMA 2-Chat. Besides, all of the open-source models exhibit relatively poor fluid performance w.r.t. GPT-3.5, especially those that are pre-trained but not instruction-tuned, such as T5 and LLaMA 2. This implies a solution to improve problem-solving

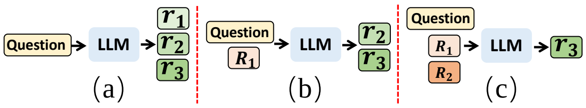

Figure 6: Comparing knowledge-enhanced baselines $\mathcal{M}$+R_1 (b), $\mathcal{M}$+R_1+R_2 (c) and the original setting (a).

performance, *i.e.* boosting the efficacy of knowledge utilization. Section 3.2 presents a knowledge-enhanced method to demonstrate this solution.

Both the model size and the quality of fine-tuning dataset affect the capabilities of LLMs. Firstly, backbone models play a critical role in building superior models. For example, as shown in Table 4, fine-tuned on the same dataset, LLaMA-based Alpaca performs better than T5-based Flan-Alpaca, and the larger scale proprietary models show a greater advantage over other open-source models. In addition, scaling open-source models does improve both language and cognitive capability. As shown in Figure 4, problem-solving performance across various capabilities increases as the model size increases, and 65B achieves the best performance. In particular, the level of formal knowledge of 65B is close to that of GPT-3.5. Last, but not least, there is no significant performance difference among various open-source instruction-tuning datasets whether it is comprised of human-written instruction or not. As illustrated in Figure 5, there is not a single best instruction tuning dataset across all tasks, indicating different datasets bring different benefits to LLMs' capabilities. This finding is consistent with the recent success of a mixture of instruction-tuning datasets or expert LLMs (Jiang et al., 2024; Xia et al., 2024).

3.2 Boosting LLMs with Injected Knowledge

Based on the above analysis showcasing the limitations of the crystallized performance of existing LLMs, as illustrated in Figure 6, we propose a knowledge-enhanced approach. Specifically, for each given instance with a question and answer, the first baseline, denoted as $\mathcal{M}$+R_1, append the first reference rationale, *i.e.* R_1, to the input question with string concatenation. Then the augmented input is fed into the model with the same instruction as the examined model. Note that we also remove the first triplet of ⟨ `[thought]`, `[action]`, `[answer]` ⟩ in the input demonstrations for the $\mathcal{M}$+R_1 baseline because we have provided the corresponding reference rationale. As a comparison, following a similar procedure, we also construct another baseline by incorporating both R_1 and R_2 into the model, denoted as $\mathcal{M}$+R_1+R_2.

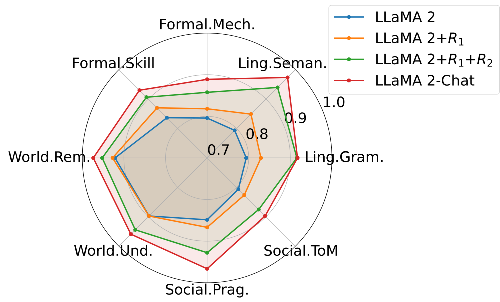

Figure 7: A 8-dimensional capability map of LLaMA 2 model when augmented with different knowledge text. The score is re-scaled through max-min normalization among each capability for clarity.

Taking LLaMA 2, performing moderately in Table 4, as the backbone model, the multifaceted results are summarized in Figure 7. We can observe that explicit injected rationales both R_1 and R_2 can substantially improve the problem-solving performance, and R_2 results in more improvements than R_1. Specifically, R_1 contributes slightly to language-related capabilities, such as linguistic semantics and formal skills, while R_2 brings about significant improvements to cognition-related ones, especially social modeling. Overall, the knowledge-enhanced LLaMA 2 baseline can achieve approximately 90% performance compared to the corresponding instruction-tuned variant.

4 Related Works

Evaluation of LLMs. LLMs are initially assessed on various understanding (Wang et al., 2018, 2019) and generation tasks (Pilault et al., 2020; Thompson and Post, 2020). With the increasing emphasis on the trustworthiness of models, dedicated benchmarks are proposed to evaluate robustness (Yang et al., 2022; Wang et al., 2023b), hallucination (Li et al., 2023; Belyi et al., 2024), and generalizability (Wang et al., 2023a). Recent-emerged works evaluate LLMs using another evaluator LLM (Kim et al., 2024a,b) or on a holistic benchmark (Mialon et al., 2023; Rein et al., 2023). Chia et al. (2023) conduct evaluation from problem-solving, writing, and human alignments, while Ye et al. (2024) annotates a single instance with a set of skills, including logical thinking, background knowledge, problem handling, and user alignment. Although they provide fine-grained analysis of LLMs' capability, they suffer from the limited number of testing instances, *e.g.* 1,700 of Ye et al. (2024) and overlook the language-related low-level capabilities.

Cognition-inspired intelligence evaluation. How to define and evaluate intelligence is widely investigated by both cognitive science and AI benchmarking (Cattell, 1963; Rogers et al., 2023). In the MRC evaluation, Chollet (2019) describes intelligence as skill-acquisition efficiency, while Sugawara et al. (2020), Wang et al. (2022a) and Ray Choudhury et al. (2022) propose to benchmark MRC through reasoning skills and steps a system would be "reading slowly". As for the LLMs, Mahowald et al. (2023) summarize extensive neuroscience evidence of human language and propose a conceptual framework with formal and functional competencies, which largely motivated the design of $\mathtt{FAC^2E}$. Compared to Mahowald et al. (2023), $\mathtt{FAC^2E}$ introduces more concrete NLP tasks with a more comprehensive capability system and leverages stepwise evaluation to improve the accuracy of assessments.

5 Conclusion

We present $\mathtt{FAC^2E}$, defining a fine-grained capability evaluation framework for LLMs, which decomposes each capability into sub-steps to assess performance of knowledge recalling and utilization as well as problem-solving. $\mathtt{FAC^2E}$ reveals the limitations of existing LLMs in knowledge utilization and provides a knowledge-enhanced remedy for it. Empirical results demonstrate its effectiveness.

Limitations

Our work proposes a fine-grained and cognition-grounded capability evaluation framework for LLMs, namely $\mathtt{FAC^2E}$, which is based on the dissociated relationship of language and cognition, and evaluating the intermediate reasoning steps of LLMs. The limitations are two-fold, including data quality and domain generalizability.

On the other hand, motivated by a variety of empirical evidence from both neuroscience and probing experiments of LLMs, we formulate $\mathtt{FAC^2E}$ as four capability dimensions, then re-formulate instances from multiple existing benchmarks and conduct stepwise evaluation. Although we try to ensure that the employed datasets are as consistent and targeted with the defined dimensions as possible, the kind of evaluation data construction might not reflect the required skills accurately. For example, an instance can cover more than one language or cognition capabilities As discussed in Section 3.1, different capabilities are correlated to each other to some extent, Besides, the reference rationales, *i.e.* the gold standard of the intermediate reasoning step, are based on the human annotations from original benchmarks, leading to the inconsistency of reference answers and a limited number of available data, which might bias the evaluation results. One remedy to these incidental issues could be building a new holistic benchmark with fine-grained annotations following our proposed schema. We regard it as our future work and deem designing a new annotation specification a promising direction.

On the other hand, our $\mathtt{FAC^2E}$ only examined LLMs on general domains and English input, ignoring the domain-specific and multilingual application. In particular, the reasoning process of LLMs may expose social bias encoded in these models, such as race and gender (Lucy and Bamman, 2021). Therefore, additional evaluation protocols considering potential risks to user safety are left for our future work.

Ethics Statement

We introduce $\mathtt{FAC^2E}$, a fine-grained and cognition-grounded capability evaluation framework for LLMs, and conduct evaluation experiments on publicly available datasets which are widely used in related research. Although LLMs have the potential to cause harm at the individual and societal levels (Gonen and Goldberg, 2019), our $\mathtt{FAC^2E}$ aims to provide a deep understanding of the capabilities and limitations of LLMs, potentially making the risks from the LLMs more predictable.

Acknowledgements

This work is supported by the Canada CIFAR AI Chair Program and the Canada NSERC Discovery Grant (RGPIN-2021-03115).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ralph Adolphs. 2009. The social brain: neural basis of social knowledge. *Annual review of psychology*, 60:693–716.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. 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 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Masha Belyi, Robert Friel, Shuai Shao, and Atindriyo Sanyal. 2024. Luna: An evaluation foundation model to catch language model hallucinations with high accuracy and low cost. *arXiv preprint arXiv:2406.00975*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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.

Raymond B Cattell. 1963. Theory of fluid and crystallized intelligence: A critical experiment. *Journal of educational psychology*, 54(1):1.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. *arXiv preprint arXiv:2308.13198*.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

John Duncan. 2010. The multiple-demand (md) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences*, 14(4):172–179.

Evelina Fedorenko and Rosemary Varley. 2016. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369(1):132–153.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Dirk Geeraerts. 2009. *Theories of lexical semantics*. OUP Oxford.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Google. 2023. bard.

Zihui Gu, Xingwu Sun, Fengzong Lian, Zhanhui Kang, Chengzhong Xu, and Ju Fan. 2024. Dingo: Towards diverse and fine-grained instruction-following evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18108–18116.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1152–1157, San Diego, California. Association for Computational Linguistics.

Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models. *Preprint at https://arxiv. org/abs/2302.02083*.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models. *arXiv e-prints*, pages arXiv–2301.

Kamil Malinka, Martin Peresíni, Anton Firc, Ondrej Hujnák, and Filip Janus. 2023. On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, pages 47–53.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Erika Petersen and Christopher Potts. 2023. Lexical semantics with large language models: A case study of English "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490–511, Dubrovnik, Croatia. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models "understand" language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Terri L Scott, Jeanne Gallée, and Evelina Fedorenko. 2017. A new fun and robust version of an fmri localizer for the frontotemporal language system. *Cognitive neuroscience*, 8(3):167–176.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

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.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023a. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*.

Jindong Wang, HU Xixu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, et al. 2023b. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Longyue Wang, Zefeng Du, Donghuai Liu, Cai Deng, Dian Yu, Haiyun Jiang, Yan Wang, Leyang Cui, Shuming Shi, and Zhaopeng Tu. 2023c. Disco-bench: A discourse-aware evaluation benchmark for language modelling. *arXiv preprint arXiv:2307.08074*.

Xiaoqiang Wang, Bang Liu, Fangli Xu, Bo Long, Siliang Tang, and Lingfei Wu. 2022a. Feeding what you need by understanding what you learned. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5858–5874, Dublin, Ireland. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023d. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

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

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. 2019. Structural supervision improves learning of non-local grammatical dependencies. 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 3302–3312, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2022. Glue-x: Evaluating natural language understanding models from an out-of-distribution generalization perspective. *arXiv preprint arXiv:2211.08073*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. 2024a. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024b. Unveiling linguistic regions in large language models. *arXiv preprint arXiv:2402.14700*.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, et al. 2023. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022a. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022b. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

A Implementation Details

All of the examined open-source models are based on HuggingFace Transformers package (Wolf et al., 2020). Their model cards, *i.e.* checkpoints consist of:

- T5 (t5-11b),
- Flan-T5 (google/flan-t5-xxl),
- Flan-Alpaca (declare-lab/flan-alpaca-xxl),
- LLaMA [1],
- Alpaca and LLaMA on Alpaca (allenai/open-instruct-stanford-alpaca-13b),
- Vicuna (lmsys/vicuna-7b-v1.1),
- TÜLU 1 (allenai/tulu-7b, allenai/tulu-13b, allenai/tulu-30b, allenai/tulu-65b),
- LLaMA on Flan V2 (allenai/open-instruct-flan-v2-13b),
- LLaMA on ShareGPT (allenai/open-instruct-sharegpt-13b),
- LLaMA 2 (meta-llama/Llama-2-7b-hf),
- LLaMA 2-Chat (meta-llama/Llama-2-7b-chat-hf)

For the response generation of each target model, as suggested by Wei et al. (2022b); Zhou et al. (2022a), we employ 4-shot instruction-following settings, *i.e.* 4 in-context demonstrations in the input prompt, set the temperature to 0.7 and set the max length of generated sequences as 1024. For automatic metrics, we leverage the official implementation of BARTScore (Yuan et al., 2021).

After collecting instances from various benchmarks as summarized in Table 2, we remove those instances where the input length is longer than 2048, maximal context length during training except T5, Flan-T5, and Flan-Alpaca,

We conduct evaluation experiments on 2 A100 GPUs and report the average results of a total of ten runs for each model on each benchmark. For the capabilities involving multiple benchmarks, the overall score are calculated as the arithmetic mean of crystallized performance (s_1), fluid performance (s_2), or problem-solving performance (s_3).

[1] `https://huggingface.co/docs/transformers/main/en/model_doc/llama`

B Instruction Design

See Figure 8 and Figure 9 for full version example of capability-specific instruction when evaluating the analogical reasoning and grammaticality.

Instruction: Solve a question-answering task by conducting analogical reasoning between words.
Given three words, i.e. A, B, and C in a format of A:B::C:?, which means that A implies B by some relationship, reason with this relationship and predict a word D such that C implies D by the same relationship. In other words, A:B is a reference pair in some relationship, complete the pair of C:D in the same relationship as A:B.
Please solve the task by interleaving Thought, Action, and Answer steps. Thought can reason about the current situation, and Action can be the following two types:
(1) Follow-up Question[question], which returns a sub-question with a single answer that helps solve the original question.
(2) Finish[], which means no more sub-questions. The final answer should be generated in the following line.
[Demonstration Question]: throw:fly::aspire:?
[Thought 1]: "throw" and "fly" are related in some way.
[Action 1]: Follow-up Question[By what relationship "throw" implies "fly"?]
[Answer 1]: "throw" is the action that leads to flying "fly".
[Thought 2]: To conduct analogical reasoning between words, the words implied by "aspire" in the cause-effect relationship should be considered.
[Action 2]: Follow-up Question[What is an effect of "aspire"?]
[Answer 2]: The result or outcome of "aspire" is to attain or achieve the desired goal.
[Thought 3]: "aspire" can imply "attain" by the same relationship as "throw:fly".
[Action 3]: Finish[]
[Answer 3]: attain.
[More Demonstration Questions] [...]
[Input Question]: listen:hear::drop:?

Figure 8: Full version example of the capability-specific instruction.

Instruction: Solve a question-answering task judging which one of the minimal pairs is acceptable and grammatical.
The minimal pairs consist of two sentences that differ by a few words, one of them is grammatical, but another is ungrammatical.
Please solve the task by interleaving Thought, Action, and Answer steps. Thought can reason about the current situation, and Action can be the following two types:
(1) Follow-up Question[question], which returns a sub-question with a single answer that helps solve the original question.
(2) Finish[], which means no more sub-questions. The final answer should be generated in the following line.
[Demonstration Question]: Which sentence of the following two sentences is grammatical?
FirstSentence[No author that no senators liked has had any success.]
SecondSentence[The author that no senators liked has had any success.]
[Thought 1]: Both of the two sentences use the word "any". in their most common uses, it can only be used in an appropriate syntactic-semantic-environment.
[Action 1]: Follow-up Question[In what syntactic-semantic-environment can the word "any" be used?]
[Answer 1]: To a first approximation, it can be only in the scope of negation.
[Thought 2]: If a sentence does not contain a negation structure to match the word "any", it will be ungrammatical.
[Action 2]: Follow-up Question[Which sentence does not contain a negation structure?]
[Answer 2]: SecondSentence. Although it contains a negation structure of "no senators" in the subordinate clause, it does not contain a negation structure in the main clause to match the word "any" in the main clause.
[Thought 3]: The SecondSentence of minimal pairs lacks a negation structure, so it is ungrammatical.
[Action 3]: Finish[]
[Answer 3]: FirstSentence is grammatical and SecondSentnce is ungrammatical.
[More Demonstration Questions] [...]
[Input Question]:

Figure 9: Full version example of the capability-specific instruction.