None of the Others: a General Technique to Distinguish Reasoning from Memorization in Multiple-Choice LLM Evaluation Benchmarks

Anonymous ACL submission

Abstract

In LLM evaluations, a common strategy to probe cognitive abilities—beyond simple recall or memorization—involves introducing variations to multiple-choice questions, often by altering numbers in math tasks.In contrast, we propose a general variation method that fully dissociates the correct answer from any previously seen tokens or concepts, encouraging reasoning over memorization. Using this method, we evaluate state-of-the-art proprietary and open-source LLMs on two datasets in English and Spanish: the public *MMLU* benchmark and the private *[anonymous dataset name]*. All models show substantial accuracy drops under our variation, averaging 56% on MMLU and 51% on [anonymous dataset], with losses ranging from 10% to 93%. Notably, the most accurate model (OpenAI-o3-mini) is not the most robust (DeepSeek-R1-70B), suggesting that top performers in standard benchmarks may lack stronger reasoning abilities. We also observe larger drops on public datasets and original-language questions (vs. manual translations), pointing to contamination and the role of memorization in current LLMs' performance.

1 Introduction

Large Language Models (LLMs) currently achieve remarkable performance across diverse natural language tasks and even rival humans on general knowledge benchmarks. Yet a fundamental question remains: to what extent do these models truly understand and reason, rather than merely recall patterns from training data? This is especially relevant in benchmarks based on multiple-choice questions, one of the most common methods for evaluating LLMs. While models like OpenAI's (OpenAI, 2024a,c) report state-of-the-art results on *reasoning-heavy tasks* (e.g., GPQA diamond (Rein et al., 2024)), doubts persist that their success may still hinge more on memorization than on the kind of flexible reasoning that characterizes general intelligence—a crucial capacity for tasks requiring logical inference.

To assess reasoning robustness, recent studies employ multi-prompt evaluations, introducing small changes to questions or modifying numerical values in math problems. These strategies test models on structurally similar but novel inputs, yet often focus on narrow domains like mathematics (Srivastava et al., 2024; Mirzadeh et al., 2025; Huang et al., 2025) or rely on manually crafted variations, limiting scalability and generality (Wang et al., 2021).

Our main goal is to examine to what extent LLMs answer general multiple-choice questions by retrieving information from previously seen (compressed) content, versus truly acquiring knowledge and understanding the questions. This leads to three research questions: **RQ1 [Reasoning vs. Memorization]**: How do models respond when questions are reformulated to require reasoning rather than recall? **RQ2 [Contamination and translation biases]**: To what extent does prior exposure (i.e., dataset contamination) affect reasoning? And how does translation impact robustness, given that translated questions are less likely to appear verbatim in training data? **RQ3 [Robustness predictors]**: Can performance drops be explained solely by model size and reference accuracy, or do other factors beyond scaling laws affect reasoning?

Our **main contributions** are: (i) we propose a simple, fully automatic method to rewrite multiple-choice questions from any domain so that the correct answer cannot be retrieved from previously seen texts, requiring genuine understanding; (ii) we show that all models suffer significant performance drops under our variation (average loss above 50%), though the magnitude varies across models; (iii) we provide evidence that models rely partly on memorization, as drops are smaller on private, contamination-free datasets and on translated questions, where recall is less effective; (iv)

we show that **high accuracy does not imply robustness**: for example, Claude-3.5-Sonnet excels on original questions but drops up to 52%, while DeepSeek-R1-70B performs worse initially but degrades far less under NOTO, indicating stronger reasoning; and (v) we find that **robustness does not correlate with model size**: the smallest drop comes from a medium-sized model (DeepSeek-R1-70B), and overall, the most robust models are the latest ones, especially those optimized for reasoning, such as o3-mini, GPT-4o, and DeepSeek.

2 Related work

Recent advances in LLMs have sparked debate over whether their apparent reasoning stems from genuine understanding or mere memorization. This section reviews prior work on evaluating LLM reasoning, including general capabilities, contamination issues, benchmarking practices, robustness testing, and variation methods.

2.1 Benchmarking approaches in LLMs

LLMs are commonly evaluated using question-answer datasets—often in multiple-choice format—or through *LLM arenas*, where users pose questions and compare responses across models (Chiang et al., 2024). Benchmarks span a wide range of tasks, from commonsense reasoning to code generation, with exam-style evaluations becoming increasingly prominent (e.g., MMLU (Hendrycks et al., 2021a), GSM-8k (Cobbe et al., 2021), AGIEval (Zhong et al., 2024), and GPQA (Rein et al., 2024)). However, these benchmarks largely prioritize overall accuracy, which may not directly reflect reasoning capabilities or generalization beyond training data—though some recent efforts aim to fill this gap by explicitly targeting reasoning.

2.2 LLMs and *emergent reasoning capabilities*

While benchmarks aim to assess reasoning, their results often conflate genuine inference with pattern matching. Models like GPT-4 and Claude-3 have been shown to exhibit *emergent capabilities*—behaviours that scale with model size and appear to mimic reasoning—yet many argue these are grounded in memorized patterns and statistical associations, particularly on familiar tasks. This limits their ability to generalize to out-of-distribution problems that require more robust reasoning. Smeaton (2024) suggests that such abilities emerge not solely from scaling, but also from novel training techniques that enable phenomena like grokking. This underscores two ongoing challenges: understanding the internal mechanisms of LLMs, and designing evaluations that reliably measure reasoning ability.

2.3 Data contamination and out-of-distribution generalization

A key limitation in evaluating LLM reasoning is data contamination, which can inflate performance by allowing models to rely on memorized content. Genuine reasoning can only be assessed on truly unseen inputs, making out-of-distribution (OOD) generalization a central challenge (Yang et al., 2023a). Razeghi et al. (2022) argue that evaluations neglecting pretraining exposure are difficult to interpret, calling for a reevaluation of current benchmarking practices. To detect contamination, researchers use heuristics such as checking dataset release dates, conducting web searches, or prompting models to reveal memorized content (Jiang et al., 2024c; Dong et al., 2024; Golchin and Surdeanu, 2023; Sainz et al., 2023; Yang et al., 2023b; Samuel et al., 2025). However, these techniques remain limited due to indirect data leakage and frequent model updates (Ahuja et al., 2023; Balloccu et al., 2024). Alternative strategies include searching through known training corpora or using adversarial setups to test robustness. One such strategy is to introduce small variations—e.g., synonyms, reordering, or typos—to assess whether models are relying on memorization or actual understanding. Yet automating these perturbations without changing the question's meaning remains difficult (Wang et al., 2021).

2.4 Content variation methods in reasoning evaluations

Content variation is a common strategy for evaluating reasoning and detecting contamination, particularly in mathematical tasks due to their structured nature. Srivastava et al. (2024) propose "functional variants" of the MATH dataset (Hendrycks et al., 2021b), defining the *reasoning gap* as the accuracy drop between static and functional variants. Similarly, Mirzadeh et al. (2025) insert irrelevant details into GSM-8k (Cobbe et al., 2021), causing larger drops than simple numeric changes, while Hong et al. (2025) apply semi-automatic perturbations to math and coding tasks, revealing low robustness to minor edits.

Other studies target compositional reasoning.

Hosseini et al. (2024) evaluate performance on multi-step math word problems with interdependent sub-tasks, finding significant gaps, especially in smaller or math-specialized models. Zhu et al. (2024) show that typos degrade math performance, and synonym substitutions impact sentiment classification.

Beyond math, content variation has been used to test generalization across less contaminated domains. Wu et al. (2024) generate counterfactual variants in coding and chess; Lewis and Mitchell (2024, 2025) focus on analogical reasoning; and Yan et al. (2024) assess logical inference. In line with our goals, Nezhurina et al. (2024) design a simple commonsense reasoning task where even minor modifications lead to major performance fluctuations and strong overconfidence in incorrect answers. Similarly, (Elhady et al., 2025) replaces a random option with "None of the above," but since the correct answer often remains, the link with the question is not fully broken, leading to smaller drops than in our stricter setup.

2.5 Reasoning and robustness evaluations

Standard accuracy metrics often miss the nuances of reasoning and robustness in LLMs, prompting calls for more targeted evaluations. Some researchers propose reclassifying advanced models like o1 (strawberry) (OpenAI, 2024c) as *Large Reasoning Models* (LRMs), emphasizing the need for reasoning-specific benchmarks (Valmeekam et al., 2024). Robustness—understood as the ability to handle unfamiliar or unexpected inputs—is crucial for real-world reliability (Wang et al., 2024, 2022). Yet models often underperform in these settings: McCoy et al. (2024) show that LLMs struggle with unseen tasks, and Razeghi et al. (2022) find that GPT-based models disproportionately succeed on arithmetic problems involving frequent training numbers.

Several studies explore the mechanisms behind this behaviour. Nikankin et al. (2025) identify neuron-level circuits involved in arithmetic, concluding that LLMs use heuristic pattern matching rather than robust algorithms or pure memorization. Broader limitations have also been observed: Jiang et al. (2024b) report strong dependence on token-level biases and poor logical inference; Asgari et al. (2024) use multi-answer formats and novel metrics to expose shortcut learning; and Dziri et al. (2023) show that models handle simple tasks well but fail to generalize in complex, multi-step problems, often relying on superficial patterns.

Even techniques designed to promote reasoning, such as Chain of Thought (CoT), are not exempt. Prabhakar et al. (2024) characterize CoT as probabilistic, memorization-influenced reasoning, suggesting that LLMs blend shallow generalization with latent recall in ways that limit robust inference.

Taken together, these studies reveal key limitations in current reasoning evaluations—ranging from contamination and shortcut learning to a reliance on surface patterns—and highlight the need for more robust, generalizable benchmarks. This motivates our proposed variation method, designed to isolate reasoning from recall and enable more reliable assessment across domains and languages.

3 None of the others (NOTO) variation

We propose a variation of multiple-choice questions—referred to as NOTO —in which the correct answer is replaced with *"None of the other answers"*. This becomes the new correct choice, as all remaining options are incorrect by design. With this substitution, the correct answer is no longer terminologically or conceptually tied to the question, thereby reducing the effectiveness of memorization, i.e., recalling associations from pretraining data. To succeed, the model must eliminate all other options and infer that "none of the others" is correct.

Notably, while "none of the others" is a common distractor in multiple-choice formats, it is rarely the correct answer. As a result, models relying on shallow heuristics or frequency-based priors may be reluctant to choose it, potentially yielding performance below chance. Although this strategy does not entirely eliminate memorization effects, it introduces a substantially more challenging setting that places greater demands on reasoning.

To ensure compatibility, we filtered out questions that already included options such as "None of the above", "All of the above", or similar constructions, as these would interfere with the intended variation. This was done automatically using regular expressions. In addition, we applied a classifier trained by Elhady et al. (2025) to detect and exclude questions with potential multiple correct answers. This model[1]—based on a fine-tuned BERT architecture—is designed to identify whether a multiple-choice question has a single cor-

[1] https://huggingface.co/ahmedselhady/bert-base-uncased-sba-clf

rect option, helping ensure that our reformulation remains logically valid.

4 Experimental setup

This section outlines the datasets, models, hyperparameters, prompting strategy and evaluation metrics used to assess performance and robustness.

4.1 Datasets

We experiment with two bilingual datasets. The first is the **MMLU** benchmark (Hendrycks et al., 2021a), which includes English questions across 57 tasks ranging from high school to professional and graduate levels, along with a professional manual translation into Spanish (OpenAI, 2024b). After filtering out questions incompatible with the "none of the others" substitution (using regular expressions and the SBA classifier), 10,270 questions remain.

The second is **[anonymous dataset name]**, comprising 1,003 Spanish questions across 11 university-entry-level subjects, with professional translations into English. Unlike MMLU, this dataset has never been publicly released, making contamination effects unlikely. We applied the same filtering procedure, with one exception: since most psychology items included "None of the other answers" as a fourth option, we removed that option unless it was correct—if so, we discarded the question entirely. This yielded a final set of 923 questions.

4.2 Models and prompting configuration

We evaluated 16 instruction-tuned generative models: five proprietary and eleven open-source. Proprietary models were accessed via API—GPT-4o (OpenAI, 2024a), GPT-4-Turbo (Achiam et al., 2023), GPT-3.5-Turbo (Brown et al., 2020), o3-mini (OpenAI, 2025), and Claude-3.5-Sonnet (Anthropic, 2024). Open-source models were run locally using Hugging Face or Ollama library[2]. These include LLaMA-2 (Touvron et al., 2023) and LLaMA-3 (Meta, 2024), Gemma-7B and Gemma-2-27B (Mesnard et al., 2024; Gemma Team, 2024), Mistral-7B (Jiang et al., 2023), Mixtral-8x7B and Mixtral-8x22B (Jiang et al., 2024a), DeepSeek-R1-70B (DeepSeekAI, 2025), and two Spanish-aligned models: Leniachat-Gemma-2B[3] and Salamandra-7B[4].

Only instruction-tuned models were included to ensure consistent behavior in zero-shot settings and avoid discrepancies in prompt adherence or output format. Each model received the question in its original language (English or Spanish), using a standardized three-part prompt:

- **System prompt**
  ES: `Eres un sistema experto en responder preguntas de exámenes.`
  EN: `You are an expert system for answering exam questions.`
- **User prompt**
  ES: `Responde a la siguiente pregunta de la asignatura {}, tan solo con la letra de la respuesta correcta. Pregunta: {}`
  EN: `Answer the following question of the subject {} only with the letter of the correct answer. Question: {}`
- **Assistant prompt**
  ES: `Letra de la respuesta correcta:`
  EN: `Letter of the correct answer:`

Following standard evaluation practices, all models were prompted in the same language as the question (English or Spanish) (Zhang et al., 2023; Achiam et al., 2023). A zero-shot setup was used throughout, as it mirrors realistic usage and enhances reproducibility, particularly benefiting recent models trained for strong zero-shot performance (DeepSeekAI, 2025). The temperature was set to 0 for all models to ensure deterministic outputs, except for o3-mini, which does not allow temperature control. For open models, prompts were adapted to each model's formatting requirements, based on their official model cards. Outputs were post-processed to extract the predicted letter and discard any justifications or additional text.

4.3 Evaluation metrics

We report **Accuracy**, defined as the proportion of correct answers (C) over total responses (N), and complement it with **Cohen's Kappa coefficient** (McHugh, 2012), which accounts for chance-level performance and varying numbers of answer choices across subjects:

$$\text{Kappa} = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}} = \frac{\frac{C}{N} - \frac{1}{M}}{1 - \frac{1}{M}}$$

Here, M is the number of choices, and expected accuracy is that of random guessing (e.g., $1/3$ or $1/4$). Kappa scores normalize correctness so that

[2] https://ollama.com/

[3] https://huggingface.co/LenguajeNaturalAI/leniachat-gemma-2b-v0

[4] https://huggingface.co/BSC-LT/salamandra-7b-instruct

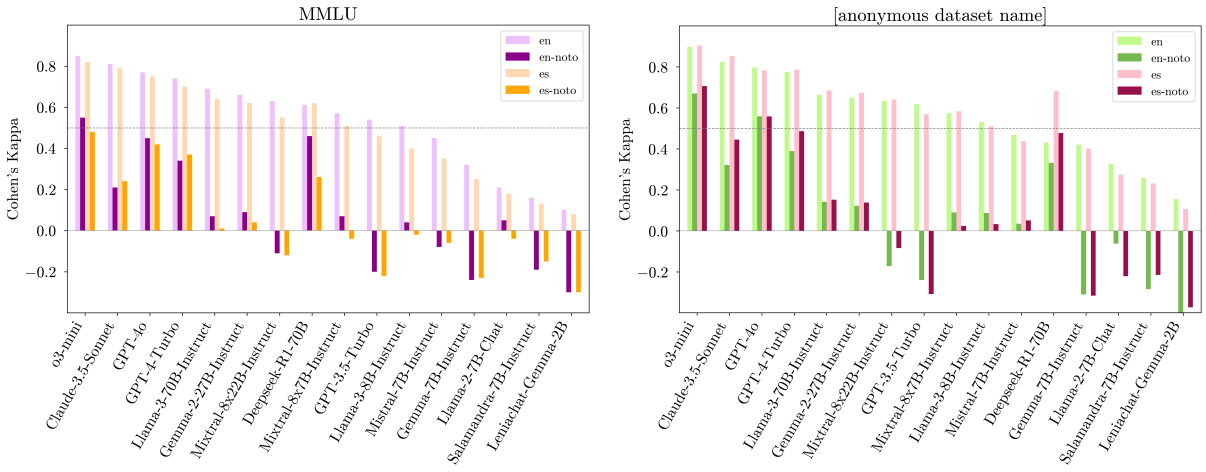

Figure 1: Performance on MMLU and [anonymous dataset name] (original questions and *none of the others* variation). Results per model and language are averaged across all subjects and expressed in terms of **Cohen's Kappa**.

random responses yield a Kappa of zero. Values range from $-1/2$ to 1, with negative values indicating performance worse than chance. Final results are averaged across subjects (not questions) to avoid bias from dataset imbalances.

To quantify performance degradation under our proposed variation, we compute the *drop* as the percentage decrease in Accuracy.[5]

5 Results

Each question from MMLU and [anonymous dataset name] is evaluated under four conditions: original formulation in English and Spanish, and a modified version where the correct answer is replaced with "None of the other answers" in both languages. This setup enables direct comparison between standard multiple-choice performance and the reasoning challenge introduced by the exclusion option.

Figure 1 provides a visual representation of performance in terms of Cohen's Kappa, while Table 1 presents accuracy scores and performance drop between the original questions and their NOTO variations.

5.1 RQ1: Performance vs. Reasoning

To evaluate the impact of the exclusion option on model performance, we analyse both effectiveness (Cohen's Kappa and Accuracy) and robustness (in terms of the performance drop) across datasets in English and Spanish.

Performance with NOTO : Figure 1 depicts results in terms of Cohen's Kappa (where random guessing always gets zero regardless of the number of choices)[6]. All models exhibit a substantial drop in performance under the NOTO variation. In multiple cases, models are even worse than random answers, which suggests that they rely almost purely on memorization, and they probably learnt in the pre-training phase that "None of the others" is statistically less likely than any other option. With the NOTO variation, only o3-mini (the top-performing model overall) exceeds the 0.5 passing threshold in MMLU, for one language (English). Note that, with the use of appropriate question variations, the MMLU dataset is far from being saturated. For [anonymous dataset name], two models pass both in English and Spanish (o3-mini and GPT-4o).

Performance drop: The accuracy drop (Table 1) varies drastically across models (from 10% to 92.5%) in all four datasets, highlighting substantial differences in robustness. Some mid-sized models such as Mixtral-8x22B and GPT-3.5-Turbo suffer particularly steep drops comparable to much smaller models, and scores well below random chance in the NOTO setting. The same applies to somewhat more modern models, such as Llama-3-70B and Gemma-2-27B, which fall drastically to near random-chance performance. Among the top performing models, Claude-3.5-Sonnet expe-

[5] Cohen's Kappa coefficient is not in a ratio scale (the origin is not zero) and therefore percentages cannot be computed directly.

[6] Note that the slightly lower MMLU results compared to previous studies are primarily due to our use of Cohen's Kappa. Additional differences may stem from our zero-shot setup (versus few-shot in other works), prompt formulation, or the quantization of Ollama models.

	MMLU (English)			MMLU (Spanish)			[anonym] (English)			[anonym] (Spanish)		
	base	NOTO	drop %	base	NOTO	drop %	base	NOTO	drop %	base	NOTO	drop %
DeepSeek-R1-70B	0.71	0.60	15.49	0.71	0.45	36.62	0.60	0.54	10.0	0.78	0.63	19.23
OpenAI-o3-mini	**0.89**	**0.67**	24.72	**0.86**	**0.61**	29.07	**0.92**	**0.76**	17.39	**0.93**	**0.79**	15.05
GPT-4o	0.83	0.59	28.92	0.81	0.57	29.63	0.86	0.69	19.77	0.85	0.69	18.82
Llama-2-7B-Chat	0.41	0.29	29.27	0.38	0.22	42.11	0.53	0.26	50.94	0.49	0.15	69.39
GPT-4-Turbo	0.80	0.51	36.25	0.77	0.53	31.17	0.84	0.57	32.14	0.85	0.64	24.71
Claude-3.5-Sonnet	0.86	0.41	52.33	0.84	0.43	48.81	0.88	0.53	39.77	0.90	0.61	32.22
Mixtral-8x7B-Instruct	0.68	0.30	55.88	0.63	0.22	65.08	0.70	0.37	47.14	0.71	0.32	54.93
Llama-3-8B-Instruct	0.64	0.28	56.25	0.55	0.24	56.36	0.67	0.36	46.27	0.66	0.33	50.00
Gemma-2-27B-Instruct	0.75	0.32	57.33	0.71	0.28	60.56	0.76	0.39	48.68	0.77	0.40	48.05
Llama-3-70B-Instruct	0.77	0.30	61.04	0.73	0.26	64.38	0.77	0.40	48.05	0.78	0.41	47.44
Mistral-7B-Instruct	0.59	0.19	67.80	0.51	0.20	60.78	0.63	0.33	47.62	0.61	0.34	44.26
Salamandra-7B-Instruct	0.37	0.11	70.27	0.35	0.14	60.00	0.48	0.11	77.08	0.46	0.16	65.22
Mixtral-8x22B-Instruct	0.72	0.17	76.39	0.66	0.16	75.76	0.75	0.19	74.67	0.75	0.25	66.67
GPT-3.5-Turbo	0.65	0.10	84.62	0.59	0.09	84.75	0.73	0.14	80.82	0.70	0.09	87.14
Gemma-7B-Instruct	0.49	0.07	85.71	0.44	0.08	81.82	0.59	0.09	84.75	0.58	0.09	84.48
Leniachat-Gemma-2B	0.32	0.03	90.63	0.31	0.03	90.32	0.40	0.03	92.50	0.37	0.05	86.49

Table 1: **Accuracy** results on the original and *none of the others* configurations, and percentage decrease between scenarios. Systems are sorted by drop in English MMLU, smaller to largest.

riences the most remarkable drop: despite achieving strong performance in the original setting, its NOTO accuracy falls well below that of its peers (o3-mini, DeepSeek-R1-70B, GPT-4-Turbo, and GPT-4o).

DeepSeek's R1 case is particularly surprising: although the 70B model ranks well below the top performers on the original dataset, it exhibits the smallest accuracy drop in both English datasets, and also the lowest drop overall (only 10% in [anonymous dataset name] in English and 15.49% in English MMLU). This suggests that while DeepSeek-R1 is smaller and with less memory, it has stronger reasoning abilities.

Overall, these results reveal significant differences in how models handle scenarios that demand refined reasoning. DeepSeek-R1-70B and OpenAI-o3-mini have the smallest relative drops, which indicates a stronger, albeit imperfect, ability to validate answer options rather than rely solely on memorization. In contrast, Claude-3.5-Sonnet, despite being a high-performing model in standard conditions, suffers one of the largest drops (52.33% in English MMLU). The most affected models, such as GPT-3.5-Turbo, experience extreme accuracy degradation (over 80% drop), which points to an almost exclusive reliance on approximate matching heuristics.

5.2 RQ2: Contamination and translation biases

To investigate whether the accuracy drop is due to reasoning limitations or reliance on memorized patterns, we compare results from two angles: (1) the effect of dataset contamination, contrasting the public MMLU dataset with the private [anonymous dataset name] dataset, and (2) the effect of translations, contrasting models' performance in the original language versus manually translated versions. These aspects are closely related, as they both influence the extent to which models rely on prior exposure rather than true reasoning. If memorization plays a dominant role, we expect larger drops in public datasets and in original-language versions, as these are more likely to have been seen during pretraining, and for them approximate search may be more effective.

Contamination effects: The mean drop is higher in MMLU (55.8%) than in [anonymous dataset name] (50.88%), consistent with the expectation that MMLU, as a public dataset, is more likely to have been seen during pretraining and leads models to fail more when they are prevented from using that memorisation. In fact, the lowest absolute drop (10% from DeepSeek) is observed on the least likely contaminated dataset, the English [anonymous dataset name] (which is both private and translated from the original questions).

Translation effects: Within MMLU, the average drop is slightly greater in Spanish (57.33%) than in English (55.8%), whereas in [anonymous dataset name], the pattern reverses (51.1% in English vs. 50.9% in Spanish). With the original questions, models perform better in each dataset's original language: all models (16/16) achieve higher

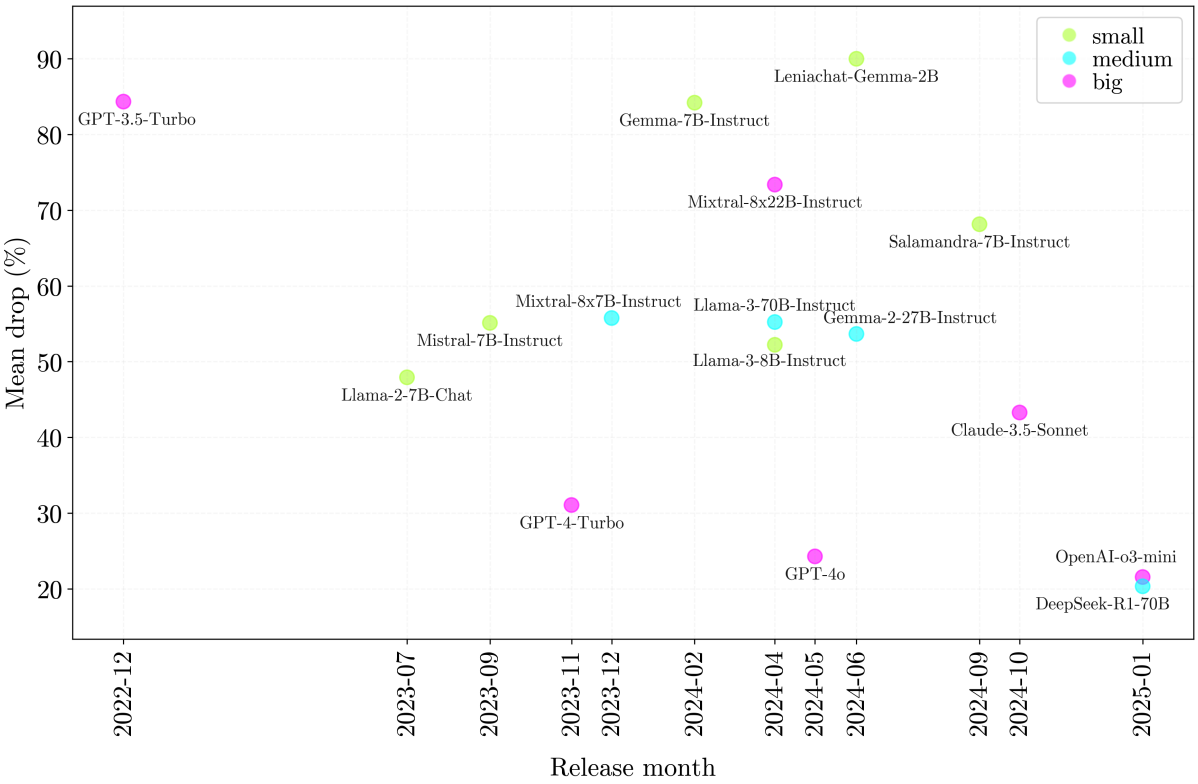

Figure 2: Mean drop across release dates and model sizes.

accuracy in English for MMLU, while in [anonymous dataset name], 8 models perform better in Spanish. This trend still holds in the NOTO scenario: in MMLU, 11 models perform better in English, and in [anonymous dataset name], 11 models now perform better in Spanish. When considering passing thresholds, more models pass in English for MMLU (11 vs. 8), and in Spanish for [anonymous dataset name] (11 vs. 10). This pattern holds in NOTO : 1 vs. 0 in MMLU and 2 vs. 2 in [anonymous dataset name].

These are signs of contamination, since, in other words, (i) models fall more in the public dataset, which is likely more contaminated and (ii) models fall more in the original versions than in the translated versions, with which models are probably less familiar (the Spanish MMLU is newer and less likely to be contaminated, and even if [anonymous dataset name] is private, it is less likely that models have seen the English questions since they are manual translations and have never been released).

Overall, these findings confirm that the NOTO substitution exposes reliance on memorized content, and lead us to the conclusion that models experiencing the highest drops are those most likely answering with their memorization skills, rather than with true reasoning. Results with the NOTO configuration are a better indication of models' true capabilities, show that with a little twikering the datasets are far from being saturated, and reveal comparative differences between the reasoning capabilities of models that are hidden in the evaluation with the original questions. In particular, we have seen that the performance difference between the most recent *reasoning* models (Deepkseek, o3-mini) and other state-of-the-art ones such as Claude-3.5 is much wider than can be measured with the original questions.

5.3 RQ3: Robustness predictors

	correlation	***p*-value**
MMLU (English)	-0.50	0.0480
MMLU (Spanish)	-0.58	0.0182
[anonym] (English)	-0.59	0.0167
[anonym] (Spanish)	-0.77	0.0005

Table 2: Pearson's correlation between accuracy results on the base configuration and the drop.

Here we examine whether base performance predicts robustness under the NOTO variation. As shown in Table 2, the correlation is weak to moderate in most cases, and only becomes strong and

highly significant in our private Spanish dataset ($r = -0.77$, $p = 0.0005$). This suggests that, when contamination is minimal, higher base accuracy may better reflect genuine reasoning ability. However, in more exposed datasets like MMLU, base performance is a poor predictor of robustness.

Figure 2 shows the the mean drop in performance across models, sorted by release date and classified into three groups according to their size: small (less than 10B parameters), medium (10-100B) and large (over 100B). Note that the drop does not correlate well with model size, as there are large models with large drops (GPT-3.5-Turbo, Mixtral-8x22B and Claude-3.5-Sonnet), and the smallest average drop is for a medium-sized model (DeepSeek-R1-70B): size alone is insufficient to ensure robust reasoning. There is a noticeable trend, though, where newer models tend to exhibit smaller drops, with some exceptions. The oldest model, GPT-3.5-Turbo, is a mid performer with the original datasets, but stands out as one of the worst in terms of performance drop. In the period since ChatGPT's debut, the generalisation capabilities of models seems to have improved widely and consistently, and this improvement does not necessarily come with increased model sizes. Finally, the newest proprietary models and DeepSeek-R1 are the ones that show smaller performance drops; this suggests that robustness in reasoning is influenced more by advanced model architectures and training strategies rather than sheer model size.

6 Conclusions

Our results show that the proposed NOTO variation poses a major challenge for LLMs, and provides a useful signal to distinguish answers based on recall/memorization from genuine knowledge and reasoning. While many models perform well when retrieving memorized information, their performance plummets when the correct answer is disconnected from memory associations and they are required to verify and reject each candidate answer. The NOTO variation consistently reveals reasoning gaps, exposing limitations that remain hidden in standard multiple-choice settings (RQ1). Dataset contamination further complicates the evaluation of reasoning: while prior exposure may artificially inflate accuracy in base scenarios, its impact diminishes in NOTO , where models cannot rely on memorized answers. Similarly, models perform better on original (and likely more contaminated) datasets, while translated versions mitigate this effect, reinforcing the role of memorization in standard benchmarks (RQ2).

Unlike accuracy, which scaling laws correlate with model size (Kaplan et al., 2020), we have seen that robustness is not strictly correlated with model size. High-performing models such as Claude-3.5 suffer severe drops, and some mid-sized models (e.g., GPT-3.5-Turbo, Mixtral) degrade to below-random performance. The most robust model in our experimentation, DeepSeek-R1-70B, is mid-sized, suggesting that architectural advancements and training strategies, rather than sheer scale, play a greater role in reasoning robustness (RQ3). Remarkably, the two so-called *reasoning models* in our sample (o3-mini and DeepSeek-R1) are indeed the ones that better resist the NOTO variation.

In short, our experimentation is a direct confirmation that LLMs remain far from true reasoning, but also that progress is being made towards that goal. Our findings emphasize the need for models that can reliably handle question reformulations without relying on surface-level heuristics, and show that classic datasets that appear to be saturated, such as MMLU, may still be useful for LLM evaluation under appropriate transformations.

Limitations

The NOTO variation offers a useful approximation for testing reasoning beyond memorization, but it does not eliminate recall effects entirely. Still, it provides a complementary perspective to standard evaluations, especially on benchmarks that may be saturated. Focusing on multiple-choice questions allows for consistent comparisons, but naturally limits the scope of reasoning assessed. Exploring open-ended or real-world tasks could provide additional insights.

We rely on two datasets—MMLU and a smaller, private bilingual set—which offer valuable contrast but do not cover all domains or task types. Future work will extend this analysis to larger and more diverse benchmarks.

All models were tested in zero-shot settings for consistency and reproducibility. Other prompting strategies, such as few-shot or Chain-of-Thought, may affect outcomes and should be examined.

While we do not include detailed per-model analyses, we observe that some newer and reasoning-oriented models tend to be more robust. This suggests that training choices may play an important

role, though more work is needed to understand these effects fully.

Lastly, while humans also use heuristics and pattern recognition (Lampinen et al., 2024), their reasoning differs in important ways. Comparing human and model responses under NOTO conditions remains an open and promising direction.

Despite these limitations, we hope this method contributes to more nuanced evaluations of LLMs and encourages further exploration of reasoning robustness.

Acknowledgments

[removed for anonymity]

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and Aleman et al. 2023. Gpt-4 technical report.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Saeid Asgari, Aliasghar Khani, and Amir Hosein Khasahmadi. 2024. MMLU-pro+: Evaluating higher-order reasoning and shortcut learning in LLMs. In *Neurips Safe Generative AI Workshop 2024*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *arXiv preprint*. ArXiv:2403.04132 [cs].

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems.

DeepSeekAI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pages 70293–70332. Curran Associates, Inc.

Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. WiCkeD: A simple method to make multiple choice benchmarks more challenging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1183–1192, Vienna, Austria. Association for Computational Linguistics.

Google DeepMind Gemma Team. 2024. Gemma 2: Improving open language models at a practical size.

Shahriar Golchin and Mihai Surdeanu. 2023. Data contamination quiz: A tool to detect and estimate contamination in large language models. *CoRR*, abs/2311.06233.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Pengfei Hong, Navonil Majumder, Deepanway Ghosal, Somak Aditya, Rada Mihalcea, and Soujanya Poria. 2025. Evaluating LLMs' mathematical and coding competency through ontology-guided interventions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22811–22849, Vienna, Austria. Association for Computational Linguistics.

Arian Hosseini, Alessandro Sordoni, Daniel Kenji Toyama, Aaron Courville, and Rishabh Agarwal. 2024. Not all LLM reasoners are created equal. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan Zhang, and Mengdi Wang. 2025. MATH-perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. In *Forty-second International Conference on Machine Learning*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, and Florian others Bressand. 2023. Mistral 7B. *arXiv preprint*. ArXiv:2310.06825 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, and Devendra Singh Chaplot et al. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J Su, Camillo Jose Taylor, and Dan Roth. 2024b. A peek into token bias: Large language models are not yet genuine reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4722–4756, Miami, Florida, USA. Association for Computational Linguistics.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024c. Investigating data contamination for pre-training language models. *ArXiv*, abs/2401.06059.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Andrew K Lampinen, Ishita Dasgupta, Stephanie C Y Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233.

Martha Lewis and Melanie Mitchell. 2024. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *Preprint*, arXiv:2402.08955.

Martha Lewis and Melanie Mitchell. 2025. Evaluating the robustness of analogical reasoning in large language models. *Transactions on Machine Learning Research*.

R McCoy, Shunyu Yao, Dan Friedman, Mathew Hardy, and Thomas Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences of the United States of America*, 121:e2322420121.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, and Morgane Rivière et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.

Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. 2024. Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. *arXiv preprint*. ArXiv:2406.02061 [cs].

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *The Thirteenth International Conference on Learning Representations*.

OpenAI. 2024a. Gpt-4o system card.

OpenAI. 2024b. Multilingual massive multitask language understanding (mmmlu).

OpenAI. 2024c. Openai o1 system card.

OpenAI. 2025. Openai o3-mini.

Akshara Prabhakar, Thomas L. Griffiths, and R. Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3710–3724, Miami, Florida, USA. Association for Computational Linguistics.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Vinay Samuel, Yue Zhou, and Henry Peng Zou. 2025. Towards data contamination detection for modern large language models: Limitations, inconsistencies, and oracle challenges. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5058–5070, Abu Dhabi, UAE. Association for Computational Linguistics.

Alan F. Smeaton. 2024. Understanding foundation models: Are we back in 1924? In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 66–72.

Saurabh Srivastava, Annarose M. B, Anto P V, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. 2024. Functional Benchmarks for Robust Evaluation of Reasoning Performance, and the Reasoning Gap. *arXiv preprint*. ArXiv:2402.19450 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, and Lachaux et al. 2023. LLaMA: Open and Efficient Foundation Language Models.

Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. 2024. Llms still can't plan; can lrms? a preliminary evaluation of openai's o1 on planbench. *Preprint*, arXiv:2409.13373.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Advances in Neural Information Processing Systems*.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. 2024. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *IEEE Data Eng. Bull.*, 47(1):48–62.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Junbing Yan, Chengyu Wang, Jun Huang, and Wei Zhang. 2024. Do large language models understand logic or just mimick context? *Preprint*, arXiv:2402.12091.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023a. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12731–12750, Toronto, Canada. Association for Computational Linguistics.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023b. Rethinking benchmark and contamination for language models with rephrased samples. *CoRR*, abs/2311.04850.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2024. Promptbench: a unified library for evaluation of large language models. *J. Mach. Learn. Res.*, 25(1).