Probing GPT-3's Linguistic Knowledge on Semantic Tasks

Lining Zhang
New York University
lz2332@nyu.edu

Mengchen Wang
New York University
mw3723@nyu.edu

Liben Chen
New York University
lc4438@nyu.edu

Wenxin Zhang
New York University
wz2164@nyu.edu

Abstract

GPT-3 has attracted much attention from both academia and industry. However, it is still unclear what GPT-3 has understood or learned especially in linguistic knowledge. Some studies have shown linguistic phenomena including negation and tense are hard to be recognized by language models such as BERT. In this study, we conduct probing tasks focusing on semantic information. Specifically, we investigate GPT-3's linguistic knowledge on semantic tasks to identify tense, the number of subjects, and the number of objects for a given sentence. We also experiment with different prompt designs and temperatures of the decoding method. Our experiment results suggest that GPT-3 has acquired linguistic knowledge to identify certain semantic information in most cases, but still fails when there are some types of disturbance happening in the sentence. We also perform error analysis to summarize some common types of mistakes that GPT-3 has made when dealing with certain semantic information.

1 Introduction

GPT-3 (Brown et al., 2020) is a large neural language model (NLM) released in 2020, and it has realized state-of-the-art performance on various language tasks. Disregarding its achievement in recent years, however, few pieces of literature interpret well what would happen inside GPT-3, as well as the knowledge it has acquired or represented. This is also true for understanding linguistic phenomena, which represent all features and grammar that a linguist should study (Bhatt et al., 2011). Based on recent studies, like the task-agnostic methodology named CheckList (Ribeiro et al., 2020), it is revealed that NLP models have a high failure rate in testing linguistic phenomena, though they may perform well in many other language tasks. This paper contributes to the existing literature by making an effort on understanding GPT-3's knowledge of the linguistic phenomenon, especially with a focus on semantic information including tense and singularity or plurality of the number of subject and object of a given sentence.

The SentEval probing tasks (Conneau et al., 2018) introduce 10 probing tasks covering the aspects of surface information, syntactic information, and semantic information. In our study, we want to evaluate GPT-3's knowledge and understanding of linguistic phenomena, thus we focus on the aspect of semantic information. Specifically, we apply the semantic tasks (Tense, SubjNum, and ObjNum) to test GPT-3's linguistic knowledge to understand tense and singularity or plurality of the number of subject and object, which do not involve any replacement or inversion of source corpus. (See section 3.1 for details of these semantic tasks.)

For our experiment, we design zero-shot and few-shot prompts separately, which means different numbers of examples from the dataset appear in the prompt. We also set the binary choice question like "Is the number of the subject of the sentence singular or plural?" as the default prompt style, while designing another general prompt that allows GPT-3 to give its own answer to the general question. More details about prompt design can be found in Table 1. For the decoding method, we set temperature as 0, 0.5, 0.7, and 0.9 accordingly, where lower temperature means GPT-3 will take fewer risks when making the prediction. We test the semantic tasks from SentEval on GPT-3 with combinations of the above prompt and temperature, and calculate accuracy for each type of linguistic phenomena in the probing task.

Based on our experiment result, we find that GPT-3 has acquired some linguistic knowledge to understand semantic information like tense and singularity or plurality of subject and object, though it may be disturbed in some cases. Besides, we notice that designing the prompt with the general question might lead to model performance degradation. The model tends to provide irrelevant answers, since

Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 297 - 304

Table 1: Examples of Different Prompt Design.

Prompt Design	Example
zero-shot prompt with **default** style	Is the sentence "It senses your movement." present or past?
zero-shot prompt with **general** style	What is the tense of the sentence "It senses your movement."?
few-shot prompt with **default** style	Is the sentence "He messed with you." present or past? ⇒ past Is the sentence "It senses your movement." present or past? ⇒?

no expected choice is provided as in the default prompt style. We also find that variation in temperature has a minor impact on GPT-3's performance. Further, it is unexpected that more examples in the few-short prompts confuse and hurt the model in some tasks, rather than providing more hints.

Our work contributes to the stream of the work on probing the large language models, which helps us better understand what linguistic properties the model has acquired or represented. Specifically, we provide better insights on GPT-3's linguistic knowledge of certain semantic information.

2 Related Work

In recent years, although pre-trained language models like BERT (Devlin et al., 2019) have achieved state-of-the-art performance in many NLP tasks, it is still difficult to figure out what linguistic information is learned by the language representations.

Probing tasks are designed to test whether language models have encoded linguistic phenomena in learned representations by training a probing classifier on these representations. In an early study of machine translation, Shi et al. (2016) convert source sentences into encoded representations by the neural machine translation model, and train a logistic regression model on these representations to predict syntactic labels. In another study, Adi et al. (2017) design tasks to measure what extent the sentence representation from CBOW (Mikolov et al., 2013) and LSTM auto-encoder encodes its length, the identities of words within it, and word order. Their results indicate that the probing task is an effective way to evaluate the language model's ability to learn linguistic information.

Thus, many recent works have made some efforts to profile neural language models (NLMs) (Marvin and Linzen, 2018; Warstadt et al., 2019; Miaschi et al., 2020). For example, Marvin and Linzen (2018) test whether the language model assigns a higher probability to the grammatical sentence than the ungrammatical ones, showing that the performance of the language model lags behind the human performance in recognizing the grammaticality of the sentence. Warstadt et al. (2019) also assess the NLM's ability on learning grammatical knowledge and show that the BERT has significant knowledge of some grammatical features in sentences. Miaschi et al. (2020) test the model's ability to understand linguistic features, such as sentence length and part-of-speech tagging (POS tagging). It reveals that "the more NLM stores readable linguistic information of a sentence, the stronger its predictive power". Many other works also focus on understanding the attention mechanism of NLMs (Tang et al., 2018; Jain and Wallace, 2019; Clark et al., 2019). For example, Clark et al. (2019) conduct an analysis on BERT's attention and show that "certain attention heads correspond well to linguistic notions of syntax and coreference".

Previous work has provided evidence of NLM's ability to learn linguistic knowledge from the data. Some work tries to understand whether the learned linguistic knowledge has a particular structure (e.g. hierarchical structure) (Belinkov et al., 2018; Lin et al., 2019). These works have developed important probing tasks that profile the different aspects of the linguistic knowledge of NLMs. We follow the approach to conduct our own experiments of probing tasks on exploring GPT-3's ability to understand linguistic phenomena.

3 Experiment

We test the semantic tasks (Tense, SubjNum, and ObjNum) from SentEval (Conneau et al., 2018) on GPT-3 with combinations of different prompt designs and temperatures, and calculate accuracy for each type of linguistic phenomena in the probing task.

3.1 Dataset

We use the SentEval dataset with a focus on probing tasks of semantic information. Specifically, we apply the semantic tasks of Tense, SubjNum, and ObjNum to test GPT-3's linguistic knowledge.

The Tense task is a binary classification task that predicts whether the tense of the main verb of a sentence is present (PRES) or past (PAST). The SubjNum task is also a binary classification task that predicts whether the number of the subject of a sentence is singular (NN) or plural (NNS). The ObjNum task is almost the same as the SubjNum task, but it predicts the number of the object of a sentence instead.

The original SentEval dataset has over 100 thousand of records for each probing task. Given the computational efficiency, we randomly sample a subset of 500 records for each semantic task of Tense, SubjNum, and ObjNum to run our experiments. The datasets and codes are available at https://github.com/lining-zhang/GPT-3_Linguistic.

3.2 Experimental Design

Baseline Experiment For our baseline experiment, we use the prompt from the OpenAI API ("QA prompt")[1], with some modifications on the instruction part of the prompt. This makes our default prompt zero-shot, which means no examples from the SentEval probing dataset appear in the prompt. We also design the question in the default prompt to directly specify the labels that GPT-3 should choose from. For the decoding method, we set the temperature to 0, which means GPT-3 will take fewer risks when making the prediction. For the engine, we use "text-davinci-002"[2] for all experiments, which is the most capable GPT-3 model for all kinds of tasks.

Default Prompt vs General Prompt For the default prompt style, we set the binary choice question like "Is the number of the subject of the sentence singular or plural?" to appear in the prompt. This default style specifies the exact answers that GPT-3 is expected to choose from. To investigate the effect of the prompt design, we also design another general prompt that allows GPT-3 to directly give its own answer to the general question like "What is the number of the subject of the sentence?". This general prompt style gives GPT-3 more freedom to generate its own answer without restriction to certain choices, but still with the risk that it may not be able to find the expected answer. The experiments of the default prompt and general prompt are all zero-shot. Examples of different prompt designs can be found in Table 1.

Temperature Variations To test the influence of temperature on model performance, we measure GPT-3's linguistic knowledge on the semantic tasks with the temperature variation of 0, 0.5, 0.7, and 0.9 accordingly. Both default and general prompts are tested.

Few-shot Experiment To test whether the model benefits from more examples, we provide randomly selected examples of the linguistic phenomena to create the few-shot prompt. Based on the assumption that the number of examples in the few-shot prompt might also have an effect on the model's performance, we vary the number of examples provided, while keeping all few-shot prompts in the default style and setting the temperature to 0. We experiment with two examples and five examples in the few-shot prompt separately. See appendix A for more few-shot learning examples.

Evaluation To evaluate GPT-3's performance on each semantic task, we compare the response returned by GPT-3 with the true label. If GPT-3 predicts the true label correctly, we will assign a new label of response type with a value of 1. If GPT-3 predicts the true label adversely, we will assign the label of response type with a value of 2. If the response GPT-3 returned doesn't hit any of the true labels, or even doesn't make sense given the context, we will assign the label of response type with a value of 3. Then we calculate the ratio corresponding to each label of response type to show GPT-3's performance on each type of linguistic phenomena in the semantic probing task.

3.3 Results

Based on the answers returned from GPT-3, we categorize the responses into three response types which indicate their prediction as correct, adverse, or irrelevant. Detailed proportions for each case can be found in Table 2 and the corresponding visualization can be found in Figure 1. Considering the case that GPT-3 cannot detect linguistics phenomena at all and tends to give responses simply by random guess, then the ratio of each label of response type would all be approximately 0.33.

In terms of the default prompt style, which provides options like "Is the tense of the sentence past or present?", we find that GPT-3 has acquired some linguistic knowledge to understand semantic information like tense and singularity or plurality of sub-

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

[2] https://beta.openai.com/docs/models/gpt-3

Table 2: Experiment Results for Combinations of Different Prompt Design and Temperature.

Experimental Setting	Task Name	Correct Answer (Label 1)	Adverse Answer (Label 2)	Irrelevant Answer (Label 3)
Default Prompt Temperature=0	Tense	0.712	0.288	0
	SubjNum	0.74	0.254	0.006
	ObjNum	0.608	0.392	0
Default Prompt Temperature=0.5	Tense	0.718	0.28	0.002
	SubjNum	0.698	0.276	0.026
	ObjNum	0.596	0.404	0
Default Prompt Temperature=0.7	Tense	0.698	0.3	0.002
	SubjNum	0.684	0.3	0.016
	ObjNum	0.6	0.398	0.002
Default Prompt Temperature=0.9	Tense	0.698	0.294	0.008
	SubjNum	0.662	0.3	0.038
	ObjNum	0.584	0.408	0.008
General Prompt Temperature=0	Tense	0.668	0.308	0.024
	SubjNum	0.044	0.03	0.926
	ObjNum	0.26	0.19	0.55
General Prompt Temperature=0.5	Tense	0.67	0.306	0.024
	SubjNum	0.062	0.04	0.898
	ObjNum	0.212	0.174	0.614
General Prompt Temperature=0.7	Tense	0.678	0.29	0.032
	SubjNum	0.048	0.042	0.91
	ObjNum	0.208	0.144	0.648
General Prompt Temperature=0.9	Tense	0.662	0.288	0.05
	SubjNum	0.06	0.058	0.882
	ObjNum	0.208	0.134	0.658
Five-shot examples Temperature=0	Tense	0.67	0.33	0
	SubjNum	0.718	0.282	0
	ObjNum	0.674	0.326	0
Two-shot examples Temperature=0	Tense	0.7	0.3	0
	SubjNum	0.72	0.28	0
	ObjNum	0.702	0.298	0

ject and object. However, when general prompts are provided, we notice that it degrades GPT-3's performance slightly regarding the tense query, and heavily regarding the singularity or plurality query. This issue results from the fact that GPT-3 sometimes cannot distinguish "What is the number of subject/object of the sentences" from "What is the subject/object of the sentences". Thus, GPT-3 tends to choose irrelevant answers in this situation.

For the variation of temperature, we find that it has a minor impact on GPT-3's performance regardless of which type of prompt is given. For the Tense task, the highest ratio reached by the correct answer (Label 1) happens when the temperature is 0.5 with the default prompt style and 0.7 with the general prompt style. For SubjNum and ObjNum tasks in whatever prompt style, the ratio of Label 1 tends to decrease slightly or fluctuate as the temperature increases.

Besides, GPT-3 performs better in identifying the singularity or plurality of the subject than that of the object given the default prompt style. This circumstance may result from the fact that GPT-3 can infer the subject's singularity or plurality not only based on the subject itself, but also on the predicate of the sentence. On the other hand, the structure of the object of sentences can be more confusing than the subject's most of the time, introducing more challenges to GPT-3 in syntactic parsing.

We first conduct the experiment with five examples in a few-shot prompt for each semantic task. The experiment is also in default prompt style with the temperature = 0. We notice that GPT-3's performance degrades for the Tense and SubjNum tasks but increases for the ObjNum task. This phenomenon matches the observation that identifying the singularity or plurality of the object is more difficult compared to the subject given a zero-shot prompt in the default style. Thus, in the few-shot

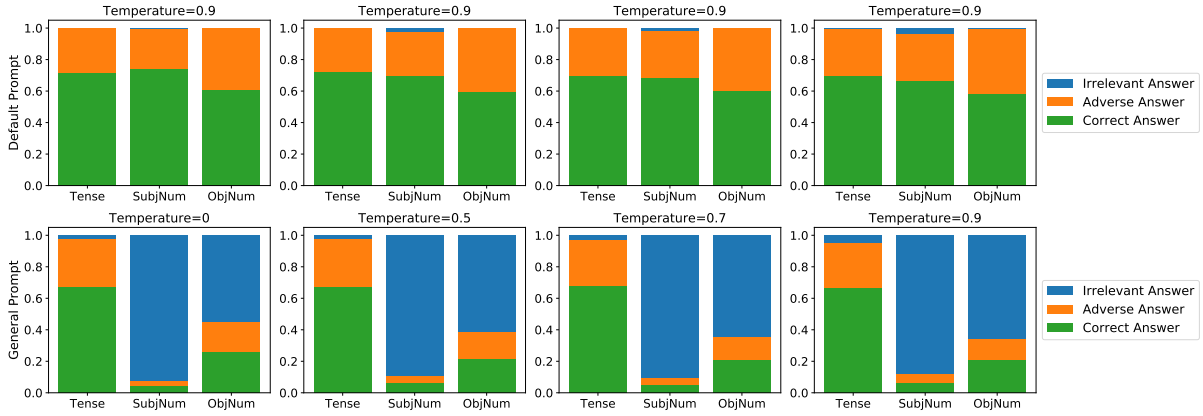

Figure 1: The Proportion of Different Answers Returned by GPT-3 for Each Combination of Prompt and Temperature

prompt, more examples are provided to support the ObjNum task which results in the increase. Besides, GPT-3 tends to be more confident in getting relevant answers, as the only percentage of the irrelevant answer which is not zero from the SubjNum task also goes to zero after several examples are provided in the few-shot prompt. After observing the degradation of model performance for the Tense and SubjNum tasks, we suspect that more examples obfuscate the model and reduce the number of examples to two in the few-shot prompt. We then observe an increase in model performance for all tasks, compared to results in the experiment with five examples. However, GPT-3's performance in the few-shot prompt still cannot rival the one from the baseline experiment for the Tense and SubjNum tasks, but improves for the ObjNum task.

4 Error Analysis

We perform the error analysis on records that GPT-3 does not answer correctly, which indicates the response is either adverse or irrelevant with Label 2 and Label 3 separately. We manually go through some records that get incorrect responses from GPT-3 to mark them with potential reasons and categorize them into several types of mistakes. However, this process requires a large amount of human annotations for each scenario to identify mistake types precisely, which may not be feasible in our case. Thus, this analysis may not exhaust all possibilities that GPT-3 might make mistakes when identifying certain linguistic phenomena and is not able to quantify the corresponding proportions, but it still provides some insights into how GPT-3 understands linguistic knowledge to some extent. Below is a brief summary of some common types of mistakes that GPT-3 has made when returning the response. Examples of each mistake type can be found in Table 3.

Disturbance of Quotation Mark For sentences that have partial content inside quotation marks as part of the dialogue, if the tense of the main verb is in past but the tense of the content inside quotation marks is in present, then GPT-3 will predict the tense as "present" incorrectly.

Disturbance of Concomitant Adverbial If the sentence has the present participle as the concomitant adverbial, but the tense of the main verb is in past, then GPT-3 will be disturbed by the adverbial and predict the tense as "present" incorrectly.

Identification of Negation If the sentence contains negation and the main verb followed by negation like "didn't" is in the present form, GPT-3 will ignore the context and return an incorrect "present" label for the whole sentence, focusing only on the form of the verb partially.

Disturbance of Clause If the sentence has a clause with a singular object, then GPT-3 will have difficulty identifying the object of the main sentence and its number.

Subject or Object Found, Not Its Number In some cases, GPT-3 finds the subject/object of the sentence, instead the number of the subject/object (singular or plural) as asked in the prompt.

5 Conclusion and Future Work

Based on our experiments and analysis, we find GPT-3 has acquired some linguistic knowledge to understand semantic information like tense and singularity or plurality of subject and object. Moreover, the variation in temperature does not have a big impact on GPT-3's performance, but design-

Table 3: Examples of Certain Error Types

Error Type	Task	Example	True Label	Predicted Label
Disturbance of Quotation Mark	Tense	"Beauty fades, but dumb is forever" Scarlet countered.	PAST	PRES
Disturbance of Concomitant Adverbial	Tense	Fake Mira commanded, pointing at Jace.	PAST	PRES
Identification of Negation	Tense	As if she truly didn't care whether or not someone loved her, as long as he at least pretended to.	PAST	PRES
Disturbance of Clause	ObjNum	Since the kiss that morning, Neal hadn't renewed his attentions.	NNS	NN
Subject or Object Found, Not Its Number	SubjNum	The rope around your waist will protect you if you fall.	NN	- (subject returned)

ing the prompt with the general question might lead the model to provide irrelevant answers. We also notice that the performance of identifying the number of the subject is commonly better than the performance in identifying objects, which explains why the ObjNum task benefits from the few-shot prompt. However, the few-shot learning experiment has a relatively degraded result for the Tense and SubjNum tasks, since more examples may obfuscate the model but the answer tends to be more relevant.

There are still some further works we could do based on the previous analysis. First, besides the baseline prompt and general prompt, there are still more combinations of different prompt designs and temperatures that we could test, suggesting that there might be more explorations when we analyze GPT-3's linguistic knowledge. Besides, our study mainly focuses on the semantic information of linguistic phenomena, which is restricted to a limited amount of probing tasks to test the model. A more exhaustive list of probing tasks or a carefully designed benchmark based on the error analysis could be created to better test the language model's linguistic knowledge in the future. Moreover, further human annotations could be applied in identifying mistake types for each scenario, which provides the quantitative measurement for each phenomenon where GPT-3 makes a mistake.

6 Acknowledgments

This work is finished as the course project for DS_GA 1012 Natural Language Understanding and Computational Semantics in Spring 2022 semester at New York University. We would like to thank our professor Samuel R. Bowman and our teaching assistant Richard Yuanzhe Pang for the suggestions for our paper, as well as Ruiqi Zhong for his initial proposal of this idea for this course.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *ICLR 2017*.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2018. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *arXiv preprint arXiv:1801.07772*.

Rajesh Bhatt, Owen Rambow, and Fei Xia. 2011. Linguistic phenomena, analyses, and representations: Understanding conversion between treebanks. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1234–1242.

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

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What

you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: getting inside bert's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. *arXiv preprint arXiv:1810.07595*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating bert's knowledge of language: Five analysis methods with npis. *arXiv preprint arXiv:1909.02597*.

A Appendix

We use the below examples to create the few-shot prompt. The binary value for the column of "Include in two-shot" indicates whether this example will be included in the few-shot prompt with two examples. By default, all the examples are included in the few-shot prompt with five examples.

Task Name	Include in two-shot	Example
Tense	1	Q: Is the tense of the sentence "He grunted And climbed to his feet, still holding me" present or past? A: Past
	1	Q: Is the tense of the sentence "It senses your movement" present or past? A: Present
	0	Q: Is the tense of the sentence "With a beer in his door hand and the window open to yell endlessly at everyone, he steered and shifted with the other hand" present or past? A:Past
	0	Q: Is the tense of the sentence, "His nostrils flare in reaction" present or past? A: Present
	0	Q: Is the tense of the sentence "Jack rolled and took me with him, capturing me on top of him, my head fitting perfectly into the hollow of his shoulder" present or past? A: Past
SubjNum	1	Q: Is the number of the subject of the sentence "Romulus was unreadable As ever" singular or plural? A: Singular
	1	Q: Is the number of the subject of the sentence "The wolves circled restlessly, their glowing yellow eyes fixed on the driver's door" singular or plural? A: Plural
	0	Q: Is the number of the subject of the sentence "There were several drips of whatever it was" singular or plural? A: Plural
	0	Q: Is the number of the subject of the sentence "An ape like Amy was not a cheap and stupid version of a human worker" singular or plural? A: Singular
	0	Q: Is the number of the subject of the sentence "Things were going even better than he had planned and it was all because of Misty" singular or plural? A: Plural
ObjNum	1	Q: Is the number of the object of the sentence "Practically purring with contentment, she rubbed her slightly bulging belly" singular or plural? A: Singular
	1	Q: Is the number of the object of the sentence "He flexed his biceps, and I groaned" singular or plural? A: Plural
	0	Q: Is the number of the object of the sentence "I served beers on autopilot" singular or plural? A: Plural
	0	Q: Is the number of the object of the sentence "The big man made a vague gesture" singular or plural? A: Singular
	0	Q: Is the number of the object of the sentence "The old woman could see my indecision" singular or plural? A: Singular

Table 4: Examples for Few-shot Prompt.