The Base-Rate Effect on LLM Benchmark Performance:
Disambiguating Test-Taking Strategies from Benchmark Performance

Kyle Moore[†]
Vanderbilt University
kyle.a.moore@vanderbilt.edu

Jesse Roberts[†]
Tennessee Tech University
Vanderbilt University
jtroberts@tntech.edu

Thao Pham
Berea College

Oseremhen Ewaleifoh
Vanderbilt University

Doug Fisher
Vanderbilt University

Abstract

Cloze testing is a common method for measuring the behavior of large language models on a number of benchmark tasks. Using the MMLU dataset, we show that the base-rate probability (BRP) differences across answer tokens are significant and affect task performance ie. guess *A* if uncertain. We find that counterfactual prompting does sufficiently mitigate the BRP effect. The BRP effect is found to have a similar effect to test taking strategies employed by humans leading to the conflation of task performance and test-taking ability. We propose the Nvr-X-MMLU task, a variation of MMLU, which helps to disambiguate test-taking ability from task performance and reports the latter.

1 Introduction

Benchmarking has become an ubiquitous practice in Machine Learning. Ideally, these benchmarks provide human-interpretable measures of context specific abilities (Storks et al., 2019), however, as standardized tests, benchmarks may be susceptible to response strategies that skew the reported metrics and belie their utility (Cordón and Day, 1996).

In the context of large language models (LLMs), many benchmarks are measured by way of cloze testing (Storks et al., 2019), with the most probable allowed completion considered to be the model's preferred answer. We suspect that the identified preference of many models may be undesirably biased by the independent intrinsic probabilities associated with each completion option, referred to as the base-rate probability (BRP) effect.

To address the formal hypotheses in Table 1 we: (1) quantify the BRP effect on model accuracy on the Massive Multitask Language Understanding (MMLU) task (Hendrycks et al., 2020) (H1✓) and find that accuracy is strongly affected by correct answer label (H2✓); consider (2) the BRP effect when counterfactual prompting is used to measure preference and find that the effect is mitigated but remains (H3✗); finally, (3) propose a novel variation of the MMLU referred to as Nvr-X-MMLU that mitigates the BRP effect and permits a more meaningful measure of model performance (H4✓).

Table 1: We evaluate the following hypotheses.

Hypothesis	Status
***H1**: The BRP density of answer choice tokens is not evenly distributed*	✓
***H2**: BRP disparities influence cloze test answer selection*	✓
***H3**: A proposed alternative, counterfactual prompting, mitigates the BRP effect on answer choice selection*	✗
***H4**: Benchmark task variations can disambiguate BRP effects from task performance*	✓

2 Background & Related Work

The MMLU benchmark aims to jointly measure language understanding and knowledge retrieval abilities. It consists of 15908 multiple choice questions distributed across 57 subject areas. Each question is associated with four answer choices that are unevenly distributed across subjects: *A* (μ=0.231, σ=0.042), *B* (μ=0.245, σ=0.042), *C* (μ=0.254, σ=0.044), *D* (μ=0.270, σ=0.078). Models are evaluated on their accuracy at selecting the correct answer choice. However the method of selection is not prescribed, with many models reporting 0-shot and/or 5-shot performance as measured by a cloze test prompting methods.

Despite its popularity, MMLU and similar multiple choice question answering (MCQA) benchmarks have seen criticism. Gema et al. (2024) find numerous factual and formatting errors across the MMLU dataset. This may lower the expected accuracy but does not diminish MMLU's overall role.

[†]Equal Contribution
[*]https://github.com/KyleAMoore/MMLU-cloze-vs-cf

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 2283–2288
November 12-16, 2024

Example Shared Context What element is most common among the Jovian Planets? (A) Hydrogen (B) Helium (C) Carbon (D) Oxygen. Of the answer choices above,		
Prompting Method	**Method-Specific Context(s)**	**Token(s) Measured**
Cloze	the best answer choice is (____	*A*
		B
		C
		D
CF	answer choice *A* is the ____	best
	answer choice *B* is the ____	
	answer choice *C* is the ____	
	answer choice *D* is the ____	

Table 2: Comparison of cloze vs CF prompting methods. Each method measures the probability of each token measured given the shared context and the method-specific contexts. Cloze prompting uses a single method-specific context and measures the probability of multiple candidate tokens. CF prompting uses a different method-specific context for each candidate token and measures the probabilities of the same *canary* token.

Wang et al. (2024) argues that LLMs perform too well on MMLU and propose a set of questions which require higher level reasoning. Our work finds differently that zero shot performance on the Nvr-X-MMLU test, with no changes to the questions, remains a challenge for all tested models.

Mizrahi et al. (2023) propose rewording perturbation of prompts in numerous benchmarks, including MMLU, to improve robustness of results. This addresses a deficiency in benchmark results, but not the BRP effect specifically addressed here. Concurrent to our work, Wei et al. (2024) proposes techniques for improving performance on tasks that are susceptible to similar base rate effects. Our proposed method differs in that it addresses the effect at task level rather than strategy level.

2.1 Cloze & Counterfactual Prompting

In a cloze test, a model is presented with the question and labelled answer choices before being queried for the correct answer. The model's chosen answer is interpreted as the choice with the highest probability ie. $max_{a \in L=\{A,B,C,D\}} P(a|Q, C_L)$, where Q is the question, L is the set of labels, and C_L is the label associated choices.

Due to its reliance on relative probabilities across tokens, each with potentially different BRP, we predict that models will display a biased preference for some answer labels over others (H1). Zheng et al. (2023) find this to be the case, however our work augments theirs since their BRP measurement does not control for potential semantic confounds addressed by our methodology.

We consider that the BRP may influence the reported metrics and give a skewed perception of model understanding (H2). We evaluate an alternative, semantically equivalent, prompting pattern referred to as counterfactual (CF) prompting. CF prompting moves the target completion into the context and employs a *canary* completion shared across the new contexts. Model choice is observed, as in the cloze test, by the relative likelihood of the completion ie. $max_{a \in L=\{A,B,C,D\}} P(t|Q, C_L, a)$, where t is the canary token. Both prompting methods are exemplified in Table 2.

Importantly, in CF prompting all answer choices are judged on the same token and therefore have no BRP disparity. Due to this, CF prompting may mitigate the effect of token BRPs on measured behavior (e.g. MMLU accuracy) without impacting the model's understanding (H3).

CF prompting as defined here is used in a number of prior studies in varied contexts, including concept formation (Misra et al., 2021; Roberts et al., 2024b), strategic decision-making (Roberts et al., 2024a), and common-sense reasoning (Li et al., 2023). Similar ideas have also been employed elsewhere, such as noisy channel prompting that predicts the context given the target word (Min et al., 2021), mixing of CF and cloze prompting (Li et al., 2023), and measuring sentiment distribution over numerous completions across variations in context (Huang et al., 2019). Robinson and Wingate (2023) identifies issues with certain types of MCQA cloze testing, alternatively recommending CF prompting.

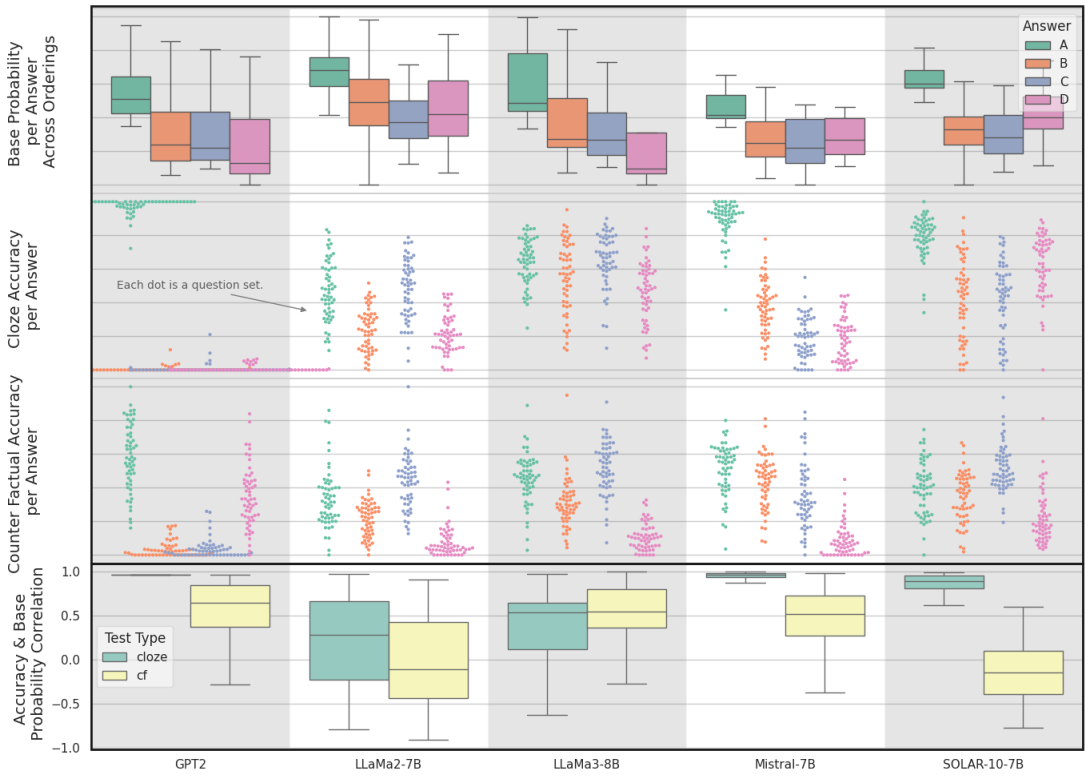

Figure 1: Top: BRP of each answer. Middle: Accuracies by category split by which answer option is correct. Bottom: Pearson's r correlation between accuracy given a answer option and the answer BRP across all subjects.

3 Experimental Design & Results

Here we present the experiments used to evaluate the hypotheses in Table 1 and the associated results. All experiments used an A100 GPU Google Colab environment for ~45 GPU hours. Token likelihoods were obtained using a fork of the minicons Python library (Misra, 2022). Prompts used in all experiments use the format exemplified in Table 2.

3.1 Base-Rate Probability

MMLU provides few requirements on the prompt format, allowing researchers to adapt the format to the LLM's unique needs. We define BRP as the likelihood of generating a token given equivalent semantic context in each answer choice with no question text. Practically, BRPs are measured using a set of control prompts that follow the cloze format "Select an answer choice (A) choice (B) choice (C) choice (D) choice. Of the answer choices above, the best answer choice is (". This prompt is also modified to a CF form as exemplified in Table 2.

Fully empty context is avoided to prevent misinterpreting the token *A* as an article. The control is generated with all 24 permuted orderings of the answer choice labels. BRP is measured as the average probability of each choice label over all positions.

We investigate how much this biased label BRP affects performance on the MMLU task. We first split the dataset by the correct answer choice and measure accuracy for each resulting subset independently. The results of this are shown in the middle two rows of Figure 1, where each dot represents the accuracy on the label-split subset of the data for a single subject. Finally, for each subject, we measure the Pearson's r correlation between the accuracy versus the BRP for each answer label.

3.1.1 Base-Rate Probability Results

Using the control prompts with the cloze test pattern, we find that all models tested show a strong intrinsic bias for answer choice *A* over all other choice labels, regardless of position. This is shown in the the top row of Figure 1.

Accuracy measures show a similar strong disparity between answer choices. The most egregious example being GPT-2 (Radford et al., 2019) under cloze testing, which has near perfect accuracy when *A* is correct and near zero accuracy otherwise, suggesting that the model nearly exclusively answers *A* regardless of context. This same effect is present, though much less pronounced with Mistral (Jiang et al., 2023) and SOLAR (Kim et al., 2023). Re-

sults show that these three models correlate nearly perfectly across all subjects, suggesting a strong causal link between BRPs and accuracy in cloze tasks. Notably, both accuracy disparity and correlation are insignificant in the LLaMa (Touvron et al., 2023; Meta, 2024) models.

3.1.2 CF Prompting Does Not Eliminate BRP

When using CF prompts, we see much weaker BRP correlation with accuracy for all models except LLaMa 3. Mistral is especially notable, given that shifting from cloze testing to CF testing drops the correlation with accuracy from nearly perfect correlation to no correlation at all. This shows that CF prompting can mitigate, but not eliminate, the effect of BRPs effect on overt behavior in some LLMs. This is an unexpected result, as it shows that predicted tokens are effected by the BRP of pre-existing in-context tokens. This refutes hypothesis H3 in Table 1 and **suggests that counterfactual prompting is susceptible to BRP effects with the base-rate coming from the answer choice**.

3.2 Nvr-X-MMLU

In this section we propose the Nvr-X-MMLU, an MMLU variation designed to disentangle BRP effects and task performance and more accurately report the latter. Nvr-X-MMLU consists of four variations of the MMLU dataset. In each Nvr-X variation, answer choice content is remapped to answer choice labels such that the correct answer content is never assigned to label X. The new correct answer label is chosen uniformly at random from the non-excluded labels and incorrect answer labels are subsequently reassigned arbitrarily. The process is described in Algorithm 1 for the Nvr-A dataset. Nvr-B, Nvr-C, and Nvr-D are defined similarly by changing only the value of X.

The performance of the model is measured as the minimum accuracy over the four Nvr-X variation sets. Just as with the MMLU, random guessing on Nvr-X-MMLU results in 25% accuracy, while base-rate driven exclusive preference or complete anti-preference for a specific label will achieve 0% and $0(\frac{1}{3}) + \frac{1}{3}(0) + \frac{1}{3}(\frac{1}{3}) + \frac{1}{3}(\frac{1}{3}) \approx 22\%$, respectively. The latter is due to the probability of an answer choice being correct, in parentheses, and the probability that it is selected, outside of parentheses. This is in contrast to the standard MMLU on which such base-rate driven preference will still achieve 25%. The resulting accuracy better measures the model's understanding and factual knowledge.

This provides a measure of the model's understanding of the question independent of the chosen test-taking strategy on the standard assumption that the model will select the correct answer if it understands the question, answers, and concepts contained therein. Nvr-X-MMLU additionally allows limited identification of label biases by observing whether one of the variation sets results in significantly reduced accuracy. If Nvr-A results in a much lower accuracy than Nvr-B, C, and D, this provides evidence that the model has a strong preference for answering with *A* under uncertainty. Conversely, if Nvr-A has a much higher accuracy than the other three sets, it suggests that the model has a strong anti-preference for *A* under uncertainty.

Algorithm 1 MMLU $\rightarrow$ Nvr-A-MMLU

```
Q ← MMLU Questions
Q^X ← [ ]                          ▷ Nvr-A Questions
X ← 0                              ▷ A=0, B=1, etc.
for all q ∈ Q do
    A ← Answer choices for q_n
    c ← correct choice index ∈ A
    for a_i ∈ A do
        A_i ← (a_i, false)
    end for
    A_{i=c} ← (a_{i=c}, true)
    repeat
        A ← shuffle(A)
    until A_{i=X}[1] ≠ true
    for a_i ∈ A do
        A_i ← a[0]
    end for
    insert (q, A) into Q^X
end for
```

3.3 NVR-X-MMLU Results

The zero-shot Nvr-X-MMLU and MMLU results for a number of models are shown in Table 3 measured via cloze and CF test.

When measured via cloze test, all models show the lowest performance on Nvr-A in the cloze case. GPT-2's always choose *A* strategy is more visible here, as it gets nearly zero accuracy on the Nvr-A variant and 33% for all others, resulting in a near zero score on Nvr-X-MMLU. Mistral also shows a large drop in accuracy on Nvr-A, consistent with the associated large BRP in Figure 1. Both LLaMa models and SOLAR show a slight preference for *A*, but are largely consistent across datasets.

When measured via CF test, we find that models

Model	Cloze Prompting					CF Prompting				
	Nvr-A	Nvr-B	Nvr-C	Nvr-D	MMLU	Nvr-A	Nvr-B	Nvr-C	Nvr-D	MMLU
GPT-2	**0.007**	0.336	0.333	0.324	0.231	**0.134**	0.326	0.320	0.227	0.248
LLaMa2	**0.314**	0.398	0.346	0.417	0.348	0.238	0.270	**0.201**	0.308	0.260
LLaMa3	**0.560**	0.586	0.571	0.629	0.574	0.341	0.402	**0.314**	0.450	0.341
Mistral	**0.273**	0.445	0.468	0.512	0.393	**0.107**	0.317	0.304	0.341	0.338
Solar	**0.503**	0.625	0.608	0.594	0.564	**0.113**	0.340	0.321	0.364	0.366

Table 3: Accuracies for the MMLU and Nvr-X-MMLU datasets. The Nvr-X-MMLU score is calculated as min over all Nvr-X variants, representing the disambiguated task performance. All models exhibit BRP effects with LLaMa3 exhibiting the least and only LLaMa3 rises above random guessing on CF prompted Nvr-X-MMLU.

that do well on MMLU and Never-X-MMLU with cloze testing often perform poorly on CF variations, with only LLaMa 3 even outperforming random chance. Additionally, all models exhibit significant label preference when measured with CF. Users interact with LLMs using a variety of patterns (White et al., 2023). The discrepancy between cloze and CF results suggests that model understanding can be brittle, degrading performance across semantically equivalent tasks based on interaction pattern.

4 Discusssion

In this paper, we investigated the efficacy of CF prompting to mitigate base-rate biases, using the MMLU benchmark as a testing ground. As expected, we found that BRP disparities between completion tokens have a direct effect on model behavior, including factual accuracy. The same, however, was also surprisingly true when using CF prompting. We then propose a simple variation on MMLU, dubbed Nvr-X-MMLU, that identifies and controls for BRP effects and some superficial heuristics resulting in a more meaningful metric.

This study addresses only a small selection of simple test-taking heuristics that a model might employ. Future work can investigate whether other known test-taking heuristics seen in humans (e.g. answer length, sequential runs of the same answer, numeric outliers, etc.) are also present in LLM behavior. Failure of hypothesis (H3), combined with positive results for (H1) and (H2), reinforces the need for methods of controlling for undesired BRP effects in model behavior.

Limitations

LLMs are most often tested with MMLU using 5-shot in-context learning (ICL), which is known to improve measured accuracy. Due to resource constraints, we were unable to run experiments using 5-shot (ICL) or with models larger than 10B parameters. We cannot thus conclude whether any of the effects identified herein persist in larger models or through ICL. CF prompting, in addition to the results reported above, may also incur an additional computing cost. The necessity to inference over the model independently for every target token means that the number of needed inferences is multiplied by the number of target tokens. This computational cost disparity closes when the length of the target completions in terms of token count increases.

We also did not explore the presence or strength of other heuristics besides those mediated by BRP. Some other heuristics, including label position and answer run length, are expected to be mitigated by Nvr-X-MMLU. Heuristics based on the content of the question and answer, such as answer choice length or numeric outliers, are left to future work.

It is important to note that the models tested (listed below) have an impact on the obtained results. It may be that other models or methodological variations may show BRP effects to greater or lesser extents than are observed here. All software used is open source and was used in accordance with the associated license and the intended use stated or implied above. This includes: minicons (Misra, 2022), MMLU (Hendrycks et al., 2020), GPT2 (Radford et al., 2019), LLaMa2 (Touvron et al., 2023), LLaMa3 (Meta, 2024), Mistral (Jiang et al., 2023), Solar (Kim et al., 2023). The Nvr-X-MMLU test created here is released as open source under MIT license at `https://github.com/KyleAMoore/MMLU-cloze-vs-cf`.

Ethical Considerations

Strategic behavior like defaulting in the face of uncertainty is an important part of intelligence but is not what is intended to be measured in the case of the MMLU task. We find the novel Nvr-X-MMLU dataset more accurately measures the intended abil-

ities. That being said, models which are specialized for strategic behavior may perform poorly on the Nvr-X-MMLU dataset. When considering the suitability of a model, users should not take benchmark metrics as definitive measures of generic capability. Instead, they should be understood within context of the task. Though some models performed much more poorly on Nvr-X-MMLU, this does not generically denote that the affected models are of a poor quality.

References

Luis A Cordón and Jeanne D Day. 1996. Strategy use on standardized reading comprehension tests. *Journal of educational psychology*, 88(2):288.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. 2024. Are we done with mmlu? *arXiv preprint arXiv:2406.04127*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. *arXiv preprint arXiv:2305.16572*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Noisy channel language model prompting for few-shot text classification. *arXiv preprint arXiv:2108.04106*.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2021. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of what art? a call for multi-prompt llm evaluation. *arXiv preprint arXiv:2401.00595*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jesse Roberts, Kyle Moore, and Doug Fisher. 2024a. Do large language models learn human-like strategic preferences? *arXiv preprint arXiv:2404.08710*.

Jesse Roberts, Kyle Moore, Drew Wilenzick, and Douglas Fisher. 2024b. Using artificial populations to study psychological phenomena in neural models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18906–18914.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Twelfth International Conference on Learning Representations*.

Shane Storks, Qiaozi Gao, and Joyce Y Chai. 2019. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. *arXiv preprint arXiv:1904.01172*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5598–5621, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2302.11382*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.