GPTEval: A Survey on Assessments of ChatGPT and GPT-4

Rui Mao♠, Guanyi Chen♣, Xulang Zhang♠, Frank Guerin♦, and Erik Cambria♠
♠Nanyang Technological University, Singapore
♣Central China Normal University, China
♦University of Surrey, United Kingdom
rui.mao@ntu.edu.sg; g.chen@ccnu.edu.cn; xulang001@e.ntu.edu.sg;
f.guerin@surrey.ac.uk; cambria@ntu.edu.sg

Abstract

The emergence of ChatGPT has generated much speculation in the press about its potential to disrupt social and economic systems. Its astonishing language ability has aroused strong curiosity among scholars about its performance in different domains. There have been many studies evaluating the ability of ChatGPT and GPT-4 in different tasks and disciplines. However, a comprehensive review summarizing the collective assessment findings is lacking. The objective of this survey is to thoroughly analyze prior assessments of ChatGPT and GPT-4, focusing on its language and reasoning abilities, scientific knowledge, and ethical considerations. Furthermore, an examination of the existing evaluation methods is conducted, offering several recommendations for future research.

Keywords: LLM Evaluation, ChatGPT, GPT4

1. Introduction

ChatGPT (OpenAI, 2023b) has generated significant scholarly interest across various disciplines due to its impressive dialogue-based task-processing capabilities. This has enabled users to explore and evaluate its performance across a wide range of tasks and disciplines, thereby sparking considerable enthusiasm in the field of Artificial Intelligence (AI). While many researchers have concentrated on evaluating ChatGPT and GPT-4 (OpenAI, 2023a) within their specific domains of expertise, a comprehensive review encompassing the assessments in multiple tasks and disciplines can offer a holistic understanding of the strengths and limitations of these GPT models. We focus on ChatGPT and GPT-4, because they are state-of-the-art (SOTA) large language models (LLMs). The scope of our survey encompasses quantitative evaluations carried out on ChatGPT or GPT-4, specifically focusing on their language proficiency, scientific knowledge, and ethical considerations. Our main findings are summarized as follows:

a) ChatGPT and GPT-4 are strong in language understanding and generation, adeptly engaging in user interactions through dialogues, enabling them to tackle diverse NLP tasks and provide explanatory outputs. However, their current status falls short of being a comprehensive AI, as their performance lags behind expert models in numerous domains involving domain-specific knowledge.

b) ChatGPT performs satisfactorily in general science knowledge and can answer science questions that desire open responses. However, it can also make mistakes, especially for questions that require multi-step reasoning. The exceptional language proficiency poses challenges for users in assessing the accuracy of factual information, giving rise to a range of ethical concerns.

c) Existing evaluation methods may be unreliable. The current evaluation methods heavily depend on prompt engineering and benchmark datasets. Varying prompts can yield disparate evaluation results. Additionally, the comparison of expert systems often relies on (in-domain) datasets that were utilized for training those systems. It remains uncertain if the examined data, such as public datasets and scientific knowledge, have been inadvertently exposed during the training of ChatGPT and GPT-4. These factors may contribute to an unfair comparison between LLMs and their respective baselines.

The contributions of this work are threefold: (1) We conduct a comprehensive survey of recent assessments focusing on the language proficiency and scientific knowledge of ChatGPT and GPT-4. (2) We compare their assessment results across various tasks and disciplines to highlight the strengths and weaknesses of the GPT models. (3) We critically analyze the existing assessment methods employed, offering recommendations for future evaluation studies and delivering our ethical considerations associated with the GPT models.

2. Language and Reasoning Ability

2.1. Classic NLP Tasks

Dialogue. Cabrera and Neubig (2023) compared several chatbots with a novel LLM evaluation toolkit, termed Zeno Build. The performance was evaluated by Critique metrics[1] such as ChrF (a character-

[1] `https://docs.inspiredco.ai/critique/`

LREC-COLING 2024, pages 7844–7866

and word n-grams similarity-based metric, Popović, 2015), BERTScore (a BERT embedding similarity-based metric, Zhang et al., 2019), and UniEval Coherence (a coherence probability-based metric, Zhong et al., 2022). They found that ChatGPT surpassed all the baselines, e.g., GPT-2 (Radford et al., 2019), LLaMa (Touvron et al., 2023), Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), MPT-Chat (MosaicML, 2023), and Cohere Command[2] in the three evaluation metrics. However, Cabrera and Neubig (2023) also highlighted that ChatGPT exhibited vulnerabilities that were noticeable in various aspects, such as occurrence of hallucinations, inadequate exploration for additional information, and repetition of content. Although ChatGPT exceeded other LLMs, Bang et al. (2023) observed that SOTA models still outperformed ChatGPT on task-oriented dialogue, and open-domain knowledge-grounded dialogue, based on automatic evaluation metrics.

Generation was almost evaluated on text-to-text generation tasks. For machine translation (MT), an early assessment suggested SOTA MT systems could defeat ChatGPT by a large margin (Bang et al., 2023). Later on, by testing on more language pairs, more datasets, and better prompts, Hendy et al. (2023); Jiao et al. (2023) found ChatGPT yielded competitive performance for high-resource languages, but still had limited capabilities for low-recourse languages. Through a large-scale human evaluation and error analysis by expert translators, Karpinska and Iyyer (2023) found that, when doing paragraph-level translation, ChatGPT's translations were overwhelmingly preferred compared to those from Google Translate and it largely reduced errors, including mistranslation, grammatical errors, inconsistency errors, and more.

For summarization, by testing on multiple summarization datasets in multiple languages, Qin et al. (2023); Bang et al. (2023); Lai et al. (2023) observed that ChatGPT largely underperformed SOTA systems in doing either abstractive or extractive summarization. Zhang et al. (2023a) found that its performance, especially the faithfulness, improved if we asked ChatGPT to first extract salient sentences and then generate the summarization, based on the extracted sentences, although it still lost to SOTA. However, similar to what happened when assessing ChatGPT's translation ability, This conclusion should be further confirmed by larger populations and rigorous experiments. Later on, a small human evaluation found that annotators could not distinguish summaries generated by ChatGPT from those by humans (Soni and Wade, 2023), which, however, was then falsified by a much larger manual assessment study (Pu et al., 2023).

A few assessments on data-to-text generation suggested that both ChatGPT and GPT-4 underperformed to fine-tuned T5 and BART (Ren and Liu, 2023; Yuan and F"arber, 2023), which was probably due to the complexities in expressing relations between data and text in prompts. Although ChatGPT is often considered to perform well in generation tasks that need creativity, there have not been many related assessments. Jentzsch and Kersting (2023) asked ChatGPT to produce jokes and suggested that while ChatGPT generated jokes, it struggled to produce "new" jokes. Approximately 90% of the generated jokes were repetitions of the same 25 jokes. Chu and Liu (2023) conducted reader experiments and found that ChatGPT wrote more engaging and persuasive short stories than its human counterparts. However, the conclusion was the opposite if the aim was long stories. They suggested that ChatGPT might focus on retaining information from the instruction when writing a long story, which limited its flexibility.

Affective Computing. Amin et al. (2023) compared ChatGPT and GPT-4 to supervised learning models on 13 affective computing domains. The baselines were trained with task-specific datasets using different embedding representations. They found that the RoBERTa embedding-based model exceeded ChatGPT and GPT-4 on 22 out of 34 evaluation tasks. Qin et al. (2023); Bang et al. (2023); Kocoń et al. (2023) also found that fine-tuned models exceeded ChatGPT on English sentiment analysis and emotion detection tasks.

Information Retrieval. Wei et al. (2023) examined ChatGPT on relation extraction, named entity recognition (NER), and event extraction tasks, showing that the performance of the basic version of ChatGPT was much weaker than supervised methods, while largely exceeding 50 shot or fewer shot baselines. Then, they introduced a multi-turn question answering (QA) framework, where the modified querying process did not help ChatGPT to exceed the supervised baseline. Qin et al. (2023) observed that ChatGPT yielded much weaker performance than fine-tuning-based models on CoNLL03 NER dataset (Sang and De Meulder, 2003). Sun et al. (2023) evaluated ChatGPT and GPT-4 on passage re-ranking tasks, showing that GPT-4 outperformed SOTA supervised baselines across all three datasets, including a multilingual dataset with 10 languages. ChatGPT slightly fell behind SOTA methods, while largely exceeding BM25. Bubeck et al. (2023) found that GPT-4 (77.4% accuracy) significantly outperformed the SOTA model (40.8%, Payne, 2020) on a personally identifiable information detection task.

GPT as Human Annotator. Wang et al. (2023c) compared ChatGPT with SOTA natural language generation evaluation metrics, e.g., BERTScore,

[2] https://docs.cohere.com/docs/command-beta

BARTScore (Yuan et al., 2021), ROUGE (Lin, 2004), and more. ChatGPT slightly outperformed the strongest BARTScore-based setup on the SummEval dataset (Fabbri et al., 2021) in coherence, relevance, consistency, and fluency dimensions, while was surpassed by BARTScore on the NewsRoom dataset (Grusky et al., 2018). ROUGE-1 yielded the highest correlation scores on the sample- and dataset-level evaluations on the RealSumm dataset (Bhandari et al., 2020). The above three datasets were used for evaluating text summarization, although presenting inconsistent results. ChatGPT achieved the largest improvements on the OpenMEVA-ROC (Guan et al., 2021) story generation dataset. ChatGPT and baselines were comparable on the BAGEL dataset (Mairesse et al., 2010) for informativeness, naturalness, and quality evaluations. Liu et al. (2023c) found that GPT-4 could exceed other metrics on the dialogue generation dataset from Mehri and Eskenazi (2020). Kocmi and Federmann (2023) evaluated ChatGPT-generated machine translation accuracy scores with its former versions and other SOTA scoring systems, finding that ChatGPT was less accurate than those baselines. Gilardi et al. (2023) observed that ChatGPT achieved higher intercoder agreements than MTurk crowd-workers and educated annotators, and maintained the highest annotation accuracy in tweet frame and stance annotation tasks. The relevance annotation accuracy of ChatGPT was comparable to humans, while topic annotations were not as useful as MTurk.

2.2. Multilingualism

Chinese Linguistic Test. SuperCLUE benchmarks (Xu et al., 2023) compared ChatGPT-like foundation models in regard to their basic language ability, professional ability, and Chinese-featured ability. By 18th May 2023, they reported that GPT-4 and ChatGPT achieved the second (76.67) and third best (66.18) results after humans (96.50), exceeding other LLMs. GPT-4 and ChatGPT had better basic language ability than the other two metrics in Chinese. Both models achieved human-like accuracy on role-playing, chit-chatting, and coding. However, their ability to understand Chinese poetry, literature, classical Chinese, and couplets was far inferior to that of humans. Huang et al. (2023) also ranked GPT-4 and ChatGPT as top-2 on a multi-discipline (52 subjects) Chinese evaluation.

Multilingual NLP Tasks were examined by Lai et al. (2023), e.g., multilingual part-of-speech (PoS) tagging, NER, relation classification, NLI, QA, commonsense reasoning, and summarization. The researchers analyzed 36 languages and discovered that task-specific fine-tuned models outperformed ChatGPT in the majority of examined tasks, except PoS tagging. ChatGPT exhibited superior performance in English tasks compared to tasks in other languages; for low- and extremely low-resource languages, ChatGPT performed significantly worse than baselines. Noticeably, despite the use of non-English languages in the target tasks, ChatGPT improved its performance with English prompts. Wei et al. (2023) found that direct usage of ChatGPT yielded unsatisfying results in Chinese information extraction. Wang et al. (2023b) tested ChatGPT and GPT-4 on English-to-Chinese and English-to-German summarization, showing that although ChatGPT and GPT-4 exceeded other LLM baselines on a zero-shot setup, they fell behind a fine-tuned mBART-50 (Tang et al., 2021) on most of the examined datasets. Bang et al. (2023) argued that ChatGPT generally yielded weak performance on low-resource languages in language understanding and generation, while achieving higher proficiency in comprehending non-Latin scripts compared to its proficiency in generating them.

2.3. Reasoning

Logical Reasoning was tested by Bang et al. (2023). They found 56 out of 60 answers correct (with appropriate prompts) for deductive reasoning, i.e. applying general rules to specific situations or cases. This was stronger than other types of reasoning. 26 out of 30 were scored for abductive reasoning, i.e. forming plausible explanations or hypotheses, based on limited evidence or incomplete information. 33 out of 60 were scored for inductive reasoning, i.e., drawing generalized conclusions from examples or specific observations.

Commonsense Reasoning. Bang et al. (2023) tested ChatGPT via three commonsense datasets, showing that 80 out of 90 of ChatGPT's predictions were correct. ChatGPT was able to give good explanations of the reasoning steps to support its answer. However, Qin et al. (2023); Laskar et al. (2023) showed that the commonsense reasoning accuracy of ChatGPT fell behind fine-tuned baselines. Davis (2023) found significant flaws in common benchmarks for common sense, including the CommonsenseQA dataset used by Bang et al. (2023), which he explicitly addressed. Davis (2023) listed several examples of commonsense and particularly physical reasoning failures that had been found shortly after the release of ChatGPT, and pointed to others. However, there does not exist a thorough assessment of the GPT models' commonsense reasoning ability. More generally, Davis (2023) pointed out that "many important aspects of commonsense reasoning and commonsense knowledge are not tested in existing benchmark". Bubeck et al. (2023) probed a small number of their own real-world physical reasoning tasks with GPT-4, finding that it had good knowledge and concluded it was able to learn an understanding of the real-world environment.

Causal Reasoning. Bang et al. (2023) found that 24 out of 30 causes or effects could be correctly identified. Gao et al. (2023) systematically evaluated event causality identification, causal discovery, and causal explanation generation. Compared to SOTA models, ChatGPT and GPT-4 yielded lower scores in causality identification. They outperformed baseline models on the causal discovery, although the compared models, e.g., BERT- (Devlin et al., 2019) and RoBERTa-base (Liu et al., 2019) were relatively weak. The generation of causal explanations yielded inconsistent findings in terms of AVG-BLEU and ROUGE-l metrics, while the human evaluation affirmed that both GPT models attained a level of accuracy comparable to that of human performance. Kıcıman et al. (2023) examined ChatGPT and GPT-4 on the causal discovery, counterfactual reasoning, and actual causality inferring, finding that they outperformed other LLMs and SOTA models largely on the first two tasks.

Psychological Reasoning is the ability of humans to reason about other's unobservable mental states (a.k.a Theory of Mind (ToM)). Kosinski (2023) and Moghaddam and Honey (2023) designed sets of False-Belief questions and quantified results suggested that both ChatGPT and GPT-4 had ToM ability, but that was still inferior to a human's. However, Marcus and Davis (2023) pointed out flaws in the Kosinski (2023) study because the test material was in the training data. Holterman and van Deemter (2023) tested GPT-3 and GPT-4 on more ToM tasks summarized in Kahneman (2000). They acknowledged the potential problem of the test material being in the training data, so they substituted various nouns in the scenario. However, this is unlikely to be adequate to stop a neural model from generalising from those examples. Borji (2023) found that chatGPT failed on a variant of a classic 'Sally-Anne Test' (used to test children). Bubeck et al. (2023) found that chatGPT answered correctly on a similar variant Sally-Anne, and they further tested on a range of more advanced ToM scenarios, with probing questions, e.g. to infer the counterfactual impact of actions on mental states. They found that GPT-4 had superior abilities and suggested that GPT-4 had a very advanced level of ToM.

Task-Oriented Reasoning. Qin et al. (2023) evaluated dialogue, logical reasoning, complex yes/no QA, symbolic reasoning (last letter concatenation and coin flip), date understanding-, and tracking shuffled objects-oriented logical reasoning. However, ChatGPT underperformed fine-tuned baselines on most of the tasks, excluding the logical reasoning tasks. To ascertain whether ChatGPT relies on profound comprehension of truth and logic in their reasoning or merely exploits shallow memorized patterns, Wang et al. (2023a) proposed a dialectical evaluation task, finding that despite displaying high confidence, ChatGPT demonstrated an inability to hold its belief in the truth in a wide range of reasoning tasks, e.g., mathematics, first-order logic, commonsense, and generic reasoning.

Natural Language Inference (NLI) aims to examine if a statement can be inferred, contradicted, or neutral, compared to another statement. Liu et al. (2023b) compared ChatGPT and GPT-4 with RoBERTa. However, both models encountered difficulties when dealing with novel and out-of-distribution data. They yielded relatively modest performance on NLI that needed logical reasoning. Qin et al. (2023) also proved that the NLI ability of ChatGPT was lower than that of supervised models. Ambiguity is one of the difficulties of NLI. For example, whether "*John and Anna are not a couple*" contradicts "*John and Anna are married*" depends on whether "*married*" means "*both married*" or "*married to each other*". Given an ambiguous NLI premise, Liu et al. (2023a) asked ChatGPT and GPT-4 to either generate disambiguations of a premise with respect to the hypothesis or recognize disambiguation (i.e., deciding whether the disambiguation is an interpretation of an ambiguous premise). Their human evaluation showed that for the first task, GPT-4 achieved correctness at 32%, while, for the second task it was at the level of random guessing, suggesting that resolving tricky ambiguity remained challenging for ChatGPT.

3. Scientific Knowledge

3.1. Formal Science

Mathematics. Frieder et al. (2023) proposed a mathematical benchmark at the graduate level to test ChatGPT's mathematical reasoning, related to textbook exercises, Olympiad problems, proof completion, algebra and probability theory problem solving, and theorem-proof and definition understanding. They found that ChatGPT only achieved a passing grade (50% of points) on 6 out of 17 testing sets. ChatGPT performed badly on mathematical problem solving, e.g., Olympiad problems, textbook exercises, algebra, and probability theory, while presenting comparatively better grades in definition understanding and proof completion. It seems ChatGPT is not good at solving mathematical problems that are weakly related to language memory or generation, because compared to the textual understanding, e.g., theorem-proofs and definitions, the other takes depend on multi-step reasoning. Qin et al. (2023) evaluated ChatGPT in arithmetic reasoning, finding that ChatGPT achieved the highest score on 3 out of 6 datasets, while fine-tuning-based methods exceeded ChatGPT on the rest of the 3 datasets. Bubeck et al. (2023) showed that GPT-4 largely exceeded the

Minerva expert model (Lewkowycz et al., 2022), while ChatGPT fell behind Minerva on average. From a user interaction perspective, Collins et al. (2023) evaluated the mathematical capabilities of InstructGPT, ChatGPT and GPT-4, and identified several weaknesses, including their performance in algebraic manipulations, tendency towards verbosity, and reliance on memorized solutions.

Computer Science. Bordt and von Luxburg (2023) developed an exam with 10 different exercises. ChatGPT and GPT-4 achieved 20.5 and 24 points out of 40 full points, respectively. GPT-4 slightly exceeded the average score (23.9) of 200 students. Bordt and von Luxburg (2023) believed that passing the exam by ChatGPT should not be misconstrued as an indication of its comprehension of computer science, because numerous topics addressed in the exam were extensively documented and readily available online. The coding skills of ChatGPT and GPT-4 largely exceeded the SOTA baselines (Bubeck et al., 2023). GPT-4 even achieved human-level accuracy in LeetCode tests, while ChatGPT yielded lower accuracy, reaching only half of the average human performance. Li et al. (2023) compared in-context learning (ICL)-based ChatGPT to SOTA models on Text-to-SQL tasks. To improve the performance of the examined ICL models, Chain-Of-Thought (COT) and extra knowledge evidence sentences were also incorporated. ChatGPT (40.08%) exceeded the strongest baseline Codex (36.47%, Chen et al., 2021b), while largely lagging behind humans (92.96%) in execution accuracy.

3.2. Natural Science

Physics. ChatGPT was evaluated as if it is a college student who needs to finish homework, clicker questions, programming exercises, and exams in the first-year Calculus-based Physics (Kortemeyer, 2023). Overall, ChatGPT achieved 53.05% after weighing different testing modules. This score met the minimum requirement for course credit, yet it adversely affected the overall grade-point average, falling below the necessary threshold for graduation. ChatGPT showed outstanding performance in clicker and programming questions, achieving scores higher than 90%. However, its performance in homework and exams was subpar. Additionally, ChatGPT's mathematical difficulties in the field of physics lowered its overall score.

Chemistry. Clark (2023) asked ChatGPT to finish two real chemistry exams with closed- and open-response questions. 44% closed-response questions were correctly answered by ChatGPT, although this is lower than the student average score (69%). Conversely, when it came to open-response questions, ChatGPT's performance was even lower than that of the least successful student.

Medicine. Gilson et al. (2023); Kung et al. (2023) showed that ChatGPT achieved college student level on the United States Medical Licensing Examination (USMLE). Antaki et al. (2023) found ChatGPT yielded low accuracy on neuro-ophthalmology and high accuracy on general medicine, indicating it had not grasped specialized medical knowledge well. Hirosawa et al. (2023) examined ChatGPT's diagnosis with 30 clinical vignettes, showing that the rate of top-10 differential suggestions covering the correct diagnosis reached 93.3%, while the accuracy of top-1 suggestions was just 53.3% (physicians were 93.3%). Rao et al. (2023) suggested that ChatGPT achieved an impressive final diagnosis accuracy (76.9%), while its initial diagnosis accuracy was just 60.3%. Mehnen et al. (2023) analyzed ChatGPT and GPT-4 diagnoses by 40 common and 10 rare clinical vignettes. They found that GPT-4 achieved 100% accuracy on the common cases within three possible diagnostic suggestions. ChatGPT reached about 95% with the same number of suggested diagnoses.

3.3. Social Science

Education. Dahlkemper et al. (2023) aimed to investigate the extent to which students possessed accurate perception regarding the scientific accuracy and the linguistic quality of ChatGPT's responses. They provided 102 physics students with ChatGPT answers and (masked) expert answers to evaluate the scientific accuracy and linguistic quality of both parties, perceived by the students, based on three physics questions. Despite the fact that all responses generated by ChatGPT in their study were incorrect, imperfect, or misleading, the evaluation results indicated that when confronted with a difficult question, students perceived ChatGPT's scientific accuracy to be on par with that of the expert solution, attributing this to the higher linguistic quality of ChatGPT. However, in cases where the questions were relatively easy, both the scientific accuracy and linguistic quality of the expert's answers was perceived to surpass ChatGPT.

Law. ChatGPT was tested with four law class exams by Choi et al. (2023), including Constitutional Law, Employee Benefits, Taxation, and Torts. ChatGPT performed better on 12 essay questions than on the 95 multiple-choice questions. Although it passed all four exams, it ranked almost the lowest among law school students in each class. Liu et al. (2023b) examined ChatGPT and GPT-4 in Law School Admission Test (LSAT) and the Chinese Civil Service Examination (CCSE). Compared with RoBERTa, the advantage of GPT-4 was much larger than that of ChatGPT. GPT-4 even exceeded the average level of a human on LSAT. Nevertheless, when compared to the human ceiling, significant disparities persisted in both models.

Economics. ChatGPT was examined with the Test of Understanding of College Economics (TUCE, Geerling et al., 2023). It demonstrated proficiency by providing accurate responses to 19/30 microeconomics questions and 26/30 macroeconomics questions. Such performance placed it in the upper echelons, ranking within the top 9% and top 1% among 3,255 and 2,789 college students who had successfully finished a full semester of microeconomics and macroeconomics. Xie et al. (2023) tested ChatGPT in stock predictions. ChatGPT showed weaker performance than traditional machine learning algorithms in numerical feature-based predictions. Through error analysis, it was determined that ChatGPT was also weak in comprehending investor sentiment expressed in text.

4. Ethical Considerations

Fairness asks models to be fair in terms of gender, race, language, culture, and more. Seghier (2023); Yong et al. (2023b) found that ChatGPT's responses were much worse in languages other than English (see § 2.2). Zhuo et al. (2023) assessed two types of social biases, namely gender bias and race bias. They concluded that, for text generation, although biases still existed, ChatGPT had mitigated them to a large extent compared to its predecessors, and that, for dialogue generation, ChatGPT could generate unbiased responses.

Robustness requires models to maintain their performance when the inputs are different from the training data. Such inputs could be noisy data, outliers and attacks. For noisy data, Zhuo et al. (2023) tried two NLP datasets while Ye et al. (2023) tried another 12 datasets. They found that although ChatGPT defeated preceding LLMs, it was still far from perfection. Wang et al. (2023d) assessed ChatGPT on outliers, with similar results. Besides generating bad responses, the unsuccessful handling of such inputs may sometimes cause more serious consequences, for example, data leakage or Denial of Service attacks. Peng et al. (2023) assessed whether ChatGPT would suffer from SQL injection if it served as a text-to-SQL interface (Li et al., 2023), while the answer was positive.

Reliability requires the generated text to be faithful. Neural text generators hallucinate (Ji et al., 2023) and, thus, produce unfaithful texts. ChatGPT/GPT-4 is no exception (Bang et al., 2023). Zhuo et al. (2023) assessed ChatGPT on fact-based question-answering datasets and concluded that it was not improved compared to its predecessors. This makes ChatGPT unreliable in tasks where faithfulness is vital. For example, it might make up references when writing scientific articles (Athaluri et al., 2023) and make up legal cases when serving in the legal domain (Deroy et al., 2023).

Toxicity asks models not to generate harmful, offensive and pornographic content. The GPT models were designed to normally refuse to generate toxic content. Nevertheless, Derner and Batistič (2023) found that through role-playing ChatGPT still produced offensive content. A quantitative study by Zhuo et al. (2023) showed that, by feeding toxic prompting, only 0.5% of the ChatGPT responses were toxic. Nonetheless, this ability was not very robust. In line with Derner and Batistič (2023), they also found that it was susceptible to prompt injections achieved by role-playing.

Domain-specific concerns were raised in cybersecurity (Sebastian, 2023), marketing (Rivas and Zhao, 2023), politics (Rozado, 2023; Motoki et al., 2023), legal (Deroy et al., 2023), education (Sallam, 2023; Lee, 2023; Kasneci et al., 2023), academic writing (Athaluri et al., 2023), recommendation (Zhang et al., 2023b) and tourism (Carvalho and Ivanov, 2023). These concerns highlighted the potential misuse of the GPTs, the spread of misinformation, and the erosion of privacy.

5. Discussion

5.1. Comparing GPT versus Humans

For NLP tasks with rich training resources, the GPTs may not perform as well as expert models (see Tables 2 and 4 in Appendix A). There is also a certain distance from humans (see Tables 3 and 5). However, when it comes to scientific knowledge, extensive multidisciplinary testing has showcased their superiority compared to earlier models (see Table 6). Notably, in computer science and law exams, GPT-4 has achieved a level of accuracy that closely matches or even surpasses the average human performance (see Table 7). However, when we are quoted a study that shows an AI system performing at a similar level to humans, these are typically based on average performance over many cases. This average hides the fact that the AI is typically outperforming on some specialist knowledge that is difficult for humans, and underperforming on some examples that are relatively easy for humans. The same has been shown in computer vision (Russakovsky et al., 2015). Thus, human-level average accuracy achieved by AI is not equivalent to human performance and intelligence.

The GPTs may exhibit instances of hallucinations and false information (Cabrera and Neubig, 2023; Bang et al., 2023). This may be attributed to the fact that the "next word prediction"-based pre-training only taught GPTs "what is right" while neglecting "what is wrong". To learn to generate a correct last word, e.g., "bird" after the context "if an animal has wings and can fly, it is likely a", GPTs have to learn logic, commonsense, linguistics, and science.

However, without learning "if an animal has wings and can fly, it is a penguin" is wrong, GPTs may struggle to distinguish penguins from other birds by flight ability and yield penguins' flying hallucinations, because GPTs likely have learned "penguins are birds, having wings" from corpora. In reality, incorrect examples are far more than the negations we can see from corpora. False information may be exhibited in ambiguous cases if GPTs do not know the boundary between positive and negative examples. In contrast, humans learn knowledge from both right and wrong applications to sidestep obvious fallacies (NASEM, 2018).

The inference process of GPT models is also dissimilar to humans. Humans are known to have two types of reasoning: "thinking fast" and "thinking slow" (Kahneman, 2000). However, GPT models seem to only have the thinking fast-like inference mechanism, e.g., they do a feedforward propagation without thinking twice (Bubeck et al., 2023). Humans in contrast can spend longer deliberating for hard questions. This requires a longer process of iterative inference, not necessarily dictated by a fixed number of iterations, but instead iterating until the system achieves a satisfactory state (van Bergen and Kriegeskorte, 2020). Nevertheless, it is clear that GPT models can do a certain type of reasoning. Especially, it has been shown that adding "Let's think step by step" to a prompt can allow GPT to use its own output as a sort of scratchpad to help it chain together multiple steps to arrive at a solution (Bubeck et al., 2023). In this way GPT is simulating "thinking slow", however, it is limited to "linear" sequences of thought and has severe limitations in tasks that require planning (ibid.); it does not backtrack to try other possible alternatives. Such backtracking would require some sort of short-term memory or workspace to remember what has been tried and what is yet to be tried. These are well-known shortcomings of neural models (Minsky, 1991).

Bubeck et al. (2023) noted the context-dependence of GPT's mathematical knowledge; "changes in the wording of the question can alter the knowledge that the model displays." The general phenomenon of sensitivity to input phrasing is well known from other language models also (Mao et al., 2023b). Lai et al. (2023) found a huge drop in commonsense knowledge when testing in languages other than English. This also explains why assessments by different papers can arrive at contradictory conclusions about the GPT's knowledge. It demonstrates that the internal representation in GPT suffers from severe entanglement (Bengio et al., 2013). While more extensive training can cause it to give correct responses in more contexts, it cannot change the fundamental fact that knowledge is not disentangled from the text contexts in which it appears; therefore there will always be the possibility that a change to the text of a question, that preserves the semantics, will elicit a factually incorrect response. Fixing this issue may require a hybrid system with separate facts. A commonsense-based neurosymbolic AI framework, such as the one proposed by Cambria et al. (2022) for sentiment analysis, moreover, can help increase the explainability of the reasoning processes required for decision-making, which is crucial for sensitive applications involving ethics, privacy and health.

5.2. Evaluation

Concerns about the validity of NLP evaluations in general (Belz et al., 2023) are applicable to this paper also. The following points are worth noting:

1) ChatGPT as a commercial product is updated periodically (Tu et al., 2023). This means the "flaws" identified in early studies are fixed in later versions and different studies about the same task are not fully comparable. Additionally, since the repair may be done by including datasets from previous assessments in fine-tuning, the data leakage may make the assessment unfair (Min et al., 2022; OpenAI, 2023a). This has been well-evidenced in Aiyappa et al. (2023) and de Wynter et al. (2023).

2) The design of prompts highly influences the results, leading to biased comparisons. Take MT as an example, by seeking better prompts, Hendy et al. (2023) and Jiao et al. (2023) had very different conclusions compared to Bang et al. (2023). In future studies, it is necessary to ensure transparency in prompt design and facilitate a fair comparison between LLMs and baselines.

3) The factors that matter in previous NLP evaluations are still valid, which may include the choice of evaluation corpora and metrics, the design of human evaluations, the task formulation and so on. For instance, Martínez (2023) argued that OpenAI's assessment of GPT-4 (OpenAI, 2023a) on the Uniform Bar Exam is misleading because they incorrectly included test-takers who re-took the exam in their comparison. Zhang et al. (2023c) misconcluded that GPT-4 had solved all MIT math and computer science curricula because they improperly used GPT-4 as the judge.

4) The assessments of some key abilities, e.g., creativity and logical reasoning, lack either objective criteria or large-scale benchmarks. Consequently, the evaluation can only capture a portion of the overall capabilities of GPT. People seem to commonly believe that AI will cause mass unemployment (Hatzius et al., 2023). However, in these fields, the evaluation is also very limited. It would be expected to see what kind of leading results the GPT models achieved compared with humans in the field of work that can be replaced by AI.

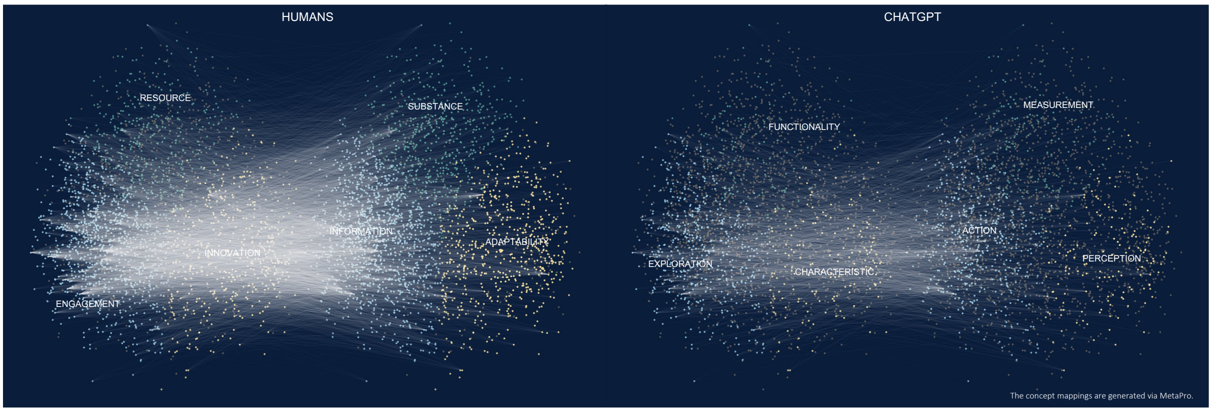

Figure 1: The concept mapping patterns between humans (left) and ChatGPT (right) from Mao et al. (2024a). Each cluster on the left represents target concepts, while on the right, the cluster represents source concepts. Bright and grey dots denote activated and unactivated concepts, respectively. The capitalized terms represent key activated concepts within a cluster.

5.3. Ethics

Several works have found that human perception about the reliability of ChatGPT's output can be misled by its seemingly scientific language style (Gudibande et al., 2023; Dahlkemper et al., 2023). Given AI-generated content, it would be necessary to allow users to retrieve source references that were generated by humans to improve response reliability. Caution should be also exercised when employing AI-generated data for training new AI models, as this practice carries the risk of introducing irreparable flaws into the resultant models (Shumailov et al., 2023). Mao et al. (2024a) comparatively analyzed the cognitive patterns of ChatGPT and humans via metaphoric concept mappings. They used MetaPro (Mao et al., 2023a) to extract target and source concepts from parallel answers provided by ChatGPT and humans to the same questions. Then, they plotted the concept distribution for each subject. In Figure 1, compared to the diverse activated concepts from humans, ChatGPT primarily activated the blue concepts, leaving many concepts greyed out (unactivated) in other areas, indicating concepts and concept mappings that were not utilized in its generated text. It shows the concept preference of ChatGPT, suggesting potential cognition biases from its generated text. Consequently, caution is advised when utilizing AI-generated text for training subsequent AI models, as it may propagate biases to these models.

Even if the Reinforcement Learning from Human Feedback (RLHF, OpenAI, 2023b) can somewhat mitigate biased, inaccurate, and toxic responses, and enhance human-centric output preferences, RLHF may be misled by human-biased feedback, e.g., system gaming (Leike et al., 2017), positive reward cycles (Ho et al., 2015), and tangled social norms and cultural context (Liu, 2023). There are also concerns regarding the non-transparency of the training set of ChatGPT, which could potentially lead to data leakage, including trade secrets or personal privacy, as user inputs might be utilized for fine-tuning (Min et al., 2022). These limitations appearing in the training and inferring processes may raise additional biases and ethical concerns.

6. Conclusion and Recommendations

In this work, we reviewed the latest assessments of ChatGPT and GPT-4, including their language and reasoning ability, scientific knowledge, and relevant ethical considerations. The former assessments showed that the GPTs have demonstrated strong capability in language understanding and generation, as well as general scientific knowledge. On the other hand, we also observed that the GPTs fall behind expert systems in many conventional NLP tasks; Their multi-step reasoning skills could benefit from further development; Ethical considerations in fairness, robustness, reliability, toxicity, and different application domains remain; The current evaluation tasks can be further improved with better transparency in GPT training corpora and evaluation methodology, as well as broader testing domains. In addition to implementing effective measures to address the aforementioned issues, the following aspects are also recommended:

Task-agnostic evaluation is desirable. Our survey indicates that many evaluations of ChatGPT and GPT-4 have primarily relied on benchmark data or task-specific inquiries. However, this method has limitations, e.g., the risk of data contamination during the pre-training phase, the variability of LLM performance across different prompts and task formulations, and the narrow scope of the evaluation.

To address these limitations, task-agnostic evaluation approaches offer a promising solution. In psychology, a common practice for assessing mental health involves using concept mappings derived from indirect questions and answers that reflect emotional states and cognition. These indirect tests, such as word-association tests, thematic apperception tests, and the Rorschach test (Rapaport et al., 1946), elicit responses that reflect the subconscious mind, which is believed to influence the majority of brain activities (Pally, 2007) and human behaviors (Pradeep, 2010). Similarly, for non-open-source LLMs, such as ChatGPT and GPT-4, researchers can develop indirect tasks to detect potential issues by considering LLMs as subjects. For example, cognitive bias analysis can be conducted by comparing concept mappings derived from human-generated and ChatGPT-generated text, particularly in response to questions from other domains (Mao et al., 2024a). This comparative analysis can reveal discrepancies that may indicate biases or limitations in the model's understanding. For open-source neural networks, neuro-activation variation analysis can also be employed to identify potential deficiencies in pre-trained language models (Chen et al., 2021a).

Fundamental research is still valuable. The landscape of NLP has been largely influenced by the emergence of ChatGPT-like LLMs. Researchers are obsessed with using them to process high-level application tasks. Comparatively, conventional computational linguistics tasks, e.g., syntactic and semantic processing (Zhang et al., 2023d; Mao et al., 2024b), have taken a backseat in recent times. This shift is driven by the belief that the direct utility of these foundational techniques is less pronounced in enhancing model performance on high-level tasks. For instance, semantic knowledge can now be acquired through extensive pre-training, obviating the need for explicitly distinguishing parts of speech or word senses.

However, the significance of computational linguistics research extends beyond the development of NLP applications. It also offers insights into how humans perceive and use language. Fundamental research in computational linguistics remains valuable, as in many resource-rich areas, "expert models" still lead by some distance, compared to the GPTs. However, the scope of the fundamental tasks must be broadened to consider more than simple classification based on input text, encompassing, e.g., broader narrative (Cambria and White, 2014) or cognitive (Ge et al., 2023) understanding. The value of studying fundamental tasks is also manifested in the spirit of task decomposition – one of the important pillars of AI (Cambria et al., 2023). Failing to break down a complex task into its component subtasks effectively requires the model to implicitly address numerous subtasks for which it has not been specifically trained. This approach undermines the establishment of trustworthiness and accountability.

AI-generated content should be regulated. To keep pace with the rapid development of generative AI, regulatory frameworks must also evolve swiftly to prevent the misuse of AI and AI-generated content. The success of current LLMs largely stems from their exposure to diverse human-generated content from various sources and perspectives. This diversity enables models like GPTs to learn from a wide range of ideas and generate output that aligns with the expectations of the majority.

However, the speed at which machines can generate content far exceeds that of humans. As shown in Figure 1, humans presented greater diversity in metaphorical concepts and concept mappings compared to ChatGPT, reflecting the richness of human thought. Without proper controls, a large volume of AI-generated content may blend with human-generated content, potentially diluting the diversity of human thought. This scenario could lead to individuals perceiving opinions from a single major source – machines. Moreover, the misuse of AI-generated content for training could introduce biases into subsequent AI (Mao et al., 2024a). Given the aforementioned ethical considerations and the limitations of current evaluation methods, it is necessary to properly regulate AI-generated content before knowing how to use them well. Otherwise, as Cambria et al. (2023) argued: "AI could very much end up being like plastic: a great invention that made our life easier about a century ago but is now threatening our own existence."

The Future of LLMs. The major advances in recent LLMs have largely come from scaling up. At this time so much text training data is already being used that the potential for scaling up further is limited. Recently interest has moved to multimodal models, for example, those that use images as well as text (Yang et al., 2023; McKinzie et al., 2024), and other modalities (Wang et al., 2023e). This opens a much wider range of human tasks that these models could potentially be applied to. Some of the major shortcomings of LLMs such as hallucination or inaccuracy, and inability to do "thinking slow" reasoning, seem unlikely to be solved at a foundational level in the near future. Instead, we see LLMs using external resources to compensate for their deficiencies, e.g. they can fact-check using provided resources, or the Internet, or they can check calculations with a calculator, or check code with an interpreter. Furthermore, multimodal models can potentially be proactive in going to websites and performing tasks using the graphical interface as a human would. LLMs would be conceptualized as the kernel process of a new operating system, facilitating interactions with various resources to accomplish tasks.

Acknowledgments

This research/project is supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005). Guanyi Chen is supported by the Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning and the National Language Resources Monitoring and Research Center for Network Media of Central China Normal University in Wuhan, China.

7. Bibliographical References

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can we trust the evaluation on ChatGPT? *arXiv preprint arXiv:2303.12767*.

Mostafa M Amin, Rui Mao, Erik Cambria, and Björn W Schuller. 2023. A wide evaluation of ChatGPT on affective computing tasks. *arXiv preprint arXiv:2308.13911*.

Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. 2023. Evaluating the performance of ChatGPT in ophthalmology: An analysis of its successes and shortcomings. *Ophthalmology Science*, page 100324.

Sai Anirudh Athaluri, Sandeep Varma Manthena, VSR Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. 2023. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*, 15(4).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Anja Belz, Craig Thomson, and Ehud Reiter. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359.

Sebastian Bordt and Ulrike von Luxburg. 2023. ChatGPT participates in a computer science exam. *arXiv preprint arXiv:2303.09461*.

Ali Borji. 2023. A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Alex Cabrera and Graham Neubig. 2023. Zeno chatbot report. Accessed: 18/05/2023.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. 2023. Seven pillars for the future of Artificial Intelligence. *IEEE Intelligent Systems*, 38(6):62–69.

Erik Cambria and Bebo White. 2014. Jumping NLP curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57.

Inês Carvalho and Stanislav Ivanov. 2023. ChatGPT for tourism: Applications, benefits and risks. *Tourism Review*.

Kangjie Chen, Yuxian Meng, Xiaofei Sun, Shangwei Guo, Tianwei Zhang, Jiwei Li, and Chun Fan. 2021a. BadPre: Task-agnostic backdoor attacks to pre-trained NLP foundation models. In *International Conference on Learning Representations*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021b. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023.

Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. Accessed: 16/05/2023.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. ChatGPT goes to law school. *Journal of Legal Education (Forthcoming)*.

Haoran Chu and Sixiao Liu. 2023. Can AI tell good stories? Narrative transportation and persuasion with ChatGPT. *PsyArXiv*.

Ted M Clark. 2023. Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education*.

Katherine M. Collins, Albert Qiaochu Jiang, Simon Frieder, Li Siang Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B. Tenenbaum, William Hart, Timothy Gowers, Wen-Ding Li, Adrian Weller, and Mateja Jamnik. 2023. Evaluating language models for mathematics through interactions. *arXiv preprint arXiv:2306.01694*.

Merten Nikolay Dahlkemper, Simon Zacharias Lahme, and Pascal Klein. 2023. How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality. *arXiv preprint arXiv:2304.05906*.

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

Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. 2023. An evaluation on large language model outputs: Discourse and memorization. *arXiv preprint arXiv:2304.08637*.

Erik Derner and Kristina Batistič. 2023. Beyond the safeguards: Exploring the security risks of ChatGPT. *arXiv preprint arXiv:2305.08005*.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How ready are pre-trained abstractive models and LLMs for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier, and Julius Berner. 2023. Mathematical capabilities of ChatGPT. *arXiv preprint arXiv:2301.13867*.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? A comprehensive evaluation. *arXiv preprint arXiv:2305.07375*.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56:1829–1895.

Wayne Geerling, G Dirk Mateer, Jadrian Wooten, and Nikhil Damodaran. 2023. ChatGPT has aced the test of understanding in college economics: Now what? *The American Economist*, page 05694345231169654.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. 2023. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. NEWSROOM: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 708–719.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6394–6407.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey

Levine, and Dawn Song. 2023. The false promise of imitating proprietary LLMs. *arXiv preprint arXiv:2305.15717*.

RJan Hatzius, Joseph Briggs, Devesh Kodnani, and Giovanni Pierdomenico. 2023. The potentially large effects of artificial intelligence on economic growth. Technical report, Goldman Sachs.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Takanobu Hirosawa, Yukinori Harada, Masashi Yokose, Tetsu Sakamoto, Ren Kawamura, and Taro Shimizu. 2023. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained Transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. *International Journal of Environmental Research and Public Health*, 20(4):3378.

Mark K Ho, Michael L Littman, Fiery Cushman, and Joseph L Austerweil. 2015. Teaching with rewards and punishments: Reinforcement or communication? In *CogSci*.

Bart Holterman and Kees van Deemter. 2023. Does ChatGPT have theory of mind? *arXiv preprint arXiv:2305.14020*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-EVAL: A multi-level multi-discipline Chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Daniel Kahneman. 2000. *Thinking Fast and Slow*. Farrar.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

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. *arXiv preprint arXiv:2302.10724*.

Gerd Kortemeyer. 2023. Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19(1):010132.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, 2(2):e0000198.

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 Xiangji Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *arXiv preprint arXiv:2305.18486*.

Hyunsu Lee. 2023. The rise of ChatGPT: Exploring its potential in medical education. *Anatomical Sciences Education*.

Jan Leike, Miljan Martic, and Shane Legg. 2017. Learning through human feedback. Technical report, Google DeepMind.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*.

Jinyang Li, Binyuan Hui, Ge Qu, Binhua Li, Jiaxi Yang, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, et al. 2023. Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs. *arXiv preprint arXiv:2305.03111*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023a. We're afraid language models aren't modeling ambiguity. *arXiv preprint arXiv:2304.14399*.

Gabrielle Kaili-May Liu. 2023. Perspectives on the social impacts of reinforcement learning with human feedback. *arXiv preprint arXiv:2303.02891*.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023b. Evaluating the logical reasoning ability of ChatGPT and GPT-4. *arXiv preprint arXiv:2304.03439*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-EVAL: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

François Mairesse, Milica Gasic, Filip Jurcicek, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561.

Rui Mao, Guanyi Chen, Xiao Li, Mengshi Ge, and Erik Cambria. 2024a. A comparative analysis of metaphorical cognition in ChatGPT and human minds. Technical report, Nanyang Technological University.

Rui Mao, Kai He, Xulang Zhang, Guanyi Chen, Jinjie Ni, Zonglin Yang, and Erik Cambria. 2024b. A survey on semantic processing techniques. *Information Fusion*, 101:101988.

Rui Mao, Xiao Li, Kai He, Mengshi Ge, and Erik Cambria. 2023a. MetaPro Online: A computational metaphor processing online system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, volume 3, pages 127–135.

Rui Mao, Qian Liu, Kai He, Wei Li, and Erik Cambria. 2023b. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

Gary Marcus and Ernest Davis. 2023. How not to test GPT. Accessed: 29/05/2023.

Eric Martínez. 2023. Re-evaluating GPT-4's bar exam performance. *Available at SSRN 4441311*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. 2024. MM1: Methods, analysis insights from multimodal llm pre-training.

Lars Mehnen, Stefanie Gruarin, Mina Vasileva, and Bernhard Knapp. 2023. ChatGPT as a medical doctor? A diagnostic accuracy study on common and rare diseases. *medRxiv*, pages 2023–04.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Marvin L Minsky. 1991. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI magazine*, 12(2):34–34.

Shima Rahimi Moghaddam and Christopher J. Honey. 2023. Boosting theory-of-mind performance in large language models via prompting. *arXiv preprint arXiv:2304.11490*.

NLP Team MosaicML. 2023. Introducing MPT-7B: A new standard for open-source, commercially usable LLMs. Accessed: 18/05/2023.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: Measuring ChatGPT political bias. *Available at SSRN 4372349*.

NASEM, National Academies of Sciences, Engineering, and Medicine and others. 2018. *How people learn II: Learners, contexts, and cultures*. National Academies Press.

OpenAI. 2023a. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2023b. Introducing ChatGPT. Accessed: 11/05/2023.

Regina Pally. 2007. The predicting brain: Unconscious repetition, conscious reflection and therapeutic change. *The International Journal of Psychoanalysis*, 88(4):861–881.

Brad Payne. 2020. Privacy protection with AI: Survey of data-anonymization techniques. Accessed: 03/06/2023.

Xutan Peng, Yipeng Zhang, Jingfeng Yang, and Mark Stevenson. 2023. On the security vulnerabilities of text-to-SQL models. *arXiv preprint arXiv:2211.15363*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the 10th Workshop on Statistical Machine Translation*.

Anantha Krishnan Pradeep. 2010. *The buying brain: Secrets for selling to the subconscious mind*. John Wiley & Sons.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

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*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Arya S Rao, Michael Pang, John Kim, Meghana Kamineni, Winston Lie, Anoop K Prasad, Adam Landman, Keith Dryer, and Marc D Succi. 2023. Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv*.

David Rapaport, Merton Gill, and Roy Schafer. 1946. *Diagnostic psychological testing: The theory, statistical evaluation, and diagnostic application of a battery of tests: Volume II.* The Year Book Publishers.

Xuan Ren and Lingqiao Liu. 2023. You can generate it again: Data-to-text generation with verification and correction prompting. *arXiv preprint arXiv:2306.15933*.

Pablo Rivas and Liang Zhao. 2023. Marketing with ChatGPT: Navigating the ethical terrain of GPT-based chatbot technology. *AI*, 4(2):375–384.

David Rozado. 2023. The political biases of ChatGPT. *Social Sciences*, 12(3):148.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of ChatGPT. *arXiv preprint arXiv:2304.07333*.

Malik Sallam. 2023. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887.

Erik Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Glorin Sebastian. 2023. Do ChatGPT and other AI chatbots pose a cybersecurity risk?: An exploratory study. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, 15(1):1–11.

Mohamed L Seghier. 2023. ChatGPT: Not all languages are equal. *Nature*, 615(7951):216–216.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Mayank Soni and Vincent Wade. 2023. Comparing abstractive summaries generated by ChatGPT to real summaries through blinded reviewers and text classification algorithms. *arXiv preprint arXiv:2303.17650*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? Investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. Technical report, Stanford Center for Research on Foundation Models.

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

Shangqing Tu, Chunyang Li, Jifan Yu, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2023. ChatLog: Recording and analyzing ChatGPT across time. *arXiv preprint arXiv:2304.14106*.

Ruben S van Bergen and Nikolaus Kriegeskorte. 2020. Going in circles is the way forward: The role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65:176–193.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can ChatGPT defend the truth? Automatic dialectical evaluation elicits LLMs' deficiencies in reasoning. *arXiv preprint arXiv:2305.13160*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Zero-shot cross-lingual summarization via large language models. *arXiv preprint arXiv:2302.14229*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023c. Is ChatGPT a good NLG evaluator? A preliminary study. *arXiv preprint arXiv:2303.04048*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. 2023d. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. 2023e. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205*.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023. The Wall Street neophyte: A zero-shot analysis of ChatGPT over multimodal stock movement prediction challenges. *arXiv preprint arXiv:2304.05351*.

Liang Xu and others from SuperCLUE team. 2023. SuperCLUE: A benchmark for foundation models in Chinese. Accessed: 18/05/2023.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision).

Wentao Ye, Mingfeng Ou, Tianyi Li, Yipeng chen, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, and Junbo Zhao. 2023. Assessing hidden risks of LLMs: An empirical study on robustness, consistency, and credibility. *arXiv preprint arXiv:2305.10235*.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023a. Low-resource languages jailbreak GPT-4. *arXiv preprint arXiv:2310.02446*.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Samuel Cahyawijaya, Holy Lovenia, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Long Phan, Yin Lin Tan, and Alham Fikri Aji. 2023b. Prompting multilingual large language models to generate code-mixed texts: The case of South East Asian languages. *arXiv preprint arXiv:2303.13592*.

Shuzhou Yuan and Michael F"arber. 2023. Evaluating generative models for graph-to-text generation. *arXiv preprint arXiv:2307.14712*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Extractive summarization via ChatGPT for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023b. Is ChatGPT fair for recommendation? Evaluating fairness in large language model recommendation. *arXiv preprint arXiv:2305.07609*.

Sarah J. Zhang, Samuel Florin, Ariel N. Lee, Eamon Niknafs, Andrei Marginean, Annie Wang, Keith Tyser, Zad Chin, Yann Hicke, Nikhil Singh, Madeleine Udell, Yoon Kim, Tonio Buonassisi, Armando Solar-Lezama, and Iddo Drori. 2023c. Exploring the MIT mathematics and EECS curriculum using large language models. *arXiv preprint arXiv:2306.08997*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Xulang Zhang, Rui Mao, and Erik Cambria. 2023d. A survey on syntactic processing techniques. *Artificial Intelligence Review*, 56:5645–5728.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multidimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming ChatGPT via jailbreaking: Bias, robustness, reliability and toxicity. *arXiv preprint arXiv:2301.12867*.

A. ChatGPT and GPT-4 benchmark

We visualize the performance gaps between the surveyed GPT models and baselines on different tasks in Tables 2-7. Task abbreviations and meanings can be viewed in Table 1. The visualization illustrates different levels of disparity in performance between ChatGPT or GPT-4 and the most robust baseline presented in an evaluation paper. Darker colors represent greater disparities, while lighter colors indicate smaller differences. The degree of discrepancy is measured by $g/b-1$, where g is the average score of a GPT model on the major metric of an evaluation task; b is the average score of a baseline on the same major metric. The color red indicates instances where the GPT models outperformed the baselines, while blue indicates cases where the GPT models lagged behind. The teal color represents the gaps between the GPT models and the ground truth, while gray indicates a lack of comparison. The summary of the surveyed assessment papers can be viewed in Tables 8 and 9.

Abbr.	Meaning	Abbr.	Meaning
ACI	Actual Causality Inferring	Natu.S	Natural Science
Asp.E	Aspect Extraction	NER	Named Entity Recognition
Asp.PD	Aspect Polarity Detection	NLI	Natural Language Inference
Caus.D	Cause Discovery	Offe.D	Offensiveness Detection
Caus.R	Causal Reasoning	OKEP	Ophthalmic Knowledge Assessment Program
CBQA	Complex Boolean Question Answering	Opin.E	Opinion Extraction
CCSE	Chinese Civil Service Examination	Overa.	Overall
CEG	Causal Explanation Generation	Pass.R	Qassage Re-ranking
CFA	Chinese-Featured Ability	Pers.D	Personality Detection
Chem.Exam	Chemistry Exam	Phys.Exam	Physics Exam
Com.R	Commonsense Reasoning	PoS	Part-of-Speech Tagging
Coun.R	Counterfactual Reasoning	Prag.	Pragmatic Processing Tasks
CS Exam	Computer Science Exam	Prof.A	Professional Ability
D2TE	Data-to-Text Generation Evaluation	QA	Question Answering
DGE	Dialogue Generation Evaluation	RE	Relation Extraction
Diagno.	Diagnosis	Reas.	Reasoning
Dial.S	Dialogue Systems	Sarc.D	Sarcasm Detection
DLR	Dialogue Logic Reasoning	Sem.	Semantic Processing Tasks
Edu.Ling.	Linguistic Quality in Education	Sent.A	Sentiment Analysis
Edu.Sci.	Scientific Accuracy in Education	Sent.R	Sentiment Ranking
EE	Event Extraction	SGE	Story Generation Evaluation
Emoj.P	Emoji Prediction	Soci.S	Social Science
Emot.Re	Emotion Recognition	SpamD	Spam Detection
Emot.Ra	Emotion Ranking	Stan.D	Stance Detection
Enga.A	Engagement Analysis	STEM	Science, Technology, Engineering, and Mathematics
Eval.	Evaluation	Subj.D	Subjectivity Detection
Form.S	Formal Science	Suic.D	Suicide Detection
Gene.	Generation Tasks	Summ.	Summarisation
Huma.	Humanity	Synt.	Syntactic Processing Tasks
Humo.R	Humour recognition	T-Lo.R	Task-Oriented Logical Reasoning
Info.R	Information Retrieval	T-Sy.R	Task-oriented Symbolic Reasoning
LeetCo.	LeetCode	T.Acc	Tweet Annotation Accuracy
LID	Language Identification	T.Agre	Tweet Annotation Agreement
Ling.A	Linguistic Acceptability	T2SQL	Text-to-SQL
Logi.R	Logical Reasoning	Toxi.D	Toxicity Detection
LSAT	Law School Admission Test	TSE	Text Summarisation Evaluation
Mach.T	Machine Translation	TUCE	Test of Understanding of College Economics
Med.Kno.	Medical Knowledge	USMLE	United States Medical Licensing Examination
Medi.S	Medical Science	WBA	Well-being Analysis
Misi.D	Misinformation Detection	WSD	Word Sense Disambiguation
MVQE	Medical and Vocational Qualification Examinations		

Table 1: Task abbreviations and meanings.

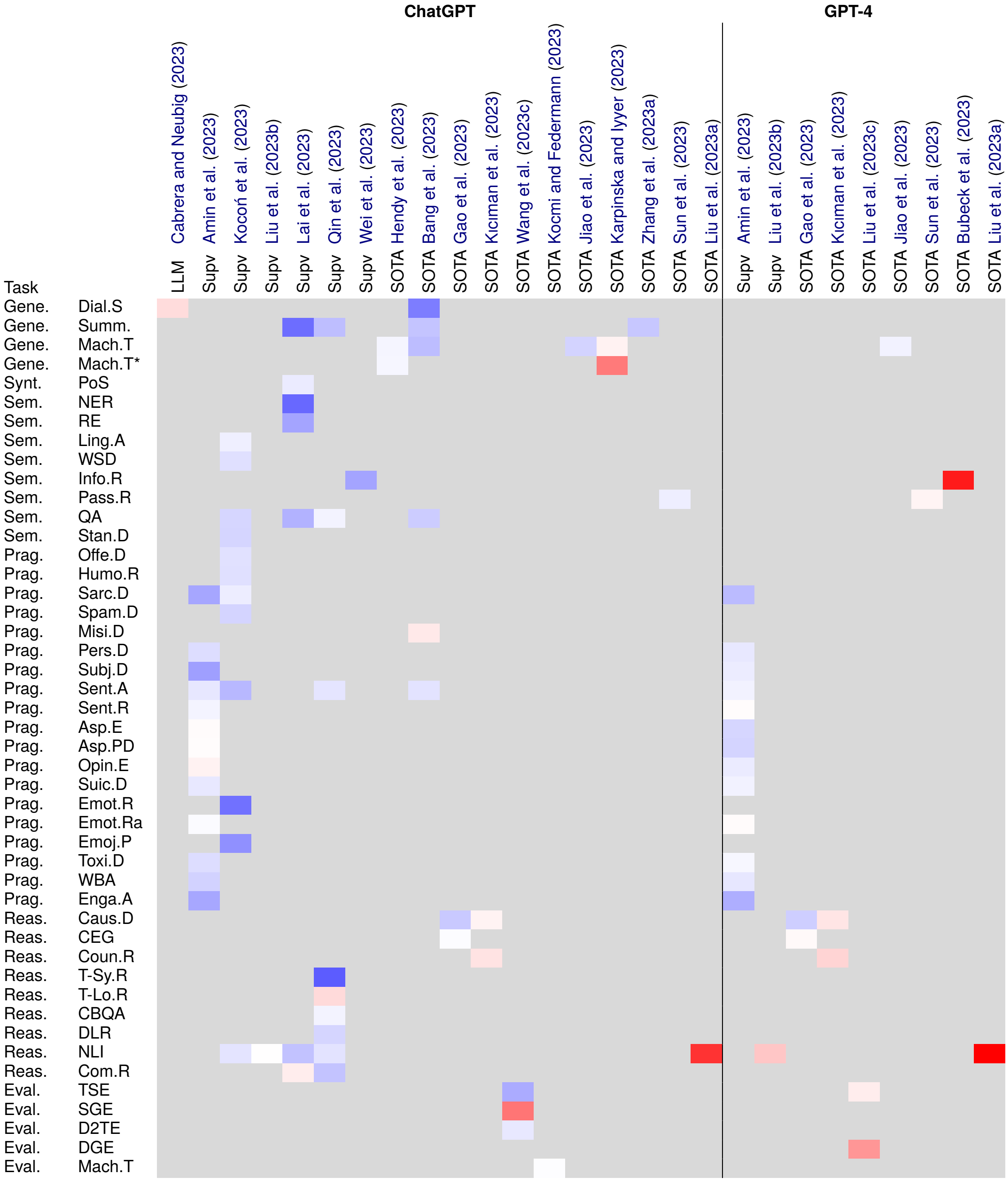

Table 2: ChatGPT and GPT-4 performance on linguistic and reasoning tasks, compared to baseline models. Supv denotes supervised models. The column names denote the type of the strongest baseline, introduced in an evaluation paper. The row names indicate evaluation tasks.

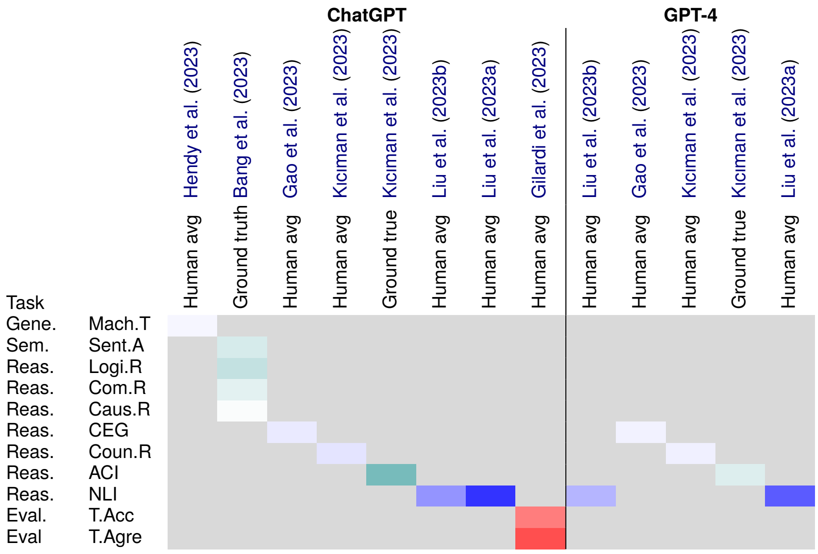

Table 3: ChatGPT and GPT-4 performance on linguistic and reasoning tasks, compared to humans or ground truth.

Task		ChatGPT LLM Xu et al. (2023)	ChatGPT LLM Huang et al. (2023)	ChatGPT Supv Lai et al. (2023)	ChatGPT Supv Wei et al. (2023)	ChatGPT SOTA Wang et al. (2023b)	GPT-4 LLM Xu et al. (2023)	GPT-4 LLM Huang et al. (2023)	GPT-4 SOTA Wang et al. (2023b)
en2zh	Summ.								
en2de	Summ.								
ru	PoS								
ru	NER								
ru	RE								
ru	NLI								
ru	QA								
ru	Com.R								
ru	Summ.								
de	PoS								
de	NER								
de	RE								
de	NLI								
de	QA								
de	Com.R								
zh	PoS								
zh	NER								
zh	RE								
zh	EE								
zh	NLI								
zh	QA								
zh	Com.R								
zh	Summ.								
zh	Ling.A								
zh	Prof.A								
zh	CFA								
zh	STEM								
zh	Soci.S								
zh	Huma.								
zh	MVQE								
fr	PoS								
fr	RE								
fr	NLI								
fr	Com.R								
fr	Summ.								
es	PoS								
es	NER								
es	RE								
es	NLI								
es	QA								
es	Com.R								
es	Summ.								
it	PoS								
it	NER								
it	RE								
it	Com.R								
nl	PoS								
nl	RE								
nl	Com.R								
pl	PoS								
pl	RE								
pl	Com.R								
pt	RE								
pt	Com.R								

		ChatGPT Supv Lai et al. (2023)
jp	Com.R	
vi	PoS	
vi	NLI	
vi	QA	
vi	Com.R	
tr	PoS	
tr	NER	
tr	NLI	
tr	QA	
tr	Summ.	
sv	RE	
ar	PoS	
ar	RE	
ar	NLI	
ar	QA	
ar	Com.R	
ar	Summ.	
fa	NER	
fa	RE	
ko	NER	
ko	RE	
el	PoS	
el	NLI	
el	QA	
th	PoS	
th	NLI	
th	QA	
uk	RE	
bg	PoS	
bg	NLI	
hi	PoS	
hi	NLI	
hi	QA	
hi	Com.R	
hi	Summ.	
bn	NER	
bn	Com.R	
bn	Summ.	
ta	Com.R	
ur	PoS	
ur	NLI	
ur	Com.R	
ml	Com.R	
mr	Com.R	
te	Com.R	
gu	Com.R	
my	Summ.	
kn	Com.R	
sw	NLI	
sw	Com.R	
pa	Com.R	
ky	Summ.	
or	Com.R	
as	Com.R	

Table 4: ChatGPT and GPT-4 performance on multi-lingual tasks, compared to baseline models. Languages are chunked by high (en2zh-vi), medium (tr-hi), low (bn-kn), and extremely low (sw-as) language resources, respectively, categorized by Lai et al. (2023).

Task		ChatGPT Human Xu et al. (2023)	ChatGPT Ground truth Bang et al. (2023)	G4 Human Xu et al. (2023)
zh	Ling.A			
zh	Prof.A			
zh	Overa.			
id	Sent.A			
id	LID			
jw	Sent.A			
jw	LID			
bug	Sent.A			
bug	LID			

Table 5: ChatGPT and GPT-4 performance on multi-lingual tasks, compared to humans or ground truth.

Task		ChatGPT Old GPT Gilson et al. (2023)	ChatGPT Supv Liu et al. (2023b)	ChatGPT Supv Qin et al. (2023)	ChatGPT SOTA Li et al. (2023)	ChatGPT SOTA Xie et al. (2023)	ChatGPT Expert LLM Kung et al. (2023)	ChatGPT SOTA Bubeck et al. (2023)	GPT-4 Supv Liu et al. (2023b)	GPT-4 SOTA Bubeck et al. (2023)
Form.S	T2SQL									
Form.S	Coding									
Form.S	LeetCo.									
Form.S	Math									
Soci.S	CCSE									
Soci.S	LSAT									
Soci.S	Finance									
Medi.S	Med.Kno.									

Table 6: ChatGPT and GPT-4 performance on scientific knowledge, compared to baselines.

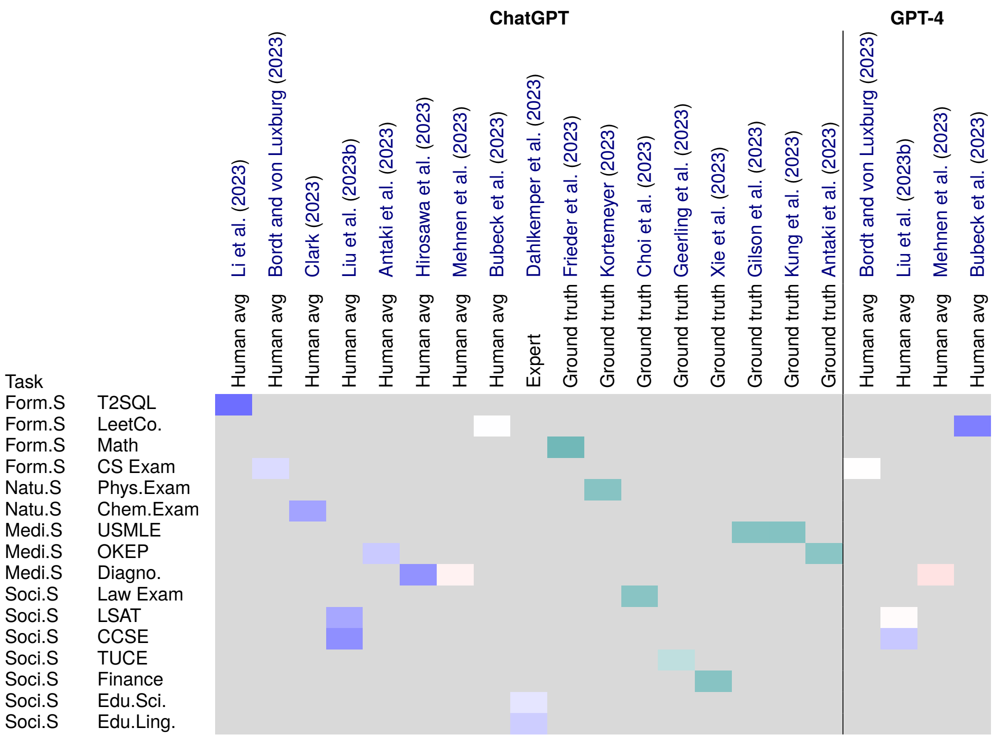

Table 7: ChatGPT and GPT-4 performance on scientific knowledge, compared to humans or ground truth.

Reference	Evaluation task	Version	Date	Evaluation method	Evaluation method novelty
Cabrera and Neubig (2023)	Dialogue fluency	CG, G4	18/05	Compare to zero-shot LLMs on benchmark data	Unified interface for open-source LLMs; Online browsing and analysis
Bang et al. (2023)	Summ., Mach.T, Sent.A, QA, Misi.D	CG	28/02	Compare to SOTA on benchmark data	Multitask, multilingual, multimodal evaluation
Hendy et al. (2023)	Machine translation	CG	18/02	Compare to SOTA and other LLMs on benchmark data	Evaluating by prompting strategies, languages, document-level and domain robustness
Jiao et al. (2023)	Machine translation	CG, G4	19/03	Compare to commercial MT systems on benchmark data	Testing robustness of MT systems on domain biases and noise
Karpinska and Iyyer (2023)	Machine translation	CG, G4	22/05	Compare to commercial MT systems on literacy texts	Conducting large scale human evaluation by experts and error analysis
Qin et al. (2023)	NLP and reasoning tasks	CG	15/02	Compare to other LLMs and supervised models on benchmark data	A wide range of task-oriented reasoning tasks
Zhang et al. (2023a)	Summarisation	CG	09/04	Compare to SOTA on benchmark data	Exploring several ways to enhance the performance
Soni and Wade (2023)	Summarisation	CG	30/04	Compare to humans on benchmark data	Humans and a fine-tuned classifier try to distinguish GPT generated summary from references
Jentzsch and Kersting (2023)	NLG	CG	07/06	Analyzed by case study and statistics	Evaluated ChatGPT in joke generation, detection and explanation
Amin et al. (2023)	13 affective computing tasks	CG, G4	28/08	Compare to naive supervised models on benchmark data	Novel prompts for evaluating models on classification and regression tasks
Kocoń et al. (2023)	25 sem. and prag. tasks	CG	21/02	Compare to supervised models on benchmark data	Analyzing ChatGPT by task difficulties, contexts, and explainability
Wei et al. (2023)	RE, NER, EE	CG	20/02	Compare to supervised models on benchmark data	Comparing to few-shot learning baselines with diverse shot values
Sun et al. (2023)	Document ranking	CG, G4	19/04	Compare to BM25 and SOTA on benchmark data	Testing three ranking strategies, namely query-, relevance- and permutation-generation
Wang et al. (2023c)	NLG evaluation	CG	25/04	Compare to SOTA on benchmark data	A wide range of NLG tasks
Liu et al. (2023c)	NLG evaluation	G4	06/04	Compare to SOTA on benchmark data	A wide range of NLG tasks
Kocmi and Federmann (2023)	Machine translation evaluation	CG	28/02	Compare to SOTA and other LLMs on benchmark data	Testing four distinctive prompts for machine translation evaluation
Gilardi et al. (2023)	Text classification annotation	CG	27/03	Compare to humans on annotation tasks	Comparing ChatGPT with crowd-workers in annotation accuracy and agreements
Xu et al. (2023)	Multi-level language abilities	CG, G4	18/05	Compare to other LLMs on benchmark data	Evaluating language abilities from different aspects
Huang et al. (2023)	Multi-discipline knowledge in Chinese	CG, G4	17/05	Compare to other LLMs on benchmark data	Scientific knowledge benchmarking at multi-levels
Lai et al. (2023)	Multilingual NLP	CG	12/04	Compare to supervised models on benchmark data	A wide range of multilingual benchmarking tasks with 13 languages
Wang et al. (2023b)	Multilingual summarization	CG, G4	06/04	Compare to fine-tuned mBART-50 and other LLMs on benchmark data	Comparing different prompting methods
Borji (2023)	Reasoning, linguistics, and perception.	CG	03/04	Analyzed by case study	Collecting ChatGPT failures cases from different tasks
Davis (2023)	Commonsense	CG	22/02	Analyzed by case study	Reviewing commonsense benchmarks in multiple modalities
Gao et al. (2023)	Causal reasoning	CG, G4	25/04	Compare to SOTA, other LLMs and humans on benchmark data	Fine-grained causal reasoning benchmarking tasks
Kıcıman et al. (2023)	Causal reasoning	CG, G4	08/05	Compare to SOTA and other LLMs on benchmark data	Fine-grained causal reasoning benchmarking tasks
Kosinski (2023)	Theory of Mind	CG, G4	14/03	Compare to ground truth on self-developed questions	Using a number of manmade false and belief tasks
Moghaddam and Honey (2023)	Theory of Mind	CG, G4	26/04	Compare to ground truth on self-developed questions and humans	Using a number of manmade false and belief tasks
Holterman and van Deemter (2023)	Theory of Mind	CG, G4	23/05	Compare to ground truth on self-developed questions	Considering 4 different types of Theory of Mind tasks
Wang et al. (2023a)	Dialectical evaluation	CG	22/05	Measure the failure rates on benchmark data	A novel ChatGPT's deficiency evaluation task with diverse reasoning setups
Liu et al. (2023b)	Reading comprehension and NLI	CG, G4	05/05	Compare to fine-tuned RoBERTa base and humans on benchmark data	Developing a toolkit for prompt-based LLM testing
Liu et al. (2023a)	NLI	CG, G4	27/04	Compare to SOTA on benchmark data	A novel NLI task to evaluate the abilities of ChatGPT and GPT-4 in detecting ambiguous language
Bubeck et al. (2023)	Coding, math, perception abilities	G4	13/04	Compare to SOTA and humans on benchmark data and analyze by cases	Thoroughly evaluating GPT-4 abilities in diverse scenarios
Li et al. (2023)	Text-to-SQL	CG	04/05	Compare to SOTA models and humans on benchmark data	Developing a large and difficult text-to-SQL benchmarking dataset
Bordt and von Luxburg (2023)	Math	CG	22/05	Compare to ground truth on an exam	Blind evaluation together with 200 participating students.
Frieder et al. (2023)	Math	CG	31/01	Compare to ground truth on benchmark data	Providing a large natural-language mathematics testing-oriented dataset
Collins et al. (2023)	Math	CG, G4	05/11	Interactive rating by humans and analyzed by experts' case study	Providing an interactive evaluation platform for LLM evaluation in mathematics
Kortemeyer (2023)	Physics	CG	11/05	Compare to ground truth on multiple testing modules	Evaluating ChatGPT by homework, clicker, programming exercises, and exams
Clark (2023)	Chemistry	CG	24/04	Compare to humans on multiple testing modules	Evaluating ChatGPT by closed- and open-response questions
Dahlkemper et al. (2023)	Education	CG	12/05	Compare to experts on 3 physics questions	Evaluating student perceptions about ChatGPT and expert answers by scientific and language accuracy
Choi et al. (2023)	Law	CG	23/01	Compare to ground truth on an exam	Blind evaluation based on four law class exams
Geerling et al. (2023)	Economics	CG	28/03	Compare to ground truth on an exam	Blind evaluation based on TUCE
Xie et al. (2023)	Finance	CG	28/04	Compare to SOTA on benchmark data	Evaluating ChatGPT by single- and multi-module setups in stock prediction
Gilson et al. (2023)	Medicine	CG	08/02	Compare to ground truth on an exam	Analyzing by question difficulty and response quality.
Kung et al. (2023)	Medicine	CG	09/02	Compare to ground truth on an exam	Analyzing by concordance and density of insight
Antaki et al. (2023)	Medicine	CG	11/06	Compare to ground truth on an exam	Analyzing by examination sections, cognitive level, and question difficulty
Hirosawa et al. (2023)	Medicine	CG	15/02	Compare to humans on 30 clinical vignettes	Comparing ChatGPT with physicians by top 5 and top 1 suggestions
Rao et al. (2023)	Medicine	CG	26/02	Compare to ground truth on 36 clinical vignettes	Evaluating ChatGPT throughout the entire clinical workflow
Mehnen et al. (2023)	Medicine	CG	27/04	Compare to ground truth on 50 clinical vignettes	Evaluating ChatGPT by common and rare cases
Chu and Liu (2023)	Generation	CG	05/04	Analyzed by questionnaires	A psychological study of stories generated by ChatGPT

Table 8: The summary of our surveyed works (Part I). The date denotes the available online date in 2023. CG denotes ChatGPT. G4 denotes GPT-4.

Reference	Evaluation task	Version	Interesting findings
Cabrera and Neubig (2023)	Dialogue fluency	CG, G4	A chat-tuned model should be leveraged; As the context increases, the utility of prompt engineering became less prominent; Hallucinations and repetitive content
Bang et al. (2023)	Summ., Mach.T, Sent.A, QA, Misi.D	CG	ChatGPT had basic multimodality ability in drawing, albeit with remarkable improvements when provided with a self-generated textual description
Hendy et al. (2023)	Machine translation	CG	ChatGPT had competitive performance for high-resource languages but had limited capabilities for low-resource languages
Jiao et al. (2023)	Machine translation	CG, G4	ChatGPT performed badly on low resource or distant languages and had lower robustness. GPT-4 could narrow the gap
Karpinska and Iyyer (2023)	Machine translation	CG, G4	ChatGPT overwhelmingly outperformed Google Translation when doing paragraph-level translation if evaluated by expert translators
Qin et al. (2023)	NLP and reasoning tasks	CG	ChatGPT was good at NLP tasks with intensive reasoning, while weak in sequence tagging tasks; ChatGPT was weaker than GPT-3.5 in commonsense, symbolic, and logical reasoning tasks
Zhang et al. (2023a)	Summarisation	CG	An extract-then-generate pipeline with ChatGPT yielded significant performance improvements over abstractive baselines in terms of summary faithfulness
Soni and Wade (2023)	Summarisation	CG	While text classification algorithms could distinguish between real and generated summaries, humans were unable to distinguish between real summaries and those produced by ChatGPT
Jentzsch and Kersting (2023)	NLG	CG	ChatGPT likely generated repetitive jokes. Out of all the 1008 generated jokes, 90.1% of them matched one of the top 25 jokes
Amin et al. (2023)	13 affective computing tasks	CG, G4	GPTs excel in conventional sentiment-related tasks but struggle with implicit affective computing tasks requiring deeper psychological understanding
Kocoń et al. (2023)	25 sem. and prag. tasks	CG	More user context annotations led to better ChatGPT performance
Wei et al. (2023)	RE, NER, EE	CG	Multi-turn QA improved ChatGPT information extraction accuracy
Sun et al. (2023)	Document ranking	CG, G4	Permutation generation strategy was more supportive for ChatGPT and GPT-4 on ranking tasks
Wang et al. (2023c)	NLG evaluation	CG	The ChatGPT evaluator exhibited sensitivity towards prompts; The presence of inherent biases of evaluation datasets could potentially weaken ChatGPT evaluator
Liu et al. (2023c)	NLG evaluation	G4	The GPT-4 evaluator exhibited sensitivity towards prompts; LLM-based metrics preferred LLM-generated text to human-generated text
Kocmi and Federmann (2023)	Machine translation evaluation	CG	ChatGPT frequently yielded evaluation scores with explanations
Gilardi et al. (2023)	Text classification annotation	CG	ChatGPT yielded higher agreement rates than crowd-workers
Xu et al. (2023)	Multi-level language abilities	CG, G4	GPT-4 and ChatGPT achieved human-level accuracy on five basic language ability tests
Huang et al. (2023)	Multi-discipline knowledge in Chinese	CG, G4	CoT did not consistently enhance outcomes across various subjects, particularly in cases where the subjects did not heavily rely on reasoning abilities
Lai et al. (2023)	Multilingual NLP	CG	Despite processing tasks in other languages, English prompts helped ChatGPT achieve better results
Wang et al. (2023b)	Multilingual summarization	CG, G4	Certain multi-lingual and bilingual LLMs exhibited constrained zero-shot cross-lingual summarization capabilities
Borji (2023)	Reasoning, linguistics, and perception.	CG	It is necessary to use a standardized set of questions to continuously evaluate the progress of ChatGPT over time, thereby moving away from subjective opinions
Davis (2023)	Commonsense	CG	ChatGPT's commonsense understanding showed more flaws after "pressing on its weakness"
Gao et al. (2023)	Causal reasoning	CG, G4	Suffer causal hallucinations; Strong in causal interpretation, but weak in causal reasoning
Kıcıman et al. (2023)	Causal reasoning	CG, G4	ChatGPT and GPT-4 generally surpassed SOTA causal algorithms in graph discovery and counterfactual inference
Kosinski (2023)	Theory of Mind	CG, G4	GPT-3.5 solved 90% of false-belief tasks, at the level of seven-year-olds. GPT-4 solved nearly all the tasks
Moghaddam and Honey (2023)	Theory of Mind	CG, G4	Appropriate prompting enhanced LLM ToM reasoning, and they underscored the context-dependent nature of LLM cognitive capacities
Holterman and van Deemter (2023)	Theory of Mind	CG, G4	ChatGPT-4 outperformed ChatGPT-3 overall on six different ToM principles
Wang et al. (2023a)	Dialectical evaluation	CG	The belief (and disbelief) of ChatGPT lacked robustness and was susceptible to being perturbated by users
Liu et al. (2023b)	Reading comprehension and NLI	CG, G4	RoBERTa largely outperformed ChatGPT and GPT-4 on the MNLI dataset; GPT-4 exceeded the average level of human LSAT
Liu et al. (2023a)	NLI	CG, G4	Language ambiguity detection was a very difficult task for LLMs
Bubeck et al. (2023)	Coding, math, perception abilities	G4	GPT-4 achieved human-level accuracy, and it consistently outperformed previous models across various evaluation metrics
Li et al. (2023)	Text-to-SQL	CG	Execution accuracy was less than half that of humans; Wrong schema linking and misunderstanding database content
Bordt and von Luxburg (2023)	Math	CG	ChatGPT achieved a minimum passing score on the exam, while GPT-4 slightly surpassed the performance of average students
Frieder et al. (2023)	Math	CG	ChatGPT could not demonstrate consistent capability in providing reliable proofs or calculations in mathematics. It varied by the complexity of the mathematical questions
Collins et al. (2023)	Math	CG, G4	The GPTs have weaknesses in algebraic manipulations, tendency towards verbosity, and reliance on memorized solutions
Kortemeyer (2023)	Physics	CG	ChatGPT just passed the physics exam with a weakness in mathematics
Clark (2023)	Chemistry	CG	ChatGPT's chemistry knowledge was lower than the average on closed-response questions and ranked last in open-response questions
Dahlkemper et al. (2023)	Education	CG	ChatGPT's scientific accuracy was rated similarly to that of the expert solution in a hard question, despite the presence of misleading answers
Choi et al. (2023)	Law	CG	ChatGPT passed all four law classes, yet its performance ranked among the lowest within each class
Geerling et al. (2023)	Economics	CG	ChatGPT ranked top 9% and top 1% among thousands of college students in microeconomics and macroeconomics, respectively
Xie et al. (2023)	Finance	CG	ChatGPT's performance in multi-modal stock price prediction tasks was constrained, exhibiting lower effectiveness compared to conventional methods in single-modal tests
Gilson et al. (2023)	Medicine	CG	ChatGPT exhibited a performance that aligned with a third-year medical student
Kung et al. (2023)	Medicine	CG	ChatGPT achieved a minimum passing score on the exam
Antaki et al. (2023)	Medicine	CG	ChatGPT did not grasp specialized knowledge well in ophthalmology
Hirosawa et al. (2023)	Medicine	CG	ChatGPT's top 10 differential-diagnosis lists were acceptable, while the accuracy of its top 1 result significantly lagged behind that of physicians
Rao et al. (2023)	Medicine	CG	ChatGPT achieved higher accuracy on the final diagnoses, compared to the initial diagnoses
Mehnen et al. (2023)	Medicine	CG	GPT-4 needed to propose more than 8 suggestions to diagnose 90% rare cases
Chu and Liu (2023)	Generation	CG	ChatGPT wrote more engaging and persuasive short stories than their human counterparts. However, the conclusion was the opposite if the aim was long stories

Table 9: The summary of our surveyed works (Part II).