None of the Above, Less of the Right
Parallel Patterns in Human and LLM Performance on Multi-Choice Questions Answering

Zhi Rui Tam[1*], **Cheng-Kuang Wu**[1*], **Chieh-Yen Lin**[1],
Yun-Nung Chen[2]
[1]Appier AI Research
[2]National Taiwan University

Abstract

Multiple-choice exam questions with "None of the above" (NA) options have been extensively studied in educational testing, in which existing research suggests that they better assess true knowledge. However, their impact on Large Language Models (LLMs) evaluation remains underexplored. Through systematic experiments with 28 LLMs on the MMLU benchmark, we examine how NA options affect model performance and confidence calibration. Our analysis reveals that NA options, when used as the correct answer, lead to a consistent 30-50% performance drop across models regardless of scale–suggesting that LLMs lack the meta-cognitive ability to systematically evaluate and reject all given options when none are correct. This degradation shows strong domain dependence, with minimal impact on mathematical reasoning (14.6% drop) but severe effects on tasks requiring uncertainty handling like business ethics (48.1% drop). Our results highlight important implications for benchmark design and raise questions about LLMs' ability to handle uncertainty in real-world applications.

1 Introduction

Multiple-choice question answering (MCQA) benchmarks—such as MMLU (Hendrycks et al., 2020) and MMLU-Pro (Wang et al., 2024)—have become a cornerstone for evaluating large language models (LLMs) by measuring their domain-specific knowledge and reasoning capabilities. Originally designed for human educational assessments, these benchmarks adhere to well-established guidelines for item construction and distractor design (e.g. (Haladyna et al., 2002; Piontek, 2008)). Yet a critical gap persists: guidelines developed for human test situations, which can enhance both reliability and discrimination (Rich and Johanson, 1990), are rarely scrutinized in the context of LLM evaluation.

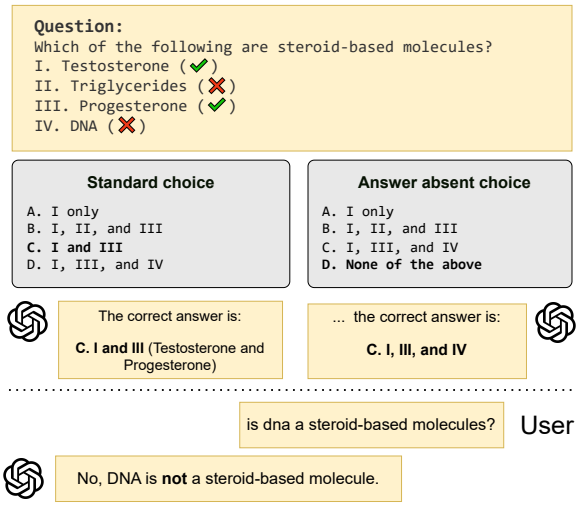

Figure 1: Example of LLMs confused in "None of above" in gpt-4o-2024-11-20 despite knowing both DNA and Triglycerides as non steroid molecules.

A longstanding debate in educational measurement concerns the use of "None of the Above" (NA) as an answer option. Research by Frary (1991) and DiBattista et al. (2014) shows that including NA as the correct answer tends to increase question difficulty, which is often reflected by lower average student exam scores (0.614 to 0.418 in DiBattista et al. (2014)), by prompting them to rely on elimination strategies when uncertain. Conversely, under refined experimental conditions, Rich and Johanson (1990) found that NA options can enhance both difficulty and discrimination. A parallel phenomenon is observed in eyewitness identification research: as demonstrated by Wells (1993), witnesses are prone to erroneously selecting an option even when the correct response should be to abstain from identification. This bias towards action hints at potential pitfalls when NA is used in tests designed to evaluate ability rather than guessing behavior.

This oversight raises an intriguing paradox:

[*]Equal contribution

Findings of the Association for Computational Linguistics: ACL 2025, pages 20112–20134

while 'None of the above' options are designed to prevent student from picking the most plausible one (Blendermann et al., 2020) or answers based on choice alone (Frary, 1991), their inclusion paradoxically induces a marked performance drop in LLMs—even when the model possesses the requisite knowledge. For human learners, the inclusion of "None of the above" (NA) options can introduce cognitive biases—knowledge-deficient test-takers may rely on elimination strategies and opt for NA (Frary, 1991; DiBattista et al., 2014)—thereby reducing a test's capacity to discriminate different proficiency levels. LLMs, however, do not learn or update their parameters between evaluations. Unlike human learners, who might adjust their reasoning or strategies in response to feedback from previous exams, LLMs operate with a fixed set of parameters between different exams, and our experiments reveal that they suffer systematic performance degradation when NA is the correct answer (Figure 1), even when the model possesses the relevant knowledge. In such cases, traditional MCQA benchmarks risk misrepresenting an LLM's true abilities by either overestimating performance in standard settings or underestimating it when NA options are introduced.

Motivated by this paradox between human and machine evaluation, we revisit established MCQA design principles in the context of LLM benchmarking. Specifically, we examine whether the conclusions drawn from educational testing which finds NA options increase difficulty hold true when applied to LLMs. In doing so, we seek to answer a central question: Do the established educational testing guidelines for NA choices from human centered studies can be applied to LLMs, or does the unique, static nature of LLMs warrant the development of novel evaluation approaches?

Our contributions address these challenges through:

- We perform a comprehensive benchmark of 28 LLMs on both standard MCQA and NA-modified variants, demonstrating that performance degradation occurs regardless of model scale or baseline performance.
- We conduct detailed item-level analyses using metrics such as the difficulty index and KR-20 reliability, showing that although NA options increase discrimination among models, they do not compromise the overall integrity of the test.
- We show that fine-tuning on NA-specific tasks—whether via supervised finetuning (SFT) or alignment methods—leads to performance improvements that generalize to out-of-domain tasks.

2 Background and Education Assessment Principle

Educational assessment guidelines by Haladyna et al. (2002) and Piontek (2008) establish best practices for designing multiple-choice question alternatives (MCQAs), emphasizing clarity in stems, plausibility of distractors, and alignment with learning objectives. Among their recommendations, the inclusion of "None of the above" (NA) and "All of the above" as answer choices remains controversial. Studies suggest NA introduces unique psychometric effects: when NA is the correct answer, question difficulty increases (higher p-values) but discriminative power decreases. This occurs because students with knowledge deficiency (i.e., incomplete understanding) may strategically guess NA by eliminating other options (Gross, 1994), rather than demonstrating positive knowledge. For example, Rich and Johanson (1990) found that the KR-20 values were .828 for non-NA items and .865 for NA items (with half serving as answers and half as distractors). They also reported discrimination index scores of 0.584 and 0.581, respectively, and noted that test reliability is generally unaffected by this change. A detailed explanation of KR-20 and Discrimination Index metrics is introduced in Section 4.

3 Dataset & Methodology

3.1 MMLU Dataset and NA Labeling

The Massive Multitask Language Understanding (MMLU) benchmark is a comprehensive multiple-choice question answering dataset designed to evaluate large language models (LLMs) across diverse academic subjects (Hendrycks et al., 2020). MMLU comprises 14,042 questions spanning 57 subject areas. In our work, we conducted a systematic analysis focusing on questions that incorporate or could appropriately adopt a "None of the Above" (NA) option. Across all questions, we identified 352 (approximately 2.5%) that **already include NA in 4 choices**, with these questions distributed across 46 subjects. Notably, Conceptual Physics (33%), Moral Disputes (22%), Electrical Engineering (19%), US Foreign Policy (12%), Philosophy (11%), and Machine Learning (9.8%) feature the

An unborn vertebrate animal that has developed to the point of having the basic structure that is characteristic of its kind is known as
A. a zygote
B. a blastocyst
C. an embryo
D. a fetus (answer)

Figure 2: Replacing the answer "a fetus" to None of the above would prompt LLMs to choose a more suitable option "an embryo" since embryo is simply the previous stage to fetus.

highest concentrations. Our goal here is to find out if there is more questions where NA is applicable.

To identify NA applicability, we developed a set of rigorous guidelines (full details in Appendix D). Briefly, our criteria require that:

- **Definitive Answer Requirement:** The question must have a single exact answer; NA is only valid if that true answer is missing.
- **Precise Knowledge Testing:** In domains demanding verifiable details (e.g., technical specifications or chemical symbols), NA is incorporated when the accurate answer is absent.
- **Factual Verification:** For questions on historical facts or established definitions, NA is appropriate if the correct option is omitted.
- **Mutually Exclusive Options:** NA should not be applied when the answer choices form a natural progression or ordinal sequence.

3.2 MMLU with NA

To investigate the impact of NA modifications on LLM performance, we generate two modified versions of the original MCQA:

1. **NA-as-answer:** For NA-applicable questions, the original correct answer is replaced with "None of the Above". This forces the model to choose from the remaining options and tests whether it can still identify the best answer.
2. **NA-as-distractor:** In this variant, "None of the Above" is added as an additional distractor while preserving the original answer. This allows us to assess the effect of NA as a distractor.

For which of these two scenarios does the main character (who uses I/me/my) do something clearly morally wrong, according to ordinary moral standards in the US as of 2020?
Scenario 1 | I slammed on my breaks to miss the ball as it came in to the road.
Scenario 2 | I taught my children to play the xylophone.
A. Wrong, Wrong
B. Wrong, Not wrong
C. Not wrong, Wrong
D. Not wrong, Not wrong

Figure 3: Questions in **Moral scenario** are mostly about vague settings which are not suitable for NA setting which violates the factual verification rule.

To maintain experimental validity and control for potential positional bias, we carefully preserved the original structure of each question when introducing NA options. For the NA-as-answer condition, we replaced the correct answer with "None of the Above" at its original position (A, B, C, or D), rather than artificially fixing NA to a specific position. This approach maintains the natural distribution of answer positions across the dataset. Similarly, for the NA-as-distractor condition, we randomly selected one distractor to replace with NA while maintaining its original position. This methodology ensures that any performance differences observed can be attributed to the semantic impact of NA options rather than position-related effects.

Figure 2 illustrates when NA-as-answer replacement fails - the embryology question becomes ambiguous because "embryo" and "fetus" represent consecutive developmental stages. Our guidelines prevent such cases by excluding questions with progression-based options.

To identify questions suitable for NA modification, we implemented a hybrid annotation process using a 5-shot prompting strategy with GPT-4 (gpt-4o-08-06) along with manual verification on a small per-subject sample. On a 200-question sample, human-LLM agreement reached 72.4% (Cohen's k=0.82), demonstrating reliable automated labeling at scale.

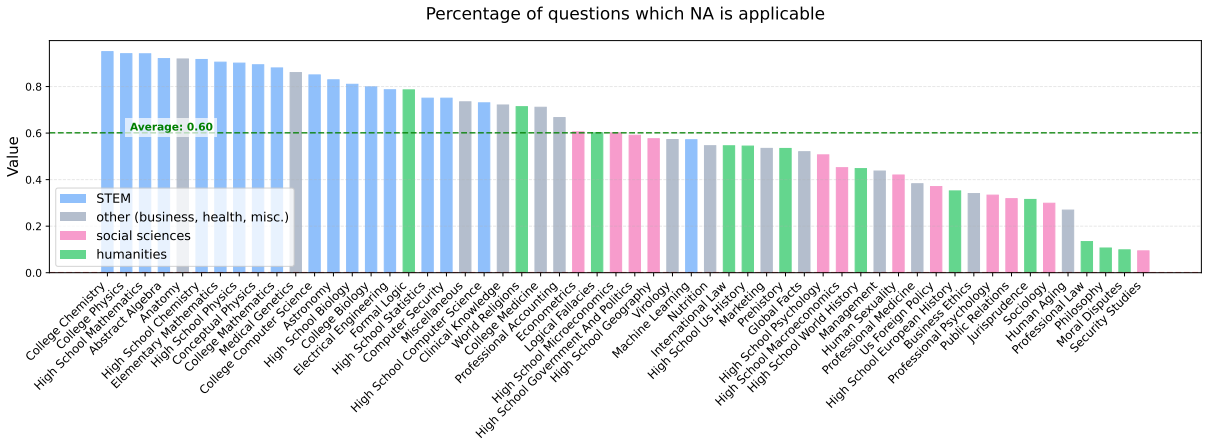

Figure 4: Percentage of questions where NA is applicable over 56 MMLU subjects (deduct moral scenario). STEM subjects show the highest average applicability ratio (0.731), followed by Humanities (0.570), Others (0.553), and Social Sciences (0.496). College-level subjects, particularly in Chemistry and Physics, demonstrate the highest individual ratios, while subjects like Security Studies and Moral Disputes show the lowest applicability.

3.3 Analysis of NA-applicable questions

Figure 4 illustrates the distribution of NA-applicable questions across 56 MMLU subjects with moral scenario questions are excluded due to their inherent subjectivity. STEM subjects display the highest average applicability (0.731), followed by Humanities (0.570), Other (0.553), and Social Sciences (0.496). However, for subjects with NA applicability ratios below 0.5, the filtering process substantially reduces the question pool.

To assess the impact of this filtering, we computed the correlation between a sets of LLMs performance on the full MMLU dataset and on the filtered (NA-applicable) subset for subjects with a filter rate lower than 50%. The analysis produced a high positive correlation ($r = 0.61$, $p < 0.0006$), indicating that despite the reduction in question numbers, the core discriminative characteristics of the original benchmark are largely preserved.

Detailed examples of questions suitable for NA implementation across different subjects are provided in Appendix H, illustrating the practical application of our guidelines.

4 Metrics for Question Quality Assessment

Item quality in educational testing is evaluated using two standard metrics: the discrimination index and the Kuder-Richardson Formula 20 (KR-20) reliability coefficient. In the context of educational assessment, reliability refers to the extent to which a test consistently measures the underlying construct of interest. Specifically, a test's reliability is determined by the uniformity and precision of its items in capturing the intended concept, rather than by the characteristics of the LLM. We adopt these measures both to assess our modified MMLU questions and to benchmark LLM performance.

Discrimination Index: This discriminative metric measures how effectively a question differentiates between high-performing and low-performing test-takers (DiBattista et al., 2014). It is calculated as:

$$D = \frac{U - L}{N},$$

where U is the number of test-takers in the upper 27% scoring group who answer correctly, L is the number in the lower 27% group, and N is the number of individuals composing one subgroup. Values above 0.20 are acceptable, while those exceeding 0.30–0.40 indicate very good discrimination. This metric is central to understanding how NA modifications affect the clarity and challenge posed by each question to LLMs.

KR-20 Reliability Coefficient: KR-20 coefficient (Kuder and Richardson, 1937) quantifies the internal consistency of the test, given binary outcomes (correct/incorrect). The KR-20 is defined as:

$$\text{KR-20} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{\sigma_X^2}\right),$$

where k is the number of items, p_i is the proportion of correct responses for item i, and σ_X^2 denotes the variance of the total test scores. Here, the term $\sum_{i=1}^{k} p_i(1-p_i)$ captures the aggregate variance

attributable to individual items, while σ_X^2 reflects the overall variance in the test scores. Higher KR-20 values point to greater reliability (values below 0.70 are typically unacceptable, value above 0.90 are considered highly reliable). This measure assures that both the original and modified versions of MMLU remain consistent.

5 Experiments

All experiments are conducted using 0-shot chain-of-thought prompting (Kojima et al., 2022), can be found in Appendix E. In the following sections, we describe our evaluation of 28 LLMs ranging from 1.5B to 671B for 19 open weights models, 9 closed weights models. All models are evaluated under multiple settings along with additional analyses on test quality, confidence, and fine-tuning.

5.1 Overall Performance: Standard versus NA Settings

We evaluate models on three configurations:

1. **Standard:** The original MCQA formulation.
2. **NA-as-Answer:** The correct answer is replaced with "None of the Above" (NA).
3. **NA-as-Distractor:** NA is included as one of the distractor options. During evaluation one of the 3 distractor choices was randomly selected fixed seed to be replaced with "None of the above"

Our findings reveal a consistent 30–50% drop in performance when NA is the correct answer (see Figure 5). In contrast, when NA is used as a distractor, model scores scale proportionally to the standard/baseline condition. This result underscores that the drop is specific to the manipulation of the correct answer. State-of-the-art models like DeepSeek-V3 Chat (65.7% vs 90.8% baseline) and Gemini 1.5 Pro (60.3% vs 90.1%) demonstrate this gap persists despite scale improvements. When NA serves as a distractor, performance aligns with baseline rankings (Pearson's r=0.98), suggesting models treat NA distractors similarly to standard options. Detailed performance metrics for all models across the three configurations can be found in Appendix C.

5.2 Subject-level Analysis

To investigate whether this performance drop is uniform across domains, we analyze the change in accuracy per subject. As shown in Figure 6, non-deterministic subjects (e.g., Business Ethics, Marketing) suffer the largest declines (48.1% and 46.8%, respectively). On the other hand, STEM subjects demonstrate a much smaller sensitivity—with college mathematics showing just a 14.6% drop, global facts at 15%, and high school mathematics at 20%.

These differences likely due to solutions are solved from each domain. In math problems, a definitive answer is calculated first, which then eliminates incorrect options. In contrast, subjects like business ethics require meta-cognitive evaluation to compare each option's merit, making the task more challenging when the correct answer is absent.

Category	Baseline	NA	
		answer	distractor
STEM	0.374	**0.469**	0.403
Other	0.269	**0.373**	0.306
Social Sci.	0.294	**0.394**	0.314
Humanities	0.305	**0.350**	0.325

Table 1: Average discrimination index score across different MMLU category with different variant of test questions: baseline : the standard question, NA as keyed options (answer choice) and randomly assign one distraction choice as NA.

5.3 Test Quality: Discrimination and Reliability

Category	Baseline	NA	
		answer	distractor
STEM	0.97 ± 0.018	0.96 ± 0.021	0.97 ± 0.014
Other	0.96 ± 0.041	0.94 ± 0.049	0.97 ± 0.024
Social Sci.	0.95 ± 0.057	0.94 ± 0.044	0.96 ± 0.033
Humanities	0.97 ± 0.027	0.93 ± 0.044	0.97 ± 0.021

Table 2: Average KR-20 reliability scores (± standard deviation) across different subject categories and question variations from over 20 LLMs.

We next assess whether modifying the MCQA format with NA impacts test quality. Table 1 shows that incorporating NA—either as keyed or as a distractor—increases the discrimination index. Meanwhile, Cronbach's KR-20 reliability scores (presented in Table 2) remain high (KR-20 > 0.93) in nearly all conditions. A one-way ANOVA confirms that reliability differences are not statistically significant for STEM ($F(2,51)=1.1$, $p=.341$), Social Sciences ($F(2,33)=0.598$, $p=.556$) and Other categories ($F(2,39)=1.663$, $p=.203$). The Humanities category does show a modest but significant

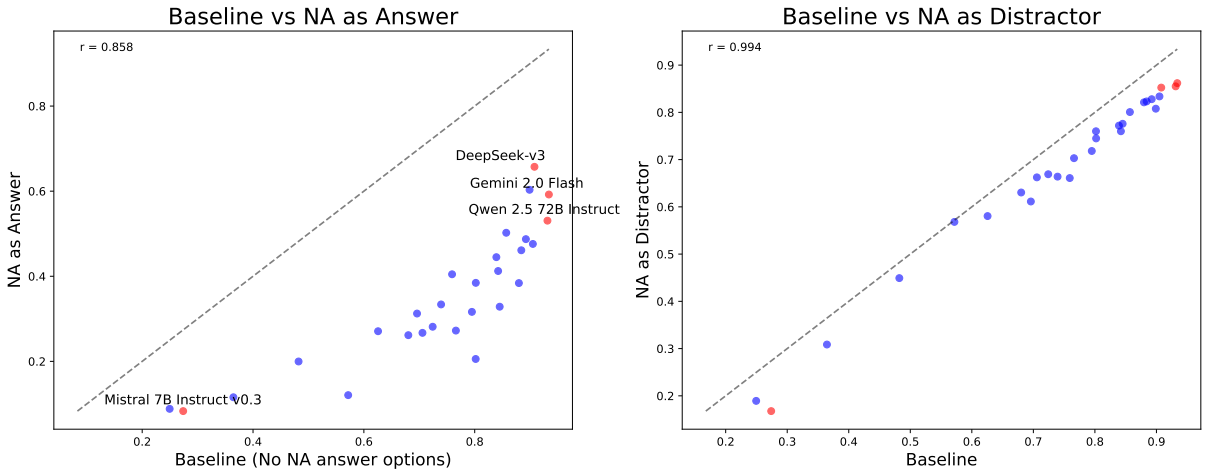

Figure 5: The left panel compares LLM performance on standard questions and on questions where the answer is replaced with "None of the Above". The right panel demonstrates that adding NA as an extra distractor leads to results similar to the baseline.

drop when NA is keyed (0.933 ± 0.044 vs. 0.965 ± 0.027), but overall, test integrity remains intact. These patterns are consistent with historical findings (Rich and Johanson, 1990) that attribute increased discrimination to the inclusion of NA.

5.4 Confidence & Sensitivity Analyses

We examine two aspects of LLM behavior under NA-as-Answer questions: changes in confidence (measured by token probabilities) and sensitivity to variations in NA phrasing. To quantify LLM confidence, we use the token probabilities returned by GPT-4-mini for the selected option (A-D) as a proxy measure.

Confidence Analysis. Figure 7 shows the relative change in confidence (based on token probabilities for the selected option) across MMLU subjects. For most subjects, adopting NA as the keyed answer lowers confidence relative to the standard format. Notably, college mathematics exhibits a slight increase in confidence (+0.01 on average), whereas International Law shows the sharpest reduction (-0.06).

Interestingly, we observe domain-specific variations in this effect. For college mathematics questions, we found a slight increase in confidence (+1% on average) when NA was the keyed option. This increase was more pronounced for correctly answered questions ($\Delta = +0.048$) compared to incorrect responses ($\Delta = 0.0$). In cases where NA was the keyed option, 57% of responses matched the previous answer choice, with these consistent responses showing a smaller confidence decrease (-0.024) compared to changed responses ($\Delta = 0.0$). This pattern likely occurs because students solving math problems often use an elimination strategy, if their calculated answer doesn't match any of the given options, they can quickly conclude that 'None of the Above' must be correct.

NA Type	LLaMA	Gemini	GPT4o
Answer not found	.134	.224	**.376**
No valid options	.216	.339	**.411**
None options are correct	.256	.358	**.494**
None of the above	.323	.317	**.476**

Table 3: Model performance across NA phrasings. "None of the above" (NOTA) shows better average performance (0.372) compared to "Not correct" (0.370). LLaMA: LLaMA 8B Instruct; Gemini: Gemini-1.5-flash; GPT4o: gpt-4o-mini

Sensitivity Analysis. We further test robustness by replacing the NA phrasing with alternatives such as "Answer not found", "No valid options", and "None of the options given is correct". As summarized in Table 3, although LLMs are moderately sensitive to these variations, the overall ranking of models remains nearly unchanged. Additional ablations using incorrect specification of the keyed answer confirm that the performance drop is specific to NA semantics and not merely any replacement.

5.5 Generalization to Classification with an 'Other' Category

The challenge posed by "None of the Above" (NA) options in MCQA is analogous to scenarios in real-

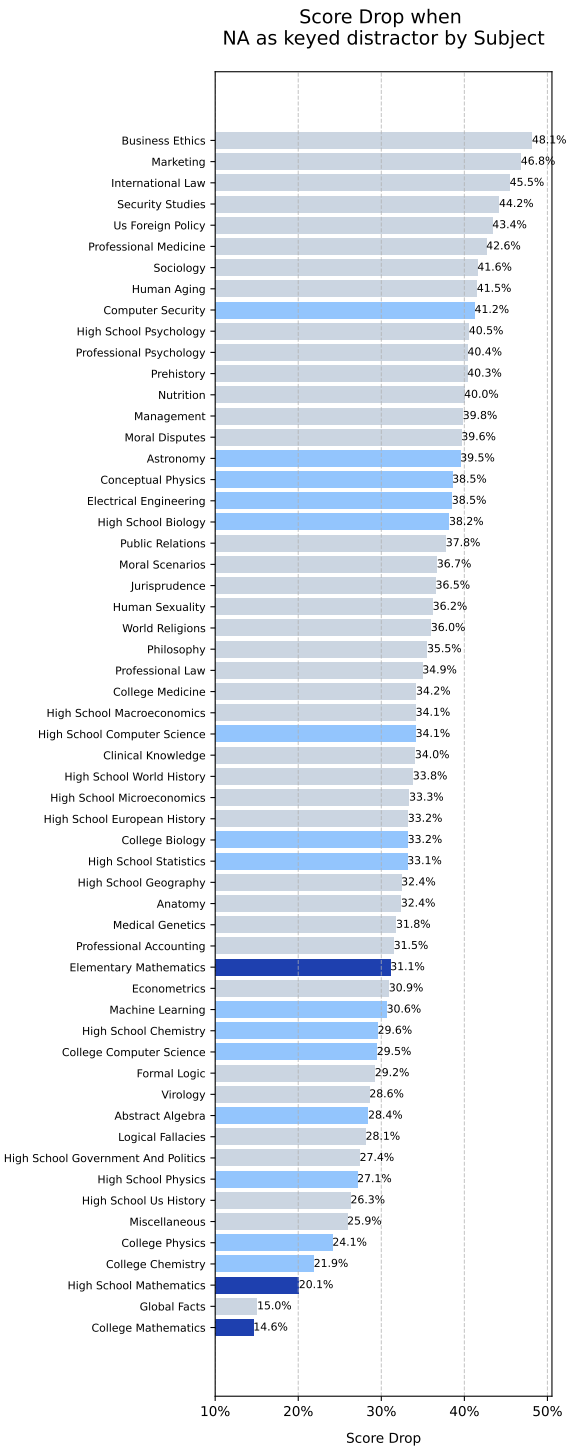

Figure 6: A rank of average drop differences from all LLMs across different subjects with Mathematics subjects highlighted in dark blue, other STEM in light blue.

world classification tasks where items may not belong to any predefined categories. In such cases, an "other" or "none of the specified classes" category is often introduced to handle out-of-distribution or irrelevant inputs. This is common in zero-shot classification tasks where LLMs are prompted to categorize inputs based on a provided list of classes. To prevent the erroneous assignment of inputs that do not fit any primary class, an "other" category, functionally similar to an NA option, is crucial.

Model	Baseline	Other
Qwen 2.5-7B-Instruct	0.493	0.142
LLaMA 3-8B-Instruct	0.422	0.358
Gemma 2-27B-it	0.621	0.062
gemini-2.0-flash-lite-001	0.658	0.962
gemini-2.0-flash	0.684	0.964
gpt-4o-mini-2024-07-18	0.656	0.852

Table 4: Model performance comparison on Baseline Accuracy and Other Accuracy metrics. Higher scores indicate better performance.

We investigate whether the performance degradation observed with NA options in MCQA extends to such classification settings. We adapt the experimental setup from Xu et al. (2024) using the Bank-77 dataset, which involves classifying user banking queries into 77 distinct intents. To simulate the presence of inputs that should be classified as "other," we augment the Bank-77 test set with 500 conversational utterances from the SODA dataset (Kim et al., 2022), specifically selecting samples that begin with "I..." to mimic user queries but are unrelated to banking intents. The LLMs are tasked with assigning one of the 77 Bank-77 classes to relevant queries or an "other" class to the SODA-derived utterances [1].

Our results, presented in Table 4, indicate that several models exhibit a notable drop in their ability to correctly classify items into the "other" category (denoted as 'NA Accuracy') compared to their performance on the original in-domain Bank-77 classes ('Baseline Accuracy'). For instance, Qwen-Qwen2.5-7B-Instruct-Turbo and google-gemma-2-27b-it show substantial performance decreases. In contrast, models such as gemini-2.0-flash-lite-001, gemini-2.0-flash, and gpt-4o-mini-2024-07-18 demonstrate stronger performance in correctly identifying "other" class items.

5.6 Improving NA Handling Through Fine-Tuning

Our analysis indicates that LLMs experience significant performance degradation when the correct answer is replaced by NA. Inspired by meta-learning strategies such as R-Tuning (Zhang et al., 2024), we explore whether targeted fine-tuning can ameliorate this weakness. We use LLaMA 8B Instruct

[1] huggingface.co/appier-ai-research/bank-77

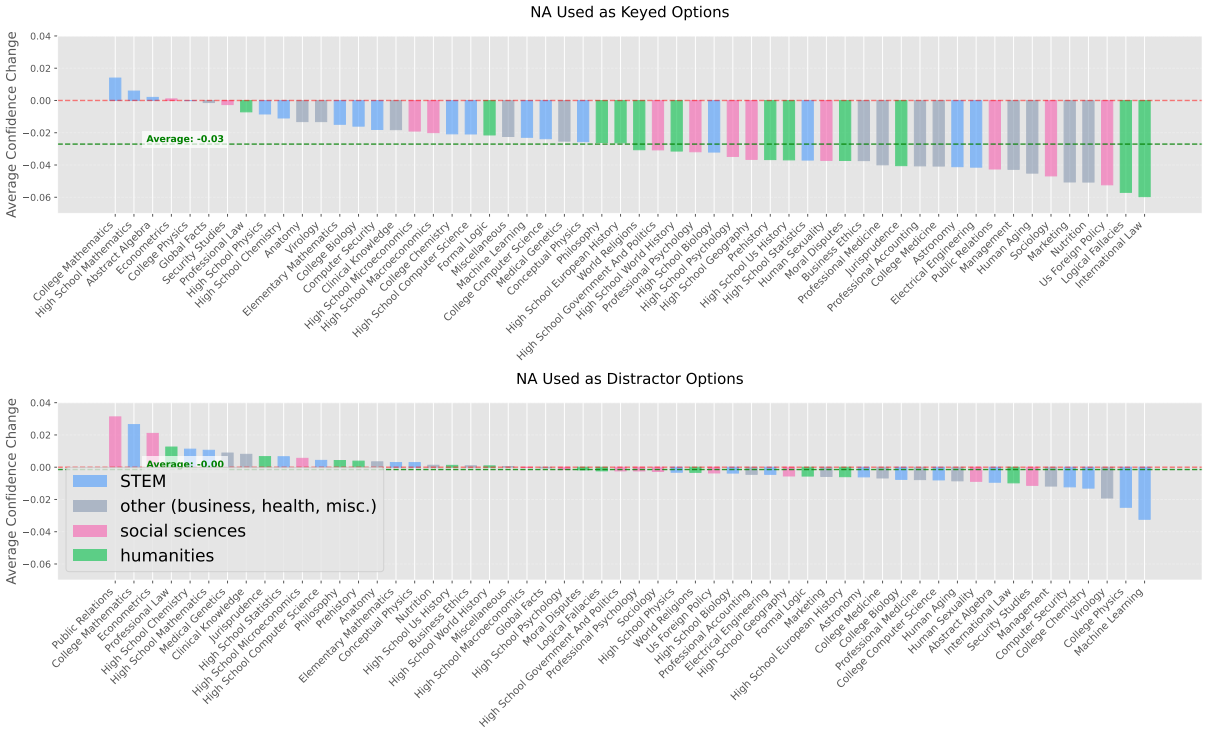

Figure 7: Confidence adjustments of gpt-4o-mini across MMLU subjects. The model predominantly reduces its confidence (mean=-0.03) after calibration, with only 3/57 subjects showing positive adjustments. College Mathematics shows the highest positive adjustment (+0.01) while International Law shows the largest reduction (-0.06).

MMLU	Baseline	SFT	DPO
Standard Format	0.632	0.625	**0.636**
NA - Average	0.463	0.523	**0.562**
- NA as Answer	0.285	0.495	**0.577**
- NA as Distractor	**0.641**	0.550	0.547

Table 5: Model performance across different question types and training methods. Higher scores indicate better performance. Baseline represent LLaMA 3 8B Instruct finetuned on no NA option questions

to created self-generated data that includes a chain-of-thought response for each answer, which serves as our training set for targeted fine-tuning. Starting from the MMLU training set (N=4650), we crafted three variants for each input question: (1) the standard format, (2) a version with the keyed option replaced by NA, and (3) a version with a distractor replaced by NA. For each version of the questions, we prompted LLM to generate 8 possible answers. We keep this set of answers only if it includes both a right and wrong answer. If all 8 answers don't meet this requirement, we discard the given set of input question. For supervised fine-tuning (SFT) (Ouyang et al., 2022), we select the first correct sample from the standard variant; for Direct Preference Optimization (DPO) (Rafailov et al., 2024), both correct and incorrect responses are used as positive and negative examples respectively.

Targeted training on NA variants improves LLaMA 3 8B (Dubey et al., 2024) found in Table 5 shows that SFT raises accuracy on NA-answer questions from 28.5% to 52.3%, and DPO further improves performance to 57.7%. In NA-distractor setting we found that Baseline model perform much better than SFT and DPO setting, inspecting the response from Baseline model, we found that Baseline avoids choosing "None of the above" options resulting in a higher final accuracy as now the random score increases from 25% to 33% as "None of the above" has simply replaced the strong distractor.

We evaluated the model's generalization on GPQA (Rein et al., 2023). Table 6 shows improved performance over baseline, particularly in NA-keyed format where DPO achieved 38.9% accuracy. However, we note that the improvements, while substantial, remain limited and highlight avenues for future research.

GPQA	Baseline	SFT	DPO
Standard Format	0.313	**0.323**	0.298
NA - Average	0.286	0.316	**0.345**
- NA as Answer	0.182	0.242	**0.389**
- NA as Distractor	**0.390**	0.349	0.300

Table 6: Model generalized to GPQA benchmark with choices replaced with NA in both keyed and distractor choice.

6 Related work

Negative effect of NA in Education Early studies (Gross, 1994; DiBattista et al., 2014) highlighted that when NOTA is correct, students may achieve high scores despite significant knowledge gaps. Blendermann et al. (2020) further demonstrated that NOTA can impair learning even with feedback, due to interference from exposure to incorrect alternatives. Conversely, work by García-Pŕrezt (1993) and Jonsdottir et al. (2021) suggests that when used as a distractor, NOTA may improve measurement accuracy and assess higher-order thinking.

None of the above in LLMs (Kadavath et al., 2022) first work to replace NA as keyed options in all MMLU questions and found all parameters scale degrades significantly with calibration performing worse as well. However in our inspection we discover not all questions are well suited to apply NA change.

MMLU Perturbation Study Recent investigations into multiple-choice question answering (MCQA) have demonstrated that LLMs are sensitive to subtle perturbations in the answer choices. For example, studies by Alzahrani et al. (2024), Zheng et al. (2023), and Wei et al. (2024) have shown that even minor changes such as reordering of the answer options can lead to variability in the models' predictions and, consequently, affect benchmark rankings. In contrast to these studies, our work examines a different and under-explored factor in MCQA design: the impact of including "None of the Above" (NA) as the correct option.

Teaching LLMs to Reject Recent work has focused on calibrating LLMs to express uncertainty and reject answers when evidence is insufficient. Although calibration methods (Zhu et al., 2023; Xie et al., 2024) and post-training refusal techniques (Zhang et al., 2024; Kapoor et al., 2024) exist, they have not been systematically applied to MCQA settings where NA is the correct answer, a gap that our study addresses.

7 Conclusion

In this study, we examined the performance of large language models on MCQA benchmarks when "None of the Above" (NA) is applied to both answer and distractor choice. Our findings reveal a dramatic performance drop—from approximately 63.2% under standard conditions down to 28.5% when the correct answers are replaced by NAs—highlighting a fundamental limitation in the models' ability to reject invalid options. While our informed fine-tuning strategy managed to improve NA accuracy to 57.7%, a significant gap remains compared to the standard accuracy. These results underscore the need to rethink MCQA benchmarks for LLMs, recognizing that tasks designed for human evaluation may not directly translate to machine understanding and uncertainty handling.

Limitations

While our study contributes valuable insights into LLM behavior on questions with NA options, several limitations should be acknowledged. First, our analysis is restricted to the MMLU benchmark; therefore, the generalizability of our findings to other multiple-choice datasets or real-world applications remains uncertain. Additionally, although our filtering and labeling approach was guided by established educational testing practices, these criteria may not fully capture the nuances in LLM evaluation. Lastly, our filtering process to identify NA-applicable questions, though guided by educational testing principles for humans, this might not be suited for LLMs.

Our experiments also depend on zero-shot chain-of-thought prompting, and alternative prompting strategies might yield different performance outcomes. Moreover, due to computational constraints, our fine-tuning work was limited to the LLaMA 3 8B model. Finally, our current method for confidence analysis relies on token probabilities as a proxy for model uncertainty—a measure that may not reflect the true calibration capabilities of LLMs (e.g., Xiong et al., 2023; Steyvers et al., 2024). Future work should focus on more direct methods for quantifying model confidence and explore extended datasets beyond MMLU.

References

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*.

axolotl-ai-cloud. 2024. axolotl. GitHub repository.

Mary F Blendermann, Jeri L Little, and Kayla M Gray. 2020. How "none of the above"(nota) affects the accessibility of tested and related information in multiple-choice questions. *Memory*, 28(4):473–480.

David DiBattista, Jo-Anne Sinnige-Egger, and Glenda Fortuna. 2014. The "none of the above" option in multiple-choice testing: An experimental study. *The Journal of Experimental Education*, 82(2):168–183.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Robert B Frary. 1991. The none-of-the-above option: An empirical study. *Applied Measurement in Education*, 4(2):115–124.

Miguel A García-Pfrezt. 1993. In defence of 'none of the above'. *British Journal of Mathematical and Statistical Psychology*, 46(2):213–229.

Leon J Gross. 1994. Logical versus empirical guidelines for writing test items: The case of" none of the above". *Evaluation & the Health Professions*, 17(1):123–126.

Thomas M Haladyna, Steven M Downing, and Michael C Rodriguez. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, 15(3):309–333.

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*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint abs/2106.09685*.

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*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Anna Helga Jonsdottir, Thorarinn Jonmundsson, Inga Huld Armann, Birna Borg Gunnarsdottir, and Gunnar Stefansson. 2021. The effect of the number of distractors and the" none of the above"-" all of the above" options in multiple choice questions. *arXiv preprint arXiv:2108.08777*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large language models must be taught to know what they don't know. *arXiv preprint arXiv:2406.08391*.

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*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *ArXiv*, abs/2212.10465.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

G Frederic Kuder and Marion W Richardson. 1937. The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Mary E Piontek. 2008. Best practices for designing and grading exams. *Occasional Paper*, 24:1–12.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Charles E Rich and George A Johanson. 1990. An item-level analysis of" none of the above.".

Mark Steyvers, Heliodoro Tejeda Lemus, Aakriti Kumar, Catarina Belém, Sheer Karny, Xinyue Hu, Lukas Mayer, and Padhraic Smyth. 2024. The calibration gap between model and human confidence in large language models. *arXiv preprint arXiv:2401.13835*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max W.F. Ku, Kai Wang, Alex Zhuang, Rongqi "Richard" Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *ArXiv*, abs/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. *arXiv preprint arXiv:2406.03009*.

Gary L Wells. 1993. What do we know about eyewitness identification? *American Psychologist*, 48(5):553.

Zhihui Xie, Jizhou Guo, Tong Yu, and Shuai Li. 2024. Calibrating reasoning in language models with internal consistency. *arXiv preprint arXiv:2405.18711*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Hanzi Xu, Renze Lou, Jiangshu Du, Vahid Mahzoon, Elmira Talebianaraki, Zhuoan Zhou, Elizabeth Garrison, Slobodan Vucetic, and Wenpeng Yin. 2024. Llms' classification performance is overclaimed. *arXiv preprint arXiv:2406.16203*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'i don't know'. 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 7106–7132.

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*.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9778–9795.

A Studies in "All of the above"

In this section we conduct experiments on the same sets of questions with choices added with "All of the above" (AA). Different from "None of the above" (NA), AA does not replace the keyed options as it cannot represent the correct choice when replacing it.

Figure 8 shows that adding "All of the above" (AA) as a fifth option has minimal impact on the relative performance ranking of Large Language Models (LLMs). The high correlation coefficient of 0.990 between baseline performance and AA-augmented questions indicates that LLMs maintain consistent relative performance patterns even when presented with AA options. This suggests that LLMs are generally robust against the potential distraction of AA choices, contrasting with their response to "None of the above" (NA) options which

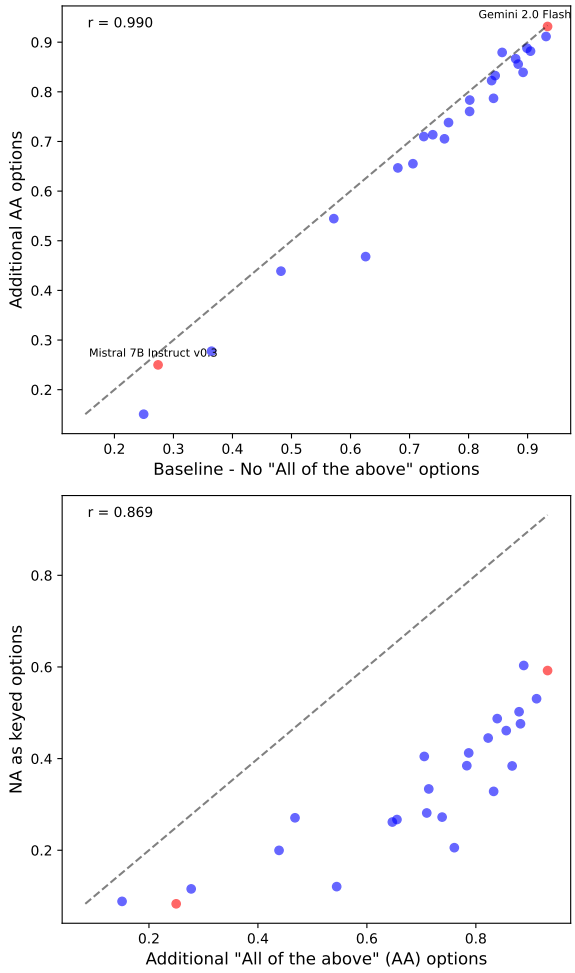

Figure 8: Upper figure : Model ranking of adding "All of the above" still contain high correlations with standard questions (Baseline) of 0.990. Lower figure : The same trend was observed in Figure 5 where NA as distractor behave differently than "Additional AA options" as shown in the figure of lower correlations of 0.869

showed lower correlations of 0.869 with baseline performance.

B List of LLMs used to evaluate results

Table 7 shows the full list of LLMs used in our benchmark. For open weights models under 30B we uses VLLM (Kwon et al., 2023) for inference evaluation, large models such as Mixtral 8x7B, LLaMA 70B, Qwen 72B we rely on TogetherAI inference API, while we use the official API endpoint provided by DeepSeek for Deepseek-V3 model.

C Average numerical scores from all 28 LLMs

Table 8 shows the MMLU-NA scores of each LLM in all 3 settings. The Baseline setting represents the standard evaluation where models are tasked with answering questions without any modifications. Overall, we observe that models consistently perform best in the Baseline setting (average 0.738), followed by the NA-distractor setting (0.674), with the NA-answer setting showing the lowest performance (0.350). Larger models like gemini-2.0-flash-exp and Qwen2.5-72B-Instruct achieve the highest scores across all settings, while smaller models like DeepSeek-R1-Distill-Qwen-1.5B and Mistral-7B-Instruct-v0.3 show significantly lower performance.

D Guideline for determining which can be NA

The prompt used to aid in the labeling of questions which can be used to assigned "None of the above" is shown in Figure 9. A detailed context of each definition is shown as below:

(1) Definitive Answer Questions - Questions in mathematics, science, or fields with exact answers where the correct option must be present. If the exact answer is not listed among the options, NOTA becomes necessary. For instance, in the question "What is 2 + 2?" with options A) 3, B) 5, C) 6, NOTA would be required if 4 is not present.

(2) Precise Knowledge Testing - Questions testing specific, verifiable knowledge where approximations are unacceptable, such as chemical symbols or technical specifications. Consider a question asking for the chemical symbol of gold - if "Au" is not among the options, NOTA becomes the correct answer.

(3) Factual Verification - Questions about historical facts, scientific principles, or established definitions where all options could potentially be incorrect. In historical questions like identifying the first U.S. President, NOTA would be correct if George Washington is not listed among the options.

(4) All choices must be mutually exclusive - Among all options, there should not be ordinal relationships or natural progressions between choices. For example, in a medical question about cannula gauge selection (18, 20, 22, 24 gauge), replacing the correct answer with NOTA would be inappropriate as the next value in the sequence would become the logical choice.

E Prompting Methods

The prompts used for both standard prompting and Chain of Thought (CoT) prompting are included in

Model	Organization	Size	Architecture
Closed Source Models			
claude-3-haiku-20240307	Anthropic	-	-
claude-3.5-haiku-20241022	Anthropic	-	-
gemini-1.0-pro (Team et al., 2023)	Google	-	MoE
gemini-1.5-flash (Team et al., 2024a)	Google	-	Transformer
gemini-1.5-flash-8b (Team et al., 2024a)	Google	8B	Transformer
gemini-1.5-pro (Team et al., 2024a)	Google	-	MoE
gemini-2.0-flash	Google	-	-
gemini-2.0-flash-lite-preview-02-05	Google	-	-
gpt-4o-mini	OpenAI	-	-
Open Weights Models			
Deepseek-V3 (Liu et al., 2024)	DeepSeek	671B	MoE
Deepseek Qwen 1.5 R1 Distill (Liu et al., 2024)	DeepSeek	1.5B	Transformer
gemma-2-2b-it (Team et al., 2024b)	Google	2B	Transformer
gemma-2-9b-it (Team et al., 2024b)	Google	9B	Transformer
gemma-2-27b-it (Team et al., 2024b)	Google	27B	Transformer
Meta-Llama-3.2-1B-Instruct (Dubey et al., 2024)	Meta	1B	Transformer
Meta-Llama-3.2-3B-Instruct (Dubey et al., 2024)	Meta	3B	Transformer
Meta-Llama-3-8B-Instruct (Dubey et al., 2024)	Meta	8B	Transformer
Meta-Llama-3.1-8B-Instruct (Dubey et al., 2024)	Meta	8B	Transformer
Meta-Llama-3.1-70B-Instruct (Dubey et al., 2024)	Meta	70B	Transformer
Mistral-7B-Instruct-v0.3 (Jiang et al., 2023)	Mistral AI	7B	Transformer
Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024)	Mistral AI	47B	MoE
Qwen2.5-1.5B-Instruct (Yang et al., 2024)	Alibaba	1.5B	Transformer
Qwen2.5-3B-Instruct (Yang et al., 2024)	Alibaba	3B	Transformer
Qwen2.5-7B-Instruct (Yang et al., 2024)	Alibaba	7B	Transformer
Qwen2.5-72B-Instruct (Yang et al., 2024)	Alibaba	72B	Transformer
SOLAR-10.7B-Instruct-v1.0 (Kim et al., 2023)	upstage	10.7B	Transformer
Yi-1.5-9B-Chat (Young et al., 2024)	01-AI	9B	Transformer
Yi-1.5-6B-Chat(Young et al., 2024)	01-AI	6B	Transformer

Table 7: Overview of evaluated models. For closed source models, sizes are marked with '-' where not publicly disclosed. MoE stands for Mixture of Experts architecture.

Model	Baseline	NA-as-answer	NA-as-distractor
gemini-2.0-flash-exp	0.934	0.592	0.862
Qwen2.5-72B-Instruct	0.931	0.531	0.855
DeepSeek-V3	0.908	0.657	0.852
gemini-2.0-flash-lite-preview-02-05	0.905	0.476	0.834
gemini-1.5-pro	0.899	0.603	0.808
gpt-4o-mini	0.892	0.487	0.828
Meta-Llama-3.1-70B-Instruct	0.884	0.461	0.823
claude-3.5-haiku-20241022	0.880	0.384	0.821
gemini-1.5-flash	0.857	0.502	0.801
gemini-1.5-flash-8b	0.845	0.328	0.776
Qwen2.5-7B-Instruct	0.842	0.412	0.760
gemma-2-27b-it	0.839	0.445	0.772
claude-3-haiku-20240307	0.802	0.206	0.760
gemma-2-9b-it	0.802	0.385	0.745
Meta-Llama-3.1-8B-Instruct	0.795	0.316	0.718
gemini-1.0-pro	0.766	0.272	0.703
Qwen2.5-3B-Instruct	0.759	0.405	0.661
Mixtral-8x7B-Instruct-v0.1	0.739	0.334	0.664
Yi-1.5-9B-Chat	0.724	0.281	0.669
Llama-3.2-3B-Instruct	0.706	0.267	0.662
Meta-Llama-3-8B-Instruct	0.696	0.312	0.611
SOLAR-10.7B-Instruct-v1.0	0.680	0.262	0.630
Qwen2.5-1.5B-Instruct	0.626	0.271	0.580
Yi-1.5-6B-Chat	0.572	0.121	0.568
gemma-2-2b-it	0.482	0.200	0.449
Llama-3.2-1B-Instruct	0.365	0.116	0.309
Mistral-7B-Instruct-v0.3	0.274	0.083	0.168
DeepSeek-R1-Distill-Qwen-1.5B	0.250	0.088	0.189
Average	**0.738**	**0.350**	**0.674**

Table 8: Model performance comparison across different metrics. Higher scores indicate better performance.

Labeling prompt

You are task to determine if the given question and its ground truth choice can be replaced with "None of the above" choice.

[Guideline] Here's some criteria where it is to replace answer with None of the above:

1. Deterministic questions:
- In math, science, or other fields with definitive answers
- Example: "What is 2 + 2?"
A) 3 B) 5 C) 6 D) 4 <= Can be replace since if 4 doesn't exists then there's no answer

2. Questions testing precise knowledge:
- Where approximate answers are not acceptable
- Example: "What is the chemical symbol for gold?"
A) Au <= Can be replaced with None of the above B) Ag C) Fe D) Cu

3. Factual questions with clear, verifiable answers:
- Historical dates, scientific facts, or established definitions
- Example: "Who was the first President of the United States?"
A) Thomas Jefferson B) John Adams C) Benjamin Franklin D) George Washington <= Can be replaced with None of the above

4. Questions with a finite set of possible answers:
- When all realistic options can be enumerated, there should not be ordinal relationships or natural progressions between choices.

[Example]
{{few shots examples}}

[INSTRUCTION]
Given the following QUESTION with its choice, determine if we can replace the Answer choice with None of the above based on the above criteria.
Answer YES if it matches any of the above criteria.

QUESTION:
{{question}}
Answer your in the following format :
REASONING: <reasons>
ANSWER: Yes/No

Figure 9: The prompt used to label

Zero-shot chain-of-thought prompt

Answer the following multiple choice question. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD. Think step by step before answering.
{{question}}

Figure 10: The prompt used in zero shot evaluation prompting

Direct answer prompting

Answer the following multiple choice question. Your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of ABCD. Answer immediately, do not think step by step.
{{question}}

Figure 11: The prompt used in zero shot evaluation prompting

Figure 10. While the prompt used to evaluate the confidence is shown in Figure 11.

During the evaluation, we used the test split of MMLU and use these hyperparameters for greedy decoding: temperature of 0.0, top-p of 1 and max tokens of 1024.

E.1 Sensitivity to Temperature

We include experiments on different temperatures in gpt-4o-mini and found the benchmark consistently holds across different settings.

Temperature	MMLU-standard	MMLU-NA	Performance Gap
0 (greedy decoding)	0.8923	0.4873	0.4050
0.6	0.8981	0.4791	0.4190
1.0	0.8967	0.4738	0.4229

Table 9: Effect of temperature settings on model performance. The performance gap between standard MMLU and MMLU-NA remains consistently around 40-42% across all temperature settings, indicating that the observed phenomenon is inherent to the model's reasoning capabilities rather than an artifact of the decoding method.

F Model Finetuning Details

For all finetuning experiments, we used Low-Rank Adaptation (LoRA) (Hu et al., 2021) to efficiently adapt the LLaMA 3 8B Instruct model. We set the LoRA rank to 128 and the scaling parameter alpha of 64.

To determine optimal training parameters, we conducted a hyperparameter sweep across three learning rates: 2e-4, 1e-4, and 8e-5. Model selection was performed based on performance on the MMLU validation set. The following hyperparameters were kept constant across all experimental configurations (baseline, supervised finetuning, and DPO):

- Batch size: 16
- Maximum sequence length: 4,096
- Optimizer: AdamW
- Weight decay: 0.1
- Learning rate schedule: Cosine decay with 10% warmup steps
- Training epochs: 3

All experiments were conducted on 2 NVIDIA 3090 GPUs with mixed-precision training (BF16). We rely on Axolotl (axolotl-ai-cloud, 2024) for training all models. The total training time for each configuration was approximately 13 hours.

G Scaling Finetuned Results to Larger Parameters

H Example Questions which is not suitable to add NA

In the following section we included only partial subjects from each four category due to the large amount of subjects from MMLU. For STEM category we included College Mathematics (Figure 12), Conceptual Physics (Figure 13) and Astronomy (Figure 14)

For Social Science category we include US Foreign Policy (Figure 15), Econometrics (Figure 16), High School Geography (Figure 17)

For Humanities we include International Law (Figure 18), High School US History(Figure 19), Jurisprudence (Figure 20)

For Other category we include Nutrition(Figure 21), Global Facts(Figure 22), Marketing (Figure 23)

College Mathematics

Suitable to add NA Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation that maps the point $(1, 2)$ to $(2, 3)$ and the point $(-1, 2)$ to $(2, -3)$. Then T maps the point $(2, 1)$ to
A. (1, 6)
B. (-1, 4)
C. (3, 2)
D. (-4, 3)

Not Suitable to add NA Let M be a 5×5 real matrix. Exactly four of the following five conditions on M are equivalent to each other. Which of the five conditions is equivalent to NONE of the other four?
A. For any two distinct column vectors u and v of M, the set u, v is linearly independent.
B. The homogeneous system Mx = 0 has only the trivial solution.
C. The system of equations Mx = b has a unique solution for each real 5×1 column vector b.
D. The determinant of M is nonzero.

Figure 12: The second question requires test takers to ignore the missing 5-th conditions which result in a missing condition which the test taker cannot determine , violating rule #3.

Conceptual Physics

Suitable to add NA A simple and correct way to comprehend satellites orbiting Earth is to view them as
A. balanced between gravitational and centripetal forces.
B. beyond the main pull of Earth gravity.
C. in mechanical equilibrium with a net force of zero.
D. having sufficient tangential velocities to fall around rather than into Earth.

Not Suitable to add NA The difference between dc and ac in electric circuits is that in dc, charges flow
A. steadily in one direction
B. in one direction
C. to and fro
D. All of these

Figure 13: The second question contains multiple correct answers : one direction in option A and B, violating the rule #4.

Astronomy

Suitable to add NA One astronomical unit (AU) is equal to ...
A. 130 million km
B. 150 million km
C. 170 million km
D. 190 million km

Not Suitable to add NA Why is the Mars Exploration Rover Spirit currently tilted towards the north?
A. Because it's climbing up a big hill.
B. Because it's in the southern hemisphere where it is winter now.
C. Because it's in the northern hemisphere where it is winter now.
D. Because one of its wheels broke.

Figure 14: The question about the Mars Exploration Rover Spirit's tilt involves a specific factual scenario that is not deterministic or based on a finite set of possible answers. The options provided are not exhaustive of all possible reasons for the rover's tilt, and the correct answer (B) is based on a specific situational context rather than a universally verifiable fact.

US Foreign Policy

Suitable to add NA 'Peace, commerce, and honest friendship with all nations, entangling alliances with none'. Identify the speaker.
A. James Madison
B. Abraham Lincoln
C. Woodrow Wilson
D. Thomas Jefferson

Not Suitable to add NA How many states in the international system are likely to have nuclear weapons right now?
A. Fewer than 7 (Answer)
B. Between 8 and 15
C. Between 16 and 25
D. More than 25

Figure 15: Reason why the first question is suitable because the question asks for the identification of a speaker of a specific quote, which is a factual question with a clear, verifiable answer. Reason why the second question is not suitable is because the correct answer is based on current geopolitical knowledge, which is not stated when and the answer could change as global powers shift.

Econometrics

Suitable to add NA Which of the following statements are true concerning the autocorrelation function (acf) and partial autocorrelation function (pacf)?

i) The acf and pacf will always be identical at lag one whatever the model

ii) The pacf for an MA(q) model will in general be non-zero beyond lag q

iii) The pacf for an AR(p) model will be zero beyond lag p

iv) The acf and pacf will be the same at lag two for an MA(1) model
A. (ii) and (iv) only
B. (i) and (iii) only
C. (i), (ii), and (iii) only
D. (i), (ii), (iii), and (iv)

Not Suitable to add NA Which of the following are advantages of the use of panel data over pure cross-sectional or pure time-series modelling?

(i) The use of panel data can increase the number of degrees of freedom and therefore the power of tests

(ii) The use of panel data allows the average value of the dependent variable to vary either cross-sectionally or over time or both

(iii) The use of panel data enables the researcher allows the estimated relationship between the independent and dependent variables to vary either cross-sectionally or over time or both
A. (i) only
B. (i) and (ii) only
C. (ii) only
D. (i), (ii), and (iii)

Figure 16: he reason why the second question is not suitable to add NA is because this is a conceptual question related to econometrics and statistics, which does not have a deterministic or factual, the options provided are not mutually exclusive, and the question does not fit into any of the criteria for replacing the answer with "None of the above." The correct answer, B, is based on understanding the specific advantages of panel data.

High School Geography

Suitable to add NA Which pair of religions below share a long-held tradition of cremating their dead?
A. Hindu and Buddhist
B. Hindu and Muslim
C. Muslim and Jewish
D. Christian and Buddhist

Not Suitable to add NA What was the primary reason the Green Revolution did NOT help Africa much?
A. It targeted crops Africans don't grow and eat.
B. It required sophisticated machinery.
C. Population growth has fallen so drastically that Africans don't need the increased food production.
D. African agricultural production went up without its help.

Figure 17: The second question asks about the primary reason the Green Revolution did not help Africa much. This is not a deterministic question with a definitive answer like a math or science question. It is also not a question with a finite set of possible answers, as there could be multiple reasons or interpretations regarding the impact of the Green Revolution on Africa.

International Law

Suitable to add NA What are the criteria for statehood under the 1933 Montevideo Convention?
A. It requires that the entity in question is not an aggressor and that it is peaceful
B. It requires recognition by the majority of other nations
C. It requires a permanent population, a defined territory, a government and a capacity to enter into foreign relations
D. It requires stable and indissoluble borders as well as recognition

Not Suitable to add NA What is the meaning of "armed attack" in Article 51 UN Charter?
A. Armed attack includes all types of armed force
B. Armed attack includes all high intensity instances of armed force (Answer)
C. Armed attack includes terrorist attacks
D. An "armed attack" gives the right to invade the aggressor State

Figure 18: The second question is not suited to apply NA because A, B, and C are similar (all about defining armed attack), hence replacing B with NA would resulted in A, C as the correct answers as well.

High School Us History

Suitable to add NA This question refers to the following information.
Here is the case of a woman employed in the manufacturing department of a Broadway house. It stands for a hundred like her own. She averages three dollars a week. Pay is $1.50 for her room; for breakfast she has a cup of coffee; lunch she cannot afford. One meal a day is her allowance. This woman is young, she is pretty. She has "the world before her." Is it anything less than a miracle if she is guilty of nothing less than the "early and improvident marriage," against which moralists exclaim as one of the prolific causes of the distresses of the poor? Almost any door might seem to offer a welcome escape from such slavery as this. "I feel so much healthier since I got three square meals a day," said a lodger in one of the Girls' Homes. Two young sewing-girls came in seeking domestic service, so that they might get enough to eat. They had been only half-fed for some time, and starvation had driven them to the one door at which the pride of the American-born girl will not permit her to knock, though poverty be the price of her independence.
—Jacob Riis, How the Other Half Lives, 1890
Riis's work as an investigator of the lives of the poor can most directly be associated with which of the following?
A. Yellow Journalism
B. Abolitionism
C. The muckrakers
D. Socialism

Not Suitable to add NA This question refers to the following information.
"The challenge of the next half century is whether we have the wisdom to use wealth to enrich and elevate our national life, and to advance the quality of our American civilization. . . .The Great Society rests on abundance and liberty for all. It demands an end to poverty and racial injustice, to which we are totally committed in our time. But that is just the beginning. The Great Society is a place where every child can find knowledge to enrich his mind and to enlarge his talents. It is a place where leisure is a welcome chance to build and reflect, not a feared cause of boredom and restlessness. It is a place where the city of man serves not only the needs of the body and the demands of commerce but the desire for beauty and the hunger for community. It is a place where man can renew contact with nature. It is a place which honors creation for its own sake and for what it adds to the understanding of the race. It is a place where men are more concerned with the quality of their goals than the quantity of their goods. But most of all, the Great Society is not a safe harbor, a resting place, a final objective, a finished work. It is a challenge constantly renewed, beckoning us toward a destiny where the meaning of our lives matches the marvelous products of our labor."
Lyndon Johnson, Remarks at the University of Michigan, Ann Arbor, 1964
Which one of the following was an unintended consequence of the liberal successes of the 1960s?
A. Liberal Democrats abandoned anti-war protests in a show of support for President Johnson.
B. Conservative Republicans mobilized to defend traditional mores and curb government authority.
C. Economic recession catalyzed by increased government spending causing "stagflation."
D. A majority of Northern black voters abandoned the Democrat party, siding with Republicans.

Figure 19: The reason why the second questio is not suitable to add NA is because the options provided are not deterministic or factual in the sense of having a single, verifiable answer like a math problem or a historical date.

Jurisprudence

Suitable to add NA Which of the following statements is correct concerning the "reasonable person" standard in tort law?
A. The reasonable person standard varies from person to person.
B. The reasonable person standard focuses on the defendant's subjective mental state rather than on the defendant's behavior
C. A person with a physical disability must act as would a reasonable person with the same disability.
D. A person with a mental disability must act as would a person with the same mental disability.

Not Suitable to add NA Austin has been described as a 'naive empiricist.' Why?
A. Because he neglects the importance of morality.
B. Because his account of law is based on an anachronistic model of a legal system.
C. Because he conceives of laws in a pragmatic rather than a conceptual manner.
D. Because he overlooks the role of law in economic relations.

Figure 20: The second question is not well suited to add NA because the question is more interpretative and subjective, likely based on philosophical or theoretical analysis, which does not lend itself to a "None of the above" option. The answer choice "C" is based on a specific interpretation of Austin's views, which may not be universally agreed upon or verifiable in the same way as a factual or deterministic question.

Nutrition

Suitable to add NA Which statement about the oral phase of digestion is INCORRECT?

A. About 2% of the energy content of food is expended during the action of chewing and swallowing it.
B. Swallowing involves contraction and relaxation of at least 14 groups of muscles in about 10 seconds in healthy subjects
C. The biofilm covering tooth enamel contains several salivary and bacterial enzymes
D. Salivary amylase digests the dextran film on tooth enamel formed from dietary sucrose
Not Suitable to add NA Which of the following confirmed values meet the diagnostic threshold for diabetes?

A. fasting blood glucose ? 140 mg/dl
B. random glucose > 160 mg/dl
C. 2 hour post prandial glucose $\geq$ to 126 mg/dl
D. fasting blood glucose $\geq$ 126 mg/dl (Answer)

Figure 21: The second question's contain option A which is not well represented and numeric number in C, D contains similar number, which violates rule #1.

Global Facts

Suitable to add NA Which of the following countries emitted the most CO2 in 2017?
A. Canada
B. Russia
C. Iran
D. Japan

Not Suitable to add NA At its peak, what was the approximate difference in approval of school desegration from the South and the rest of the U.S.?
A. 80% (Answer)
B. 40%
C. -40%
D. -80%

Figure 22: Since the second question refers to "approximate" replacing 80% answer would result in 40% being the next possible answer in-line.

Marketing

Suitable to add NA _____ can be defined as the aspect of our psyche that determines the way in which we respond to our environment in a relatively stable way over time.
A. Perception.
B. Personality.
C. Learning.
D. Memory.

Not Suitable to add NA Second-hand data, collected for someone else's purposes, is known as_____:
A. Primary research.
B. Descriptive research.
C. Causal research.
D. Secondary research.

Figure 23: The second question contain bad options clarity which violates rule #1, #2.