On Robustness of Prompt-based Semantic Parsing with Large Pre-trained Language Model: An Empirical Study on Codex

Terry Yue Zhuo[1][2] and **Zhuang Li**[2*]
Yujin Huang[2] and **Fatemeh Shiri**[2]
Weiqing Wang[2] and **Gholamreza Haffari**[2] and **Yuan-Fang Li**[2]
[1]CSIRO's Data61, Australia
[2]Monash University, Australia
{terry.zhuo, zhuang.li}@monash.edu

Abstract

Semantic parsing is a technique aimed at constructing a structured representation of the meaning of a natural-language question. Recent advances in language models trained on code have shown superior performance in generating these representations compared to language models trained solely on natural language text. The existing fine-tuned neural semantic parsers are vulnerable to adversarial attacks on natural-language inputs. While it has been established that the robustness of smaller semantic parsers can be enhanced through adversarial training, this approach is not feasible for large language models in real-world scenarios, as it requires both substantial computational resources and expensive human annotation on in-domain semantic parsing data. This paper presents the first empirical study on the adversarial robustness of a prompt-based semantic parser based on CODEX, a state-of-the-art (SOTA) language model trained on code. Our results demonstrate that the large language model of code is vulnerable to carefully crafted adversarial examples. To overcome this challenge, we propose methods for enhancing robustness without requiring substantial amounts of labelled data or intensive computational resources.

1 Introduction

Semantic parsing is a technique that transforms natural-language utterances (NLs) into machine-readable logical forms (LFs) and has been widely applied in various research fields, such as code generation, question-answering systems, and dialogue systems (Kamath and Das, 2018). Most current state-of-the-art semantic parsers are deep-learning models trained in a supervised manner using in-domain data. However, this approach requires a large amount of in-domain semantic parsing data, which can be costly to obtain (Bapna et al., 2017).

To address this issue, prompt-based semantic parsers based on large pre-trained language models, such as Codex (Chen et al., 2021) and GPT-J (Wang and Komatsuzaki, 2021), have become a new choice for semantic parsing applications. Prompt-based semantic parsers learn to solve a new task by in-context learning, instructing the parsers to generate correct LFs by constructing the prompt with a few demonstration examples. Such a method can significantly lower the cost of annotations by including only a few exemplars in the prompt and achieve comparable results to fully-supervised semantic parsers (Shin and Van Durme, 2022).

Recent studies (Huang et al., 2021; Pi et al., 2022) show that fully-supervised semantic parsers are vulnerable to adversarial attacks, which perturb input sentences into their semantic equivalent adversaries to mislead models to produce attacker-desired outputs. Hence, to mitigate such attacks, various adversarial training methods (Tramer and Boneh, 2019; Shafahi et al., 2019; Ganin et al., 2016; Shafahi et al., 2020) have been proposed to improve the adversarial robustness of the semantic parsers. In light of this, two main questions naturally arise: (1) *Do prompt-based semantic parsers based on large pre-trained language models also suffer from adversarial attacks?* (2) *If so, how can we improve the robustness of the large prompt-based semantic parsers?*

To address the former question, we evaluate the prompt-based semantic parsers on several evaluation sets built by different perturbation approaches mentioned in the AdvGLUE (Wang et al., 2021) dataset. Adopting the adversarial evaluation metrics proposed by Huang et al. (2021), it is found that the prompt-based semantic parsers are vulnerable to various types of adversarial attacks.

According to the experimental results from the first step, we perform a three-fold experiment to answer the latter questions. The first aspect of the study aims to determine if the inclusion of

---

*corresponding author

1090

additional examples within the prompt during in-context learning improves the robustness of prompt-based parsers. This hypothesis is based on prior research that has demonstrated that the increase in the size of the training data results in an enhancement of robustness in fully-supervised models (Pang et al., 2019). The second part of the study aims to determine if the integration of few-shot adversarial examples within prompts can improve the robustness of Codex. This was based on the observation that conventional adversarial training methods often include adversarial examples within the training set (Miyato et al., 2016; Tramer and Boneh, 2019). Finally, the third part of the study aims to evaluate if sampling methods other than random sampling can select more effective examples that improve the robustness of prompt-based parsers.

In this work, we perform a series of experiments to probe CODEX, a large pre-trained model trained on code, on two semantic parsing benchmarks, GeoQuery (Zelle and Mooney, 1996) and Scholar (Iyer et al., 2017). Our key findings from the above experiments are as follows:

- Prompt-based semantic parsers are vulnerable to adversarial examples, particularly the ones crafted by sentence-level perturbations.

- In-context learning with more demonstration examples in the prompt can improve the in-domain robustness of prompt-based parsers.

- Augmenting the prompt with adversarial examples has limited effect in improving the robustness of prompt-based parsers.

- The few-shot example sampling strategy with higher language complexity can result in stronger robustness for the prompt-based parsers.

2 Related Work

Prompt-based Learning. Prompt-based learning is an alternative approach to supervised learning that aims to reduce the reliance on large human-annotated datasets (Liu et al., 2021). Unlike traditional supervised models, which estimate the probability of an output given an input text, prompt-based learning models estimate the probability of the text directly. This is achieved by applying prompt functions to modify the input text into various prompt templates with unfilled slots. By filling these slots, various Natural Language Processing (NLP) tasks can be completed, such as common-sense reasoning (Kojima et al., 2022), self-rationalization (Marasović et al., 2021), and text style transfer (Suzgun et al., 2022). The development of prompt-based methods has enabled zero-shot and few-shot learning in a variety of artificial intelligence domains (Ramesh et al., 2021; Yang et al., 2022; Sanghi et al., 2022). Recent research has also evaluated the capabilities of few-shot prompt-based learning for semantic parsing (Shin and Van Durme, 2022; Roy et al., 2022a; Drozdov et al., 2022). Our contribution extends the current research by investigating the effect of prompts comprising only a limited number of examples on the robustness of prompt-based semantic parsers.

Adversarial Robustness. Neural networks have achieved impressive performance across various domains. However, as demonstrated by Szegedy et al. (2014), neural models are vulnerable to adversarial examples. Adversarial attacks in NLP normally take on various forms, including character-level manipulations (Hosseini et al., 2017; Ebrahimi et al., 2018; Belinkov and Bisk, 2018; Gao et al., 2018; Eger et al., 2019; Boucher et al., 2022), sentence-level rewriting (Iyyer et al., 2018; Ribeiro et al., 2018; Zhao et al., 2018), and adversarial word substitutions (Alzantot et al., 2018; Liang et al., 2018; Zhang et al., 2019).

There has been an increasing interest in defending against adversarial attacks in large language models via adversarial training (Yi et al., 2021; Ross et al., 2022; Bartolo et al., 2021; Guo et al., 2021). Adversarial training involves incorporating adversarial examples in the training set, thus making the model robust to such attacks. However, adversarial training can sometimes negatively impact the generalization ability of the neural models (Raghunathan et al., 2019; Min et al., 2021).

3 Robustness Evaluation for Prompt-based Semantic Parsing

This section gives an overview of our evaluation framework, including the methods of constructing the evaluation corpora and the evaluation metrics to evaluate the robustness of the prompt-based semantic parser.

Linguistic Phenomenon	Samples (S̶t̶r̶i̶k̶e̶t̶h̶r̶o̶u̶g̶h̶ = Original Text, **red** = Adversarial Perturbation)
Typo (Word-level)	**NL**: what can you t̶e̶l̶l̶**te11** me about t̶h̶e̶**th e** population of missouri
Substitution (Word-level)	**NL**: what c̶a̶n̶**will** you tell me about t̶h̶e̶**a** population of missouri
Paraphrase (Sent.-level)	**NL**: w̶h̶a̶t̶ ̶c̶a̶n̶ ̶y̶o̶u̶ ̶t̶e̶l̶l̶ ̶m̶e̶ ̶a̶b̶o̶u̶t̶ ̶t̶h̶e̶ ̶p̶o̶p̶u̶l̶a̶t̶i̶o̶n̶ ̶o̶f̶ ̶m̶i̶s̶s̶o̶u̶r̶i̶ **What information can you provide on Missouri's population?**

Table 1: **Examples from Robustness Evaluation Set**. We show 3 examples from `GeoQuery`. These examples are generated with three different perturbations, and they all can successfully change the predictions of CODEX.

3.1 Construction of the Evaluation Corpus

A robust prompt-based semantic parser should be able to parse both the utterances and their adversarial counterparts into correct LFs. As proposed by Huang et al. (2021), an adversary of an utterance for a semantic parser is defined as i) an utterance with the same semantic meanings as the original one given the human judgment and ii) an utterance on which the semantic parser cannot produce correct LF. Therefore, to evaluate the robustness of prompt-based semantic parsers, we craft the robustness evaluation sets by perturbing the original utterances in existing benchmark datasets with multiple adversarial perturbation methods. Such perturbations should not alter the semantics of the original utterances. Each example in a robustness evaluation set is a perturbed utterance paired with its ground-truth LF. Next, we introduce the details of each perturbation method and how we guarantee the perturbations do not change the semantics. Table 1 illustrates some meaning-preserved utterances after perturbation in the robustness evaluation set of `GeoQuery` based on different perturbation methods. More examples can be found in Appendix B.

3.1.1 Adversarial Perturbations

Following the principles as in Wang et al. (2021) to design adversarial attacks, we perform five word-level perturbations and two sentence-level perturbations to generate seven robustness evaluation sets for the standard evaluation set in each benchmark.

Word-level Perturbations.

- **Typo-based (TB)** uses TextBugger (Li et al., 2018) to replace two words in each utterance with the typos.

- **Random Deletion (RD)** randomly deletes two words in the utterance.

- **Random Swap (RS)** swaps the positions of two random words in each utterance.

- **Context-aware Substitution (CS)** leverages RoBERTa (Liu et al., 2019) to substitute two random words with their synonyms.

- **Context-aware Insertion (CI)** inserts two most probable words selected by RoBERTa at two random positions in each utterance.

Sentence-level Perturbations.

- **Rewriting-based (RB)** chooses Quillbot[1] (Fitria, 2021), a state-of-the-art (SOTA) commercial paraphrasing model, to rewrite the complete utterances. Quillbot has been demonstrated as an effective tool to paraphrase utterances in semantic parsing data (Shiri et al., 2022).

- **Distraction-based (DB)** appends interrogation statements to the end of each NL, inspired by StressTest (Naik et al., 2018). Specifically, we design the following interrogation statements: "who is who; what is what; when is when; which is which; where is where", in which the selected interrogative words are more likely to appear in the utterance.

3.1.2 Data Filtering

In order to ensure that the perturbed examples preserve the meaning of the original NL, we design a two-stage evaluation process:

Step1: We first generate 20 adversarial examples against the original NL for each perturbation method and choose the top 10 candidates ranked based on text similarity scores between the original and the perturbed ones, which are calculated by Sentence-BERT (Reimers and Gurevych, 2019).

---

[1] https://www.quillbot.com/paraphrasing

1092

Step2: We engage human experts to select the best one among the 10 adversarial candidates produced in **Step1**.

3.2 Evaluation Metrics

Since the output LFs of the prompt-based language models may not follow the same naming convention (Shin et al., 2021; Shin and Van Durme, 2022) as the ground truth, previous string-based evaluation metrics, including BLEU (Papineni et al., 2002) and Exact Match (Poon and Domingos, 2009), are not suitable for prompt-based semantic parsers. Therefore, we follow Rajkumar et al. (2022) to report the *execution accuracy*, which is based solely on the execution correctness of the LFs on the test sets, for the purpose of robustness evaluation.

Following Huang et al. (2021), we report the experiment results with three variants of execution accuracy, namely *standard accuracy*, *perturbation accuracy* and *robust accuracy*:

- **Standard Accuracy** is measured on the standard (original) test sets.

- **Perturbation Accuracy** tests the performance of the model on perturbed test sets.

- **Robust Accuracy** is defined as $n/|R_{eval}|$. R_{eval} denotes a subset of the perturbed test sets, and n is the number of the utterances in R_{eval} that are parsed correctly. More specifically, R_{eval} consists of the examples whose counterparts before perturbation are parsed correctly. Intuitively, Robust Accuracy estimates the quantity of cases that a parser can successfully parse before perturbation but cannot do so after perturbation, and hence shows the robustness of the parsers against adversarial perturbation.

4 Improving Robustness of Prompt-based Semantic Parsers

Instead of predicting the LF conditioned on the input utterance, large language models such as CODEX could learn to solve a specific task by *in-context learning*. During in-context learning, the parser predicts the LF conditioned on a prompt which consists of a small list of utterance-LF pairs to demonstrate the semantic parsing task and, optionally, a table schema. To defend against adversarial attacks, one seminal approach is adversarial

training. One of the most typical adversarial training methods augments the training data with adversarial examples, from which the machine learning model could learn robust features (Allen-Zhu and Li, 2022) by gradient descent. However, directly adapting conventional adversarial training is not suitable for in-context learning. First, the number of demonstration examples is limited due to the restriction on the maximum number of tokens for the pre-trained language model. As a result, we cannot include an arbitrary number of adversarial examples in the prompt, which might not include enough robust features. Second, in-context learning does not update the parameters of the language model. The model would not be optimized towards learning the robust features in the adversarial examples through gradient descent.

Given the difference, it is unclear whether in-context learning could improve the robustness of the parser as the conventional supervised training. In this paper, we conduct the first investigation on in-context learning for model robustness. More specifically, we examine the impact of variants of in-context learning and sampling methods on parser robustness.

4.1 Standard In-context Few-shot Learning

In our setting, given an input utterance x, the pre-trained language model $P(\cdot; \theta)$ predicts the LF y' conditioned on the prompt, which consists of a set of demonstration examples $\mathcal{M} = \{(x_i, y_i)\}_{i=1}^N$, and a table schema $\mathcal{T}$:

$$y' = \arg\max_{y \in \mathcal{Y}} P(y|x, \mathcal{M}, \mathcal{T}; \theta) \qquad (1)$$

For the *few-shot* setting, the number of demonstration examples N is limited by a budget size.

4.2 Adversarial In-context Few-shot Learning

In adversarial in-context learning, we include the perturbed adversarial examples, $\mathcal{M}_{adv}$, in the demonstration examples:

$$y' = \arg\max_{y \in \mathcal{Y}} P(y|x, \mathcal{M} \cup \mathcal{M}_{adv}, \mathcal{T}; \theta) \quad (2)$$

4.3 In-context Few-shot Selection

Current in-context learning assumes there is an example pool from where they can select prompting examples. However, most of the works only randomly pick examples from the pools. We argue that the way to select the examples might deeply impact

the robustness of the prompt-based semantic parser. Therefore, we examine various strategies to select in-context few-shot examples.

Random Sampling (Random). We randomly sample N utterances from the example pool.

Confidence-based Sampling (Confidence) (Duong et al., 2018). We score each utterance with the confidence of the parser on the predicted LF given the utterance and the table schema. Then we select the ones with the lowest parser confidence scores[2].

Diversity-based Sampling. Following Li et al. (2021), we partition the utterances in the utterance pool into N clusters with the K-means (Wu, 2012) clustering algorithm and select the example closest to the cluster centers. We measure the edit distance (Cluster-ED) (Wagner and Fischer, 1974), and Euclidean distances using utterance features of TF-IDF (Cluster-TF-IDF) (Anand and Jeffrey David, 2011), or Contextual Word Embedding (Cluster-CWE) encoded by Sentence-BERT (Reimers and Gurevych, 2019), between each pair of utterances for K-means.

Perplexity-based Sampling (Sen and Yilmaz, 2020). We score each utterance with the perplexity of GPT-2 on this utterance. Then we select the utterances with the highest (PPL. Asc) and lowest (PPL. Desc) perplexity scores, respectively.

5 Experiments

5.1 Setup

Evaluation Datasets. We evaluate the robustness of the prompt-based semantic parsers via the adversarial robustness sets built on top of the test sets of GeoQuery (Finegan-Dollak et al., 2018) and Scholar (Finegan-Dollak et al., 2018) with the proposed perturbation methods in Section 3. As in Finegan-Dollak et al. (2018), we choose the *query* splits of both GeoQuery and Scholar, where there is no LF template overlap among train, test, and dev sets.

Prompt-based Semantic Parser. We choose CODEX (Chen et al., 2021) as the representative prompt-based semantic parser for our evaluation.

In recent studies, CODEX has performed comparably via in-context few-shot semantic parsing to the SOTA-supervised trained neural semantic parsers (Shin and Van Durme, 2022; Roy et al., 2022b; Drozdov et al., 2022) in terms of execution accuracy.

To examine the vulnerability of large prompt-based semantic parsers against adversarial examples, we choose the `code-davinci-002` version of CODEX as it is the most powerful variant among all CODEX models, with 175B parameters. In our experiments, we sample a maximum of 200 tokens from CODEX with the temperature set to 0, with the stop token to halt generation.

Prompts. In this work, we adopt the prompt design of `Create Table + Select X` as presented in Rajkumar et al. (2022), which has been shown to be effective for semantic parsing using *static prompting*[3].

The prompt for semantic parsing on CODEX consists of `CREATE TABLE` commands, including specifications for each table's columns, foreign key declarations, and the results of executing a `SELECT * FROM T LIMIT X` query on the tables via the column headers. As described in Section 4.3, we select NL-LF pairs as in-context few-shot examples from the *train* sets.

To guide the prompt-based semantic parser, we also include the textual instruction of "`Using valid SQLite, answer the following questions for the tables provided above.`" as proposed by Rajkumar et al. (2022).

5.2 Research Questions and Discussions

Our experimental results answer the following four research questions (RQs) related to the robustness of CODEX.

RQ1: How vulnerable is the prompt-based semantic parser to adversarial examples?

Settings. To answer RQ1, we evaluate the standard accuracy and perturbation accuracy of CODEX on GeoQuery and Scholar test sets through *zero-shot* learning.

Results. The zero-shot parsing performances of CODEX are shown in Table 2. Our first observa-

---

[2]The confidence scoring parser is a zero-shot model, meaning that there are no examples present in the prompt. It operates solely based on the input utterance, instructions, and schema provided within the prompt. Please see Section 5.1 for more information on the prompt structure.

[3]In contrast to the approach of Shin et al. (2021) which involves dynamically selecting few-shot examples from an example pool, we refer to static prompting as being performed with a fixed set of examples.

Category	Pert. Strategy	GeoQuery			Scholar		
		Pert. Acc.	Std. Acc.	Δ	Pert. Acc.	Std. Acc.	Δ
Word-level	TB	53.85		-3.29	11.35		-0.86
	RD	50.55		-6.59	10.52		-1.69
	RS	37.36	57.14	-19.78	8.31	12.21	-3.90
	CS	42.31		-14.83	8.40		-3.81
	CI	38.46		-18.68	8.31		-3.90
Sentence-level	RB	**31.87**		-25.27	**5.22**		-6.99
	DB	35.71		-21.43	7.88		-4.33

Table 2: Results of perturbation accuracy (Pert. Acc.) and standard accuracy (Std. Acc.) of zero-shot performance on `GeoQuery` and `Scholar`. The zero-shot prompt only contain the table information and initial semantic parsing instruction. Perturbation accuracy is calculated based on each perturbation method.

tion is that CODEX is more vulnerable to sentence-level perturbations than to word-level perturbations, as indicated by the more significant performance gaps between standard and perturbed accuracies on the sentence-level perturbed test sets. Wang et al. (2021) observed that neural language models are vulnerable to human-crafted adversarial examples where there are complex linguistic phenomenons (e.g., coreference resolution, numerical reasoning, negation). We observe that the rewriting model trained on human paraphrase pairs also introduces such complex linguistic phenomenons.

With respect to the word-level perturbations, CODEX is most robust to typo-based perturbations, which is surprising as Wang et al. (2021) shows typo-based perturbation is the most effective attack method for large language models like BERT (Devlin et al., 2019) in the evaluation of natural language understanding tasks. However, utterances with typos drop only 3% of the accuracy of CODEX. Random Deletion is also less effective than the other word-level methods, consistent with the observations by Huang et al. (2021) on the fully-supervised semantic parsers. This phenomenon can be attributed to the fact that Random Deletion primarily makes minor modifications to the standard NL utterances, as this method often involves removing non-functional words such as articles, for example, "the" and "a."

Although CODEX is pre-trained on a considerably large dataset, it does not show robustness on the in-domain tasks. We conjecture that the reason is that zero-shot CODEX has not yet learned any in-domain knowledge on `GeoQuery` or `Scholar`. So in RQ2, we would address whether in-domain examples would improve the robustness of CODEX.

Takeaways. Zero-shot CODEX is vulnerable to adversarial examples, especially sentence-level perturbation of utterances, rather than to word-level

perturbations.

Pert. Strategy	5-shot	10-shot	20-shot	30-shot	40-shot	50-shot
TB	63.19	71.98	78.02	81.32	81.30	82.42
RD	59.34	64.29	71.43	71.98	70.33	75.27
RS	52.20	52.75	54.87	59.34	60.99	63.14
CS	54.40	56.04	60.44	63.19	65.93	67.03
CI	51.65	54.95	55.49	57.69	58.79	61.02
RB	44.51	47.80	49.45	50.55	54.23	57.27
DB	48.35	49.77	53.30	54.20	59.34	59.89
Avg. Pert. Acc.	53.38	56.80	57.50	59.49	64.42	66.58
Std. Acc.	66.48	74.37	79.67	81.87	83.52	84.62
Avg. Robust Acc.	75.67	77.28	78.74	80.44	82.07	83.22

Table 3: Few-shot performances on `GeoQuery`. We conduct {5, 10, 20, 30, 40, 50}-shot learning experiments. Average perturbation accuracy (Avg. Pert. Acc.) is the average score of execution accuracies on different perturbation sets. Average robust accuracy (Avg. Robust Acc.) is the average score of execution accuracies on the test sets perturbed by different perturbation methods.

Pert. Strategy	3-shot	5-shot	10-shot
TB	10.57	20.33	34.29
RD	12.09	25.27	31.43
RS	6.04	17.03	25.08
CS	9.34	18.13	26.03
CI	8.24	14.29	20.63
RB	3.30	8.43	18.10
DB	4.40	10.44	21.90
Avg. Pert. Acc.	7.71	16.27	25.35
Std. Acc.	14.29	23.08	40.32
Avg. Robust. Acc.	51.12	53.10	55.97

Table 4: Few-shot performances on `Scholar`. We conduct {3, 5, 10}-shot learning experiments.

RQ2: Does standard in-context few-shot learning improve the robustness?

Settings. We respectively select up to 50 and 10 examples from `GeoQuery` and `Scholar` train sets[4], with the random sampling strategy, to construct prompts for parsers. Then, for each few-shot

---

[4]Due to the larger size of table schema in `Scholar`, we could only include up to 10 examples in the prompt.

learning experiment, we measure standard accuracy, perturbation accuracy and robust accuracy on our various perturbed test sets.

Results. Tables 3 and 4 show the performance of standard in-context few-shot learning on the robustness evaluation sets perturbed by different methods. We observe that more standard examples in the prompt can evidently improve the robust accuracy of CODEX, which demonstrates the effectiveness of standard in-context few-shot learning in improving the robustness of semantic parsing. Although it performs slightly worse on the test sets perturbed by typo-based methods than the one perturbed by the random deletion in GeoQuery, we argue that this is due to the performance variance, which does not necessarily hurt the model robustness.

The performance gap between perturbation accuracy with standard accuracy is enlarged when the number of in-context shots increases. However, the robust accuracy grows slowly. This indicates that improving the generalization ability of the parser does not necessarily mean the improvement of the robustness. The trade-off between standard and robust accuracies is a long-standing problem in adversarial training. Raghunathan et al. (2019) shows that increasing the training sample size could eliminate such a trade-off. Our experiments demonstrate that in-context learning follows similar patterns as supervised adversarial training. It can be observed that both objectives can be improved with a limited number of examples when compared to the zero-shot parser. However, the extent of improvement varies.

Takeaways. With more standard in-context examples, few-shot CODEX can be guided to achieve better robustness and standard execution performance.

RQ3: Does adversarial in-context learning improve robustness?

Settings. In this work, we present the experimental results of CODEX on both GeoQuery and Scholar datasets, using 10 and 5 in-context examples, respectively. In order to assess the robustness of CODEX through adversarial in-context learning, we first augment the standard few-shot examples by incorporating examples whose utterances have been perturbed using various perturbation methods. Subsequently, for each set of augmented examples, we calculate the average robust accuracy of CODEX based on the average of the

Adv. L. Strategy	GeoQuery		Scholar	
	Avg. Robust Acc.	Std. Acc.	Avg. Robust Acc.	Std. Acc.
No Adv.	77.28	74.37	53.10	23.08
No Adv. (× 2)	78.74	79.67	55.97	40.32
TB	77.32	73.62	52.75	34.99
RD	77.40	73.64	53.11	33.65
RS	78.30	74.73	54.88	33.46
CS	78.05	74.47	53.12	34.86
CI	78.14	74.81	54.66	33.65
RB	78.47	75.51	56.58	35.85
DB	78.31	75.08	55.08	35.71

Table 5: The results of the average robust accuracy obtained through Adversarial In-context Learning (Adv. L. Strategy) with different types of perturbed few-shot examples. Additionally, we include results of applying the method with only standard examples (No Adv.), as well as with a doubled number of standard examples (No Adv. (× 2)).

parser robust accuracies on all robustness evaluation sets.

Results. The experimental results of the various perturbation strategies applied to the in-context few-shot examples are presented in Table 5. While the approach of supervised adversarial training has been widely regarded as an effective means of enhancing the robustness of machine learning models, the results indicate that on both GeoQuery and Scholar, the robust accuracies are only marginally improved through the application of adversarial in-context learning. Previous studies (Raghunathan et al., 2019; Huang et al., 2021) have pointed out that supervised adversarial training can sometimes result in a decrease in standard accuracy, even as it improves robust accuracy. However, the results of adversarial in-context learning diverge from this trend, with significant improvement in standard accuracy, from 23% to more than 33%, observed on Scholar, while robust accuracy only experiences marginal improvement. These observations indicate that adversarial in-context learning represents a distinct approach from supervised adversarial training in terms of enhancing the robustness of the model. Furthermore, the results suggest that simply incorporating adversarial examples into the prompt has a limited impact on the robustness of parsers, in contrast to supervised adversarial training.

Of all the perturbation strategies analyzed, the results indicate that CODEX achieves the best performance in terms of both standard and robust accuracy through the application of RB adversarial in-context learning, but experiences the worst performance through TB adversarial in-context learning. The hypothesis is that RB produces utterances

Dataset	Metric	Confidence	Cluster-CWE	Cluster-ED	Cluster-TF-IDF	PPL. Asc	PPL. Desc	Random
	Avg. Robust Acc.	78.77	78.02	82.40	**85.10**	70.36	50.82	73.22
GeoQuery	Avg. Pert. Acc.	69.80	68.93	74.41	**77.74**	62.11	45.71	66.58
	Std. Acc.	74.73	78.41	81.32	**85.64**	73.14	70.76	74.18
	Avg. Robust Acc.	55.31	60.24	60.68	**62.61**	53.97	47.18	55.97
Scholar	Avg. Pert. Acc.	28.55	29.81	30.28	**31.44**	22.25	7.47	25.35
	Std. Acc.	37.99	41.49	42.19	**42.97**	35.93	34.89	36.91

Table 6: The performance of standard few-shot in-context learning using various sampling methods on the `GeoQuery` and `Scholar` datasets. The average robust accuracy, average perturbation accuracy, and standard accuracy are computed for each sampling method to assess their efficiency in this learning scenario.

Dataset	LC. Metric	Confidence	Cluster-CWE	Cluster-ED	Cluster-TF-IDF	PPL. Asc	PPL. Desc	Random
	TTR ↑	7.68	7.24	8.47	**10.26**	5.94	3.22	6.44
GeoQuery	Yule's I ↑	68.55	64.37	69.59	**71.49**	62.94	43.41	58.14
	MTLD ↑	12.44	12.19	13.37	**15.58**	11.32	8.16	10.41
	TTR ↑	28.18	29.91	31.40	**33.11**	21.15	14.17	25.67
Scholar	Yule's I ↑	198.15	207.11	223.76	**262.36**	177.37	102.17	193.31
	MTLD ↑	15.68	15.63	16.34	**19.49**	11.94	13.12	14.36

Table 7: Results of the language complexity of standard NLs sampled by different sampling strategies, measured by three lexical diversity (LC.) metrics. For the ease of readability and comparison, we multiply both TTR scores and Yule's I scores by 100.

with more complex linguistic features, resulting in enhanced standard and robust accuracy during in-context learning. To test this hypothesis, the number of standard examples is doubled (No Adv.×2) to match the size of the examples augmented with the adversarial examples. The results show that the robust and standard accuracies of CODEX are higher than those obtained through adversarial in-context learning, likely due to the greater diversity of linguistic variations in the doubled standard examples.

Takeaways. The robustness of few-shot CODEX can be marginally improved by adversarial in-context learning without significant negative impacts on standard performances.

RQ4: What is the impact of sampling techniques on the robustness of parsers that utilize standard in-context few-shot learning?

Settings. In order to compare the influence of different sampling methods on the robustness of the model, we vary standard in-context examples on `GeoQuery` and `Scholar` with all 7 strategies aforementioned in Section 4.3. We choose the 50-shot setting for `GeoQuery` and 10-shot setting for `Scholar`.

Results and Takeaways. We present standard accuracies in Table 6 when varying the sampling methods for the few-shot example selection. We first observe that different sampling strategies impact the robust and standard performance of the

CODEX. Overall, the Cluster methods, which diversify the features of the selected utterances, perform better than the other sampling methods. On the other hand, PPL. Desc sampling method performs consistently poorly than the other sampling methods. In brief, we conclude that CODEX is sensitive to the few-shot example sampling strategies.

RQ4-Ablation: Why does CODEX react differently to various sampling strategies?

Settings. The findings of RQ 1 and RQ 3 indicate that linguistic complexity has an impact on the performance of CODEX. As a result, the results of RQ 4 may also be influenced by linguistic complexity. To further explore this correlation, three lexical diversity metrics, Type-Token Ratio (TTR) (Templin, 1957), Yule's I (Yule, 1944), and Measure of Textual Lexical Diversity (MTLD) (Mc-Carthy, 2005), are adopted to measure the lexical diversity of the selected NLs from `GeoQuery` and `Scholar`. The TTR is defined as the ratio of the number of unique tokens, also known as types, to the total number of tokens. The TTR is used as an indicator of lexical diversity, with a higher TTR indicating higher lexical diversity. Yule's Characteristic Constant (Yule's K) is a measure of text consistency that considers vocabulary repetition. Yule's K and its inverse, Yule's I, are considered more robust to variations in text length than the Type-Token Ratio (TTR). MTLD is computed as the average number of words in a text required to maintain a specified TTR value.

Results and Takeaways. Table 7 presents the lexical diversity of each set of NLs sampled by different approaches. The diversity scores obtained from the three metrics align with the performance of CODEX as presented in Table 6. For instance, the three metrics indicate that the examples selected using the Cluster-TF-IDF method achieve higher lexical diversity compared to those selected through the other methods. Additionally, the Cluster-TF-IDF method also produces the highest results in terms of robust and standard accuracy for CODEX. Thus, it can be inferred that an increase in the lexical diversity of the few-shot examples leads to an improvement in the robust and standard accuracy of CODEX.

6 Conclusion

This study examines the robustness of semantic parsers in the context of prompt-based few-shot learning. To achieve this objective, robustness evaluation sets were curated to evaluate the robustness of the prompt-based semantic parser, CODEX. The research aims to identify methods to improve the robustness of CODEX. The results of our comprehensive experiments demonstrate that even the prompt-based semantic parser based on a large pretrained language model is susceptible to adversarial attacks. Our findings also indicate that various forms of in-context learning can improve the robustness of the prompt-based semantic parser. It is believed that this research will serve as a catalyst for future studies on the robustness of prompt-based semantic parsing based on large language models.

Limitations

In this study, we examine the robustness of the prompt-based semantic parser, CODEX, by focusing on the impact of prompt design on its execution performance. However, there is a need for future research to investigate more alternative adversarial training strategies for prompt-based semantic parsers in order to advance this field. In addition, our focus is limited to text-to-SQL tasks, and we encourage further investigation into the robustness of semantic parsers across different datasets and LFs. Despite these limitations, we emphasize the importance of exploring more effective prompt design in order to enhance the robustness of prompt-based semantic parsers, including CODEX, which shows non-negotiable vulnerability.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2022. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896.

Rajaraman Anand and Ullman Jeffrey David. 2011. *Mining of massive datasets*. Cambridge University Press.

Ankur Bapna, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2017. Towards zero-shot frame semantic parsing for domain scaling. *arXiv preprint arXiv:1707.02363*.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible nlp attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004. IEEE.

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

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.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models. *arXiv preprint arXiv:2209.15003*.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip R Cohen, and Mark Johnson. 2018. Active learning for deep semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers), pages 43–48.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding nlp systems. 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 1634–1647.

Catherine Finegan-Dollak, Jonathan K Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-sql evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 351–360.

Tira Nur Fitria. 2021. Quillbot as an online tool: Students' alternative in paraphrasing and rewriting of english writing. *Englisia: Journal of Language, Education, and Humanities*, 9(1):183–196.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5747–5757.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Shuo Huang, Zhuang Li, Lizhen Qu, and Lei Pan. 2021. On robustness of neural semantic parsers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3333–3342.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885.

Aishwarya Kamath and Rajarshi Das. 2018. A survey on semantic parsing. In *Automated Knowledge Base Construction (AKBC)*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Zhuang Li, Lizhen Qu, and Gholamreza Haffari. 2021. Total recall: a customized continual learning method for neural semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3816–3831.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *IJCAI*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2021. Few-shot self-rationalization with natural language prompts. *arXiv preprint arXiv:2111.08284*.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Yifei Min, Lin Chen, and Amin Karbasi. 2021. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. In *Uncertainty in Artificial Intelligence*, pages 129–139. PMLR.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. 2019. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pages 4970–4979. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Xinyu Pi, Bing Wang, Yan Gao, Jiaqi Guo, Zhoujun Li, and Jian-Guang Lou. 2022. Towards robustness of text-to-sql models against natural and realistic adversarial table perturbation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2022.

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1–10.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. 2019. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*.

Nitarshan Rajkumar, Raymond Li, and Dzmitry Bahdanau. 2022. Evaluating the text-to-sql capabilities of large language models. *arXiv preprint arXiv:2204.00498*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3194–3213.

Subhro Roy, Sam Thomson, Tongfei Chen, Richard Shin, Adam Pauls, Jason Eisner, and Benjamin Van Durme. 2022a. Benchclamp: A benchmark for evaluating language models on semantic parsing.

Subhro Roy, Sam Thomson, Tongfei Chen, Richard Shin, Adam Pauls, Jason Eisner, and Benjamin Van Durme. 2022b. Benchclamp: A benchmark for evaluating language models on semantic parsing. *arXiv preprint arXiv:2206.10668*.

Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. 2022. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613.

Priyanka Sen and Emine Yilmaz. 2020. Uncertainty and traffic-aware active learning for semantic parsing. In *Proceedings of the First Workshop on Interactive and Executable Semantic Parsing*, pages 12–17, Online. Association for Computational Linguistics.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32.

Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. 2020. Universal adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5636–5643.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen Jr, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715.

Richard Shin and Benjamin Van Durme. 2022. Few-shot semantic parsing with language models trained on code. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5417–5425, Seattle, United States. Association for Computational Linguistics.

Fatemeh Shiri, Terry Yue Zhuo, Zhuang Li, Shirui Pan, Weiqing Wang, Reza Haffari, Yuan-Fang Li, and Van Nguyen. 2022. Paraphrasing techniques for maritime qa system. In *2022 25th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. *arXiv preprint arXiv:2205.11503*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Mildred C Templin. 1957. *Certain language skills in children: Their development and interrelationships*, volume 10. JSTOR.

Florian Tramer and Dan Boneh. 2019. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems*, 32.

Robert A Wagner and Michael J Fischer. 1974. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Junjie Wu. 2012. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Mingyang Yi, Lu Hou, Jiacheng Sun, Lifeng Shang, Xin Jiang, Qun Liu, and Zhiming Ma. 2021. Improved ood generalization via adversarial training and pretraing. In *International Conference on Machine Learning*, pages 11987–11997. PMLR.

GU Yule. 1944. The statistical study of literary vocabulary. cambridge, cambridge [eng.].

John M Zelle and Raymond J Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the national conference on artificial intelligence*, pages 1050–1055.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations*.

A Experiments

Datasets. The `GeoQuery` dataset contains 877 pairs of NL-LF pairs about U.S. geographical information. On the other hand, `Scholar` contains pairs of NL-SQL regarding information about academic publications. Finegan-Dollak et al. (2018) proposed a dataset split for evaluating the compositional generalization capability of semantic parsers on several datasets, including `GeoQuery` and `Scholar`. The proposed split, referred to as the query-split, presents a more challenging scenario for semantic parsing models. This paper utilizes the query-split, where the two test sets in our experiments include 182 NL-LF pairs from `GeoQuery` and 315 NL-LF pairs from `Scholar`, respectively, during the evaluation of the prompt-based semantic parser.

Hyperparameters. We sample at most 200 tokens from CODEX with temperature 0, with the following strings used as stop tokens to halt generation: "--", "\n\n", ";", "#".

Model Versioning. The version of the `code-davinci-002` model referred to in this paper is as of the midpoint of the year 2022.

B Adversarial Examples

Table 8 lists the examples generated by all perturbation strategies.

Linguistic Phenomenon	Samples (~~Strikethrough~~ = Original Text, **red** = Adversarial Perturbation)
Typo (Word-level)	**NL**: what can you ~~tell~~**te11** me about ~~the~~**th e** population of missouri
Random Deletion (Word-level)	**NL**: ~~what~~ can you tell me ~~about~~ the population of missouri
Random Swap (Word-level)	**NL**: what can you tell me ~~about~~**missouri** the population of ~~missouri~~**about**
Context-aware Substitution (Word-level)	**NL**: what ~~can~~**will** you tell me about ~~the~~**a** population of missouri
Context-aware Insertion (Word-level)	**NL**: what **what** can you tell me about the **exact** population of missouri
Rewriting (Sent.-level)	**NL**: ~~what can you tell me about the population of missouri~~ **What information can you provide on Missouri's population?**
Distraction (Sent.-level)	**NL**: what can you tell me about the population of missouri **who is who; what is what; when is when; which is which; where is where**

Table 8: **Examples from Robustness Evaluation Set**. The adversarial utterances in this Table are generated by applying various perturbation strategies to a single utterance "*what can you tell me about the population of missouri*" sampled from the `GeoQuery` dataset. All of the generated utterances can successfully alter Codex's output.