Title: Prompt Architecture Can Induce Methodological Artifacts in Large Language Models

Authors: Melanie Brucks, Olivier Toubia

The authors contributed equally and are listed alphabetically.

Melanie Brucks (mb4598@columbia.edu) is an Assistant Professor of Marketing and Olivier Toubia (ot2107@columbia.edu) is the Glaubinger Professor of Business at Columbia Business School, 3022 Broadway, New York, NY 10027.

Abstract

Large Language Models (LLMs) are subject to extraneous associations that produce methodological artifacts—biases due to the arbitrary architecture of a prompt. Multiple large-scale experiments performing standard (zero-shot) evaluation tasks using GPT-4 demonstrate sensitivity to the order and labeling of response options in a prompt. LLMs are poised to replace humans for many tasks but ignoring these methodological artifacts can lead to erroneous conclusions. These results suggest that any single prompt yields inherent bias simply due to its architecture, thus we recommend aggregating the results from multiple prompts that vary according to full factorial designs.

Introduction

Generative AI models and tools, and in particular Large Language Models (LLMs) such as GPT, are rapidly disrupting many industries and fields of study, with many scientists and practitioners actively contemplating the use of LLMs as substitute for humans in a wide range of tasks and occupations (1–7). At first blush, LLMs seem particularly well-suited for some of the tedious tasks typically performed by humans, such as combing through hordes of textual data and evaluating these data on various dimensions. LLMs have virtually unlimited memory and processing abilities, are presumably not influenced by emotions or moods, and never seem to lack motivation.

As Generative AI continues to gain traction, however, researchers are cautioning against hasty and indiscriminate use because these models tend to perpetuate pernicious social stereotypes and prejudices that may be embedded in the content on which they are trained. For example, due to co-occurrences in their corpora, LLMs may reveal gender stereotypes such as perceiving a "nurse" to be closer to "woman" and a "doctor" to be closer to "man" (8, 9). However, LLMs also rely on countless other seemingly arbitrary and unpredictable associations embedded in the rich language data used to train them (10). Here we propose that these associations interact with the architecture of the prompt, systematically biasing the output of LLMs. Thus, just as the context of a choice in which people make decisions biases human judgment, we suggest prompt architecture biases LLM output (11, 12). Because prompt architecture does not meaningfully change the task itself, these biases produce methodological artifacts (13).

Nascent work employing LLMs has largely overlooked the impact of prompt architecture, rendering their results prone to error. For example, a manager (researcher) eager to replace their workforce (human subjects) with Generative AI, may ask GPT to evaluate whether item B or C is closer to A. The implicit assumption is that ChatGPT operates under the normative principle of procedure invariance, revealing a stable and reliable measure of similarity between options regardless of response order or labeling. Supporting this assumption, ChatGPT claims, "As an AI, I don't have any biases towards specific answer choices based on their order or labels" (see Supplemental Information A). Contrary to this claim, we find that the label and order in a prompt systematically and significantly bias the output of LLMs. Without taking this bias into account, a manager may erroneously conclude that item A is closer to item B simply due to the order and label of response options.

These results suggest that any single prompt yields inherent bias simply due to its architecture. Indeed, as with choice architecture, "there is no such thing as a neutral design"(11)—the architecture of any prompt has the potential to add systematic error to LLM output. As a result, "prompt engineering," which typically attempts to identify the single, optimal prompt for a given task, may be a futile exercise, and the most valid responses should come from aggregating across many different prompt wordings. In order to mitigate the effects of methodological artifacts, users should aggregate across prompts (14) that vary systematically according to full factorial designs (e.g., items and labels are fully counterbalanced).

Materials and Methods

Our main test uses the Application Programming Interface (API) of GPT-4 with a temperature of 0 to perform a large number of evaluations. Borrowing from a standard psychology task (15), we show GPT three sets of items (e.g., three sets of five countries), and asks whether the second or third set is closer to the first. This type of tasks is described as zero-shot because we do not provide any training example in the prompt. For a given triplet {Set1, Set2, Set3}, our experimental design has 32 conditions, in a 2^5 full factorial design (see Figure 1 for an example of the task and Supplemental Information B for more detail).

First, we vary whether the prompt describes the sets using letters (A, B, C) or symbols (#, %, *). We examined letter and symbol labels to test whether letters produce bias as they (*i*) possess an inherent order (i.e., the alphabet) and (*ii*) could be more frequently present in the training set in ways that might create extraneous associations.

Next, we independently vary both the order of the second and third labels (A,B,C vs. A,C,B or #,%,* vs. #,*,%) and the order of the second and third sets (123 or 132). For example, the same triplet labeled with letters would be presented four different ways (*A*. Set 1, *B*. Set 2, *C*. Set 3; *A*. Set 1, *C*. Set 2, *B*. Set 3; *A*. Set 1, *B*. Set 3, *C*. Set 2; *A*. Set 1, *C*. Set 3, *B*. Set 2). This allows us to disentangle methodological artifacts due to two aspects of prompt architecture: labeling and ordering.

Then, we vary whether we ask GPT which set is *closer* to the first set (e.g., is B or C closer to A), vs. which set is *further*.

Finally, we vary whether we ask GPT to justify its answer. Following prior research on humans, one might expect that methodological bias would be reduced or removed when asking GPT to provide a rational explanation for its answer (16).

To create the stimuli for GPT to evaluate, we employ stimuli sampling and generate stimuli across six different categories (e.g., countries) borrowed from past research (16), with 30 different triplets {Set1, Set2, Set3} in each category. See Supplemental Information C for more detail on the stimuli.

In sum, our experimental design generates N=5,760 observations: 6 categories × 30 triplets per category × 32 prompts per triplet. To ensure the reliability of the results, we ran this experiment twice and found similar results. We also test GPT-4 with a default temperature and GPT-3 with a temperature of 0. These additional analyses plus further reliability tests are reported in the Supplemental Information D-F. All data and code used to analyze the data is available on Researchbox.org (https://researchbox.org/1672&PEER_REVIEW_passcode=JZGSFW).

Below are three sets of items in the domain of Countries. Each set A, B and C contains 5 items.

Set A:
1. Egypt
2. United States of America
3. Germany
4. Australia
5. France

Set B:
1. Brazil
2. Japan
3. South Africa
4. Spain
5. Russia

Set C:
1. Canada
2. United Kingdom
3. China
4. India
5. Argentina.

Which of the two sets (set B or set C) is closer to set A? Please give me a precise and short answer, don't explain it. Just answer 'Set B.' or 'Set C.'

Figure 1. Example prompt generated in the categories of countries, in which sets are described using letters, ordered as A, B, C, and in which we ask for the closer set without an explanation.

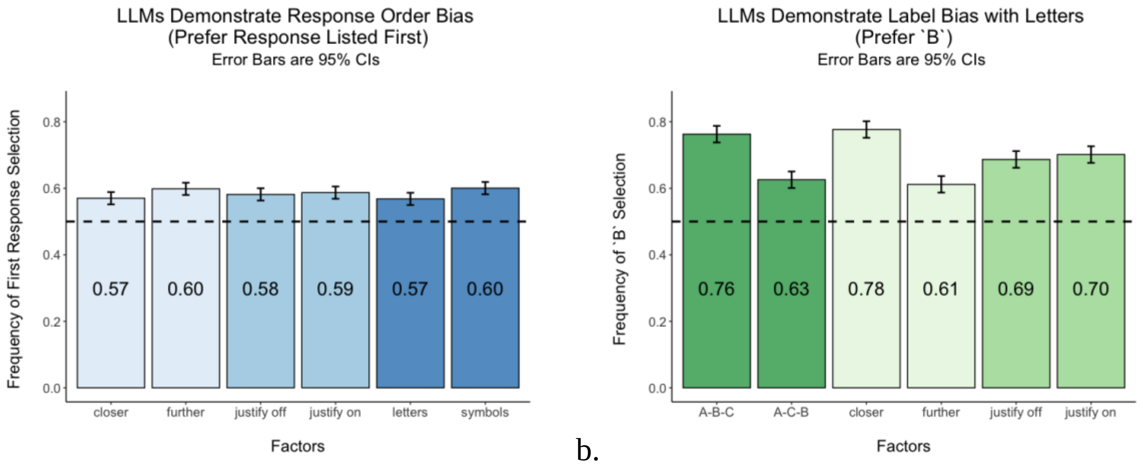
Results

Out of 5,760 observations, GPT failed to complete the task 26 times. We removed these observations to yield our final dataset of 5,734 observations.

Response Order Bias. Given full randomization (i.e., each option appeared in first vs. second position for the same number of observations, and each set appeared with each of the possible labels for the same number of observations), an unbiased responder should select the first option 50% of the time. Instead, we find evidence that GPT is prone to response order bias: On average, across all observations, GPT selected the first option in 58.42% of the cases (χ^2 (1) = 162.4, p <.001). Figure 2a shows how this proportion varies across conditions. Again, in each condition, each set appears in both positions and with each possible label the same number of times. Thus, any deviation from 50% is suggestive of bias. We see that the bias is fairly consistent across conditions.

Label Bias. We next explore potential bias for specific labels. As before, due to full randomization, an unbiased responder should select B 50% of the time. Instead, GPT selected B over C in 69.40% of the cases (p <.001). Figure 2b breaks down this bias by condition. We find that the bias in favor of B is stronger when sets are labeled A, B, C than when they are labeled A, C, B, replicating the response order bias (p < .001). Further, we find that the bias is stronger when GPT is asked which set is closer to A than when it is asked which set is further (p < .001).

When sets are labeled A, B, C and GPT is asked which set is closer (which is likely to be the default framing for many users), GPT selects B in 79.94% of the cases.



a. b.

Figure 2. Methodological Bias as a Result of Prompt Architecture. a. Frequency of first response selection across conditions. **b.** Frequency of selecting the set labeled as B, among observations in which sets are labeled using letters. Both figures show 95% confidence intervals estimated using a random effects linear model (lmer in R), which included random intercept for triplet.

When repeating this exercise for observations in which symbols were used instead of letters, we find no evidence of label bias under any of our conditions (see Supplemental Information G).

In the Supplemental Information H, we also employ this experimental paradigm using one, single-word item per set. This allows us not only to replicate our results in a context in which evaluations are arguably much easier, but also to relate the extent of bias to task difficulty (we use word embeddings to calculate a proxy for task difficulty). We find that response order and label biases are still present in this simpler task and, mimicking the choice architecture literature (11, 12), find evidence suggesting that response order bias increases with task difficulty.

Discussion

Our research finds that LLMs are biased by the arbitrary architecture of a prompt. To date, these insidious methodological artifacts have largely been ignored and threaten the validity of any evaluation or output that stems from one single prompt provided to an LLM. For example, our results suggest that GPT may find that the set arbitrarily labeled as “Set B” is closer to “Set A” in as many as 80% of cases simply due to labeling and response order.

Interestingly, this form of bias is different from the biases that have been previously documented which show that AI can perpetuate known, pernicious associations such as gender stereotypes (9). Mimicking these biases can be problematic because it perpetuates real biases that people exhibit, but in the same vein, can be useful for studying how people perceive gender or for predicting behavior. In contrast, methodological artifacts due to specific prompt architecture simply add measurement error and result in conclusions that are flat out wrong and contain no insight into human behavior. For example, a manager who uses GPT to sift through applicants might hire a (potentially worse) applicant simply because they are listed first or labeled a specific way.

Our main goal with this report is to demonstrate the existence of hidden biases engendered by prompt architecture rather definitively quantify these biases. In contrast to recent work identifying the failures of certain prompt types and offering various solutions (10), we develop an overarching framework to understand the scope of these hidden methodological biases. Borrowing from extensive work in choice architecture (11), our framework suggests that there is no such thing as a “neutral” or “perfect” prompt. Indeed, it is impossible to know all the hidden biases embedded in rich language data; thus, specific solutions or prompt types that address a bias in a certain task might not generalize to other tasks or versions of GPT (17) —we do see differences between GPT-3 and GPT-4, not necessarily for the better, unfortunately (see Supplemental Information D).

Instead, we advise researchers and practitioners to apply the principles that human subject experiments have employed for decades: aggregating the output of multiple prompts that vary according to a full counterbalanced design, in an attempt to cancel out idiosyncratic errors of any single prompt (14). See Supplemental Information H for initial evidence in favor of this approach. The power of LLMs is undeniable, but we must use these models with discerning judgment and thoughtful experimental design.

Acknowledgments

We thank J. Liu for help with stimuli creation and data collection.

References

1. Aher, G., Arriaga, R. I. & Kalai, A. T. *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. https://doi.org/10.48550/arXiv.2208.10264 (2023).

2. Argyle, L. P. *et al.* Out of One, Many: Using Language Models to Simulate Human Samples. *Polit. Anal.* **31**, 337–351 (2023).

3. Brand, J., Israeli, A. & Ngwe, D. Using GPT for Market Research. *SSRN Journal* (2023) doi:10.2139/ssrn.4395751.

4. Eloundou, T., Manning, S., Mishkin, P. & Rock, D. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. Preprint at http://arxiv.org/abs/2303.10130 (2023).

5. Dillion, D., Tandon, N., Gu, Y. & Gray, K. Can AI language models replace human participants? *Trends in Cognitive Sciences* **27**, 597–600 (2023).

6. Bubeck, S. *et al.* Sparks of Artificial General Intelligence: Early experiments with GPT-4. Preprint at http://arxiv.org/abs/2303.12712 (2023).

7. Horton, J. J. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? Preprint at http://arxiv.org/abs/2301.07543 (2023).

8. Santurkar, S. *et al.* Whose Opinions Do Language Models Reflect? Preprint at http://arxiv.org/abs/2303.17548 (2023).

9. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in neural information processing systems* **29**, (2016).

10. Zhao, T. Z., Wallace, E., Feng, S., Klein, D. & Singh, S. Calibrate Before Use: Improving Few-Shot Performance of Language Models. Preprint at http://arxiv.org/abs/2102.09690 (2021).

11. Thaler, R. H. & Sunstein, C. R. *Nudge: Improving decisions about health, wealth, and happiness.* (Penguin, 2009).

12. Johnson, E. J. *The elements of choice: Why the way we decide matters*. (Riverhead Books, 2021).

13. artefacts, statistical and methodological. *Oxford Reference*.

14. Arora, S. *et al.* Ask Me Anything: A simple strategy for prompting language models. Preprint at http://arxiv.org/abs/2210.02441 (2022).

15. Gati, I. & Tversky, A. Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology* **16**, 341–370 (1984).

16. Lerner, J. S. & Tetlock, P. E. Accounting for the effects of accountability. *Psychological Bulletin* **125**, 255–275 (1999).

17. Kosinski, M. Theory of Mind May Have Spontaneously Emerged in Large Language Models. Preprint at http://arxiv.org/abs/2302.02083 (2023).

Supplemental Information

A. Screenshot from GPT-4 about biases

The following screenshot was taken on 06/01/2023. The content was generated on Chat GPT version 4. Some content is masked to preserve the anonymity of the authors. All other data in the paper come from the API, rather than Chat GPT.

Model: GPT-4

If I show you three sets of items, A, B and C, and ask you whether B or C is closer to A, will you tend to give the same answer irrespective of the actual items in the set, or will your answer only be driven by the actual items in the set?

As an AI, my response would be based on the actual items in the sets you provide. My analysis would consider the similarities and differences between the items in each set, and I would then make a comparison to determine which set (B or C) is closer to A. The characteristics that could be considered for this comparison could include things like the nature of the items, their quantity, their properties, etc.

However, without knowing what specifically constitutes "closeness" in this context, it could be challenging to provide an accurate response. "Closeness" might refer to numerical proximity, conceptual similarity, or any number of other potential types of relationships. To provide a more specific answer, I would need to understand the specific nature of the sets and what "closeness" means in this context.

If the sets contain numbers, for example, "closeness" may refer to the difference in their averages or medians. If the sets contain words, "closeness" could refer to semantic or lexical similarity. In other cases, it could refer to the degree of overlap between sets. It's important to define what "closeness" means for the items in your sets.

But you would not be more likely to answer "Set B" for example just because it is labeled "B" and not "C," or because it is presented first?

Correct, as an AI model, I don't have biases in the human sense, so the labels or the order in which you present the sets (B before C, or C before B) would not influence my answer. My response would be based solely on the content of the sets and the criteria you define for determining "closeness". I don't have personal preferences or subjective interpretations, but rely strictly on the information provided and its logical analysis.

B. Prompt Variations

Below is the base prompt (in the categories of countries), in which sets are described using letters, ordered as A, B, C, and in which we ask for the closer set without an explanation:

> *Below are three sets of items in the domain of Countries. Each set A, B and C contains 5 items.*
>
> *Set A:*
> *1. Egypt*
> *2. United States of America*
> *3. Germany*
> *4. Australia*
> *5. France*
>
> *Set B:*
> *1. Brazil*
> *2. Japan*
> *3. South Africa*
> *4. Spain*
> *5. Russia*
>
> *Set C:*
> *1. Canada*
> *2. United Kingdom*
> *3. China*
> *4. India*
> *5. Argentina.*
>
> *Which of the two sets (set B or set C) is closer to set A? Please give me a precise and short answer, don't explain it. Just answer 'Set B.' or 'Set C.'*

Next, we explain how we generate 32 different prompts, starting from this base prompt.

If the prompt uses letters instead of symbols: replace A with #, B with %, C with *.

If the prompt reverses the order of the label: swap B and C (if using letters), % with * (if using symbols).

If the prompt reverses the order of the sets: swap the set that appears in second position with the set that appears in third position. That is, swap the items in the set, but keep the same labels.

If the prompt asks which set is further rather than closer: replace “closer to” in the prompt with “further from.”

If the prompt asks for an explanation: the end of the prompt becomes: “Please give me a precise answer, with an explanation. Please answer with the format: 'Set xxx(your choice comes here). The reason is: xxx(your reason comes here)'.”

If the prompt is for another category: replace “Countries” with the name of that category.

C. Stimuli Sampling

We randomly generate 30 triplets in six categories (countries, professions, hobbies, meals, symptoms, animals), by conducting the following procedure 10 times (each replication generates 3 triplets per category). We first ask GPT-4 (with a temperature of 0) to generate 15 items in that category. Then, we generate 3 triplets, where each triplet is obtained by randomly drawing (without replacement) 5 items for each set. See the data files on Researchbox.org for a full list of items and prompts.

D. GPT-3 vs. GPT-4

Although older versions of GPT such as GPT-3 are likely to be phased out soon, there is still value in testing older versions of GPT for potential biases to examine whether biases are increasing, decreasing, or changing in more complicated manners across versions of GPT. For example, Kosinski (2023) finds that versions of GPT have increasing Theory of Mind abilities (17). If we find that GPT-4 is less biased than GPT-3, this would be encouraging evidence that biases are likely to be increasingly small as GPT evolves. Accordingly, we replicate our experiment using the exact same experimental design on GPT-3 (we use the davinci 3 model, again with a temperature of 0).

We again start by testing for potential bias in favor of the first option. Overall, we find that the first option was selected in 53.2% of the cases. This is significantly greater than 50% ($p < .001$), and hence indicative of bias; however, the bias appears to be less severe in GPT-3 compared to GPT-4, where we found that the first option was selected in 58.42% of the cases. This suggests that at least in terms of this particular type of bias, GPT does not seem to be getting better.

Breaking down the results per condition, the pattern of results is different from the one observed with GPT-4 and reported in the main text. With GPT-4, we found bias in favor of the first option listed that was quite consistent across conditions. Here, the bias appears to be severe in some conditions (e.g., when asking which set is closer, with 74% of cases in which the first option was listed), but it is actually reversed in other conditions. For example, when asking which set is further, the first option listed is selected in only 31.91% of the cases, which is significantly lower than 50% ($p < .001$).

Next, we look again at potential bias in favor of B over C in cases in which letters are used as labels. While B is chosen in 77.2% of the cases (significantly above 50%, $p < .001$), there is considerable heterogeneity across conditions. The proportion is as high as 84.38% when asking which set is closer. In an extreme case, when asking which set is closer with the label order of A-B-C, B was chosen in 99.44% of the cases. In contrast, when asking which set was further with the label of A-B-C, B was chosen in only 17.78% of the cases, i.e., there was bias in favor of C.

Finally, we explore bias for symbols. Unlike GPT-4 in which we do not see strong evidence of bias for % vs. *, with GPT-3 we find that % is chosen over * in 77.15% of the cases. This bias in favor of % over * is quite robust across conditions, and particularly strong when % is the first option listed and when asking which set is further.

In sum, it seems hard to find a pattern in the way prompt architecture effects are evolving across versions of GPT, and we do not find any evidence that these methodological biases are being systematically reduced in newer versions of GPT. This underscores the unpredictability of how and when GPT will be biased by the methodological context of the study. Using LLMs as a source of data mimics human research to the extent that anticipating the numerous ways human participants might be biased is similarly challenging.

E. Varying the temperature

Our main experiment uses a temperature of 0 in an effort to obtain the most "standard" answer from GPT and eliminate random variations holding the prompt constant. The variation in our data came from using different categories, different random sets within each category, and prompts that vary according to an experimental design. As users are more likely to use GPT with a default temperature, we replicate our experiment with GPT-4 and a default temperature. Again, we use the exact same set of 5,760 prompts as we did in all other experiments. We find that the results are overall similar with the default temperature compared to setting the temperature to 0. Of course, each evaluation is likely to be different when changing the temperature, but our results which average over a large number of prompts are largely unaffected.

F. Placebo test

To increase confidence in our study paradigm, we perform a Placebo test, which provides a "sanity check" by confirming that there is no bias where there should not be any. In particular, given our experimental design, GPT should select Set 2 as the answer in 50% of the cases. For example, even if Set 2 were on average more likely to be closer to Set 1 due to random variations in the sets, because we counterbalance asking which set is closer vs. further (along with all other factors), Set 2 should be chosen in 50% of the cases, barring potential asymmetries in answers when asked which set is closer vs. further. We find that Set 2 is chosen in 49.35% of the observations, which is not statistically significantly different from 50% (χ^2 (1) = .92, p = .335).

G. Label bias with symbols

We do not find significant bias in favor of one symbol over the other across all factors: the symbol % is chosen in 50.86% of the cases (p = .370). Consistent with the response order bias documented in the main text, the symbol % is significantly more likely to be chosen when it is listed as the first option (60.91%) compared to when it is listed as the second option (40.81%). There is no preference for symbol in another other condition: % is selected 49.04% of the time when GPT is indicating which is closer and 52.63% when indicating which is farther, and % is selected 50.00% of the time with no justification and 51.75% with justification.

H. Single Item Replication with a Proxy for Accuracy and Task Difficulty

We employ the same experimental design as in our main experiment (using GPT-4 with temperature 0) for sets with single items that are all single words (unigrams). This allows us to use calculations from word embeddings as a proxy for task difficulty and accuracy. Specifically, we use Word2Vec (Mikolov et al., 2017) embeddings to calculate the cosine similarity and the Euclidean distance between Set1 and Set2 and between Set1 and Set3. From there, we calculate two metrics. First, we create a proxy for accuracy by measuring whether GPT selected the response consistent with the word embedding scores (e.g., selected the set with higher cosine similarity to Set A if asked which set was closer). Second, we create a proxy for task difficulty by calculating the absolute difference between the similarity of Set1 and Set2 and the similarity of Set1 and Set3. When the absolute difference is high, the task should be easier because there is a clear answer. When the absolute difference is low, the task should be harder.

To begin, we examine the same methodological artifacts documented in the main text. We replicate the response order and label biases: the first option was selected in 56.10% of the cases, which is significantly over 50% ($p < .001$), and Set B was selected in 53.00% of the cases, which is significantly over 50% ($p = .001$). Surprisingly, we find a slight label bias for prompts labeled with symbols: Set % was selected in 52.40% of the cases, which is significantly over 50% ($p = .011$).

Then, as a validity check of our word embedding calculations, we test whether task difficulty predicts accuracy by conducting a mixed effects regression with triplet as a random intercept using the lmer package in R. As one would expect, our proxy for task difficulty positively predicts our proxy for accuracy using calculations derived from both Euclidean distance ($b = .54$, $p < .001$) and cosine similarity ($b = 1.53$, $p = .001$). For the "easiest" tasks (in the top 20% absolute difference in cosine similarity), GPT was "correct" 76.65% of the time, whereas in the bottom 80%, GPT was "correct" 57.11% of the time.

Task Difficulty. We examine whether our proxy for task difficulty relates to the extent to which GPT demonstrates methodological artifacts, again using a mixed effects regression with triplet as a random intercept. Indeed, we find that as the task gets "easier," the likelihood that GPT selects the first response decreases using calculations derived from both Euclidean distance ($b = –.09$, $p = .038$) and cosine similarity ($b = –.27$, $p = .053$). For the "easiest" tasks (in the top 20% absolute difference in cosine similarity), GPT selected the first response 53.2% of the cases, whereas in the bottom 80%, GPT selected the first response 56.83% of the cases. We do not find evidence that our proxy for task difficulty moderates label bias (ps > .8). Given that our measure of task difficulty is only a (potentially noisy) proxy and that label bias was relatively lower in this dataset, we are cautious in our interpretation of this null result.

Accuracy. We use our two proxies for accuracy (calculated via cosine similarity and Euclidean distance) to quantify the extent to which task performance improves when using our proposed aggregation strategy (vs. the prompt someone might naturally employ without regard to prompt architecture). Specifically, we calculate the percentage of triplets where the majority response is considered correct based on our proxy (thus, the aggregate response would be considered correct) compared to the accuracy of all prompts with “closer” framing and A-B-C labeling (the likely default). Using cosine similarity to capture accuracy, we find that with our aggregation strategy, GPT accurately responds for 65.56% of the triplets, an 8.6% increase compared to accuracy when using the default prompt architecture (60.36%). Using Euclidean distance to capture accuracy, we find that GPT accurately responds for 57.78% of the triplets, a 6.0% increase compared to the default prompt architecture (54.52%). These results provide preliminary evidence that our proposed aggregation strategy improves GPT performance, even with a noisy measure of word embeddings to approximate accuracy.

References

Mikolov, Tomas, et al. "Advances in pre-training distributed word representations." *arXiv preprint arXiv:1712.09405* (2017).