Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs

Simone Balloccu Patrícia Schmidtová Mateusz Lango Ondřej Dušek
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
{balloccu,schmidtova,lango,odusek}@ufal.mff.cuni.cz

Abstract

Natural Language Processing (NLP) research is increasingly focusing on the use of Large Language Models (LLMs), with some of the most popular ones being either fully or partially closed-source. The lack of access to model details, especially regarding training data, has repeatedly raised concerns about data contamination among researchers. Several attempts have been made to address this issue, but they are limited to anecdotal evidence and trial and error. Additionally, they overlook the problem of *indirect* data leaking, where models are iteratively improved by using data coming from users. In this work, we conduct the first systematic analysis of work using OpenAI's GPT-3.5 and GPT-4, the most prominently used LLMs today, in the context of data contamination. By analysing 255 papers and considering OpenAI's data usage policy, we extensively document the amount of data leaked to these models during the first year after the model's release. We report that these models have been globally exposed to ∼4.7M samples from 263 benchmarks. At the same time, we document a number of evaluation malpractices emerging in the reviewed papers, such as unfair or missing baseline comparisons and reproducibility issues. We release our results as a collaborative project on https://leak-llm.github.io/, where other researchers can contribute to our efforts.

1 Introduction

The recent emergence of large language models (LLMs), that show remarkable performance on a wide range of tasks, has led not only to a dramatic increase in their use in research but also to a growing number of companies joining the race for the biggest and most powerful models. In pursuing a competitive advantage, many popular LLMs today are locked behind API access and their details are unknown (OpenAI, 2023; Thoppilan et al., 2022; Touvron et al., 2023). This includes model weights (OpenAI, 2023), training data (Piktus et al., 2023), or infrastructural details to assess model carbon footprint (Lacoste et al., 2019).

In particular, the lack of information on training data raises important questions about the credibility of LLMs performance evaluation. The data from which these models learn, typically collected automatically by scraping documents from the web, may contain training, validation, and – most critically – test sets coming from NLP benchmarks. Because of this, researchers and stakeholders may later inadvertently evaluate LLMs on the same data they were trained on. This phenomenon, known as data contamination, may not be an issue in the general use of commercial LLMs, where adherence to research principles is not mandatory, but it becomes a serious problem when these models are widely used and evaluated in research.

Unfortunately, many proprietary models are locked behind inference-only APIs, making it hard to inspect data contamination. Because of this, existing work on the matter mostly focuses on detecting extreme forms of overfitting and memorization, such as the model's ability to generate benchmarks verbatim. These approaches are not only limited but also neglect that recent proprietary LLMs get iteratively improved from user interactions. If such interactions involve benchmark data (for example when researchers evaluate LLMs against baselines), the model may, in fact, become contaminated even if it was contamination-free during its initial training. We refer to this phenomenon as *indirect data leaking*.

In this paper, we address the issue of indirect data contamination in closed-source[1] LLMs by conducting a systematic literature review. We review 255 papers and carefully detail data leakage emerging from them. We focus primarily on the models

[1] In this paper we use the terms "proprietary" and "closed-source" interchangeably to refer to these models.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 67–93

accessible through OpenAI's ChatGPT,[2] (GPT-3.5 and GPT-4[3]) as these are the most frequently used commercial LLMs in NLP research. By considering OpenAI's data usage policy, we assess how much data was reported to be sent to the models in a way that it could be used for further training, hence giving the models an unfair advantage during evaluation. We also report a series of emergent evaluation malpractices, including lack of comparison with other approaches, differences in the evaluation scale (e.g., evaluating open models on entire benchmarks while comparing to proprietary LLMs evaluated on samples only), lack of code and data access, or data leakage even in situations where it could be avoided. To our knowledge, this work is the most comprehensive and extensive quantification of the data leakage issue in LLMs to date.

In short, our contributions are as follows:

(1) We systematically analyse 255 papers evaluating OpenAI's GPT-3.5 and GPT-4 on a variety of tasks in NLP and other domains (Section 4).

(2) For each paper, we estimate the amount of data leaked in such a way that it could be used for further model training. Overall, we conclude that $\sim$42% of the reviewed papers leaked data to GPT-3.5 and GPT-4, for a total of $\sim$4.7M benchmark samples across 263 benchmarks (Section 5.1).

(3) We further analyse the evaluation protocols of the selected papers, and we reveal some critical malpractices limiting both the experiments' reproducibility and fairness (Sections 5.2 and 5.3).

(4) Based on our findings, we propose a list of suggested practices for the evaluation of closed-source LLMs (Section 6).

We believe that our work can contribute to ongoing efforts on quantifying LLM data contamination by pointing out which datasets are worthy of further investigation. We release our survey results as a collaborative repository, in the form of a webpage at https://leak-llm.github.io/. It features a list of datasets, detailing the extend of data leakage for each of them. We invite other researchers to contribute any additional known leaks to the list.

[2]https://openai.com/blog/chatgpt

[3]https://openai.com/gpt-4

2 Prior Work on LLM Data Contamination

Work on LLMs data contamination traces back to OpenAI's GPT-3 (Brown et al., 2020; Magar and Schwartz, 2022), one of the first models with API-only access and limited training data disclosure. Despite results hinting at the presence of significant data contamination (Raffel et al., 2020; Magar and Schwartz, 2022), the model has been used extensively in research and the issue was rarely taken into account when interpreting its performance. With the release of ChatGPT and following closed-source models to general public,[4] the data contamination topic became an even more pressing issue.

When a model is closed-source, it becomes implicitly complex to assess data contamination from known benchmarks. Therefore, only few practical approaches have been proposed to investigate the issue.

One notable example is the LM Contamination Index,[5] featuring a regularly updated estimate of contamination for a list of both open and proprietary models. This approach works by zero-shot prompting the model to generate instances from specific datasets, providing details on the required split and format (Sainz et al., 2023). The premise is that no model should be able to replicate specific benchmark formats without having seen them first.

More applied approaches have been proposed recently (Golchin and Surdeanu, 2023), where LLMs are prompted to complete a given sentence coming from a known benchmark. The completion is then compared with the original reference through text overlap metrics and a statistical test is used to assess if the model is contaminated.

Although these preliminary works are promising, they cannot be fully trusted and have some limitations. Most importantly, they are based on an assessment of the model's ability to generate an example from the benchmark. The recall of such methods can be affected by two issues:

(1) Some closed-source models have incorporated special filters into their decoding algorithms that prevent them from generating texts that significantly overlap with their training sets (GitHub, 2022; Ippolito et al., 2023). This

[4]Including GPT-4 (OpenAI, 2023), Google's LaMDA (Thoppilan et al., 2022) and PaLM (Chowdhery et al., 2022), Cohere's Command and Anthropic's Claude.

[5]https://hitz-zentroa.github.io/lm-contamination/

creates an additional noise for the detection methods and results in the lack of confidence that even the datasets tested negative for data leakage are not present in LLM training data.

(2) Such approaches can only detect the most extreme form of overfitting which results in (almost) complete memorization of data samples by the model. However, even a regular adjustment of the model by training on the leaked data, which does not necessarily lead to its memorization, poses a problem for fair comparisons.

3 The Issue of Indirect Data Leaking

The related work presented in Section 2 approaches the issue of data contamination mainly by backtracking models' training data. It is commonly assumed that using benchmarks available only to authorised parties, or datasets being constructed after the ChatGPT release, is a guarantee that they have not been leaked. This ignores the fact that models using reinforcement learning from human feedback (RLHF, Ouyang et al., 2022), such as those used by ChatGPT, are subject to repeated updates (Aiyappa et al., 2023) with training data also coming from user interactions. This process leads to a previously overlooked phenomenon, where new data are leaked to the model just through using it. We refer to this problem as *indirect data leaking* and consider it a new development of the issue for two main reasons:

(1) Unlike plain text scraped from the internet, data from users might be harder to inspect for contamination as it might involve model prompts, textual alterations, or truncation of benchmark samples.

(2) Users supply the data along with instructions on how to perform the task. In LLMs, this can be considered a novel form of gold-standard data for continued training, even in the absence of target labels. Model updates on such data are likely much more effective than plain in-domain text.

The issue (1) is particularly complex to trace, even with a conscious and targeted effort by the LLM vendor. When evaluating a closed-source LLM, users often feed the model with test-set samples (with or without labels) surrounded by additional text, such as instructions in the form of prompts. In some cases, especially when evaluating the LLM robustness, the test-set samples are perturbed and hence no longer an exact match of their original version. Therefore, it is unlikely that LLM vendors could effectively exclude leaked benchmarks from further model fine-tuning, especially at scale. For (2), it would be necessary to understand how the LLM vendor uses the data to improve the model. A very likely scenario is continued pretraining, where the data leaked by users is treated as an in-domain corpus (and thus given more influence than pretraining data). This procedure is known to improve models' performances in the leaked domains (Gururangan et al., 2020). Notably, Shi and Lipani (2023) find that fine-tuning a model on in-domain text enriched by textual instructions leads to an increase in the model performance even if gold labels are not shown to the model. This setup perfectly matches the kind of data shown to chat LLMs when evaluated by researchers. This means that closed-source LLMs such as GPT-3.5 and GPT-4 can make use of these gold standard examples from widely used NLP benchmarks to gain an unfair advantage over other models.

We also point out that recent work (Aiyappa et al., 2023) showed that after model updates, ChatGPT performance improved on benchmarks to which it was previously exposed (Zhang et al., 2022). With these motivations, we conduct a systematic review to quantify how much of such data the models powering ChatGPT could have obtained.

4 Methodology

Following the standard systematic review protocol from the medical domain (Khan et al., 2003), we analyse the existing work on LLMs evaluation to inspect the issue of indirect data contamination and other evaluation malpractices. We focus on OpenAI's GPT-3.5 and GPT-4 models, as they are the most prominently used in recent NLP research. We organize our work into five macro-steps, corresponding to the following subsections.

4.1 Framing questions

In reviewing the existing work evaluating the performace of GPT-3.5 and GPT-4, we pose the following research questions:

(1) Which datasets have been demonstrably leaked to GPT-3.5 and GPT-4 during the last year?

(2) Do all papers evaluating these models include a fair comparison with existing baselines?

4.2 Identifying relevant work

We employ commonly used online databases[6] and major NLP conferences proceedings (including ACL, NAACL, EMNLP, NeurIPS), considering both peer-reviewed work and pre-prints, as the interaction with LLMs happened regardless of publication status. We filter our queries on work containing the terms "ChatGPT", "GPT-4", "GPT-3.5" "OpenAI" "evaluation", "large language models", "AI" either in title, abstract, body, or all of them.

We also do not limit our search to computer science works only, as recent LLMs have been investigated by researchers from many other domains, e.g. healthcare (Kung et al., 2023), psychology (Cai et al., 2023) and education (Szefer and Deshpande, 2023). Since the ChatGPT models are our primary focus, we limit our search to works between late November 2022 (when the first model was publicly released) and early October 2023. Among all the papers, we first do a preliminary screening, assessing if they effectively run GPT-3.5 or GPT-4 in any form.[7]

4.3 Assessing quality and relevance

To assess which work effectively leaked data to ChatGPT, we refer to OpenAI's data usage policy,[8], which explicitly mentions the use of users' data for model training:

> "[...] when you use our services for individuals such as ChatGPT or DALL-E, we may use your content to train our models [...]"

It also clarifies that the user data are not used for model training if sent via API and business services:

> "[...] we don't use content from our business offerings [...] and our API Platform to train our models [...]"

Therefore, only the work interacting with the models through the web interface[9] is considered to leak data. We note that while it is possible to opt out of providing the data for model improvement purposes,[21] we found no evidence suggesting any of the surveyed papers did so.

A small number of works used both the web interface and API access.[10] We carefully review such works to calculate which portion of the data was used in the former setup. We drew our conclusions from the paper draft history on arXiv; in some cases, this information was also transparently disclosed by the authors. In the case of work with multiple drafts dating before the model release in November 2022, we consider the earliest draft that includes GPT-3.5 or GPT-4 for the calculation.

4.4 Summarizing the evidence

We inspect each surveyed paper, looking for information on the used datasets, split, and number of samples. If no mention of sampling or similar information is made, we assume that the whole dataset has been used. Similarly, if no information on the used split is provided, we assume that the authors treated the dataset as a whole. It could be argued that feeding entire datasets to ChatGPT is unrealistic because of the usage restrictions imposed by OpenAI on the web interface, and the amount of work necessary for manually inputting the data inside the chat. However, we note that quickly after ChatGPT release, many unofficial wrappers have been developed[11] for circumventing said issues, most of which are still in active use. We also point out that many of the papers we surveyed mentioned the use of such tools explicitly.

We also track secondary information relevant to the evaluation – for each work, we inspect: (1) if it has been peer-reviewed;[12] (2) if the used prompts are available; (3) if a repository to reproduce the experiment is provided; (4) if the authors used a whole dataset or a sample; (5) if GPT-3.5 or GPT-4 were compared to other open models/approaches and if the evaluation scale was the same; (6) if the version of the model used is reported.

4.5 Interpreting the findings

We report the results of our review both quantitatively and qualitatively. Specifically, we report the number of works surveyed leaking data to GPT-

[6] We query Google Scholar, Semantic Scholar, DBLP, arXiV, ACL Anthology.

[7] We encountered a small number of papers also comparing to other closed-source LLMs, such as Anthropic's Claude.

[8] https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

[9] https://chat.openai.com/

[10] Their experiments began prior to March 1st, 2023 and the authors started using the API soon after it was released.

[11] E.g. revChatGPT, PyChatGPT, and ChatGPT-to-API.

[12] We do note that part of the work we reviewed might still be under review, also see Footnote 15.

3.5 or GPT-4 in such a way that it can be used by OpenAI to further improve the model (according to their data policy). In this paper we do not distinguish between works leaking data to GPT-3.5, GPT-4, or both. This is because indirect data leaking is caused by browser access, where both models are available through the ChatGPT Plus subscription. We also note that OpenAI confirmed that creating GPT-4 involved the use of ChatGPT to some extent.[13] For this reason, we estimate the data leakage to be effectively shared across the two models and for simplicity, we refer to both models as "ChatGPT" from now on.

We also document a series of evaluation practices emerging for the work reviewed that is problematic with respect to objectiveness and reproducibility. Finally, drawing upon our results, we present a series of best practices for researchers evaluating OpenAI's and other closed-source LLMs.

5 Results

Following our methodology, in the first step we identified 255 research papers, 212 of which were found relevant[14] during the initial screening (see Sec. 4.2). Among the relevant papers, 70 ($\sim 32\%$) were peer-reviewed, while the remainder (142) consisted of pre-prints.[15] We subsequently analysed the retrieved papers to examine the problem of data contamination and the adopted evaluation practices.

5.1 Indirect data contamination

From our analysis, 90 papers ($\sim 42\%$) accessed ChatGPT through the web interface, hence providing data that OpenAI could have used to further improve its models.

We first inspected the time distribution of the reviewed works (Figure 1) to gain insight into when most data leaks happened. Unsurprisingly, the majority of the papers leaking data dates before the official release of ChatGPT API, and it can be seen

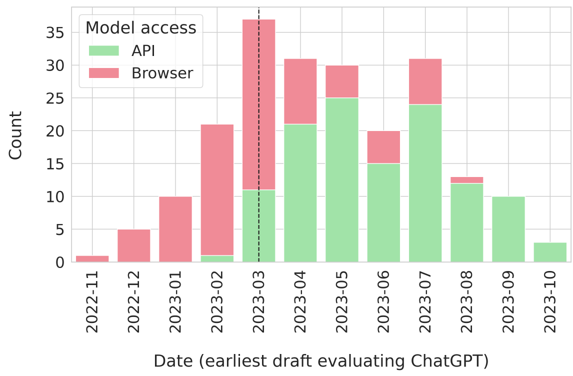

Figure 1: Distribution of the dates when papers evaluating ChatGPT were first uploaded to arXiv or published. The dotted line represents the ChatGPT API release (March 1st, 2023, dotted line in the chart) as a cutoff point. The single paper shown using the API in February is by a research group that reported having early API access.

that web interface access rapidly decreased following March 2023. However, we must note that (1) a considerable amount of work kept using the web interface to access ChatGPT until September 2023 and (2) our analysis cannot inspect the preliminary stages of prompt engineering, which are rarely reported and might still be done through the web interface because of its trial-and-error nature.

The presence of leaked data after the API release may indicate that a part of the research community is either unaware of OpenAI's data policy, or does not consider it a problem when conducting experiments. Many works, especially small case studies, also reported using the web interface for cost reasons, as it allows free access to the models.

As a second step, we quantified leak severity per dataset and split. For work specifying the amount of data used (either in the paper or through a repository), we consider the given value. For the rest, we calculate it by inspecting the actual dataset.[16] In seven papers, no number of samples used was specified, so we contacted the authors for clarification. In the two cases where the authors did not respond, we assumed the entire split of a dataset was used. We calculated both the number of instances and the percentage of the considered split (or the whole dataset when applicable).

Since a small number of datasets (18) was used in multiple papers in different amounts, we had to consider whether these should be interpreted as

[13] https://openai.com/research/gpt-4

[14] The excluded papers either were opinion pieces that minimally tested ChatGPT on certain tasks, or did not include any evaluation.

[15] We note that, during this paper's review period, 43 of the pre-prints were peer-reviewed and published. However, some of the relevant proceedings have not been released yet, making it impossible to consistently check for paper updates. We cannot rule out that some of these works leaked more data with further experiments, or addressed some evaluation malpractices.

[16] We mainly use HuggingFace Datasets, but also refer to Kaggle or other sources based on availability.

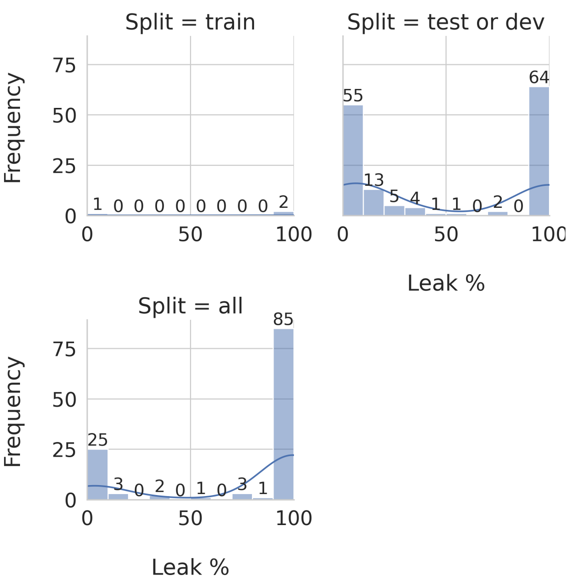

Figure 2: Data leakage distribution. We report the number of times (y) we observed a specific percentage of leaking (x) for the considered split. As some work vaguely describes the used split as "test or dev set", we merge these two values in a unique chart.

Task name	Lo	M-Lo	M-Hi	Hi
AI safety & ethics	0	0	2	0
Creative NLG	1	0	0	0
Dialogue	2	1	0	5
NLG evaluation	0	0	0	4
Machine Translation	6	4	1	1
Math	0	1	0	8
Natural language generation	2	1	0	14
Natural language inference	6	2	0	15
Language understanding	0	0	0	2
Paraphrasing	2	0	0	0
Politics	0	1	0	3
Programming	0	0	0	1
Psychology	0	0	0	1
Question answering	24	14	5	31
Commonsense reasoning	3	4	0	9
Semantic similarity	2	1	0	3
Sentiment analysis	8	9	1	8
Summarization	5	6	1	1
Text classification	1	0	0	3
Text extraction	2	1	0	7

Table 1: The number of datasets with low (Lo), moderate-low (M-Lo), moderate-high (M-Hi) and high leak severity (Hi) is reported for each task, omitting custom datasets. A more detailed table, including specific dataset names, is provided in the Appendix C.

individual separate leaks (that should be summed up) or not. We were not able to verify this from the provided data, so we adopted an "optimistic" approach and assumed that the largest leak for a given dataset is always a superset of all smaller ones.[17]

Our calculations show that the 90 papers leaked data from 263 unique datasets, for a total of over 4.7M samples (see Tables 4 to 6 in the Appendix).[18]

We find most samples ($\sim 93.8\%$) coming from datasets treated as whole (with no split), followed by test and development ($\sim 5.6\%$),[19] and training ($\sim 0.6\%$) sets. In line with what we discussed in Section 3, we can conclude that ChatGPT was exposed to millions of benchmark samples, enriched with instructions that could be considered de-facto novel gold-standard data in some cases.

We also report that several works included the examples' labels when few-shot prompting ChatGPT or using it as a reference-based evaluation metric. We consider this the worst possible case of data leaking, as it gives the model information about the desired output as well.

To classify leak severity, we examine the frequency distribution of leak sizes (Figure 2). It appears that most works either leak full splits or very small samples, with only a few works leaking intermediate amounts. With this information, we classify a portion of leaked data as *low* ($< 5\%$), *moderate-low* ($5 - 50\%$), *moderate-high* ($50 - 95\%$), or *high* ($> 95\%$).

Consequently, we categorize all leaked datasets into these 4 thresholds. Overall, we find a low leak for 66 ($\sim 25\%$) datasets, moderate-low for 47 ($\sim 18\%$), moderate-high for 10 ($\sim 4\%$) and high for 142 ($\sim 53\%$). This result is particularly worrying as the majority of datasets were almost completely leaked.

Finally, we inspect which NLP tasks are covered by the leaked data (Table 1). We find that the tasks suffering the most from high leaks are natural language inference, question answering, and natural language generation. These and other tasks include many highly popular NLP benchmarks, as well as high-quality custom datasets created ad-hoc for individual evaluations (see Tables 4 to 6 in the Appendix). To name a few, almost the entire test sets from Semeval2016 Task 6 (Mohammad et al., 2016), SAMSum (Gliwa et al., 2019), and MultiWOZ 2.4 (Ye et al., 2022) are leaked. The custom datasets were frequently phrased as an

[17]We also tried a pessimistic approach, where we assumed all the leaks were independent, but due to the small number of works covering the same data, the results are virtually identical.

[18]The survey total is 4,714,753 leaked samples.

[19]As some work vaguely describes the used split as "test or dev set", we merge these two values.

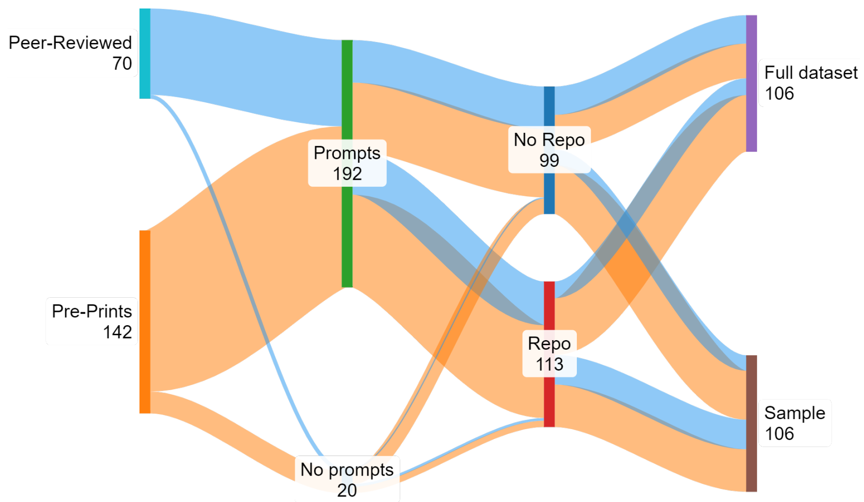

Figure 3: Evaluation reproducibility. Through the above Sankey diagram, we report facilitators and barriers to reproducing the carried-out experiments. This includes providing the used prompts, a repository with usable code and the use of sampling.

exam in a field different from NLP, e.g., medicine, physics, psychology, or law. Other custom datasets explored, for example, the LLMs' sense of humour, philosophical and political leaning, or bias. We note that not all the leaked custom datasets have been publicly released. This makes the leak even more severe, as it potentially makes OpenAI the only organisation (besides the authors) with access to such data.

5.2 Reproducibility

We assess the evaluations' reproducibility by checking whether the prompts used to query ChatGPT were provided, whether a repository containing data or code was available, and whether the datasets used were custom-made. Finally, we also check for sampling of the original data or other practices that make it impossible to exactly reconstruct the data used.

From our results (Figure 3), 192 ($\sim$ 91%) works report the prompts used to convert data into a query and possibly to instruct the model on how to perform a given task. The number of works providing a code repository is significantly smaller, at 113 ($\sim$ 53%). This figure excludes papers that provided a link to a non-existent or empty repository. Overall, 72 ($\sim$ 51%) of the pre-prints and 34 ($\sim$ 48%) peer-reviewed papers provided both prompts and a repository. We report further details on this data in Appendix B.

Another barrier to reproducibility is that most closed-source LLMs are being regularly updated.

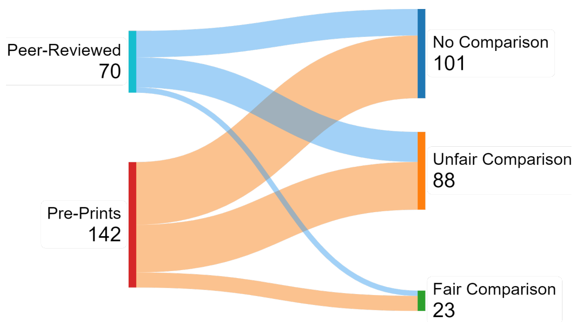

Figure 4: Evaluation fairness. Through the above Sankey diagram, we report whether the proprietary LLMs were compared against other models, and if the comparison was equal. In this context, "Unfair" comparison refers to evaluating different models on different amounts of data.

Therefore, it is crucial to report the used model version, as different versions may lead to significantly different outputs (Chen et al., 2023b). In the surveyed works, this was generally done by reporting the running period of the experiments when using the web interface, or by reporting which version of the model has been accessed via the API. Unfortunately, as regular model updates are a relatively new concept, this practice is not yet common. Only 29 (40%) of the peer-reviewed papers and 33 (23%) of the pre-prints provide this information.

5.3 Evaluation fairness

We find the evaluation of ChatGPT's performance to be often unfair. First, comparison to any open-source LLM or non-LLM-based method may be missing. Our results (Figure 4) show that this is similarly prevalent regardless of the publication status, appearing in 71 ($\sim 50\%$) of pre-prints and 30 ($\sim 43\%$) of published papers. Second, when a comparison with open models and baselines is made, 54 pre-prints ($\sim 38\%$) and 34 peer-reviewed ($\sim 49\%$) papers compare the results computed on different samples. ChatGPT is typically evaluated on a random sample of the benchmark while other models are compared on its entirety. In many works, ChatGPT's performance is measured on only a handful (10-50) of examples, which substantially lowers the expressive power of the comparison. For instance, considering a simplistic case with binary assessment of model output (correct/incorrect) on 10 examples, the difference should be more than 30% to be statistically significant,[20] which is rarely seen. Statistical analysis of results is almost never performed. We report further details on evaluation fairness in Appendix B.

Another concerning practice is how the size of the evaluation data is reported, especially when sampling is used. We find that papers often show the size of the whole evaluation dataset upfront (e.g. in a table or in the dataset description section), but they report the actual sample sizes used for evaluation only later and in a less obvious way (in footnotes, limitations sections, or appendices). This practice makes the experimental results harder to interpret.

6 Suggested Practices in Closed-source LLM Evaluation

Our survey revealed both a significant amount of data leakage in ChatGPT and many worrying trends in its evaluation. In light of this, we list a series of suggested practices that we believe could help mitigate the issues. We believe that researchers looking to objectively evaluate LLMs today should:

Access the model in a way that does not leak data The first step when planning proprietary LLMs evaluation should be reading their most up-to-date data policies, and access models accordingly (e.g. API instead of web interface for OpenAI's LLMs). We also acknowledge that in some cases this might not be viable due to budget limits, or an overly steep learning curve for the use of APIs by researchers outside of computer science.[21]

Interpret performance with caution The lack of system specifications and training details can make proprietary LLMs look like incredibly powerful tools with impressive zero-shot performance. This can often be explained by data contamination (Aiyappa et al., 2023). In our review, we documented that over 4 million samples across more than 200 NLP datasets have been leaked to these models. The performance of closed-source LLMs should always be interpreted while keeping these results in mind.

When possible, avoid using closed-source models We strongly encourage using the available open-source LLMs. While there has been discussion in the research community about proprietary models being consistently better than open-source ones, we note that (1) this is often driven by hype, while there is evidence of the opposite (Kocoń et al., 2023), (2) research done solely on closed LLMs limits scientific progress, bringing benefits mainly to the LLM vendors and (3) LLM vendors can arbitrarily make changes to the models, e.g., making previous versions unavailable, changing their behaviour in a way that may not be visible to the user (Chen et al., 2023b) or changing the data treatment policy.

Adopt a fair and objective comparison Evaluating closed-source LLMs is tied to comparing them with pre-existing approaches. Evaluating proprietary models on a limited number of samples while evaluating open ones on dramatically larger sets is scientifically dubious at best. When sampling is required (for example because of budgetary restrictions), it should be applied to all the considered approaches. We also discourage taking state-of-the-art values directly from previous work and suggest to re-run all approaches on the considered data only.

Make the evaluation reproducible In light of the known NLP evaluation reproducibility crisis (Belz et al., 2023; Thomson et al., 2024) we strongly encourage researchers to report as many details about their setup. Besides all the relevant details about the setup for reproducibility, such as random seeds, open model parameters, etc., we

[20]Assuming Fisher's exact test, typical $\alpha = 5\%$ and moderate model performance around $\hat{p} = 0.5$

[21]In such case, as of January 2024, OpenAI allows users to opt out of providing data for model improvement through the OpenAI Privacy Request Portal.

note that when the evaluation involves closed models, additional details should be disclosed. Prompts, as well as the process leading to them, should be detailed since LLMs are very sensitive to even minor changes in prompts (Lu et al., 2022). The model version and experiment running period should be mentioned as well so that further researchers can use the same model checkpoint if possible. Data, especially if sampled, should be released (ideally in a repository) to avoid potential differences in sampling.

Report indirect data leaking Indirect data leaking is a serious issue, and when it happens it should be reported. Clear information on which benchmarks have been leaked benefits research, helps other researchers orient their experiments, and ultimately leads to a more objective evaluation of proprietary LLMs. We invite all researchers to contribute to our collaborative project at https://leak-llm.github.io/.

7 Conclusion and Future Work

In this work, we present our findings based on the analysis of 255 papers evaluating the performance of GPT-3.5 and GPT-4. We investigate the problem of indirect data contamination and report that 4.7M samples coming from 263 distinct datasets have been exposed to the models in such a way that this data could be used for training by OpenAI. We also report concerning research practices with respect to reproducibility and fairness. Finally, informed by our analysis, we detailed some suggested practices for the evaluation of closed-source LLMs.

Future Work In our future work, we aim to run experiments via the OpenAI API to see the impact of leaked test data on the performance of GPT-3.5 and GPT-4 on the leaked datasets and the tasks in general.

Furthermore, we consider investigating indirect data leakage in other closed-source models, namely from Anthropic or Cohere, which appeared in a small number of papers reviewed in this work.

Limitations

We are aware the list of contaminated datasets we compiled in our work is not fully conclusive for one of several reasons:

(1) We review the information that has been publicly revealed via articles. We postulate more experiments could have revealed test set data to closed-source models but were never published.

(2) In this paper, we focus on the works that use ChatGPT or GPT-4. However, prior to March 1st, 2023, OpenAI's policy stated that they may also use data from the API to improve their models. This would imply that data sent to GPT-3 via the API could have been used for training.

(3) The number of papers investigating the performance of ChatGPT is vast, and despite our best efforts, we could have missed some works.

(4) Information on whether individual works are pre-prints or published is given at the time of writing (early October 2023). This is subject to change, especially given the freshness of many of the works reviewed.

(5) Many datasets released prior to 2021 could have been fully leaked by being a part of the models' pre-training data.

As mentioned in Section 4, in some cases the papers were not clear about some aspects of the experiments. We contacted the authors of such papers for clarification, however, two of them did not respond. Therefore, our best-judgment assumptions may be wrong for these papers.

Acknowledgements

This research was supported by the European Research Council (Grant agreement No. 101039303 NG-NLG) and by Charles University project SVV 260 698. Patrícia Schmidtová was also supported by the Women in Quant Finance Network grant awarded by G-Research. We would also like to thank Zdeněk Kasner and Dominik Macháček for their valuable feedback on the manuscript.

References

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yongyeol Ahn. 2023. Can we trust the evaluation on ChatGPT? In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 47–54, Toronto, Canada. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zhenguang G Cai, David A Haslett, Xufeng Duan, Shuqi Wang, and Martin J Pickering. 2023. Does ChatGPT resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv:2204.02311 [cs]*. ArXiv: 2204.02311.

GitHub. 2022. About github copilot. https://github.com/features/copilot.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in LLMs: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.

Khalid S Khan, Regina Kunz, Jos Kleijnen, and Gerd Antes. 2003. Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, 96(3):118–121.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

TH Kung, M Cheatham, A Medenilla, C Sillos, L De Leon, C Elepaño, M Madriaga, R Aggabao, G Diaz-Candido, J Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *Plos Digit Health*, 2:000198.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016.

SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The ROOTS search tool: Data transparency for LLMs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 304–314, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023. Did ChatGPT cheat on your test? https://hitz-zentroa.github.io/lm-contamination/blog/.

Zhengxaing Shi and Aldo Lipani. 2023. Don't stop pretraining? make prompt-based fine-tuning powerful learner. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jakub Szefer and Sanjay Deshpande. 2023. Analyzing chatgpt's aptitude in an introductory computer engineering course. *arXiv preprint arXiv:2304.06122*.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*, pages 1–10.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK. Association for Computational Linguistics.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

A Full list of the reviewed work

In this section, we list all the work that we reviewed and classified as relevant.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI.

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

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. 2023. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7716–7733, Toronto, Canada. Association for Computational Linguistics.

Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023. Will affective computing emerge from foundation models and general AI? a first evaluation on ChatGPT.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models.

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*, 3(4):100324.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023a. Longbench: A bilingual, multitask benchmark for long context understanding.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023b. Benchmarking foundation models with language-model-as-an-examiner.

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

Jonas Belouadi and Steffen Eger. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. ChatGPT is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models.

Sebastian Bordt and Ulrike von Luxburg. 2023. ChatGPT participates in a computer science exam.

Ali Borji. 2023. A categorical archive of ChatGPT failures.

Ritwik Bose, Ian Perera, and Bonnie Dorr. 2023. Detoxifying online discourse: A guided response generation approach for reducing toxicity in user-generated text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 9–14, Toronto, Canada. Association for Computational Linguistics.

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.

Ana-Maria Bucur. 2023. Utilizing ChatGPT generated data to retrieve depression symptoms from social media.

Laura Cabello, Jiaang Li, and Ilias Chalkidis. 2023. Pokemonchat: Auditing ChatGPT for pokémon universe knowledge.

Alex Cabrera and Graham Neubig. 2023. Zeno chatbot report.

Zhenguang G Cai, David A Haslett, Xufeng Duan, Shuqi Wang, and Martin J Pickering. 2023. Does ChatGPT resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models.

Mohna Chakraborty, Adithya Kulkarni, and Qi Li. 2023. Zero-shot approach to overcome perturbation sensitivity of prompts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5698–5711, Toronto, Canada. Association for Computational Linguistics.

Shreya Chandrasekhar, Chieh-Yang Huang, and Ting-Hao Huang. 2023. Good data, large data, or no data? comparing three approaches in developing research aspect classifiers for biomedical papers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 103–113, Toronto, Canada. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. 2023a. Skills-in-context prompting: Unlocking compositionality in large language models.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023b. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023c. Large language models meet Harry Potter: A bilingual dataset for aligning dialogue agents with characters.

Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023d. How robust is GPT-3.5 to predecessors? a comprehensive study on language understanding tasks.

Yanran Chen and Steffen Eger. 2022. Transformers go for the lols: Generating (humourous) titles from scientific abstracts end-to-end.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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

Mohita Chowdhury, Ernest Lim, Aisling Higham, Rory McKinnon, Nikoletta Ventoura, Yajie He, and Nick De Pennington. 2023. Can large language models safely address patient questions following cataract surgery? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 131–137, Toronto, Canada. Association for Computational Linguistics.

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*, 100(5):1905–1916.

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.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023a. AugGPT: Leveraging ChatGPT for text data augmentation.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023b. Uncovering ChatGPT's capabilities in recommender systems.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating ChatGPT's performance for multilingual and emoji-based hate speech detection.

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey.

Sanjay Deshpande and Jakub Szefer. 2023. Analyzing ChatGPT's aptitude in an introductory computer engineering course.

Sifatkaur Dhingra, Manmeet Singh, Vaisakh SB, Neetiraj Malviya, and Sukhpal Singh Gill. 2023. Mind meets machine: Unravelling GPT-4's cognitive psychology.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

George Duenas, Sergio Jimenez, and Geral Mateus Ferro. 2023. You've got a friend in ... a language model? a comparison of explanations of multiple-choice items of reading comprehension between ChatGPT and humans. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 372–381, Toronto, Canada. Association for Computational Linguistics.

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. GPT-3.5, GPT-4, or bard? evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts.

Yaxin Fan and Feng Jiang. 2023. Uncovering the potential of ChatGPT for discourse analysis in dialogue: An empirical study.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is ChatGPT a highly fluent grammatical error correction system? a comprehensive evaluation.

Sarah E. Finch, Ellie S. Paek, and Jinho D. Choi. 2023. Leveraging large language models for automated dialogue analysis. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 202–215, Prague, Czechia. Association for Computational Linguistics.

Kathleen Fraser, Svetlana Kiritchenko, Isar Nejadgholi, and Anna Kerkhof. 2023. What makes a good counter-stereotype? Evaluating strategies for automated responses to stereotypical text. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 25–38, Toronto, Canada. Association for Computational Linguistics.

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

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023a. Is ChatGPT a good causal reasoner? a comprehensive evaluation.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023b. Exploring the feasibility of ChatGPT for event extraction.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023c. Human-like summarization evaluation with ChatGPT.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023d. Enabling large language models to generate text with citations.

Yuan Gao, Ruili Wang, and Feng Hou. 2023e. How to design translation prompts for ChatGPT: An empirical study.

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*, 68(2):233–245.

Omid Ghahroodi, Seyed Arshan Dalili, Sahel Mesforoush, and Ehsaneddin Asgari. 2023. SUT at SemEval-2023 task 1: Prompt generation for visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2160–2163, Toronto, Canada. Association for Computational Linguistics.

Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2023. ChatGPT for suicide risk assessment on social media: Quantitative evaluation of model performance, potentials and limitations.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks.

Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. 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 Med Educ*, 9.

Github. 2023. Evaluation papers for ChatGPT.

Shahriar Golchin and Mihai Surdeanu. 2023. Time travel in LLMs: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*.

Srinivas Gowriraj, Soham Dinesh Tiwari, Mitali Potnis, Srijan Bansal, Teruko Mitamura, and Eric Nyberg. 2023. Language-agnostic transformers and assessing ChatGPT-based query rewriting for multilingual document-grounded QA. In *Proceedings of the Third DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 101–108, Toronto, Canada. Association for Computational Linguistics.

Wenshi Gu. 2023. Linguistically informed ChatGPT prompts to enhance japanese-chinese machine translation: A case study on attributive clauses.

Reto Gubelmann, Aikaterini-lida Kalouli, Christina Niklaus, and Siegfried Handschuh. 2023. When truth matters - addressing pragmatic categories in natural language inference (NLI) by large language models (LLMs). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 24–39, Toronto, Canada. Association for Computational Linguistics.

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

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Tahsina Hashem, Weiqing Wang, Derry Tanti Wijaya, Mohammed Eunus Ali, and Yuan-Fang Li. 2023. Generating faithful text from a knowledge graph with noisy reference text. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 106–122, Prague, Czechia. Association for Computational Linguistics.

Shreya Havaldar, Bhumika Singhal, Sunny Rai, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. 2023. Multilingual language models are not multicultural: A case study in emotion. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 202–214, Toronto, Canada. Association for Computational Linguistics.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023a. ICL-D3IE: In-context learning with diverse demonstrations updating for document information extraction.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, Haoning Ye, Zihan Li, Shisong Chen, Yikai Zhang, Zhouhong Gu, Jiaqing Liang, and Yanghua Xiao. 2023b. Can large language models understand real-world complex instructions?

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023c. MGTBench: Benchmarking machine-generated text detection.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.

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.

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).

Bart Holterman and Kees van Deemter. 2023. Does ChatGPT have theory of mind?

Ari Holtzman, Peter West, and Luke Zettlemoyer. 2023. Generative models as a complex systems science: How can we make sense of large language model behavior?

Ruixin Hong, Hongming Zhang, Hong Zhao, Dong Yu, and Changshui Zhang. 2023. Faithful question answering with Monte-Carlo planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3944–3965, Toronto, Canada. Association for Computational Linguistics.

Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. 2023a. Chain-of-symbol prompting elicits planning in large langauge models.

Nan Hu, Yike Wu, Guilin Qi, Dehai Min, Jiaoyan Chen, Jeff Z. Pan, and Zafar Ali. 2023b. An empirical study of pre-trained language models in simple knowledge graph question answering.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023c. Zero-shot clinical entity recognition using ChatGPT.

Fan Huang, Haewoon Kwak, and Jisun An. 2023a. Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023b. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023c. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023d. Look before you leap: An exploratory study of uncertainty measurement for large language models.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023e. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of ChatGPT.

Sophie Jentzsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 325–340, Toronto, Canada. Association for Computational Linguistics.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. StructGPT: A general framework for large language model to reason over structured data.

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.

Ana Jojic, Zhen Wang, and Nebojsa Jojic. 2023. Gpt is becoming a turing machine: Here are some ways to program it.

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

David Kartchner, Selvi Ramalingam, Irfan Al-Hussaini, Olivia Kronick, and Cassie Mitchell. 2023. Zero-shot information extraction for clinical meta-analysis using large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 396–405, Toronto, Canada. Association for Computational Linguistics.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating GPT-4 and ChatGPT on japanese medical licensing examinations.

Yuheun Kim, Lu Guo, Bei Yu, and Yingya Li. 2023. Can ChatGPT understand causal language in science claims? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 379–389, Toronto, Canada. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Nam Ho Koh, Joseph Plata, and Joyce Chai. 2023. Bad: Bias detection for large language models in the context of candidate screening.

Philipp Koralus and Vincent Wang-Maścianica. 2023. Humans in humans out: On GPT converging toward common sense in both success and failure.

Gerd Kortemeyer. 2023. Could an artificial-intelligence agent pass an introductory physics course?

Michal Kosinski. 2023. Theory of mind might have spontaneously emerged in large language models.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language generation models can cause harm: So what can we do about it? an actionable survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3299–3321, Dubrovnik, Croatia. Association for Computational Linguistics.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2023. Performance of ChatGPT on usmle: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):1–12.

Mu-Tien Kuo, Chih-Chung Hsueh, and Richard Tzong-Han Tsai. 2023. Large language models on the chessboard: A study on ChatGPT's formal language comprehension and complex reasoning skills.

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

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.

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.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2023. ChatGPT: A meta-analysis after 2.5 months.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023b. ChatHaruhi: Reviving anime character in reality via large language model.

Jinyang Li, Binyuan Hui, Ge Qu, Binhua Li, Jiaxi Yang, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C. C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023c. Can LLM already serve as a database interface? a big bench for large-scale database grounded text-to-sqls.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023d. HaluEval: A large-scale hallucination evaluation benchmark for large language models.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2023e. "hot" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2023f. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023g. Guiding large language models via directional stimulus prompting.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023a. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Yancheng Liang, Jiajie Zhang, Hui Li, Xiaochen Liu, Yi Hu, Yong Wu, Jiaoyao Zhang, Yongyan Liu, and Yi Wu. 2023b. Breaking the bank with ChatGPT: Few-shot text classification for finance. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 74–80, Macao. -.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S. Yu. 2023a. A comprehensive evaluation of ChatGPT's zero-shot text-to-sql capability.

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

Chang Liu and Bo Wu. 2023. Evaluating large language models on graphs: Performance insights and comparative analysis.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023c. Evaluating the logical reasoning ability of ChatGPT and GPT-4.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023d. Logicot: Logical chain-of-thought instruction-tuning data collection with GPT-4.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023e. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation.

Xin Liu, Yuan Tan, Zhenghang Xiao, Jianwei Zhuge, and Rui Zhou. 2023f. Not the end of story: An evaluation of ChatGPT-driven vulnerability description mappings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3724–3731, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023g. G-Eval: NLG evaluation using GPT-4 with better human alignment.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023h. Summary of ChatGPT-related research and perspective towards the future of large language models.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023i. MolXPT: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1606–1616, Toronto, Canada. Association for Computational Linguistics.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023j. Deid-GPT: Zero-shot medical text de-identification by GPT-4.

Zhexiong Liu, Diane Litman, Elaine Wang, Lindsay Matsumura, and Richard Correnti. 2023k. Predicting the quality of revisions in argumentative writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 275–287, Toronto, Canada. Association for Computational Linguistics.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada. Association for Computational Linguistics.

Guang Lu, Sylvia B. Larcher, and Tu Tran. 2023a. Hybrid long document summarization using c2f-far and ChatGPT: A practical study.

Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. 2023b. Toward human-like evaluation for natural language generation with error analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5892–5907, Toronto, Canada. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023c. Error analysis prompting enables human-like translation evaluation in large language models: A case study on ChatGPT.

Xin Lu, Zhuojun Li, Yanpeng Tong, Yanyan Zhao, and Bing Qin. 2023d. HIT-SCIR at WASSA 2023: Empathy and emotion analysis at the utterance-level and the essay-level. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity,* *Sentiment, & Social Media Analysis*, pages 574–580, Toronto, Canada. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. ChatGPT as a factual inconsistency evaluator for text summarization.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, Yinglong Xia, and Jiebo Luo. 2023a. LLM-Rec: Personalized recommendation via prompting large language models.

Qing Lyu, Josh Tan, Michael E. Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J. Myers, Ge Wang, and Christopher T. Whitlow. 2023b. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: Promising results, limitations, and potential.

Paula Maddigan and Teo Susnjak. 2023. Chat2vis: Generating data visualisations via natural language using ChatGPT, codex and GPT-3 large language models.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of ChatGPT and GPT-4.

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*.

Andrianos Michail, Stefanos Konstantinou, and Simon Clematide. 2023. Uzh_clyp at semeval-2023 task 9: Head-first fine-tuning and ChatGPT data generation for cross-lingual learning in tweet intimacy prediction.

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

Robert Morabito, Jad Kabbara, and Ali Emami. 2023. Debiasing should be good and bad: Measuring the consistency of debiasing techniques in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4581–4597, Toronto, Canada. Association for Computational Linguistics.

Vishvak Murahari, Ameet Deshpande, Carlos Jimenez, Izhak Shafran, Mingqiu Wang, Yuan Cao, and Karthik Narasimhan. 2023. MUX-PLMs: Pretraining language models with data multiplexing. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 196–211, Toronto, Canada. Association for Computational Linguistics.

Duan Nan. 2023. Frontier review of multimodal AI. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 110–118, Harbin, China. Chinese Information Processing Society of China.

Andrew Nedilko. 2023. Generative pretrained transformers for emotion detection in a code-switching setting. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 616–620, Toronto, Canada. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems.

Shinhyeok Oh, Hyojun Go, Hyeongdon Moon, Yunsung Lee, Myeongho Jeong, Hyun Seung Lee, and Seungtaek Choi. 2023. Evaluation of question generation needs more references. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6358–6367, Toronto, Canada. Association for Computational Linguistics.

Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. ChatGPT versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots.

Amin Omidvar and Aijun An. 2023. Empowering conversational agents using semantic in-context learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 766–771, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

Miguel Ortega-Martín, Óscar García-Sierra, Alfonso Ardoiz, Jorge Álvarez, Juan Carlos Armenteros, and Adrián Alonso. 2023. Linguistic ambiguity analysis in ChatGPT.

Lidiia Ostyakova, Veronika Smilga, Kseniia Petukhova, Maria Molchanova, and Daniel Kornev. 2023. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 242–254, Prague, Czechia. Association for Computational Linguistics.

Naoki Otani, Jun Araki, HyeongSik Kim, and Eduard Hovy. 2023. On the underspecification of situations in open-domain conversational datasets. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 12–28, Toronto, Canada. Association for Computational Linguistics.

Wenbo Pan, Qiguang Chen, Xiao Xu, Wanxiang Che, and Libo Qin. 2023. A preliminary evaluation of ChatGPT for zero-shot dialogue understanding.

Ralph Peeters and Christian Bizer. 2023. Using ChatGPT for entity matching.

Alessandro Pegoraro, Kavita Kumari, Hossein Fereidooni, and Ahmad-Reza Sadeghi. 2023. To ChatGPT, or not to ChatGPT: That is the question!

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023a. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023b. Instruction tuning with GPT-4.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023c. Towards making the most of ChatGPT for machine translation.

Denis Peskoff and Brandon Stewart. 2023. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, Toronto, Canada. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions.

Andrei Popescu-Belis, Àlex R. Atrio, Bastien Bernath, Etienne Boisson, Teo Ferrari, Xavier Theimer-lienhard, and Giorgos Vernikos. 2023. GPoeT: a language model trained for rhyme generation on synthetic data. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–20, Dubrovnik, Croatia. Association for Computational Linguistics.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver?

Ali Quidwai, Chunhui Li, and Parijat Dube. 2023. Beyond black box AI generated plagiarism detection: From sentence to document level. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 727–735, Toronto, Canada. Association for Computational Linguistics.

Abhiramon Rajasekharan, Yankai Zeng, Parth Padalkar, and Gopal Gupta. 2023. Reliable natural language understanding with large language models and answer set programming.

Aman Rangapur and Haoran Wang. 2023. ChatGPT-crawler: Find out if ChatGPT really knows what it's talking about.

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

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023b. Can ChatGPT Assess Human Personalities? A General Evaluation Framework.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2023. Unsupervised summarization re-ranking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8341–8376, Toronto, Canada. Association for Computational Linguistics.

Anton Razzhigaev, Mikhail Salnikov, Valentin Malykh, Pavel Braslavski, and Alexander Panchenko. 2023. A system for answering simple questions in multiple languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 524–537, Toronto, Canada. Association for Computational Linguistics.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation.

Saed Rezayi, Zhengliang Liu, Zihao Wu, Chandra Dhakal, Bao Ge, Haixing Dai, Gengchen Mai, Ninghao Liu, Chen Zhen, Tianming Liu, and Sheng Li. 2023. Exploring new frontiers in agricultural nlp: Investigating the potential of large language models for food applications.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages.

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

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023. Did ChatGPT cheat on your test?

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2023. Neural theory-of-mind? on the limits of social intelligence in large lms.

Jakob Schuster and Katja Markert. 2023. Nutcracking sledgehammers: Prioritizing target language data over bigger language models for cross-lingual metaphor detection. In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 98–106, Gothenburg, Sweden. Association for Computational Linguistics.

Paulo Shakarian, Abhinav Koyyalamudi, Noel Ngu, and Lakshmivihari Mareedu. 2023. An independent evaluation of ChatGPT on mathematical word problems (mwp).

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451, Toronto, Canada. Association for Computational Linguistics.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In ChatGPT we trust? measuring and characterizing the reliability of ChatGPT.

Yucheng Shi, Hehuan Ma, Wenliang Zhong, Qiaoyu Tan, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. 2023. Chatgraph: Interpretable text classification by converting ChatGPT knowledge to graphs.

Zhengxaing Shi and Aldo Lipani. 2023. Don't stop pretraining? make prompt-based fine-tuning powerful learner. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dominik Sobania, Martin Briesch, Carol Hanna, and Justyna Petke. 2023. An analysis of the automatic bug fixing performance of ChatGPT.

Mingyang Song, Haiyun Jiang, Shuming Shi, Songfang Yao, Shilong Lu, Yi Feng, Huafeng Liu, and Liping Jing. 2023. Is ChatGPT a good keyphrase generator? a preliminary study.

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

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2023a. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023b. Is ChatGPT good at search? investigating large language models as re-ranking agent.

Eugene Syriani, Istvan David, and Gauransh Kumar. 2023. Assessing the ability of ChatGPT to screen articles for systematic reviews.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023a. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023b. Can ChatGPT replace traditional kbqa models? an in-depth analysis of the question answering performance of the GPT LLM family.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, and Yong Jiang. 2023c. DAMO-NLP at SemEval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2014–2028, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023a. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023b. Does synthetic data generation of LLMs help clinical text mining?

Ruibo Tu, Chao Ma, and Cheng Zhang. 2023a. Causal-discovery performance of ChatGPT in the context of neuropathic pain diagnosis.

Shangqing Tu, Chunyang Li, Jifan Yu, Xiaozhi Wang, Lei Hou, and Juanzi Li. 2023b. Chatlog: Recording and analyzing ChatGPT across time.

Jens Van Nooten and Walter Daelemans. 2023. Improving Dutch vaccine hesitancy monitoring via multi-label data augmentation with GPT-3.5. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 251–270, Toronto, Canada. Association for Computational Linguistics.

Bhaskara Hanuma Vedula, Prashant Kodali, Manish Shrivastava, and Ponnurangam Kumaraguru. 2023. PrecogIIITH@WASSA2023: Emotion detection for Urdu-English code-mixed text. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 601–605, Toronto, Canada. Association for Computational Linguistics.

Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. ChatGPT for robotics: Design principles and model abilities.

Boshi Wang, Xiang Yue, and Huan Sun. 2023a. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023b. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023c. Is ChatGPT a good NLG evaluator? a preliminary study.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023d. Zero-shot cross-lingual summarization via large language models.

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. 2023e. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective.

Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, Zhichao Yang, and Hong Yu. 2023f. UMASS_BioNLP at MEDIQA-chat 2023: Can LLMs generate high-quality synthetic note-oriented doctor-patient conversations? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 460–471, Toronto, Canada. Association for Computational Linguistics.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 626–667, Toronto, Canada. Association for Computational Linguistics.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023g. ChatCAD: Interactive computer-aided diagnosis on medical image using large language models.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023h. Scibench: Evaluating college-level scientific problem-solving abilities of large language models.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2023i. Mint: Evaluating LLMs in multi-turn interaction with tools and language feedback.

Yuqing Wang and Yun Zhao. 2023. Metacognitive prompting improves understanding in large language models.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023j. Is ChatGPT a good sentiment analyzer? a preliminary study.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with ChatGPT.

Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023a. ChatGPT or grammarly? evaluating ChatGPT on grammatical error correction benchmark.

Qianhui Wu, Huiqiang Jiang, Haonan Yin, Börje Karlsson, and Chin-Yew Lin. 2023b. Multi-level knowledge distillation for out-of-distribution detection in text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7317–7332, Toronto, Canada. Association for Computational Linguistics.

Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023c. Do PLMs know and understand ontological knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3080–3101, Toronto, Canada. Association for Computational Linguistics.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 610–625, Toronto, Canada. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023a. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts.

Qianqian Xie, Weiguang Han, Yanzhao Lai, Min Peng, and Jimin Huang. 2023b. The wall street neophyte: A zero-shot analysis of ChatGPT over multimodal stock movement prediction challenges.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023a. Are large language models really good logical reasoners? a comprehensive evaluation and beyond.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023b. SuperCLUE: A comprehensive chinese large language model benchmark.

Zihang Xu, Ziqing Yang, Yiming Cui, and Shijin Wang. 2023c. IDOL: Indicator-oriented logic pre-training for logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8099–8111, Toronto, Canada. Association for Computational Linguistics.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023a. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond.

Wayne Yang and Garrett Nicolai. 2023. Neural machine translation data generation and augmentation using ChatGPT.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023b. Exploring the limits of ChatGPT for query or aspect-based text summarization.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023c. Mm-react: Prompting ChatGPT for multimodal reasoning and action.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023a. Zero-shot temporal relation extraction with ChatGPT.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023b. How well do large language models perform in arithmetic tasks?

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023a. Investigating chain-of-thought with ChatGPT for stance detection on social media.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Extractive summarization via ChatGPT for faithful summary generation.

Jianguo Zhang, Kun Qian, Zhiwei Liu, Shelby Heinecke, Rui Meng, Ye Liu, Zhou Yu, Huan Wang, Silvio Savarese, and Caiming Xiong. 2023c. Dialogstudio: Towards richest and most diverse unified dataset collection for conversational AI.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023d. BayLing: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023e. Sentiment analysis in the era of large language models: A reality check.

Xiyuan Zhang, Ranak Roy Chowdhury, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. 2023f. Modeling label semantics improves activity recognition.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023a. Is ChatGPT equipped with emotional dialogue capabilities?

Xiaofeng Zhao, Min Zhang, Miaomiao Ma, Chang Su, Yilun Liu, Minghan Wang, Xiaosong Qiao, Jiaxin Guo, Yinglu Li, and Wenbing Ma. 2023b. HW-TSC at SemEval-2023 task 7: Exploring the natural language inference capabilities of ChatGPT and pre-trained language model for clinical trial. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1603–1608, Toronto, Canada. Association for Computational Linguistics.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2023c. Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15590–15606, Toronto, Canada. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023a. Large language models are not robust multiple choice selectors.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023b. Why does ChatGPT fall short in providing truthful answers?

Yi Zheng, Björn Ross, and Walid Magdy. 2023c. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. Association for Computational Linguistics.

Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, and Hui Xiong. 2023d. Generative job recommendations with large language model.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023a. Can ChatGPT understand too? a comparative study on ChatGPT and fine-tuned bert.

Tianyang Zhong, Yaonai Wei, Li Yang, Zihao Wu, Zhengliang Liu, Xiaozheng Wei, Wenjun Li, Junjie Yao, Chong Ma, Xiang Li, Dajiang Zhu, Xi Jiang, Junwei Han, Dinggang Shen, Tianming Liu, and Tuo Zhang. 2023b. Chatabl: Abductive learning via natural language interaction with ChatGPT.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. RecurrentGPT: Interactive generation of (arbitrarily) long text.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT reproduce human-generated labels? a study of social computing tasks.

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

B Detail on evaluation malpractices

As the Sankey diagrams showed in Section 5.2 and Section 5.3 offer limited insights on our findings regarding evaluation reproducibility and fairness, we do provide additional details in this section. We provide concrete numbers for our assessment of reproducibility (Sec. 5.2) and evaluation (mal)practices (Sec. 5.3) in Tables 2 and 3, respectively.

C Detailed List of ChatGPT Data Leak

We show which datasets have been leaked to ChatGPT in Tables 4 and 5.

Prompts	Repo	Sampl.	Custom	n. (%)
				3 (2.11%)
			✓	1 (0.70%)
		✓		8 (5.63%)
	✓			3 (2.11%)
	✓	✓		2 (1.41%)
✓				20 (14.08%)
✓			✓	3 (2.11%)
✓		✓		27 (19.01%)
✓		✓	✓	3 (2.11%)
✓	✓			37 (26.06%)
✓	✓		✓	4 (2.82%)
✓	✓	✓		27 (19.01%)
✓	✓	✓	✓	4 (2.82%)

(a) Pre-prints

Prompts	Repo	Sampl.	Custom	n. (%)
				1 (1.43%)
	✓		✓	1 (1.43%)
	✓	✓		1 (1.43%)
✓				14 (20.00%)
✓			✓	7 (10.00%)
✓		✓		9 (12.86%)
✓		✓	✓	3 (4.29%)
✓	✓			8 (11.43%)
✓	✓		✓	4 (5.71%)
✓	✓	✓		16 (22.86%)
✓	✓	✓	✓	6 (6.57%)

(b) Peer-reviewed works

Table 2: Statistics related to the reproducibility of the work reviewed: the availability of used prompts (Prompts) and code/data repository (Repo), the usage of custom datasets (Custom), the application of random sampling or any other practice that does not allow the exact reconstruction of the data used (Sampl.).

Comp.	Scale	n. (%)
		71 (50.00%)
✓		54 (38.03%)
✓	✓	17 (11.97%)

(a) Pre-prints

Comp.	Scale	n. (%)
		30 (42.86%)
✓		34 (48.57%)
✓	✓	6 (8.57%)

(b) Peer-reviewed works

Table 3: Fairness statistics for reviewed work. Statistics related to the practices of performance comparisons between ChatGPT/GPT-4 and other open models: whether such comparisons are performed at all (Comp.) and whether they are of the same scale (Scale).

Task name	Lo	M-Lo	M-Hi	Hi
AI safety & ethics			bbq (all), bold (all)	
Creative text generation	WrintingPrompts (test)			
Dialogue	OpenDialKG (test), ProsocialDialog (test)	MultiWOZ 2.2 (test)		DSTC11 track 5 (dev), DSTC7 Track 2 (all), MultiWOZ 2.1 (test), MultiWOZ 2.4 (test), mutual (test)
Evaluation of generated texts				NEWSROOM (all), OpenMEVA (all), RealSumm (all), SummEval (all)
Machine Translation	FLORES-101 (test), WMT20 (EN-DE; Robustness Task Set 2 - EN-JA; Robustness Task Set 2 - JA-EN; Robustness Task 3; ZH-EN) (test), WMT22 (test)	NusaX (test), WMT19 Biomedical Translation Task (test), WMT 2014 News dataset (EN-FR; EN-DE) (test)	FLORES-200 (dev)	MQM annotations of the WMT 2022 task (EN-DE, EN-RU, ZH-EN) (test)
Math		NumerSense (dev)		AddSub (all), AQUA-RAT (test), DRAW-1K (all), GHOSTS (all), GSM8K (test), MultiArith (all), SingleEQ (all), SVAMP (all)
Medical text generation	DDXPlus (EN) (test), MIMIC-CXR (test)			Merck Sharpe & Dohme (MSD) clinical manual (all)
Natural Language Inference	BECEL (SNLI; RTE) (test), CommitmentBank (all), MultiNLI (dev), QNLI (dev), RTE (all), αnli (dev)	EntailmentBank (test)		MED (test), Adversarial GLUE (MNLI; QNLI; RTE) (dev), ANLI-R3 (test), SuperGLUE (AX-g; cb) (dev), ConjNLI (test), ConTRoL (logical reasoning) (test), HELP (test), mnli (test), RTE (dev), NLI4CT (SemEval 2023 - Task 7) (all), TaxiNLI (test), WNLI (dev)

Table 4: The names of datasets with low (Lo), moderate-low (M-Lo), moderate-high (M-Hi), and high (Hi) leakage, categorized according to the task. (1/3)

Task name	Lo	M-Lo	M-Hi	Hi
Natural Language Understanding				ATIS (test), SNIPS (test)
Paraphrasing	MRPC (dev), Glue (QQP) (dev)			
Politics		Covid19 (Scientific; Social) (test)		P-Stance (test), SemEval 2016 Task 6 (test), TweetEval (TweetStance) (test)
Programming				QuixBugs (all)
Psychology				Myers–Briggs Type Indicator (all)
Question answering	Custom medical dataset from AMBOSS (all), bAbI (Task 16) (test), CLUTTR (test), e-CARE (dev), FinanceZhidao (all), FreebaseQA (all), HotpotQA (dev), LCQUAD 2.0 (all), LegalQA (all), LogiQA (all), math (test), MC-TACO (dev), MedDialog (all), MKQA (all), pep-3k (all), PIQA (test), ReClor (all), SimpleQuestions (all), SpartQA (test), StepGame (test), TimeDial (test), WebQuestions (all), WebTextQA & BaikeQA (all)	bAbI (Task 15) (test), Custom dataset from BCSC Self-Assessment Program (all), Unnamed Chinese Psychological QA dataset (all), ELI5 (all), NLPCC-DBQA (all), OpenBookQA (dev), Custom dataset from OphthoQuestions (all), PIQA (dev), QASC (dev), RACE (test or dev), Social IQA (dev), SQuAD 2.0 (dev), TruthfulQA (test), WikiQA (all)	fiqa (all), GSM8K (test), KQA Pro (all), OpenBookQA (test), Custom USMLE dataset (all)	Adversarial GLUE (qqp) (dev), AR-LSAT (test), Custom QA dataset from BaiduBaike (all), BoolQ (test), BoolQ Contrast Set (test), Pre-processe version of BRON (all), CVE (2021; ATT) (all), ComplexWebQuestions (all), DBLP (all), EfficientQA (dev), GrailQA (test), GraphQuestions (all), HC3 (Chinese; English) (all), Custom dataset based on the Hofstede Culture Survey (all), LC-QuAD 2.0 (all), LogiQA 2.0 (test), MAG (all), Custom medical dataset from NBME (all), OTT-QA (all), ProtoQA (dev), QALD-9 (all), ReClor (dev), TruthfulQA (Generation subset) (test), Test of Understanding in College Economics (TUCE) (all), Wiki-csai (computer science-related concepts extracted from Wikipedia) (all), WQSP (all), YAGO (all)
Reasoning & common sense	CommonsenseQA(test), HellaSwag (dev), Letter String Analogies (Webb et al.) (all)	ARC 2018 (dev), Coin flip dataset (all), COPA (dev), WSC (dev)		CoLA (dev), CommonsenseQA (dev), Date Understanding (all), Last letter dataset (all), MATRES (test), Object counting (all), StrategyQA (all), TDDiscourse (test), TimeBank-Dense (test)

Table 5: The names of datasets with low (Lo), moderate-low (M-Lo), moderate-high (M-Hi), and high (Hi) leakage, categorized according to the task. (2/3)

Task name	Lo	M-Lo	M-Hi	Hi
Semantic similarity	STS-B (dev), TweetEval (TweetEmoji) (test or dev)	BECEL (MRPC) (test)		WSDEval (test or dev), WiC (dev), WiC(test or dev)
Sentiment analysis	ColBERT (test or dev), Flipkart Product Reviews (all), IMDb Movie Review Data (test), SST-2 (dev), UCC (test or dev), UnhealthyPer (test or dev) WikiDetox (aggression task) (test or dev), AggressionPer	GoEmotions (test or dev), GoEmoPer0-3 Implicit Hate Corpus (all), Sarcasmania (sarcasm task) (test or dev), SemEval 2023 - Task 9 (test), TweetEval - Sentiment (test or dev)	Real Toxicity Prompts (all)	AdvGLUE (SST-2) (dev), CLARIN-Emo (test or dev), ChaLearn 2016 FI (personality task) (all), Contrast Sets (IMDb) (all), PolEmo 2.0 (test or dev), Sentiment140 (all), The Suicide and Depression Dataset (all)
Summarization	CNN DailyMail (test), CrossSum (En - Zh) (test), Reddit TIFU (test), WikiLingua (En - Zh/De) (test), XSAMSum (En - Zh/De) (test)	CovidET (test), NEWTS (test), PubMed dataset (test), QMSum (test), XSum (test)	SQuALITY (test)	SAMSum (test)
Text classification	Inverse Scaling Prize (all datasets) (all) (all)			PubMed20K (train), SMS Spam Collection V1 (test or dev), Symptoms dataset (train)
Text extraction	MTSamples (all)	I2B2 2010 (all)		ACE 2005 (all), CoNLL++ (all), CoNLL 2003 (test), DuEE 1.0 (all), DuIEduie 2.0 (all), MSRA (all), NYT11-HRL (all)
Text generation		CoNLL 2014 Shared Task dataset (test)		ADVETA (ADD, RPL) , COSQL (dev), CSpider (dev), DuSQL (all), Quiz Design (all), SParC (dev), Spider (dev), Spider-CG (app, sub) (all), Spider-DK (dev), Spider-Realistic (dev), Spider-Syn (dev)

Table 6: The names of datasets with low (Lo), moderate-low (M-Lo), moderate-high (M-Hi), and high (Hi) leakage, categorized according to the task. (3/3)