INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

How ChatGPT is Robust for Spoken Language Understanding?

Guangpeng Li, Lu Chen, Kai Yu**

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering
Shanghai Jiao Tong University, Shanghai, China
{guangpengli, chenlusz, kai.yu}@sjtu.edu.cn

Abstract

Large language models (LLMs), e.g. ChatGPT, have shown super performance on various NLP tasks. There is a doubt whether these LLMs, which are trained on a large corpus of written text, can show the robustness of understanding the spoken text. Therefore in this paper, we give a detailed investigation of robust spoken language understanding (SLU) with ChatGPT. In our experiments, we evaluate ChatGPT on two sets of public datasets, Spoken SQuAD and ASR-GLUE, in which there are ASR errors in the text. Quantitative and qualitative analyses on the experimental results are conducted to show that ChatGPT not only performs very well for SLU tasks but also can recover some ASR errors with its super reason ability.

Index Terms: Spoken Language Understanding, ChatGPT, Large Language Models (LLMs)

1. Introduction

Large language models (LLMs) are language models based on text-to-text framework with a vast amounts of parameters. A number of research teams have published their own LLMs such as GPT-3 and GPT-3.5 [1], PaLM [2], OPT-175B [3], etc. According to [4], all text-based problems can be converted into a text-to-text format and can be handled by a unified text-to-text language model. On the other hand, the previous work [1, 5] have shown that as the size of the language model increases, the performance and reasoning ability of the model are significantly improved.

These two types of research laid the foundation for the bloom of large language models and their amazing performance on almost all the NLP tasks. There have been a lot of work of evaluating LLMs' performance on many NLP tasks. [6] carries out an evaluation of ChatGPT using 21 public datasets covering 8 different NLP tasks. The experimental results have shown that ChatGPT outperforms other LLMs in zero-shot setting and even outperforms fine-tuned models on some tasks. [7] proposes an approach works by chaining together reasoning steps to perform faithful multi-step reasoning via a process whose causal structure mirrors the underlying logical structure of the problem. Their work have proved that with an appropriate method, LLMs could generate humanly interpretable reasoning traces whose validity can be checked by the user. [8] demonstrates both the effectiveness and limitations of the current version of ChatGPT. Their evaluation on 20 popular NLP datasets covering 7 representative task categories has shown that ChatGPT performs well on many tasks favoring reasoning capabilities (e.g., arithmetic reasoning) while it still faces challenges when solving specific tasks such as sequence tagging and give an in-depth analysis through qualitative case studies.

*Corresponding author

Since LLMs have shown amazing performance and strong reasoning abilities on many NLP tasks, there is a natural idea that whether LLMs are robust to the text that may contain ASR errors, could we use LLMs for spoken language understanding (SLU) tasks? To answer this question, we select one of the most influential and representative LLMs ChatGPT, and evaluate its performance on SLU benchmarks. The main contributions of our paper are summarized as follows: we conducted a series of experiments on two benchmarks of spoken language understanding (SLU): Spoken SQuAD [9] and ASR-GLUE [10], covering 5 SLU tasks: spoken document question answering, sentiment analysis, semantic similarity classification, natural language inference (NLI), and recognizing textual entailment (RTE). Based on the experimental results, we give a quantitative and qualitative analysis of ChatGPT on SLU tasks.

2. Related work

2.1. Spoken Language Understanding

The main difference between normal natural language understanding (NLU) and spoken language understanding (SLU) is that SLU needs to process spoken language which is usually the result of automatic speech recognition (ASR). The "spoken language" here means not only the spoken language in dialogue but also other natural language from speech recognition such as spoken document and spoken question.

There have been a lot of work in SLU [11, 12, 13, 14]. [15] is the first one to propose spoken conversational question answering (SCQA) task, aiming at enabling the systems to model complex dialogue flows given the speech documents. They build an end-to-end system to deal with conversational questions based on the audio recordings and publish a spoken conversational question answering dataset. [16] proposes an data augmentation approach by leveraging an ASR error simulator to inject noise into the error-free text data. The experiments show that their approach improves the performance of downstream dialogue models in the presence of ASR errors. [17] proposes three empirical approaches to improve the robustness of SLU models. [18] proposes controlled paraphrase networks to generate syntactically adversarial examples that both fool pre-trained models and improve the robustness of these models to syntactic variation when used to augment their training data.

To the best of our knowledge, there are currently no existing studies that employ LLMs for SLU tasks.

2.2. Large Language Models and ChatGPT

The emergence of large language models (LLMs) have triggered a complete revolution to natural language processing (NLP) field. They can complete different NLP tasks with appropriate prompts without any gradient update. For example, if

10.21437/Interspeech.2023-1466

Table 1: *Prompts and input formats of 5 SLU tasks*

SLU tasks	**Corresponding prompts and data format**
Document Question Answering	I need you help me to do the reading comprehension task, please remember all the answers are in the original text(which means that your answer are all from the text I give to you). Now the article's title are 'Title', the text are 'Text', what's the answer for the question 'Question'?
Sentiment Analysis	I need you help me to predict the sentiment of a given sentence from movie reviews, the sentiment only contains two class: positive and negative. The sentence is 'Sentence', what's this sentence's sentiment (positive or negative)?
Semantically Equivalent Classification	I will give you a question pairs from the community question-answering website Quora, and I need you help me to determine whether a pair of questions are semantically equivalent. The question pairs are 'Question1', and 'Question2', are they semantically equivalent? (YES or NO)
Natural Language Inference	I need you help me to do natural language inference task which is a directional relation between text fragments. The relation holds whether the truth of one text fragment can be inferenced from another text. So the task means you must decide whether the given hypothesis can be inferenced from the given text or not. Now I give you a text: 'Text' and a hypothesis: 'Hypothesis', can the hypothesis be inferenced from the text? (YES or NO)
Recognizing Textual Entailment	I need you help me to do textual entailment task which is a directional relation between text fragments. The relation holds whether the truth of one text fragment follows from another text. In the TE framework, the entailing and entailed texts are termed text (t) and hypothesis (h), respectively. So the task means you must decide whether the given hypothesis is entailed to the given text or not. Now I give you a text: 'Text' and a hypothesis: 'Hypothesis', what is their relation? (YES or NO)

you want LLMs to do the machine translation task, you can use a prompt like " Translate the following sentence from English to Spanish." with a specific English sentence as an input and the model will generate the translation result. Compared to the other pre-trained language models before, LLMs usually don't need any downstream data (in zero shot setting) or just require a few samples (in few shot setting).

Since the inputs of LLMs during the inference period are mainly prompts with specific task data, the quality of the prompts greatly affects the performance of LLMs. There have been a lot of work about prompt engineering such as manually designed prompts [19] or automatically designed prompts [20, 21].

ChatGPT is the most advanced LLMs released by OpenAI a few months ago. As soon as it's introduced to public, ChatGPT has attracted the attention of the entire NLP community. Compared with other LLMs before, ChatGPT is different for its ability to communicate with users. Therefore, it could complete some complex downstream tasks through multi-turn dialogues. This "chat" ability not only achieves better performance on specific tasks but also gives users a perfect experience. Despite achieving so many accomplishments, ChatGPT also has some limitations. First, ChatGPT may generate answers that seem to be correct but acturally not true. Second, just like other LLMs, ChatGPT is sensitive to the users' prompt, so choose the right prompt could be the key factor to its performance. Finally, ChatGPT may be excessively verbose sometimes and generate some content that is not related to the answer.

Although there have been a lot of papers evaluating LLMs and ChatGPT for many NLP tasks [22, 23, 24], within our konwledge, we are the first one to evaluate ChatGPT on SLU tasks.

3. Methodology

As mentioned in Section 1, we choose one of the most advanced LLMs ChatGPT to evaluate its ability on several SLU tasks. Specifically, we use the ChatGPT API published by OpenAI[1]. To make sure that ChatGPT could understand these tasks, we have designed appropriate prompts for each task. The corresponding prompts and data formats of all the tasks are listed in Table 1.

The overall experimental procedure is as follows:

1. Given the input test X of one task and its corresponding prompt P, we feed them to ChatGPT and get its output completion $C = f(X, P)$.
2. After post-processing of completion C above, we get the final answer of ChatGPT.

For example, if we want to evaluate ChatGPT on spoken semantically equivalent classification task, we need to combine the corresponding prompt with the question pairs to get the final input like "I will give you a question pairs from the community question-answering website Quora, and I need you help me to determine whether a pair of questions are semantically equivalent. The question pairs are 'How do I become a video game maker', and 'How do I become an indie game maker', are they semantically equivalent? (YES or NO)". As soon as we get ChatGPT's output, the post-processing method will be used to extract the final answer from it.

4. Experiments

In this section, we will give a brief introduction to the datasets we used and their corresponding tasks, then present the experimental results, and finally we will give a quantitative and qualitative discussion about whether ChatGPT is robust to classic ASR errors.

[1] https://openai.com/blog/openai-api

Table 2: *Case Study on Spoken SQuAD*

Index	Excerpt from the original text	ASR transcription	Question and Answer	Response of ChatGPT
1	In 1888, the editor of Electrical World magazine, Thomas Commerford Martin ...	in eighteen eighty eight the editor of electrical world magazine thomas comer for the margin ...	Question: Who was Thomas Commerford Martin? Answer: editor of electrical world magazine	I'm sorry, but there seems to be some errors in the text you provided. The name of the editor of Electrical World magazine is actually Thomas Commerford Martin, not Thomas Comer.
2	...and that the game would be named using Arabic numerals as Super Bowl 50 as opposed to Super Bowl L.	...and that the game would be named using arabic numerals is super bowl fifty as opposed to super bowl al	Question: If Roman numerals were used in the naming of the 50th Super Bowl, which one would have been used? Answer: L	If Roman numerals were used in the naming of the 50th Super Bowl, it would have been Super Bowl L.
3	...Tesla's father, in a moment of despair, promised to send him to the best engineering school if he recovered from the illness (his father had originally wanted him to enter the priesthood).	...kesslers father in a moment of despair promise to send him to the best engineering school if he recovered from the illness his father had originally wanted him to enter the priesthood.	Question: What did Tesla's father promise him while he were bedridden? Answer: the best engineering school	I'm sorry, but the text you provided contains information that is not related to Nikola Tesla. It seems to be about someone named Kessler who contracted cholera in 1873. Can you please provide the correct text for the reading comprehension task?

Table 3: *Accuracy on Spoken SQuAD*

MODEL	Accuracy	WER
Original	94.72	-
Spoken	82.36	22.73

Table 4: *Accuracy on ASR-GLUE*

Datasets	Original	Low	Medium	High
SST-2	85.16	84.51	76.77	70.96
QQP	80.18	77.02	73.20	70.04
QNLI	88.07	86.09	73.51	71.85
SciTail	71.71	74.34	69.4	66.12
RTE	76.72	75.58	66.81	63.79

4.1. Datasets and Results

We evalutate ChatGPT on 5 SLU tasks on two benchmarks: Spoken SQuAD and ASR-GLUE. All the experiments are in zero shot setting and single-turn dialogue with ChatGPT.

4.1.1. Spoken SQuAD

Spoken SQuAD is an extraction-based spoken question answering (SQA) dataset. It contains the ASR transcriptions of the document and text form of input question. The original textual document and question are taken from another QA benchmark SQuAD. To get the final spoken document, they used Google text-to-speech system to generate the spoken version of the original document, and then get the corresponding ASR transcriptions through CMU Sphinx.

Table 5: *The average WER of our selected ASR-GLUE test set on 3 noise levels*

Datasets	Low	Medium	High
SST-2	15.01	25.66	29.94
QQP	11.74	20.49	23.05
QNLI	15.45	27.37	32.07
SciTail	11.23	22.72	26.51
RTE	21.01	35.8	42.66

For our experiment, we randomly selected 10% of the data in the test set as our test set, which contains 947 questions. The accuracy of ChatGPT on Spoken SQuAD is reported on Table 3. The reason why we only give the accuracy result without exact match (EM) result is that the outputs of ChatGPT are usually natural sentences which can not be the same as the original text though they are right. We evaluate the ChatGPT's outputs manually. Compared with original text result, the accuracy of ChatGPT in spoken text whose WER is 22.73% is reduced by 13%.

4.1.2. ASR-GLUE

ASR-GLUE is constructed on the basis of GLUE. It contains 5 typical SLU tasks and 6 corresponding datasets. They hire 6 native speakers to record all test samples in 3 different noise levels. Specially, low-level noise is the same as the original audio; medium-level noise means introducing reverberation and 15dB signal-to-noise-ratio (SNR) level noise into the original audio; high-level noise means introducing reverberation and 10dB SNR level noise into the original audio.

For ASR-GLUE, we selected 4 SLU tasks and their cor-

Table 6: *Examples of each error type*

Error Type	**Ground Truth**	**ASR Result**
Number	Liberated by Napoleon's army in 1806	liberated by napoleons army in eighty no sex
Entity	Tancred was instrumental in the conquest	ten grand was instrumental in the conquest
Similar Sound	need the city council to help fund the event	need the city council to help find the event
Liaison	What's the size of the live chat	what's the size of the lichen
Deletion	What language is this in	what languages * in
Insertion	when a person should ideally retire	when a person should ideally english tired tired tired

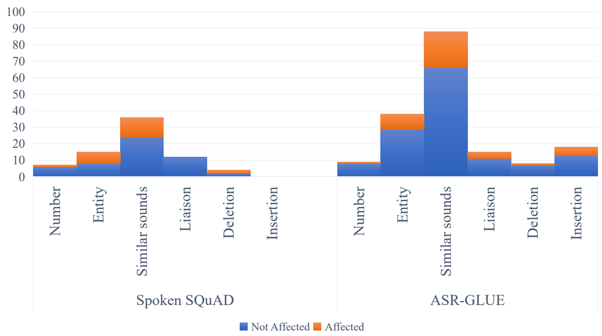

Figure 1: *Statistics on the 6 types of errors. Note that the blue means ChatGPT is not affected by ASR errors and orange means ChatGPT is affected by ASR errors. The statistics data is the specific number of each type.*

responding datasets: sentiment classification (SST-2), semantic similarity classification (QQP), question-answering and science natural language inference (QNLI, SciTail) and recognizing textual entailment (RTE). We report the accuracy of ASR-GLUE on Table 4. Note that we randomly choose 2 speakers from the original 6 speakers and report the average result.

4.1.3. Statistics on the types of ASR errors

To explore whether ChatGPT is robust to specific type of ASR errors in the selected datasets, we try to give an appropriate classification of common ASR errors. Based on the ASR error type defined in ASR-GLUE, we propose our own classification criteria: *number recognition errors, entity recognition errors, similar sound errors (except number and entity), liaison errors, deletion errors (including symbol missing) and insertion errors*. Examples of each error type are given in Table 6. Note that 95% errors appeared in the data we analyzed are covered by the 6 categories we defined.

Then we give a comprehensive statistics on the specific types of ASR errors and whether ChatGPT is affected by the errors. For Spoken SQuAD, we randomly selected 150 question-answer pairs; for 5 datasets from ASR-GLUE, we randomly selected 50 samples on medium noise level respectively, so the total number of ASR-GLUE is 200. The results of the statistics are presented in Figure 1.

Two conclusions can be drawn from the results:

1. Among all the error types, Similar Sound errors occur most frequently, which is consistent with our intuition about ASR errors. The reason why there is no insertion errors in Spoken SQuAD is that the spoken data we used contains no noise.
2. ChatGPT is robust to Number errors, Liaison errors and Insertion errors, and is not affected by more than half of the other 3 type errors. For Spoken SQuAD it is not affected by 85.71% Number errors, 53.33% Entity errors, 66.67% Similar Sound errors, 100% Liaison errors, 50% Deletion errors; for ASR-GLUE it is not affected by 88.88% Number errors, 76.31% Entity errors, 75% Similar Sound errors, 73.33% Liaison errors, 87.5% Deletion errors, 72.22% Insertion errors.

4.2. Discussion

To investigate the robustness of ChatGPT to the 6 errors mentioned above, we conduct case study to the Spoken SQuAD results. All the selected cases are list in Table 2. The case by case study is as follows:

1. For case 1, ChatGPT points out the Entity error which is "homascomer for the margin" ("Thomas Commerford Martin") and gives the right answer.
2. For case 2, ChatGPT infers the correct Roman numerals "L" by itself without influenced by the Number error "al".
3. For case 3, because of the Entity error "kesslers" instead of "Tesla", ChatGPT could not give the right answer about Tesla's father. This case demonstrates that although ChatGPT has the error correction capability, it can still lead to answer failure when the semantics of the corresponding sentence changes very significantly.

Based on the total experiments, we believe that ChatGPT could correct some ASR errors with its own knowledge or its surprising reasoning ability.

5. Conclusion

To evaluate whether ChatGPT is robust to ASR errors on SLU tasks, we conduct a series of experiments on two benchmarks: Spoken SQuAD and ASR-GLUE. We divide the classic ASR errors into 6 types and report a quantitative and qualitative analysis of ChatGPT on these ASR errors. Based on the experiments, we conclude that ChatGPT is robust enough to Number, Liaison and Insertion errors, and have the ability to correct some ASR errors. We hope our work could help the subsequent study on the application of LLMs to SLU tasks.

6. Acknowledgements

This study was supported by Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0102), and the Key Research and Development Program of Jiangsu Province, China (No.BE2022059-2).

7. References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," *CoRR*, vol. abs/2204.02311, 2022.

[3] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: open pre-trained transformer language models," *CoRR*, vol. abs/2205.01068, 2022.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[5] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *CoRR*, vol. abs/2206.07682, 2022.

[6] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *CoRR*, vol. abs/2302.04023, 2023.

[7] A. Creswell and M. Shanahan, "Faithful reasoning using large language models," *CoRR*, vol. abs/2208.14271, 2022.

[8] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is chatgpt a general-purpose natural language processing task solver?" *CoRR*, vol. abs/2302.06476, 2023.

[9] C.-H. Lee, S.-L. Wu, C.-L. Liu, and H. yi Lee, "Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension," in *Interspeech*, 2018.

[10] L. Feng, J. Yu, D. Cai, S. Liu, H. Zheng, and Y. Wang, "ASR-GLUE: A new multi-task benchmark for asr-robust natural language understanding," *CoRR*, vol. abs/2108.13048, 2021.

[11] K. Sun, L. Chen, S. Zhu, and K. Yu, "The SJTU system for dialog state tracking challenge 2," in *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*. The Association for Computer Linguistics, 2014, pp. 318–326.

[12] S. Zhu and K. Yu, "Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 5675–5679.

[13] S. Zhu, L. Chen, K. Sun, D. Zheng, and K. Yu, "Semantic parser enhancement for dialogue domain extension with little data," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 336–341.

[14] K. Sun, L. Chen, S. Zhu, and K. Yu, "A generalized rule based tracker for dialogue state tracking," in *2014 IEEE Spoken Language Technology Workshop, SLT 2014, South Lake Tahoe, NV, USA, December 7-10, 2014*. IEEE, 2014, pp. 330–335.

[15] C. You, N. Chen, F. Liu, S. Ge, X. Wu, and Y. Zou, "End-to-end spoken conversational question answering: Task, dataset and model," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1219–1232.

[16] L. Wang, M. Fazel-Zarandi, A. Tiwari, S. Matsoukas, and L. Polymenakos, "Data augmentation for training dialog models robust to speech recognition errors," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*. Online: Association for Computational Linguistics, Jul. 2020, pp. 63–70.

[17] L. Feng, J. Yu, Y. Wang, S. Liu, D. Cai, and H. Zheng, "Asr-robust natural language understanding on asr-glue dataset," in *Interspeech*, 2022.

[18] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, "Adversarial example generation with syntactically controlled paraphrase networks," in *North American Chapter of the Association for Computational Linguistics*, 2018.

[19] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8-13, 2021, Extended Abstracts*, Y. Kitamura, A. Quigley, K. Isbister, and T. Igarashi, Eds. ACM, 2021, pp. 314:1–314:7.

[20] T. Gao, A. Fisch, and D. Chen, "Making pre-trained language models better few-shot learners," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 3816–3830.

[21] T. Shin, Y. Razeghi, R. L. L. IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 4222–4235.

[22] J. Wang, Y. Liang, F. Meng, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou, "Is chatgpt a good NLG evaluator? A preliminary study," *CoRR*, vol. abs/2303.04048, 2023.

[23] Z. Luo, Q. Xie, and S. Ananiadou, "Chatgpt as a factual inconsistency evaluator for abstractive text summarization," *CoRR*, vol. abs/2303.15621, 2023.

[24] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen, "Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning," *CoRR*, vol. abs/2304.05613, 2023.