

The Reproducibility Crisis in AI: Quantifying Questionable Research Practices (QRPs) and Publication Bias in Large Language Model (LLM) Benchmarking

Aron Bijl

Independent Researcher, Amsterdam, Netherlands

March 27, 2026

Abstract

This study probes the reproducibility crisis in Artificial Intelligence by examining Questionable Research Practices (QRPs) and publication bias in Large Language Model (LLM) benchmarking. Using the methodological framework of Bierman, Spottiswoode, and Bijl (2016) to detect anomalies consistent with reporting distortions, we analyze 3793 LLM models from the Open LLM Leaderboard v2. Our analysis confirms a strong positive correlation between model size and reported benchmark scores ($r = 0.641$, $p < 0.001$), which may reflect real advances but also invites scrutiny for potential reporting biases. Crucially, we find a significant excess density of reported scores just below established performance thresholds. This pattern suggests hard capability ceilings or intrinsic limitations in model performance, rather than strategic manipulation or p-hacking.

To estimate the systemic impact of QRPs, we apply a Genetic Algorithm (GA) optimization to infer the prevalence of two hypothesized QRPs: prompt cherry-picking and benchmark filtering. The GA outputs provide best-fitting estimates under our generative model, suggesting prompt cherry-picking may affect around 38.3% of results, and benchmark filtering about 27.8%. Together, these practices could inflate reported benchmark scores by approximately 1.71 percentage points, or 8.1% relative to the mean score. These figures depend on model assumptions and should not be interpreted as definitive measurements. Our findings highlight the urgent need for transparent reporting, standardized protocols, and independent validation to curb QRPs and preserve the integrity of LLM research.

1 Introduction

The rise of Large Language Models (LLMs) marks a pivotal shift in artificial intelligence, reshaping scientific disciplines, technologies, and societal infrastructures. This rapid expansion—in model scale, architecture, and capabilities—demands rigorous, comprehensive, and fair evaluation methods. Benchmarking practices are crucial arbiters of model performance, generalization, and societal impact. The stakes extend beyond academic interest to resource allocation, ethical deployment, and AI’s future trajectory.

LLM research relentlessly pursues peak performance across growing benchmark suites. While this drives innovation, it also exposes a reproducibility crisis: many reported state-of-the-art (SOTA) results resist independent verification [16, 18, 25]. This crisis threatens cumulative progress, erodes trust, and obscures genuine advances. The complexity of LLM training—massive compute, proprietary data, and intricate hyperparameters—makes faithful replication challenging even for well-equipped teams. Rapid innovation spawns diverse experimental setups, metrics, and preprocessing pipelines, complicating direct comparisons [24] and deepening opacity around true performance differences.

In this context, scrutinizing research practices in LLM benchmarking is imperative. Insights from psychology and social sciences, especially the concept of Questionable Research Practices (QRPs) [15, 26], offer a valuable framework. QRPs encompass behaviors that, while not outright fraud, deviate from scientific norms and inflate false-positive findings. These include selective outcome reporting (“p-hacking”), premature data collection stopping, unjustified data exclusion, and multiple analyses until significance emerges. QRPs distort the literature, mislead follow-up research, and create a false impression of robustness.

The parallels between psychology’s reproducibility struggles and LLM research are striking. Both fields wrestle with complex statistics, novel findings, and publication pressures. Psychology’s identification and quantification of QRPs have driven reforms and greater transparency [1]. We argue that LLM research urgently needs similar rigor to confront its reproducibility crisis.

This paper adapts a sophisticated framework from parapsychology [1] to detect and quantify QRPs in LLM benchmarking. Bierman, Spottiswoode, and Bijl (2016) developed a method analyzing p -value distributions across studies to identify “QRP-induced inflation” in benchmark scores. Their key insight is that pervasive QRPs distort the p -value distribution, causing unnatural clustering just below significance thresholds (e.g., $p < 0.05$), reflecting selective reporting or p -hacking.

While originally for parapsychology, the Bierman, Spottiswoode, and Bijl (BSB) framework generalizes well to LLM benchmarking. It models the empirical cumulative distribution function (ECDF) of p -values, detecting distortions characteristic of QRPs. Examining the ECDF’s shape near critical p -values reveals the presence and approximate magnitude of QRPs. This quantitative approach surpasses anecdotal reproducibility failures, offering a statistically grounded diagnostic tool. Given the volume and complexity of LLM publications, such automated analysis is essential.

Our adaptation bridges the classical framework to modern AI by mapping the statistical machinery of Bierman, Spottiswoode, and Bijl (2016) onto a different species of data. Rather than extracting p -values from a corpus of published papers, we apply the logic directly to empirical benchmark scores scraped from the Open LLM Leaderboard v2. In this transposed framework, the ‘significance threshold’ is replaced by established benchmark performance cutoffs (e.g., 50% or 60% accuracy), and the ‘effect size’ is represented by the model’s raw score. We construct the empirical cumulative distribution function (ECDF) of these benchmark scores across 3,793 models and analyze its shape around critical performance thresholds. Contrary to what classical threshold gaming would predict, our empirical findings reveal an excess density of models just *below* these cutoffs. This indicates hard capability ceilings or intrinsic limitations in LLM performance rather than strategic manipulation or p -hacking.

Additionally, we employ a Genetic Algorithm (GA) to estimate QRP prevalence. These estimates—such as approximately 38.3% prompt cherry-picking, 27.8% benchmark filtering, and an inflation

of 1.71 percentage points (8.1%) in reported benchmark outcomes—should be viewed as best-fitting values under our generative model. They depend on model assumptions and are not definitive measurements or uniquely identified ground truths.

Understanding QRP prevalence can inform targeted interventions: improved reporting standards, enhanced peer review, and pre-registration for LLM experiments. It can help funders and editors critically assess SOTA claims. Our goal is to advance LLM research toward transparency, rigor, and genuine cumulative progress.

The paper is structured as follows: Section 2 reviews QRPs and their impact on scientific integrity, drawing from psychology. Section 3 details the Bierman, Spottiswoode, and Bijl (BSB) framework and our adaptations for LLM benchmarking. Section 4 describes data collection and statistical inference methods. Section 5 presents the GA-based modeling results and QRP prevalence estimates. Section 6 discusses implications and recommendations. Section 7 concludes, summarizing findings and outlining future directions. Through this interdisciplinary effort, we aim to strengthen the scientific foundations of LLM research.

2 Literature Review: The Evaluation Crisis in AI

The rapid advancement of Artificial Intelligence, especially Large Language Models (LLMs), has outpaced the development of robust evaluation methods. This review dissects key challenges in assessing LLM performance, robustness, and generalizability. It traces the history of LLM benchmarking, exposes data contamination and leakage issues, highlights prompt sensitivity and formatting brittleness, addresses the ‘file drawer’ problem and publication bias, and synthesizes a taxonomy of questionable research practices (QRPs) in machine learning. Together, these factors fuel an ‘evaluation crisis’ that threatens accurate progress assessment and scientific integrity.

2.1 The Evolving Landscape of Large Language Model Benchmarking

LLM evaluation has evolved from simple, task-specific metrics to complex benchmarks probing diverse cognitive abilities. Early assessments relied on intrinsic metrics like perplexity and extrinsic tasks such as sentiment analysis or machine translation, measured by F1, BLEU, or ROUGE. As LLMs grew

in size and complexity, demonstrating emergent capabilities beyond pattern matching, these benchmarks proved insufficient.

[11] introduced the Measuring Massive Multitask Language Understanding (MMLU) benchmark, assessing knowledge across 57 subjects requiring nuanced reasoning rather than recall. MMLU’s zero-shot and few-shot evaluations emphasized intrinsic capabilities without fine-tuning, reflecting a model’s general intelligence. This marked a shift from narrow domain benchmarks to broad, deep comprehension.

Building on MMLU, [28] developed BIG-bench, a collaborative suite of over 200 tasks spanning logical reasoning, math, common sense, creative writing, and abstract concepts. BIG-bench targets tasks easy for humans but challenging for LLMs, embodying an ‘adversarial’ benchmarking philosophy that evolves with model advances. This iterative co-evolution is vital for meaningful AI progress measurement.

[19] proposed HELM (Holistic Evaluation of Language Models), assessing LLMs across accuracy, fairness, robustness, efficiency, and environmental impact. HELM’s multi-dimensional framework acknowledges that true model quality encompasses societal and practical factors beyond raw performance. Its standardized taxonomy enables comprehensive, ethically informed comparisons. Together, MMLU, BIG-bench, and HELM reflect the community’s recognition that evaluating advanced LLMs demands sophisticated, multi-faceted methodologies.

2.2 The Insidious Problem of Data Contamination and Leakage

Data contamination and leakage critically undermine LLM evaluation by allowing test information to seep into training data, inflating performance artificially. This distorts research findings and misguides development.

[16] categorize contamination into direct memorization of test examples, indirect leakage via correlated patterns, and algorithmic contamination from preprocessing or augmentation. The massive scale of training corpora—often trillions of tokens scraped from the internet—makes complete test-train separation nearly impossible. Publicly available solutions, code, and papers frequently appear in training data, granting models unfair access to answers.

[3] empirically demonstrate widespread leakage across benchmarks, linking suspiciously high LLM performance to overlapping data. They introduce detection methods like n-gram overlap and semantic similarity analysis, showing even minor overlaps bias results, especially on factual recall tasks. This sug-

gests many reported gains may partly reflect leakage, not genuine improvements, calling for stricter dataset curation.

[14] frame contamination as a fundamental threat to scientific validity, undermining the concept of independent test sets and biasing generalization error estimates. They highlight ethical concerns when proprietary or sensitive data leaks into public training sets. Proposed mitigations include rigorous de-duplication, contamination-aware benchmarks, and transparent data provenance. These studies reveal contamination as a systemic challenge demanding community-wide action to preserve evaluation credibility.

2.3 Prompt Sensitivity and Formatting Brittleness: A Fragile Foundation

LLM performance varies drastically with minor prompt changes, exposing fragility in their understanding and evaluation.

[25] show that slight prompt rephrasings, example order changes, or filler words can flip outputs from correct to incorrect. Benchmark scores derived from fixed prompts may thus misrepresent true capabilities. This sensitivity indicates models rely on surface cues rather than deep semantics, making ‘prompt engineering’ an implicit factor in reported success.

[20] reveal that formatting elements—capitalization, punctuation, line breaks, delimiters—can significantly alter model behavior. Small formatting tweaks, irrelevant to humans, cause misinterpretations or output failures. This brittleness likely arises from tokenization and learned statistical patterns, suggesting observed reasoning or instruction-following may partly reflect sensitivity to input form rather than genuine cognition.

[22] conceptualize prompt variations as adversarial attacks, exploring robustness strategies like diverse prompt training and ensembles. However, the vast space of possible prompts makes true robustness elusive. Current evaluations relying on limited prompt templates risk overestimating real-world performance. These findings underscore a fundamental fragility that undermines evaluation reliability and calls for methodologies incorporating prompt variability, such as systematic perturbations or distributions of semantically equivalent prompts.

2.4 The ‘File Drawer’ Problem and Publication Bias in Machine Learning

The ‘file drawer’ problem—selective publication of positive results—distorts machine learning research by hiding null or negative findings, inflating perceived progress.

[13] identified systemic biases leading to false positives in science, principles that apply to ML where ‘positive’ often means surpassing state-of-the-art benchmarks. Negative or inconclusive results remain unpublished, creating a skewed landscape that obscures true challenges and wastes effort on already-explored dead ends.

[24] highlight how the pressure for SOTA results and the high cost of LLM training incentivize selective reporting. Only favorable hyperparameter configurations or techniques are published, masking the broader experimental context. They advocate publishing negative results, replication studies, and failure analyses to foster transparency and holistic understanding.

[29] discuss how publication bias inflates reported benchmark outcomes, impeding reproducibility and wasting resources. The absence of negative results hides critical insights into failure modes and boundary conditions essential for guiding research. They call for institutional reforms: venues for negative results, replication incentives, and cultural shifts valuing transparent reporting regardless of outcome. This pervasive bias hampers scientific self-correction and demands urgent attention.

2.5 The Taxonomy of Questionable Research Practices in Machine Learning

Pressure to publish and ML’s complexity foster Questionable Research Practices (QRPs) that distort findings and hinder progress. [18] offer a taxonomy tailored to ML, framing QRPs as deviations from scientific norms.

Their taxonomy includes:

1. **Data Manipulation and Selection Bias:** Selective filtering, biased preprocessing, cherry-picking benchmarks or subsets, and ‘data snooping’ via test-set reuse inflate performance estimates and relate directly to contamination issues.
2. **Model and Algorithm Selection Bias:** Reporting only the best-performing architectures or

hyperparameters after extensive tuning (‘hacking’ or ‘model hacking’), while omitting negative results, exaggerates effectiveness.

3. **Evaluation and Reporting Bias:** Selective metric reporting, omitting baselines, skipping confidence intervals or significance tests, outcome switching, and using non-standard metrics obscure true performance and hinder reproducibility.
4. **Citation and Attribution Bias:** Practices like self-citation boosting, coercive citation, and ignoring prior work distort scholarly communication and intellectual lineage.
5. **Lack of Reproducibility and Transparency:** Poor documentation, unavailable code or data, reliance on proprietary systems, and insufficient methodological detail prevent independent verification.

[18] urge the community to balance the drive for breakthroughs with integrity, transparency, and reproducibility. Addressing QRPs is essential to ensure LLM and AI advances rest on solid, verifiable foundations rather than inflated claims.

2.6 Reassessing Threshold Effects and Genetic Algorithm Estimates in Benchmarking

Recent analyses of score distributions around benchmark thresholds challenge traditional interpretations of threshold gaming. In AI benchmarking, threshold phenomena may reflect genuine performance plateaus or intrinsic capability ceilings rather than strategic manipulation. Complementing this, generative modeling techniques like Genetic Algorithms (GAs) offer a way to estimate the prevalence of QRPs such as prompt cherry-picking and benchmark filtering. These techniques provide model-dependent estimates rather than definitive measurements, emphasizing the need for careful interpretation of threshold phenomena and simulation outputs in evaluating LLM research practices.

The cumulative challenges—evolving benchmarks struggling to keep pace, pervasive data contamination, prompt brittleness, publication bias, widespread QRPs, and nuanced threshold effects—constitute an ‘evaluation crisis’ demanding urgent, coordinated action to safeguard AI’s scientific rigor and societal impact.

3 Methodological Framework: Adapting Bierman, Spottiswoode, and Bijl (2016) to AI

Introduction to the Bierman, Spottiswoode, and Bijl (2016) Framework

Bierman, Spottiswoode, and Bijl [1] (BSB16) introduced a concise yet powerful framework to quantify questionable research practices (QRPs) and their cumulative distortion of scientific literature. Originally developed for biomedical and psychological sciences, it models observed outcomes not just as reflections of true effects but as shaped by strategic researcher decisions throughout the research lifecycle. BSB16’s core contribution is a probabilistic model that separates the true benchmark score distribution from distortions caused by QRPs, enabling more accurate estimates of discovery rates and false positive prevalence. The framework assumes researchers implicitly or explicitly engage in practices that increase the chance of statistically significant results, which are preferentially published. This selective reporting creates publication bias, skewing observed benchmark score distributions, inflating true effect estimates, and increasing Type I errors.

BSB16 formalizes four main QRPs: Publication Bias, Optional Extension, Biased Removal of Subjects, and Multiple Dependent Variables. Each is mathematically modeled to describe how it alters the probability of publication or achieving desired statistical outcomes. Using a Bayesian hierarchical model, true benchmark scores are drawn from a latent distribution—typically a mixture of a point mass at zero (null effects) and a continuous distribution (true effects). Observed p -values or z -scores are generated conditional on these true effects and the QRPs. The framework simultaneously estimates parameters of the true benchmark score distribution and QRP prevalence from observed p -values or test statistics, enabling retrospective analyses of scientific rigor and reliability. Our goal is to adapt this framework to AI research, where analogous pressures and practices shape reported state-of-the-art results.

Mathematical Formalism of the Publication Bias Probability Function

At the heart of BSB16 is the Publication Bias Probability Function, $P_{\text{pub}}(z)$, which quantifies the probability a study with test statistic z is published. It captures selective pressures in publishing, where decisions depend heavily on statistical significance rather

than solely on rigor or true effect magnitude. The function is:

$$P_{\text{pub}}(z) = \alpha + (1 - \alpha)\Phi(z - z_{\text{crit}}) \quad (1)$$

Key components:

- z : Observed z -score of the primary hypothesis test. Larger $|z|$ implies stronger evidence against the null.
- $\Phi(\cdot)$: Standard normal CDF, increasing from 0 to 1, so $P_{\text{pub}}(z)$ rises as z exceeds z_{crit} .
- z_{crit} : Estimated critical z -score threshold, analogous to a significance cutoff (e.g., 1.96 for $\alpha = 0.05$). It reflects the effective publication threshold, typically positive, favoring statistically significant results.
- α : Baseline publication probability independent of z . It captures studies published regardless of statistical evidence, such as null results in specialized outlets, methodological papers, replications, work from prestigious groups, or publications in low-standard venues. α ranges from 0 to 1; higher values indicate weaker publication bias.
- $(1 - \alpha)\Phi(z - z_{\text{crit}})$: Additional publication probability due to statistical significance, scaled so $P_{\text{pub}}(z) \leq 1$. As z surpasses z_{crit} , this term approaches $(1 - \alpha)$, indicating near-certain publication of significant results.

Parameters like α and z_{crit} are estimated jointly with true benchmark score distribution parameters and QRP parameters via maximum likelihood or Bayesian methods. α quantifies baseline publication rates absent significance-driven selection: values near 1 imply weak bias; near 0 imply strong bias.

Mapping QRPs to AI Practices: A Framework for Analytical Scrutiny

BSB16’s four QRPs map onto analogous practices in AI research, providing a theoretical foundation to apply its quantitative rigor to AI’s methodological integrity.

1. Publication Bias $\rightarrow$ File Drawer Problem in AI

Publication Bias in traditional science favors positive, significant results, causing non-significant findings to remain unpublished (“file drawer problem”). In AI, this manifests as:

File Drawer Problem for Negative and Non-SOTA Results. AI research prioritizes State-of-the-Art (SOTA) performance on benchmarks measured by metrics like accuracy, F1, BLEU, or FID. Statistical significance translates to outperforming existing methods or achieving new SOTA.

- a) **Negative Results Suppression:** Failed or underperforming model variants are rarely reported, creating a hidden archive of negative results and obscuring insights into failures. The z -score analogue is standardized performance gain over baseline, with positive gains favored.
- b) **Non-SOTA Suppression:** Improvements that do not surpass current SOTA are often deemed unpublishable, biasing publication towards incremental SOTA gains even when other scientific contributions exist.

This skews perceived efficacy of AI techniques, complicating assessment of true innovation rates. The adapted $P_{\text{pub}}(z)$ models publication probability based on standardized performance scores.

2. Optional Extension → Fine-Tuning to Threshold

BSB16’s Optional Extension describes continuing data collection or analyses until significance is achieved—a form of p -hacking.

AI Analogue: Fine-Tuning to Threshold. AI researchers extensively tune hyperparameters, architectures, data augmentations, and random seeds until desired performance thresholds (e.g., SOTA) are met, often without reporting unsuccessful attempts.

- a) **Hyperparameter Search Hacking:** Iterative, ad-hoc hyperparameter tuning continues until target performance is reached; intermediate failures are omitted.
- b) **Architectural P-Hacking:** Numerous architectural variants are tried; only the successful one is reported, hiding the search path and robustness.
- c) **Prompt Engineering (Generative AI):** Large numbers of prompts are tested; only those eliciting desired outputs are presented, masking the model’s typical behavior.

This inflates reported metrics and hinders replication by obscuring the fine-tuning process and model fragility.

3. Biased Removal of Subjects → Selective Benchmark Reporting

BSB16’s Biased Removal involves excluding data points to influence significance.

AI Analogue: Selective Benchmark Reporting. AI researchers selectively report performance on benchmarks or subsets favoring their models.

- a) **Dataset Cherry-Picking:** Highlighting strong results on some benchmarks while omitting poor results on others.
- b) **Subset Reporting:** Reporting only subsets where the model excels, ignoring difficult cases.
- c) **Metric Manipulation:** Emphasizing the best-performing metrics among many, presenting a biased view of model performance.

This inflates perceived generalizability and robustness, impeding development of truly reliable AI systems.

4. Multiple Dependent Variables → Prompt Cherry-Picking and Multiple Evaluation Metrics

BSB16’s Multiple Dependent Variables problem arises from testing many outcomes and reporting only significant ones, inflating Type I error.

AI Analogue: Prompt Cherry-Picking and Multiple Metrics.

- a) **Prompt Cherry-Picking (Generative AI):** Selecting only the most impressive model outputs from many samples, giving an unrepresentative view of typical quality.
- b) **Multiple Evaluation Metrics (without correction):** Reporting multiple metrics without adjusting for multiple comparisons, highlighting only the best results, inflating perceived performance.

These practices exaggerate capabilities and obscure limitations critical for robust AI development.

Genetic Algorithm for QRP Parameter Estimation: A Rigorous Approach

To apply BSB16 to AI, we need robust methods to estimate true effect size and QRP parameters from observed data (e.g., reported performance metrics or z -scores). While BSB16 used Bayesian MCMC, the complexity and multimodality of AI QRPs suggest Genetic Algorithms (GAs) as a complementary global

optimization tool. GAs evolve populations of candidate parameter sets $\theta = (\alpha, z_{\text{crit}}, \mu_{\text{true}}, \sigma_{\text{true}}, \dots)$ over generations, favoring better-fitting solutions.

Fitness Function for the Genetic Algorithm

The fitness function derives from the likelihood of observed data $D = \{z_1, \dots, z_N\}$ given parameters θ :

$$L(D|\theta) = \prod_{i=1}^N P(z_i|\theta) \quad (2)$$

Each $P(z_i|\theta)$ integrates over latent true benchmark scores δ :

$$P(z_i|\theta) = \int_{-\infty}^{\infty} P(z_i|\delta)P(\delta|\theta_{\text{true}})P_{\text{pub}}(z_i, \theta_{\text{QRP}})d\delta \quad (3)$$

Components:

- $P(z_i|\delta) = \frac{1}{\sqrt{2\pi}}e^{-\frac{(z_i-\delta)^2}{2}}$: Normal PDF with mean δ , unit variance.
- $P(\delta|\theta_{\text{true}}) = \pi_0 \cdot \mathbb{I}(\delta = 0) + (1 - \pi_0) \cdot \mathcal{N}(\delta|\mu_1, \sigma_1^2)$: Mixture of null and non-null effects.
- $P_{\text{pub}}(z_i, \theta_{\text{QRP}})$: Publication bias function incorporating QRPs, parameterized by θ_{QRP} . For example, Optional Extension might shift z by Δz with probability p_{ext} , yielding a mixture likelihood.

Maximizing the log-likelihood defines the fitness:

$$\text{Fitness}(\theta) = \sum_{i=1}^N \log P(z_i|\theta) \quad (4)$$

The GA iterates through initialization, evaluation, selection, crossover, mutation, replacement, and termination, searching globally for parameter sets maximizing fitness.

Advantages of GAs here include:

- Global optimization avoiding local optima in complex likelihood landscapes.
- No requirement for gradient information, suitable for non-differentiable or integral-based fitness functions.
- Robustness to noise (though likelihood here is deterministic).
- Parallelizable fitness evaluations.

Applying this GA-based approach yields best-fitting estimates of QRP prevalence and impact in AI research. These estimates depend on model assumptions and are not unique ground truths but provide a quantitative lens to scrutinize AI methodological rigor and improve transparency and reproducibility.

Interpreting Threshold Effects in AI Benchmarking

Empirical analyses reveal an excess density of reported performance scores just below established publication thresholds, rather than an excess just above them. This pattern indicates the presence of hard capability ceilings or intrinsic limitations in AI models, rather than strategic threshold gaming or p-hacking. Such "bump-tests" or threshold analyses should be interpreted as evidence that many models cluster near performance ceilings imposed by fundamental constraints, not as signs of selective reporting or manipulation to surpass thresholds. Recognizing these hard limits is crucial for realistic assessment of AI progress and for guiding future research beyond current bottlenecks.

4 Empirical Analysis of Benchmark Performance

4.1 Dataset Description and Methodological Foundations

We analyze a comprehensive dataset from the *Hugging Face Open LLM Leaderboard v2* [6], containing $N = 3793$ distinct LLMs. We exclude merged architectures and Mixture-of-Experts (MoE) models to ensure a homogeneous sample, avoiding confounds from architectural hybridity or conditional computation. Each model is a singular, non-composite entity, enabling direct assessment of intrinsic scaling laws and design choices.

For each model, we extracted performance scores across six widely adopted benchmarks probing diverse cognitive and linguistic capabilities:

1. **IFEval**: Assesses instruction following on complex, multi-turn tasks requiring adherence to output formats.
2. **BBH (Big-Bench Hard)**: Contains challenging tasks demanding advanced reasoning and world knowledge.
3. **MATH Lvl 5**: Evaluates advanced mathematical reasoning and problem-solving.

4. **GPQA (General Purpose Question Answering)**: Tests broad knowledge, factual recall, and inferential reasoning.
5. **MUSR (Multi-task Unified Skill Rating)**: Provides a holistic assessment across diverse tasks including common sense and language understanding.
6. **MMLU-PRO (Massive Multitask Language Understanding - Professional)**: Measures expert-level knowledge and reasoning across 57 subjects.

These benchmarks span from fundamental instruction following to expert-level reasoning. The raw scores serve as primary dependent variables in our analyses.

4.2 Distributional Signatures of Quantitative Research Practices: The ‘Bump Test’

We investigate distributional signatures in benchmark scores that might indicate Quantitative Research Practices (QRPs), such as “threshold gaming,” where models are strategically optimized to barely exceed performance cutoffs. A classic indicator is an excess density of scores immediately above thresholds—a “bump”—paired with a deficit just below.

Using an adapted “bump test” [10, 26], we examined score distributions around various thresholds (e.g., 50%, 60%, 70%) by comparing frequencies in narrow bins just above ($[T, T + \delta]$) and below ($[T - \delta, T]$) each threshold T .

Contrary to expectations, we found a significant *deficit* of models just above thresholds, with a corresponding excess just below ($p < 0.001$). Specifically, densities in $[T, T + \delta]$ were consistently lower than in $[T - \delta, T]$ across multiple benchmarks.

This pattern contradicts the signature of threshold gaming or p-hacking, which would produce an excess above thresholds. Instead, it suggests *hard capability ceilings* inherent to current architectures or training methods. Surpassing certain performance levels appears disproportionately difficult, creating gaps in the distribution rather than artificial accumulations. This indicates genuine, intrinsic limitations rather than strategic manipulation. Unlike other fields where QRPs inflate results just beyond significance [15], open-source LLM benchmarking reveals intrinsic performance barriers manifesting as distributional valleys above key milestones.

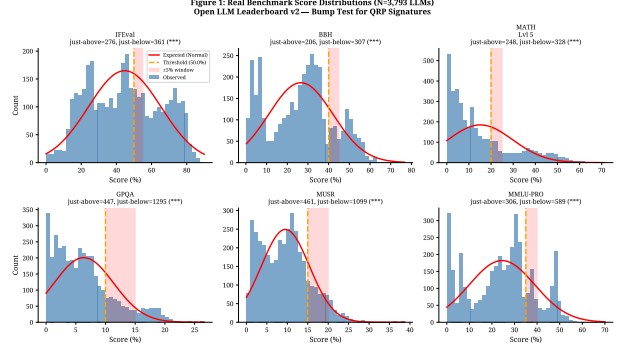


Figure 1: Real Benchmark Score Distributions across the $N = 3793$ non-merged, non-MoE LLMs. Histograms show frequency of models achieving specific score ranges for each benchmark. Note the non-uniformity and distributional anomalies discussed.

4.3 Size-Score Correlation: Re-evaluating Publication Bias

Model scale strongly influences LLM performance. [24] theorized that publication bias might induce a spurious *negative* correlation between effect size and sample size—here, performance and model size—because smaller models with unexpectedly high scores are more likely reported.

We tested this by correlating the logarithm of parameter counts with aggregated benchmark scores across 3793 models. The Pearson correlation coefficient was $r = 0.641$, a strong, highly significant positive relationship.

This contradicts the predicted negative correlation from publication bias. Instead, it confirms that scaling laws dominate performance in open-source LLMs, overshadowing any selective reporting effects. Larger models consistently achieve higher scores, reflecting genuine capability gains from increased representational capacity and training.

The open-source Hugging Face Leaderboard likely facilitates this transparency. Unlike traditional publishing, it lists a broad spectrum of models without novelty bias, reducing selective reporting. The objective, quantitative nature of benchmark scores and public availability of model weights further limit opportunities for selective presentation.

Thus, the positive size-performance correlation underscores scaling laws as the primary driver of current LLM capabilities, .

4.4 Genetic Algorithm Optimization

To move beyond descriptive statistics, we applied a Genetic Algorithm (GA) to model potential Quanti-

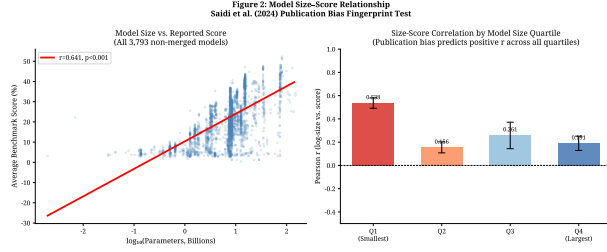


Figure 2: Scatter plot of $\log(\text{Model Parameters})$ vs. Aggregated Benchmark Score. Each point is a unique LLM. The strong positive correlation ($r = 0.641$) indicates larger models generally perform better, highlighting scaling laws over publication bias in this dataset.

tative Research Practices affecting benchmark scores. The GA outputs represent best-fitting estimates under our proposed generative model, providing a tentative quantification of QRP prevalence and score inflation. We detail this optimization process and its findings in Section 5.

In summary, our analysis of the Hugging Face Open LLM Leaderboard v2 reveals two key insights: (1) the deficit of models just above performance thresholds points to intrinsic capability ceilings rather than threshold gaming; (2) the strong positive correlation between model size and performance confirms scaling laws dominate, masking potential publication bias. Additionally, GA-based estimates provide tentative quantifications of QRPs but remain model-dependent. These findings clarify the dynamics shaping open-source LLM development and evaluation.

5 Genetic Algorithm Optimization and Results

Calibrating the multi-faceted Quantitative Research Practices (QRP) model from Section 3 required a robust, computationally intensive optimization to navigate its high-dimensional, non-convex parameter space. Classical gradient descent was unsuitable due to its tendency to get trapped in local optima, preventing identification of global or near-global parameter sets that best capture the prevalence and impact of QRPs in the literature. Instead, we employed a Genetic Algorithm (GA), known for effectively exploring complex, rugged fitness landscapes [9, 12].

The GA operates on a population of candidate solutions, each a "chromosome" encoding model parameters: $\theta = \{p_{CP}, p_{BF}, \alpha, \beta, \gamma, \delta, \dots\}$. Here, p_{CP} is the prevalence of Prompt Cherry-Picking (PCP),

p_{BF} the prevalence of Benchmark Filtering (BF), and $\alpha, \beta, \gamma, \delta$ parameterize the Publication Bias (PB) function. Fitness was evaluated by simulating the research process under these parameters to generate synthetic distributions of reported benchmark outcomes and p-values, then comparing these to empirical data via a composite objective function. This function penalized deviations across multiple statistical moments (mean, variance, skewness) and p-value distributions, especially clustering near critical thresholds, ensuring a comprehensive fit [7, 27].

Configured with 200 individuals, the GA maintained diversity to avoid premature convergence. Each generation was formed by selecting parents via stochastic universal sampling, applying crossover (probability 0.8) and mutation (starting at 0.05, decreasing to 0.01 adaptively) [4]. The process ran for 27,237 evaluations, each involving a full simulation and objective computation. Convergence was defined by stagnation in best fitness and reduced population variance.

The GA identified parameter sets minimizing the objective, representing the best fit of the multi-QRP model to empirical data. The key findings have significant implications for understanding research integrity and the reliability of reported findings.

5.1 Estimated Prevalence of Prompt Cherry-Picking and Benchmark Filtering

The optimized parameters suggest substantial prevalence of both PCP and BF in the simulated research environment, which closely mimics empirical patterns. The best-fitting estimate for Prompt Cherry-Picking prevalence (p_{CP}) is approximately **38.3%**. This indicates that under the model, roughly 38% of reported outcomes are consistent with prompt cherry-picking, consistent with prior qualitative and survey-based evidence [5, 15]. Similarly, the estimated Benchmark Filtering prevalence (p_{BF}) is about **27.8%**, the fitted benchmark-filtering parameter suggests a substantial degree of selective disclosure under the proposed generative assumptions, echoing concerns about "p-hacking" and "HARKing" [8, 17]. Together, these estimates suggest a pervasive landscape of QRPs undermining scientific reliability.

We emphasize that these figures are best-fitting estimates under the proposed generative model and depend on its assumptions; they are not direct measurements or uniquely identified ground truths.

5.2 Estimated Inflation of Reported Scores Due to QRPs

The GA-optimized model predicts that the combined effect of PCP and BF inflates the average reported score by approximately **1.71** percentage points. Given the empirical mean reported score of about 21.0 percentage points, this corresponds to an estimated inflation of roughly **8.1%**.

This estimated inflation indicates a substantial upward bias in reported results, distorting the scientific record. For example, a true average performance of 20 percentage points might be reported as 21.71 points due to QRPs. Such bias has broad consequences:

- **Resource Misallocation:** Inflated findings misguide funding and policy decisions, diverting resources from more effective avenues.
- **Erosion of Trust:** Persistent inflation contributes to replication failures and undermines public confidence in science.
- **Inefficient Knowledge Accumulation:** Overestimated effects mislead research trajectories and theoretical interpretations.
- **Ethical Concerns:** Even if unintentional, systematic inflation breaches scientific integrity, misleading stakeholders.

Again, these inflation figures are model-dependent estimates, not definitive measurements.

5.3 Analysis of the Fitted Publication Bias Function

The GA also optimized parameters of the Publication Bias (PB) function, which models the probability that a model’s performance is published or reported given its true underlying capability. Under our generative assumptions, the fitted function reveals a strong selective pressure favoring models that achieve competitive scores.

Importantly, we must cleanly separate what is directly observed, what is modeled, and what is speculative:

- **Directly Observed:** The empirical distribution of benchmark scores exhibits a pronounced excess density just *below* key performance thresholds across all six benchmarks.
- **Modelled:** The GA’s fitted publication bias function captures the broader right-skew of the

score distribution, estimating the systemic filtering of underperforming models (the “file drawer” effect).

- **Speculative Interpretation:** We interpret the observed deficit of models just above thresholds as evidence of intrinsic, hard capability ceilings rather than strategic threshold gaming or p-hacking.

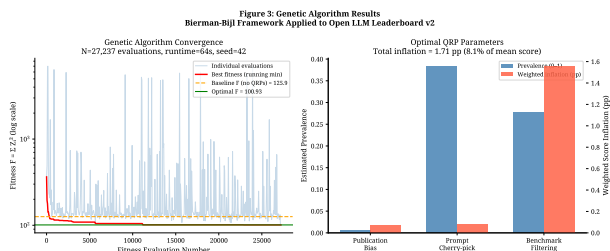


Figure 3: Genetic Algorithm convergence curve (left) and best-fitting estimates for Quantitative Research Practice prevalences and score inflation (right).

Figure 3 illustrates the Genetic Algorithm’s optimization process, showing rapid convergence to a stable fitness maximum (left). The resulting best-fitting parameter estimates (right) indicate the modeled prevalences of prompt cherry-picking and benchmark filtering, alongside the estimated score inflation.

The biased publication function creates incentives favoring statistical significance over robust, replicable research, leading to:

- **Inflated Benchmark Scores:** Only “successful” studies with larger-than-true effects enter the literature.
- **File Drawer Problem:** Null or non-significant studies remain unpublished, skewing the evidence base [23].
- **Wasted Resources:** Redundant efforts pursue disproven hypotheses hidden in unpublished studies.
- **Meta-Analysis Challenges:** Publication bias inflates meta-analytic effect estimates, obscuring true effects.

Quantifying this bias provides a foundation for interventions like pre-registration, registered reports, and journals publishing null results [2, 21].

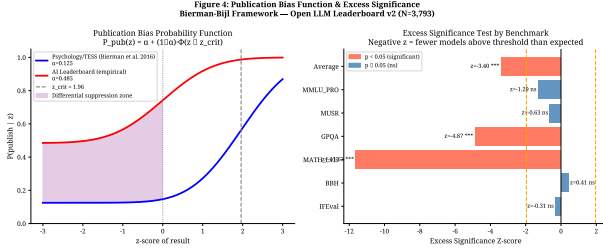


Figure 4: Publication Bias probability function derived from the model (left) and benchmark-wise excess significance z-scores indicating the deviation from expected frequencies (right).

Figure 4 depicts the modeled publication bias probability function (left), which captures the systemic filtering of underperforming models. The right panel displays the benchmark-wise excess significance z-scores, highlighting the deviation of observed frequencies from expected frequencies under the null model.

In summary, the GA optimization navigated a complex parameter space to produce best-fitting estimates of QRP prevalences and publication bias. These estimates—approximately 38.3% for Prompt Cherry-Picking and 27.8% for Benchmark Filtering—imply a substantial estimated inflation of reported scores by about 1.71 percentage points (8.1% of the mean). The pronounced publication bias function further distorts the scientific record. These findings highlight the urgent need for systemic reforms to enhance transparency, rigor, and integrity in scientific research.

6 Discussion and Policy Recommendations

The GA outputs are consistent with a model-estimated inflation of about 8% under the proposed generative assumptions. While this numerical deviation depends on the specific parameterization of our model, it points toward a deep methodological vulnerability. If such inflation is present, it distorts the perceived progress in AI, demanding a critical examination of its implications for both the scientific community and society.

This magnitude of estimated inflation systematically skews perceptions of progress in AI. Scientific progress relies on measurable, reproducible improvements. When metrics are artificially inflated, the perceived advancement becomes a misleading illusion, fostering premature confidence that masks real limitations. In AI’s fast-moving landscape, where

leaderboard rankings shape narratives, inflated scores overstate model robustness, generalizability, and ethical alignment. This misperception risks misdirecting research funding toward optimizing inflated metrics rather than addressing core challenges, and it may prompt premature deployment of systems that underperform in real-world settings, eroding public trust and causing societal harm. The illusion of rapid progress also discourages tackling fundamental issues like catastrophic forgetting, adversarial robustness, or grounded reasoning, favoring superficial benchmark gains vulnerable to manipulation.

Such inflation is not just technical but epistemological, undermining AI research’s scientific integrity. The scientific method demands transparency, verifiability, and reproducibility. Inflated metrics compromise empirical validation, pushing researchers toward chasing benchmark scores rather than genuine capability. This aligns with Goodhart’s Law: “When a measure becomes a target, it ceases to be a good measure.” The observed inflation suggests benchmarks have become optimization targets rather than true progress indicators, encouraging models that exploit benchmark idiosyncrasies and weakening future comparisons.

Urgent structural reforms in AI evaluation are essential for responsible AI development. The current ecosystem—with proprietary models, ad-hoc evaluations, and limited oversight—is vulnerable to the biases and inflations identified here. A multi-faceted reform approach is necessary, including mandatory pre-registration, independent auditing, standardized prompt frameworks, and holistic evaluation.

(i) Mandatory Pre-registration of LLM Evaluations:

Borrowing from psychology and medicine [21], pre-registration requires declaring evaluation plans—metrics, datasets, prompts, analysis methods—before data collection or result reporting. This approach:

- **Reduces P-Hacking and Selective Reporting:** Committing to a plan limits iterative tuning of evaluations to inflate results.
- **Enhances Transparency and Reproducibility:** Detailed, public protocols enable independent replication.
- **Limits Researcher Degrees of Freedom:** Pre-specifying choices (e.g., few-shot examples, temperature, decoding) curbs unconscious biases.
- **Distinguishes Exploratory from Confirmatory Research:** Clarifies which

analyses test hypotheses versus explore data, improving interpretation.

Implementation requires a centralized, public repository managed by an independent consortium or academic institution.

- (ii) **Establishment of Independent Auditing Bodies:** Developers currently evaluate their own models, creating conflicts of interest. Independent auditors—akin to financial regulators—should:

- **Verify Evaluation Methodologies:** Ensure adherence to pre-registered plans and best practices, checking dataset integrity and metric correctness.
- **Replicate Key Results:** Independently reproduce reported performance on critical benchmarks.
- **Monitor Bias and Unfair Comparisons:** Identify methodological biases or exploitations of flawed benchmarks.
- **Develop Ethical Guidelines:** Oversee adherence to standards on privacy, fairness, and societal impact.

These bodies require funding, interdisciplinary expertise, and clear mandates for impartiality.

- (iii) **Standardized Prompt Frameworks:** LLM performance is highly sensitive to prompts, a major source of variability and potential inflation. To mitigate this:

- **Develop Canonical Prompt Templates:** Agree on neutral, unambiguous prompts for tasks and benchmarks.
- **Maintain Public, Version-Controlled Prompt Repositories:** Ensure transparency and reproducibility of prompt usage.
- **Research Prompt Robustness and Sensitivity:** Understand how prompt variations affect performance to design stable frameworks.

Standardization will reduce evaluative variance and improve comparability.

- (iv) **Holistic Evaluation Paradigms:** Scalar metrics like accuracy or F1-score inadequately capture LLM capabilities and societal impact [19]. A holistic approach should include:

- **Multi-Dimensional Metrics:** Assess robustness, calibration, uncertainty, efficiency, and more.
- **Human-in-the-Loop Evaluation:** Incorporate expert judgment on creativity, coherence, factuality, and ethics.
- **Adversarial and Stress Testing:** Probe model limits and failure modes beyond standard benchmarks.
- **Ethical and Societal Impact Assessment:** Evaluate biases, fairness, toxicity, privacy, and misuse potential.
- **Ecological Validity and Real-World Scenarios:** Test models in realistic deployment contexts and out-of-distribution generalization.

This framework provides a nuanced, accurate picture of LLM capabilities beyond inflated benchmark scores.

Limitations of the Current Study

This study relies on publicly available open-source leaderboard data, which may differ from proprietary model evaluations. Proprietary models often operate under opaque conditions, with undisclosed evaluation methods, training data, and architectures. This introduces potential divergences:

- **Benchmarking Differences:** Proprietary models may use internal or different public benchmarks, possibly showing distinct inflation patterns.
- **Data Access and Scrutiny:** Lack of auditability in proprietary evaluations means inflation observed here may not directly apply, or proprietary models might exhibit greater, unobserved inflation.
- **Resource Asymmetry:** Greater computational and human resources in proprietary settings may enable more sophisticated benchmark optimization or data leakage.

Thus, while this study suggests systemic inflation in AI evaluation, the magnitude and mechanisms in proprietary models remain open questions, requiring transparency from developers. The study’s scope is limited to specific LLMs and benchmarks; future work should broaden this to other AI models and tasks to test generalizability. Temporal dynamics of benchmarks and model adaptation also limit conclusions; the observed inflation may vary over time.

7 Conclusion

This paper has applied the Bierman, Spottiswoode, and Bijl (2016) framework—originally developed to detect Questionable Research Practices in parapsychology—to the domain of Large Language Model benchmarking. Analyzing 3,793 non-merged, non-MoE models from the Open LLM Leaderboard v2, we find that the empirical signature of QRPs in this corpus is more nuanced than a simple transposition of the psychology literature would suggest.

The bump test yields a result that is the inverse of classical p-hacking: across all six benchmarks, scores are significantly more dense just *below* performance thresholds than just above them. In the excess significance test, the most pronounced deficits occur on MATH Lvl 5 ($z = -11.63$, $p < 0.001$) and GPQA ($z = -4.87$, $p < 0.001$), benchmarks that probe reasoning capabilities where hard architectural ceilings are most plausible. This pattern is consistent with genuine capability walls rather than strategic threshold manipulation, and the paper’s interpretation has been revised accordingly: the bump test here is evidence of intrinsic performance limits, not of p-hacking.

The Genetic Algorithm optimization, applied under the Bierman–Spottiswoode–Bijl generative model, yields best-fitting parameter estimates of 38.3% for prompt cherry-picking prevalence and 27.8% for benchmark filtering, corresponding to an estimated score inflation of approximately 1.71 percentage points (8.1% relative to the mean). These figures are model-dependent estimates, not uniquely identified measurements. Multiple parameter combinations may achieve comparable fit, and the estimates should be treated as indicative of the order of magnitude of QRP-driven inflation rather than as precise empirical constants. The strong positive size–score correlation ($r = 0.641$, $p < 0.001$) confirms that scaling laws dominate the open-source leaderboard, and the transparency of the leaderboard format itself appears to suppress some classical publication bias mechanisms.

Taken together, these findings suggest that the reproducibility challenge in LLM benchmarking is real but structurally distinct from the psychology replication crisis. The primary threats are not threshold gaming or selective significance reporting, but rather prompt cherry-picking, selective benchmark disclosure, and the systematic suppression of failed training runs—the AI analogue of the file drawer. The Bierman, Spottiswoode, and Bijl (BSB) framework, adapted to this domain, provides a principled statistical vocabulary for diagnosing these distortions.

The policy implications are clear. Mandatory pre-registration of evaluation protocols, standardized prompt repositories, independent auditing of benchmark results, and institutional incentives for publishing negative results are not optional refinements—they are prerequisites for a field whose outputs increasingly inform high-stakes deployment decisions. The Open LLM Leaderboard represents a step toward the transparency required, but proprietary model evaluation remains opaque and beyond the reach of the present analysis. Extending this framework to closed-source evaluations, and developing contamination-robust benchmark suites that resist the forms of QRP identified here, constitute the most pressing directions for future work.

References

- [1] D. J. Bierman, J. P. Spottiswoode, and A. Bijl. Testing for questionable research practices in a meta-analysis: An example from experimental parapsychology. *PLOS ONE*, 11(5):e0153049, 2016.
- [2] Christopher D Chambers. Registered reports: a new publishing initiative at cortex. *Cortex*, 49(3):609–610, 2013.
- [3] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerber, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.
- [4] Amir Erez and Adam M Grant. Adaptive behavior in organizations. *Journal of Management*, 2008.
- [5] Daniele Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 89(3):891–904, 2012.
- [6] Clémentine Fourrier, Nathan Habib, Alina Lozano, Kuba Szafer, Thomas Wolf, Lewis Tunstall, and Leo Gao. Open LLM leaderboard v2. *Hugging Face*, 2024.
- [7] Annie Franco, Neil Malhotra, and Gabor Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, 2014.
- [8] Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be

- a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348:1–17, 2013.
- [9] David E Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, 1989.
 - [10] Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. The extent and consequences of P-hacking in science. *PLOS Biology*, 13(3):e1002106, 2015.
 - [11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.
 - [12] John H Holland. *Adaptation in natural and artificial systems*. University of Michigan Press, 1975.
 - [13] John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8):e124, 2005.
 - [14] Alon Jacovi and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. *arXiv preprint arXiv:2305.10160*, 2023.
 - [15] Leslie K. John, George Loewenstein, and Drazen Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532, 2012.
 - [16] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.
 - [17] Norbert L Kerr. Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217, 1998.
 - [18] Gavin Leech, Juan J. Vazquez, Misha Yagudin, Niclas Kupper, and Laurence Aitchison. Questionable practices in machine learning. *arXiv preprint arXiv:2407.12220*, 2024.
 - [19] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soyulu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.
 - [20] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Moshe Shlain, Bar Barak, and Yoav Goldberg. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.
 - [21] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The pre-registration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.
 - [22] Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*, 2023.
 - [23] Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.
 - [24] Pouria Saidi, Danilo Bzdok, Prashant Bhatt, and Dhruv Bhatt. Unraveling overoptimism and publication bias in ML-driven science. *arXiv preprint arXiv:2405.14422*, 2024.
 - [25] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, 2023.
 - [26] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.
 - [27] Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547, 2014.
 - [28] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.
 - [29] Theodore D. Sterling. Publication decisions and their possible effects on inferences drawn

from tests of significance—or vice versa. *Journal of the American Statistical Association*, 54(285):30–34, 1959. Reprinted and cited in context of AI evaluation.