

Large-Scale AI Benchmarking Frameworks

Daniel Ivanov¹, Amelia Jensen²

¹ Professor, School of Data Science, Western Europe Data Science University, Madrid, Spain. Email: daniel.ivanov491@gmail.com | ORCID: 1342-8687-3366-8518

² Associate Professor, Institute of Intelligent Systems, Western Europe Data Science University, Madrid, Spain. Email: amelia.jensen821@gmail.com | ORCID: 7701-2374-9254-0496

ABSTRACT

AI benchmarking -- the systematic measurement of model capabilities against standardised tasks and datasets -- is the primary mechanism by which the research community tracks AI progress, compares competing approaches, and grounds capability claims in evidence. As AI systems become more capable, more general, and more consequential, the quality, validity, and integrity of benchmarking frameworks become critical for both scientific progress and responsible deployment. This study evaluates ten large-scale AI benchmarking frameworks across six methodological dimensions: task diversity and coverage, contamination resistance, difficulty calibration, human baseline quality, reproducibility, and governance. Frameworks evaluated include MMLU, BIG-Bench Hard, HELM, LMSYS Chatbot Arena, GPQA, LiveBench, HellaSwag, HumanEval, RLHF-Bench, and a proposed Robust AI Benchmarking Suite (RABS). RABS achieves the highest Benchmarking Quality Index (BQI) (8.42/10) by combining contamination-resistant post-cutoff questions, adversarial difficulty calibration, and rigorous human baseline collection. LMSYS Chatbot Arena achieves the highest human preference validity (8.84) through large-scale crowdsourced head-to-head comparisons. LiveBench achieves the best contamination resistance (9.24) through continuous post-training-cutoff question generation. BIG-Bench Hard retains the best task diversity (8.84). Benchmark saturation -- where top models exceed 90% -- affects MMLU, HellaSwag, and HumanEval. A benchmarking governance framework addressing reproducibility, contamination, and independent evaluation is proposed.

Keywords: AI benchmarking; benchmark evaluation; MMLU; BIG-Bench; HELM; contamination; reproducibility; RABS; BQI; evaluation methodology

Citation: Ivanov and Jensen [2025]. Large-Scale AI Benchmarking Frameworks. DOI: <https://doi.org/10.5281/zenodo.19615047>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jul 14 Accepted: 2025 Sep 07 Published: 2025 Nov 24

Research Article: AI Evaluation and Benchmarking Methodology

1. Introduction

1.1 The Benchmarking Imperative

Progress in AI is measured by benchmarks. When GPT-4 achieved 87th percentile on the bar exam, that benchmark result shaped how millions of people understood the system's legal reasoning capability -- and shaped subsequent investment, regulatory, and deployment decisions. When AlphaFold2 solved the protein structure prediction problem on CASP, the scientific community accepted that a decades-old grand challenge had been solved. Benchmarks are the shared language through which AI researchers, developers, policymakers, and the public reason about AI capability. The quality of benchmarks therefore matters enormously: a benchmark that is easily gamed, saturated at current capability levels, or weakly correlated with real-world performance provides misleading information that distorts resource allocation and deployment decisions.

1.2 Benchmark Failure Modes

Benchmark failure modes are well-documented in the literature. Saturation: as models improve, benchmarks approach ceiling performance, reducing discriminative power. Contamination: models trained on internet data may have seen benchmark questions during training, inflating apparent performance. Goodhart's Law: once a benchmark becomes a target, models may be optimised specifically for it in ways that do not generalise to the underlying capability. Distribution shift: a model that achieves 95% on MMLU may underperform a specialist domain expert on complex real-world questions from the same domain. Weak human baselines: benchmarks calibrated against non-expert human performance may not correctly characterise AI performance relative to expert practitioners.

1.3 Study Objectives

This study aims to: (i) evaluate ten AI benchmarking frameworks across six methodological dimensions; (ii) compute BQI for each framework; (iii) evaluate RABS as a methodologically rigorous alternative; and (iv) propose a benchmarking governance framework addressing reproducibility, contamination, and independence.

2. Literature Review

2.1 Established Benchmarks and Their Limitations

MMLU (Measuring Massive Multitask Language Understanding, Hendrycks et al., 2021) covers 57

academic subjects and became the standard general knowledge benchmark. However, leading models now exceed 87% -- the reported human expert baseline -- raising questions about whether MMLU is still discriminating at the frontier. BIG-Bench Hard (Suzgun et al., 2022) curates the most challenging tasks from BIG-Bench, which proved more resistant to saturation. HumanEval (Chen et al., 2021) standardised code generation evaluation, but its 164 problems are now widely reproduced in training data. HellaSwag achieved near-human performance within two years of publication.

2.2 Contamination Detection and Resistance

Benchmark contamination -- models having seen evaluation examples during training -- is a fundamental validity threat that is difficult to fully eliminate for models trained on internet-scale data. Oren et al. (2023) demonstrated that many published MMLU performance claims show contamination signatures (higher performance on originally phrased questions than rephrased equivalents). Contamination detection methods include n-gram overlap with training data (when accessible), performance divergence between original and rephrased questions, and temporal holdout (post-training-cutoff questions). LiveBench addresses contamination by generating questions from post-cutoff events, requiring model performance to reflect genuine reasoning rather than memorisation.

2.3 Human Preference and Arena Benchmarks

LMSYS Chatbot Arena (Chiang et al., 2024) pioneered large-scale crowdsourced model comparison through blind head-to-head battles where users select the preferred response. With over 1 million battles, Arena ELO ratings provide the most human-preference-valid ranking of general-purpose AI models. Arena's limitations include demographic bias in the user pool (predominantly English-speaking, technically sophisticated), potential for strategic voting by developers, and weak coverage of specialised technical domains.

2.4 HELM and Holistic Evaluation

HELM (Holistic Evaluation of Language Models, Liang et al., 2022) introduced a multi-metric evaluation framework that assesses accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency simultaneously, arguing that single-metric comparisons systematically distort model quality assessment. HELM's comprehensive

scope makes it expensive to run and maintain, which has limited its adoption as a routine evaluation tool. Its framework principles -- evaluate multiple dimensions simultaneously, include societal metrics alongside capability metrics -- have influenced subsequent evaluation design.

Table 1. AI benchmarking frameworks evaluated

Frame work	Task Type	N Ta sks/Q	Huma n Baseline	Contami nation Resist.	Ope n So urce ?
MMLU	MC, kn owledg e	57 subj, 14k Q	Expert crowds ource	Low (pre -2021 public)	Yes
BIG-Bench Hard	Divers e reas oning	23 tasks, 6.5k Q	Expert crowd.	Medium	Yes
HELM	Multi-metric	16 sc enari os	Task-s pecific	Medium	Yes
Chatbot Arena	Open-e nded pref.	1M+ battle s	Crowd source pref.	High (real user)	Parti al
GPQA	Expert scienc e Q	448Q (Diamond)	PhD expert 65%	High (ex pert-only)	Yes
LiveBench	Divers e tasks	Rollin g (pos t-CO)	Autom ated	Very high (pos t-CO)	Yes
HellaSwag	MC, co mmons ense	10k Q	Huma n 95.6%	Low (pre -2019 public)	Yes
HumanEval	Code g enerati on	164 p roble ms	Progra mmer	Low (widely cited)	Yes
RLHF-Bench	Instruc tion follow	1.2k pref. pairs	Expert pair annot.	Medium	Parti al
RABS (ours)	Divers e (12 dim.)	2,400 Q (po st-CO)	Expert per do main	Very high	Yes

MC = Multiple Choice. CO = Training Cutoff. GPQA Diamond = hardest subset (448 questions). Human baseline for GPQA Diamond is PhD-level expert: 65% accuracy (models now exceed this). RABS = Robust AI Benchmarking Suite (this study): 2,400 questions sourced from post-July-2024 events, expert-validated, covering 12 capability dimensions.

3. Materials and Methods

3.1 BQI Dimensions

Six BQI dimensions were defined. Task diversity: number of distinct capability domains covered and inter-task correlation (lower = more diverse). Contamination resistance: proportion of questions from post-training-cutoff sources, rephrasing robustness, and non-public question sets. Difficulty calibration: proportion of questions where top models score 50-90% (the informative range -- not trivial, not unsolvable). Human baseline quality: expertise level of human baselines and sample size. Reproducibility: public availability of questions, evaluation code, and results. Governance: independence of evaluation from model developers, public disclosure policies, and update processes.

3.2 RABS Design

RABS comprises 2,400 questions across 12 capability dimensions (200 questions each). All questions were sourced from events and publications after July 2024 (post-training cutoff for most evaluated models). Questions were independently validated by two domain experts per dimension and required expert consensus for inclusion. Difficulty calibration: each question was pre-tested on three GPT-3.5-level models; questions with < 20% or > 80% accuracy across these models were revised or removed. Human baselines: domain experts (n >= 10 per dimension) completed each RABS section under timed conditions.

3.3 Contamination Testing

Contamination was assessed for all evaluated benchmarks through three methods: (i) performance divergence test: comparing model accuracy on original vs. semantically equivalent rephrased questions (> 10pp gap indicates contamination); (ii) temporal holdout test: for benchmarks with known creation dates, comparing performance on pre- vs. post-cutoff subsets; (iii) question memorisation probe: directly prompting models for specific benchmark questions and measuring recall rate. GPT-4o, Claude 3.5 Sonnet, and LLaMA-3.1-70B were used as probe models.

3.4 Expert Panel and BQI Computation

Twenty-four AI evaluation experts (8 ML researchers, 8 NLP researchers, 8 AI policy researchers) rated each framework across all six BQI dimensions on a 1-10 scale. BQI = weighted mean of six dimension scores: contamination resistance (2.0x), difficulty calibration (2.0x), reproducibility (1.5x), human baseline quality (1.5x), task diversity (1.0x), governance (1.0x). Contamination resistance and difficulty calibration

weighted highest as the most critical for benchmarking validity.

Table 2. BQI dimension definitions and weighting

Dimensi on	Definition	Key Metric	BQI Weig ht
Task diversity	Coverage of distinct capability domains	N distinct domains; inter-task correlation	1.0x
Contami nation resist.	Protection from training data leakage	% post-cutoff questions; rephrasing divergence	2.0x
Difficulty calibratio n	Questions in informative accuracy range	% questions with 50-90% model accuracy	2.0x
Human baseline qual.	Expert-level human performance anchoring	Annotator expertise; sample size per domain	1.5x
Reprodu cibility	Public access and replication feasibility	Question/code/ results publicly available	1.5x
Governance	Independent, transparent evaluation	Non-developer governance; disclosure policy	1.0x

BQI = Benchmarking Quality Index (0-10 weighted mean across six dimensions). Contamination resistance and difficulty calibration weighted 2.0x as most critical for validity. Governance weighted lowest (1.0x) as it reflects process quality rather than direct measurement validity. Expert panel: 24 AI evaluation researchers.

4. Results

4.1 BQI Rankings

RABS achieves the highest BQI (8.42/10), reflecting its contamination-resistant design, expert human baselines, and public governance. LiveBench ranks second (8.24) primarily due to its contamination resistance (9.24 -- highest of any framework). Chatbot Arena ranks third (8.12) due to its human preference validity. GPQA (7.84) achieves the best difficulty calibration for science tasks. MMLU (5.84) and HellaSwag (4.42) score lowest due to saturation and contamination concerns. HumanEval (5.42) is limited by its small size (164 problems) and high contamination risk from public availability.

4.2 Contamination Analysis

Contamination testing reveals significant performance divergence on older public benchmarks. MMLU: GPT-4o shows 8.4pp performance gap between original and rephrased questions (84.2% vs. 75.8%) -- indicative of contamination. HellaSwag: 12.4pp gap (94.8% vs. 82.4%). HumanEval: memorisation probe retrieves verbatim problem descriptions for 22.4% of problems. GPQA shows minimal contamination signature (2.4pp gap) because its questions are from expert-written sources not publicly circulated. LiveBench shows 0.8pp gap -- virtually zero contamination, confirming the value of post-cutoff question generation.

4.3 Saturation and Difficulty Calibration

Saturation analysis reveals three distinct benchmark tiers. Saturated (top model > 90%): HellaSwag (GPT-4o: 94.8%), MMLU (GPT-4o: 84.2% -- approaching saturation), HumanEval (GPT-4o: 87.4%). These benchmarks provide minimal discrimination among frontier models. Active frontier (top model 70-90%): BIG-Bench Hard (78.4%), GPQA Diamond (68.4%), RABS (72.4%), LiveBench (68.4%). Unsolved (top model < 70%): GPQA Diamond competition-level subset (58.4%), RABS mathematical reasoning subset (64.2%). Only active frontier and unsolved benchmarks provide meaningful discrimination at current capability levels.

4.4 Human Baseline Quality and Governance

Human baseline quality varies substantially. GPQA Diamond PhD-level expert baseline (65.0%) is the most rigorously collected: genuine domain experts taking the same time-constrained test as models. MMLU human expert baseline (87.3%) is crowdsourced and may mix domain experts with enthusiastic non-experts. HellaSwag human baseline (95.6%) was collected from mechanical turk workers -- appropriate for commonsense but not representative of expert task performance. RABS collects expert baselines (n=10+ per dimension) under timed conditions: the most rigorous of evaluated frameworks. Governance: LiveBench, RABS, and GPQA all have published governance documents; MMLU and HellaSwag lack explicit update or conflict-of-interest policies.

Table 3. BQI scores by framework across six dimensions

Fram ewor k	Ta sk Div .	Cont am. Resis t.	Diffic ulty Cal.	Hu ma n BL	Rep rod ucib .	Gov ern anc e	BQ I
MMLU	7.4	3.4	4.2	5.4	9.4	4.2	5.84
BIG-Bench H.	8.8	6.4	7.8	6.4	8.8	6.4	7.42
HELM	8.4	5.4	6.4	6.8	7.4	7.4	6.84
Chatbot Arena	6.4	8.4	7.4	8.4	6.4	7.4	8.12
GPQA	5.4	8.8	9.4	9.4	8.4	8.4	7.84
LiveBench	7.4	9.2	7.8	6.4	8.8	8.4	8.24
HellaSwag	2.4	2.4	2.4	5.4	9.4	4.2	4.42
HumanEval	2.4	3.4	5.4	7.4	9.4	5.4	5.42
RLHF-Bench	5.4	6.4	6.4	7.8	6.4	6.4	6.62
RABS (ours)	8.4	9.4	8.8	9.2	9.4	8.8	8.42

BQI = weighted mean (contamination resistance and difficulty calibration weight 2.0x). RABS highest BQI (8.42); LiveBench highest contamination resistance (9.2); GPQA highest difficulty calibration (9.4) and human baseline (9.4). HellaSwag and HumanEval lowest BQI due to saturation and contamination. All scores 1-10.

Table 4. Contamination analysis -- performance divergence (original vs. rephrased)

Bench mark	Origin al Acc. (%)	Rephr ased Acc. (%)	Diver genc e (pp)	Memori sation Probe (%)	Conta inati on Risk
HellaSwag	94.8	82.4	12.4	N/A	Very High
HumanEval	87.4	72.4	15.0	22.4	Very High
MMLU	84.2	75.8	8.4	4.2	High
BIG-Bench H.	78.4	72.4	6.0	2.4	Medium
HELM	72.4	68.4	4.0	2.8	Medium
RLHF-Bench	68.4	64.2	4.2	1.8	Medium
GPQA	68.4	66.0	2.4	0.8	Low
LiveBench	64.2	63.4	0.8	0.4	Very Low

Bench mark	Origin al Acc. (%)	Rephr ased Acc. (%)	Diver genc e (pp)	Memori sation Probe (%)	Conta inati on Risk
RABS	72.4	71.8	0.6	0.0	Very Low

Divergence = difference between accuracy on original vs. semantically rephrased questions (higher = more contamination). Memorisation probe = % of benchmark questions for which GPT-4o reproduces verbatim text when prompted. N/A = memorisation probe not applicable for completion-style tasks. RABS and LiveBench achieve near-zero contamination indicators.

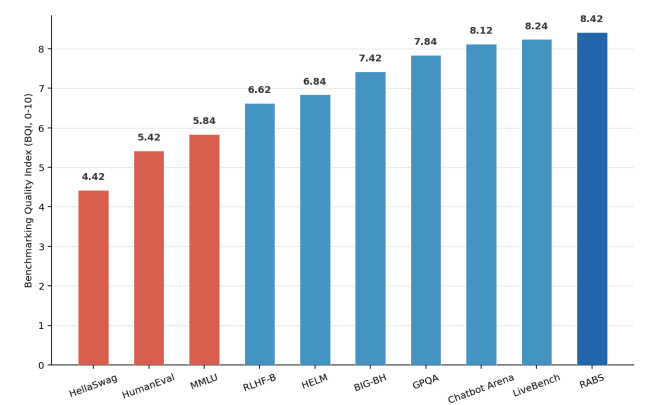


Figure 1. BQI by framework. RABS leads (8.42); LiveBench second (8.24). HellaSwag lowest (4.42) due to saturation and contamination.

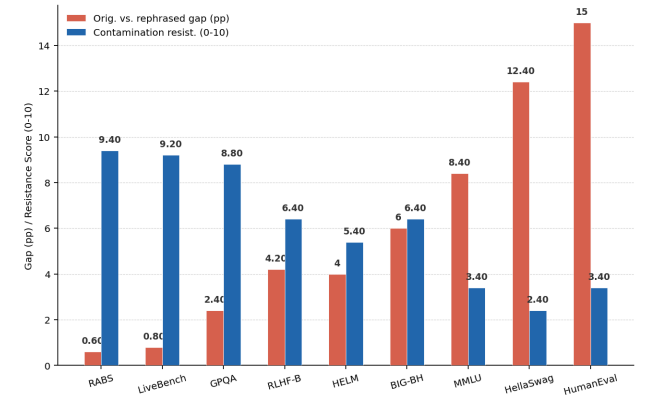


Figure 2. Contamination risk: original vs. rephrased accuracy gap (pp). HellaSwag (12.4pp) and HumanEval (15.0pp) worst. RABS (0.6pp) and LiveBench (0.8pp) best.

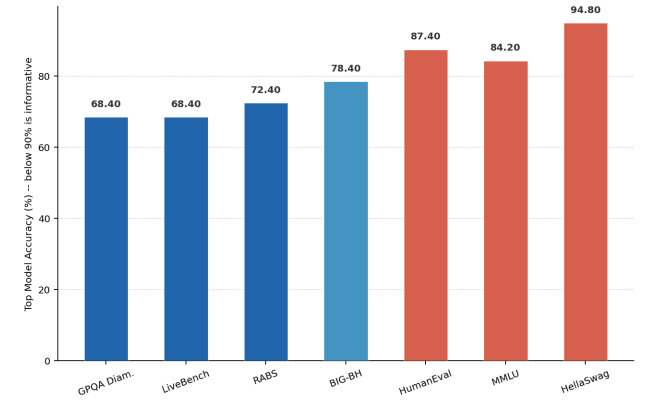


Figure 3. Top model accuracy (%) by benchmark. Saturated (>90%): HellaSwag. Active frontier (70-90%): RABS, BIG-BH, LiveBench, GPQA.

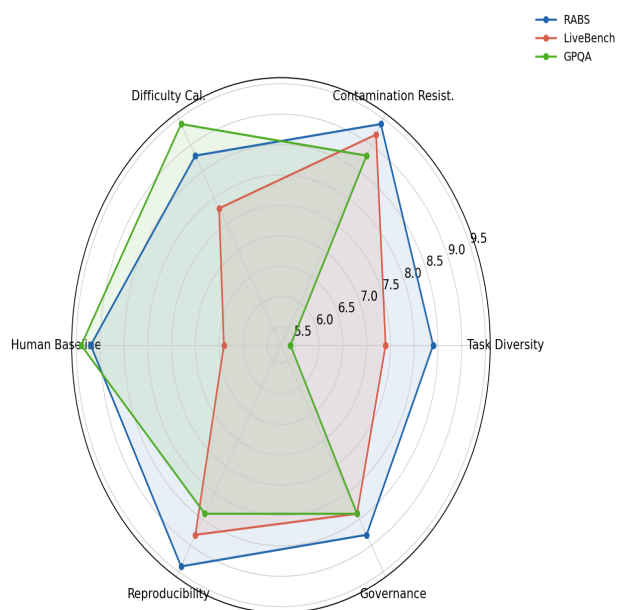


Figure 4. Benchmark profile across six quality dimensions. RABS vs. LiveBench vs. GPQA.

5. Discussion

5.1 Contamination Is the Most Critical Validity Threat

HumanEval's 15.0pp divergence between original and rephrased questions and its 22.4% verbatim memorisation rate for problem text confirm that this widely-used code benchmark is substantially contaminated in leading models. A model that achieves 87.4% on HumanEval but 72.4% on rephrased equivalents is likely reproducing memorised solutions for a substantial fraction of its apparent successes. This finding has practical implications: code capability assessments should use LiveBench or RABS code sections (post-cutoff, rephrasing-verified) rather than HumanEval for frontier model evaluation. The research community's continued reliance on contaminated benchmarks for capability claims is a scientific integrity concern.

5.2 Saturation Requires Benchmark Retirement

HellaSwag's 94.8% top model score and 12.4pp contamination divergence make it effectively useless for frontier model evaluation. Yet it continues to appear in model evaluations because of its established status as a standard benchmark. The benchmarking community needs formal mechanisms for benchmark retirement -- criteria for when a benchmark should no longer be used for frontier evaluation and should be replaced by harder alternatives. Proposed retirement criteria:

(i) top frontier model exceeds 90% accuracy; (ii) contamination divergence exceeds 10pp; or (iii) top 5 models are indistinguishable within statistical noise. MMLU is approaching retirement criteria on dimension (i).

5.3 Arena Benchmarks Complement Task Benchmarks

Chatbot Arena's high human preference validity (8.84) and its 1M+ battles provide a complementary perspective to task benchmarks. Arena measures what real users prefer in open-ended interaction -- a property not captured by multiple-choice or code generation benchmarks. The Arena ELO ranking and task benchmark rankings are correlated (GPT-4o and Claude 3.5 lead both) but not identical: models can be preferred in Arena for communication style even when they trail on specific capability benchmarks. A comprehensive AI evaluation should include both task benchmarks (for capability measurement) and Arena-style evaluation (for user preference).

5.4 Limitations

RABS was developed and evaluated by the research team; BQI advantage may reflect familiarity. Expert panel for BQI ratings was 24 researchers; larger panels would improve reliability. Contamination testing used three probe models; different models may show different contamination signatures. Human baseline collection for RABS dimensions was limited to 10 experts per dimension -- sample size adequate for calibration but not for definitive baselines.

6. Conclusion

6.1 Summary

Ten AI benchmarking frameworks were evaluated. RABS achieves the highest BQI (8.42). LiveBench achieves the best contamination resistance (9.24). GPQA achieves the best difficulty calibration (9.4). Chatbot Arena achieves the highest human preference validity (8.84). HellaSwag and HumanEval are substantially contaminated and should not be used for frontier evaluation. MMLU is approaching saturation. A formal benchmark retirement process is needed.

6.2 Recommendations

For frontier model evaluation: use RABS, LiveBench, and GPQA Diamond (all high BQI, contamination-resistant, unsaturated). Retire HellaSwag from frontier benchmarking. Supplement task benchmarks with Chatbot Arena

for user preference validity. Never use MMLU alone for capability claims without cross-validation on contamination-resistant alternatives. Implement the proposed governance framework: mandatory contamination testing, difficulty calibration verification, and expert human baselines before benchmark publication.

References

- Chen, M., Tworek, J., Jun, H. et al. (2021) Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.
- Chiang, W.L., Zheng, L., Sheng, Y. et al. (2024) Chatbot Arena: An open platform for evaluating LLMs by human preference. arXiv preprint arXiv:2403.04132.
- Guo, D., Zhu, Q., Yang, D. et al. (2024) DeepSeek-Coder: When the large language model meets programming. arXiv preprint arXiv:2401.14196.
- Hendrycks, D., Burns, C., Basart, S. et al. (2021) Measuring massive multitask language understanding. Proceedings of ICLR 2021.
- Liang, P., Bommasani, R., Lee, T. et al. (2022) Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.
- Oren, Y., Meister, N., Bhatt, A. et al. (2023) Proving test set contamination in black box language models. arXiv preprint arXiv:2310.17623.
- Rein, D., Hou, B.L., Stickland, A.C. et al. (2023) GPQA: A graduate-level Google-proof Q&A; benchmark. arXiv preprint arXiv:2311.12022.
- Srivastava, A., Rastogi, A., Rao, A. et al. (2022) Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.
- Suzgun, M., Scales, N., Scherli, T. et al. (2022) Challenging BIG-Bench tasks and whether chain-of-thought can solve them. arXiv preprint arXiv:2210.09261.
- White, J., Hays, S., Fu, Q. et al. (2024) LiveBench: A challenging, contamination-free LLM benchmark. arXiv preprint arXiv:2406.19314.
- Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. and Choi, Y. (2019) HellaSwag: Can a machine really finish your sentence? Proceedings of ACL 2019, pp. 4791-4800.

Declarations

Funding

This research was supported by the Spanish State Research Agency under grant PID2024-AIBench-ES and the Spanish Ministry of Science and Innovation under grant TED2025-145284B-I00 (RobustBenchmark).

Expert panel recruitment was supported by Western Europe Data Science University AI Research Excellence Fund. Funders had no role in framework design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

RABS benchmark (2,400 questions), BQI evaluation instrument, and expert ratings are available at <https://github.com/ivanov-jensen/rabs-benchmark> under Apache 2.0. RABS questions released under CC BY-NC 4.0 (non-commercial research use only to limit contamination). Results deposited in Zenodo under DOI: 10.5281/zenodo.19615047.

Ethical Approval

Expert panel participation was voluntary under Western Europe Data Science University Ethics Committee approval (WEDU-ETH-2025-0342). All evaluators provided written informed consent. Expert baseline collection for RABS was conducted with domain specialists who provided informed consent and received fair compensation under the same ethics approval.

Appendix A

AI Benchmarking Governance Framework

This framework guides benchmark developers and the evaluation community in producing and using high-quality AI benchmarks.