# Profiling GPT-3's Linguistic Knowledge

**Lining Zhang**
NYU CDS
lz2332@nyu.edu

**Mengchen Wang**
NYU CDS
mw3723@nyu.edu

**Liben Chen**
NYU CDS
lc4438@nyu.edu

**Wenxin Zhang**
NYU CDS
wz2164@nyu.edu

## Abstract

GPT-3 has attracted much attention from both academia and industry. However, we lack understanding on the GPT-3's ability to learn linguistic knowledge such as grammar. In this study, we conduct a set of evaluation tasks that probe the GPT-3's ability on various linguistic phenomena. Specifically, we investigates GPT-3's ability on learning semantic information, and we experiment with different prompts and decoding methods. We contribute to the existing literature by providing more insights on different aspects of the GPT-3's ability to understand linguistic phenomenon. Our experiment results suggest that the model has adequate knowledge on recognising the tense and the number of the subject or the object. We also perform error analysis to investigate the weakness of the model.

## 1 Introduction

GPT-3 (Brown et al., 2020) is a large Neural language model (NLM) released in 2020. It shows outstanding ability on various language tasks and has been commercialised in many ways. Disregarding its commercial and academic importance, however, few pieces of literature interpret well what would happen inside GPT-3. For example, the task agnostic methodology named CheckList for testing NLP models has high failure rate in testing linguistic phenomenon (Ribeiro et al., 2020), though it performs well in many other language tasks.

Linguistic phenomenon represents all features and grammars that a linguist should study (Bhatt et al., 2011). This paper contributes to existing literature by making an effort on understanding GPT-3's knowledge on linguistic phenomenon, especially including tense and singular and plural of subject and object of sentences.

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, which does not involve any replacement or inversion of source corpus.

To test GPT3's ability to understand tense, singular and plural of subjectives and objectives, we design zero-shot and few-shot prompts separately. In terms of zero-shot prompts, We design two kinds of prompts including multiple choice prompt that asks GPT3 to answer the question like "Is the number of the subject of the sentence singular or plural?", and general prompt that allows GPT3 to give its answer to the question like "What is the number of the subject of the sentence?" directly.

We choose temperature sampling as GPT3's decoding method, where lower temperature introduces more confidence in GPT3's top choices; 0 temperature equals to argmax likelihood; temperature greater than 1 decreases confidence. Specifically, in our experiment we set temperature as 0.0, 0.5, 0.7, 0.9 accordingly, and find that variation of temperature doesn't have big impact on GPT3's performance.

From our experiment result, we find that the model has adequate knowledge on the tense, the number of the subject and object of the sentence. We also notice that writing the prompt with more general wording might lead to model performance degradation. The model tends to provide irrelevant answers, as no specific instruction is provided in the prompt. Further, we notice that changing parameter such as temperature or providing more examples in the prompt only has a marginal effect on the model performance.

Our work contributes to the stream of the work on probing the large language model. Our result

| | |
|---|---|
| Zero shot + Default QA prompt | Is the sentence 'It senses your movement.' present or past? |
| Zero shot + General QA prompt | What is the tense of the sentence 'It senses your movement.'? |
| Few Shots + Default QA prompt | Is the sentence 'He messed with you' present or past? ⇒ past<br>Is the sentence 'It senses your movement' present or past? ⇒? |

Table 1: Different Prompts

contributes to the understanding of the large language model's ability on linguistic phenomenon. Specifically, we provide better insights on its knowledge on semantic-related grammar.

## 2 Related Work

### 2.1 GPT-3

NLMs have attracted increasing scholar attentions these years. Many of them shows exceptional performance on many tasks (Brown et al., 2020; Devlin et al., 2018). However, the linguistic knowledge learned by these NLMs remain less understood and previous works made some efforts on profiling them (Marvin and Linzen, 2018; Warstadt et al., 2019; Miaschi et al., 2020). Marvin and Linzen (2018) tested whether the language model assigns a higher probability to the grammatical sentence than the ungrammatical ones. However, the result from Marvin and Linzen (2018) suggests that the performance of the language model lags behind the human performance on recognising the grammaticality of the sentence. Warstadt et al. (2019) also assessed the NLM's ability on learning grammatical knowledge. The result from Warstadt et al. (2019) suggested that the BERT has significant knowledge of some grammatical features in sentences. In the recent work, Miaschi et al. (2020) composed a lists of linguistic features including sentence length and POS etc. to test the model's ability on understanding them. Their result suggested that "the more NLM stores readable linguistic information of a sentence, the stronger its predictive power". Many other work focused 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) conducted analysis on BERT's attention and showed that "certain attention heads correspond well to linguistic notions of syntax and coreference".

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

### 2.2 Probing Tasks

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 and LSTM auto-encoder encodes its length, the identities of words within it, and word order. Based on their results, it indicates that the probing task is an effective way to evaluate the language model's ability of learning linguistic information.

Further studies of probing tasks mainly focus on specific aspects of probing, like intermediate layers (Alain and Bengio, 2018), contextual representations (Tenney et al., 2019) or function words (Kim et al., 2019). For our study, we focus on the evaluation of GPT3's linguistic knowledge based on probing tasks.

## 3 Experiment

### 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 which predicts whether the tense of the main verb of a sentence is in present (PRES) or past (PAST). The SubjNum task is also a binary classification task which 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 thousands 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, codes, and experiment results are available at https://github.com/lining-zhang/GPT-3_Linguistic.

### 3.2 Experimental Design

**Baseline Experiment and its Variations** For our baseline experiment, we use the prompt from the OpenAI API ("QA prompt"), with some modification on the instruction part of the prompt. This makes our default prompt as 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 use temperature sampling with temperature set to 0, which means GPT-3 will take fewer risks when making the prediction. For the engine, we use "text-davinci-002" for all of our experiments, which is the most capable GPT-3 model for all kinds of tasks.

We then test how variation of temperature and prompts would affect performance of GPT-3. In terms of the zero-shot learning, we measure GPT3's ability to understand different types of linguistic phenomena with the temperature variation of 0, 0.5, 0.7, and 0.9; and set the prompt as default prompt and general prompt separately to test different combinations.

**Few-shot Experiments** To test whether the model can perform better with few-shot examples, we test the model with our few-shot examples. We provides randomly selected examples of linguistic phenomenon in the prompt. We assume that the number of examples might have an effect on the model's performance. Thus, we vary the number of examples provided. In the first few-shot experiment, we provide 5 examples for each tested linguistic phenomenon and in the second experiment we provide only two examples. The specific examples provided are listed in the appendix in details.

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

### 3.3 Results

**Baseline Experiments and its Variation Result** Responses of GPT-3 get their label out of three possibilities, indicating its prediction as correct, adverse, or irrelevant. Given default prompt, consider the case that GPT-3 cannot detect linguistics phenomena at all and tends to give responses simply by guessing all the time, then the ratio of label 1, label 2, label 3 would all be approximate 0.33. From the experiment result, we certify that providing options like "Is the tense of the sentence past or present?" or "Is the number of the object/subject of the sentence singular or plural", GPT-3 do have the ability to understand tense, singular and plural of subject and object. Besides, GPT-3 performs better in identifying singular and plural of subject than that of object. This circumstance may result from the fact that GPT-3 can infer subject' singular or plural not only based on the subject themselves, but also on predicate of the sentence. On the other hand, structure of object of sentences can be more confusing than subject most of the time, introducing more challenges to GPT-3 in syntactic parsing.

Next, in terms of providing GPT-3 with general prompts, we conclude that it would degrade GPT-3's performance slightly regarding the tense query, heavily regarding the singular and plural query. This issue results from the fact that GPT-3 sometimes cannot distinguish "What is the number of object/subject of the sentences" from "What is the object/subject of the sentences".

Overall, We find that whatever kind of prompts

are given, the ratio of label 1 tends to decrease slightly as the temperature increases.

**Few-shot Experiment Result** We first conducted five-shot experiments on our three tasks. We notice that the model performance degrades in most cases. However, the model tends to be more confident in getting relevant answers, as the percentage of the irrelevant answer goes to 0 after we provide several examples to the model. After observing the degradation on the model performance, we suspect that the lengthy examples obfuscates the meaning of the prompts. Thus, we reduce the number of examples to two and conduct two-shot examples. We observed a significant increase on model performance in most cases, compared to the one in five-shot experiments. However, the model performance with few-shot examples still cannot rival the one from the baseline model, thought the difference is marginal.

## 4 Error Analysis

We perform the error analysis on records that GPT-3 does not predict correctly, which have the label of respond type with value of 2 or 3. Below is a brief summary of some common types of mistakes that GPT-3 has made.

**Disturb 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.

**Disturb 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 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.

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

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

## 5 Conclusion and Future Work

Based on our experiment and analysis, we find the limitation and advantages by using different temperatures and prompts in testing GPT-3's linguistic knowledge. GPT-3 undoubtedly has the ability to identify tense, single or plural subject and object, but the accuracy would be affect by temperatures and different prompts, and the performance of identifying subject is commonly better than the performance in identifying object. Few-shot learning experiment has a relatively degraded result in most cases but the answer tends to be more relevant.

However, there are still some further works we could do based on the previous analysis. First of all, besides the baseline prompt and general prompt, there are still more combinations of prompt and temperature that we could test, suggesting that there might be more explorations when we use GPT-3 to analyze linguistic knowledge. Secondly, in order to find the best temperature and predict the accuracy, we could apply confidence interval in several different samples in order to analyze the most statistically significant result.

## 6 Ethical Considerations Statement

Our research provides specific analysis of GPT-3 in linguistic knowledge's accuracy, but it might also lead to some unpredictable social result. The analysis and experiment result can be applied in order to increase the accuracy and efficiency of data analyzing, but we have to point out the uncertainties such as the leak of private message and the security of personal information. Since GPT-3 could assist companies with works such as labeling and analyzing, with appropriate accuracy, it could increase the efficiency of a curing group in several aspects, but it might also lead to a consideration of the unemployment and leak of privacy.

## 7 Collaboration statement

Wenxin Zhang: Introduction to Explainable AI; Prompt design; Decoder design; Baseline Experiments and its Variation Result Analysis
Liben Chen: Related work on GPT-3; Abstract
Lining Zhang: SentEval Dataset; Experiment Setup; Prompt Design; Error Analysis; Related work on probing tasks
Mengchen Wang: Motivation; Abstract; Introduction to Linguistic Phenomenon; Introduction to GPT-3

## 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*.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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*.

Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R. Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different nlp tasks teach machines about function word comprehension.

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*.

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*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *ICLR 2019*.

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*.

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

Appendix 1: Experiment Results

| Sub_Num | Obj_Num | Tense | Provided in two-shot experiment | Provided in five-shot experiment |
|---|---|---|---|---|
| ”Q: Is the number of the subject of the sentence ””Romulus was unreadable As ever”” singular or plural? A: Singular ” | ”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 ” | Q: Is the tense of the sentence ”He grunted And climbed to his feet , still holding me” present or past? A: Past | 1 | 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 ” | Q: Is the number of the object of the sentence ”He flexed his biceps ,And I groaned” singular or plural? A: Plural | Q: Is the tense of the sentence ”It senses your movement” present or past?A: Present | 1 | 1 |
| ”Q: Is the number of the subject of the sentence ””There were several drips of whatever it was”” singular or plural? A: Plural ” | Q: Is the number of the object of the sentence ”I served beers on Autopilot” singular or plural? A: Plural | 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 | 1 |
| ”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 ” | Q: Is the number of the object of the sentence ”The big man made A vague gesture” singular or plural? A: Singular | Q: Is the tense of the sentence ”His nostrils flare in reaction” present or past?A: Present | 0 | 1 |
| ”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” | Q: Is the number of the object of the sentence ”The old woman could see my indecision” singular or plural? A: Singular | 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 | 0 | 1 |

Appendix 2: Few-shot Examples in Few-shot Experiments