Discovering Hidden Patterns in the Data: The Value of Input and Output Analysis in Optimizing LLM Prompt Chains

Isabel Deibel

Abstract

This study investigates how prompt engineering can optimize for distractor plausibility in GPT-generated fill-in-the-blank language exercises by comparing output from three different prompt chains. The findings suggest that clarity and conciseness in prompt chains may outperform more complex ways of prompting, and that linguistic patterns in the input and output provide insightful data that may be crucial for better prompting success. These insights help us to understand the impact of prompt engineering on complex prompt chains and to adjust prompting strategy in order to generate more optimal outputs.

Background

Optimizing prompts to produce the highest quality output is at the center of recent interest in the usage of Large Language Models (LLMs). While it's a known fact that rewriting a prompt to vary the wording, layout or conceptual presentation of the task can result in significant positive or negative changes to the model's performance on a given task, it can be a tedious manual task when different prompts cannot be easily compared to each other at scale and when, as a result, the return on investment for a human to change the prompt structure is unclear (s. Zhou et al. 2023). This issue is exacerbated by the fact that, for many applications at scale, a simple prompt may not suffice to produce optimal outputs in light of a high number of crucial constraints on the output, leading prompt engineers to craft more complex prompts, which in turn become harder to assess for areas of optimization.

The current paper sets out to provide insight into how prompts utilizing different state-of-the-art prompt engineering techniques may affect LLM-generated output by focusing on one step within a complex prompt chain aimed at automatically creating fill-in-the-blank language education content (Example 1). In particular, this analysis explores prompt optimization in an area that appears to be challenging for LLMs like OpenAI's GPT-4 "out-of-the-box". As a transformer-based model, GPT-4 is "pre-trained to predict the next token" (OpenAI 2023: 1). When used to create fill-in-the-blank exercises, in order to create distractor options that do not fit the exercise as likely answers (which would be unacceptable for educational content), GPT has to essentially predict something that – counter its training – is not the next most likely token (s. Appendix 1 for evidence from zero-shot prompts that can't reliably produce acceptable exercises).

Example 1 showing the effects of plausibility in a (human-made) fill-in-the-blank exercise. Possible plausibility is indicated by "?" (s. Appendix 2 for further discussion):
Fill in the blank sentence: I'm going to the ___ tonight because I want to buy some groceries.
Correct Answer: store
Distractors: bedroom, plant, ?office

In light of the limitations that simple prompts exhibit for this problem, the current study compares output elicited from three prompt chains employing current state-of-the-art prompting strategies, to explore whether and how prompt engineering can optimize for distractor plausibility. The analysis shows that considering only acceptance rates and input variables as predictors for optimal output does not tell the whole story and that it may be necessary to conduct more fine-grained data analysis or mining to discover additional patterns that could inform prompting strategy as a whole (i.e., how to improve prompts and prompt chains, which prompts to utilize and how to prioritize inputs for optimal output).

Method

In order to achieve optimal output for a task that GPT struggles to complete when prompted with simple prompts, the three prompt chains that were used in the current study all employ chain-of-thought prompting that breaks down the creation of fill-in-the-gap exercises into smaller incremental steps, combined with few-shot examples with and without labels (Wei et al. 2022; Brown et al. 2020; Min et al. 2022). Appendix 3 shows the high-level make-up of the three prompt chains utilized in the current study. In order to keep prompt variations constrained, only prompts 3 and 4 (the prompts that elicit distractors) within these chains were edited for each of the prompt versions.

- Prompt Chain 1: describes distractors concisely as “implausible”
- Prompt Chain 2: describes distractors as semantically unlikely in the blank and provides labels for likelihood
- Prompt Chain 3: describes distractors as semantically unlikely in the blank but does not provide labels

For each prompt chain, output was elicited utilizing GPT-4-0314 between November 6-10, 2023, for a total of 405 fill-in-the-blank observations. All observations were manually rated by a human (the author) and auto-rated by GPT (with a few-shot prompt) as plausible or implausible in order to calculate agreement in ratings. In addition to agreement, the full data set was coded for the following variables: target word, context, target word part of speech (POS), prompt chain type, blanked word and blanked word part of speech (s. Appendix 4 for more information on the materials).

Given the subjective nature of plausibility (s. Appendix 2), the working hypothesis for this study was that Prompt Chain 2 (containing a clear semantic description and labels) would be the best candidate to create implausible distractors and Prompt Chain 1 the worst due to the lack of further definition of plausibility.

Results

Overall, there was a 60% (245/405) overlap between the human judgment and GPT judgments as concerns acceptable items. The human rater accepted significantly more items than GPT ($X^2(1) = 29.05$, $p < .05$), but was reviewing the distractor sets as a whole (finding an answer that’s clearly best, s. Appendix 2) while GPT had no clear instructions along those lines. While cross-tabulations of observations showed no significant difference in agreement for the three different prompt chains ($X^2(2) = 3.99$, $p < .14$), this does not account for any lexical factors that might be operating “under the hood” of each prompt for actual input variables.

Analyzing input variables

A mixed effects logistic regression analysis modeling the impact of the input variables (prompt type, target word, target word POS and context) on human/GPT agreement indicates that, with all fixed effect predictors at their reference levels (Prompt Chain Type: 1, Target Word POS: noun), the estimated odds of disagreement on acceptability of an item (i.e plausibility) are smaller than the odds of agreement (i.e., no plausibility). The odds of disagreement are significantly increased when a different Prompt Chain Type 3 or a verb as the target input word are used to generate fill-in-the-blank exercises (s. Appendix 5 for all results).

A savvy reader may notice that the target word in the input is only indirectly related to the plausibility of the distractor set given that GPT places the blank during generation. The result that the input word POS significantly impacts the ratings, thus, justifies further examination of the output itself if we want to describe the data fully.

Analyzing input and output variables together

A mixed effects logistic regression analysis modeling the impact of the input variables (target word, target word POS and context) as well as output variables (blanked word and blanked word POS) on human/GPT agreement indicates that, with all fixed effect predictors at their reference levels (Target Word POS: noun, Blanked Word POS: verb), the estimated odds of disagreement on acceptability of an item (i.e plausibility) are smaller than the odds of agreement (i.e., no plausibility). The odds of disagreement are significantly increased when a verb or adjective as the target input word are used to generate fill-in-the-blank exercises as well as when the word in the gap is an adjective. Thus, the odds of agreement are highest when the target word is a noun and the blanked word is a verb (s. Appendix 6). Including prompt chain type as a fixed effect in this model did not lead to convergence.

Discussion and Conclusion

The data presented here indicates that for certain highly-optimized prompt chains, conciseness and clarity trumps (Prompt chain 1) more accurate yet complex descriptions of plausibility (Prompt chains 2 and 3) although labels help GPT complete the task in the latter case. More importantly, however, the current data shows that linguistic patterns in the input and output may actually be more insightful than success/failure counts, suggesting that we may not be able to fully identify optimal prompts without analyzing or mining outputs for additional patterns and that we run the risk of missing the big picture and discarding good prompts if we base decisions on success/failure counts only. Input variables can hint at underlying patterns in the data but deeper linguistic output analysis is needed in order to identify areas of highest prompting success and adjust prompting strategy. In the current case, we should consider optimizing prompts for distractor placement and POS. More research is needed to compare the present results with prompt chains that are less optimized for optimal output.

Appendix

Appendix 1: Zero-shot prompts fail to reliably create fill-in-the-gap-exercises

Utilizing simple zero-shot prompts demonstrates that GPT cannot reliably produce acceptable exercises (Table 1).

Table 1: Using Zero-shot prompts to generate fill-in-the-blank exercises.
Example 2a: Prompt: "Using the following context and target word, generate a sentence less than 100 characters long, blank out a word that is not the target word, and provide three distractors. Context: going to the mall to buy clothing Target Word: t-shirt" GPT-generated Output (generated on November 17, 2023): Sentence: I'm going to the mall to buy a new t-shirt and a pair of _____. Blanked Word: jeans Distractors: shoes, shorts, sunglasses
Example 2b: Prompt: "Using the following context and target word, generate a sentence less than 100 characters long, blank out a word that is not the target word, and provide three distractors. The distractors should fit into the sentence grammatically but not semantically. They should also be parallel in form to the correct answer. Context: going to the mall to buy clothing Target Word: t-shirt" GPT-generated Output (generated on November 17, 2023): Sentence: When I went to the mall, I bought a blue t-shirt and a pair of ________. Blanked Word: jeans Distractors: shovels, pancakes, rainbows

As we can see, given these prompts, GPT apparently selects distractors based on what would best complete "a pair of" (example 2a). This is obviously not ideal for educational purposes, as all answer options can clearly be chosen as correct answers. In example 2b, providing further instructions helps to create more implausible responses but "shovels" could still be bought in a pair.

GPT has explicitly acknowledged its limitations when prompted for fill-in-the-blank distractors that are implausible: "I apologize for the misunderstanding. Currently, GPT-3 and GPT-4 models have limitations when it comes to understanding complex constraints and producing specific types of output, like generating semantically incorrect distractors that fit grammatically in a sentence. The models are designed to generate the most plausible and coherent text, which makes the task of generating semantically incorrect but grammatically fitting distractors quite challenging" (GPT-4, queried on November 8, 2023).

Appendix 2: Plausibility in fill-in-the-blank exercises is subjective

As indicated in example 1, the distractors in fill-in-the-blank exercises tend to fall on a subjective spectrum of plausibility, ranging from more to less plausible. "Bedroom" and "plant" are perhaps trending in the more implausible direction if one assumes that groceries can't typically be bought in bedrooms or near plants. Along these lines, "office" is not necessarily a place where you typically can buy groceries, but the perspective that groceries might be sold on the way to the office is not entirely implausible. Notice, however, that if we define plausibility based on whether there is one option within the set that is clearly the best answer, rather than whether each individual distractor is plausible, that may mean that we would lean towards accepting this exercise as implausible given that “store” is clearly the best place to purchase groceries out of all options. Either way, given the subjective nature of plausibility, we will leave the decision of whether to consider this exercise too plausible to be shown to a learner up to the reader. However, if our goal is to fully autogenerate such exercises with GPT, accounting for plausibility is crucial, as otherwise exercises may not be usable for (clear) teaching. Potential negative ramifications would be that this prompting structure does not scale well and results in negative business impact (i.e., learners experiencing frustration over unsolvable exercises).

Appendix 3: High-level prompt chain structure

Table 2 shows the high-level prompt chain structure for all three prompt chains used in the current study. Notice that the three prompt chains share the same basic structure: The task of writing the exercise is split into different subtasks that are prompted for separately. The entire prompt chain is also split into two separate parts: a modeling and generation part. The first four prompts serve as a conversation model for the computer; after the modeling part, all user prompts are repeated for the actual word in question until prompt 3. The first generation occurs after prompt 4 when the first user prompt is repeated with the actual target word we want to generate for. Instructions differ in prompts 3 in order to assess whether different instructions can lead to less plausible distractors.

Table 2: High-level prompt chain structure

Prompt Chain 1 “Implausible”	**Prompt Chain 2 “Semantic Likelihood & Label”**	**Prompt Chain 3 “Semantic Likelihood, no Label”**
System Instructions: Giving general instructions for GPT's role and the main task		
User (prompt 1 – model): Prompt to write a sentence of 100 characters or less that includes the target word.		
Computer (non-generated response / model): a placed example sentence		
User (prompt 2 – model): Prompt to select a semantically related word to the target word as the correct answer		
Computer (non-generated response / model): a placed example sentence and correct answer		
User (prompt 3 – model): Prompt to add 3 implausible distractors	**User (prompt 3 – model):** Prompt to add 3 distractors based on low semantic likelihood	**User (prompt 3 – model):** Prompt to add 3 distractors based on low semantic likelihood
Computer (non-generated response / model): A placed example sentence, correct answer and implausible distractor	**Computer (non-generated response / model):** A placed example sentence, correct answer and implausible distractor; added labels of likelihood of each answer	**Computer (non-generated response / model):** A placed example sentence, correct answer and implausible distractor
User (prompt 4 – model): Prompt for additional examples		
Computer (non-generated response – model): 4 additionally placed examples that follow the same layout as in the previous computer step [—this is the end of the modeling part—]		
User (prompt 1 – real generation): Prompt to select a semantically related word to the target word as the correct answer **for a new context and target word**		
Computer (generated response): a sentence		
… Conversation continues for the new context and target word with all responses now generated until prompt 3 at which point the exercise is fully assembled …		

The instructions that were inserted in prompts 3 were as follows:

Prompt Chain 1 (“Implausible”):

User: Add 3 Distractors to this exercise following the following rules for them:

- Distractors have the same grammatical form as the Correct Answer (e.g., all nouns, verbs, gerunds, etc.).
- Distractors should have a high grammatical likelihood of following the word immediately prior to the Correct Answer. This means that if a distractor requires a specific form of the word prior to the Correct Answer to be considered grammatically correct, you need to pick a different one. A common pitfall is a Correct Answer that starts with a vowel, which co-occurs with "an" prior to it. That means that all Distractors should also be able to follow "an" and be grammatically correct.
- Distractors should be implausible when inserted in the sentence.

In summary, when Distractors replace the Correct Answer in the sentence, the result is an illogical or impossible meaning but a grammatically correct sentence.

Computer:
Sentence: Maria is getting a few groceries because she's going to try a new recipe tonight!
Correct Answer: try
Distractor 1: fly
Distractor 2: be
Distractor 3: ride

Prompt Chain 2 (“Semantic Likelihood & Label”):

User: Add 3 Distractors to this exercise following the following rules for them:

- Distractors have the same grammatical form as the Correct Answer (e.g., all nouns, verbs, gerunds, etc.).
- Distractors should have a low semantic likelihood of following the word immediately preceding the Correct Answer but overall a high grammatical likelihood of following the word immediately prior to the Correct Answer. This means that if a distractor requires a specific form of the word prior to the Correct Answer to be considered grammatically correct, you need to pick a different one. A common pitfall is a Correct Answer that starts with a vowel, which co-occurs with "an" prior to it. That means that all Distractors should also be able to follow "an" and be grammatically correct.

In summary, when Distractors replace the Correct Answer in the sentence, the result is an illogical or impossible meaning but a grammatically correct sentence. For all distractors that you add, add the semantic likelihood of them being inserted into the sentence instead of the Correct Answer.

Computer:
Sentence: Maria is getting a few groceries because she's going to try a new recipe tonight!
Correct Answer: try
Distractor 1: fly
Distractor 2: be
Distractor 3: ride

Distractor 1 likelihood of replacing the Correct Answer: low
Distractor 2 likelihood of replacing the Correct Answer: low
Distractor 3 likelihood of replacing the Correct Answer: low

Prompt Chain 3 ("Semantic Likelihood, no Label"):

User: Add 3 Distractors to this exercise following the following rules for them:

- Distractors have the same grammatical form as the Correct Answer (e.g., all nouns, verbs, gerunds, etc.).
- Distractors should have a low semantic likelihood of following the word immediately preceding the Correct Answer but overall a high grammatical likelihood of following the word immediately prior to the Correct Answer. This means that if a distractor requires a specific form of the word prior to the Correct Answer to be considered grammatically correct, you need to pick a different one. A common pitfall is a Correct Answer that starts with a vowel, which co-occurs with "an" prior to it. That means that all Distractors should also be able to follow "an" and be grammatically correct.

In summary, when Distractors replace the Correct Answer in the sentence, the result is an illogical or impossible meaning but a grammatically correct sentence. For all distractors that you add, add the semantic likelihood of them being inserted into the sentence instead of the Correct Answer.

Computer:
Sentence: Maria is getting a few groceries because she's going to try a new recipe tonight!
Correct Answer: try
Distractor 1: fly
Distractor 2: be
Distractor 3: ride

Appendix 4: Materials and Coding

Table 3 shows all input variables. Five observations were collected from each input variable (e.g., five fill-in-the-blank exercises were generated with the target word "dress" in the context "going to the mall to buy clothing". Each context yielded a total of 45 observations per prompt.

Table 3: Input variables (target words by part of speech for each context)

Context	**Target Words by Part of Speech**
Going to the mall to buy clothing	**Nouns:** dress, hat, t-shirt **Verbs:** try on, pay for, need **Adjectives:** cheap, expensive, elegant
You're on vacation at the beach with your family	**Nouns:** kids, sand castle, ship **Verbs:** swim, play, relax **Adjectives:** noisy, warm, excited
You're at the supermarket	**Nouns:** cash, dinner, fruit **Verbs:** look for, make, read **Adjectives:** discounted, favorite, fresh

Each observation was coded for the following variables (reference levels presented first):

- Target Word POS: noun, verb, adjective
- Context: "going to the mall to buy clothing", "you're on vacation at the beach with your family", "you're at the supermarket"
- Prompt Chain Type: 1 ("implausible"), 2 ("Semantic Likelihood & Label"), 3 ("Semantic Likelihood, no Label")
- Agreement (a binary variable to represent where human labels and GPT labels overlapped that items were acceptable or disagreed): acceptable, disagree
- Blanked Word POS: verb, noun, other (16 adjectives, 1 preposition)

Appendix 5: Analyzing input variables

A generalized linear mixed effects (logistic) regression was fit in R using the lme-4 package with agreement (agree to accept/disagree) in whether the exercise was acceptable/ plausible as the dependent variable, prompt type and target word part of speech as independent variables, and context and target word as random intercepts (Table 4) (R Core Team 2012; Bates et al. 2015).

The results in Table 4 indicate that, with all predictors at their reference levels (Prompt Chain: 1, Target Word POS: noun), the estimated odds of disagreement on acceptability of an item (i.e plausibility) are smaller than the odds of agreement (i.e., no plausibility) [exp(-1.34) = 0.26]. Positive estimates of the coefficients indicate a higher likelihood of disagreement on acceptability (i.e., plausibility), that is, the odds of disagreement are significantly increased when Prompt Chain 3 or a verb as the target input word are used to generate fill-in-the-blank exercises.

Table 4: Results of a generalized linear mixed effects (logistic) regression with agreement (agree to accept/ disagree) in whether the exercise was acceptable/ plausible as the dependent variable, prompt type and target word part of speech as independent variables, and context and target word as random intercepts

	Estimate	Std. Error	z-value	p-value
Intercept	-1.34	0.4	-3.35	< .001
Target Word POS (adjective)	0.78	0.42	1.86	< .07
Target Word POS (verb)	0.82	0.42	1.96	< .05
Prompt Chain (2, labeled)	0.33	0.27	1.22	0.2
Prompt Chain (3, unlabeled)	0.57	0.27	2.13	< .04

Comparison of this model against the null model with only random effects and without fixed effects shows a marginally better fit that is not statistically significant (AIC (reported model) = 524.9; $X^2(4)$ = 9.14, p<.06; Variance (Context) = 0.12; Variance (Target word) = 0.44). Comparisons against models with either fixed effect showed similar results (Only Prompt Type as fixed effect: $X^2(4)$ = 4.55, p<0.2; Only Word POS as fixed effect: $X^2(4)$ = 4.55, p>0.1; $X^2(4)$ = 4.6, p<0.2). This suggests that, when analyzing GPT output, it is not sufficient to only consider the input variables.

In sum, this data appears to suggest that conciseness and clarity (Prompt Chain 1) trumps more complex descriptions of plausibility in the prompt instructions (Prompt Chains 2 and 3). If a more complex description of the task is chosen, using labels to support GPT completing the task (Prompt Chain 2) outperforms the prompt version without labeling (Prompt Chain 3). However, as stated above, this does not seem to tell the whole story.

Appendix 6: Analyzing input and output variables

A generalized linear mixed effects (logistic) regression was fit in R using the lme-4 package with agreement in whether the exercise was acceptable/plausible as the dependent variable, target word part of speech and part of speech of the blanked word as independent variables, and context, blanked word and target word as random intercepts (Table 5) (R Core Team 2012; Bates et al. 2015).

The results in Table 5 indicate that, with all predictors at their reference levels (Target word POS: noun, Blanked Word POS: verb), the estimated odds of disagreement on acceptability of an item (i.e plausibility) are smaller than the odds of agreement (i.e., no plausibility) [exp(-1.19) = 0.3]. Positive estimates of the coefficients indicate a higher likelihood of disagreement on acceptability (i.e., plausibility), that is, the odds of disagreement are significantly increased when Prompt Chain 3 or a verb as the target input word are used to generate fill-in-the-blank exercises.

Table 5: Results of a generalized linear mixed effects (logistic) regression with agreement (agree to accept/ disagree) in whether the exercise was acceptable/ plausible as the dependent variable, target word part of speech and part of speech of the blanked word as independent variables, and context, blanked word and target word as random intercepts.

	Estimate	Std. Error	z-value	p-value
Intercept	-1.19	0.45	-2.63	< .01
Target Word POS (adjective)	0.87	0.35	2.49	< .02
Target Word POS (verb)	0.81	0.4	2.03	< .05
Blanked Word POS (noun)	0.03	0.41	0.08	0.9
Blanked Word POS (other)	2.59	0.88	2.93	< .01

Comparison of this model against the null model with only random effects and without fixed effects shows significantly better fit (AIC (reported model) = 503.6; $X^2(4)$ = 18.5, p<.005; Variance (Context) = 0.21; Variance (Target word) = 0.03; Variance (Blanked Word) = 1.23). This model also showed significantly better fit than models with only one fixed effect (Blanked Word POS: $X^2(2)$ = 6.23, p<.05; Target Word POS: $X^2(2)$ = 11.97, p<.005). As concerns the random effects, inclusion of the blanked word significantly increased fit ($X^2(1)$ = 14.8, p<.005), similar for context ($X^2(1)$ = 4.03, p<.05), but not for target word ($X^2(1)$ = 0.04, p>0.8). Given that target word part of speech increased fit as a fixed effect, the final model retains the random effect for target word.

Models with an added interaction for target word part of speech and blanked word part of speech as well as with an added fixed effect for Prompt Type failed to converge.

It is worth addressing that the variance for the random effect for blanked word is quite large. Recall that GPT selected the words in the blanks based on whether they were semantically related to the target word. There were no other restrictions on selecting these words, so high variability is somewhat expected.

References

Bates, D.; Mächler, M.; Bolker, B. J.; and Walker, S. C. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48. doi:10.18637/jss.v067.i01.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33*:*1877–1901. https://arxiv.org/pdf/2005.14165.pdf

Min, S.; Lyu, X.; Holtzman, A.; Artetxe, M.; Lewis, M.; Hajishirzi, H.; Zettlemoyer, L. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. https://arxiv.org/pdf/2202.12837.pdf

OpenAI. 2023. GPT-4 Technical Report. https://arxiv.org/pdf/2303.08774.pdf

R Core Team. 2023. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at http://www.R-project.org/, accessed November 17, 2023.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS.* https://arxiv.org/abs/2201.11903

Zhou, Y.; Muresanu, A. I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; Ba, J. 2023. Large Language Models Are Human-Level Prompt Engineers. https://arxiv.org/abs/2211.01910