Are the predictions and explanations provided by ChatGPT on Entity Recognition task robust under Adversarial Attack?

Anonymous ACL submission

Abstract

ChatGPT has been increasingly used not only as a productivity tool but also to evaluate its performance on various NLP tasks. While prior works have vouched for its language understanding and generation capabilities, limited efforts have been made to assess the robustness of ChatGPT under adversarial perturbations. This work aims to evaluate the effect of input perturbations on the accuracy for prediction; quality of explanation for the prediction, and confidence in the prediction, for the most fundamental task of Information Extraction (IE) *i.e.,* Named Entity Recognition (NER). We present a systematic evaluation of the robustness of ChatGPT (under both zero-shot and few-shot setups) on two NER datasets using both automatic and human evaluations. Our findings suggest: ChatGPT is more brittle on **Drug** or **Disease** replacements (rare entities) as compared to the perturbations on widely known **Person** or **Location** entities; the quality of explanations for the same entity *considerably differ* under different types of "Entity-Specific" and "Context-Specific" perturbations, and it is overconfident for the majority of the incorrect predictions which could misguide the end-users, potentially breaching their trust in predictions.

1 Introduction

The rapidly evolving field of natural language processing (NLP) witnesses the upsurge of large language models (LLMs) (GPT3 (Brown et al., 2020b), LaMDA (Thoppilan et al., 2022) and PaLM (Chowdhery et al., 2022), etc.). Prompting these models has emerged as a widely adopted paradigm, given their superior zero-shot learning capability (Min et al., 2022). Moreover, with a proper instruction (Hegselmann et al., 2023; Ma et al., 2023), these LLMs achieve better performances on downstream NLP tasks. ChatGPT[1] is one of such powerful LLM that has attracted a huge volume of users ever since its inception. However, it is not clear whether it is *reliable* in the realistic applications in which entities or context words can be out of distribution of the training data, thereby calling attention to gauge its robustness. While previous efforts have evaluated various aspects of ChatGPT in law (Choi et al., 2023), ethics (Shen et al., 2023), education (Khalil and Er, 2023), verifiability (Liu et al., 2023) and reasoning (Bang et al., 2023), we focus on its robustness (Bengio et al., 2021) to adversarial input perturbations, which has not been thoroughly evaluated yet.

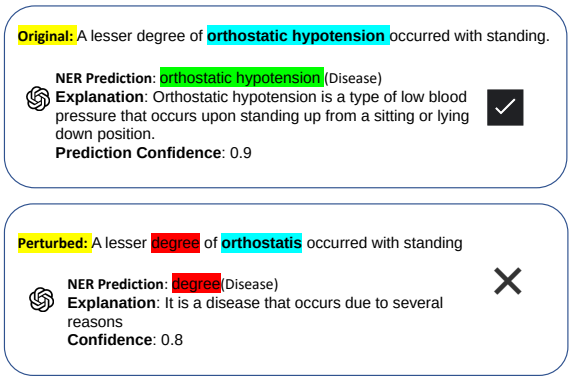


Figure 1: An example of sentence from BC5CDR in which the disease entity **orthostatic hypotension** has been perturbed with a synonym **orthostatis**. Before perturbation, the disease was correctly predicted and explained in the right way with high confidence (90%). After perturbation, *degree* has been incorrectly predicted as a disease entity with a wrong explanation. However, ChatGPT is nearly equally confident (80%) as the situation when it made a correct prediction.

Since ChatGPT is a black-box model that hardly provides any information about its training details, the generated responses can significantly influence user's trust (Deshpande et al., 2023; Huang et al., 2023). Hence the evaluation based on its sensitivity to input perturbations should also involve gauging the *reliability* of responses **under the light of robustness** by investigating its **prediction confi-**

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

dence and the **rationale**[2] behind its prediction.

In order to assess that, we focus on an elementary IE task, *i.e.*, Named Entity Recognition (NER). We make adversarial changes in the input data at both *"Entity-level"* (by replacing target entities with other entities of the **same semantic class in Wikidata**, **typo**, **alias**, **random string**"); and at the *"Context-level"* (by using pre-trained language models (e.g., BERT (Devlin et al., 2019)) to generate contextual **verb substitutions**). In this paper, we investigate the *reliablity* of ChatGPT's prediction under both zero-shot and few-shot settings along three dimensions, including RQ1) **Performance Shift under Attack:** How does the robustness of ChatGPT vary with domains and types of perturbations? Does ChatGPT make incorrect predictions for the examples which are easy for humans? RQ2) **Difference in Explanation Quality under Attack:** Is ChatGPT better at explaining its predictions on a *local-level* (grounded in the input) or a *global-level* (grounded in world knowledge) and how does that vary under attack under zero-shot setup? We also examine if the perturbation in the target entity causes a change in semantic similarity of other non-target entities explanation before and after attack. We have also assessed the difference in explanation quality of the same entities (*local-level* or *global-level*) before and after attack using both automatic and human evaluation. RQ3) **Variation in Confidence Calibration under Attack**: Is there any difference between prediction confidence between correct and incorrect predictions under attack?

In a nutshell, our contributions are four fold:

1. To the best of our knowledge, we are the first to comprehensively analyze the effect of adversarial perturbations on ChatGPT's predictions and rationale behind its prediction.

2. Our automatic evaluation reveals that under the light of robustness ChatGPT's predictions and faithfulness of explanations are less reliable on domain-specific entities compared to popular entities; and quality of explanations for the overlapping entities which are predicted both before and after attack also considerably vary, indicating less *reliability*.

3. Human evaluation further validates our findings from automatic evaluation and we throw some light on human's notion of informativeness of explanations, ease of entity prediction under perturbations and how does that correlate with the behavior of ChatGPT.

4. Even though ChatGPT is overconfident for incorrect predictions, its overconfidence can be significantly reduced using in-context learning; the quality of explanations (containing both local and global cues) also improve under few-shot setup.

2 Can we automatically generate 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 1. 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.[3]

[2] We use the terms *rationale* and *explanation*, *robustness* and *reliability* interchangably

[3] 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

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 1: 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.

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).

3 Experimental Setup

Datasets: We evaluate the explainability and NER capability of ChatGPT on CONLL-2003 (Tjong Kim Sang and De Meulder, 2003) and BC5CDR (Li et al., 2016) datasets by prompting ChatGPT (see A.2 for prompt) to obtain the predicted entities and corresponding explanations in a structured format. We only consider two types of entity predictions (PERSON, LOCATION) from the CONLL-2003 dataset (See A.1).

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.

Evaluation Criteria: We provide a comprehensive understanding of how we approach our research questions mentioned in Section 1 and evaluate the robustness of ChatGPT under adversarial perturbations. On a high-level, we define the following evaluation criteria to measure the same:

1. Performance Difference under Attack: Motivated by (Wang et al., 2021b; Mondal, 2021), we comprehensively evaluate the overall performance of ChatGPT on NER task and compare it when the inputs are adversarially perturbed. By examining the change in its performance across two situations, we seek to provide a detailed understanding of ChatGPT's reliability under adversarial attack.

2. Difference in Quality of Explanations under Attack: The explainability of ChatGPT is crucial for its application in real-world scenarios (Aghajanyan et al., 2021; Rajani et al., 2019); hence we ask ChatGPT to provide reasons for its predictions before and after adversarial attack.

3. Confidence Calibration under Attack: Calibration measurement is a crucial aspect to determine the predictive uncertainty of a model (Guo et al., 2023; Ulmer et al., 2022); a well-calibrated classifier must have predictive scores that reflect the probability of its correctness (Minderer et al., 2021; Thulasidasan et al., 2019). Here we aim to identify the uncertainties and confidence of ChatGPT (by prompting it to provide confidence score in the range of 0 to 1) in predicting named entities before and after adversarial perturbations.

Zero-shot and Few-shot Setup: To conduct a thorough evaluation of ChatGPT's capabilities on

NER task, we first measure its performance in the **Zero-shot scenario**. Then, we investigate how a **Few-shot Approach** or in-context learning (ICL) approach affects its performance. First, we manually design different zero-shot prompts since ChatGPT is sensitive to different prompts, then we choose the ones which provide maximally correct output on the non-perturbed sentences (S). Then we construct few-shot ICL prompts (See A.3 by selecting zero-shot prompt and randomly adding some samples from the corresponding training set.

How are the prompts designed? The prompts designed for zero-shot settings consist of the following integral main elements: the task instruction, candidate target labels, output format description and the input text. The task instruction describes the specific IE sub-task where we ask the model to provide confidence of its prediction and the explanation behind its prediction; candidate target labels are the types of target information, such as entity types; the output format description specifies the format of outputs to facilitate easy parsing. In the few-shot setting, we also provide some demonstration examples, which can also provide the chain-of-thought explanation and confidence of prediction. Since we assume a combination of local+global explanations are the most useful ones for NER prediction, we combine both wikipedia description of the entities and local contextual cues in the explanation behind predicting it as entity. (See Appendix A.2, A.3 for prompts used)

3.1 Implementation Details

We use "*gpt-3.5-turbo*" 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**[4]. To eliminate the randomness of predicted samples, we set the temperature to 0.

4 How to estimate *Reliability*?

We perform both **automatic** and **manual** evaluations of the predictions and generated explanations separately for target (t) (in 4.1), non-target (nt) (in 4.1), and overall entities (T)[5]. Based on **automatic** evaluation, we come up with the following evaluation metrics that align well with answering our research questions (as laid out in §1):

4.1 Automatic Evaluation

Is there any effect of ChatGPT's NER prediction on the target entity? We hypothesize that after each type of perturbation (§2), we can observe some differences in the reliability of predicting target entities along with explanations provided in support of its predictions. In this case, we generate S' by perturbing T_E with T'_E, and evaluate if ChatGPT can predict T'_E correctly in S', since from the contextual cues a smart human can predict the entity correctly instead of just predicting based on prior knowledge about the entity.

A: Accuracy Before and After Attack: For each type of perturbation, we measure the difference in the accuracy of predicting the target entity T_E before prediction and T'_E after perturbation (Δ Accuracy). This is measured with respect to gold annotated entities in train split. Lesser the Δ Accuracy, higher is the robustness.

B: Faithfulness of Explanation Before and After Attack: Ideally, entity prediction based on contextual cues should be *faithful* to the input context even after adversarial perturbation. Thus, we measure the difference in the faithfulness (*local-level* explanation measured in terms of cosine similarity of explanation with the input query) of explanation for the target entity prediction before (T_E) and after (T_E) perturbation (Δ Faithfulness).

C: Similarity of Explanation Before and After Attack: We measure the cosine similarity of the explanation generated for the prediction of the target entity before (T_E) and after (T_E) perturbations.

Is there any effect of ChatGPT's NER prediction on the non-target entities? Here we aim to analyze whether after perturbing T_E with T'_E, ChatGPT's predictions on O_E alters. In other words, our primary goal is to verify if ChatGPT can successfully ignore the target entity perturbation and generate similar predictions and explanations for the other entities in S and S'. Here also a smart human can predict the other non-perturbed entities based on the contextual cues.

A: F1-Score Before and After Attack: We measure the difference in the F1 of the prediction of

[4]Only 1 perturbation since it cannot have much variations

[5]We consider entities to be case-sensitive for accuracy computation as NER can be considered as a span (grounded in input) prediction task

	Effect on Target Entity			Effect on non-target Entities		Overall Effect
	Δ Accuracy	Δ Faithfulness	Similarity	Δ F1	Δ Faithfulness	Δ F1
Entity-Level						
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
Context-Level						
Verb Substitution	-	-	-	0.01 / 0.01	0.01 / 0.01	0.02 / 0.02

Table 2: Assessment of Robustness of NER predictions, Faithfulness of its predictions to input (extrinsic) and similarity between the explanation generated for the original and perturbed instances in the form of (zero-shot / few-shot) prediction performances on the BC5CDR Dataset.

	Effect on Target Entity			Effect on non-target Entities		Overall Effect
	Δ Accuracy	Δ Faithfulness	Similarity	Δ F1	Δ Faithfulness	Δ F1
Entity-Level						
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
Context-Level						
Verb Substitution	-	-	-	0.01 / 0.01	0.02 / 0.01	0.01 / 0.02

Table 3: Assessment of Robustness of NER predictions, Faithfulness of its predictions to input (extrinsic) and similarity between the explanation generated for the original and perturbed instances in the form of (zero-shot / few-shot) prediction performances on the CONLL dataset.

non-target entities (O_E) (with respect to gold standard annotations) before and after perturbation (Δ F1). Lesser the Δ F1, higher is the robustness.

B: Faithfulness of Explanation Before and After Attack: We measure the difference in the faithfulness (measured in terms of cosine similarity of explanation with the input query) of explanation for the target entity (t) prediction before and after perturbation (Δ Faithfulness). For a certain type of perturbation in §2, if there are x inputs each containing n entities on average and each entity has k different perturbations, then each of the above-mentioned metrics is reported using the weighted average rule: $(n * k)/x$. We use these metrics to answer our **RQ1** in §1. We **approximate how the explanation of an entity is grounded to world knowledge (*global-level*)** by obtaining the entity description from wikipedia[6] and calculating the similarity of generated explanation with respect to the summary. Then **we analyze the effect of perturbations on generating *global* and *local* explanations** for the common non-target entities which are predicted both before and after perturbations in order to answer **RQ2**. Here we assume that whenever explanation's faithfulness (both *local* or *global*) score changes for the same entity as before, we try to measure when there is an increase or decrease and enumerate those for both zero-shot and few-shot approach in Table 4.

Confidence Calibration We estimate confidence in terms of a probability value (0-1) indicating the likelihood of belonging to a specific category, for both the correct and incorrect samples. We estimate overconfidence as the difference between confidence scores on correct and incorrect predictions (ΔC). We aim to evaluate how ΔC varies for different perturbations and if that gets reduced due to in-context learning. After that, **Manual** analysis of the explanations is done for the target and non-target entities before and after perturbation. We sample 5 inputs and their perturbations for each of the four possibilities (correct/incorrect prediction before/after the perturbation) to answer **RQ3** and further confirm findings obtained for **RQ2** from automatic evaluation.

5 Automatic Evaluation

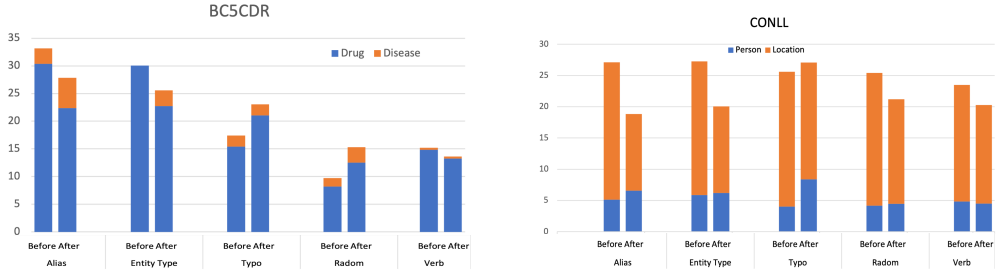
How sensitive is k in k-shot Learning? We observe that one example for each type of perturbation and for each entity type is imperative for achieving better robustness.

Robustness depends on perturbation type and domain of perturbing entities. Table 2 and Table 3 show that under **zero-shot scenario** ChatGPT is more brittle on **Drug** or **Disease** replacements (rare entities) compared to the perturbations on widely known **Person** or **Location**

[6] https://pypi.org/project/wikipedia/

	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↓↓
BC5CDR								
Alias	**0.54**	0.26	0.17	0.02	**0.57**	0.20	0.18	0.03
Same Entity Type	**0.61**	0.22	0.13	0.02	**0.48**	0.29	0.16	0.06
Typo	**0.36**	0.26	0.26	0.10	**0.46**	0.19	0.30	0.03
Random	0.39	**0.43**	0.17	0.00	**0.46**	0.15	0.19	0.19
Verb	0.24	**0.48**	0.24	0.02	**0.48**	0.28	0.20	0.02
CONLL								
Alias	0.21	**0.58**	0.06	0.13	**0.60**	0.24	0.05	0.11
Same Entity Type	0.21	**0.48**	0.15	0.16	**0.46**	0.23	0.15	0.16
Typo	0.24	**0.40**	0.15	0.20	**0.34**	0.30	0.15	0.20
Random	0.11	**0.63**	0.15	0.10	**0.60**	0.20	0.20	0.11
Verb	0.22	**0.56**	0.07	0.13	**0.56**	0.22	0.07	0.13

Table 4: shows the change in the generated explanations due to the predictions of common entities before and after attack. Here ↑↑ and ↓↓ indicate increase and decrease after perturbation respectively.



(a) BC5CDR (b) CONLL

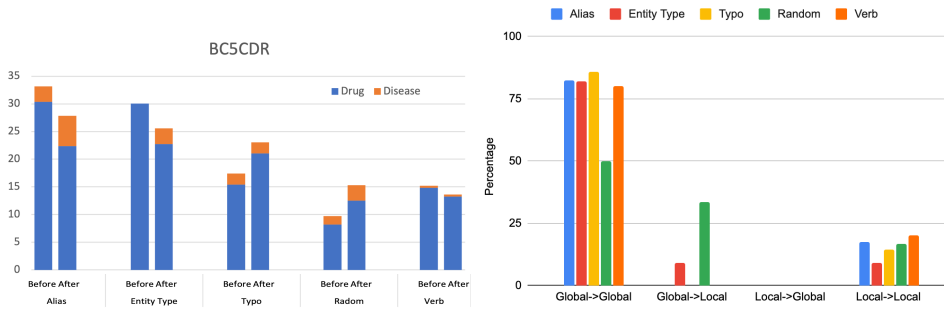
Figure 2: Percentage of examples **Before** and **After** attack for which the explanations are less informative such as "refers to a country/person", "it is a chemical compound/substance" for BC5CDR and CONLL datasets.

	Confidence of Correct+Incorrect Predictions			
	Zero-Shot		Few-Shot	
	ΔC BA	ΔC AA	ΔC BA	ΔC AA
BC5CDR				
Alias	0.12	0.05	0.18	0.09
Typo	0.10	0.08	0.15	0.13
Random	0.11	0.08	0.15	0.11
Same Type	0.07	0.11	0.14	0.18
Verb	0.04	0.08	0.01	0.03
CONLL				
Alias	0.05	0.03	0.08	0.06
Typo	0.12	0.08	0.15	0.11
Random	0.21	0.18	0.23	0.20
Same Type	0.07	0.11	0.15	0.17
Verb	0.05	0.06	0.01	0.05

Table 5: shows the change in the average confidence scores between the correct and incorrect predictions before attack (BA) and after attack (AA) in both zero-shot and few-shot predictions.

entities in CONLL in terms of Δ Accuracy and Δ Faithfulness. Besides, Typo and Random entity substitution seems too brittle in terms of both these metrics. Using human evaluation, we wanted to confirm if the incorrectly predicted examples are also difficult to be identified by the humans. However, we notice that under **few-shot scenario**, ΔAccuracy gradually decreases for almost all the perturbations in both the datasets, indicating high robustness.

Transition of global and local explainability for same entity prediction under attack. Based on the **zero-shot** results in Table 4, we observe that overall, the globality of explanations decreases while faithfulness 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 models always prefer looking at contextual cues. Since while performing **few-shot** experiments, our goal has been to increase both locality and globality in all the explanations of the predicted entities (G↑↑L↑↑), we notice that the performance improves significantly under **few-shot as shown in Table 4**. Sample output predictions for sentences containing target entities in order to show the difference in the quality of explanations under zero-shot and



(a) Target Entities (b) Non-Target Entities

Figure 3: Percentage of (input, perturbed input) pairs with change in type of explanations for (i) target and (ii) non-target entities in BC5CDR.

few-shot setup are shown in Table 6.

Overconfidence of ChatGPT on incorrect predictions can be alleviated to some extent using In-context Learning. Table 5 shows the difference in the average confidence scores of the correct and incorrect predictions before attack (BA) and after attack (AA) in both zero-shot and few-shot predictions. It can be observed that under zero-shot scenario, ChatGPT is highly overconfident on the incorrect predictions, causing very less difference between correct and incorrect predictions. However, for all types of perturbations, the few-shot setup improves reliability in predictions even after adversarial attack by increasing the gap between correct and incorrect predictions (except verb substitutions) (ΔC). Moreover, ΔC is being reduced after pertubations, indicating more overconfidence in incorrect predictions due to attack.

6 Manual Evaluation

We manually evaluate explanations for a subset of examples in BC5CDR and CONLL.

Global vs. Local explanation. Figure 3a presents the change in the type of explanation under attack for target entities (see Figure 3b for non-target entities). While, under zero-shot scenario, majority of the explanations are grounded in world knowledge (global) before and after the attack across all the perturbation types, we observe that 33% of the explanations (BC5CDR) and 24.45% (CONLL) change from global to local-level for Random perturbations showing that local context is required for predictions in such cases. E.g., the explanation generated for *"Recently, we found that therapy with **r30s1k0** and L-dopa was associated with selective systolic orthostatic hypotension which was abolished by withdrawal of **r30s1k0**."* is *"R30s1k0 is a chemical compound used in therapy."* while it was *"This is a medication used to treat Parkinson's disease."* before the perturbation. This connects to our zero-shot findings for automatic evaluation in 5. Besides, we found that majority of content in explanations generated under few-shot setup contain both globality and locality cues.

Prediction of target entities from context is easier for humans than ChatGPT. While ChatGPT is able to correctly predict 67.8% and 75% of perturbed examples that are easy for humans for Alias, and Same Entity Type, respectively, only 45%, and 47% are predictable for Random and Typo perturbations. This indicates that ChatGPT finds it harder to perform contextual predictions which are easier for humans. E.g., it is easy for humans to understand that the typo **legilinese** for **selegiline** is a chemical in *"CONCLUSION: This study confirms our previous finding that **legilinese** in combination with L-dopa is associated with selective orthostatic hypotension."*, however, the model is not able to predict this entity. Besides, in case of random perturbation, it is easy for humans to understand from context that **8076mhq** is a type of disease in *"METHODS: The cardiovascular responses to standing and head - up tilt were studied repeatedly in **8076mhq** patients receiving selegiline and as the drug was withdrawn."* however, ChatGPT fails to predict this entity. This shows that ChatGPT is not able to understand the type of entity from the context which humans can do easily.

Informativeness of the explanations change for the same non-target entity under attack. We define informativeness of explanations based on its source (world knowledge or from input) and amount of information, rated by a human on the basis of Likert Scale (1-5) where the input-grounded

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. **Confidence**: 0.8	**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. **Confidence**: 0.8
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. **Confidence**: 0.9	**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. **Confidence**: 0.9

Table 6: Sample output predictions for sentences containing target entities (T_E)[perturbed entities (T'_E)] in order to show the difference in the quality of explanations under zero-shot and few-shot setup. We only show predictions for the target entities. The model is equally confident in the prediction irrespective of the informativeness (more informative in few-shot) of the explanation. We show the explanations after perturbation in Table 8.

explanations are considered more informative than global explanations, and a combination of these as the most informative. E.g. when Israel is replaced with 'Mount lebanon' in *"Israel's Channel Two television said Damascus had sent a "calming signal" to Israel."*, the explanation for 'Damascus' changes from 'refers to the capital city of Syria' to less informative 'refers to a geographical location'. We observe (Figure 2) that % of least informative explanations (such as 'France is a country') increases (or comparable) for person (disease) type in CONLL (BC5CDR) after attack for all types of attacks while it decreases for location (drug) type except for **Typo** and **Random** attacks in BC5CDR.

Generalizability of method on other LLMs Our methodology is generalizable for analyzing robustness of any LLMs. We ran trials on some other LLMs: OPT 176-B9 (Zheng et al., 2022), Flan-T5-xxl (Chung et al., 2022), GPT-3 (Brown et al., 2020a) using prompts described in A.2.

7 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., 2020a), 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 and Qian, 2020), and financial challenges (Sun, 2023; Li et al., 2023). 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) and other NLP tasks (Bian et al., 2023). A number of studies have been done in order to evaluate and improve the robustness of LLMs (Chen and Durrett, 2021; Awadalla et al., 2022; Wang et al., 2021a, 2022). 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 and 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 quality of explanations vary due to perturbations, which we try to fill up in this work.

8 Conclusion

We perform automatic and manual evaluation of explainability and IE capabilities of ChatGPT under the light of robustness before and after perturbations in the input. We find that ChatGPT is more brittle on domain-specific entity perturbations compared to the ones on widely known entities. Besides, we observe that the quality of explanations for the same entity *considerably differ* under different types of perturbations and the quality can be significantly improved using in-context learning. Even though ChatGPT is overconfident for incorrect predictions, its overconfidence can be significantly reduced using in-context learning. To the best of our knowledge, we are the first to comprehensively analyze the effect of adversarial perturbations on ChatGPT's predictions and rationale behind its prediction on an IE Task.

Limitations

While we analyze the faithfulness of the explanations with respect to the input, we do not evaluate if the global explanations are factual. During the manual evaluation, we observe that some of the perturbations resulted in invalid sentences or changed the meaning of the input, leaving this investigation for future work.

Ethics Statement

Our method does not include any content that has potential risks or harms as we are anlayzing the outputs of an exisitng model, ChatGPT. However, we acknowledge and condemn the malicious use of outputs of such AI systems to alter the opinions of the stakeholder and that these systems might generate biased outputs that needs to be considered before using them for real-world applications.

Acknowledgements

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTEX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. 2022. Exploring the landscape of distributional robustness for question answering models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5971–5987, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Gabriel Bernier-Colborne and Phillippe Langlais. 2020. HardEval: Focusing on challenging tokens to assess robustness of NER. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1704–1711, Marseille, France. European Language Resources Association.

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.

Sravan Bodapati, Hyokun Yun, and Yaser Al-Onaizan. 2019. Robustness to capitalization errors in named entity recognition. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 237–242, Hong Kong, China. 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. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jifan Chen and Greg Durrett. 2021. 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*, pages 1251–1263, Online. Association for Computational Linguistics.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3694, Online. Association for Computational Linguistics.

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

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

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

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

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

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.

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

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

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

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

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. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

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

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. 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*, pages 3728–3737, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models.

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc.

Ishani Mondal. 2021. 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*, pages 5378–5384, Online. Association for Computational Linguistics.

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. 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. *arXiv preprint arXiv:2304.08979*.

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. 2022. Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model. *CoRR*, abs/2201.11990.

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

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.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in nlp: A study on the impact of method data scarcity.

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

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021b. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. *CoRR*, abs/2111.02840.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. 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*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

	Accuracy (BC5CDR)	Accuracy (CONLL)
OPT	0.45	0.64
Flan-T5-xxl	0.55	0.67
GPT3	0.73	0.77
ChatGPT	0.78	0.83

Table 7: Generalizability of our approach (using accuracy of entity predictions) on three other LLMs except chatgpt. Stoked by the best performance of **GPT3.5-turbo**, we conduct all our experiments in the main paper using that model.

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

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

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. 2022. Emergent abilities of large language models.

A Example Appendix

A.1 Why did we consider only Person and Location Types?

Our rationale behind experimenting with only PER and LOC was from the perspective of evaluation. In our paper, evaluating the quality of predictions and the explanation behind its predictions is an important contribution. We wanted to evaluate the difference in the faithfulness of explanations before and after predictions, and we classify the explanations into local and global types. We approximate how the explanation of an entity is grounded to world knowledge (global-level) by obtaining the entity description from wikipedia and calculating the similarity between the generated explanation with respect to the summary. However, for the ORG and MISC types, there is no way of approximating the global knowledge, as we cannot achieve the Wikipedia descriptions/definitions of these entity types in a straightforward way. In order to make fair comparisons with all the completed experiments of our robustness evaluation framework, we omit these two types from our evaluation set.

A.2 Dealing with Prompt Sensitivity

Some of the seed prompts that we started with are:

1. "Find the named entities of type ""person"" or ""location"" in text. You should format your response as a list of JSON objects with keys as ""type"", ""entity"", ""explanation"", ""confidence"" and values as ""type of the identified entity"", ""identified entity"", ""explanation about tagging it as that type of entity"", and ""your confidence in identifying the entity as its type"", respectively. Ensure that the identified entities can only be words or phrases present in the provided text. Confidence is a real value between 0 and 1. Text: """""""text"""""""

2. "What are the different ""person"" or ""location"" entities present in the text? You should format your response as a list of JSON objects with keys as ""type"", ""entity"", ""explanation"", ""confidence"" and values as ""type of the identified entity"", ""identified entity"", ""explanation about tagging it as that type of entity"", and ""your confidence in identifying the entity as its type"", respectively. Ensure that the identified entities can only be words or phrases present in the provided text. Confidence is a real value between 0 and 1. Text: """""""text"""""""

3. "What are the different ""person"" or ""location"" entities present in the text? You should format your response as a list of JSON objects with keys as ""type"", ""entity"", ""explanation"", ""confidence"" and values as ""type of the identified entity"", ""identified entity"", ""explanation about tagging it as that type of entity"", and ""your confidence in identifying the entity as its type"", respectively. Ensure that the identified entities can only be words or phrases present in the provided text. Confidence is a real value between 0 and 1. Text: """""""text"""""""

4. "Identify the different ""person"" or ""location"" entities present in the text? Person is any individual, corporation, company, foundation, association, labor organization, firm, partnership, society, joint stock company, group of organizations, or state or local government and Location refers to the actual physical address. You should format your response as a list of JSON objects with keys as ""type"", ""entity"", ""explanation"", ""confidence"" and values as ""type of the identi-

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. **Confidence**: 0.9	**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. **Confidence**: 0.9
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. **Confidence**: 0.8	**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. **Confidence**: 0.8

Table 8: Sample output predictions for sentences containing target entities (T_E)[perturbed entities (T'_E)] in order to show the difference in the quality of explanations under zero-shot and few-shot setup. We only show predictions for the target entities. The model is equally confident in the prediction irrespective of the informativeness of the explanation.

fied entity"", ""identified entity"", ""explanation about tagging it as that type of entity"", and ""your confidence in identifying the entity as its type"", respectively. Ensure that the identified entities can only be words or phrases present in the provided text. Confidence is a real value between 0 and 1. Text: """""""text"""""""

We have paraphrased our prompts using sentence-transformer based paraphrasing tool, then we post-edited using humans. Finally, we ran pilot studies with these prompts and finalized one prompt upon manual examination of the output quality and structure. The output quality was being judged on a scale of 1-5 using a Likert-Scale based human judgement technique.

Our finally chosen prompt looks like:
"Identify named entities of type ""person"" or ""location"" in the below text delimited by triple quotes. Format your response as a list of JSON objects with keys as ""type"", ""entity"", ""explanation"", ""confidence"" and values as ""type of the identified entity"", ""identified entity"", ""explanation of why it is an entity of that type"", and ""your confidence in identifying the entity as its type"", respectively. Ensure that the identified entities can only be words or phrases present in the provided text. Confidence is a real value between 0 and 1. Text: """""""text""""""""" An example text: "Only France and Britain backed Fischler 's proposal ."

A.3 Few-shot Prompts

We have included 4 examples in the few-shot prompt for each perturbation experiment. One example with original input and another with its perturbed version for all the entity types considered in that particular dataset. For example if we consider the perturbation type as "Typo replacement" in BC5CDR dataset where we want to evaluate the quality of predictions and explanations for entity types: "Disease and Drug", a sample few-shot prompt will be like: prompt=f""" 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", "explanation", "confidence" and values as "type of the identified entity", "identified entity", "explanation of why it is an entity of that type", and "your confidence in identifying the entity as its type", respectively. Ensure that the identified entities can only be words or phrases present in the provided text. Confidence is a real value between 0 and 1. Use the following examples as a guide:

EXAMPLE 1: Text: ¨¨N̈one 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.", "confidence": 0.7

EXAMPLE 2: Text: ¨¨N̈one of the patients had decompensated ilevr disease¨¨. Output: "entity": "ilevr" , "type": "disease", "explanation": "ilevr disease (it is a typo) is a disease because it a widely known disease and in the sentence it is mentioned that patients did not have decompensated this disease.", "confidence": 0.8

EXAMPLE 3: Text: ¨¨Ïn conclusion , any disease can occur in patients receiving continuous infusion of 5 - FU.¨¨Ötput: "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.", "confidence": 0.8

Example 4: Text: ¨¨Ïn conclusion , any disease can occur in patients receiving continuous

infusion of F-5 U. ¨¨Öutput: "entity": "F-5 U" , "type": "chemical", "explanation": "F-5 U (probably a typo) 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.", "confidence": 0.7

====== Text: ¨¨ẗext¨¨Öutput: """

Here, we manually examine the examples and the order in which we should place the few-shot examples as in-context prompts, ran pilot studies with these prompts and finalized one prompt upon manual examination of the output quality and structure.

B 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.