Prompting *ChatGPT* to Draw Morphological Connections for New Word Comprehension

Bianca-Madalina Zgreaban and **Rishabh Suresh**
Institute for Language Sciences, Utrecht University, Utrecht, Netherlands
madalinazgreaban0@gmail.com and r.suresh@students.uu.nl

Abstract

Though more powerful, Large Language Models need to be periodically retrained for updated information, consuming resources and energy. In this respect, prompt engineering can prove a possible solution to re-training. To explore this line of research, this paper uses a case study, namely, finding the best prompting strategy for asking ChatGPT to define new words based on morphological connections. To determine the best prompting strategy, each definition provided by the prompt was ranked in terms of plausibility and humanlikeness criteria. The findings of this paper show that adding contextual information, operationalised as the keywords 'new' and 'morpheme', significantly improve the performance of the model for any prompt. While no single prompt significantly outperformed all others, there were differences between performances on the two criteria for most prompts. ChatGPT also provided the most correct definitions with a *persona*-type prompt.

1 Introduction

In recent years, *Large Language Models* (henceforth LLMs) have become increasingly more powerful (Bender et al., 2021). However, in terms of training, LLMs have large power consumption (Bender et al., 2021). This results in large expenditures of energy and resources if the models are periodically re-trained to have updated information, such as an updated vocabulary. A potential alternative to re-training is prompt engineering, namely designing better input prompts to get heightened performance (Ekin, 2023; Zhao et al., 2021). Through prompt engineering, considered even a programming method by Zhao et al. (2021), a model applies previous acquired knowledge to new tasks, thereby increasing its performance.

To research the advantages of prompt engineering in terms of updating vocabulary, the current study looks for the best prompting strategy in a case study, namely employing morphological connections for new word comprehension. We define morphological connections as extrapolating the definition of unfamiliar words by use of morphological knowledge, i.e. morphemes.

Thus, our research question is "What prompting strategy best allows an LLM to make correct morphological connections in a context of zero-shot learning?".

This study is not relevant only for novel word comprehension, but also for the task of morphological decomposition. For example, Kim and Smolensky (2021) showed evidence that BERT can abstract over grammatical categorisations, while McCoy et al. (2023) showed some models, such as GPT-2, use derivational morphemes in the generation of novel words, being able to generalise and use morphemes in a compositional manner. However, though LLMs might hold knowledge on abstract morphological categories, they might not always exploit this knowledge (Mahowald et al., 2023), given such models are not task-specific like decompositional models (for a systematic review on unsupervised learning in morphology, see Hammarström and Borin, 2011).

Thus, the current article specifically concerns prompting agnostic-task models to exploit morphological knowledge, i.e. individual morphemes and their functional roles, for morphological connection.

The chosen model was ChatGPT (OpenAI, 2021), a GPT-3.5 fine-tuned model. We compared statistically how the model performed given different types of prompts, and varying amounts of contextual information. Here, 'context' was operationalised in terms of keywords (i.e. 'new' or 'morpheme') or information about the task the model needed to perform. Given the model's training to avoid hallucinations, we expected it to classify new words as non-existent in the absence of context. We also used English as our test case, primarily due

Proceedings of the Student Research Workshop associated with RANLP-2023, pages 98–107, Varna, Bulgaria, Sep 4–6, 2023.
https://doi.org/10.26615/issn.2603-2821.2023_011

to being one of the languages for which ChatGPT performs the best in various tasks (Lai et al., 2023).

We show that depending on the criteria considered (i.e. plausibility or humanlikeness) different prompting strategies were better. Additionally, in line with our expectations, regardless of the criteria used, adding more contextual information increases the model's chance of offering a correct word definition.

Our results prove useful specifically for the users of ChatGPT, as despite the spread in its use, limited studies have been conducted on ChatGPT's neologism comprehension (Lenci, 2023). Even more, our results also prove important for research on novelty and LLMs, especially given this is overall a neglected research topic according to (McCoy et al., 2023). Moreover, this study presents a systematic methodology to test an LLM's ability to identify morphological connections or decompositions, which can be applied in the future to investigate other languages and language models as well.

The article is organized as follows: Section 2 outlines out methods, including the materials, procedure, scoring, and analysis, Section 3 describes our results, and Sections 4 and 5 include our discussion and conclusions.

2 Methods

2.1 Prompts

We adopted an approach advocated by White et al. (2023) in defining six basic prompts to be tested. We also manipulated the amount of contextual information in each prompt by designing four conditions: -morpheme, -new, as in (1); -morpheme, +new, as in (2); +morpheme, -new, as in (3); and +morpheme, +new, as in (4).

(1) Define...
(2) Define the *new* word ...
(3) Define the word ... considering its *morphemes*.
(4) Define the *new* word ... considering its *morphemes*.

Table 1: 'Define' prompt pattern in all four conditions

White et al.'s (2023) paper proved useful not only in offering examples of prompts, but also in presenting them systematically through what they call *prompt patterns*. They argue that when defining a prompt, it is important to specify its intent, motivation, structure, key ideas and consequences, along with an example of it. Such characteristics might help users design better prompts in view of their expectations, what they predict to be necessary information for the task, and possible consequences. An illustration of the prompt pattern for the prompt 'Define' can be seen in the following sentences:

> **Name**: Define
>
> **Intent**: Make morphological connections.
>
> **Motivation**: In case of use of novel word outside training data, the user needs to prompt the model for the new word.
>
> **Structure and key ideas**:
>
> Define **[word]**.
>
> Define **[the new word]**.
>
> Define **[the word]** considering its morphemes.
>
> Define **[the new word]**, considering its morphemes.
>
> **An example**: Define signatorily.
>
> **Consequences**: The model can generate hallucinations as it accepts the 'new word' as truly existent.

We adopted four prompts directly from White et al. (2023) with two variants of one pattern (i.e. the 'Persona' pattern), and defined one of our own (i.e. the 'Define' pattern presented above).

In another paper (Ekin, 2023), efficient prompt engineering entails, among other things, clear instructions, explicit constraints, varying contexts or examples, but also considering the type of task the system has to achieve (i.e. based on analysis or recall). The same paper also recommends other practices such as testing iteratively more prompts and designing prompts for specific domains. The chosen prompts from White et al.'s (2023) paper achieve these recommendations to various degrees. For example, the prompts' instructions vary in detail ('define words' vs. 'define new words'), have different contexts (e.g.. with or without 'new'), and various constraints (i.e. no constraints vs. following a pattern). Though the prompts do not provide any examples, given our zero-shot learning task, some of them can be regarded as domain-specific. For example, the *persona*-type prompt (see Appendix) by default restrains the domain of the output.

2.2 Test Items

For the words to test the model on, we decided to create a set of non-existent, but plausible English words, comprised of an existing English word, with a derivational affix. In making this choice, we followed McCoy et al. (2023) to ensure that we were testing the capacity of morphological connection, and not just recall of the model. In a previous study on the role of derivational morphology in lexical acquisition by Scandinavian children (Bertram et al., 2000), the authors found that the productivity of a given affix was strongly predictive of how well the participants performed on new word comprehension. They argued that more productive affixes were acquired earlier because they are used more regularly, and the form-meaning mapping may thus be easier to learn for children. In the case of LLMs, in may be argued analogously that the less productive a derivational affix is, the less is recombined with different lexical roots. As a consequence, the chances the model considers that affix as an independent meaningful unit are lower.

Thus, the novel words were created with either productive or unproductive suffixes to see if productivity had any systematic influence on the model's performance. The specific suffixes were selected from a paper by Ford et al. (2010) who tested word recognition by human participants based on morpheme frequency and productivity. They provide a list of derivational suffixes which are classified as either productive or unproductive based on three criteria: "hapax legommena of an affix, the type and token frequency of an affix and dictionary citation dates for neologisms" (Ford et al., 2010, p.120). Five productive suffixes: *-ly*, *-less*, *-ish*, *-able*, and *-ify*, and five unproductive suffixes: *-ise*, *-ous*, *-some*, *-ary*, and *-en* were selected from that paper to create the non-words.

In total, 10 new words were created with an equal number of productive and unproductive suffixes: *signatorily*, *assemblyless*, *benchish*, *delveable*, *lunchify*, *palatialise*, *violinous*, *musksome*, *containary*, *shallowen*. The complete list included nouns, verbs, adjectives, and adverbs, though adjectives were the most frequent. It was important to ensure that all test words were absent from the model's training data, so for every word, a browser search using *Microsoft Bing* was carried out to make sure that there were no matching hits. Additionally, we also verified that the words differed from a familiar word by a maximum of one derivational morpheme, in order to limit the number of morphological connections that the model would have to draw.

2.3 Procedure

The testing was conducted on three separate occasions within a two-week period on two separate devices, due to the number of prompts per hour per user. For each of the 24 prompts, a single attempt to define each word was given for the model. Additionally, a new chat was used for each word, given that Ortega-Martín et al. (2023) showed that repeated use of the same chat leads to lack of inference from the model due to its cache memory. In this way, we ensured wrong answers are not generated due to lack of inference, while we also ruled out the possibility that ChatGPT could improve its performance as it "learns" what is expected of it. The total number of trials was 240, and for each trial the output was saved for coding and analysis.

2.4 Scoring and Coding

Every output text produced by the model was scored on two criteria, *plausibility* and *humanlikeness*, being coded a 1 if the given definition was a 'success' (i.e. was plausible or humanlike) or 0 otherwise. If the model failed to provide any definition at all, this was coded as 0.

The need for two criteria was based on the understanding that expectations of the model may differ based on the downstream task. For example, if morphological connection is used in sentiment analysis of perfume reviews, we would prefer the model to make more humanlike inferences about one of our fictional words, i.e. *musksome*, has more on earthy, strong smell.

A definition was deemed plausible if it specified the correct word class of the derived word, and if it was related to the definition of the root word.

To score the output on humanlikeness, we conducted a survey with a sample of 11 native English-speakers, asking participants to provide hypothetical definitions for our words. In this survey, participants were asked to provide an intuitive definition of words taking advantage of the fact that the non-words were related to existing words. It was assumed that participants would not have knowledge of the linguistic definition of a *morpheme*. The participants were also requested to provide a words class for the words, along with hypothetical example sentences. The 11 definitions for each word were then analysed to identify a set of common characteristics based on which to define *humanlikeness*. An example of such characteristics is provided below for the word *violinous*.

> **violinous**: Adjective, violin-like quality in sound or appearance, of music/of an object/of a composition.

In general, definitions provided by human participants were assumed to be plausible themselves as long as they involved correct identification of the component morphemes. Human definitions that were not based on the morphological structure of a word were not used in formulating the criteria for humanlikeness. For example, one definition of the word *shallowen* provided by a participant was a "shallow Halloween", which clearly did relate to the meaning of the morpheme *-en*. In sum, a definition was considered humanlike if it had the common characteristics of the definitions in our survey.

2.5 Analysis

Two generalised linear mixed-effect regression models were created for humanlikeness and plausibility using R (Team, 2021), in RStudio (RStudio Team, 2020). The models contained a three-way interaction between the presence in the prompt of the word *new*, that of the word *morpheme* (both coded as binary variables), and the productivity of the suffixes. By-type (prompt type) random intercepts were also included in the models. The type of the prompt was left out of the model given it did not improve its fit. The effects *new*, *morpheme*, and *productive* had orthogonal contrasts set.

3 Results

The descriptive statistics for the number of plausible and humanlike definitions provided by the model for any given prompt are shown in table 2 below. In principle, a prompt could score anywhere between 0 and 10 on each criterion.

Plausibility			
Mean	***SD***	***Min.***	***Max.***
6.58	2.89	1	10
Humanlikeness			
Mean	***SD***	***Min.***	***Max.***
5.38	2.70	0	9

Table 2: Descriptive statistics for plausibility and humanlikeness scores for the model across prompts.

3.1 Plausibility

In terms of plausibility, two tested prompts offered plausible definitions for all words, such as the 'Persona' prompt word generator, conditions +new, +morpheme and -new, +morpheme, and the 'Context manager' prompt, condition +new, +morpheme.

From the model we built, we found significant main effects of adding the word *new* (β = 1.5645, z-score = 4.47, p-value < 0.05), and the word *morpheme* (β = 1.3083, z-score = 3.76, p-value < 0.05), as well as a significant main effect of suffix productivity (β = 1.1030, z-score = 3.18, p-value < 0.05). Thus, ChatGPT was over four and a half times more likely to provide a plausible definition if the prompt included the word *new*, over three and a half times more likely if the prompt included the word *morpheme*, and performed approximately three times better for words with a productive suffix than an unproductive one.

3.2 Humanlikeness

Considering humanlikeness, no prompt achieved correct definitions for all words. However, the 'Persona' prompt word generator, in condition -new, -morpheme and condition +new, +morpheme had 9 out of 10 correct definitions. Note that that the 'Persona' prompt word generator, condition +new, +morpheme, scores high for both measurements.

We found significant main effects of adding the word *new* (β = 1.81369, z-score = 5.92, p-value < 0.05), and the word *morpheme* (β = 0.78839, z-score = 2.60, p-value < 0.05), and a significant interaction effect of *new* with the productivity of the suffix (β = -1.39553, z-score = -2.3, p-value < 0.05). Thus, ChatGPT was approximately six times more likely to provide a humanlike definition if the prompt included the word *new*, approximately twice as likely to provide a humanlike definition if the prompt included the word *morpheme*, and the effect of adding the word *new* was almost four times greater for words with an unproductive suffix, than for words with productive suffixes.

4 Discussion

While our analysis did not identify one prompt that was significantly better than all others in humanlikenss or plausibility, we found that the model always performed better when the prompt provided additional contextual information in the form of the words *new* and *morpheme*. As expected, the more specific the task is for the language model, the better it performs. Thus, our results also reinforce Ekin's (2023) suggestions on giving specific details about the task.

Better performance on plausibility for words with a productive affix suggests that ChatGPT may be better at identifying form-meaning mappings for productive suffixes than for unproductive ones. This finding is in line with what authors like Bertram et al. (2000) found for language acquisition in children, and might be, similarly, caused by frequency of the affix. For example, the more productive the morpheme, the more often it may occur in the training data, and the more information the model could have regarding the contexts it appears in. Eventually this might result in a more accurate representation of the meaning of the morpheme.

In contrast, productivity had no significant effect on humanlike outputs of the models, which might indicate that high probabilities counts of productive morphemes do not encompass characteristics of human definitions. This might be because human definitions rely often on aspects that cannot be inferred from probabilities counts, such as personal experience or world knowledge when defining new words, e.g. defining *musksome* as having a certain kind of smell.

However, though productivity had no effect on humanlike definitions, there was a significant interaction between productivity and the word *new*. More exactly, adding the word *new* to the prompt always improves it. However, the impact of the word 'new' on the humanlikeness of the generated definition is less strong if the morpheme is productive. Thus, for words with unproductive morphemes, the model may improve more because it may be more likely to "look for" a morphological decomposition as a heuristic when given the additional context.

This interaction effect between productivity and the word 'new' can also be influenced by the root morphemes of the words. This might be because the most salient word meaning differs significantly for humans and ChatGPT: given the model was trained on data for a numerous specialised tasks, it sometimes regards specialised definitions very likely. This was clearly shown for the word *assemblyless*. We found that while almost all human participants defined the word in relation to an *assembly* as a 'group or gathering of people' or 'the process of assembling something', ChatGPT most frequently drew the connection to a programming language called 'assembly'. This resulted in less humanlike definitions for the model, showcasing humanlike definitions were always plausible, but not the other way around, as shown in the example below:

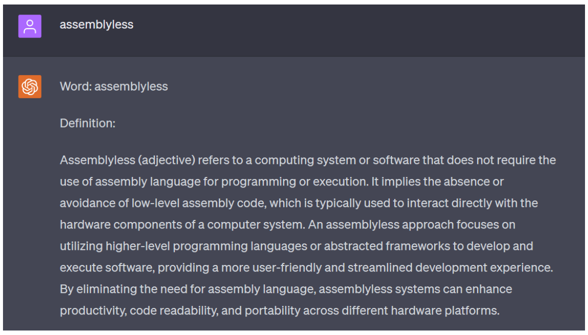

Figure 1: Definition for *assemblyless* provided by ChatGPT using the 'Word generator' prompt template.

Even more, productive morphemes are defined in Ford et al. (2010) in terms of appearance in hapax legomena, which might indicate they form more specialised words. Thus, the probability of generating specialised definitions can increase because our definition of productive suffixes is defined, to some extent, in relation to specialised words. This would overall increase chance of generating specialised word definitions if the root morpheme can also have a specialised meaning, like *assembly*.

Even for words without a root morpheme with specialised meanings, productive derivational morphemes might often be associated with various root morphemes with diverse meanings, which can lead to bigger perplexity and, consequently, less specific or humanlike output. Comparatively, unproductive morphemes have less chances to be associated with more words, which would make the model stick to a few definitions that might also be more specific.

Thus, depending if the user intends to obtain more humanlike or just plausible definitions, the prompt that scored highest in one of the criteria can be used. The statistical results can also be used to further create new prompts by addition of the two contextual words, i.e. 'new' and 'morpheme', depending on the tasks morphological connection will be used in. Note that the both conditions of the best humanlike prompt either have no context, or both context words. This indicates some successful prompts follow an overall opposite effect, i.e. they lack context words but perform well.

Lastly, if the reader wants to achieve a balance in the chosen prompt, i.e. to be as humanlike and as plausible as possible, the 'Persona' word generator prompt, condition + new, + morpheme may be the best choice, given its high scores in both criteria, averaged at 9.5 correct definitions for morphological connection.

5 Conclusion

Our study aimed to identify an optimal prompting strategy for ChatGPT to draw morphological connections while producing word definitions for unfamiliar words. We conducted a word definition task using different prompts, and compared the model's performance on plausibility and humanlikeness for definitions that it provided for a set of morphologically complex nonce words.

We found that irrespective of the prompt template, adding the words *new* and *morpheme* to a prompt, significantly improved the performance of ChatGPT. Thus, the users looking to obtain definitions of unfamiliar morphologically complex words from the model can apply the current findings by including such keywords in their prompts.

In terms of best prompting strategies, our results found two prompts that had the maximal performance on plausibility, i.e. 'Persona' prompt word generator, condition + new, + morpheme and condition - new, + morpheme, and 'Context manager' prompt, condition + new, + morpheme. In terms of humanlikeness, our results show one prompt that had the best performance, i.e. 'Persona' prompt word generator, condition - new, - morpheme, and condition + new, + morpheme. These results also indicate the 'Persona' prompt word generator, condition + new, + morpheme scores, on average, the highest. However, note that the statistical analysis did not point to an overall best prompt.

We found evidence that ChatGPT's response is however also always modulated by factors such as the nature of its training data and world knowledge, which can lead it to produce definitions which while plausible, may not be humanlike. This is something that may be of interest to researchers in the future.

Limitations

The first limitation to our study is small sample sizes in terms of the number of prompting strategies tested, the number of words tested, the number of human participants to define *humanlike* performance, and the number of contextual factors compared. Given that this was designed as a pilot study, we restricted our sample in a number of ways, so expanding the testing material in any direction would be a useful direction for future research.

We also recognise that one prominent difference between the way that we prompted the model and the human participants is that we specifically suggested to the latter that they provide examples with their definitions, and we did not do so for ChatGPT. This meant that especially for humanlikeness, a very brief or vague definition from the model was difficult to categorise, since we would not necessarily obtain direct evidence that the model was exploiting morphological relatedness in a humanlike way. Including a request for examples in the prompt text in future research might help in the coding of outputs.

With regards to the test words, we also wish to point out that for all the nonce words, the lexical root morpheme could always also be used as a free morpheme. That is to say, we did not look at how allomorphy might affect the performance of the model, a point which would be of great interest in the context of the form-meaning mapping question. If we had used a nonce word like *proficiate*, a hypothetical combination of the lexical root in *proficiency*, and the verbalising suffix *-ate*, how successful might the model be at parsing the word into its component morphemes and extrapolating a definition? In addition to that, all the nonce words differed from a familiar word by only a single morpheme, so it would be worth looking at how the model performs when the number of derivational morphemes involved increases. Future studies should also consider controlling for meaning of root morphemes, i.e. for specialised meanings, so as to rule the possibility of the model to choose more unlikely humanlike definitions.

According to previous studies (Ortega-Martín et al., 2023), ChatGPT generalises knowledge with time. To test this, the study initially planned to have prompts run a second time. However, due to limitations of time and because the prompting was done in 3 rounds, testing for generalisation could not be systematic, and therefore, would have not been relevant anymore. It would be interesting in the future to investigate if generalisation exists and if the model can become increasingly specialised on the topic only by prompting. If true, this might be an alternative to fine-tuning to some extent.

Finally, as we pointed out in the introduction, we focused on English as our test case, despite the observation that other languages have far more extensive morphological systems. This reduces the generalisability of our conclusions to the performance of the model with languages with richer morphological systems. Morphological decomposition is especially important for languages with rich or agglutinative morphology. For example, rich morphology has been anticipated to be problematic since Creutz and Lagus (2002), where high morpheme productivity would lead to a irrationally high number of distinct types, which, in turn, would lead to poor comprehension abilities. Languages rich in morphology still raise difficulties today. For example, Belinkov et al. (2017) trained a classifier for morphological prediction on word feature representations from a machine translation model and showed that representations learned in the English model do better when predicting morphology, than those for languages richer in morphology such as Hebrew. As the authors remark, one could expect models for languages with richer morphologies to encode more morphological knowledge. However, the inherent difficult nature of translating into a language with rich morphology might make the model perform overall worse, which would result in weaker features representations. In this respect, future studies might also look to expand the number of test cases to improve the validity of our findings, but to also test if our methodology can improve morphological features of models by forcing better decomposition of words.

Lastly, future work should consider automatic methods in designing prompts too. As Wang et al. (2023) remarks, manually designing prompts is expensive and time-consuming. Thus, future studies in prompting for morphological connection might benefit from deploying models in designing prompts, i.e. seeing the task as sequence-to-sequence generation task (for a better review, see Wang et al., 2023).

Acknowledgements

We would like to thank Dr. Denis Paperno for his supervision and advice during the project, as well as our peers at Utrecht University for their input and feedback. We would also like to acknowledge the contributions of the two anonymous reviewers whose comments helped us refine this article.

References

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Raymond Bertram, Matti Laine, and Minna Maria Virkkala. 2000. The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology*, 41(4):287–296.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, MPL '02, pages 21–30, USA. Association for Computational Linguistics.

Sabit Ekin. 2023. Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices. *TechRxiv*.

Michael A. Ford, Matt H. Davis, and William David Marslen-Wilson. 2010. Derivational morphology and base morpheme frequency. *Journal of Memory and Language*, 63(1):117–130.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Najoung Kim and Paul Smolensky. 2021. Testing for grammatical category abstraction in neural language models. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 467–470, Online. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

Alessandro Lenci. 2023. Understanding natural language understanding systems. a critical analysis. *ArXiv*, abs/2303.04229.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective.

R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN. *Transactions of the Association for Computational Linguistics*, 11:652–670.

OpenAI. 2021. ChatGPT 3.5. OpenAI Blog. `https://openai.com/blog/chatgpt`.

Miguel Ortega-Martín, Óscar García-Sierra, Alfonso Ardoiz, Jorge Álvarez, Juan Carlos Armenteros, and Adrián Alonso. 2023. Linguistic ambiguity analysis in chatgpt.

RStudio Team. 2020. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, Enze Shi, Yi Pan, Tuo Zhang, Dajiang Zhu, Xiang Li, Xi Jiang, Bao Ge, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang. 2023. Review of large vision models and visual prompt engineering.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

A Appendix A: Prompts

Define Pattern

1. Define...
2. Define the *new* word ...
3. Define the word ... considering its *morphemes*.
4. Define the *new* word ... considering its *morphemes*.

Context Manager Pattern

1. I want you to generate a definition for the word I provide.
2. I want you to generate a definition for the *new* word I provide.
3. I want you to generate a definition for the word I provide. When analysing the word, especially consider its *morphemes*.
4. I want you to generate a definition for the *new* word I provide. When analysing the word, especially consider its *morphemes*.

Template Pattern

1. Provide definitions for words. I am going to provide a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders that I list. Please preserve the formatting and overall template that I provide as WORD, means DEFINITION.
2. Provide definitions for words. I am going to provide a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders that I list. Please preserve the formatting and overall template that I provide as *NEW* WORD, means DEFINITION.
3. Provide definitions for words. I am going to provide a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders that I list. Please preserve the formatting and overall template that I provide as WORD, with *morphemes* MORPHEMES, means DEFINITION.
4. Provide definitions for words. I am going to provide a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders that I list. Please preserve the formatting and overall template that I provide as *NEW* WORD, with *morphemes* MORPHEMES, means DEFINITION.

Word Generator Persona

1. You are going to pretend to be a word generator. When I type in a word, you are going to output the corresponding definition that the word generator would produce.
2. You are going to pretend to be a word generator that can generate *new* words. When I type in a word, you are going to output the corresponding definition that the word generator would produce.
3. You are going to pretend to be a word generator that can generate words only by considering the *morphemes* of words. When I type in a word, you are going to output the corresponding definition that the word generator would produce.
4. You are going to pretend to be a word generator that can generate *new* words only by considering the *morphemes* of words. When I type in a word, you are going to output the corresponding definition that the word generator would produce.

Lexicographer Persona

1. From now on, act as a lexicographer when providing definitions of words. Provide outputs that a lexicographer would regarding words.
2. From now on, act as a lexicographer when providing definitions of *new* words. Provide outputs that a lexicographer would regarding words.
3. From now on, act as a lexicographer when providing definitions of words. Pay close attention to the *morphemes* of any word that we talk about. Provide outputs that a lexicographer would regarding words.
4. From now on, act as a lexicographer when providing definitions of *new* words. Pay close

attention to the *morphemes* of any word that we talk about. Provide outputs that a lexicographer would regarding words.

Infinite Generation Pattern

1. From now on, I want you to generate a definition for the word I provide until I say stop. I am going to provide a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders that I list. Please preserve the formatting and overall template that I provide at WORD, means DEFINITION

2. From now on, I want you to generate a definition for the *new* word I provide until I say stop. I am going to provide a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders that I list. Please preserve the formatting and overall template that I provide at WORD, means DEFINITION.

3. From now on, I want you to generate a definition for the word I provide until I say stop. I am going to provide a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders that I list. Please preserve the formatting and overall template that I provide at WORD, with *morphemes* MORPHEMES, means DEFINITION

4. From now on, I want you to generate a definition for the *new* word I provide until I say stop. I am going to provide a template for your output. Everything in all caps is a placeholder. Any time that you generate text, try to fit it into one of the placeholders that I list. Please preserve the formatting and overall template that I provide at WORD, with *morphemes* MORPHEMES, means DEFINITION.