International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
https://ijcjournal.org/index.php/InternationalJournalOfComputer/index

Unified Benchmark for Evaluating Performance, Bias, and Consistency in LLM Binary Question Answering

Olesia Khrapunova*

Senior AI/ML Engineer, Paris, France
Email: khrapunova.ml@gmail.com

Abstract

Binary question answering is central to many real-world applications of large language models (LLMs), such as fact-checking or decision-making support. Yet, despite its prevalence and the high stakes of getting a binary judgment wrong (where an error yields the exact opposite outcome), there are no recent comprehensive benchmarks dedicated to evaluating LLM behavior on this task. To address this gap, we introduce a unified benchmark for assessing binary QA across three dimensions: performance, bias, and consistency. The benchmark is supported by a five-domain dataset augmented with new controlled reformulations of each question, including paraphrases, negations, and answer option variations. Across fifteen state-of-the-art LLMs, we find strong overall performance on the task, with larger and reasoning-optimized models showing better results than the smaller variants. At the same time, we observe pervasive No-leaning bias, universally weak consistency when handling semantically opposite questions, and substantial cross-domain variation. Reading comprehension and multi-hop reasoning topics are handled reliably, whereas numerical reasoning, ethical judgment, and, especially, translation evaluation remain challenging. These findings reveal both the strengths and shortcomings of current LLMs on binary QA, providing researchers with a basis for targeted future improvements while also helping practitioners make informed choices when deploying the models in binary decision contexts.

Keywords: Benchmarking; Bias; Binary Question Answering; Consistency; Large Language Models; Performance Evaluation.

Received: *10/17/2025*
Accepted: *12/17/2025*
Published: *12/27/2025*

---

* *Corresponding author*.

1. Introduction

1.1. Problem Description

Large language models (LLMs) are increasingly integrated into everyday workflows, powering tools used for drafting content, task automation, and even decision support. Many of these interactions ultimately require the model to resolve a binary question, such as whether a statement is correct or whether a user should take a particular action. Therefore, their ability to answer Yes/No reliably becomes an essential component of effective downstream system behavior and user trust.

Binary questions, however, present a unique challenge because they compress the decision space into only two polar opposite outcomes: Yes or No. In borderline cases, where evidence is ambiguous or phrasing is vague, the system must still commit to a single categorical answer, even when its underlying uncertainty is high. This challenge is further compounded by the fact that LLMs are inherently statistical systems: their outputs reflect the patterns and biases present in their training data. When applied to binary question answering, these tendencies may surface in critical ways.

This raises several practical questions for real-world deployment. How accurate are LLMs when making binary judgments? Do they reliably preserve their answers when binary questions are paraphrased or otherwise changed? And do they exhibit inherent tendencies, such as being more inclined to say No or Yes, that could systematically skew outcomes? Understanding these behaviors is crucial for designing safe and predictable LLM-based systems.

Yet, to the best of our knowledge, there is currently no unified benchmark to make this assessment. This paper addresses this gap by introducing the first systematic, multi-dimensional evaluation framework that jointly measures performance, bias, and consistency of LLMs on the binary question answering task. To support this framework, we assemble a curated dataset of 1,000 cross-domain Yes/No questions, augmented with paraphrases and structured negations, to test LLM behavior in various contexts. We then apply the framework across 15 state-of-the-art models to provide a transparent snapshot of current models' binary QA abilities and patterns.

All in all, this work offers a new, comprehensive benchmark designed specifically to evaluate binary QA behavior in LLMs. It establishes a foundation for greater understanding of AI systems that rely on these judgments and a clear roadmap for necessary LLM improvements.

1.2. Related Work

In the past few years, the rapid expansion of LLM capabilities has driven a corresponding surge in evaluation benchmarks, supporting quantitative assessment of these models. A recent meta-analysis by Ni and colleagues [1] cataloged 283 benchmarks spanning general-purpose evaluations, domain-specific assessments, and target-specific tasks. Notably, despite its breadth, the survey did not identify any benchmarks explicitly dedicated to binary question answering (binary QA). This suggests that, although binary decisions are deeply embedded in real-world applications, they remain an under-examined part within the LLM evaluation landscape. And although there are some earlier datasets that focus on Yes/No questions (most prominently, BoolQ [2]), they remain rare and insufficient to support systematic benchmarking of binary decision-making across diverse scenarios.

This absence is increasingly important as emerging work reveals that binary formats introduce systematic biases in LLM behavior that can meaningfully distort model outputs in downstream applications. Recent work by Yu and colleagues [3] demonstrated the existence of negative bias in LLMs (specifically, in mathematical and logical reasoning tasks) through the introduction of the Negative Attention Score (NAS). Using NAS, they identified dedicated, query-agnostic negative attention heads that disproportionately attend to negative tokens in a prompt during binary decision tasks. They showed that this increased focus on negative tokens is correlated with a higher confidence in the negative answer. In other words, LLMs tend to output negative responses more frequently and with greater confidence, while being more cautious with positive responses. This leads to a skewed, risk-averse decision behavior detrimental in scenarios that require balanced Yes/No judgments.

Lu and colleagues [4] extended this mechanistic view by demonstrating that the binary format itself (i.e., 'Yes/No') significantly amplifies negative bias relative to continuous format (i.e., 'on a scale from X to Y'), with consistent effects observed across value judgments and sentiment analysis tasks. Further strengthening the empirical evidence, Braun conducted experiments on legal texts, showing that LLMs, once again, display bias towards choosing No [5]. Similarly, Cheung and colleagues shared observations of LLMs preferring No in moral dilemmas and even flipping their decision based on how the question was worded (i.e., answering No even if the reworded question suggested the opposite choice) [6].

This pattern of changing the answer suggests that bias in binary judgment surfaces together with broader LLM consistency issues, which have been observed on a variety of tasks. However, in a binary scenario, their implications are especially critical, as inconsistency does not just shift a nuance in response but can invert the decision entirely. For this reason, we briefly review prior findings on LLM inconsistency here.

An early consistency benchmark, BECEL [7], showed that traditional language models (e.g., BERT, T5) often violate logical relations across semantically paraphrased, negated, symmetric, and transitive variants of the same input, with negation being especially challenging. That is, when logically equivalent forms of a question are presented, the model's predictions can diverge significantly, revealing fundamental instability in its reasoning. Later analysis of ChatGPT and GPT-4 confirmed that although improvements have been made, the models continue to make mistakes that violate logical properties, with considerable frequency [8]. Similar findings have been made by Ahn and Yin, who observed inconsistency in LLMs handling of opposite prompts [9], and Atil and colleagues who observed inconsistencies in LLMs responding to the same prompt several times [10]. Moreover, Labruna and colleagues demonstrated that in binary question answering, models may change their selection solely due to the ordering of response options (especially in high uncertainty situations) [11].

Taken together, these findings reveal that binary decision-making is a vulnerable area for current LLMs. It is shaped by systematic negative bias and further weakened by inconsistencies that can invert a model's decision under even small input changes. Yet, despite the clear practical importance of binary judgments and the growing evidence that LLMs struggle with them, no existing benchmark is designed to systematically evaluate this capability. This gap underscores the need for a dedicated, comprehensive framework for assessing LLM behavior on binary QA that jointly measures performance, bias, and consistency, to enable robust comparisons across models and guide progress in future LLM development. In this paper, we propose such a framework.

2. Materials and Methods

Our approach centers on constructing a comprehensive multi-domain binary question answering dataset, augmenting this dataset with systematic reformulations, and deploying a standardized evaluation protocol for it.

2.1. Main Evaluation Dataset

To capture a broad range of binary question answering settings, we constructed a 1,000-sample dataset, spanning five domains: reading comprehension, multi-hop reasoning, numerical reasoning, machine translation error detection, and ethical judgment. Each domain contributes 200 items. All samples were normalized to a binary Yes/No format, with label distributions balanced within each domain (i.e., 100 Yes samples and 100 No samples). Below, we describe the sourcing and transformation process for each domain in detail.

Reading comprehension. We randomly sampled 200 instances from the BoolQ dataset [2]. This is a question answering dataset that contains naturally occurring binary questions. Each question can be answered by extracting information from the associated passage and does not require complex reasoning. The only transformation applied to this data was converting the True/False labels to Yes/No. The class balance was enforced by sampling an equal number of Yes and No items.

Multi-hop reasoning. We sampled 200 questions from HotpotQA dataset [12]. This dataset is based on Wikipedia articles and focuses on the question answering task designed specifically for multi-hop reasoning. Each question in the dataset requires integrating information from multiple supporting documents, encouraging systems to reason over scattered evidence rather than relying on a single-passage lookup. The dataset includes two subsets: 'distractor', which provides a fixed context of ten paragraphs (two relevant paragraphs and eight distractors), and 'full wiki', which requires retrieval over the paragraphs of the entire Wikipedia corpus. For our experiments, we selected items exclusively from the 'distractor' subset. To align with our binary classification setting, we included only the questions with Yes/No answers, eliminating the need for additional reformulation. All selected questions belong to the 'comparison' category, which requires comparing two entities along a shared attribute. Of the 200 examples, exactly 100 had the correct answer Yes, and the other 100 had the correct answer No. Because HotpotQA spans multiple difficulty levels, we constructed a balanced sample comprising 20% easy, 30% medium, and 50% hard questions within each label group to ensure adequate coverage of the dataset's challenge spectrum.

Numerical reasoning. We sampled 200 examples from the DROP dataset [13]. This dataset requires a system to analyze the provided paragraph, extract relevant information, and perform discrete operations (e.g., addition or counting) over it to answer the question. It requires a comprehensive understanding of the context and the ability to perform mathematical operations. Because DROP does not inherently contain binary-style questions, we manually converted each randomly selected item into a Yes/No format. Questions were reformulated either as threshold-based queries (e.g., transforming *'How many years did Dhanraj Pillay's career span?'* into *'Did Dhanraj Pillay's career span over 14 years?'*) or as exact number verification queries (e.g., reframing *'How many TD runs were there in the game?'* as *'Were there 2 TD runs in the game?'*). To maintain a balanced label

distribution, 100 items were rephrased such that the correct answer was Yes, and 100 such that the correct answer was No, with a 50/50 split between threshold and exact number styles within each label category.

Machine translation error detection. We sampled 200 items from the WMT (Workshop on Machine Translation) data [14], specifically the 2020 WMT dataset with new translations from English to German. It contains human-generated translation error annotations following the MQM (Multidimensional Quality Metrics) framework, in which each error is labeled with both a type and a severity level. It requires that the model draw on its inherent linguistic knowledge and evaluate whether the target sentence is an appropriate translation of the source. For our purposes, we simplified the labeling scheme by retaining only the binary presence or absence of errors, disregarding type and severity. Given that translation error detection is somewhat subjective, we only included samples where all annotators agreed about the presence/absence of an error in the segment. Each sample was then converted into a Yes/No question of the form *'Is there an error in the translation?'*, where the corresponding passage contained the translation to be evaluated. If the annotators agreed on the presence of an error in the segment, the correct answer would be Yes; otherwise, No. We included 100 segments with an error and 100 segments with no errors.

Ethical judgment. We curated 200 scenarios from the ETHICS dataset [15]. It includes subsets that require the model to make judgments regarding justice, common sense, deontology, utilitarianism, and virtue. For our dataset, we only included common sense and justice subtypes (100 samples of each), as these categories reflect forms of everyday moral judgment that LLMs are likely to have acquired during pretraining. To align these scenarios with our binary framework, we reformulated each item into a Yes/No query. Justice scenarios were posed as *'Is this fair/just according to ordinary morality in usual circumstances?'*, while common sense scenarios were phrased as *'Is this acceptable according to ordinary morality in usual circumstances?'*. Within each subset, we balanced the labels by ensuring a 50/50 distribution of Yes and No answers, for a total of 100 Yes and 100 No samples.

2.2. Dataset Reformulations

To evaluate the consistency of model outputs, we generated several controlled reformulations for each question in the dataset. These variations modify the form of the question while preserving its underlying meaning and ground-truth label (or, in the case of negative reformulations, intentionally reversing the label to reflect the logical change). To minimize unintended sources of variance, each reformulation type followed a predefined transformation pattern. All variants were first produced using GPT-4.1 [16] in a one-shot prompting setup, where each reformulation type was illustrated with a single example. The generated rewrites were then manually reviewed by the author to ensure correctness. We applied three categories of reformulations: positive question reformulations, negative question reformulations, and answer option reformulations.

Positive question reformulations. These preserve the meaning and the true answer of the original question, while introducing controlled variations. These transformations are designed to test whether models remain stable under paraphrasing in the form of lexical or stylistic shifts. We applied two types of positive transformations:

1. Synonym Replacement: substitutes key words or phrases with close equivalents (e.g., *'Is caffeine found*

in black tea?' → 'Is caffeine present in black tea?')

2. Politeness/Hedging: incorporates pragmatic markers, such as polite or tentative expressions (e.g., *'Is Paris the capital of France?' → 'Would you say Paris is the capital of France?'*)

Negative question reformulations. These create a question with an opposite meaning, thereby requiring the correct label to flip. These transformations evaluate whether models can reason accurately under negation and stay logically consistent when the same question is framed in a contradictory way. We applied two forms of negative transformations:

1. Contradictory Reformulation: rewrites the question around an opposite hypothesis, including through the use of negation words (e.g., *'Does the sun rise in the east?' → 'Does the sun not rise in the east?'*)
2. Explicit 'False' Framing: adds the explicit negation expression (e.g., *'Is the moon a star?' → 'Is it false that the moon is a star?'*)

Answer option reformulations. This reformulation is not applied on a per-question basis but instead changes the core prompt uniformly across all examples. Whenever a question is passed to an LLM, we specify the available answer options it can return. Therefore, this reformulation examines whether models change their answer solely due to the order of the response options presented to them. We evaluated the LLMs on two versions of the prompt, one with the Yes/No order (*'Choose one of the following answers: Yes/No'*) and one with the No/Yes order (*'Choose one of the following answers: No/Yes'*).

2.3. Binary Question Answering Evaluation Framework

To systematically assess model abilities and patterns in binary question answering, we designed an evaluation framework built around four complementary metric groups: Performance, Bias, Consistency, and Cross-Domain Stability. The first three categories capture the core aspects of LLM behavior this benchmark set out to measure. The Cross-Domain Stability metrics show how much these behaviors vary across different topics, helping us see if they are general or domain-dependent. All these components are then combined into a unified Model Binary Score (MBS) to make model comparisons more straightforward.

Performance metrics. Assess the model's overall effectiveness on the binary question answering task, with all metrics computed on the original (non-reformulated) dataset. They determine whether the model is fundamentally capable of performing the task and could be useful to generate predictions in a binary setting.

- **F1 (weighted)** - an overall quality measure that adjusts for label distribution between Yes and No classes.

Bias metrics. Quantify whether the model's performance differs between Yes and No classes, with all metrics computed on the original (non-reformulated) dataset. This matters because a model that performs well overall may still have divergent results on specific labels, affecting fairness and reliability of the downstream decision-making.

- **∆Recall (No - Yes)** - difference in recall of the two classes.

- ∆**Precision (No - Yes)** - difference in precision of the two classes.

Consistency metrics. Evaluate how consistently a model answers a question when its phrasing is altered in ways that do not change its underlying logical meaning; computed using the dataset reformulations described above. This matters because a reliable model should not change its decision simply because the question is phrased differently. If meaning-preserving reformulations cause the model to switch the answer, this reveals fragility and suggests that the model might be relying on superficial patterns rather than genuine reasoning.

- **Positive reformulation consistency (PRC)** - measures the frequency with which the answer stays the same across multiple semantically equivalent versions of the same question; uses positive question reformulations (synonym replacement and politeness/hedging). Equation 1 shows how sample-level positive reformulation consistency indicator is calculated, where $a_{j,0}$ is the model's answer to the original question j, $a_{j,i}$ is the model's answer to its i-th reformulation of question j, and $PRC_j = 1$ only if answers to all the question variants match the original.

$$PRC_j = 1[a_{j,i} = a_{j,0} \; \forall \; i \in \{1, \dots, n\}] \quad (1)$$

Equation 2 shows how the dataset-level *PRC* score is calculated, where N is the number of samples in the dataset and PRC_j is the sample-level positive reformulation consistency indicator for question j.

$$PRC = \frac{1}{N}\sum_{j=1}^{N} PRC_j \quad (2)$$

- **Negative reformulation consistency (NRC)** - measures how consistently the model handles question negations, i.e., whether it provides the correct opposite answer when the question is rewritten to express the contradictory form of the same claim; uses negative question reformulations (contradictory reformulation and explicit 'false' framing). Equation 3 shows how sample-level negative reformulation consistency indicator is calculated. Dataset-level *NRC* score is calculated in the same way as *PRC* score but using NRC_j indicator instead of PRC_j indicator.

$$NRC_j = 1[a_{j,i} \neq a_{j,0} \; \forall \; i \in \{1, \dots, n\}] \quad (3)$$

- **Answer reformulation consistency (ARC)** - measures the frequency with which the model response remains consistent when the question stays the same and only the presentation of the answer options is changed; uses answer option reformulations (changes in order). It is calculated in the same way as *PRC*.
- **Self-consistency (SC)** - measures the frequency with which the answer stays the same when the identical question is asked several times; we ask the same question 3 times to get the measure (note that we kept the temperature parameter at 0 to minimize the randomness). It is calculated in the same way as *PRC*.

Cross-domain stability metrics. Assess whether a model's performance, bias, and consistency patterns hold across the different domains in our dataset. As each domain poses distinct demands, comparing metric values across them reveals whether observed effects are general or tied to specific contexts. This helps distinguish

fundamental LLM tendencies from domain-driven phenomena.

- **Stdev of the performance metrics** - provides a quantitative measure of how widely performance varies between domains; for each metric, we compute its value separately for each domain and then calculate the standard deviation across these values.
- **Stdev of the bias metrics** - provides a quantitative measure of how widely bias varies between domains; for each metric, we compute its value separately for each domain and then calculate the standard deviation across these values.
- **Stdev of the consistency metrics** - provides a quantitative measure of how widely consistency varies between domains; for each metric, we compute its value separately for each domain and then calculate the standard deviation across these values.

2.4. Composite Model Binary Score

We combine all previously defined metrics into a single composite measure, Model Binary Score (MBS), which provides a concise overall assessment of model behavior on the binary task. While the individual metrics reveal specific strengths and weaknesses, the composite score consolidates performance quality, bias level, consistency under reformulations, and cross-domain stability into one value. This allows all key dimensions of binary question answering quality to be represented collectively in a single summary measure. To create MBS, each metric is first normalized to a common 0–100 scale based on its theoretical minimum and maximum values, where higher values always indicate better performance. Each normalized metric is then weighted according to its predefined importance, and the weighted values are added together to produce the composite score.

Metric normalization. Because different metrics have different natural ranges and interpretations, we apply range-specific normalization rules as described below:

- **Symmetric metrics on -1 to 1 scale** - metrics with the range of [−1,1], where 0 represents the best case and −1/1 represent the worst case (ΔRecall, ΔPrecision) are normalized as: $\widetilde{M} = 100 * (1 - |M|)$
- **Increasing metrics on 0 to 1 scale** - metrics bounded by [0,1], where larger values correspond to better performance (F1, PRC, NRC, ARC, SC) are normalized as: $\widetilde{M} = 100 * M$
- **Decreasing metrics on 0 to 1 scale** - metrics bounded by [0,1], where smaller values indicate better performance (standard deviation of ΔRecall, ΔPrecision) are normalized by inverting the scale: $\widetilde{M} = 100 * (1 - M)$
- **Decreasing metrics on 0 to 0.5 scale** - metrics with the range of [0, 0.5], where 0 is the best and 0.5 is the worst (standard deviation of F1, PRC, NRC, ARC, SC) are normalized as: $\widetilde{M} = 100 * (1 - 2M)$

Weight assignment. Each normalized metric contributes to the composite Model Binary Score according to a predefined weighting scheme that reflects the relative importance of the four major evaluation components. Below are the weights we selected for the experiments based on our assessment of the criticality of each aspect:

- **Performance metrics (40%)** - represent the core model capability on the binary task and therefore

receive the largest share of the overall weight (all assigned to F1, the single metric in this category).

- **Bias metrics (20%)** - capture a core, but less important evaluation dimension. The total allocation is distributed equally between ΔRecall (10%) and ΔPrecision (10%).
- **Consistency metrics (30%)** - constitute the second most important category as they capture model reliability under variation occurring with normal use. This weight is distributed equally among all metrics in the category (i.e., 7.5% for each).
- **Cross-domain stability (10%)** - receive the remaining part of the score. To keep the influence of each cross-domain metric aligned with the importance of the corresponding core metric, each stdev weight is assigned proportionally to the weight of its associated main metric. In other words, if w_i is the weight of a core metric M_i, and the total weight of all cross-domain metrics is 0.1, then the weight of its cross-domain stability metric is: $w_{stdev\ i} = 0.1 * \frac{w_i}{1-0.1}$

2.5. Large Language Models Evaluated

We applied the benchmark to evaluate a set of 15 LLMs. To ensure relevance, we used the most recent releases from each model family, providing an up-to-date snapshot of current LLM capabilities. The selection spans multiple sizes and includes both open-weight and closed-weight models. Whenever available, we also included reasoning-optimized variants. The chosen models were:

1. **o3, OpenAI:** Closed-weight, reasoning-optimized model; state-of-the-art for structured reasoning [17].
2. **GPT-4o, OpenAI:** Large, closed-weight model; flagship and highly performant general-purpose system Reference [18].
3. **GPT-4o mini, OpenAI**: Small, closed-weight model; cost-efficient and fast variant of GPT-4o [19].
4. **Gemini 2.5 Pro, Google:** Large, closed-weight model; Google's high-capability flagship system with strong reasoning capabilities [20].
5. **Gemini 2.5 Flash, Google:** Small, closed-weight model; efficient, fast-response variant of the Pro model Reference [20].
6. **Claude Opus 4.1, Anthropic:** Closed-weight, reasoning-focused model; strongest in Anthropic's lineup for specialized reasoning tasks [21].
7. **Claude Sonnet 4.5, Anthropic:** Large, closed-weight model; Anthropic's main general-purpose system Reference [22].
8. **Claude Haiku 4.5, Anthropic:** Small, closed-weight model; lightweight, fast variant of Claude [23].
9. **Magistral Medium 1.2, Mistral AI:** Medium-sized model optimized for reasoning tasks, with weights provided on request [24].
10. **Mistral Medium 3.1, Mistral AI:** Mistral's main frontier-class model with weights available upon request; medium-sized and designed to deliver strong general-purpose performance [25].
11. **Mistral Small 3.2, Mistral AI:** Small, open-weight model; efficient and publicly accessible [26].
12. **Llama 4 Maverick, Meta AI:** Large, open-weight model; the latest addition by Meta with a best-in-class performance-to-cost ratio [27].
13. **Llama 3.1 405B Instruct, Meta AI:** Large, open-weight model; top-tier performance with broad

capabilities [28].

14. **Llama 3.3 70B Instruct, Meta AI:** Medium-sized, open-weight model; delivers strong performance, even approaching the 405B variant in some cases [28].
15. **Llama 3.1 8B Instruct, Meta AI:** Small, open-weight model; Meta's efficient general-purpose system Reference [28].

In addition, we evaluated a **random baseline** that predicts Yes and No with equal probabilities of 1/2, offering a simple chance-level reference for comparing various model performance results to.

3. Results

3.1. General LLM Trends

Table 1: Overall benchmark results, core metrics (best 2 highlighted)

Model	Performance	Bias		Consistency				
	F1	ΔRecall	ΔPrecision	PRC	NRC	ARC	SC	MBS
Baseline	0.46	0.01	0.00	0.27	0.26	0.48	0.24	57.4
o3	**0.88**	0.07	-0.06	**0.90**	**0.87**	0.96	0.94	90.2
GPT-4o	0.83	0.09	-0.06	0.87	0.67	0.95	0.97	86.0
GPT-4o mini	0.77	0.06	-0.04	0.85	0.67	0.95	0.97	83.9
Gemini 2.5 Pro	**0.88**	0.05	-0.04	**0.90**	**0.86**	0.96	0.95	**90.5**
Gemini 2.5 Flash	0.86	0.07	-0.05	0.87	0.82	0.95	0.98	89.0
Claude Opus 4.1	**0.88**	**0.02**	**-0.01**	**0.91**	0.77	**0.98**	0.98	**91.0**
Claude Sonnet 4.5	**0.87**	0.05	-0.03	**0.90**	0.69	**0.97**	0.99	89.4
Claude Haiku 4.5	0.85	0.14	-0.10	0.88	0.72	**0.97**	**1.00**	86.9
Magistral Medium 1.2	0.85	-0.05	0.04	0.83	0.77	0.93	0.90	87.3
Mistral Medium 3.1	0.80	0.13	-0.08	0.89	0.49	0.93	**0.99**	83.0
Mistral Small 3.2	0.78	0.15	-0.09	0.83	0.58	0.95	0.98	82.4
Llama 4 Maverick	0.84	0.05	-0.03	0.87	0.68	0.95	0.93	86.5
Llama 3.1 405B	0.81	**-0.03**	**0.02**	0.85	0.67	**0.98**	**0.99**	86.4
Llama 3.3 70B	0.82	0.13	-0.09	0.82	0.76	0.96	0.96	85.0
Llama 3.1 8B	0.68	0.28	-0.12	0.80	0.42	0.94	**1.00**	74.4

Overall, as shown in Table 1, all LLMs in the benchmark performed well on the binary question answering task, substantially outperforming the random baseline on the calculated metrics. The main outlier is Llama 3.1 8B, which achieved the lowest scores across most dimensions, although still performed above chance. Importantly, this model is likely the smallest in the set. For example, Mistral Small 3.2, another small open-weight model, is 3 times larger, with 24B parameters [26]. While the exact sizes of the OpenAI, Google, and Anthropic models are not publicly disclosed, it is reasonable to assume that even their small offerings are considerably larger as well. This raises the possibility that model size plays a role in model behavior on the binary QA task, an idea we explore further in a separate section. For the weighted F1 statistic, most models achieved scores greater than 0.80, reflecting solid overall performance on the task. Nevertheless, the absence of near-perfect results indicates that the task retains inherent difficulty and that performance can still be improved.

In addition, for nearly all models (except for Magistral Medium 1.2 and Llama 3.1 405B) we observe an imbalance between the Yes and No classes, where the No class shows higher recall but lower precision than the Yes class. This pattern suggests that, across the board, current LLMs (especially the smaller variants) have a bias towards answering No more readily than Yes when faced with binary decision-making scenarios.

In terms of consistency, the models show strong self-consistency. When the same question was repeated three times, they almost always produced the same output, with all the models reaching 90% on the metric and many exceeding 97%. Interestingly, the reasoning-optimized models (i.e., o3, Magistral Medium 1.2) showed slightly lower self-consistency, likely because their greater flexibility and broader search space introduced more variation in more ambiguous cases. The LLMs also demonstrated high consistency with respect to answer option reformulations. When the order of answer options was changed, all the models maintained consistency in over 93% of cases, with most being consistent in 95%+ cases. This indicates that simple ordering shifts have minimal impact on their binary predictions. Consistency under positive question reformulations, such as replacing words with synonyms or adding polite expressions, was also quite good, but noticeably weaker. Most models remained consistent in the mid-to-high 80s range, with only the strongest reaching around 90% stability, indicating that even small shifts in wording can influence their outputs. Answering consistently to negative reformulations, such as contradictory questions or explicit 'false' framings, proved to be the most challenging. In these cases, models were expected to flip the returned label while preserving the same underlying logical judgment. However, most models did not do this consistently. They provided the correct logical answer in only about 67–77% of cases, with several performing even lower. Only two models achieved notably higher consistency, reaching 86%–87% on this metric (o3, Gemini 2.5 Pro), highlighting that handling negation remains a substantial weakness for current LLMs.

Finally, examining the standard deviation of each metric across domains (shown in Table 2) reveals a clear pattern: most measures shift noticeably depending on the domain. For bias metrics, the variation can even be large enough for the direction of the effect to reverse entirely, with some domains showing a move from a No-leaning tendency to a Yes-leaning one. In contrast, consistency under repeated questioning and answer option reformulation remains largely strong across models and domains. Overall, these results show that aggregate scores can mask important cross-domain differences, which we examine in more detail in a dedicated section.

Table 2: Overall benchmark results, standard deviation of the metrics (best 2 highlighted)

Model	Performance	Bias		Consistency			
	F1	ΔRecall	ΔPrecision	PRC	NRC	ARC	SC
Baseline	0.01	0.03	0.00	0.03	0.04	0.03	0.03
o3	**0.09**	0.11	0.06	0.06	**0.04**	0.03	0.06
GPT-4o	**0.09**	0.15	0.07	0.07	0.10	0.03	0.02
GPT-4o mini	0.12	0.19	0.07	0.06	0.15	**0.02**	0.02
Gemini 2.5 Pro	**0.09**	**0.03**	**0.03**	**0.04**	**0.05**	0.03	0.04
Gemini 2.5 Flash	0.11	**0.05**	**0.03**	0.07	0.06	0.04	**0.01**
Claude Opus 4.1	**0.09**	0.06	0.04	**0.05**	0.12	**0.02**	0.02
Claude Sonnet 4.5	**0.10**	**0.05**	0.04	**0.05**	0.08	**0.02**	**0.00**
Claude Haiku 4.5	0.11	0.14	0.05	**0.05**	0.13	0.03	**0.00**
Magistral Medium 1.2	**0.09**	0.06	**0.02**	0.09	0.08	0.06	0.08
Mistral Medium 3.1	**0.09**	0.17	0.09	0.07	0.10	0.05	**0.01**
Mistral Small 3.2	**0.09**	0.16	0.06	0.08	0.12	0.03	0.02
Llama 4 Maverick	0.11	0.09	**0.03**	0.07	0.23	0.04	0.06
Llama 3.1 405B	0.11	0.12	0.06	0.09	0.13	**0.01**	0.02
Llama 3.3 70B	0.12	0.18	0.06	0.10	0.07	**0.02**	0.02
Llama 3.1 8B	0.17	0.30	0.12	0.12	0.13	0.05	**0.00**

3.2. Overall Best Models

Taking all the metrics together and looking at the cumulative score, **Claude Opus 4.1** and **Gemini 2.5 Pro** show the best results, with 91.0 and 90.5 MBS points, respectively. In terms of performance, both models achieve the highest overall F1 score on the task (0.88). When it comes to bias, Claude Opus 4.1 shows the better balance between recall and precision (ΔRecall 0.02; ΔPrecision –0.01). Gemini 2.5 Pro exhibits a slightly stronger bias (ΔRecall 0.05; ΔPrecision –0.04) but still maintains a reasonable Yes/No balance in responses. Across the consistency dimensions, the models behave similarly, providing stable responses when asked the same question multiple times (SC), when presented with different answer option formats (ARC), and when given semantically equivalent questions (PRC). Claude Opus 4.1 is somewhat more consistent on these dimensions (SC 0.98 vs. 0.95; ARC 0.98 vs. 0.96; PRC 0.91 vs. 0.90). In contrast, Gemini 2.5 Pro performs markedly better when handling negations (NRC), scoring 0.86 compared to Claude's 0.77. When looking across domains, Gemini 2.5 Pro demonstrates the most uniform performance, showing less variation across the different areas tested. Claude Opus 4.1, although strong overall, exhibits more domain-by-domain variability, particularly in consistency metrics such as NRC (stdev of 0.12), suggesting greater fluctuation in handling negative reformulations across contexts.

Overall, both models are strong choices for binary question answering, delivering reliable and high-quality performance. Claude Opus 4.1 may be the better option when minimizing Yes/No bias is critical, while Gemini 2.5 Pro is preferable for scenarios involving frequent negations or a wider variety of binary question domains.

3.3. Importance of Size and Reasoning Optimization

Figure 1 shows a consistent pattern across all model families: larger models and reasoning-optimized variants achieved higher composite MBS scores. Reasoning-oriented models, such as OpenAI's o3 or Anthropic's Claude

Opus 4.1, appear at the top of the intra-family rankings. This makes sense, given that many questions in the benchmark required a degree of reasoning about the presented information, making models optimized for this better positioned to perform well on the task. Moreover, we can see that model size correlates reliably with performance: the largest models outperform medium and small variants, while the smallest models fall behind. This trend appears across all the families, indicating that it is not an artifact of a single model provider but a general property. Taken together, these results show that both scale and reasoning specialization enhance LLM ability to answer binary questions reliably, and smaller, non-reasoning models are facing systematic limitations on this task.

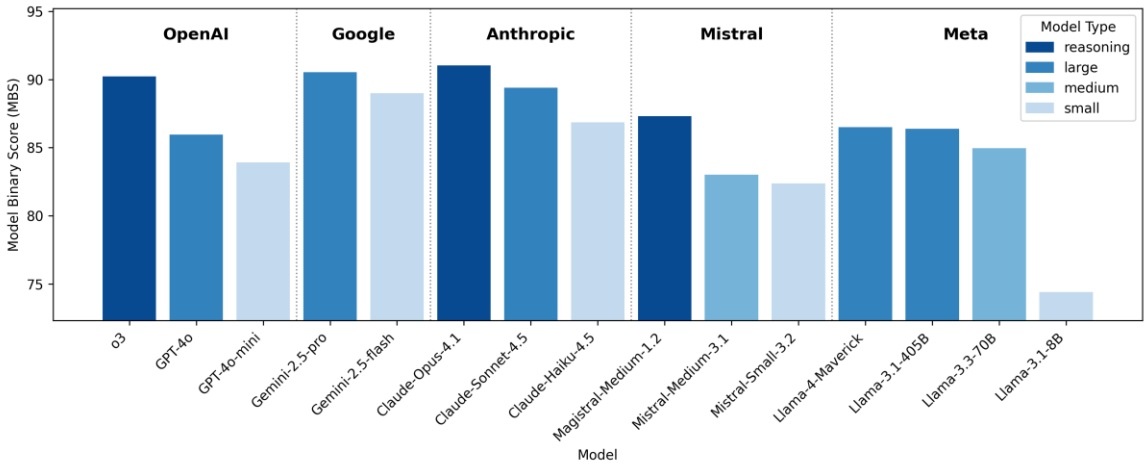

Figure 1: Model Binary Score comparison by LLM family and type

3.4. Cross-domain Breakdown

As seen in Table 3, performance of the LLMs on the binary QA task differs substantially between domains included in our benchmark. The models perform strongly on reading comprehension and multi-hop reasoning. Numerical reasoning shows a mixed but generally solid pattern, with larger and reasoning-optimized models mostly achieving higher F1 scores. Ethics falls into a mid-range band, with the notable exception of closed-source reasoning models (o3 and Claude Opus 4.1), which reach or exceed 0.90 F1. The translation quality evaluation is where the models struggle the most, with F1 scores largely in the high-60s to low-70s range. Overall, the hardest domains ended up being those that require capabilities beyond simple reasoning over the information provided in the prompt. Numerical reasoning adds the need for discrete computational skills on top of logical inference. Ethical questions rely on a sufficiently rich, norm-aligned world model to answer appropriately. Translation error detection calls for additional cross-lingual interpretation abilities and sensitivity to linguistic details. The extra requirements make these domains substantially more challenging than those based on factual retrieval and basic structured reasoning. And, as our analysis reveals, current LLMs are not yet equipped to handle them reliably in binary settings.

Table 3: Cross-domain performance metric, F1
(≥ 0.90 dark green; 0.89-0.80 green; 0.79-0.70 yellow; 0.69-0.60 red; ≤ 0.59 dark red)

Model	Reading	Reason	NumReason	Translation	Ethics
o3	0.91	0.95	0.93	0.71	0.92
GPT-4o	0.91	0.92	0.79	0.66	0.83
GPT-4o mini	0.91	0.89	0.64	0.63	0.78
Gemini 2.5 Pro	0.93	0.95	0.93	0.71	0.85
Gemini 2.5 Flash	0.92	0.95	0.92	0.66	0.84
Claude Opus 4.1	0.91	0.94	0.94	0.71	0.90
Claude Sonnet 4.5	0.91	0.95	0.95	0.68	0.86
Claude Haiku 4.5	0.92	0.95	0.94	0.66	0.78
Magistral Medium 1.2	0.90	0.93	0.91	0.67	0.86
Mistral Medium 3.1	0.90	0.92	0.71	0.73	0.72
Mistral Small 3.2	0.87	0.90	0.73	0.65	0.73
Llama 4 Maverick	0.90	0.93	0.90	0.63	0.82
Llama 3.1 405B	0.91	0.94	0.76	0.65	0.78
Llama 3.3 70B	0.92	0.90	0.87	0.58	0.80
Llama 3.1 8B	0.87	0.85	0.59	0.56	0.46

When it comes to bias, the cross-domain patterns in Table 4 show that most models tend to favor No over Yes, with only a few isolated Yes-leaning exceptions emerging in specific model–domain combinations. The magnitude of this No tendency, however, varies considerably across domains. Translation shows the strongest imbalance, followed by ethics and numerical reasoning. At the same time, some domains, particularly multi-hop reasoning, remain comparatively well balanced. These results suggest that although models generally tend to choose No more often, the strength of this inclination (and, at times, even its existence) is highly domain dependent.

Table 4: Cross-domain bias metrics
(orange: No-leaning; purple: Yes-leaning; white: balanced/minimal difference)

Model	Reading		Reason		NumReason		Translation		Ethics	
	ΔRecall	ΔPrecision	ΔRecall	ΔPrecision	ΔRecall	ΔPrecision	ΔRecall	ΔPrecision	ΔRecall	ΔPrecision
o3	-0.01	0.01	-0.02	0.02	0.12	-0.01	0.27	-0.13	0.02	-0.02
GPT-4o	0.05	-0.04	-0.01	0.01	-0.06	0.03	0.35	-0.14	0.15	-0.10
GPT-4o mini	0.01	-0.01	-0.07	0.06	-0.15	0.04	0.38	-0.12	0.15	-0.09
Gemini 2.5 Pro	0.04	-0.04	-0.01	0.01	0.08	-0.07	0.08	-0.03	0.05	-0.04
Gemini 2.5 Flash	0.04	-0.03	0.01	-0.01	0.11	-0.09	0.14	-0.05	0.06	-0.04
Claude Opus 4.1	0.05	-0.04	0.03	-0.03	0.10	-0.09	-0.07	0.03	-0.02	0.02
Claude Sonnet 4.5	0.02	-0.02	-0.02	0.02	0.10	-0.09	0.02	-0.01	0.11	-0.08
Claude Haiku 4.5	0.04	-0.03	0.01	-0.01	0.06	-0.05	0.38	-0.15	0.19	-0.11
Magistral M 1.2	-0.01	0.01	0.00	0.00	-0.03	0.02	-0.16	0.06	-0.06	0.04
Mistral M 3.1	0.07	-0.06	0.04	-0.03	0.16	-0.07	-0.05	0.02	0.44	-0.25
Mistral Small 3.2	0.10	-0.07	0.05	-0.04	-0.02	0.01	0.44	-0.19	0.17	-0.08
Llama 4 Maverick	-0.03	0.03	0.00	0.00	0.02	-0.01	0.23	-0.07	0.00	0.00
Llama 3.1 405B	0.06	-0.05	-0.01	0.01	-0.20	0.11	-0.12	0.04	0.14	-0.07
Llama 3.3 70B	-0.04	0.03	-0.01	0.01	0.07	-0.05	0.47	-0.13	0.18	-0.11
Llama 3.1 8B	0.10	-0.08	0.01	-0.01	0.11	-0.02	0.34	-0.05	0.85	-0.34

For the cross-domain consistency analysis, we included only the positive (PRC) and negative (NRC) reformulation aspects. As the LLMs showed high performance with low variance in self- (SC) and answer reformulation (ARC) consistency, these dimensions were less valuable to review in detail. Table 5 shows clear domain-dependent patterns in positive reformulation consistency (PRC) that align with earlier per-domain findings. The models are most stable in reading and multi-hop reasoning, where they tend to return the same answer even when exposed to different paraphrased versions of the same question. Stability drops in numerical reasoning, translation evaluation, and ethical judgment, though a few models stand out as exceptions (e.g., Mistral Medium 3.1 on translation). For NRC, we have already established the overall difficulty that the models face with handling negations. Interestingly, the table shows that consistency under this type of reformulation varies more by model than by domain. Moreover, in many cases, a domain that one model handles the best, another model handles the worst. For example, Gemini 2.5 Flash has the highest NRC of 0.90 in numerical reasoning, while for GPT-4o mini this is the most challenging domain with NRC of 0.41. This suggests that negative reformulation consistency depends less on domain properties and more on model-specific strengths and weaknesses.

To consolidate all the cross-domain findings, Table 6 summarizes the general trends we observed across LLMs. On reading comprehension and multi-hop reasoning, the LLMs consistently achieve the strongest results across criteria, while on ethical judgment they show mid-level performance, and struggle on translation evaluation. Numerical reasoning displays substantial variability, with outcomes highly model dependent. Negative reformulation consistency is excluded from this summary, as performance on this metric was uniformly weak across models and did not reveal meaningful domain-level distinctions.

Table 5: Cross-domain consistency metrics
(≥ 0.90 dark green; 0.89-0.80 green; 0.79-0.70 yellow; 0.69-0.60 red; ≤ 0.59 dark red)

Model	Positive Consistency (PRC)					Negative Consistency (NRC)				
	Reading	Reason	NumReason	Translation	Ethics	Reading	Reason	NumReason	Translation	Ethics
o3	0.92	0.98	0.94	0.81	0.87	0.92	0.86	0.92	0.83	0.84
GPT-4o	0.92	0.96	0.76	0.88	0.86	0.82	0.56	0.58	0.64	0.74
GPT-4o mini	0.92	0.92	0.78	0.80	0.84	0.81	0.74	0.41	0.59	0.77
Gemini 2.5 Pro	0.90	0.97	0.88	0.85	0.88	0.93	0.85	0.92	0.80	0.81
Gemini 2.5 Flash	0.89	0.96	0.91	0.77	0.84	0.84	0.83	0.90	0.73	0.81
Claude Opus 4.1	0.91	0.97	0.94	0.84	0.89	0.76	0.84	0.87	0.84	0.55
Claude Sonnet 4.5	0.88	0.97	0.94	0.84	0.88	0.69	0.57	0.78	0.79	0.62
Claude Haiku 4.5	0.92	0.94	0.90	0.80	0.85	0.79	0.76	0.70	0.85	0.47
Magistral Med 1.2	0.90	0.94	0.81	0.68	0.80	0.83	0.71	0.86	0.65	0.77
Mistral Med 3.1	0.92	0.94	0.76	0.96	0.88	0.64	0.56	0.35	0.50	0.40
Mistral Small 3.2	0.92	0.91	0.77	0.72	0.84	0.75	0.62	0.54	0.38	0.61
Llama 4 Maverick	0.89	0.97	0.85	0.76	0.89	0.87	0.74	0.80	0.23	0.70
Llama 3.1 405B	0.90	0.96	0.81	0.70	0.89	0.74	0.81	0.55	0.49	0.74
Llama 3.3 70B	0.90	0.93	0.80	0.64	0.83	0.85	0.78	0.62	0.76	0.78
Llama 3.1 8B	0.87	0.84	0.67	0.64	0.94	0.65	0.40	0.27	0.33	0.43

Table 6: Cross-domain LLM behavior summary

	Performance	Bias	Consistency (except NRC)
Reading comprehension	good	slight No-bias	good
Multi-hop reasoning	good	balanced	good
Numerical reasoning	mixed	No-bias	mixed
Translation evaluation	poor	No-bias	medium
Ethical judgement	medium	No-bias	medium-good

4. Discussion

4.1. Discussion of the Results

This study set out to provide a new multi-dimensional assessment of how contemporary large language models handle binary question answering. Our evaluation of 15 latest models across five domains shows that while modern LLMs are generally competent in binary QA, this competence is accompanied by non-negligible patterns of No-leaning bias and inconsistency, especially under negation. This indicates that while LLM can be useful in binary QA settings, they should be applied to the task with caution.

Our findings complement and extend recent studies of binary bias in LLMs. Lu and colleagues [4], Braun [5], and Cheung and colleagues [6] provide empirical evidence of No-leaning behavior of some LLMs in value judgment, sentiment analysis, legal texts, and moral decision-making, while Yu and colleagues [3] highlight "negative attention heads" as a possible architectural explanation for this phenomenon. The present study replicates these observations at a larger scale, spanning more domains and models, and showing that this No bias is not limited to specific topics or model families but is a widespread feature of current LLM behavior. At the same time, by showing that most systems achieve 0.80+ F1 score on binary QA, we demonstrated that this bias co-exists with strong overall task performance rather than arises from general incompetence. In addition, the consistency dimension of our framework situates binary QA within the broader landscape of LLM instability documented in earlier work. Prior research highlighted various LLM consistency issues, such as failure to preserve logical relations across paraphrases [8] and meaning reversals [9], sensitivity to prompt formatting [11], and, though less common, tendency to provide different answers when the same query is repeated [10]. Our consistency analysis advances this line of inquiry by systematically measuring these instabilities specifically in the context of binary QA. We find that models are almost perfectly stable under repetition and answer format (order) changes, moderately stable under paraphrasing, but quite fragile when handling negative reformulations. These results reaffirm earlier findings while also demonstrating meaningful progress: in binary QA, contemporary LLMs now handle repetitions and formatting changes with high reliability yet remain sensitive to rephrasing and, especially, negations.

Importantly, our cross-domain analysis also highlights that binary QA is not a single, uniform skill. The evaluated models handle reading comprehension and multi-step reasoning well, while numerical reasoning questions and ethical decisions are harder but still manageable. Translation error detection stands out as the weakest area, suggesting that cross-lingual evaluation places additional strain on model capabilities and urgently needs further LLM improvements. Notably, across domains, greater task difficulty seems to amplify all model weaknesses at

once: the harder the task and the lower the performance, the stronger the No-leaning bias and the more frequent the reformulation inconsistencies. All in all, these differences make clear that LLM behavior on binary QA must be evaluated in the context of specific domains rather than assumed to reflect a universal level of competence.

Finally, our benchmark has revealed that, when evaluated holistically, reasoning-optimized and larger models deliver the strongest overall binary QA performance. Their consistently higher composite scores point to a meaningful advantage in decision reliability, reinforcing the idea that scale and specialized reasoning training are key drivers of dependable binary QA behavior, especially in reasoning-dependent contexts.

Taken together, these results have several implications for practitioners who rely on LLMs for binary decision support:

- **Treat No as a conservative default.** As most models show higher recall but lower precision on No class, this indicates that No tends to be over-produced relative to the true label distribution. This implies that false positives (false Yes) in binary QA may be relatively rare, but false negatives (false No) will be more common. System designers should therefore treat No as an indication of caution and consider downstream workflows that allow for human verification when No comes with a high cost of error (e.g., application or health screenings).
- **Be cautious with negative phrasings.** Our negative reformulation consistency metric shows that models often fail to flip their answer consistently when the question meaning is inverted through, for example, use of negation words and expressions. For system designers who control how binary prompts are formulated, it is safer to avoid constructions that rely on negation, as these increase the likelihood of inconsistent answers. When negative wording truly cannot be avoided, choosing models with stronger negation consistency (e.g. o3, Gemini 2.5 Pro) becomes especially important.
- **Prefer larger and/or reasoning-optimized models for critical binary decisions.** The association between model size, reasoning optimization, and composite binary score suggests that, where resources permit, larger and/or reasoning-tuned models should be used for high-impact binary question answering tasks. Smaller models, while attractive for cost and latency reasons, often show both lower F1 and amplified biases and inconsistency. A practical deployment strategy is to use a lightweight model for triage: determining whether the input is genuinely a binary decision task and then routing confirmed binary questions to a larger, more reliable model. This approach balances efficiency with accuracy while reducing the risk that sensitive decisions are made by weaker models.
- **Benchmark models for your specific scenarios before deployment.** Reading comprehension and multi-hop reasoning questions in our benchmark were handled reliably by many models, while translation evaluation and ethical judgement questions were marked by far weaker performance and more inconsistencies. Practitioners should avoid assuming that a model's results on one binary task or specific dataset will transfer to others. Whenever possible, model selection should be based on domain-specific evaluation, ideally using a framework similar to ours that explicitly measures performance, bias, and consistency.

4.2. Limitations and Future Work

Several limitations of our study should be acknowledged. First, although our dataset was sufficiently large to draw meaningful conclusions about LLM behavior, it was still modest compared to the diversity of real-world binary decision problems. This limitation applies both to the depth of coverage within each domain and to the overall breadth of domains represented. Because we observe substantial cross-domain variation in performance, bias, and consistency, a broader and more fine-grained dataset could reveal additional patterns not captured in the current benchmark. The second limitation concerns the transformations that were required to create the benchmark dataset. Some examples (particularly in the numerical reasoning domain) needed manual rewriting, while our positive and negative reformulations relied on LLM-generated paraphrases. Although each step was standardized and carefully reviewed, any transformation process can introduce subtle shifts in difficulty or ambiguity. As a result, the dataset might have contained small inconsistencies in complexity across items. However, given the scale of the dataset and the rigorous human review checks implemented during its generation, these effects are unlikely to have been large enough to alter the overall trends we observed. Finally, although we tested a wide range of current models, the LLM landscape is evolving quickly. New releases may show different behavior, so the results presented here should be viewed as a snapshot of present capabilities rather than a definitive ranking.

Our benchmark also points to several clear directions for future work. First, expanding it to additional domains (such as clinical decision-making or legal compliance) would help test whether the binary QA patterns observed in this study generalize to other settings. A second avenue is multilingual and cross-lingual evaluation. Our findings in translation evaluation suggest that reasoning across languages poses unique challenges, so building a fully multilingual binary QA benchmark would allow us to further examine model behavior on the task in multilingual contexts, potentially revealing new important model weaknesses. Beyond broader coverage, adding uncertainty measures to the benchmark could provide a deeper insight into LLM consistency behavior. It would be useful to see if the areas of greater instability align with points of increased model uncertainty, thereby clarifying whether inconsistencies reflect areas of LLM ambiguity or indicate deeper reasoning limitations. Future work should also move from diagnosing the issues to mitigating them, involving both training and inference time strategies, and focusing on the most problematic areas first (i.e., consistency under negation). This extension would help translate this benchmark from an analytic tool into a foundation for building more robust binary decision systems.

5. Conclusion

In this paper, we introduced a new dedicated benchmark for comprehensively evaluating large language models on the binary question answering task. It combines performance, bias, and consistency into a single, interpretable score and is supported by a five-domain dataset augmented with systematic reformulations. Applied to fifteen latest models, this benchmark shows that, overall, LLMs are good binary reasoners, with larger and reasoning-optimized models reliably outperforming smaller ones. Yet, LLMs' competence is accompanied by a No-leaning bias, fragility under negative question reformulations, and clear domain-dependent variations (with especially pronounced weaknesses in translation evaluation).

These findings have direct implications for the design of LLM-based decision systems. In practice, No should be treated as a conservative default; prompts should avoid negative phrasings; and models should be selected and optimized on a per-domain basis, with a preference for larger and reasoning-optimized ones in critical binary contexts. By sharing the benchmark, we aim to provide a reusable framework for tracking progress in improving LLM performance and tackling bias and inconsistency in binary question answering settings.

References

[1] Ni, S., Chen, G., Li, S., Chen, X., Li, S., Wang, B. et al. "A Survey on Large Language Model Benchmarks," arXiv:2508.15361v1 [cs.CL], Aug. 2025.

[2] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. "BoolQ: Exploring the Surprising Difficulty of Natural Yes/No questions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers)*, Jun. 2019, pp. 2924-2936.

[3] S. Yu, J. Song, B. Hwang, H. Kang, S. Cho, J. Choi et al. "Correcting Negative Bias in Large Language Models through Negative Attention Score Alignment," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Apr. 2025, pp. 9979-10 001.

[4] Y.-L. Lu, C. Zhang, and W. Wang. "Systematic Bias in Large Language Models: Discrepant Response Patterns in Binary vs. Continuous Judgment Tasks," arXiv:2504.19445v1 [cs.CL], Apr. 2025.

[5] D. Braun. "Acquiescence Bias in Large Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Nov. 2025, pp. 11 341-11 355.

[6] V. Cheung, M. Maier, and F. Lieder. "Large language models show amplified cognitive biases in moral decision- making," in *Proceedings of the National Academy of Sciences*, vol. 122, no. 25, Jun. 2025.

[7] M. Jang, D. S. Kwon, and T. Lukasiewicz. "BECEL: Benchmark for Consistency Evaluation of Language Models," in *Proceedings of the 29th International Conference on Computational Linguistics*, Oct. 2022, pp. 3680-3696.

[8] M. Jang and T. Lukasiewicz. "Consistency Analysis of ChatGPT," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Dec. 2023, pp. 15 970-15 985.

[9] J. J. Ahn and W. Yin. "Prompt-Reverse Inconsistency: LLM Self-Inconsistency Beyond Generative Randomness and Prompt Paraphrasing," arXiv:2504.01282v2 [cs.CL], Jul. 2025.

[10] B. Atil, S. Aykent, A. Chittams, L. Fu, R. J. Passonneau, E. Radcliffe, et al. "Non-Determinism of 'Deterministic' LLM Settings," arXiv:2408.04667v5 [cs.CL], Apr. 2025.

[11] T. Labruna, S. Gallo, and G. D. S. Martino. "Positional Bias in Binary Question Answering: How Uncertainty Shapes Model Preferences," arXiv:2506.23743v2 [cs.CL], Jul. 2025

[12] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, et al. "HotpotQA: A Dataset

for Diverse, Explainable Multi-hop Question Answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2018, pp. 2369-2380.
[13] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs," in *Proceedings of the 2019 Conference of the North American Chapter of the Association of Computational Linguists: Human Language Technologies (Volume1: Long and Short Papers)*, Jun. 2019, pp. 2368-2378.
[14] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey. "Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation," in *Transactions of the Association for Computational Linguistics (Volume* 9), 2021, pp. 1460-1474.
[15] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, et al. "Aligning AI With Shared Human Values," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
[16] OpenAI. "Introducing GPT-4.1 in the API." Internet: https://openai.com/index/gpt-4-1/, Apr. 2025 [Nov. 7, 2025].
[17] OpenAI. "OpenAI o3 and o4-mini System Card." Internet: https://openai.com/index/o3-o4-mini-system-card/, Apr. 2025 [Nov. 19, 2025].
[18] OpenAI. "GPT-4o System Card." Internet: https://openai.com/index/gpt-4o-system-card/, Aug. 2024 [Nov. 7, 2025].
[19] OpenAI. "GPT-4o mini: advancing cost-efficient intelligence." Internet: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, Jul. 2024 [Nov. 8, 2025].
[20] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon et al. "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities," arXiv:2507.06261v5 [cs.CL], Oct. 2025.
[21] Anthropic. "Claude Opus 4.1." Internet: https://www.anthropic.com/news/claude-opus-4-1, Aug. 2025 [Nov. 15, 2025].
[22] Anthropic. "Introducing Claude Sonnet 4.5." Internet: https://www.anthropic.com/news/claude-sonnet-4-5, Sep. 2025 [Nov. 19, 2025].
[23] Anthropic. "Introducing Claude Haiku 4.5." Internet: https://www.anthropic.com/news/claude-haiku-4-5, Oct. 2025 [Nov. 13, 2025].
[24] Mistral-AI: A. Rastogi, A. Q. Jiang, A. Lo, G. Berrada, G. Lample, J. Rute et al. "Magistral," arXiv:2506.10910v1 [cs.CL], Jun. 2025.
[25] Mistral-AI. "Medium is the new large." Internet: https://mistral.ai/news/mistral-medium-3, May 2025 [Nov. 2, 2025].
[26] Mistral-AI. "Mistral Small 3.2." Internet: https://docs.mistral.ai/models/mistral-small-3-2-25-06, Jun. 2025. [Nov. 2, 2025].
[27] Meta-AI. "The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation." Internet: https://ai.meta.com/blog/llama-4-multimodal-intelligence/, Apr. 2025 [Nov. 20, 2025].
[28] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, et al. "The Llama 3 Herd of Models," arXiv:2407.21783v3 [cs.AI], Nov. 2024.