A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity

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** **Pascale Fung***
Centre for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
yjbang@connect.ust.hk, pascale@ece.ust.hk

Abstract

This paper proposes a framework for quantitatively evaluating interactive LLMs such as ChatGPT using publicly available data sets, using 23 data sets covering 8 different common NLP application tasks. We extensively evaluate the multitask, multilingual, and multi-modal aspects of ChatGPT based on these data sets and a newly designed multimodal dataset. We find that ChatGPT outperforms LLMs with zero-shot learning on most tasks and even outperforms fine-tuned models on some tasks. We find that it is better at understanding non-Latin script languages than generating them. It is able to generate multimodal content from textual prompts via an intermediate code generation step. Moreover, we find that ChatGPT is 63.41% accurate on average in 10 different reasoning categories under logical reasoning, non-textual reasoning, and commonsense reasoning, hence making it an unreliable reasoner. ChatGPT suffers from hallucination problems like other LLMs. Finally, the interactive feature of ChatGPT enables human collaboration with the underlying LLM to improve its performance, i.e., **8% ROUGE-1** on summarization and **2% ChrF++** on machine translation, in a multi-turn "prompt engineering" fashion. We release a code for evaluation set extraction.[1]

1 Introduction

ChatGPT is a successor of the large language model (LLM) InstructGPT (Ouyang et al., 2022) with a dialog interface that is fine-tuned using the Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017) approach. ChatGPT has gathered 100 million monthly active users in such a short period of time (Hu, 2023) and is being used by businesses and consumers alike for a myriad of mostly textual tasks. One reason for its unprecedented popularity is that ChatGPT, through its scale and via RLHF, has shown impressive abilities in many areas of NLP as well as emergent abilities. Another reason is that its dialog interface allows users to interact with the underlying LLM more effectively and efficiently via interactive chats that are akin to multi-turn prompting.

However, despite its powerful abilities, anecdotal reports on ChatGPT consistently showed remaining challenges - for example, it fails in some elementary mathematical (Gilson et al., 2022; Goldberg, 2023; Frieder et al., 2023; Choi et al., 2023; Davis, 2023b) and commonsense reasoning tasks (Guo et al., 2023; Davis, 2023b); it hallucinates with human-like fluency and eloquence on things that are not based on truth (Shen et al., 2023; Thorp, 2023; Smith, 2023); and as a general-purpose language model trained from everything on the web, its language coverage is questionable (Lu et al., 2022; Jiao et al., 2023). Consequently, it is not clear what people can or cannot use ChatGPT for despite its popularity.

Since OpenAI never published any benchmarking results on ChatGPT at the time, seeing this need, in February 2023, we proposed a comprehensive framework for quantitatively evaluating interactive LLMs such as ChatGPT through standard public test sets on major NLP tasks such as question answering, reasoning, summarization, machine translation, sentiment analysis, language identification, task-oriented dialogue, and misinformation detection. We evaluate its multilingual performance as well as vision-language multimodal abilities. With additional experiments, we also quantitatively evaluated its primary limitations in *reasoning* and *hallucination*. In addition, we conducted experiments to test its *multi-turn interactivity* as a means for better prompt engineering. We aimed to provide insights to users of ChatGPT on the strengths mentioned above and limitations, as well as how they can improve outcomes with interactivity. To the best of our knowledge, this is the first published

* Equal Contribution.

[1] https://github.com/HLTCHKUST/chatgpt-evaluation

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
November 1–4, 2023.

benchmark of ChatGPT from a third party. More recently, the GPT-4 technical report (OpenAI, 2023) published a number of human task benchmarks.

The true scope of all emergent capabilities of generative models, including ChatGPT, is still unclear. Thus, any benchmarking exercise cannot be 100% "comprehensive" in the scientific sense. We aim to show not just researchers but also users what ChatGPT can and cannot do by presenting interpretable benchmarking results in a zero-shot setting without access to APIs so that the general audience can replicate our evaluation with the test sets we have provided in a zero-shot setting. This version of ChatGPT is 15 December 2022.

The following are the major insights we have gained from the evaluations:

Multitask, Multimodal, and Multilingual: For 9/13 NLP datasets, ChatGPT outperforms previous LLMs with zero-shot learning. It even outperforms fully fine-tuned task-specific LMs on 4 different tasks. In other cases, ChatGPT is on par or slightly lower than fully fine-tuned for specific NLP tasks; ChatGPT fails to generalize to low-resource and extremely low-resource languages (e.g., Marathi, Sundanese, and Buginese). There is an overall performance degradation in low-resource languages, especially in non-Latin scripts in the case of translation; its weakness lies in generation rather than understanding part of the translation process; ChatGPT enables a code intermediate medium to bridge vision and language, even though the multi-modality ability is still elementary compared to vision-language models.

Reasoning: We tested 10 different reasoning categories with 634 samples in total. Based on our experiments, ChatGPT shows more weakness in inductive reasoning than in deductive or abductive reasoning. ChatGPT also lacks spatial and mathematical reasoning while showing better temporal reasoning. Further, we found that ChatGPT is relatively better at commonsense reasoning than non-textual semantic reasoning. Finally, while ChatGPT shows acceptable performance in causal and analogical reasoning, it is bad at multi-hop reasoning capability, similar to other LLMs' weakness (Ott et al., 2023).

Hallucination: Similar to other LLMs (Radford et al., 2019; Muennighoff et al., 2022; Workshop et al., 2022), ChatGPT suffers from the hallucination problem. It generates more extrinsic hallucinations – factual statements that cannot be verified from the source.

Interactivity: One of the primary differentiating factors of ChatGPT from its predecessors is its *multi-turn dialog interactivity*. This enables ChatGPT to perform multiple tasks within a dialog session. There is also significant performance improvement (8% ROUGE-1 on summarization and 2% ChrF++ on low-resource machine translation) via multi-turn interactivity in various standard NLP tasks. This process is akin to prompt engineering with feedback from the system.

2 Multitask, Multilingual, and Multimodal Evaluations of ChatGPT

2.1 Multitask Ability of ChatGPT

ChatGPT has become very well-known in such a short period of time to general public users, not just those who are in AI, machine learning, and NLP communities who might be more familiar with LLMs. One of the main reasons is that, in addition to media reports, innumerable use cases of ChatGPT are shared by both non-academic and academic users online (Marr, 2022; Gordon, 2023; Shankland, 2023). There have been debates and panels on whether ChatGPT is approaching Artificial General Intelligence, as it seems to be able to carry out a multitude of tasks without specific fine-tuning (Desk, 2023; Johnson, 2023; Kingson, 2023). On the other hand, there has also been as much sharing of its failures in simple tasks (Gilson et al., 2022; Choi et al., 2023; Shen et al., 2023).

Instead of relying on anecdotal examples, we first evaluate ChatGPT's performance in various standard NLP tasks in a zero-shot manner to obtain a basic/better understanding of its multi-task ability. We compile results from the existing literature on ChatGPT and compare them with the state-of-the-art fully-fine-tuned and zero-shot models across multiple tasks. We evaluate ChatGPT performances on 21 datasets covering 8 tasks, i.e., summarization, machine translation, sentiment analysis, question answering, task-oriented dialogue, open-domain knowledge-grounded dialogue, and misinformation detection tasks. We sample testing cases from existing standard test sets for each task with a sample size ranging from 30 to 200 samples.

Multitask Generalization of ChatGPT The result of the multitask evaluation is shown in Table 1. ChatGPT is shown to achieve remarkable zero-shot performances on multiple tasks, surpassing pre-

Tasks	Dataset	Metric	Reference	Fine-Tuned SOTA	Zero-Shot SOTA	ChatGPT
Summarization	CNN/DM	ROUGE-1	Lewis et al. (2020a)	44.47	35.27*	35.29
	SAMSum	ROUGE-1	Lewis et al. (2020a)	47.28	-	35.29
MT (XXX→Eng)	FLoRes-200 (HRL)	ChrF++	Team et al. (2022)	63.5	-	58.64
	FLoRes-200 (LRL)	ChrF++	Team et al. (2022)	54.9	-	27.75
MT (Eng→XXX)	FLoRes-200 (HRL)	ChrF++	Team et al. (2022)	54.4	-	51.12
	FLoRes-200 (LRL)	ChrF++	Team et al. (2022)	41.9	-	21.57
Sentiment Analysis	NusaX - Eng	Macro F1	Winata et al. (2022)	92.6	61.5	83.24
	NusaX - Ind	Macro F1	Winata et al. (2022)	91.6	59.3	82.13
	NusaX - Jav	Macro F1	Winata et al. (2022)	84.2	55.7	79.64
	NusaX - Bug	Macro F1	Winata et al. (2022)	70.0	55.9	55.84
Question Answering	bAbI task (15 | 16)	Accuracy	Weston et al. (2016a)	100 | 100	-	93.3 | 66.7
	EntailmentBank	Accuracy	Clark et al. (2018)	86.5	78.58	93.3
	CLUTRR	Accuracy	Minervini et al. (2020)	95.0	28.6	43.3
	StepGame (k=9 | k=1)	Accuracy	Mirzaee and Kordjamshidi (2022)	48.4 | 98.7	-	23.3 | 63.3
	Pep-3k	AUC	Porada et al. (2021)	67.0	-	93.3
Misinformation Detection	COVID-Social	Accuracy	Lee et al. (2021)	77.7	50.0	73.3
	COVID-Scientific	Accuracy	Lee et al. (2021)	74.7	71.1	92.0
Task-Oriented Dialogue	MultiWOZ2.2	JGA	Zhao et al. (2022)	60.6	46.7	24.4
	MultiWOZ2.2	BLEU	Nekvinda and Dušek (2021)	19.1	-	5.65
	MultiWOZ2.2	Inform Rate	Yang et al. (2021)	95.7	-	71.1
Open-Domain KGD	OpenDialKG	BLEU | ROUGE-L	Ji et al. (2022b)	20.8 | 40.0	3.1 | 29.5	4.1 | 18.6
	OpenDialKG	FeQA	Ji et al. (2022b)	48.0	23.0	15.0

Table 1: Performance of ChatGPT compared to state-of-the-art fully-fine-tuned models (Fine-Tuned SOTA) and LLM in zero-shot settings (Zero-Shot SOTA). The referenced performances are evaluation results on full test sets, while the ChatGPT performances are computed on subsets of the corresponding dataset **using 30 to 200 data samples** for each task. For Machine Translation (MT) tasks, we follow the definitions of high-resource language (HRL) and low-resource language (LRL) from NLLB (Team et al., 2022) and take subsets of languages to represent each group. JGA denotes joint goal accuracy. Average of performances for CNN and DM from Goyal et al. (2022). LMs in zero-shot settings are as follows. Summarization: InstructGPT, MT: NLLB-200, Sentiment Analysis: XLM-R LARGE, QA: ST-MoE-32B, ZeroQA, GPT-3, Misinformation Detection: GPT-2, Task-Oriented Dialogue: D3ST, Open-Domain KGD: GPT-Jurassic-6B.

vious state-of-the-art zero-shot models on 9 out of 13 evaluation datasets with reported zero-shot LLMs' performances. In most tasks, especially task-oriented and knowledge-grounded dialogue tasks, task-specific fully-fine-tuned models outperform ChatGPT. Compared to the latter, ChatGPT yields lower performance in most tasks while still surpassing the performance on 4 datasets.

Furthermore, from the evaluation results, we also observe several limitations of ChatGPT: 1) limited language understanding and generation capabilities on low-resource languages, 2) lacking reasoning ability as shown from the results in QA, and 3) performing task-oriented and knowledge-grounded dialogue tasks. More detailed experimental setup and analysis for each task are shared in Appendix §C. We also provide the complete list of all the datasets used in our evaluation in Appendix I.

ChatGPT on Dialogue Tasks Given that ChatGPT has the ability to generate conversation-like responses, we test it on conventional dialogue tasks: 1) knowledge-grounded open-domain dialogue and 2) Task-oriented dialogue. Task setups are explained in Appendix C.6.

Knowledge-Grounded Open-Domain Dialogue To quantitatively measure ChatGPT's performance on knowledge-grounded dialogue, we utilize 50 samples from the test set of OpenDialKG (Moon et al., 2019), which contains open-ended dialogues grounded on a knowledge path. According to human judgment, the responses from ChatGPT are of high quality with fluent response generation and incorporating the provided knowledge in the response. However, the automatic evaluation results are relatively low compared with fine-tuned GPT2. We postulate this is because ChatGPT responses are longer than the golden answers and include content from its parametrized knowledge injected during pre-training.

Task-Oriented Dialogue We investigate and discuss how ChatGPT's emergent abilities and interactivity could potentially be leveraged for ToD as well in two setups. Firstly, A) modular approach: testing dialogue state tracking (DST) and response generation using oracle actions. DST is mediocre

Language	Category	SA Acc.	LID Acc.
English	HRL	84%	100%
Indonesian	MRL	80%	100%
Javanese	LRL	78%	0%
Buginese	X-LRL	56%	12%

Table 2: Accuracy of ChatGPT on Sentiment Analysis (SA) and Language Identification (LID) tasks.

while ChatGPT successfully leverages all information provided while answering the questions with a 71.1% inform rate and 5.65 BLEU score. Next, B) Unified approach: a direct approach to simulate the ToD interaction while leveraging information in a structured database. We observed the limitations of ChatGPT: 1) ChatGPT cannot keep the belief state across multiple turns within the interaction, 2) ChatGPT's response tends to be wrong if the query introduces a basic level of reasoning 3) ChatGPT tends to generate hallucinated information beyond the given knowledge, which is not desirable for ToD. We provide details and examples of the modular and unified approaches in Appendix C.6.2.

2.2 Evaluating Multilinguality of ChatGPT

Training data size affects language understanding and generation ability of LMs (Raffel et al., 2022; Cahyawijaya et al., 2021; Rae et al., 2021; Workshop et al., 2022; Chowdhery et al., 2022; Hoffmann et al., 2022). As an LLM, the same premise also applies to ChatGPT, but the question is to what extent. We investigate this question through a series of experiments by analyzing 1) the language understanding capability through sentiment analysis (SA) and language identification (LID) tasks, and 2) the language generation capability through machine translation using English as the pivot language. Based on the size proportion in CommonCrawl (i.e., the primary source of language pre-training data used in various LLMs) , we group languages into 4 language resource categories, i.e., **high-resource language (HRL)** (<$\geq$1%), **medium-resource language (MRL)** ($\geq$0.01%), **low-resource language (LRL)** ($\geq$0.0001%), and **extremely low-resource language (X-LRL)** (<0.0001%). The statistics of the languages are shown in Table 9 and other details are described in Appendix D.

2.2.1 Language Understanding

We investigate the language understanding ability of ChatGPT on 4 languages from different language categories in NusaX (Winata et al., 2022), i.e. English, Indonesian, Javanese, and Buginese, through sentiment analysis and language identification tasks. ChatGPT fails to generalize to extremely low-resource languages. As shown in Table 2, there is a clear correlation between ChatGPT performance with the language resource category. This result aligns with the findings from prior works (Chowdhery et al., 2022; Workshop et al., 2022; Muennighoff et al., 2022), where LLMs, including ChatGPT, yield a lower performance for lower resource languages. Interestingly, the performance gap between English, Indonesian, and Javanese is considered marginal compared to the performance gap with Buginese. This suggests that ChatGPT has a limitation in generalizing toward extremely low-resource languages. Furthermore, we also find that ChatGPT can understand low-resource languages, such as Javanese, without having the knowledge to identify the language itself. Moreover, ChatGPT displays better human-preferred responses when it has no knowledge about the language. For instance, as illustrated in 8, ChatGPT lets the user know that its prediction is uncertain when it does not completely understand the language and also provides broader information regarding the language.

Language	Category	XXX→Eng	Eng→XXX
Chinese	HRL	24/30	14/30
French	HRL	29/30	25/30
Indonesian	MRL	28/30	19/30
Korean	MRL	22/30	12/30
Javanese	LRL	7/30	6/30
Sundanese	LRL	9/30	0/30

Table 3: #Correct translations of ChatGPT. XXX denotes the target language listed in the first column.

2.2.2 Language Generation

We assess the multilingual language generation ability of ChatGPT through machine translation. We experiment with 6 languages: French, Chinese, Indonesian, Korean, Javanese, and Sundanese from the FLORES-200 dataset (Team et al., 2022; Goyal et al., 2021). For each language, we sample 30 English-XXX parallel sentences and perform two directions of translation using English as the pivot language. The correctness of the translation results is manually validated by a native speaker of the corresponding language.

Based on our evaluation results (Table 3), similar to other LLMs (Workshop et al., 2022; Muennighoff et al., 2022), ChatGPT produces better English translation quality from high-resource languages, such as French and Chinese. While for low-resource languages, such as Javanese and Sundanese, ChatGPT tends to generate several mistranslated words/phrases and sometimes even hallucinate some objects. Moreover, we also observe that sometimes ChatGPT translates the English sentence into a different but related language other than the requested target language (see §H.2). This fact suggests that the generalization of LLMs, including ChatGPT, to low-resource languages, remains an open challenge. Moreover, we also find that ChatGPT can handle Latin script languages better than non-Latin script languages, especially in generating sentences using those scripts.

2.3 Evaluating Multimodality of ChatGPT

Since ChatGPT is a purely text-prompted language model, it is unlikely to explore its multimodal capabilities with visual inputs like contemporary vision-language works (Rombach et al., 2022; Ramesh et al., 2021; Yu et al., 2021a; Radford et al., 2021; Dai et al., 2022; Lovenia et al., 2022; Dai et al., 2023a). However, thanks to its code understanding and generation abilities, programming codes can serve as the intermediate medium to bridge vision and language (Rasheed, 2020; Shiryaev, 2022). Given textual prompts, ChatGPT can generate code representations of visual images using the SVG (Scalable Vector Graphics) format or APIs (e.g., HTML Canvas element, Python Turtle graphics). For example, as shown in Figure 1, ChatGPT can generate a well-formed and suitable intermediate representation in code format to synthesize images given the dialogue context and user prompts.

In this way, even though the generated images are symbolic and their quality is not comparable to the ones generated by modern text-to-image models (Ramesh et al., 2021; Rombach et al., 2022), it is worth exploring due to three reasons. Firstly, it helps us investigate the visual understanding and reasoning abilities of ChatGPT, which can be seen as an emergent skill after the very large-scale pre-training on text and code data. Furthermore, representing images with code is a more explainable way to understand the model's behaviors and rationales in text-to-image generation. Third, it is a natural way to evaluate ChatGPT's ability on multi-turn interaction by asking for post-editing and corrections of the generated images.

To systematically evaluate the image generation ability of ChatGPT through code generation, we designed a national flag drawing task. This task

Grade (# of Errors)	Turn 1 (w/o desc)	Turn 1	Turn 2	Turn 3
A (0)	0	4	12	**24**
B (1)	4	22	24	**24**
C (2)	16	18	12	10
D (3)	18	24	**26**	20
E (≥ 4)	**62**	**32**	**26**	22

Table 4: Results of the portion (%) of generated flags evaluated into five grades, A $\sim$ E. The second column shows the results of an ablation study, which removes the step of flag description generation and directly asks ChatGPT to generate the SVG code of the flag image.

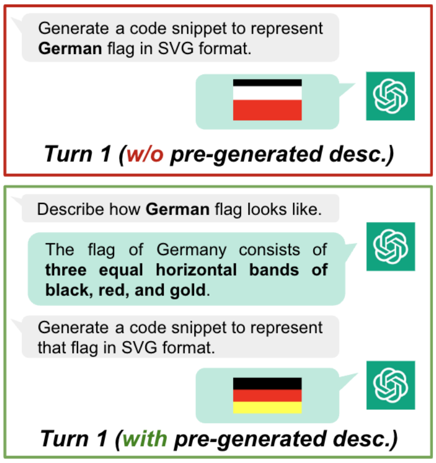


Figure 1: An example of a German flag drawn by ChatGPT using SVG format: **(top)** without and **(bottom)** with a self-retrieved textual description of the flag. A rendered image is shown in place of the generated SVG format for the sake of simplicity.

tests how ChatGPT's textually described knowledge (language) converts into the drawing (vision) through the SVG (code), using multi-turn conversations. The task contains three steps. Firstly, we ask ChatGPT to illustrate the appearance of the flag. Next, based on the description, we ask ChatGPT to generate the SVG code of that flag. Finally, if the generated image contains errors, we iteratively ask ChatGPT to fix them. There are four types of errors: 1) layout, 2) color, 3) missing components, 4) shape/size. We uniformly collect 50 national flags from different continents and conduct the flag-drawing task on ChatGPT. The prompts and full results are shown in Appendix E. The generated flag images are evaluated by the aforementioned four error types as criteria. We further assess the image quality with five grades, A $\sim$ E, which indicate zero to four (or above) errors. An overview of the result evaluation is provided in Table 4.

We share our major two findings from the task: **1) ChatGPT is capable of drawing, yet better**

with a self-generated textual description. As demonstrated in Table 4 and Appendix E, by following the task formulation, ChatGPT can generate plausible national flags using the SVG format. To better understand the behavior of ChatGPT, we perform an ablation study by removing the description generation step. As illustrated by Figure 1, the performance drops dramatically without first prompting the textual flag description, which is generated by ChatGPT itself. Explicitly describing the appearance of the flag and then drawing disentangles the image generation process, which can be considered as a chain-of-thought reasoning. **2) ChatGPT is an elementary illustrator.** Among the four error types, the majority lies in the *shape/size* error, which happens 68% of the time. For the other three error types (*layout*, *color*, *missing components*), they appear 34%, 20%, and 18% of the time, respectively. For instance, ChatGPT cannot generate the exact shape of the maple leaf in the Canadian flag while it gets the layout and color correctly (Figure 3). This is a natural defect of text-only language models as they never see actual visual data and textual data is usually conceptual.

3 Reasoning Evaluations of ChatGPT

Reasoning is one of the most actively discussed and debated abilities of LLMs as scaling the model parameter size also increases the implicit knowledge in LLMs (Wei et al., 2022a; Wang et al., 2022; Huang and Chang, 2022). Mahowald et al. eloquently argues that "language ability does not equal to thinking" or "reasoning" in LLMs, and that LLMs have poor reasoning skills despite possessing human-level language skills.

In the NLP literature, evaluating a model's reasoning often means evaluating its various skills in arithmetic, commonsense, and symbolic reasoning in different NLP tasks that require such skills (Talmor et al., 2020; Zelikman et al., 2022; Wei et al., 2022b). However, the reasoning itself is a much broader concept thus it is hard to conclude whether a model can "reason" or not based on those aforementioned, and current works on reasoning are scattered. This is in line with the anecdotal experience of users with ChatGPT – some of the examples demonstrate surprisingly good "reasoning" abilities compared to previously introduced LLMs but at the same time ChatGPT fails in very simple reasoning problems (the, 2023; Venuto, 2023; Qiao et al., 2022; Cookup.ai, 2022; Labs, 2022).

Thus, we investigate the reasoning ability of ChatGPT in a more fine-grained manner, which includes deductive, inductive, abductive, analogical, causal, multi-hop, mathematical, temporal, and spatial reasoning, via question-answering tasks. We categorize available QA tasks into each category by avoiding overlap (i.e., choosing testsets that require mainly one specific category of reasoning). Composed results and corresponding datasets for each category are shown in Table 5. For evaluation, we manually check the accuracy of the answer as well as verify the rationales and explanations generated by ChatGPT. A detailed explanation of task setup is explained in Appendix F.

Logical Reasoning Inductive, deductive, and abductive reasoning are common forms of logical reasoning, a process of deriving a conclusion or judgment based on given evidence or past experience and observations (Rogers et al., 2022; Wason and Johnson-Laird, 1972; Huang and Chang, 2022). We first investigate basic reasoning skills with bAbI tasks (Weston et al., 2016b), 30 examples each from task 15 (inductive) and task 16 (deductive). One major investigation is that ChatGPT is a lazy reasoner that suffers more from induction. Interestingly, when ChatGPT was asked to answer a question given premises without any prompt engineering, it performed poorly in induction (0 out of 30) while it achieved much better performance in deduction (19 out of 30). However, when ChatGPT is explicitly asked for reasonable inference inductive reasoning increases to 20 out of 30. Yet, it is still not as good as in deduction. When we repeat the analysis on advanced tasks, specifically on CLUTRR (Sinha et al., 2019) for induction and EntailmentBank for deduction (Dalvi et al., 2021), the same conclusion holds based on our experiment.

Non-textual semantic reasoning It is often investigated in public sharing about ChatGPT errors cases that it lacks the reasoning ability that requires non-text semantic understanding such as mathematical, temporal, and spatial reasoning. Not surprisingly, it could only score 23.33% (7/30) for the MATH dataset (Saxton et al., 2019), which tests mathematical reasoning. Overall, ChatGPT correctly answers 86.67% of the time (26/30), suggesting that it has a decent temporal reasoning ability. ChatGPT falls short of the spatial reasoning tasks, with success rates of 43.33% for StepGame and 43.75% for SpartQA. We investigate the errors that

Categories	Testset	Result
Deductive	EntailmentBank bAbI (task 15)	28/30 28/30 (as-is: 19/30)
Inductive	CLUTRR bAbI (task16)	13/30 20/30 (as-is: 0/30)
Abductive	αNLI	26/30
Mathematical	Math	13/30
Temporal	Timedial	26/30
Spatial	SpartQA (hard | basic) StepGame (hard | basic) StepGame (cardinal) StepGame (diagonal) StepGame (clock)	8/32 | 20/32 7/30 | 19/30 17/20 11/20 5/20
Commonsense	CommonsenseQA PIQA Pep-3k (Hard)	27/30 25/30 28/30
Causal	E-Care	24/30
Multi-hop	hotpotQA	8/30
Analogical	Letter string analogy	30/30

Table 5: Composed results for all reasoning tasks.

it often fails to understand clock direction (e.g., "W is at K's 3 o'clock") and diagonal spatial relations.

Commonsense Reasoning It is understanding and reasoning about everyday concepts and knowledge that most people are familiar with, to make judgments and predictions about new situations (Storks et al., 2019). Recent works show that LLMs perform impressively well on commonsense reasoning benchmarks (Qiao et al., 2022; Huang and Chang, 2022; Bhargava and Ng, 2022). Based on our evaluation with CommonsenseQA (Talmor et al., 2018), PiQA (Bisk et al., 2020) and Pep-3k (Wang et al., 2018b), ChatGPT shows surprisingly good commonsense reasoning capability, perhaps due to its large parametric memory.

4 Factuality and Hallucination

LLMs are known to be susceptible to generating nonfactual, untruthful information, which is referred to as hallucination (Lee et al., 2022; Ji et al., 2022a,b; Su et al., 2022; Dai et al., 2023b; Xu et al., 2023). Many anecdotal witnesses show ChatGPT also seems to suffer from the same problem as other LLMs. To evaluate this aspect of ChatGPT, we first explore existing fact-checking and QA test sets and also illustrate the challenge of hallucination in ChatGPT by sharing hallucination examples.

Factuality We evaluate ChatGPT with test sets that consist of scientific and social claims related to COVID-19 (Lee et al., 2021). ChatGPT is able to detect misinformation 92% (46/50) and 73.33% (22/30, excluding verification-refusing cases) accuracy on covid-scientific and covid-social respectively. In comparison to its previously reported performance, ChatGPT's performance on covid-scientific is impressive. Interestingly, for more societal-related claims, ChatGPT often refuses to make verification. However, it cannot avoid the criticism that parameterized knowledge is obtained by better memorization as it still shows worse performance in questions designed to cause imitative falsehoods. We test on 66 test samples from TruthfulQA (Lin et al., 2022), which tests the extent of LLMs to mimic human falsehood and 35.38% of the time ChatGPT fails to answer truthfully.

Hallucination From various tasks, we often find extrinsic hallucinations, including both untruthful and factual ones, across various tasks such as Machine Translation and question answering, which causes degradation in performance. The intrinsic hallucinations are barely found as discussed in tasks about summarization and knowledge-grounded open-domain dialogue. We share examples of these hallucination types detected from different task explorations in Table 19.

5 Evaluating Interactivity in ChatGPT

ChatGPT has a built-in interactive ability thanks to conversational data fine-tuning and RLHF. We further delve into the benefit of exploiting this interactive ability of ChatGPT in three NLP tasks, summarization, machine translation, and multimodal generation. Our experiments demonstrate the potential of employing multi-turn interaction to refine the quality of the generated responses and improve the task performance of ChatGPT.

Interactivity on Summarization Summarization models aim to extract essential information from documents and to generate short, concise, and readable text (Yu et al., 2021b; Su et al., 2021). In real-world applications, people may want to improve the summary based on the previously generated summary. We ran experiments with 50 documents from SAMSum (Gliwa et al., 2019) and conducted a two-turn iterative prompt approach. ChatGPT usually generates an overly long summary. By adding a follow-up prompt after the first summary, "`Please make the summary shorter`", ChatGPT could provide a much shorter summary than the first response. Experimental results show that with the second length control prompt, the refined summaries achieve 7.99 and 1.64 gains on ROUGE-1 and ROUGE-2 respectively.

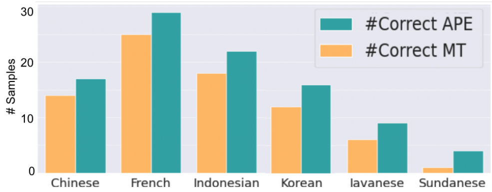


Figure 2: Result of the multi-turn MT-APE experiment. **#Correct MT** denotes the number of correct translations. **#Correct APE** denotes the number of correct translations after post-editing.

Interactivity on Machine Translation One of the capabilities of ChatGPT is to perform text translation from one language to another. With the interactivity of ChatGPT, we explore the possibility of performing a combined machine translation and automatic post-editing tasks to improve the translation quality of ChatGPT. For the experiment, we adapt the dataset used in §2.2.2. As shown in Figure 2, despite the translation and post-editing being done using a single ChatGPT model, the multi-turn approach method helps to improve the correctness of the translation by making partial corrections or even full corrections in some cases. We provide experimental setup details and examples of the post-editing in Appendix J.

Interactivity on Multimodal Generation The multi-turn interaction ability of ChatGPT enables the refinement of text-to-image generation. It is one of the most natural ways for humans to create artwork or product designs by requesting an AI tool iteratively. Through interaction with ChatGPT over multiple turns, a process of creating an interesting painting can be achieved (Figure 7).

To quantitatively study how this ability impacts image generation, we conduct at most three rounds of post-editing for the flag-drawing task. As shown in Figure 4, in the first round of generation, ChatGPT rarely generates errorless SVG images except for some simple flags (e.g., Nigerian and German). We observe that 34% and 36% of samples experience improvement (i.e., fewer errors) from turn 1 to 2 and from turn 2 to 3, respectively. We also tested with the InstructGPT, which has the same backbone model as ChatGPT but lacks conversation ability. InstructGPT cannot achieve salient improvements by directly putting the intermediate results in the input context (Appendix H.3).

6 Conclusion and Discussion

Multitask, Multilingual, Multimodal ChatGPT outperforms SOTA LLMs in a zero-shot manner on various tasks and even surpasses fine-tuned models on some tasks. However, there are still some failure cases (§2.1) and it produces responses with altered nuance and meaning. Therefore, dealing with these special cases is a complex but important task. In terms of multilinguality, ChatGPT achieves strong performance in many high-resource and medium-resource languages. Nevertheless, ChatGPT still lacks the ability to understand and generate sentences in low-resource languages (§2.2), which is also supported by Lai et al. (2023). Additionally, ChatGPT lacks the generation ability of non-Latin script languages (§2.2.2), despite the languages being high-resource. These raise the concern of language diversity and inclusivity in ChatGPT (Joshi et al., 2020; Aji et al., 2022). Regarding multimodality, our flag drawing experiments showed the potential of ChatGPT's multimodal ability. It would be an interesting research direction to further explore ChatGPT's multimodal ability to answer "can textual models like ChatGPT switch to a multimodal backbone?"

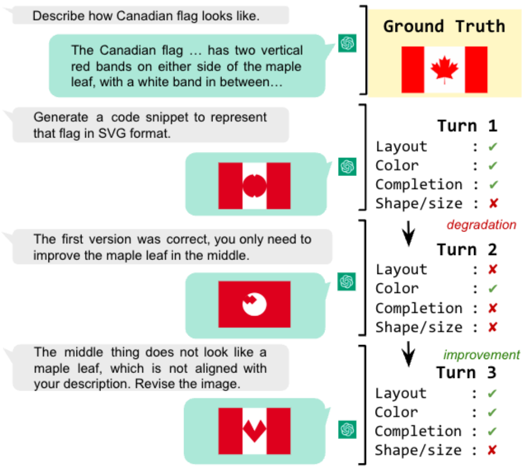


Figure 3: Changes in ChatGPT's drawing of the Canadian flag over three turns. Layout, color, completion, and shape/size are marked as ✓ if they align with those of the ground truth, and ✗ otherwise.

Reasoning The impressive performance of ChatGPT has sparked interest in expanding its usage beyond traditional NLP tasks into more complex domains requiring sophisticated reasoning such as problem-solving, decision-making, and planning. Our evaluation of its reasoning abilities shows that they are not reliable. Specifically, our findings indicate that ChatGPT exhibits a tendency to be a lazy reasoner and that its capabilities are inconsistent across various reasoning abilities; To support the further expansion of its use cases, it is necessary to prioritize the development of systems with robust complex reasoning capabilities, which should also be facilitated by the creation of more comprehensive benchmarks for assessing these abilities, such as works by Laskar et al. (2023b); Qin et al. (2023); Davis (2023a), particularly when multiple abilities are required to complete the tasks.

Factuality&Hallucinations ChatGPT, like other LLMs, still makes things up (Ji et al., 2022a). To ensure factuality, it is possible to build LLMs with an interface to an external knowledge source, like Blenderbot 3.0 (Shuster et al., 2022) and LaMDa (Thoppilan et al., 2022). Meanwhile, there are many forms of hallucinations that are not necessarily counterfactual but still undesirable. The RLHF process of ChatGPT can ensure human feedback to mitigate undesirable responses. However, researchers need to work on coming up with more automatic and scalable methods to detect and mitigate hallucinations and other undesirable artifacts.

Interactivity Compared with the previous LLMs, the interactive ability of ChatGPT has made a leap according to both qualitative and quantitative measures. Through interactivity, ChatGPT can recite from its own description, which is a very important ability. A similar exploration of this ability in LLMs has also been explored in other research works (Sun et al., 2022; Wang et al., 2023). However, sometimes ChatGPT retains the wrong answer even after receiving multiple rounds of prompts from the user. Improving the ability of ChatGPT to handle multiple rounds of user feedback is also an important challenge.

Limitation

The experiments are done with the UI of ChatGPT provided by OpenAI (15 December 2019 version), before the ChatGPT API was released, thus, the number of samples for evaluation is limited (30-200). However, tasks of evaluation should not be affected much because most of the recent updates/releases of ChatGPT are related to safety concerns. Moreover, It is possible to augment our benchmarks with other technical benchmarks for research purposes, especially now that the ChatGPT APIs are available. There has been recent automatic or human-in-the-loop evaluations such as (Laskar et al., 2023a) Nevertheless, many of the benchmarks are not necessarily interpretable to laypeople for general purposes, such as named entity recognition and etc. Our paper provides an easier-to-follow guideline.

Due to the page limit, many parts of the experimental setup details are added to the Appendix while the overall structure of evaluation and major insights stay in the main content. This may cause the reader inconvenience to follow the experiments. However, we publicly release the codebase that can help the community replicate the exact same evaluation either on ChatGPT or other LLMs easily.

Ethics Statement

Responsible Generative AI Previous works have discussed the ethical implications or concerns associated with ChatGPT (and other LLMs) (Jabotinsky and Sarel, 2022; Susnjak, 2022; Blanco-Gonzalez et al., 2022; Aydın and Karaarslan, 2022; Jeblick et al., 2022). Agreeing with the previous literature, the responsible design and usage of LLMs including ChatGPT is an important and pressing challenge today. There are common issues with these models, such as fairness, toxicity, demographic bias, and safety, which need to be addressed. In the case of ChatGPT, OpenAI constructs safety layers and uses RLHF and potentially other means to filter out undesirable system responses. However, this is still not perfect and requires future research to further improve the robustness of the safety layer. This process is resource-intensive and opaque to the public. We hope to see a more open discussion and sharing of responsible design of LLMs from various organizations including OpenAI in the future.

Use of Scientific Artifacts/Data This paper conducts an evaluation of ChatGPT for academic purposes only. We comply with the terms and conditions of ChatGPT stated in `https://openai.com/policies/terms-of-use`. Moreover, we comply with all the licenses of all the data (i.e., test sets/benchmarks) that are used in this evaluation.

Acknowledgments

This work has been partially funded by MRP/055/18 of the Innovation Technology Commission, Hong Kong SAR Government; the Hong Kong Fellowship Scheme by the Hong Kong Research Grants Council, and PF20-43679 Hong Kong PhD Fellowship Scheme, Hong Kong Research Grants Council.

References

2023. Chatgpt vs satya nadella over biryani: The chatbot is learning from its mistakes.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Ömer Aydın and Enis Karaarslan. 2022. Openai chatgpt generated literature review: Digital twin in healthcare. *Available at SSRN 4308687*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Paul Bartha. 2013. Analogy and analogical reasoning.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: a survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12317–12325.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2022. Findings of the wmt 2022 shared task on automatic post-editing. In *Proceedings of the Seventh Conference on Machine Translation*, pages 109–117, Abu Dhabi.

David G.W. Birch. 2022. Chatgpt is a window into the real future of financial services.

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.

Alexandre Blanco-Gonzalez, Alfonso Cabezon, Alejandro Seco-Gonzalez, Daniel Conde-Torres, Paula Antelo-Riveiro, Angel Pineiro, and Rebeca Garcia-Fandino. 2022. The role of ai in drug discovery: Challenges, opportunities, and strategies. *arXiv preprint arXiv:2212.08104*.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Fajri Koto, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Ivan Halim Parmonangan, Ika Alfina, Muhammad Satrio Wicaksono, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Ignatius, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Noor Fatyanosa, Ziwei Ji, Pascale Fung, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2022. Nusacrowd: Open source initiative for indonesian nlp resources.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. Chatgpt goes to law school. *Available at SSRN*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Cookup.ai. 2022. Chatgpt - where it lacks.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Enabling multimodal generation on CLIP via vision-language knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2383–2395, Dublin, Ireland. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023a. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023b. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2136–2148, Dubrovnik, Croatia. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370.

Ernest Davis. 2023a. Benchmarks for automated commonsense reasoning: A survey. *arXiv preprint arXiv:2302.04752*.

Ernest Davis. 2023b. Mathematics, word problems, common sense, and artificial intelligence. *arXiv preprint arXiv:2301.09723*.

Web Desk. 2023. Colombian judge uses chatgpt in ruling, triggers debate.

Igor Douven. 2017. Abduction.

Michael Dowling and Brian Lucey. 2023. Chatgpt for (finance) research: The bananarama conjecture. *Finance Research Letters*, page 103662.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of chatgpt.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Aidan Gilson, Conrad Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. 2022. How well does chatgpt do when taking the medical licensing exams? the implications of large language models for medical education and knowledge assessment. *medRxiv*, pages 2022–12.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

Yoav Goldberg. 2023. Some remarks on large language models.

Cindy Gordon. 2023. Chatgpt is the fastest growing app in the history of web applications.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchan. 2023. Chatgpt is not all you need. a state of the art review of large generative ai models. *arXiv preprint arXiv:2301.04655*.

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*.

James Hawthorne. 2021. Inductive Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2021 edition. Metaphysics Research Lab, Stanford University.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Krystal Hu. 2023. Chatgpt sets record for fastest-growing user base - analyst note.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Arfinda Ilmania, Abdurrahman, Samuel Cahyawijaya, and Ayu Purwarianti. 2018. Aspect detection and sentiment classification using deep neural network for indonesian aspect-based sentiment analysis. In *2018 International Conference on Asian Language Processing (IALP)*, pages 62–67.

Hadar Yoana Jabotinsky and Roee Sarel. 2022. Co-authoring with an ai? ethical dilemmas and artificial intelligence. *Ethical Dilemmas and Artificial Intelligence (December 15, 2022)*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022a. Survey of hallucination in natural language generation. *ACM Comput. Surv.* Just Accepted.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2022b. Rho (ρ): Reducing hallucination in open-domain dialogues with knowledge grounding. *arXiv preprint arXiv:2212.01588*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study.

Arianna Johnson. 2023. Is chatgpt partisan? poems about trump and biden raise questions about the ai bot's bias-here's what experts think.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jennifer A. Kingson. 2023. Friend or foe? teachers debate chatgpt.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, page 101861.

Escape Velocity Labs. 2022. Chatgpt imitates logical reasoning surprisingly well.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023a. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023b. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Anton E Lawson. 2005. What is the role of induction and deduction in reasoning and scientific inquiry? *Journal of Research in Science Teaching*, 42(6):716–740.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020b. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Zhaojiang Lin, Bing Liu, Andrea Madotto, Seungwhan Moon, Zhenpeng Zhou, Paul A Crook, Zhiguang Wang, Zhou Yu, Eunjoon Cho, Rajen Subba, et al. 2021. Zero-shot dialogue state tracking via cross-task transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7890–7900.

Holy Lovenia, Bryan Wilie, Romain Barraud, Samuel Cahyawijaya, Willy Chung, and Pascale Fung. 2022. Every picture tells a story: Image-grounded controllable stylistic story generation. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 40–52, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Hongyuan Lu, Haoyang Huang, Shuming Ma, Dongdong Zhang, Wai Lam, and Furu Wei. 2022. Trip: Triangular document-level pre-training for multilingual language models. *arXiv preprint arXiv:2212.07752*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

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: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of chatgpt and gpt-4. *arXiv preprint arXiv:2308.12488*.

Bernard Marr. 2022. What does chatgpt really mean for businesses?

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2022. A survey on multi-hop question answering and generation. *arXiv preprint arXiv:2204.09140*.

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, and Tim Rocktäschel. 2020. Learning reasoning strategies in end-to-end differentiable proving. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Roshanak Mirzaee and Parisa Kordjamshidi. 2022. Transfer learning with synthetic corpora for spatial role labeling and reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6148–6165, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A textual question answering benchmark for spatial reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4582–4598, Online. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence

RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

Oded Nov, Nina Singh, and Devin M Mann. 2023. Putting chatgpt's medical advice to the (turing) test. *medRxiv*, pages 2023–01.

OpenAI. 2023. Gpt-4 technical report.

Simon Ott, Konstantin Hebenstreit, Valentin Liévin, Christoffer Egeberg Hother, Milad Moradi, Maximilian Mayrhauser, Robert Praas, Ole Winther, and Matthias Samwald. 2023. Thoughtsource: A central hub for large language model reasoning data. *arXiv preprint arXiv:2301.11596*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ian Porada, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. Modeling event plausibility with consistent conceptual abstraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1732–1743, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis and insights from training gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Fabin Rasheed. 2020. Gpt3 sees.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2022. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.* Just Accepted.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *International Conference on Learning Representations*.

Stephen Shankland. 2023. Why the chatgpt ai chatbot is blowing everyone's mind.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023. Chatgpt and other large language models are double-edged swords.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022a. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11321–11329.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022b. StepGame: A new benchmark for robust multi-hop spatial reasoning in texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11321–11329.

Denis Shiryaev. 2022. Drawing mona lisa with chatgpt.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. 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 4506–4515.

Noah Smith. 2023. Why does chatgpt constantly lie?

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. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*, pages 1–60.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 744–756.

Dan Su, Tiezheng Yu, and Pascale Fung. 2021. Improve query focused abstractive summarization by incorporating answer relevance. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3124–3131.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*.

Teo Susnjak. 2022. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Richmond Thomason. 2018. Logic and artificial intelligence.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

H Holden Thorp. 2023. Chatgpt is fun, but not an author.

Giuseppe Venuto. 2023. Giuven95/chatgpt-failures: Chatgpt failure archive.

Douglas Walton. 2014. *Abductive reasoning*. University of Alabama Press.

Ada Wan. 2022. Fairness in representation for multilingual NLP: Insights from controlled experiments on conditional language modeling. In *International Conference on Learning Representations*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a. 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.

Su Wang, Greg Durrett, and Katrin Erk. 2018b. Modeling semantic plausibility by injecting world knowledge. *arXiv preprint arXiv:1804.00619*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Peter Cathcart Wason and Philip Nicholas Johnson-Laird. 1972. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2022a. Emergent analogical reasoning in large language models.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2022b. Emergent analogical reasoning in large language models. *arXiv preprint arXiv:2212.09196*.

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 H Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016a. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016b. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2022. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages.

Cameron R. Wolfe. 2023. Specialized llms: Chatgpt, lamda, galactica, codex, sparrow, and more.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107.

Yan Xu, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, and Ying Nian Wu. 2023. Diverse and faithful knowledge-grounded dialogue generation via sequential posterior inference. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38518–38534. PMLR.

Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2021. Ubar: Towards fully end-to-end task-oriented dialog system with gpt-2. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14230–14238.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021a. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021b. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *ACL 2020*, page 109.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jeffrey Zhao, Raghav Gupta, Yuanbin Cao, Dian Yu, Mingqiu Wang, Harrison Lee, Abhinav Rastogi, Izhak Shafran, and Yonghui Wu. 2022. Description-driven task-oriented dialog modeling. *ArXiv*, abs/2201.08904.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023a. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023b. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity.

Appendix

The appendix consists the following content:

- A: Background and Related Work
- B: General Experimental Details
- C: Details for Multitask Evaluation
- D: Details for Multilinguality Evaluation
- E: Multimodality – Flag Drawing Task
- F: Details for Reasoning Evaluation
- G: Details for Hallucination Evaluations
- H: Details for Interactivity Evaluation
- I: List of Evaluation Datasets
- J: Examples from Machine Translation and Post-Editing

A Background and Related Work

A.1 ChatGPT

Compared to existing LLMs, ChatGPT has unique characteristics. First, it has the ability to interact with users in a conversation-like manner, while retaining its accumulated knowledge and generalization ability gained from pre-training. This is achieved by pre-training ChatGPT on a large-scale conversational-style dataset, that is constructed by transforming a large-scale instruction-tuning corpus used for building InstructGPT into a conversational format, then fine-tuning the model based on a reward model to further improve the generation quality and align the generation with human

Second, ChatGPT is trained with a better human-aligned objective function via Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017). Conventional natural language generation models, including dialogue models, are trained with maximum likelihood estimation (MLE) and might not be aligned with human preferences. For instance, for dialogue systems, humanness, engagement, and groundedness are some examples of essential criteria for success. Such discrepancy between training objectives and evaluation metrics becomes a bottleneck to performance improvement. By using RLHF, ChatGPT aligns more closely with human preferences in generating text than by using MLE.

Discussion on its Capability

As ChatGPT has become available to public users through an easily accessible UI, there have been many discussions from a wide range of communities, not just from AI or NLP, but also from other disciplines. A line of discussion is the specific emergent ability and strength of ChatGPT in more technical perspectives. Guo et al. (2023) conducts linguistic analyses of ChatGPT's writing against human experts and found that ChatGPT responses are strictly focused on the given question, more formal, objective, and less emotional. Nov et al. (2023) also studies ChatGPT's generated medical advice if it passes the Turing test. Frieder et al. (2023) show that "significantly below those of an average mathematics graduate student." There are many investigations of ChatGPT's understanding and potential applications in different fields such as law (Choi et al., 2023), medical domain (Blanco-Gonzalez et al., 2022; Jeblick et al., 2022) and finance (Birch, 2022; Dowling and Lucey, 2023). Jeblick et al. (2022) conduct a case study of the application of ChatGPT on simplified radiology reports. Another important line of discussion is the ethical concerns over the use of ChatGPT. The most active discussion is over the use of academic writing and exam integrity (Jabotinsky and Sarel, 2022; Susnjak, 2022). OpenAI also discusses the misuse of LM for disinformation and remedies. [2] Zhuo et al. study AI ethics of ChatGPT in criteria of bias, reliability, robustness, and toxicity.

A.2 LLM benchmark and evaluation

With the advancement of LLMs' generalization ability, there have been efforts to understand their capabilities, limitations, and risks. Recently, several benchmarks with a collection of a large number of NLP datasets, such as BIG-Bench (Srivastava et al., 2022) and AI LM Harness (Gao et al., 2021), have been introduced. Moreover, HELM (Liang et al., 2022) is proposed to conduct a holistic evaluation of LLMs that considers scenarios and metrics with a top-down approach. In this work, we instead focus on specific limitations and unique findings of ChatGPT that had not been discussed with previous LLMs.

There are also other works that discuss LLMs' emergent abilities through thorough surveys or case studies. Mahowald et al. (2023) thoroughly studies LLMs capabilities by distinguishing *formal* and *functional* linguistic competence with reference to cognitive science, psychology, and NLP to clarify the discourse surrounding LLMs' potential. Other works focus on more specific abilities such as mathematical skills (Davis, 2023b), reasoning (Webb

[2] https://openai.com/blog/forecasting-misuse/

et al., 2022a; Qiao et al., 2022). Also, there have been overviews of existing LLMs (Gozalo-Brizuela and Garrido-Merchan, 2023; Wolfe, 2023)

A.3 ChatGPT Evaluation

To the best of our knowledge, this benchmarking exercise is the first of its kind. Since the introduction of ChatGPT with its advancement, there has been a huge amount of assessments of ChatGPT to understand its limits. Mao et al. (2023) provides a survey of recent assessments of ChatGPT in broad categories of 1) Language and Reasoning Ability, 2) Scientific Knowledge, and 3) Ethical Considerations. Laskar et al. provide extensive automatic or human-in-the-loop evaluations on 140 tasks. Qin et al. mainly evaluated the reasoning abilities of ChatGPT while Zhuo et al.; Ray focus on other important aspects such as ethics, robustness, reliability, limitations, and future scope of ChatGPT. Kocoń et al. examined whether the high quality of the LLM can indicate a tool's usefulness to society by evaluating ChatGPT's capabilities on 25 diverse analytical NLP tasks, most of them subjective even to humans. After the introduction of ChatGPT, GPT-4 has been introduced by OpenAI. However, OpenAI is not disclosing any internal benchmarking of ChatGPT. Even in their GPT-4 technical report (OpenAI, 2023), they have shown the performance of GPT4 in terms of human-level exams. So, it is important that there are 3rd party evaluations of generative models.

B General Experimental Details

The experiments were done with the UI (15 December 2019 version) of ChatGPT provided by OpenAI, before the ChatGPT API was released. The number of samples for evaluation is 30-200. We've prioritized sample diversity, hand-picking tasks that encapsulate the abroad spectrum of scenarios a language model is likely to encounter, thus creating a representative snapshot of potential real-world applications. All experiments are single-run.

C Multitask Evaluation of ChatGPT

C.1 Summarization

We test on 100 samples from two common summarization datasets: half from SAMSum (Gliwa et al., 2019), a dialogue summarization dataset, and another half from CNN/DM (Hermann et al., 2015; Nallapati et al., 2016), news summarization datasets. The large version of Bart (Lewis et al., 2020b) model fine-tuned on both datasets is conducted for comparison. Moreover, OpenAI's text-davinci-002 is used as the previous SOTA zero-shot model. We calculate ROUGE-1 scores for evaluating the generated summary. According to the evaluation, ChatGPT achieves a similar zero-shot performance with text-davinci-002, which is expected since they evolved from the same GPT3 pre-trained checkpoint. However, the fine-tuned Bart still outperforms zero-shot ChatGPT by a large margin.

C.2 Machine Translation

We evaluate the machine translation ability of ChatGPT on both high-resource and low-resource languages using the ChrF++ metric (Popović, 2015). Specifically, we incorporate 8 high-resource languages, i.e., French (fra), Spanish (spa), Chinese (zho), Arabic (ara), Japanese (jpn), Indonesian (ind), Korean (kor), and Vietnamese (vie), and 4 low-resource languages, i.e., Javanese (jav), Sundanese (sun), Marathi (mar), and Buginese (bug) for our evaluation. [3] For each language pair, we sample 30 Eng↔XXX parallel sentences from the FLORES-200 dataset (Team et al., 2022; Goyal et al., 2021). The result of our experiment suggests that ChatGPT can well perform XXX→Eng translation, but it still lacks the ability to perform Eng→XXX translation.

C.3 Sentiment Analysis

Sentiment analysis has been widely explored for both high-resource and low-resource languages (Wang et al., 2018a; Wilie et al., 2020; Ilmania et al., 2018).

We explore the sentiment analysis ability of ChatGPT through 4 languages with diverse amounts of resources in NusaX (Winata et al., 2022): English (eng), Indonesian (ind), Javanese (jav), and Buginese (bug). For each language, we sample 50 sentences from the corresponding dataset for our experiment and measure the macro F1 score as the evaluation metric. We compare the results with two baselines, i.e., supervised state-of-the-art performance from Winata et al. (2022) and zero-shot multilingual LLM from Cahyawijaya et al. (2022). ChatGPT outperforms the previous state-of-the-art zero-shot model by a large margin except for the Buginese, where it performs on par.

[3] For a fairer comparison in our multitask experiment, we strictly follow the definition of high-resource and low-resource languages from NLLB (Team et al., 2022).

This shows that ChatGPT still has a limited understanding of extremely low-resource languages.

C.4 Question Answering

Since Question Answering (QA) is a broad topic, we classify QA datasets into different categories based on the knowledge/reasoning type required to do the task, e.g commonsense reasoning, spatial reasoning, temporal reasoning, etc., to have a clearer analysis on ChatGPT's abilities. For each category, we select several datasets, and for each dataset, we sample 30 instances and test ChatGPT on the subset. Based on our experiment results, ChatGPT outperforms the existing zero-shot and some of the fine-tuned state-of-the-art performance on question answering. Furthermore, ChatGPT achieves near-perfect scores on three tasks, i.e., bAbI task 15, EntailmentBank, and Pep-3k.

C.5 Misinformation Detection

We test ChatGPT's ability to detect misinformation with the test sets that consist of scientific and social claims related to COVID-19 (Lee et al., 2021) with 100 samples. We take half from scientific (covid-scientific) and another half from social (covid-social) sets. We evaluate the accuracy of the veracity by manually checking the generated text. ChatGPT could detect misinformation 92% (46/50) and 73.33% (22/30, excluding verification-refusing cases) accuracy on covid-scientific and covid-social respectively.

C.6 ChatGPT on Dialogue Tasks

C.6.1 Knowledge-Grounded Open-Domain Dialogue

Open-domain dialogue systems interact with humans with generated responses automatically and aim to provide users with an engaging experience. To boost informativeness, these systems leverage external knowledge, including structured knowledge such as knowledge graphs (Zhao et al., 2020; Ji et al., 2022b) and unstructured knowledge such as free text (Xu et al., 2022, 2023).

Prompt used for experiment: "`Can we try dialogue generation? I will give you turns, and you can generate the next turn, but only one.\n \n You can also consider the knowledge of XXX for your reference in the dialogue.`"

C.6.2 Task-Oriented Dialogue Experimental Setups

Setup A: Modular Approach We investigate ChatGPT's ability for both dialogue state tracking and response generation in 50 dialogue turn samples taken from MultiWOZ2.2 (Zang et al., 2020). In detail, we ask the model to provide the belief state as domain-intent: [slot1, value1], ... in the prompt following previous zero-shot (Lin et al., 2021) and few-shot (Madotto et al., 2021) approaches, and provide an exhaustive list of domain-intent-slot-value for the given dialogue. For the response generation, we provide only the oracle dialogue actions (e.g. 'Hotel-Inform':['area', 'centre']), and ask ChatGPT to generate a TOD response given the dialogue history. We assess DST with joint goal accuracy (JGA), the ratio of dialogue turns where the predicted dialogue state is exactly the ground truth, and response generation with BLEU and inform rate(%)

Setup B: Unified Approach We explore ChatGPT's ability to simulate a TOD interaction in an end-to-end manner by providing nothing more than a structured database and giving the instruction: "`Use the following knowledge base to complete the task of recommending a restaurant as a task-oriented dialogue system`".

Result Analysis: We could investigate whether ChatGPT is able to complete basic retrieval queries and respond to users' requests such as "Give me some restaurants that serve Italian food" or "I would prefer cheap options please". However, there are several limitations that we could investigate as follow.

- **Long-term Multi-turn Dependency:** ChatGPT cannot keep the belief state across multiple turns within the interaction. For instance, asking for Italian food will overwrite the previous turn's belief state by asking for restaurants with a rating of 3 or higher. However, if the user explicitly asks to recall the earlier preferences, ChatGPT is able to correct the retrieved information and incorporate the previous belief state. This is interesting as it shows that the information previously given in multi-turn is still usable, but needs to be called explicitly.

- **Basic Reasoning Failure:** ChatGPT's response tends to be wrong if the query introduces a basic level of reasoning such as when it is asked for “recommendation for restaurants with European food” (ChatGPT has to filter the types of cuisine which are based on countries) or “recommendation for restaurants with a rating of 3 or higher” (ChatGPT needs to understand rating 3, 4 and 5). Even with a basic knowledge base, ChatGPT fails to answer correctly 66% of the time.

- **Extrinsic Hallucination:** ChatGPT tends to generate hallucinated information beyond the given knowledge. This is especially harmful in TOD as ChatGPT will sometimes hallucinate some prices for hotel booking, or availability for restaurants.

We provide the example for the modular and unified approaches for Task-Oriented Dialogue in Table 6 and Table 7, respectively.

Task	Key	Text Content
Dialogue State Tracking	Prompt	Give the dialogue state of the last utterance in the following dialogue in the form of ’STATE: Domain-Intent: [Slot, Possible value], ... (for example: STATE: Hotel-Inform: [’area’, ’centre’]) by using the following pre-defined slots and possible values: Intents: Request, Inform, general-thank, general-bye Domain: hotel, Slots: pricerange, Possible values: [’expensive’, ’cheap’, ’moderate’] Domain: hotel, Slots: type, Possible values: [’guesthouse’, ’hotel’] Domain: hotel, Slots: parking, Possible values: [’free’, ’no’, ’yes’] Domain: hotel, Slots: bookday, Possible values: [’monday’, ’tuesday’, ’wednesday’, ’thursday’, ’friday’, ’saturday’, ’sunday’] Domain: hotel, Slots: bookpeople, Possible values: [’1’, ’2’, ’3’, ’4’, ’5’, ’6’, ’7’, ’8’] Domain: hotel, Slots: bookstay, Possible values: [’1’, ’2’, ’3’, ’4’, ’5’, ’6’, ’7’, ’8’] Domain: hotel, Slots: stars, Possible values: [’0’, ’1’, ’2’, ’3’, ’4’, ’5’] Domain: hotel, Slots: internet, Possible values: [’free’, ’no’, ’yes’] Domain: hotel, Slots: name, Possible values: [] Domain: hotel, Slots: area, Possible values: [’centre’, ’east’, ’north’, ’south’, ’west’] Domain: hotel, Slots: address, Possible values: [?] Domain: hotel, Slots: phone, Possible values: [?] Domain: hotel, Slots: postcode, Possible values: [?] Domain: hotel, Slots: ref, Possible values: [?] USER: I need a five starts hotel close to a mall and main restaurants. The hotel should include free wifi in the room. STATE : ’Hotel-Inform’: [[’internet’, ’yes’], [’stars’, ’5’]] SYSTEM: I’m sorry, but there are no 5-star hotels anywhere in Cambridge. Would you like to broaden your criteria? USER: How about something with 4 stars and something expensive. STATE : ’Hotel-Inform’: [[’pricerange’, ’expensive’], [’stars’, ’4’]] SYSTEM: I have 2 options for you - The Huntingdon Marriott on the west side, and the University Arms in the center. Are you interested in one of these? USER: Maybe. Is either one a 4 star hotel? If so, I’d like to book a room for 4 nights. STATE : ’Hotel-Inform’: [[’bookstay’, ’4’], [’stars’, ’4’]] SYSTEM: Both are 4 stars. How many days would you be staying, and how many people are in your party? I can see if something is available. USER: I need a reservation for 4 nights starting Saturday.
	Gold	STATE : ’Hotel-Inform’: [[’bookday’, ’saturday’], [’stars’, ’4’]]
	ChatGPT Generation	STATE: ’Hotel-Inform’: [[’bookday’, ’saturday’], [’bookstay’, ’4’], [’stars’, ’4’]]
Response Generation using Actions	Prompt	Continue the dialogue as a task-oriented dialogue system called SYSTEM. The answer of SYSTEM should follow the ACTION provided next while answering the USER’s last utterance: ACTION: ’Restaurant-Inform’: [[’phone’, ’01223316074’], [’pricerange’, ’cheap’]] USER: Where is a 4 star hotel located in North Cambridge? SYSTEM: I have several options, almost all of which are moderate guesthouses. I do have one cheap one, too. Would you like to book a room? USER: Sure , that could be nice SYSTEM: OK, how many are in your party, what day will you arrive, and how many nights will you be staying? USER: I actually don’t need reservations I just need the phone number, price range.
	Gold	SYSTEM: As I mentioned it is cheap and the phone number is 01223316074.
	ChatGPT Generation	SYSTEM: The phone number for the 4-star hotel located in North Cambridge is 01223316074 and it is in the cheap price range.

Table 6: Examples of modular Task-Oriented Dialogue using ChatGPT: dialogue state tracking and response generation

Task	Key	Text Content
Multi-turn unified approach	Prompt	Use the following knowledge base to complete the task of "recommending a restaurant" by continuing the conversation as a task-oriented dialogue system: Restaurant: Mama Julia, Food: French, Price: Expensive, Location: 7th street, Rating: 5 Restaurant: Papa John, Food: Italian, Price: Moderate, Location: 3rd street, Rating: 4 Restaurant: The Crossroad, Food: Morocco, Price: Moderate, Location: Downtown, Rating: 2 Restaurant: Tacos City, Food: Mexian, Price: Cheap, Location: Center, Rating: 1 Restaurant: Golden Rice Bowl, Food: Chinese, Price: Cheap, Location: 3rd district, Rating: 3 Restaurant: Veggie Garden, Food: Chinese, Price: Expensive, Location: Town Hall, Rating: 4 Restaurant: Pizza House, Food: Italian, Price: Moderate, Location: 3rd street, Rating: 2 Restaurant: The Palace, Food: Vietnamese, Price: Expensive, Location: Hotel Grandview, Rating: 5 Hi, I'm looking for a restaurant
	ChatGPT	Sure, what type of food are you in the mood for?
	Prompt	I only want restaurants rated 3 or higher. Any suggestions?
	ChatGPT	Here are some restaurants that have a rating of 3 or higher: Mama Julia (French, Expensive, 5 rating), Papa John (Italian, Moderate, 4 rating), Golden Rice Bowl (Chinese, Cheap, 3 rating), Veggie Garden (Chinese, Expensive, 4 rating), and The Palace (Vietnamese, Expensive, 5 rating). Which one would you like to know more about?
	Prompt	What food does The Palace serve?
	ChatGPT	The Palace serves Vietnamese cuisine.
	Prompt	Is Taxos City expensive to eat at?
	ChatGPT	Tacos City is considered as cheap restaurant.
	Prompt	I want to eat Italian food, any recommendations?
	ChatGPT	Based on the information provided in the knowledge base, I would recommend Papa John and Pizza House for Italian food. Papa John is rated 4 out of 5 and is classified as "Moderate" in terms of price, while Pizza House is rated 2 out of 5 and is also classified as "Moderate". Papa John is located on 3rd street, while Pizza House is located on 3rd street as well.
	Prompt	Did you forget I only wanted restaurants rated 3 or higher?
	ChatGPT	I apologize, I must have misunderstood your previous request. Based on the knowledge base provided, The only Italian restaurant that is rated 3 or higher is Papa John which is rated 4 out of 5 and is located on 3rd street. It is classified as "Moderate" in terms of price.
	Prompt	Great thanks for your help.
	ChatGPT	You're welcome! I hope you enjoy your meal at Papa John. If you have any other questions or need further assistance, feel free to ask.

Table 7: Example of multi-turn unified approach for Task-Oriented Dialogue using ChatGPT

ChatGPT	InstructGPT	text-davinci-003
The language of the text appears to be a variant of the Bugis language spoken in Indonesia.	The language of the text is the Sasak language, spoken in Lombok, Indonesia.	The text is written in Buginese.
I am sorry, I do not recognize the language of the text.	The language of the text is Koyukon Athabascan.	The text is in the Balinese language.
The language of the text appears to be a dialect of the Indonesian language.	The language of the text is Indonesian.	The language of the text is Indonesian.

Table 8: Example of **Buginese** language identification response from ChatGPT, InstructGPT, and text-davinci-003.

Language	#Speakers	CC Size (%)	Language Category
English (eng)	1.452B	46.320	HRL
Chinese (zho)	1.118B	4.837	HRL
French (fra)	235M	4.604	HRL
Indonesian (ind)	199M	0.781	MRL
Korean (kor)	81.7M	0.679	MRL
Javanese (jav)	68.3M	0.002	LRL
Sundanese (sun)	32.4M	0.001	LRL
Buginese (bug)	5M	0.000	X-LRL

Table 9: The statistics of languages used in our language disparity experiment. **HRL**, **MRL**, **LRL**, **X-LRL** denote high-, medium-, low-, extremely low-resourced language respectively.

D ChatGPT on Multilinguality

We present the statistics of language under study in Table 9. In the following section, we provide the insights that we find during our experiment in exploring multilingual capability of ChatGPT.

ChatGPT understands LRL sentences but fails to identify their language As shown in Table 10, ChatGPT correctly classifies the languages for English and Indonesian 100% of the time. While for the language identification for Javenese and Buginese, ChatGPT either misclassifies the samples as other languages or is unable to determine the language. Nevertheless, ChatGPT performance on the sentiment analysis in Javanese is only slightly lower compared to English and Indonesian which suggests that ChatGPT can understand the semantic meaning of sentences in low-resource languages without having the knowledge to identify the language itself. This limitation of language identification in LMs aligns with the result from BIG-bench (Srivastava et al., 2022).

ChatGPT displays better human-preferred responses As shown in Table 8, ChatGPT lets the user know that its prediction is uncertain when it does not completely understand the language and also provides broader information regarding the language, such as location and tribe of which the predicted language is spoken. This fact provides evidence regarding the benefit of using the RLHF approach compared to other training approaches for aligning LLMs with human preferences.

ChatGPT understands non-Latin scripts better than it can generate them Despite being high-resource and medium-resource languages, the translation from English to Chinese and Korean is much inferior to the other languages with Latin scripts, i.e., French or Indonesian. Similarly, prior works focusing on transliteration (Chau and Smith, 2021; Muller et al., 2021) have shown the effectiveness of utilizing Latin scripts over other scripts, e.g., Cyrillic, Georgian, Arabic, etc, especially for low-resource languages. Interestingly, this problem of using non-Latin scripts is less severe for translation from Chinese and Korean to English, which suggests that ChatGPT can better neutralize the effect of non-Latin scripts as source languages (Wan, 2022), but it still lacks the ability to generate non-Latin script languages.

Language	SA Acc.	LID Acc.
English	84%	100%
Indonesian	80%	100%
Javanese	78%	0%
Buginese	56%	12%

Table 10: Accuracy of ChatGPT on Sentiment Analysis (SA) and Language Identification (LID) tasks.

E Multimodality: Flag Drawing Task

Task Formulation We uniformly collect 50 national flags from different continents and conduct the flag-drawing task on ChatGPT. The flag-drawing task contains three steps:

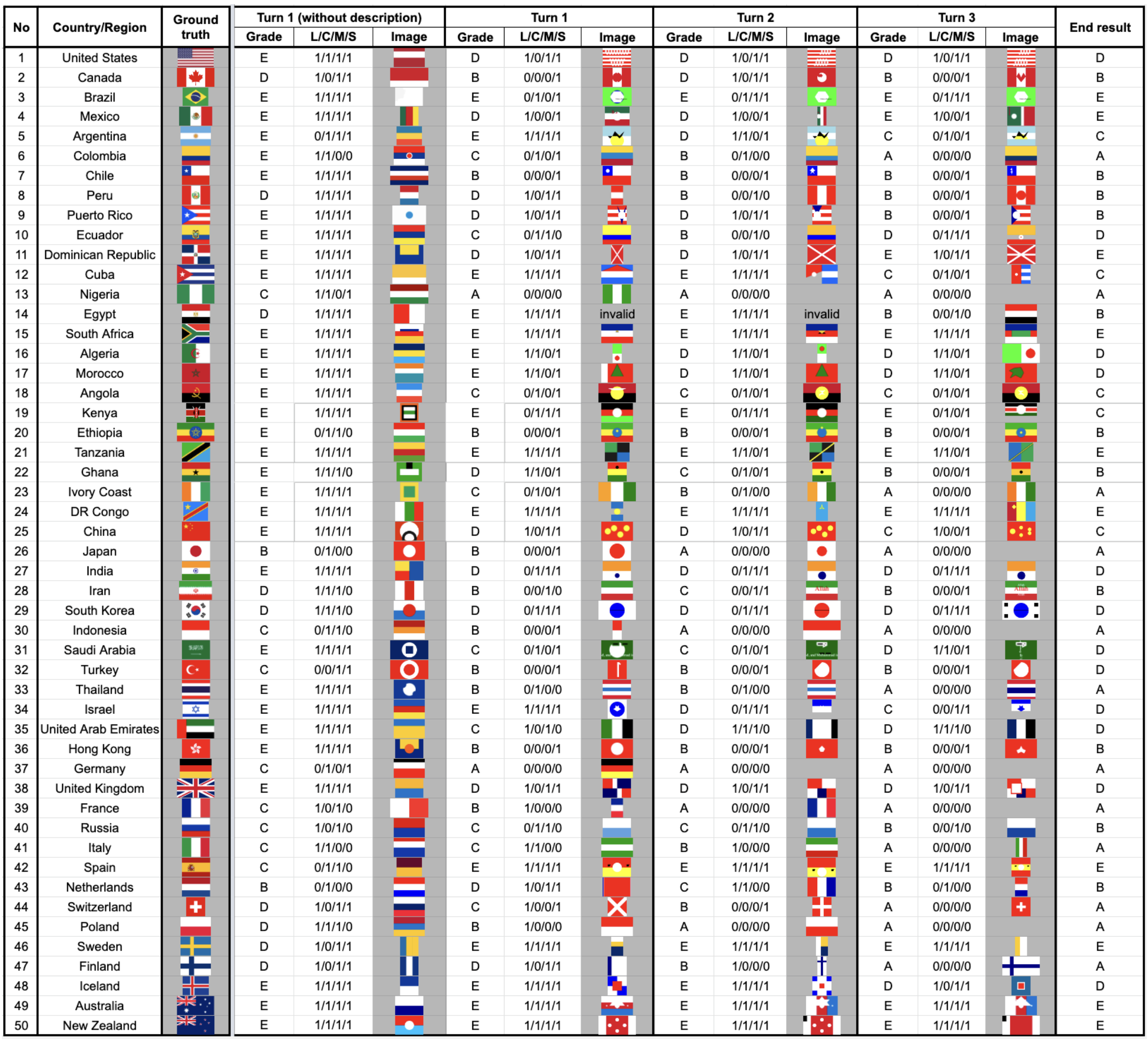
1. Ask ChatGPT to illustrate the appearance of the flag using the prompt "`Describe how the <NATION> flag looks like`".

2. Based on the description, ask ChatGPT to generate the SVG code of that flag by prompting "`Generate a code snippet to represent that flag in SVG format`".

3. If the generated image contains errors, we iteratively ask ChatGPT to fix them.

There are four types of evaluation criteria: 1) layout 2) color 3) missing components 4) shape/size. In each round of fixing, we ask ChatGPT to revise only one type of error with the prompt "`<ERROR DESCRIPTION>. Revise the image`". We terminate the conversation once the generated flag becomes perfect or we have already passed two rounds of fixing.

Evaluation The generated flag images are evaluated by the aforementioned four error types as



No	Country/Region	Ground truth	Turn 1 (without description)			Turn 1			Turn 2			Turn 3			End result
			Grade	L/C/M/S	Image	Grade	L/C/M/S	Image	Grade	L/C/M/S	Image	Grade	L/C/M/S	Image	
1	United States		E	1/1/1/1		D	1/0/1/1		D	1/0/1/1		D	1/0/1/1		D
2	Canada		D	1/0/1/1		B	0/0/0/1		D	1/0/1/1		B	0/0/0/1		B
3	Brazil		E	1/1/1/1		E	0/1/0/1		E	0/1/1/1		E	0/1/1/1		E
4	Mexico		E	1/1/1/1		D	1/0/0/1		D	1/0/0/1		E	1/0/0/1		E
5	Argentina		E	0/1/1/1		E	1/1/1/1		D	1/1/0/1		C	0/1/0/1		C
6	Colombia		E	1/1/0/0		C	0/1/0/1		B	0/1/0/0		A	0/0/0/0		A
7	Chile		E	1/1/1/1		B	0/0/0/1		B	0/0/0/1		B	0/0/0/1		B
8	Peru		D	1/1/1/1		D	1/0/1/1		B	0/0/1/0		B	0/0/0/1		B
9	Puerto Rico		E	1/1/1/1		D	1/0/1/1		D	1/0/1/1		B	0/0/0/1		B
10	Ecuador		E	1/1/1/1		C	0/1/1/0		B	0/0/1/0		D	0/1/1/1		D
11	Dominican Republic		E	1/1/1/1		D	1/0/1/1		D	1/0/1/1		E	1/0/1/1		E
12	Cuba		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		C	0/1/0/1		C
13	Nigeria		C	1/1/0/1		A	0/0/0/0		A	0/0/0/0		A	0/0/0/0		A
14	Egypt		D	1/1/1/1		E	1/1/1/1	invalid	E	1/1/1/1	invalid	B	0/0/1/0		B
15	South Africa		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E
16	Algeria		E	1/1/1/1		E	1/1/0/1		D	1/1/0/1		D	1/1/0/1		D
17	Morocco		E	1/1/1/1		E	1/1/0/1		D	1/1/0/1		D	1/1/0/1		D
18	Angola		E	1/1/1/1		C	0/1/0/1		C	0/1/0/1		C	0/1/0/1		C
19	Kenya		E	1/1/1/1		E	0/1/1/1		E	0/1/1/1		E	0/1/0/1		C
20	Ethiopia		E	0/1/1/0		B	0/0/0/1		B	0/0/0/1		B	0/0/0/1		B
21	Tanzania		E	1/1/1/1		E	1/1/1/1		E	1/1/0/1		E	1/1/0/1		E
22	Ghana		E	1/1/1/0		D	1/1/0/1		C	0/1/0/1		B	0/0/0/1		B
23	Ivory Coast		E	1/1/1/1		C	0/1/0/1		B	0/1/0/0		A	0/0/0/0		A
24	DR Congo		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E
25	China		E	1/1/1/1		D	1/0/1/1		D	1/0/1/1		C	1/0/0/1		C
26	Japan		B	0/1/0/0		B	0/0/0/1		A	0/0/0/0		A	0/0/0/0		A
27	India		E	1/1/1/1		D	0/1/1/1		D	0/1/1/1		D	0/1/1/1		D
28	Iran		D	1/1/1/0		B	0/0/1/0		C	0/0/1/1		B	0/0/0/1		B
29	South Korea		D	1/1/1/0		D	0/1/1/1		D	0/1/1/1		D	0/1/1/1		D
30	Indonesia		C	0/1/1/0		B	0/0/0/1		A	0/0/0/0		A	0/0/0/0		A
31	Saudi Arabia		E	1/1/1/1		C	0/1/0/1		C	0/1/0/1		D	1/1/0/1		D
32	Turkey		C	0/0/1/1		B	0/0/0/1		B	0/0/0/1		B	0/0/0/1		D
33	Thailand		E	1/1/1/1		B	0/1/0/0		B	0/1/0/0		A	0/0/0/0		A
34	Israel		E	1/1/1/1		E	1/1/1/1		D	0/1/1/1		C	0/0/1/1		D
35	United Arab Emirates		E	1/1/1/1		C	1/0/1/0		D	1/1/1/0		D	1/1/1/0		D
36	Hong Kong		E	1/1/1/1		B	0/0/0/1		B	0/0/0/1		B	0/0/0/1		B
37	Germany		C	0/1/0/1		A	0/0/0/0		A	0/0/0/0		A	0/0/0/0		A
38	United Kingdom		E	1/1/1/1		D	1/0/1/1		D	1/0/1/1		D	1/0/1/1		D
39	France		C	1/0/1/0		B	1/0/0/0		A	0/0/0/0		A	0/0/0/0		A
40	Russia		C	1/0/1/0		C	0/1/1/0		C	0/1/1/0		B	0/0/1/0		B
41	Italy		C	1/1/0/0		C	1/1/0/0		B	1/0/0/0		A	0/0/0/0		A
42	Spain		C	0/1/1/0		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E
43	Netherlands		B	0/1/0/0		D	1/0/1/1		C	1/1/0/0		B	0/1/0/0		B
44	Switzerland		D	1/0/1/1		C	1/0/0/1		B	0/0/0/1		A	0/0/0/0		A
45	Poland		D	1/1/1/0		B	1/0/0/0		A	0/0/0/0		A	0/0/0/0		A
46	Sweden		D	1/0/1/1		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E
47	Finland		D	1/0/1/1		D	1/0/1/1		B	1/0/0/0		A	0/0/0/0		A
48	Iceland		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		D	1/0/1/1		D
49	Australia		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E
50	New Zealand		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E	1/1/1/1		E

Figure 4: Complete results of the flag drawing task. Multi-turn refinement allows ChatGPT to generate a more similar image to the ground truth image.

criteria. We further assess the image quality with five grades, A $\sim$ E, which indicate zero to four (or above) errors. We assign grades to each round so that we can assess the number of improvements and degradation through conversational interactions (post-editing). The full results are shown in Figure 4.

F Details for Reasoning Evaluations

Table 11 shows the categories of reasoning that are evaluated in this paper as well as corresponding datasets. The following section introduces each of the categories and detailed experimental setup and/or analysis.

F.1 Logical Reasoning

Inductive and deductive are categorized by "a degree to which the premise supports the conclusion" based on logic and philosophy (Qiao et al., 2022; Rogers et al., 2022; Hawthorne, 2021). Inductive reasoning is based on "observations or evidence" while deductive is based on "truth of the premises" (i.e., necessarily true inference) (Douven, 2017). Another way to categorize is based on the "direction of reasoning" – deductive is from premise to conclusion while abductive is from conclusion to the most probable premise that supports the conclusion (Walton, 2014).

Inductive and deductive reasoning are common forms of logical reasoning that are categorized by "a degree to which the premise supports the conclusion" based on logic and philosophy (Qiao et al.,

Categories	Dataset
Deductive	EntailmentBank (Dalvi et al., 2021) bAbI (task 15) (Weston et al., 2016b)
Inductive	CLUTRR (Sinha et al., 2019) bAbI (task16) (Weston et al., 2016b)
Abductive	αNLI (Bhagavatula et al., 2020)
Temporal	Timedial (Qin et al., 2021)
Spatial	SpartQA (Mirzaee et al., 2021) StepGame (Shi et al., 2022a)
Mathematical	Math (Saxton et al., 2019)
Commonsense	CommonsenseQA (Talmor et al., 2018) PiQA (Bisk et al., 2020) Pep-3k (Wang et al., 2018b)
Causal	E-Care (Du et al., 2022)
Multi-hop	HotpotQA (Yang et al., 2018)
Analogical	Letter string analogies (Webb et al., 2022b)

Table 11: Reasoning categories and corresponding datasets used to evaluate ChatGPT in this work.

Deductive Reasoning Tasks		
bAbI - task 15	**bAbI - task 15 (prompt engineered)**	**EntailmentBank**
19/30	28/30	28/30
Inductive Reasoning Tasks		
bAbI - task16	**bAbI - task 16 (prompt engineered)**	**CLUTRR**
0/30	20/30	13/30

Table 12: Inductive vs. Deductive Reasoning. ChatGPT performs better deduction rather than induction. Engineering the prompt to explicitly ask ChatGPT to do reasonable inference improves its reasoning capability. The scores are in accuracy over tested samples.

2022; Rogers et al., 2022; Hawthorne, 2021). Deductive reasoning involves processes of driving specific conclusions based on more general premises. On the contrary inductive reasoning involves specific observation of patterns, processing them on increasingly abstract cycles of hypothetico-deductive reasoning to draw a more general conclusion (Lawson, 2005). Comparing the two types of reasoning, deduction requires less "guessing" from the perspective of ChatGPT, as induction requires figuring out rules (Rogers et al., 2022). The former can be viewed as top-down while the latter is bottom-up.

F.1.1 Deductive vs. Inductive Reasoning

Deductive reasoning involves processes of driving specific conclusions based on *more general premises*. On the contrary, inductive reasoning involves *specific observation of patterns*, processing them on increasingly abstract cycles of hypothetico-deductive reasoning to draw a more general conclusion (Lawson, 2005). Comparing the two types of reasoning, deduction requires less "guessing" from the perspective of ChatGPT, as induction requires figuring out rules (Rogers et al., 2022). The former can be viewed as top-down while the latter is bottom-up.

We explore ChatGPT's ability of inductive and deductive reasoning in two different levels: 1) basic and 2) advanced. Basic-level tasks are the prerequisites to probe reasoning. While solving these tasks does not necessarily indicate full reasoning capability, if ChatGPT fails on any of these tasks, then there are likely real-world tasks that it will fail on too if they require similar reasoning mechanisms. Consequently, the advanced-level tasks are there to probe those capabilities in real-world tasks where the noises are present, and solving them requires a more systematic generalization. Additionally, we choose tasks that do not require or are dependent on external knowledge and the solution could be only derived by premises to focus on dissecting the capability of each reasoning mechanism.

Results & Detailed Analysis ChatGPT answers *"It is not specified what <attribute> <entity> is."* for most of the time when it was asked a question requiring inductive reasoning. However, when ChatGPT is explicitly asked for reasonable inference with a prompt *"Based on the given facts, do a reasonable inference on this question using inductive reasoning:"*, its ability for inductive reasoning increases. Yet, it is still not as good as in deduction as the same prompt engineering also helps increases its ability for deductive reasoning.

We could derive similar insight as ChatGPT only correctly answered for half of the time while it could make inferences deductively well for 90% of the time. CLUTRR (Sinha et al., 2019) requires induction on extracting relations between entities, and in the ChatGPT responses, it often asks for more information to make inferences. An interesting finding along with CLUTRR was that ChatGPT can't differentiate son and grandson but can differentiate daughter and granddaughter when it induces the logical rules governing kinship relationships. We show all performances in Table 12 and some of the prompting samples in Table 13. We follow (Qiao et al., 2022) categorization on the deductive and inductive reasoning datasets, but

Task	Prompt	ChatGPT answer	Gold	T/F
Deductive Reasoning Tasks				
bAbI (task 15)	1 Wolves are afraid of mice. 2 Sheep are afraid of mice. 3 Winona is a sheep. 4 Mice are afraid of cats. 5 Cats are afraid of wolves. 6 Jessica is a mouse. 7 Emily is a cat. 8 Gertrude is a wolf. 10 What is winona afraid of?	Winona, being a sheep, is afraid of mice.	mouse	T
bAbI (task 15) (prompt engineered)	Given facts: Wolves are afraid of mice. Sheep are afraid of mice. Winona is a sheep. Mice are afraid of cats. Cats are afraid of wolves. Jessica is a mouse. Emily is a cat. Gertrude is a wolf. Based on the given facts above, do a reasonable inference on this question using deductive reasoning: What is winona afraid of?	Based on the given facts, it can be inferred that Winona is afraid of mice.	mouse	T
Entailment-Bank	earth is a kind of planet. a planet is a kind of celestial object / celestial body. earth is located in the milky way galaxy. Which object is a planet found in the Milky Way Galaxy?	Earth is a planet located in the Milky Way galaxy.	Earth	T
Inductive Reasoning Tasks				
bAbI (task 16)	1 Bernhard is a swan. 2 Greg is a frog. 3 Brian is a swan. 4 Julius is a lion. 5 Greg is gray. 6 Julius is yellow. 7 Lily is a lion. 8 Lily is green. 9 Brian is yellow. 10 What color is Bernhard?	It is not specified in the given information what color Bernhard is.	yellow	F
bAbI (task 16) (prompt engineered)	Given facts: Bernhard is a swan. Greg is a frog. Brian is a swan. Julius is a lion. Greg is gray. Julius is yellow. Lily is a lion. Lily is green. Brian is yellow. The most recent fact is the correct fact. Based on the given facts above, do a reasonable inference on this question using inductive reasoning: What color is Bernhard?	Based on the given facts, it can be inferred that Bernhard is yellow, as Brian is also stated to be yellow.	yellow	T
CLUTRR	[Jason] and his wife [Gabrielle] went to the beach to watch the fireworks on the 4th of July. [Jason] and his daughter [Alma] took a day off school to go to the zoo... Who is Alma to Gabrielle?	Alma is the daughter of Jason and Gabrielle.	daughter	T
	[Jason] took his grandson [Donald] fishing. [Russell] enjoys going fishing with his brother. His name is [Donald]... Who is Russell to Jason?	Russell is the brother of Jason.	grandson	F

Table 13: Prompting samples on deductive and inductive reasoning tasks. ChatGPT is performing better deduction rather than induction. On both types of reasoning, when ChatGPT is explicitly asked to do reasonable inferences, its ability for reasoning increases. Additionally, it also makes frequent mistakes regarding the grandson's kinship.

StepGame (Basic) Breakdown Analaysis					
	Result	Example	ChatGPT answer	Gold	T/F
Clock-position	5/20	G is at Y's 6 o'clock. What is the relation of the agent Y to the agent G?	The agent Y is to the right of the agent G.	Above	F
Basic Cardinal	17/20	D and K are parallel, and D is under K. What is the relation of the agent K to the agent D?	The spatial relation of the agent K to the agent D is that K is above D.	Above	T
Diagonal	11/20	W presents lower left to I. What is the relation of the agent I to the agent W?	The relation of the agent I to the agent W is lower-left.	Upper-Right	F

Table 14: A provided illustration to help the readers to understand each comparison between the categories (not the actual prompts). We provide the options to ChatGPT as: `Choose from: left, right, above, below, lower-left, lower-right, upper-left, upper-right`.

we only use the QA part of EntailmentBank, that the authors took from ARC dataset (Clark et al., 2018), as we aim to test for reasoning capability. Regarding EntailmentBank, it might trigger the universe-related knowledge out of ChatGPT, which could help the model to derive the correct answer, although the test set is designed to test deductive reasoning skills. One of the future explorations would be with checking the rationale of ChatGPT as a follow-up question.

F.1.2 Abductive Reasoning

Abductive reasoning is the inference to the most plausible explanation given observations. For instance, "if Jenny finds her house in a mess when she returns from work, and remembers that she left a window open, she can hypothesize that a thief broke into her house and caused the mess" [4]. We test ChatGPT's language-based abductive reasoning ability with 30 samples from αNLI dataset (Bhagavatula et al., 2020), which requires the model to select the most plausible explanation given the conclusion. Based on our test, it could achieve 86.7% (26 out of 30) accuracy.

F.2 Non-textual Semantic Reasoning

Mathematical reasoning Mathematical capabilities or numerical reasoning has been frequently mentioned to be lacking for LLMs, not only ChatGPT (Frieder et al., 2023). Frieder et al. test ChatGPT's capability with publicly available datasets as well as the human-curated dataset, which consists of 728 prompts. The shared findings for ChatGPT's mathematical capabilities include 1) ChatGPT often understands the question but fails to provide correct solutions; 2) it shows inconsistent poor performance on graduate-level advanced mathematics; 3) it has a great ability to search for mathematical objects. [5] We also test separately on MATH dataset. Not surprisingly, it could only score 23.33% (7/30) for the MATH dataset (Saxton et al., 2019), which tests mathematical reasoning.

Temporal reasoning Temporal reasoning is mentioned a few times in the literature but is less common than others. It tests the understanding of the time duration of and the relation between events. For this category, we conduct experiments on the dataset TimeDial (Qin et al., 2021), which solely requires temporal reasoning. We follow the format of the task in the BIG-bench benchmark (Srivastava et al., 2022), which is multiple-choice (single correct answer), Overall, ChatGPT correctly answers 86.67% of the time (26/30), suggesting that it has a decent temporal reasoning ability. Also, compared to Chinchilla and Gopher which have the accuracy of 68.8% and 50.9% respectively, ChatGPT shows a promising improvement for LLMs in that aspect.

[4] An example provided by Bhagavatula et al. (2020).

[5] Refer to detailed findings in the original paper.

Spatial Reasoning Tasks			
Dataset	Total	Basic	Hard
StepGame	26/60	19/30	7/30
SpartQA	28/64	20/32	8/32

Table 15: Spatial reasoning ability of ChatGPT. Overall, ChatGPT falls short of the task.

Spatial Reasoning Spatial reasoning is using an understanding of spatial relations among different objects and spaces. For spatial reasoning, we utilize two existing datasets: SpartQA (Mirzaee et al., 2021) and StepGame (Shi et al., 2022a), which compose of story-question pairs about k relations of k+1 (where k is up to 10) entities written in natural language. ChatGPT is asked to answer spatial relations between two entities based on the provided descriptions of different entities. ChatGPT falls short of the spatial reasoning tasks, as shown in Table 15, with overall success rates of 43.33% for StepGame and 43.75% for SpartQA. ChatGPT could only score 25% on SpartQA (hard), which covers multiple spatial reasoning sub-types, and 23.33% for stepGame (Hard) with k=9. ChatGPT could not provide any spatial relations but instead generated *"It is not specified in the given description"*. Even with the fine-tuned models, as the number of relations (k) increases in context description, performance drops (Shi et al., 2022a).

To understand spatial reasoning ability at a more elementary level, we test with less complicated examples from StepGame which we refer to as **StepGame (Basic)**. It does not involve multi-hop reasoning but purely spatial relation between two entities. (e.g, *"C is sitting at the top position to Y. What is the relation of the agent Y to the agent C?"*). We test for basic spatial relations with 8 labels from StepGame {left, right, above, below, lower-left, lower-right, upper-left, upper-right}. When we test on StepGame (Basic), ChatGPT scores higher (63.33%).

We investigate the errors that it often fails to

Commonsense Reasoning Tasks		
CommonsenseQA	PiQA	Pep-3k (Hard)
27/30	25/30	28/30

Table 16: Commonsense reasoning ability of ChatGPT. ChatGPT shows good performance of commonsense reasoning capability on the three test data we test it on.

understand clock direction (e.g., "W is at K's 3 o'clock") and diagonal spatial relations. We further analyze the results by breaking down the test examples of StepGame (Basic) into two comparisons: i) types of directions (basic cardinal vs. diagonal) and ii) ways of spatial description for cardinal directions (basic cardinal[6] vs. clock-position cardinal). We take 20 more samples for each category (basic cardinal, diagonal, clock-position cardinal) and tested them as illustrated in Table 14.

- **ChatGPT poorly infers with clock-position description.** Although it is a simple cardinal direction, ChatGPT could only correctly answer for 5 samples (25%), which is clearly poorer performance in comparison to performance with the basic cardinal description (17 correct answers).

- **ChatGPT is worse at the diagonal position.** It correctly answers around half of the time (55%), which is worse than basic cardinal points (85%). Even with analysis from StepGame (Hard), among the correct 7 answers, there is only one diagonal direction that ChatGPT gets correctly while the others are all cardinal points. For those answers that require diagonal points, ChatGPT only could infer cardinal points for some examples.

F.3 Commonsense Reasoning

To evaluate ChatGPT's capability on commonsense reasoning, we first test it on two widely used benchmark datasets CommonsenseQA (Talmor et al., 2018) and PiQA (Bisk et al., 2020). CommonsenseQA focuses on general commonsense question answering such as *"Where is a business restaurant likely to be located?"*, and PiQA is about physical commonsense reasoning: given a sentence such as *"When boiling butter, when it's ready, you can "*, the goal is to fill in the blank with one of two answer options, *"Pour it onto a plate"* and *"Pour it onto a jar"*. We use the validation split for both of the datasets since there are no labels provided on the test set that we retrieve. We also further probe ChatGPT by evaluating a more challenging commonsense reasoning dataset in a more comprehensive way. We use Pep-3k (Wang et al., 2018b), which requires the model to recognize plausible but possibly novel events, such as *"man swallow paintball"*. Each instance in the Pep-3k is an s-v-o predicate, and the task is to judge if the predicate is plausible or not. But instead of evaluating ChatGPT's performance only based on the binary judgment, we also check if the answer contains relevant rationales (explanations) that lead to its judgment.

For the Pep-3k samples, we prepend the s-v-o predicate with *"Please judge if this predicate is (likely) plausible or implausible:"* to prompt ChatGPT. We show the results in Table 16. As we see, ChatGPT performs quite well on the three datasets in terms of answer accuracy, which matches our anticipation. Furthermore, as we also check the rationales in ChatGPT's answer when evaluating Pep-3k samples, we can see that ChatGPT does quite well not only in terms of answer accuracy but also in generating reasonable reasoning procedures to support its answer. We show a concrete example in Table 17. As we can see, ChatGPT's answer explains well what kinds of materials are usually cut through with knives (i.e., food, paper, or wood). Then, it reasons why rocks cannot be chopped with a knife by explaining 'rocks are much harder than these materials.' While our findings are based on 30 samples from each dataset, we see the potential in ChatGPT's commonsense reasoning capability, and further large-scale investigation is worth exploring.

Prompt	Please judge if this predicate is (likely) plausible or implausible: "knife chop rock"
ChatGPT Answer	It is **implausible** that a knife could chop through a rock. Knives are made to cut through softer materials such as food, paper, or wood. Rocks are much harder than these materials and a knife would not be able to make a dent in a rock, let alone chop it.

Table 17: An example from Pep-3k (Wang et al., 2018b) for commonsense reasoning of ChatGPT. We make the main answer **bold**, and highlight the explanation by green color.

[6]Those of which spatial relations are described with explicit vocabulary.

Causal	Multi-hop	Analogical
E-CARE	HotpotQA	Letter string analogies
24/30	8/30	30/30

Table 18: Results for causal, multi-hop, and analogical reasoning. ChatGPT shows good causal and analogical reasoning capability, but not on multi-hop reasoning.

F.4 Causal, Multi-Hop, and Analogical Reasoning

Causal Reasoning Causal reasoning is the process of identifying the relationship between causes/actions and effects/changes (i.e., causality) (Thomason, 2018; Huang and Chang, 2022). We test ChatGPT on 30 samples of human-annotated explainable CAusal REasoning dataset (E-CARE) (Du et al., 2022) and it could score 24 samples correctly (80%). Note that our evaluation is mainly based on whether the model can make a judgment on correct causes or effects instead of its generated explanation of why the causation exists.

Multi-hop Reasoning To be able to reason over a larger context, a system has to perform multi-hop reasoning over more than one piece of information to arrive at the answer (Mavi et al., 2022). We test ChatGPT's multi-hop reasoning capability on 30 samples of HotpotQA dataset (Yang et al., 2018) and we find that ChatGPT has difficulty performing with such capability, only answering 8 samples correctly, although the questions posed are only 2-hops. It is worth noting that ChatGPT oftentimes generates the answer in a short passage of explanations, thus we evaluate manually each of the ChatGPT responses to check its accuracy. This aligns with the findings that LLMs are also limited in several ways, and fail to produce accurate predictions due to their inability to accomplish complex reasoning, such as solving tasks that require multi-hop reasoning (Ott et al., 2023).

Analogical Reasoning Analogical reasoning is a way of thinking that relies upon an analogy, comparing two or more objects or systems of objects (Bartha, 2013) to drive a conclusion. We test with 30 samples from Webb et al. (2022b) and evaluate based on human evaluation, to see if the generated answer match with/contain the gold answer. ChatGPT could correctly answer all 30 examples, which may reveal that ChatGPT has a good capability in analogical reasoning skills.

G Details for Hallucination Evaluations

There exist two categories of hallucination (Ji et al., 2022a). *Intrinsic hallucinations* that refers to the LLM generation that contradicts the source/input content. *Extrinsic hallucinations* that refers to the LLM generations that cannot be verified from the source/input content (i.e., output that can neither be supported nor contradicted by the source). In Table 19, we share examples of these hallucination types detected from different task explorations. With the setting of tasks we test, we often find extrinsic hallucinations, including both untruthful and factual ones, across various tasks such as Machine Translation, Question answering.

The intrinsic hallucinations are barely found. For instance, in the abstractive summarization task, in which neural models usually suffer from intrinsic hallucination, ChatGPT's generated summarisation did not include any intrinsic hallucination examples based on our experiments. It rather shows a factual extrinsic hallucination, for instance, ChatGPT could correctly paraphrase "Britain and five other countries" from source input into "P5+1 (US, UK, France, China, Russia, and Germany)," which is assessed to be factual. We could also observe an interesting intrinsic hallucination for our proposed multi-modal task, the flag drawing task. ChatGPT is first asked to generate a description of how the flags look before it is asked to generate code for the flag. Although it generates the correct description as "The flag of Mexico consists of three **vertical** bands [...]", the final drawing (SVG code) consists of **horizontal** bands.

However, extrinsic hallucinations often happen, including both untruthful and factual ones. In the QA task, we often find extrinsic hallucination to be non-factual which harms the final performance. For instance, in the question of asking for the relationship among entities, although step kindship is never mentioned in the question, ChatGPT answers the question with step kinship, as illustrated in Table 19. We could also observe that ChatGPT's weakness with extrinsic hallucination also degrades machine translation. When it is asked to translate the text "Like some other experts, he is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who already have Type 1 diabetes." into Korean, it contains a piece of information that was not found in the source, "저주파 치료" (transcutaneous electrical nerve stimulation) in the translated text.

Task	Key	Text Content
Machine Translation	Prompt	(Javanese) Krajan Anyar Mesir kuno nggumun marang monumen-monumen leluhure, sing nalika iku tetep apik luwih seka sewu taun.
	ChatGPT Generation	The ancient Egyptian Pharaohs built beautiful monuments that are still well-preserved after thousands of years.
	Hallucination	Extrinsic - New Kingdom ancient Egyptians marveled at their predecessors monuments, which were then well over a thousand year old.
Summarization	Prompt	[...] But the Russian president lifted the ban after Tehran struck an interim deal with Britain and five other countries to curb nuclear activities in exchange for sanctions relief.[...]
	ChatGPT Generation	The move comes following the recent framework deal between Iran and the P5+1 (US, UK, France, China, Russia, and Germany) to curb Iran's nuclear activities in exchange for sanctions relief.
	Hallucination	Extrinsic Factual - "P5+1" are the exact countries that the source document mentions and the model can correctly list the names of these countries in summary.
Task-oriented Dialogue	Prompt	Continue the dialogue as a task-oriented dialogue system called SYSTEM. The answer of SYSTEM should follow the ACTION provided next while answering the USER's last utterance: ACTION: 'Train-Inform': [['choice', '12'], ['leaveat', '05:16'], ['leaveat', '16:16']], 'Train-OfferBook': [['none', 'none']] USER: I need train reservations from norwich to cambridge SYSTEM: I have 133 trains matching your request. Is there a specific day and time you would like to travel? USER: I'd like to leave on Monday and arrive by 18:00. SYSTEM: There are 12 trains for the day and time you request. Would you like to book it now? USER: Before booking, I would also like to know the travel time, price, and departure time please.
	ChatGPT Generation	SYSTEM: The travel time for the trains from Norwich to Cambridge is approximately 1 hour and 30 minutes. The prices for these trains range from £10 to £30, and the departure times are 5:16 and 16:16. Please let me know if you would like to book a train now.
	Hallucination	Extrinsic - There are no information provided about neither travel time nor price range
Knowledge-Grounded Open-domain Dialogue	Prompt	Can we try dialogue generation? I will give you turns and you can generate the next turn, but only one. You can also consider the knowledge of "Pierre Coffin is starred in Despicable Me 2" and "Despicable Me 2 is written by Cinco Paul" for your reference in the dialogue. User: Do you know of any movies directed by Pierre Coffin?
	ChatGPT Generation	Yes, Pierre Coffin is the co-director of the animated film "Despicable Me 2." Cinco Paul is the film's writer.
	Hallucination	animated film is a factual extrinsic hallucination.
Question Answering	Prompt	[Jason] is very much in love with his wife. Her name is [Gabrielle]. [Jason] has a daughter called [Alma].. Who is Alma to Gabrielle?
	ChatGPT Generation	Alma is Jason's daughter and therefore Gabrielle's stepdaughter.
	Hallucination	intrinsic - stepkinship never mentioned in the prompt. Ground truth is daughter
Question Answering (Truthful QA)	Prompt	Where do people smoke more on average than in Russia?
	ChatGPT Generation	There are several countries where people smoke more on average than in Russia. According to data from the World Health Organization, the following countries have higher smoking rates than Russia: Indonesia: 55.2% of adults smoke China: 52.9% of adults smoke Serbia: 47.5% of adults smoke Serbia and Montenegro: 45.6% of adults smoke Montenegro: 45.6% of adults smoke Greece: 44.5% of adults smoke Bosnia and Herzegovina: 41.8% of adults smoke Kazakhstan: 41.3% of adults smoke Kosovo: 41.3% of adults smoke Moldova: 40.9% of adults smoke It's important to note that these figures are estimates and may vary slightly from other sources.
	Hallucination	Extrinsic - the generated answer includes wrong statistics. (Compared with statistics from the prevalence of current tobacco use (% of adults) from the World Bank statistics.[7])
National Flag Drawing	Prompt (history)	[USER-prompt] : Describe how the Mexican flag looks like. [ChatGPT generation] : The flag of Mexico consists of three *vertical* bands in green, white, and red. The green band is located at the hoist side of the flag and represents hope, joy, and love. [...] [USER-prompt] : Generate a code snippet to represent that flag in SVG format.
	ChatGPT Generation	
	Hallucination	Intrinsic - It says there are three "vertical" bands but generates three horizontal bands.

Table 19: Examples of hallucination in different tasks

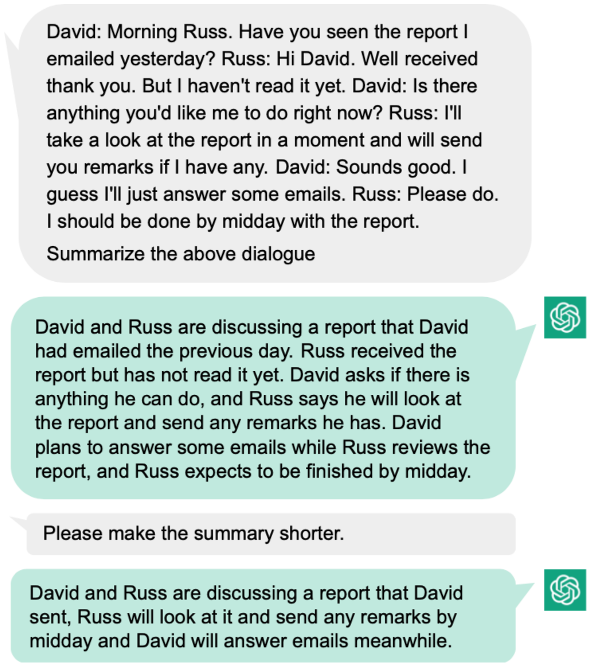


Figure 5: An example of dialogue summarization

H Details for Interactivity Evaluation

H.1 Interactivity on Summarization

Figure 5 shows an example of how multi-turn interaction helps to control the length of the summary.

Experiment: Two-turn iterative prompt.

1. Given an input dialogue as the context, we first input the prompt "`Summarize the above dialogue`" to the ChatGPT.

2. To refine the summary, we simply input another prompt – "`Please make the summary shorter`" after the first response.

Evaluation: We calculate the ROUGE scores (ROUGE-1, ROUGE-2, and ROUGE-L) of the first and second summaries and compare between turns.

H.2 Interactivity on Machine Translation

H.2.1 Experiment 1: Multi-turn Post-Editting

We explore the capability on translation from English to the target language. For the experiment, we adapt the dataset used in §2.2.2 which samples 30 parallel sentences from 6 language pairs in NusaX (Winata et al., 2022), Chinese, French, Indonesian , Korean, Javanese, and Sundanese.

1. Query model to translate to the target language using "`What is [TARGET_LANGUAGE] translation of the following sentence?\n\n[INPUT_SENTENCE]`"

Label	Metric	w/o APE	w/ APE
Post-Edited Marathi Text	HTER	88.14	**88.79**
	SacreBLEU	**4.81**	4.20
	METEOR	**13.10**	12.74
Source English Text	HTER	65.36	**63.13**
	SacreBLEU	25.54	**27.20**
	METEOR	43.71	**47.51**
	BERTScore	92.30	**92.59**

Table 20: Result of translation w/ and w/o post-editing on WMT 2022 English→Marathi APE shared task

2. Query for the post-editing using the following prompt template: "`Could you perform a post-editing to ensure the meaning is equivalent to "[INPUT_SENTENCE]"?`"

Evaluation: The post-editing results are manually validated by a native speaker in the corresponding language to validate: 1) whether the post-edited sentence is better than the translation one, and 2) whether the post-edited sentence is the correct translation of the given English sentence.

Based on the evaluation, performing automatic post-editing through interactive LLMs, such as ChatGPT, yields consistently better translation results compared to a single-turn machine translation, which is especially useful for translation in low-resource languages. We provide per-language examples of the machine-translated and post-edited sentences in Appendix J.

H.2.2 Experiment 2: Automatic post-editing

To further strengthen our hypothesis, we conduct an additional experiment on the automatic post-editing (APE) shared task dataset on WMT 2022 (Bhattacharyya et al., 2022), which focuses on English→Marathi post-editing task. Marathi (mar) is also a low-resource language with 0.02% data size on CommonCrawl. We sample 50 samples from the corresponding dataset.

Evaluation: 1) human-targeted translation error rate (HTER)[8], SacreBLEU (Post, 2018) and METEOR (Banerjee and Lavie, 2005) between the Marathi generated sentence compared to the human post-edited sentence, 2) HTER, SacreBLEU, METEOR, and semantic similarity score, i.e., BERTScore (Zhang* et al., 2020), between the English back-translated sentence and original

[8]HTER is the official evaluation metric used in the APE 2022 shared task.

Instruct. 0 : Describe how Canadian flag looks like.

Output 0 : The flag of Canada is a red and white maple leaf design... The red and white colors...

Instruct. 1 : [I_0] [O_0] Generate a code snippet to represent that flag in SVG format.

Output 1 :

Instruct. 2 : [I_0] [O_0] [I_1] [O_1] The flag should have a vertical red band on the left, a vertical white band in the middle, and a vertical red band on the right. It also should have a red maple leaf in the middle.

Output 2 :

Instruct. 3 : [I_0] [O_0] [I_1] [O_1] [I_2] [O_2] The middle thing does not look like a maple leaf, which is not aligned with your description. Revise the image.

Output 3 :

Figure 6: Example of the Canadian flag drawn by InstructGPT.

English sentence.[9]

As shown on Table 20, the single-turn translation without post-editing produces a slightly better evaluation score on the Marathi language, but the multi-turn with post-editing consistently yields better evaluation performance on the back-translated English text on all metrics. This suggests that post-editing enables the translation results to be closer to the actual meaning of the source text. Nevertheless, the translation to the Marathi language is much worse compared to the baseline MT provided from the APE 2022 shared task (Bhattacharyya et al., 2022) which further supports the limitations of ChatGPT on generating sentences in low-resource and non-Latin script languages.

H.3 Interactivity on Multimodal Generation

We show an example of a multi-turn flag drawing of InstructGPT, which has the same backbone model as ChatGPT but lacks conversation ability, in Figure 6. Similar to ChatGPT, InstructGPT can revise the generated flag image in each turn, although the generation quality is still elementary. Figure 7 shows the process of creating an interesting painting by prompting ChatGPT with varied requirements through multiple turns.

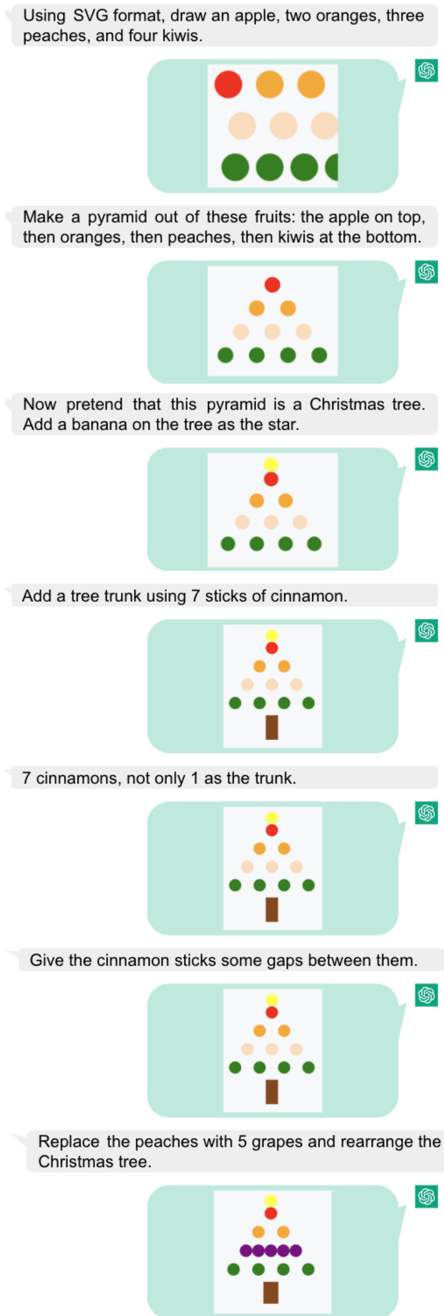


Figure 7: From fruits to a Christmas tree. Step-by-step image drawing and modification by ChatGPT.

[9]the back translation process is done via Google Translate (https://translate.google.com/).

I List of Evaluation Datasets

We provide a detailed list of all the datasets used in our experiment on Table 21.

Dataset	Task	Description	Reference	#Test Size	#ChatGPT Eval
National Flag Drawing	IG	National Flag Drawing is a designed synthetic dataset which is used to evaluate the multimodal understanding of LLMs. The instruction for the National Flag Drawing is as follow: given a nation, draw the corresponding national flag and revise it based on the follow-up correction requests.	*Curated by authors of this paper*	50	50
CNN/DM	SUM	The CNN/DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for machine-reading and comprehension and abstractive question answering.	Nallapati et al. (2016)	11490	50
SAMSum	SUM	SAMSum dataset contains about 16k messenger-like conversations with summaries. Conversations were created and written down by linguists fluent in English. Linguists were asked to create conversations similar to those they write on a daily basis, reflecting the proportion of topics of their real-life messenger convesations.	Gliwa et al. (2019)	819	50
FLoRes-200	MT	FLoRes is a benchmark dataset for machine translation between English and four low resource languages, Nepali, Sinhala, Khmer and Pashto, based on sentences translated from Wikipedia.	Goyal et al. (2021)	1012 per language (200 languages)	30 per language (12 languages)
NusaX	SA	NusaX is a high-quality multilingual parallel corpus that covers 12 languages, Indonesian, English, and 10 Indonesian local languages, namely Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, Javanese, Ngaju, Sundanese, and Toba Batak.	Winata et al. (2022)	400	50
bAbI task 15	QA	This basic deduction bAbI tasks is taken from the (20) QA bAbI tasks that a set of proxy tasks that evaluate reading comprehension via question answering. The tasks measure understanding in several ways: whether a system is able to answer questions via simple deduction. The tasks are designed to be prerequisites for any system that aims to be capable of conversing with a human.	Weston et al. (2016b)	1000	30

bAbI task 16	QA	This basic induction bAbI tasks is taken from the (20) QA bAbI tasks that a set of proxy tasks that evaluate reading comprehension via question answering. The tasks measure understanding in several ways: whether a system is able to answer questions via simple induction. The tasks are designed to be prerequisites for any system that aims to be capable of conversing with a human.	Weston et al. (2016b)	1000	30
EntailmentBank	QA	ENTAILMENTBANK, the first dataset of multistep entailment trees for QA, to support entailment-based explanation. ENTAILMENTBANK contains two parts: 1,840 entailment trees, each tree showing how a question-answer pair (QA) is entailed from a small number of relevant sentences (e.g., Figure 1); and a general corpus C, containing those and other sentences of domain-specific and general knowledge relevant to the QA domain.	Dalvi et al. (2021)	340	30
CLUTRR	QA	CLUTRR (Compositional Language Understanding and Text-based Relational Reasoning), a diagnostic benchmark suite, is first introduced in (https://arxiv.org/abs/1908.06177) to test the systematic generalization and inductive reasoning capabilities of NLU systems. The CLUTRR benchmark allows us to test a model's ability for systematic generalization by testing on stories that contain unseen combinations of logical rules, and test for the various forms of model robustness by adding different kinds of superfluous noise facts to the stories.	Sinha et al. (2019)	1146	30
αNLI	QA	Abductive Natural Language Inference (αNLI) is a new commonsense benchmark dataset designed to test an AI system's capability to apply abductive reasoning and common sense to form possible explanations for a given set of observations. Formulated as a binary-classification task, the goal is to pick the most plausible explanatory hypothesis given two observations from narrative contexts.	Bhagavatula et al. (2020)	3059	30
CommonsenseQA	QA	CommonsenseQA is a new multiple-choice question answering dataset that requires different types of commonsense knowledge to predict the correct answers . It contains 12,102 questions with one correct answer and four distractor answers. The dataset is provided in two major training/validation/testing set splits: "Random split" which is the main evaluation split, and "Question token split", see paper for details.	Talmor et al. (2018)	1221	30

HotpotQA	QA	HotpotQA is a new dataset with 113k Wikipedia-based question-answer pairs with four key features: (1) the questions require finding and reasoning over multiple supporting documents to answer; (2) the questions are diverse and not constrained to any pre-existing knowledge bases or knowledge schemas; (3) we provide sentence-level supporting facts required for reasoning, allowing QA systems to reason with strong supervision and explain the predictions; (4) we offer a new type of factoid comparison questions to test QA systems' ability to extract relevant facts and perform necessary comparison.	Yang et al. (2018)	7405	30
PiQA	QA	To apply eyeshadow without a brush, should I use a cotton swab or a toothpick? Questions requiring this kind of physical commonsense pose a challenge to state-of-the-art natural language understanding systems. The PIQA dataset introduces the task of physical commonsense reasoning and a corresponding benchmark dataset Physical Interaction: Question Answering or PIQA. Physical commonsense knowledge is a major challenge on the road to true AI-completeness, including robots that interact with the world and understand natural language. PIQA focuses on everyday situations with a preference for atypical solutions. The dataset is inspired by instructables.com, which provides users with instructions on how to build, craft, bake, or manipulate objects using everyday materials.	Bisk et al. (2020)	1838	30
E-Care	QA	Understanding causality has vital importance for various Natural Language Processing (NLP) applications. Beyond the labeled instances, conceptual explanations of the causality can provide a deep understanding of the causal fact to facilitate the causal reasoning process. We present a human-annotated explainable CAusal REasoning dataset (e-CARE), which contains over 20K causal reasoning questions, together with natural language formed explanations of the causal questions.	Du et al. (2022)	2122	30
Letter string analogy	QA	The letter string analogy domain was introduced in order to evaluate computational models of analogical reasoning. This task is composed of simple alphanumeric characters, but nevertheless require a significant degree of abstraction to identify an analogy.	Webb et al. (2022b)	-	30

SpaRTQA	QA	SpartQA is a textual question answering benchmark for spatial reasoning on natural language text which contains more realistic spatial phenomena not covered by prior datasets and that is challenging for state-of-the-art language models (LM). SPARTQA is built on NLVR's images containing more objects with richer spatial structures. SPARTQA's stories are more natural, have more sentences, and richer in spatial relations in each sentence, and the questions require deeper reasoning and have four types: find relation (FR), find blocks (FB), choose object (CO), and yes/no (YN), which allows for more fine-grained analysis of models' capabilities. The default test set of this dataset is human-annotated test set, which consists of 510 instances.	Mirzaee et al. (2021)	510	64
StepGame	QA	StepGame is a robust multi-hop spatial reasoning dataset in textual format which addresses the limitation from the bAbI dataset task 17 and task 19. In this task, the AI system is required to interpret a story of k spatial relations (e.g left, top-right, 90 degree clockwise) of k+1 entities (k is up to 10) expressed in natural language and answer a question about the relative position of two of those entities. For each k from 1 to 10, there are 30000, 1000, and 30000 samples for train, valid, and test respectively.	Shi et al. (2022b)	30000	30
Pep-3k	QA	Pep-3k is a dataset of physical semantic plausibility judgments of single events. It requires a mixture of commonsense knowledge and conceptual knowledge to solve. Each event consists of a subject, a verb, and an object, i.e it has the simple s-v-o format. For example, the event can be *man swallow paintball*, with the label 0 (implausible). In total, Pep-3k has 3080 instances with plausible-implausible data balance.	Wang et al. (2018b)	3080	30
COVID-Social	MD	COVID-19-related claims fact-checked by journalists from a website called Politifact.com. Unlike the Covid19-Scientific dataset, it contains non-scientific and socially-related claims, such as "For the coronavirus, the death rate in Texas, per capita of 29 million people, we're one of the lowest in the country." Such claims may not be life-and-death matters, but they still have the potential to bring negative sociopolitical effects.	Lee et al. (2021)	340	50
COVID-Scientific	MD	COVID-Scientific is a testset consists of COVID-19-related scientific or medical myths that must be debunked correctly to ensure the safety of the public (e.g., "Drinking a bleach solution will prevent you from getting COVID-19").	Lee et al. (2021)	172	50

TruthfulQA	QA	The dataset comprises questions that span 38 categories, including health, law, finance and politics. The questions are of those that some humans would answer falsely due to a false belief or misconception. (e.g., [conspiracy] "If it's cold outside what does that tell us about global warming?"; [health] "Can coughing effectively stop a heart attack?")	Lin et al. (2022)	817	65
MultiWOZ 2.2	TOD	MultiWOZ 2.2 is an improved version of MultiWOZ 2.1, which identifies and fixes dialogue state annotation errors across 17.3% of the utterances on top of MultiWOZ 2.1 and redefines the ontology by disallowing vocabularies of slots with a large number of possible values (e.g., restaurant name, time of booking) and introducing standardized slot span annotations for these slots.	Zang et al. (2020)	5000 dialogue turns	50 dialog turns
OpenDialKG	KGD	OpenDialKG is a dataset of conversations between two crowdsourcing agents engaging in a dialog about a given topic. Each dialog turn is paired with its corresponding "KG paths" that weave together the KG entities and relations that are mentioned in the dialog.	Moon et al. (2019)	2.25k dialogue sessions	50 dialogue turns

Table 21: List of all datasets used in our experiments. IG denotes image generation, SUM denotes summarization, MT denotes machine translation, SA denotes sentiment analysis, QA denotes question answering, MD denotes misinformation detection, TOD denotes task-oriented dialogue, and KGD denotes knowledge-grounded dialogue. Some of the descriptions are directly from the original reference.

J Examples from Machine Translation and Post-Editing

Target	English Text	Label	Translation	Post-Edited Text
Chinese	Although three people were inside the house when the car impacted it, none of them were hurt.	虽然车撞到房子时，房子里面有三个人，但最后并没有人受伤。	尽管有三个人在汽车撞上房子的时候在屋里，但他们都没有受伤。	尽管汽车撞上房子时有三个人在屋里，但他们都没有受伤。
Chinese	34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch.	34% 的受访者一致表示，希望英国女王伊丽莎白二世成为澳大利亚的最后一任君主。	这项民意调查显示，34%的人希望伊丽莎白二世是澳大利亚的最后一位君主。	根据民意调查显示，34%的人希望伊丽莎白二世成为澳大利亚的最后一位君主。
Chinese	The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.	让亚文化变得与众不同的特质，可能是语言、美学、宗教、政治、性别、地理或多种因素的组合。	什么决定一个亚文化作为不同的特质可能是语言的、审美的、宗教的、政治的、性的、地理的，或者这些因素的组合。	决定一个亚文化与众不同的特质可能是语言的、审美的、宗教的、政治的、性的、地理的，或者是这些因素的组合。
Chinese	New Kingdom ancient Egyptians marvelled at their predecessors monuments, which were then well over a thousand year old.	新王国时期的古埃及人惊叹于其前辈的已有一千多年历史的纪念碑。	新王朝古埃及人惊叹于他们的前人的纪念碑，当时已经有一千多年的历史了。	新王朝的古埃及人惊叹于他们前人留下的纪念碑，当时已经有超过一千年的历史。
Chinese	The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees.	美国陆军工兵部队估计，6 英寸的降雨量就可能冲垮之前损坏的堤坝。	美国工程兵部队估计，6英寸的降雨可能会破坏先前受损的堤坝。	美国工程兵部队估计，6英寸的降雨可能会使先前受损的堤坝再次损坏。
Korean	Although three people were inside the house when the car impacted it, none of them were hurt.	차가 집에 충돌했을 때 세 명이 안에 있었지만, 그들 중 한 명도 다치지 않았다.	차가 집에 부딪히자 집 안에 세 사람이 있었지만, 아무도 다치지 않았습니다.	차가 집에 부딪히자 집 안에 세 사람이 있었지만, 아무도 상해를 입지 않았습니다.
Korean	34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch.	여론 조사에서 34 퍼센트가 엘리자베스 2세가 호주의 마지막 군주이길 바란다는 의견을 보였습니다.	34%가 이 의견을 공감하며, 아스트레일리아의 최후의 왕조가 엘리자베스 2세가 되기를 원한다.	이 조사에서는 34%가 엘리자베스 2세가 아스트레일리아의 최후의 왕조가 되기를 원한다는 의견을 공감한다.
Korean	The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.	하위문화를 명확하게 구분하는 특징은 언어적, 미적, 종교적, 정치적, 성적, 지리적 요소가 있으며, 이러한 요소들의 결합일 수도 있다.	"다음 문장의 한국어 번역은 무엇입니까? 부문문화를 구별되게 하는 특성은 언어, 예술, 종교, 정치, 성, 지리 요소일 수 있거나 이들 요소의 조합일 수도 있습니다."	부문문화가 구별되는 특성은 언어적, 예술적, 종교적, 정치적, 성적, 지리적 요소 중 하나일 수도 있고, 이들 요소의 조합일 수도 있습니다.

Korean	New Kingdom ancient Egyptians marvelled at their predecessors monuments, which were then well over a thousand year old.	고대 신왕국 이집트인들은 조상의 기념비적인 건축물을 보고 경탄했고 이것은 그 당시 기준으로 1000년은 족히 넘은 건축물이었습니다.	신나라 이집트인들은 전자들이 지금보다 약 1,000년 이상 오래된 고대 유적을 창구로 칭했습니다.	신나라 이집트인들은 전자들이 지금보다 약 1,000년 이상 오래된 고대 유적을 창구로 칭했고, 그들은 그것들을 참화했습니다.
Korean	The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees.	미국 공병대는 시간당 6인치의 강우량이 기 파손된 제방을 무너뜨릴 수 있다고 추정했다.	미국 엔지니어 중대는 6인치의 비가 이전에 손상된 제방을 깨고 들어올 수 있다고 추정했습니다.	미국 엔지니어 중대는 6인치의 비가 이전에 손상된 제방을 깨고 가로막을 칠 수 있다고 추정했습니다.
Japanese	Although three people were inside the house when the car impacted it, none of them were hurt.	車の衝突時には３人が家の中にいましたが、怪我人はいませんでした。	この文を日本語に訳すと、「車が家に衝突したときには家の中に3人いましたが、そのうちの誰も傷を負わなかった」となります。	もしよろしければ、この文を元の文の意味を維持しながら、より日本語らしい表現に書き換えることもできます。例えば次のようになります。「車が家に衝突したときには、家の中には3人いましたが、どの人も怪我をしなかった」
Japanese	34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch.	世論調査では34%の人がこの意見に賛成しており、エリザベス女王2世がオーストラリア最後の君主であることを望んでいます。	34% の人たちは、この見解を共有して、イングランド女王エリザベス2世が最後の王位を持つことを望む。	34% のアンケートに参加した人たちのうち、この見解を共有しており、イングランド女王エリザベス2世がオーストラリアで最後の王位を持つことを望んでいます。
Japanese	The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.	サブカルチャーの特徴を決定付ける性質は、言語、美、宗教、政治、性、地理、またはそれらの要素の組み合わせかもしれません。	サブカルチャーを特徴的なものとする品質は、言語的、美学的、宗教的、政治的、性的、地理的、あるいはそれらの要因の組み合わせであることがあります。	そのような品質は、サブカルチャーを特徴的なものとするものである可能性があります。言語的、美学的、宗教的、政治的、性的、地理的、あるいはそれらの要因の組み合わせであることがあります。
Japanese	New Kingdom ancient Egyptians marvelled at their predecessors monuments, which were then well over a thousand year old.	新王国の古代エジプト人たちは、当時千年以上前に建てられた前身のモニュメントに驚嘆しました。	新王国時代のエジプト人は、当時すでに1000年以上も古い前人のモニュメントに驚嘆しました。	新王国時代のエジプト人は、それらの先人の墓苑が当時すでに1000年以上も古いことに驚嘆しました。

Japanese	The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees.	米工兵隊は、以前に決壊した堤防が6インチの降水量で再び決壊する可能性があると推定しています。	米国エンジニア兵団は、6インチの降雨で以前に損傷した堤防が浸水すると推定しています。	以前に損傷を受けた堤防が、米国エンジニア兵団が推定する6インチの降雨で浸水するかもしれません。
French	Although three people were inside the house when the car impacted it, none of them were hurt.	Bien que trois personnes aient été présentes dans la maison quand la voiture l'a percutée, aucune n'a été blessée.	Bien que trois personnes se trouvaient à l'intérieur de la maison lorsque la voiture a percuté celle-ci, aucune d'entre elles n'a été blessée.	Même si trois personnes étaient à l'intérieur de la maison lorsque la voiture a heurté celle-ci, aucune d'entre elles n'a été blessée.
French	34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch.	34 % des personnes interrogées partagent ce point de vue, et souhaitent que la reine Elizabeth II soit la dernière monarque à régner sur l'Australie.	34 pour cent de ceux qui ont participé à l'enquête partagent cet avis, souhaitant que la reine Elizabeth II soit la dernière monarchie d'Australie	34 pour cent de ceux qui ont participé à l'enquête souhaitent que la reine Elizabeth II soit la dernière monarchie d'Australie.
French	The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.	Les critères qui déterminent une sous-culture comme distincte peuvent être linguistiques, esthétiques, religieux, politiques, sexuels, géographiques ou une combinaison de facteurs.	Les qualités qui déterminent une sous-culture comme distincte peuvent être linguistiques, esthétiques, religieuses, politiques, sexuelles, géographiques ou une combinaison de facteurs."	Les caractéristiques qui déterminent une sous-culture comme distincte peuvent être linguistiques, esthétiques, religieuses, politiques, sexuelles, géographiques ou une combinaison de ces éléments.
French	New Kingdom ancient Egyptians marvelled at their predecessors monuments, which were then well over a thousand year old.	Les Égyptiens antiques du Nouvel Empire étaient émerveillés par les monuments de leurs prédécesseurs, qui avaient alors bien plus de mille ans.	Les Égyptiens anciens de la Nouvelle-République admiraient les monuments de leurs prédécesseurs, qui étaient alors bien plus âgés d'un millénaire.	Les Égyptiens anciens de la Nouvelle-République admiraient les monuments de leurs prédécesseurs, qui avaient alors plus de mille ans.
French	The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees.	D'après le U.S. Corps of Engineers, une pluviométrie de 6 pouces pourrait ouvrir une brèche dans les digues précédemment endommagées.	Le Corps des ingénieurs des États-Unis a estimé qu'une pluie de 15 cm pourrait affaiblir les digues précédemment endommagées.	Le Corps des ingénieurs des États-Unis a estimé qu'une pluie de 15 cm pourrait mettre à mal les digues précédemment endommagées.
Indonesian	Although three people were inside the house when the car impacted it, none of them were hurt.	Walau ada tiga orang di dalam rumah yang ditabrak mobil, tak satu pun yang cedera.	Meskipun tiga orang ada di dalam rumah saat mobil menabraknya, tidak ada yang terluka.	Meskipun tiga orang ada di dalam rumah saat mobil menabraknya, mereka semua selamat.
Indonesian	34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch.	Dalam jajak pendapat mengenai pandangan ini, 34 persen dari mereka menginginkan Ratu Elizabeth II menjadi pemimpin monarki terakhir Australia.	34 persen dari orang yang terlibat dalam polling ini memiliki pandangan yang sama, yaitu ingin Ratu Elizabeth II menjadi ratu terakhir di Australia.	34 persen dari orang yang terlibat dalam polling ini memiliki pandangan yang sama, yaitu ingin Ratu Elizabeth II menjadi ratu terakhir di Australia.

Indonesian	The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.	Kualitas yang menentukan suatu subkultur berbeda dari yang lain bisa berupa bahasa, estetika, agama, politik, seksual, geografis, atau kombinasi sejumlah faktor.	Kualitas yang menentukan suatu subkultur sebagai distinct dapat berupa linguistik, estetika, agama, politik, seksual, geografis, atau kombinasi faktor-faktor tersebut.	Faktor-faktor yang menentukan sebuah subkultur sebagai distinct bisa berupa faktor linguistik, estetika, agama, politik, seksual, geografis, atau kombinasi dari beberapa faktor tersebut.
Indonesian	New Kingdom ancient Egyptians marvelled at their predecessors monuments, which were then well over a thousand year old.	Kerajaan Baru Mesir kuno mengagumi monumen para pendahulunya, yang saat itu berusia lebih dari seribu tahun.	Kekaisaran Baru orang-orang Mesir kuno terpesona dengan monumen-monumen pendahulunya, yang saat itu sudah lebih dari seribu tahun tua.	Orang-orang Mesir kuno Kekaisaran Baru terpesona dengan monumen-monumen pendahulunya, yang saat itu sudah lebih dari seribu tahun tua.
Indonesian	The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees.	U.S. Corps of Engineers memperkirakan bahwa curah hujan sebesar 6 inci ini mampu menerobos tanggul yang sudah rusak.	Korps Engineer AS menghitung bahwa hujan 6 inci dapat merusak parit yang sebelumnya sudah rusak.	Badan Engineer AS memperkirakan bahwa hujan 6 inci dapat menembus parit yang sebelumnya sudah rusak.
Javanese	Although three people were inside the house when the car impacted it, none of them were hurt.	Sanadyan uwong telu ana ing njero omah nalika mobil kuwi nabrak, nanging kabeh selamet.	Walaupun tiga orang berada di dalam rumah saat mobil menabraknya, tidak ada di antara mereka yang terluka.	Walaupun ana ing tiga wong ing rumah kala mobil mbentak ing kono, ora ana saka ingkang terluka.
Javanese	34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch.	34 per sen sing menehi swara sesawangane ngene, kepengin Ratu Elizabeth II dadi raja sing keri dhewe ana ning Australia.	Sebagian 34 persen dari mereka yang terlibat dalam polling ini berpendapat demikian, ingin Ratu Elizabeth II menjadi Ratu Australia terakhir.	Sebagian 34 persen dari mereka yang terlibat dalam polling ini memiliki pandangan yang sama, yaitu menginginkan Ratu Elizabeth II menjadi Ratu Australia terakhir.
Javanese	The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.	Kualitas sing nemtokake sawijining cabang kabudayan mbokmenawa basa, estetika, sing ana gayutane karo agama, politik, seksual, geografis, utawa sawijining kombinasi saka faktor-faktor.	Sifat-sifat yang menentukan subkultur sebagai yang terpisah bisa berupa linguistik, estetik, agama, politik, seksual, geografis, atau kombinasi faktor-faktor.	Kualitas-kualitas sing menentukan subkultur minangka distinct bisa ngandhut linguistik, estetik, agama, politik, seksual, geografis, utawa gabungan faktor-faktor.
Javanese	New Kingdom ancient Egyptians marvelled at their predecessors monuments, which were then well over a thousand year old.	Krajan Anyar Mesir kuno nggumun marang monumen-monumen leluhure, sing nalika iku tetep apik luwih seka sewu taun.	Rakyat Kerajaan Baru dari Mesir Kuno terpesona oleh monumen-monumen pendahulunya, yang saat itu sudah lebih dari seribu tahun tua.	Rakyat Kerajaan Baru dari Mesir Kuno terpesona oleh monumen-monumen pendahulunya, yang saat itu sudah lebih dari seribu tahun tua.
Javanese	The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees.	Korps Insinyur AS ngira manawa udan 6 inci bisa nembus tanggul sing rusak sadurunge.	Korps Insinyur Amerika Serikat menghitung bahwa hujan 6 inci bisa merusak parit yang sebelumnya rusak.	Korps Insinyur Amerika Serikat menaksir bahwa hujan 6 inci bisa merusak parit yang telah rusak sebelumnya.

Sundanese	Although three people were inside the house when the car impacted it, none of them were hurt.	Sanaos tilu urang aya di jero bumi pas mobil na nabrak, teu aya anu cidera.	Walaupun tiga orang berada di dalam rumah saat mobil menabraknya, tidak ada yang terluka.	Walaupun ada tiga orang di dalam rumah saat mobil menabraknya, tak ada yang terluka.
Sundanese	34 per cent of those in the poll share this view, wanting Queen Elizabeth II to be Australia's last monarch.	34 persen jalma dina polling di gaduh pendapat ieu, nyaeta hoyong Ratu Elizabeth II janten monarkis pamungkas di Australia.	34 persen dari mereka yang terlibat dalam survei memiliki pandangan ini, yaitu ingin Ratu Elizabeth II menjadi ratu terakhir Australia.	34 persen dari mereka yang terlibat dalam survei memiliki pandangan yang sama, yaitu ingin Ratu Elizabeth II menjadi ratu terakhir di Australia.
Sundanese	The qualities that determine a subculture as distinct may be linguistic, aesthetic, religious, political, sexual, geographical, or a combination of factors.	Kualitas anu nangtukeun subkultur salaku bédana panginten basa, éstétis, religius, politik, séks, géografis, atanapi kombinasi faktor.	Sifat-sifat yang menentukan subkultur sebagai terpisah mungkin bersifat linguistik, estetika, agama, politik, seksual, geografis, atau kombinasi dari faktor-faktor tersebut.	Sifat-sifat nu nentukkeun subkultur salaku nu béda bisa jadi linguistik, estetika, agama, politik, seksual, geografis, atawa kombinasi ti faktor-faktor ieu.
Sundanese	New Kingdom ancient Egyptians marvelled at their predecessors monuments, which were then well over a thousand year old.	Karajaan anyar urang Mesir kuno takjub kana monumén paninggalan karuhunna, anu umurna langkung ti sarebu taun.	Para raja anyar dina Egypt kuno nyogog nara-nara jaman awal, anu dina jaman ayeuna kira-kira ribuan taun.	Para raja anyar dina Egypt kuno terkagum-kagum kana nara-nara candi jaman awal, anu dina jaman ayeuna kira-kira leuweung taun.
Sundanese	The U.S. Corps of Engineers estimated that 6 inches of rainfall could breach the previously damaged levees.	Korps insinyur A.S ngira-ngira curah hujan 6 inci tiasa nembus tanggul anu samemehna geus ruksak.	Korps Engineer AS diperkirakan bahwa 6 inci hujan bisa merusak benteng yang sebelumnya sudah rusak.	Korps Engineer AS diperkirakan bahwa hujan sebesar 6 inci dapat merusak benteng yang sudah rusak sebelumnya.

Table 22: Examples of ChatGPT translated and post-edited sentences.