ON THE ROBUSTNESS OF CHATGPT UNDER INPUT PERTURBATIONS FOR NAMED ENTITY RECOGNITION TASK

Ishani Mondal*, **Abhilasha Sancheti**
Department of Computer Science
University of Maryland, College Park
{imondal}@umd.edu

ABSTRACT

We present a systematic evaluation of the robustness of ChatGPT (in both zero- and few-shot settings) under input perturbations for Named Entity Recognition (NER) task using automatic evaluation. Our findings suggest: (1) ChatGPT is more brittle on **Drug** or **Disease** entity perturbations (rare entities) as compared to those on widely known **Person** or **Location** entities, and (2) the quality of explanations for the same entity *considerably differ* under various *Entity-specific* and *Context-specific* perturbations; the quality significantly improves using in-context learning.

1 INTRODUCTION

ChatGPT[1] has attracted many users ever since its inception. However, its *reliablility* in the realistic applications wherein entities or context words can be out of distribution of the training data is not clear yet. While previous efforts(§B) have evaluated various aspects of ChatGPT in law (Choi et al., 2023), ethics (Shen et al., 2023), education (Khalil & Er, 2023), verifiability (Liu et al., 2023) and reasoning (Bang et al., 2023), we focus on robustness (Bengio et al., 2021) of its predictions and their explanations to input perturbations for the fundamental task of Named Entity Recognition (NER).

2 METHODOLOGY AND EXPERIMENTAL SETUP

We thoroughly evaluate ChatGPT's robustness (§C and E) to input perturbations on its predictions and their explanations in the **zero- and few-shot scenario** for NER task. We manually design different zero-shot prompts and choose the ones with the maximum accuracy on the original inputs (S). The prompts contain task instruction, candidate target labels, output format description, and input text (see §D for used prompts). For few-shot experiments, we add 4 example (input, output) pairs per entity type for both original and perturbed inputs. The output contains the correct prediction and an explanation for it. Following Lin et al. (2021), we generate *Entity-specific* perturbations by replacing target entities[2] (*i.e.,* perturbed entity, T_E) with other entities (a) of the same type occurring in other sentences (**same entity type**); (b) obtained by rotating the target entity string to obtain natural-looking typos (**typo**); (c) obtained by using Wikidata API to link T_E from its surface to canonical form in Wikidata with a unique identifier (**alias**, and) (d) obtained by randomly generating a string **random string**). *Context-specific* perturbations are generated by using pre-trained language models (*e.g.,* BERT Devlin et al. (2019)) to generate contextual **verb substitutions**. For a sentence (S), we replace target entity by a perturbating entity $T_{E'}$ to generate perturbed sentence (S') (Table 3 for examples). Entities other than target entity in a sentence are referred to as *non-target entities*.

Evaluation: We experiment with CONLL-2003[3] (Tjong Kim Sang & De Meulder, 2003) and BC5CDR (Li et al., 2016) datasets. We **automatically** measure (1) the difference in the accuracy

*https://ishani-mondal.github.io/
[1]https://openai.com/blog/chatgpt
[2]We perform perturbation of 1 target entity or verb at a time to generate S' before for controlled evaluation.
[3]We only consider PERSON and LOCATION entity types for the ease of generating perturbations.

		Effect on Target Entity			Effect on non-target Entities		Overall Effect
		Δ Accuracy ↓	Δ Faithfulness	Similarity	Δ F1 ↓	Δ Faithfulness	Δ F1 ↓
BC5CDR	Alias Perturbation	0.16 / 0.03	0.10 / 0.05	0.69 / 0.81	-0.13 / 0.01	0.01 / 0.01	0.01 /0.01
	Entity Type Perturbation	0.10 / 0.15	0.09 / 0.08	0.58 / 0.74	0.03 / 0.02	0.03 / 0.03	0.03 / 0.02
	Typo Perturbation	0.30 / 0.13	0.21 / 0.15	0.63 / 0.76	0.01 / 0.01	0.01 / 0.01	0.04 / 0.03
	Random Perturbation	0.38 / 0.20	0.27 / 0.15	0.49 / 0.79	0.02 / 0.01	0.01 / 0.01	0.08 / 0.06
	Verb Substitution	-	-	-	0.01 / 0.01	0.01 / 0.01	0.02 / 0.02
CONLL	Alias Perturbation	0.06 / 0.03	0.03 / 0.02	0.77 / 0.78	0.01 / 0.01	0.03 / 0.03	0.03 / 0.03
	Entity Type Perturbation	0.06 / 0.04	0.06 / 0.05	0.75 / 0.82	0.01 / 0.005	0.02 / 0.01	0.02 / 0.01
	Typo Perturbation	0.54 / 0.33	0.46 / 0.24	0.37 / 0.46	0.03 / 0.02	0.01 / 0.01	0.05 / 0.04
	Random Perturbation	0.23 / 0.11	0.15 / 0.09	0.60 / 0.64	0.02 / 0.02	0.02 / 0.02	0.07 / 0.07
	Verb Substitution	-	-	-	0.01 / 0.01	0.02 / 0.01	0.01 / 0.02

Table 1: Results for automatic evaluation (zero-shot/few-shot) of the predictions and explanations per perturbation type for target, non-target, and all the entities for BC5CDR (top) and CONLL (bottom).

	Global vs Local Explanations (Zero-shot)				Global vs Local Explanations (Few-shot)			
	G↑↑L↑↑	G↓↓L↑↑	G↑↑L↓↓	G↓↓L↓↓	G↑↑L↑↑	G↓↓L↑↑	G↑↑L↓↓	G↓↓L↓↓
Alias	**0.54**/0.21	0.26/**0.58**	0.17/0.06	0.02/0.13	**0.57/0.60**	0.20/0.24	0.18/0.05	0.03/0.11
Same Entity Type	**0.61**/0.21	0.22/**0.48**	0.13/0.15	0.02/0.16	**0.48/0.46**	0.29/0.23	0.16/0.15	0.06/0.16
Typo	**0.36**/0.24	0.26/**0.40**	0.26/0.15	0.10/0.20	**0.46/0.34**	0.19/0.30	0.30/0.15	0.03/0.20
Random	0.39/0.11	**0.43/0.63**	0.17/0.15	0.00/0.10	**0.46/0.60**	0.15/0.20	0.19/0.20	0.19/0.11
Verb	0.24/0.22	**0.48/0.56**	0.24/0.07	0.02/0.13	**0.48/0.56**	0.28/0.22	0.20/0.07	0.02/0.13

Table 2: Shows change in type of explanations (BC5CDR/CONLL) due to predictions of common entities before and after perturbation. ↑↑ (↓↓) indicate increase (decrease) after perturbation.

(against gold entity types) of predicting the target entity T_E and perturbating entity $T_{E'}$ (**Δ Accuracy**) to assess the robustness to input perturbations. For the non-target entities, we compute the change in F1 score to assess the impact on their prediction because of target entity perturbation; (2) the difference in faithfulness (**Δ Faithfulness**) of explanations to the input sentence (localness) by measuring the cosine similarity between the explanation and the input sentence; (3) the cosine **similarity** between the explanations generated for the target entity under perturbation; and (4) the change in globalness and localness of explanations under perturbation. We **approximate how the explanation of an entity is grounded to world knowledge (*globalness*)** by obtaining the entity description from wikipedia[4] and calculating the similarity of generated explanation with the summary.

3 Results and Findings

Robustness depends on perturbation type and domain of perturbed entities. Table 1 show that in **zero-shot scenario** ChatGPT is more brittle on **Drug** or **Disease** (BC5CDR) perturbations (rare entities) as compared to that on widely known **Person** or **Location** entities in CONLL in terms of Δ Accuracy and Δ Faithfulness. Typo and Random entity substitutions seem too brittle indicated by high scores. However, in **few-shot scenario**, ΔAccuracy gradually decreases for almost all the perturbations in both datasets, indicating high robustness.

Transition of global and local explainability for the same entity prediction under perturbation. We observe (see Table 2) that overall, the globalness of explanations decreases while faithfulness (localness) to input increases due to perturbation. This provides us with an insight that when an entity is being perturbed, ChatGPT relies more on local context cues to detect entities. This holds true for all types of perturbations in CONLL since person or location names are widely popular, hence before perturbation major predictions were pivoted on world knowledge. However, for Alias, Entity Type, Typo perturbations in BC5CDR, the explanations were more global and local before attack. Thus for the well-known entity types, the model chooses either local or global explanations, whereas after random perturbations, the model always prefer looking at contextual cues. Since the goal of **few-shot** experiments is to increase both localness and globalness in all the explanations of the predicted entities (G↑↑L↑↑), we notice that the performance improves significantly under **few-shot as shown in Table 2**. Sample output predictions for sentences containing target entities are shown in Table 4.

[4] `https://pypi.org/project/wikipedia/`

4 CONCLUSION

We perform automatic evaluation of robustness of ChatGPT's predictions and explanations to input perturbations for the task of NER. We show that ChatGPT is more brittle on domain-specific entity perturbations compared to those on widely known entities. We also observe that the quality of explanations for the same entity *considerably differ* under various perturbations and the robustness and quality significantly improves using in-context (few-shot) learning.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. Exploring the landscape of distributional robustness for question answering models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5971–5987, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.441`.

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

Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Communications of the ACM*, 64(7):58–65, 2021.

Gabriel Bernier-Colborne and Phillippe Langlais. HardEval: Focusing on challenging tokens to assess robustness of NER. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1704–1711, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.211`.

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

Sravan Bodapati, Hyokun Yun, and Yaser Al-Onaizan. Robustness to capitalization errors in named entity recognition. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 237–242, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5531. URL `https://aclanthology.org/D19-5531`.

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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Jifan Chen and Greg Durrett. Robust question answering through sub-part alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1251–1263, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.98. URL `https://aclanthology.org/2021.naacl-main.98`.

Zhuang Chen and Tieyun Qian. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3685–3694, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.340. URL https://aclanthology.org/2020.acl-main.340.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023. URL https://arxiv.org/abs/2301.07597.

Mubin Ul Haque, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. ”i think this is the most disruptive technology”: Exploring sentiments of chatgpt early adopters using twitter data, 2022.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*, 2023.

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

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine, 2023.

Mohammad Khalil and Erkan Er. Will chatgpt get you caught? rethinking of plagiarism detection. *arXiv preprint arXiv:2302.04335*, 2023.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. The moral authority of chatgpt, 2023.

Nghia T. Le and Alan Ritter. Are large language models robust zero-shot coreference resolvers?, 2023.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*, 2023a.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

Xianzhi Li, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? an examination on several typical tasks, 2023b.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. Triggerner: Learning with entity triggers as explanations for named entity recognition. *arXiv preprint arXiv:2004.07493*, 2020.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. RockNER: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3728–3737, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.302. URL `https://aclanthology.org/2021.emnlp-main.302`.

Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*, 2023.

Stephen Mayhew, Nitish Gupta, and Dan Roth. Robust named entity recognition with truecasing pretraining, 2019.

Ishani Mondal. BBAEG: Towards BERT-based biomedical adversarial example generation for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5378–5384, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.423. URL `https://aclanthology.org/2021.naacl-main.423`.

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

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*, 2023.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zheng, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model. *CoRR*, abs/2201.11990, 2022. URL `https://arxiv.org/abs/2201.11990`.

Zhongxiang Sun. A short survey of viewing large language models in legal aspect, 2023.

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. Lamda: Language models for dialog applications, 2022.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL `https://aclanthology.org/W03-0419`.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. Infobert: Improving robustness of language models from an information theoretic perspective, 2021.

Xuezhi Wang, Haohan Wang, and Diyi Yang. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4569–4586, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.339. URL `https://aclanthology.org/2022.naacl-main.339`.

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

Zhen Wang, Hongyi Nie, Wei Zheng, Yaqing Wang, and Xuelong Li. A novel tensor learning model for joint relational triplet extraction. *IEEE transactions on cybernetics*, PP, 2023b.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

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. Zero-shot information extraction via chatting with chatgpt, 2023a.

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

Perturbations	Original Sentence (S)	Perturbed Sentence (S')
Same Entity Type	We tested the sulfated polysaccharide fucoidan , which has been reported to reduce inflammatory brain damage , in a rat model of intracerebral hemorrhage induced by injection of bacterial collagenase into the caudate nucleus .	We tested the sulfated polysaccharide fucoidan , which has been reported to reduce inflammatory chorioretinal atrophy , in a rat model of intracerebral hemorrhage induced by injection of bacterial collagenase into the caudate nucleus .
Alias	CONCLUSION : This study confirms our previous finding that selegiline in combination with L - dopa is associated with selective orthostatic hypotension .	CONCLUSION : This study confirms our previous finding that l-deprenalin in combination with L - dopa is associated with selective orthostatic hypotension .
Typo	China on Thursday accused Taipei of spoiling the atmosphere for a resumption of talks across the Taiwan Strait with a visit to Ukraine by Taiwanese Vice President Lien Chan this week that infuriated Beijing .	China on Thursday accused Taipei of spoiling the atmosphere for a resumption of talks across the Taiwan Strait with a visit to Ukraine by Taiwanese Vice President en ChanLi this week that infuriated Beijing .
Random	Rabinovich is winding up his term as ambassador	I3qk2ia is winding up his term as ambassador
Verb	Speaking only hours after Chinese state media said the time was right to engage in political talks with Taiwan , Foreign Ministry spokesman Shen Guofang told Reuters : " The necessary atmosphere for the opening of the talks has been disrupted by the Taiwan authorities . "	Speaking only hours after Chinese state media announced the time was right to engage in political talks with Taiwan , Foreign Ministry spokesman Shen Guofang told Reuters : " The necessary atmosphere for the opening of the talks has been disrupted by the Taiwan authorities . "

Table 3: Examples of original sentences containing target entities (T_E) and the corresponding sentences with perturbed entities (T'_E) for both "*Entity-Specific*" and "*Context-Specific*" cases. These sentences are interpolated from CONLL and BC5CDR train datasets.

A Appendix

B Background and Related Work

Pre-trained language models such as BERT Devlin et al. (2019), BART Lewis et al. (2020), etc., have shown their power to solve a wide variety of NLP tasks. Several large generative models have been proposed, such as GPT-3 Brown et al. (2020), LaMDA Thoppilan et al. (2022), MT-NLG Smith et al. (2022), PaLM Chowdhery et al. (2022). LLMs usually exhibit amazing capabilities Wei et al. (2022) that enable them to achieve good performance in zero-shot and few-shot scenarios Kojima et al. (2022); Wang et al. (2023b).

Since ChatGPT does not reveal its training details, it imperative to evaluate privacy concerns; concerns that involve ethical risks Haque et al. (2022); Krügel et al. (2023), fake news Jeblick et al. (2022); Chen & Qian (2020), and financial challenges Sun (2023); Li et al. (2023b). For its capabilities, researchers evaluate the performance of ChatGPT on different tasks, including machine translation Peng et al. (2023); Jiao et al. (2023), sentiment analysis Wang et al. (2023a), question-answering Guo et al. (2023), coreference resolution Le & Ritter (2023) and other NLP tasks Bian et al. (2023). In addition, Wei et al. (2023a) propose a two-stage framework, ChatIE, to use ChatGPT for zero-shot information extraction, and evaluate its performance on English and Chinese.

A number of studies have been done to evaluate and improve the robustness of LLMs Chen & Durrett (2021); Awadalla et al. (2022); Wang et al. (2021; 2022); Li et al. (2023a); Wei et al. (2023b); Hu et al. (2023). Since this paper centers around evaluation of robustness for NER tasks, it is worthy to mention that prior researchers have assessed the NER model's robustness on token replacement Bernier-Colborne & Langlais (2020), noisy or uncertain casing Mayhew et al. (2019) and capitalization Bodapati et al. (2019). However, there has not been any comprehensive work in evaluating ChatGPT's robustness on NER and how the quality of explanations vary due to perturbations, which we are trying to fill up in this work.

C Illustrating the method of Generating Adversarial Perturbations

Inspired by Lin et al. (2021), we generate high-quality adversarial examples for evaluating the robustness of ChatGPT on the task of NER by perturbing both the entities (*"Entity-specific"*) and contexts (*"Context-specific"*) of original examples. We refer to the perturbed entity as *"target entity"* (T_E). In a sentence (S) of length n, we denote a target entity as T and it is replaced by a perturbating entity T'_E, thereby generating perturbed sentence (S'). Besides, target entity there could be other possible k entities (O_E = O_{E_1}, O_{E2}..........O_{E_k}) (where $k < n$). Some samples of

adversarial sentences are presented in Table 3. It is important to note that, we perform perturbation of 1 target entity or verb at a time to generate S' before checking NER prediction by ChatGPT.
A. Entity-Specific: In this case, we are generating the following perturbations of entities present in the sentences (containing T_E), and asking ChatGPT to predict named entities for the perturbed sentences (containing T'_E).
a) Alias Replacement: We use Wikidata API to link the target entity T_E in original examples from its surface to canonical form in Wikidata with a unique identifier (**Entity Typing**) and generate p aliases (T_{Ea_1}, T_{Ea_2}... T_{Ea_p}) of those entities.
b) Same Entity Type Replacement: We perturb T_E with another entity of similar semantic class (For instance, a disease replaced by another disease). For this, we retrieve p additional entities occurring in other input sentences. Then we perform p replacements.
c) Typo Replacement: We also consider perturbing the target entity T_E with natural-looking typos, such as rotation of characters in the token of T_E.
d) Random Entity Replacement: We also replace target entity T_E with one randomly generated string and hypothesize that the model would be able to detect the entity based on contextual cues.[5]

B. Context-Specific: Here we generate perturbations of the context around target entities, and ask ChatGPT to predict named entities for the perturbed sentences which contain T_E, and perturbed contextual cues.
Verb substitution with synonyms: We generate *context-specific* attacks by perturbing the main verb v in the sentence with three synonyms (v'_s1, v'_s2, v'_s3) predicted by a pre-trained masked language model like BERT Devlin et al. (2019).

D PROMPT

D.1 ZERO-SHOT

Identify named entities of type “disease” or “chemical” in the below text delimited by triple quotes. Format your response as a list of JSON objects with keys as “type”, “entity”, and “explanation”, and values as “type of the identified entity”, “identified entity”, and “explanation of why it is an entity of that type”, respectively. Ensure that the identified entities can only be words or phrases present in the provided text. Text: “““text”””

D.2 FEW-SHOT

Your task is to identify the named entities of type “disease” or “chemical” in the given text delimited by triple quotes. Format your response as a list of JSON objects with keys as“type”,“entity”, and “explanation”, values as“type of the identified entity”,“identified entity”, and“explanation of why it is an entity of that type”, respectively. Ensure that the identified entities can only be words or phrases present in the provided text. Use the following examples as a guide:

EXAMPLE 1: Text: “““None of the patients had decompensated liver disease”””. Output: ”entity”: ”liver disease”, ”type”: ”disease”, ”explanation”: ”It is a widely known disease and in the sentence it is mentioned that patients did not have decompensate this disease.”

EXAMPLE 2: Text: “““None of the patients had decompensated Measles.”””. Output: ”entity”: ”Measles”, ”type”: ”disease”, ”explanation”: ”Measles is a disease as it is a highly contagious, serious airborne disease caused by a virus that can lead to severe complications and death and in the sentence it is mentioned that patients did not have decompensate this disease.”

[5]One might argue that typo and random perturbations might not guarantee a known entity type by just looking at the names. However, Person, Location names are proper nouns, and the vocabulary of these names are ever-expanding. An intuitive agent (just like humans) should ideally infer the entity-type from its context, instead of memorizing names of the person or location types from the pre-training corpora. This type of capability, usually possessed by humans, will capture the needs of an ever-growing number of different entity instances for a specific entity type. Therefore, we use these standard perturbations (as used by Lin et al. (2021), Mondal (2021)) that are designed to evaluate if context is also considered by these models in predicting the type of the entity, since, in most of the cases entity type should be predicted from the context itself. To evaluate if these unnatural perturbations lead to prediction difficulties by humans as well, we have conducted manual evaluation.

Original Sentence[Perturbation]	**Predictions Before (Zero-shot)**	**Predictions Before (Few-shot)**
Selegiline [L-deprenyl]-induced postural hypotension in Parkinson's disease: a longitudinal study on the effects of drug withdrawal.	**Chemical**: Selginline **Explanation**: Selegiline is a medication used to treat Parkinson's disease.	**Chemical**: Selginline **Explanation**: Selegiline is a chemical since it is a medication used to treat Parkinson's disease and it is mentioned in the sentence as a cause of postural hypotension.
Orthostatic hypotension [Orthostasis] was ameliorated 4 days after withdrawal of selegiline and totally abolished 7 days after discontinuation of the drug.	**Disease**: Orthostatic hypotension **Explanation**: a medical condition characterized by a sudden drop in blood pressure when standing up from a sitting or lying down position.	**Disease**: Orthostatic hypotension **Explanation**: Orthostatic hypotension is a type of low blood pressure that occurs upon standing up from a sitting or lying down position. It is mentioned in the sentence as a condition that was affected by the withdrawal and discontinuation of the drug.
Original Sentence[Perturbation]	**Predictions After (Zero-shot)**	**Predictions After (Few-shot)**
Selegiline [L-deprenyl]-induced postural hypotension in Parkinson's disease: a longitudinal study on the effects of drug withdrawal.	**Chemical**: L-deprenyl **Explanation**: L-deprenyl is a chemical substance used in the treatment of Parkinson's disease.	**Chemical**: L-deprenyl **Explanation**: L-deprenyl is a chemical since it is a medication used to treat Parkinson's disease and in the sentence it has been mentioned that it induced postural hypotension.
Orthostatic hypotension [Orthostasis] was ameliorated 4 days after withdrawal of selegiline and totally abolished 7 days after discontinuation of the drug.	**Disease**: Orthostasis **Explanation**: a medical condition characterized by lightheadedness or dizziness when standing up.	**Disease**: Orthostasis **Explanation**: Orthostasis is a medical condition where a person experiences lightheadedness or dizziness upon standing up. It is mentioned in the sentence that it was ameliorated after withdrawal of selegiline.

Table 4: Sample output predictions before (top) and after (bottom) perturbation for sentences containing target entities (T_E)[perturbed entities $(T_{E'})$] to show the difference in the quality of explanations under zero-shot and few-shot setup. We only show predictions for the target entities.

EXAMPLE 3: Text: """"In conclusion , any disease can occur in patients receiving continuous infusion of 5 - FU."""" Output: "entity": "5 - FU", "type": "chemical", "explanation": "5 - FU is a chemical since it is a cytotoxic chemotherapy medication used to treat cancer and in the sentence it has been mentioned that any disease can occur because of its continuous infusion."

Example 4: Text: """"In conclusion , any disease can occur in patients receiving continuous infusion of paracetamol. """" Output: "entity": "paracetamol", "type": "chemical", "explanation": "paracetamol is a chemical since it is a medication used to treat fever and mild to moderate pain and in the sentence it has been mentioned that any disease can occur because of its continuous infusion."

====== Text: """"text"""" Output:

E Implementation Details

We use "*gpt-3.5-turbo*(2023-03-15-preview)" model using OpenAI API key to obtain predictions for named entities and corresponding explanations for examples from the train-split for which triggers were collected by Lin et al. (2020). For each of the examples, we generate 3 perturbations per ground truth entity for **Alias, Verb, and Same Entity Type**, and 1 for **Random Entity, and Typo**[6]. To eliminate the randomness of predicted samples, we set the temperature to 0.

F Additional Results

Predicted entities may not be grounded in the input. We observe a few predictions wherein the predicted entities are not even present in the input but are relevant given the context. E.g. ChatGPT predicts 'schizophrenia' as one of the entities for *"NRA0160 and clozapine antagonized locomotor hyperactivity induced by methamphetamine (**Hcxd8rf**) in mice."* as 'clozapine' is used to treat schizophrenia.

G Results and Findings

Robustness depends on perturbation type and domain of perturbed entities. Table 1 show that in **zero-shot scenario** ChatGPT is more brittle on **Drug** or **Disease** (BC5CDR) perturbations (rare entities) as compared to that on widely known **Person** or **Location** entities in CONLL in terms of Δ Accuracy and Δ Faithfulness. Typo and Random entity substitutions seem too brittle indicated by high scores. However, in **few-shot scenario**, ΔAccuracy gradually decreases for almost all the perturbations in both datasets, indicating high robustness.
Transition of global and local explainability for the same entity prediction under perturbation. We observe (see Table 2) that overall, the globalness of explanations decreases while faithfulness (localness) to input increases due to perturbation. This provides us with an insight that when an entity is being perturbed, ChatGPT relies more on local context cues to detect entities. This holds

[6]Only 1 perturbation since it cannot have many variations

true for all types of perturbations in CONLL since person or location names are widely popular, hence before perturbation major predictions were pivoted on world knowledge. However, for Alias, Entity Type, Typo perturbations in BC5CDR, the explanations were more global and local before attack. Thus for the well-known entity types, the model chooses either local or global explanations, whereas after random perturbations, the model always prefer looking at contextual cues. Since the goal of **few-shot** experiments is to increase both localness and globalness in all the explanations of the predicted entities (G↑↑L↑↑), we notice that the performance improves significantly under **few-shot as shown in Table 2**. Sample output predictions for sentences containing target entities are shown in Table 4.

H Conclusion

We perform automatic evaluation of robustness of ChatGPT's predictions and explanations to input perturbations for the task of NER. We show that ChatGPT is more brittle on domain-specific entity perturbations compared to those on widely known entities. We also observe that the quality of explanations for the same entity *considerably differ* under various perturbations and the robustness and quality significantly improves using in-context (few-shot) learning.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. Exploring the landscape of distributional robustness for question answering models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5971–5987, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.441.

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

Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for ai. *Communications of the ACM*, 64(7):58–65, 2021.

Gabriel Bernier-Colborne and Phillippe Langlais. HardEval: Focusing on challenging tokens to assess robustness of NER. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1704–1711, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.211.

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

Sravan Bodapati, Hyokun Yun, and Yaser Al-Onaizan. Robustness to capitalization errors in named entity recognition. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pp. 237–242, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5531. URL https://aclanthology.org/D19-5531.

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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc.,

2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Jifan Chen and Greg Durrett. Robust question answering through sub-part alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1251–1263, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.98. URL https://aclanthology.org/2021.naacl-main.98.

Zhuang Chen and Tieyun Qian. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3685–3694, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.340. URL https://aclanthology.org/2020.acl-main.340.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023. URL https://arxiv.org/abs/2301.07597.

Mubin Ul Haque, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. "i think this is the most disruptive technology": Exploring sentiments of chatgpt early adopters using twitter data, 2022.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*, 2023.

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

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine, 2023.

Mohammad Khalil and Erkan Er. Will chatgpt get you caught? rethinking of plagiarism detection. *arXiv preprint arXiv:2302.04335*, 2023.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. The moral authority of chatgpt, 2023.

Nghia T. Le and Alan Ritter. Are large language models robust zero-shot coreference resolvers?, 2023.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL `https://aclanthology.org/2020.acl-main.703`.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*, 2023a.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

Xianzhi Li, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? an examination on several typical tasks, 2023b.

Bill Yuchen Lin, Dong-Ho Lee, Ming Shen, Ryan Moreno, Xiao Huang, Prashant Shiralkar, and Xiang Ren. Triggerner: Learning with entity triggers as explanations for named entity recognition. *arXiv preprint arXiv:2004.07493*, 2020.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. RockNER: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3728–3737, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.302. URL `https://aclanthology.org/2021.emnlp-main.302`.

Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*, 2023.

Stephen Mayhew, Nitish Gupta, and Dan Roth. Robust named entity recognition with truecasing pretraining, 2019.

Ishani Mondal. BBAEG: Towards BERT-based biomedical adversarial example generation for text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5378–5384, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.423. URL `https://aclanthology.org/2021.naacl-main.423`.

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

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*, 2023.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zheng, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model. *CoRR*, abs/2201.11990, 2022. URL `https://arxiv.org/abs/2201.11990`.

Zhongxiang Sun. A short survey of viewing large language models in legal aspect, 2023.

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. Lamda: Language models for dialog applications, 2022.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL `https://aclanthology.org/W03-0419`.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. Infobert: Improving robustness of language models from an information theoretic perspective, 2021.

Xuezhi Wang, Haohan Wang, and Diyi Yang. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4569–4586, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.339. URL `https://aclanthology.org/2022.naacl-main.339`.

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

Zhen Wang, Hongyi Nie, Wei Zheng, Yaqing Wang, and Xuelong Li. A novel tensor learning model for joint relational triplet extraction. *IEEE transactions on cybernetics*, PP, 2023b.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

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. Zero-shot information extraction via chatting with chatgpt, 2023a.

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