

What Current AI Benchmarks Leave Unmeasured: Modality, Search, Citations, and Implications (for Safety Evaluations)

Ro Encarnación¹, Tina Behzad², Emma Lurie¹ Danaé Metaxa¹

¹University of Pennsylvania

²Stony Brook University

rone@engineering.upenn.edu, tbehzad@cs.stonybrook.edu, ewlurie@engineering.upenn.edu, metaxa@engineering.upenn.edu

Abstract

Large language model (LLM) benchmark evaluations are routinely used to support claims about model safety, reliability, and deployment readiness. Yet most evaluations rely on a single access modality (model APIs), perform a single run per prompt, and report accuracy as the primary outcome metric, without accounting for conditions such as web search that may have effects on model behavior in deployment. We audit these assumptions for one of the most widely-used LLMs, comparing two modalities, ChatGPT’s chat UI and OpenAI’s API, with and without web search enabled. We use a stratified total sample of 401 prompts from two popular benchmarks, BBQ and SafetyBench, collecting 4,812 total responses across three repeated runs per prompt. Beyond standard performance measures, we evaluate model output dimensions including response consistency, response text similarity, citation grounding, and abstention behavior. For instance, chat UI responses were less accurate than API responses on both benchmarks with search disabled. Enabling web search reduced accuracy by up to 8 percentage points, and even reversed the direction of modality performance trends for one benchmark. Repeated runs of the same prompt produced inconsistent responses in up to 21% of prompts. The two modalities also grounded answers in different citations, and abstention behavior was also inconsistent across both modalities. These results illustrate that, even within a model family, reporting only simple accuracy metrics can obscure important forms of model behavioral variation relevant to AI safety assessments. We argue that AI safety evaluations should systematically account for modality, multi-run consistency, search conditions, and response-level behaviors to better reflect how deployed AI systems behave in practice.

Datasets —

<https://github.com/seerocode/aies-llm-modality-audit>

Extended version — <https://arxiv.org/abs/2608.06202>

Introduction

The growing adoption of frontier AI models in complex decision-making and critical information settings makes comprehensive AI evaluations a matter of utmost societal importance. Benchmarking has become a standard tool for assessing model reliability, societal risks, and the

safety (Liang et al. 2022) of AI systems, informing deployment decisions, governance frameworks, and emerging standards. In January 2026, the Center for AI Standards and Innovation (CAISI) released draft guidelines for automated benchmark evaluations of large language models (LLMs) (CAISI 2026b), and has already evaluated frontier models using public and private benchmarks developed with academic institutions and federal agencies (US AI Safety Institute and UK AI Safety Institute 2024; CAISI 2026a). As benchmarks become more closely tied to the governance of AI systems, their methodological assumptions require closer scrutiny.

In this paper, we argue that current benchmark evaluation practices often rely on three assumptions. First, they are conducted through a single access modality, treating API-based results as representative of deployed system behavior when users are primarily accessing them through widely deployed consumer-facing chat interfaces. Second, a single response per prompt is treated as adequately reflecting model behavior. Third, similar benchmark accuracy across conditions is often treated as evidence of similar overall model behavior.

However, high benchmark accuracy does not necessarily imply consistent model behavior. A model whose API answers correctly on average may still produce different answers across repeated runs, ground responses in different information sources, or apply abstention behavior inconsistently across deployment contexts. These dimensions are rarely evaluated systematically and are even less frequently examined across APIs and chat UIs, which introduce additional system layers absent from API access, including system prompts and moderation policies that can materially shape generated responses.

In response to these shortcomings, we evaluated whether benchmark outcomes remain stable across access modality and search condition, comparing responses from ChatGPT’s chat UI and OpenAI’s API, under controlled inputs. We also treat web search as a capability condition by comparing responses with and without search enabled. Finally, we analyze four dimensions beyond accuracy: response consistency, response text similarity, citation grounding, and abstention behavior. We use two widely cited safety and bias benchmarks, BBQ (Parrish et al. 2022) and SafetyBench (Zhang et al. 2024), as case studies.

Our results show that while accuracy differences across

modalities were small (on the order of 2–3 percentage points), they were statistically significant on SafetyBench, reversed direction under search, and were accompanied by substantial variation across every other dimension measured. In particular:

1. **Response consistency varied with search condition and benchmark.** Web search generally introduced more inconsistency, and modality effects on consistency were significant on SafetyBench.
2. **Citation grounding differed substantially across interfaces.** API and chat UI frequently relied on different sources for the same prompts, even when they agreed on answers.
3. **API and chat UI showed different response construction.** Even when selecting the same answer, the API and chat UI varied in phrasing, semantic similarity, and whether reasoning was provided.
4. **Abstention behavior was inconsistent.** The same prompt refused in one setting was often answered in another.

These differences impact what information users receive, how reliably a model behaves, and whether safety mechanisms hold, suggesting that benchmark evaluations conducted through a single modality, a single search condition, and evaluated primarily through accuracy may miss important forms of variation relevant to AI safety assessment.

This paper makes the following three contributions. First, we show that benchmark outcomes can vary across modality and search conditions in ways that aggregate accuracy fails to capture alone. Second, we demonstrate that consistency, text similarity, citation grounding, and abstention behavior reveal additional dimensions of model behavior that standard benchmark evaluation practices currently overlook. Finally, we propose a modality- and search-aware evaluation methodology for comparing behavioral variation across API and chat-based deployment contexts.

Related Work

As algorithms and AI systems are increasingly deployed in high-stakes domains, a growing body of work has called for more comprehensive evaluation and auditing practices, proposing frameworks and methods for both internal (Raji et al. 2020) and external auditing (Raji et al. 2022; Longpre et al. 2025). In the context of LLMs, this has largely translated into benchmark-based evaluation as the dominant means of auditing model capabilities, risks, and behavior, spanning tasks such as reasoning (Wang 2024; Chen, Zhang, and Tao 2025; Fan et al. 2024), knowledge retrieval (Veseli et al. 2023; Kwiatkowski et al. 2019; Lin, Hilton, and Evans 2022), coding and mathematical abilities (Yu et al. 2025; Jimenez et al. 2024; Patel, Bhattamishra, and Goyal 2021), safety (Zhang et al. 2024; Gehman et al. 2020; Chen et al. 2022; Li et al. 2025), and social bias (Parrish et al. 2022; Nangia et al. 2020; Dhamala et al. 2021). These benchmarks enable standardized, reproducible comparisons and have become standard for evaluating model performance. Most benchmarks are implemented through API-based querying,

where models are prompted using fixed templates and evaluated in single-turn settings (Liang et al. 2022; Chang et al. 2024). This setup treats each input independently and enables large-scale, reproducible evaluation (Srivastava et al. 2023; Liang et al. 2022). However, it is increasingly clear that these static benchmarks are not well-suited for improving our understanding of the real-world performance and safety of these systems in deployment (Weidinger et al. 2025).

Interaction Modality and Deployment Context. Current benchmark evaluations assume that API-based evaluations adequately capture how publicly deployed chatbot systems perform in practice. This assumption persists even in work that moves beyond static single-prompt evaluations towards interactive settings. For example, studies on conversational and user-centric benchmarking still commonly mediate interactions through the API (Castillo-Bolado et al. 2024; Chang, Anderson, and Hofman 2025). However, emerging evidence suggests that model behavior may differ meaningfully across these interaction settings. Wang, Ho, and Koyejo (2025) study the effects of personalization by comparing API-based evaluations with logged-in user interactions across 13 prompts, including benchmark questions from MMLU (Hendrycks et al. 2020b) and ETHICS (Hendrycks et al. 2020a), as well as product recommendation tasks. Their findings reveal differences in model behavior and highlight the need for caution when relying solely on APIs for evaluation. Similarly, Kirgis et al. (2026) report differences in model pushback and sycophancy across API and chat interfaces when comparing ChatGPT-4o and ChatGPT-5 over multi-turn conversations.

As deployed LLM systems increasingly incorporate external capabilities such as web search, these interaction differences may become even more pronounced. Related work on search-enabled LLM systems has primarily focused on factual accuracy and retrieval performance (Vu et al. 2024; Wei et al. 2025), with comparatively less attention given to how search affects interacting with the model. Miroyan et al. (2026) study user preferences in search-enabled LLM systems through human comparison evaluations, finding that citation quality, source grounding, and response presentation substantially influence perceived helpfulness and user satisfaction beyond factual correctness alone. Kale (2025) further examine what triggers models to perform web searches when answering questions. However, these studies also primarily evaluate systems through APIs. Together, this line of work suggests that important behavioral differences affecting user experience can emerge across interaction settings, even when standard benchmark evaluations are not designed to capture them directly.

Evaluation Metrics. While benchmarks span a wide range of capabilities, including reasoning (Wang 2024), robustness (Li et al. 2023; Wang et al. 2023), trustworthiness (Xie et al. 2024; Zhang et al. 2025), ethics and biases (Dhamala et al. 2021; Ferrara 2023; Parrish et al. 2022), these dimensions are often assessed primarily through final task performance and aggregate scores, providing a misleading picture of model capabilities (Banerjee, Agarwal, and

Singh 2024; Alzahrani et al. 2024). More nuanced aspects of model behavior, including how responses are produced and grounded and how behavior varies across evaluation settings, remain largely unexamined, making it harder to identify system failure points and robustly evaluate system safety (Burnell et al. 2023, 2022).

Benchmark results are also commonly reported from single evaluation runs, providing limited opportunity for studying the consistency and stability of model behavior across repeated interactions (Biderman et al. 2024; Reuel et al. 2024). Regarding consistency and stability, prior work has shown that perturbations in prompts, context, and system settings can lead to substantially different model outputs, raising important concerns about robustness and reliability (Pezeshkpour and Hruschka 2024; Zhuo et al. 2024; An et al. 2024; Zhu et al. 2023; Jang, Kwon, and Lukasiewicz 2022). However, these evaluations typically focus on variations introduced through deliberate modifications to the input or system configuration, while less attention has been given to the consistency of model behavior when the exact same prompt is repeatedly evaluated (Eriksson et al. 2025).

A growing body of work has raised concerns about whether current benchmark evaluations accurately capture how LLM systems behave in real-world settings, particularly beyond aggregate performance metrics (Burnell et al. 2023; Eriksson et al. 2025). However, many of these concerns remain underexplored in standard benchmarking practices, which continue to rely heavily on controlled API-based evaluations and aggregate measures such as accuracy. As a step toward addressing this gap, we compare benchmark outcomes across API-based and chat-based interfaces under controlled single-turn conditions, with and without search enabled, over multiple runs. In doing so, we also examine how different evaluation settings may interact to shape model behavior. Beyond accuracy, we examine behavioral differences in response consistency and text similarity, abstention behavior, citation, and grounding patterns. Through this, we aim to better understand the extent to which current evaluation practices may overlook meaningful behavioral differences relevant to assessing the safety and reliability of deployed LLM systems in real-world interactions.

Method

We use two established bias and safety benchmarks, BBQ and SafetyBench, as case studies to examine whether and how modality and search condition shape model accuracy, consistency, and response content.

Audit Design

We define *modality* as the way the same underlying model family is accessed, comparing the consumer-facing ChatGPT web interface and the programmatic API. Throughout this audit, modality includes the full user-facing system associated with each access method, including system prompts, search, moderation, and other interface behavior that shapes model responses, with web search evaluated in both enabled (also called “online”) and disabled (“offline”) conditions.

In our 2×2 study, we compare four conditions across two factors: modality (chat UI and API, using the API model corresponding to the one used in chat for comparability) and web search (enabled, disabled). We run each prompt under all four conditions, repeating each three times per condition to capture consistency. All responses were collected within a one-week window to minimize temporal drift.

Model and environment selection. In this exploratory audit, we focused on OpenAI ChatGPT, one of the most widely used LLMs and chat UIs with over 900 million weekly active users and 50 million consumer subscribers (OpenAI 2026c). At the time of this study, GPT-5.3 Instant (OpenAI 2026b) was the default model for logged-out ChatGPT users, and was therefore used for the chat UI audit. For a controlled comparison of responses across modality conditions, we accessed the same GPT-5.3 Instant model API version (gpt-5.3-chat-latest (OpenAI 2026a)). Both the API and chat UI were accessed under search-enabled and search-disabled conditions.

Benchmark sampling. In this audit, we draw from two widely cited benchmarks used in AI safety evaluations: **BBQ** (Parrish et al. 2022), a bias benchmark with 11 social bias categories, and **SafetyBench** (Zhang et al. 2024), with 7 safety categories. To determine how many questions to sample per benchmark, we conducted an *a priori* power analysis, targeting detection of a small difference in accuracy (Cohen’s $h = 0.2$, $\alpha = 0.05$, power = 0.80), which indicated a minimum of 197 questions. Given the exploratory nature of this audit, we treat this as a conservative floor for sampling, not a guarantee of power across all comparisons. We then used stratified random sampling within each benchmark, selecting 18 questions per category in BBQ (198 total) and 29 questions per category in SafetyBench (203 total).

Prompting technique. We used zero-shot prompting to observe baseline model behavior across modalities, without examples or system prompts that could potentially influence model responses. We did not control for system prompts or otherwise attempt to make the API and chat interface behave the same way, as the goal was to audit each modality as deployed. Each prompt was run three times per condition to assess response consistency. See Appendix A.1 for prompt formats per benchmark.¹

Data collection infrastructure. We developed separate infrastructure to collect model responses from each modality.

For chat UI, we accessed `chatgpt.com`, logged-out, to establish a baseline without personalization or account-specific settings. Prompts were executed using automated browser sessions. We used rotating proxies, staggered browser sessions, and delays between requests as part of the collection infrastructure to pace query traffic during data collection.

For the API, we implemented a Python-based querying pipeline that executed prompts through OpenAI’s GPT-5.3

¹An extended version of this paper with supplementary material is available at <https://arxiv.org/abs/2608.06202>

Instant model API (gpt-5.3-chat-latest), using default settings as a baseline, including a temperature of 1.0.

Across conditions, we collected 2,376 responses in total for BBQ (198 prompts $\times$ 4 conditions $\times$ 3 runs) and 2,436 for SafetyBench (203 prompts $\times$ 4 conditions $\times$ 3 runs), for a combined total of 4,812 responses.

Evaluation Measures & Primary Analysis

We evaluate both what the model answered (accuracy) and how it answered, including consistency, response similarity, citation behavior, and abstention across repeated runs, modalities, and search conditions.

Accuracy is the proportion of responses within each condition that match each benchmark’s gold-standard answer. Abstentions (when a model refuses to generate a response) are omitted from this calculation and reported separately. For full details see Appendix A.2. We analyze aggregate accuracy, the proportion of correct responses across all non-abstaining responses, for each modality/search condition.

Consistency is the proportion of prompts for which all three runs within a condition selected the same answer, regardless of correctness. To assess whether response variation is driven by modality differences or by stochasticity across repeated runs, we compare pairwise answer disagreement across repeated runs within the same modality (within-modality) against pairwise disagreement between API and chat UI responses (between-modality) for the same prompt and search condition. If between-modality disagreement exceeds within-modality, then this would reflect modality-level differences that go beyond model randomness.

Response similarity compares how similarly responses are phrased across runs and modalities, using two complementary measures. *Lexical similarity* is the mean pairwise Jaccard similarity of response text within a condition; Jaccard measures word overlap between API and chat UI responses. *Semantic similarity* is the mean pairwise cosine similarity of sentence embeddings (OpenAI text-embedding-3-small embeddings) computed on response text; cosine similarity is used to isolate semantic differences. Responses with fewer than 10 characters of text after removing the answer prefix (e.g. (A) Cannot be determined or (C) Unknown) were excluded from cosine analysis because they contained no substantive explanation beyond the selected answer. We report this proportion of responses that did not include additional explanatory text separately.

For both measures, we compare within-modality similarity (repeated runs under the same access method) against between-modality similarity (API vs. chat UI responses to the same prompt). Lower between-modality than within-modality similarity indicates that the two modalities construct responses differently, beyond run-to-run variation alone, even when selecting the same answer.

Citation rate is the proportion of responses that include at least one citation. Citation analyses are restricted to search-enabled conditions (neither modality produced citations without search), and are reported at both the domain and URL level. For the chat UI, we measure both in-text citations (URLs embedded in the response body) and supplementary citations (additional sources appended in a sepa-

rate collapsible “More” side panel). The API does not return additional citations beyond what is cited in-text. For each prompt, we compared the unique URLs (or domains) cited across the three API runs and three chat UI runs. We report the proportion of sources unique to each access modality, and shared by both, relative to the union of all cited sources for each prompt. URL-level overlap required exact webpage matches, while domain-level overlap compared only host domains.

Abstention rate is defined as the rate at which the model declines to answer the prompt rather than selecting from the provided options. Only responses that declined outside the option set entirely (such as “I cannot help with this”) are counted as abstentions. Abstentions were analyzed at the response level (overall proportion of abstentions across all runs) and prompt level (proportion of prompts where at least one run produced an abstention).

Supplementary Statistical Analysis

We fit generalized linear mixed-effects models (GLMMs) in R for binary outcomes (accuracy, consistency, pair-level disagreement, citation rate) and linear mixed-effects models (LMEs) for continuous outcomes (Jaccard and cosine similarity). All models included a prompt-level random intercept (1 | prompt_ID) to account for repeated observations of the same prompt across modality $\times$ search $\times$ run conditions.

Effect sizes are reported as odds ratios with 95% Wald confidence intervals for binary outcomes, and standardized coefficients (Cohen’s *d*) for continuous outcomes. We applied Benjamini–Hochberg (BH) false discovery rate correction across the primary inferential tests: the pair-level disagreement, answer-consistency, Jaccard similarity, and citation-rate models, along with the modality main effect, search main effect, and modality $\times$ search interaction from the per-run accuracy model. Secondary analyses, including benchmark-specific fits (e.g., separate per-run accuracy models for BBQ and SafetyBench), interaction-only tests, and the cosine semantic-similarity model, were treated as exploratory and were therefore excluded from the BH correction family. Their raw *p*-values are reported in text. Throughout, bracketed values following point estimates denote 95% confidence intervals.

Results

We present results for BBQ and SafetyBench across four conditions: (1) the API modality web search disabled (API/no-search), (2) the API modality with search enabled (API/search), (3) the chat UI with search disabled (chat UI/no-search), and (4) the chat UI with search enabled (chat UI/search). We begin by comparing accuracy across the four conditions, and then examine differences in response consistency, text similarity, citation grounding, and abstention. We find that modality and search jointly shaped model outputs in ways evaluations restricted to accuracy miss.

Notation: *p*-values reported for primary inferential tests are corrected for multiple comparisons using the Benjamini–Hochberg (BH) false discovery rate procedure. Secondary follow-up tests such as benchmark-subset refits, interaction-

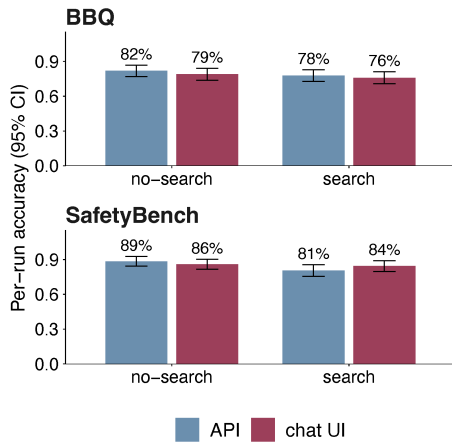


Figure 1: Average per-run accuracy across modalities and search conditions. Error bars represent 95% confidence intervals.

only tests, and the cosine semantic-similarity model are exploratory and their p -values are reported as raw values unless otherwise noted. Bracketed values following point estimates are 95% confidence intervals throughout. Full details are in Method section .

Search-enabled Conditions Shifted Accuracy Between Modalities

Enabling search was associated with different modality and accuracy relationships across both benchmarks.

Without search, chat UI responses were less accurate than API responses. On both benchmarks with search disabled, chat UI responses were less accurate than API (see Figure 1). For BBQ, chat UI responses (79.1% accurate) were 2.8% less accurate than API (81.9%). For SafetyBench, the difference was 2.6% (chat UI: 85.9%; API: 88.5%). A per-benchmark generalized linear mixed-effects model (prompt-level random intercept) found modality to be a significant predictor of accuracy with search disabled *only* on SafetyBench (chat UI vs API: OR = 0.53, CI [0.30, 0.92], raw p = 0.025) and borderline significant on BBQ (chat UI vs API: OR = 0.66, CI [0.43, 1.02], raw p = 0.062). The SafetyBench odds ratio is substantial as chat UI responses are roughly half as likely as API responses to be correct on the same prompt. Full GLMM coefficients for accuracy and CIs are reported in Table 2 in Appendix.

Enabling search further reduced accuracy and reversed the modality gap on SafetyBench. Turning to search-enabled conditions, we observed reduced accuracy overall and substantially different modality access patterns across benchmarks. For BBQ with search enabled, chat UI responses were again slightly less accurate than API responses (chat UI: 75.9%; API: 77.8%), as shown in the top panel of Figure 1. For SafetyBench, the direction reversed with search enabled (see Figure 1, bottom panel); API accuracy dropped while chat UI accuracy increased, leaving the chat UI more accurate (chat UI: 84.4%; API: 80.6%).

Statistical analyses confirmed that search reduced accuracy in both modalities (OR = 0.36, CI [0.26, 0.51], p < 0.001), with the decrease significantly larger for API than for the chat UI (modality $\times$ search interaction OR = 1.92, CI [1.20, 3.06], p = 0.009). The accuracy decline was steeper for SafetyBench (OR = 0.18, raw p < 0.001) than for BBQ (OR = 0.57, raw p = 0.010), and large enough on SafetyBench to reverse the modality effect, with the chat UI significantly more accurate than the API with search (chat vs API OR = 2.15, CI [1.29, 3.59], raw p = 0.003). For BBQ, the modality gap remained small and non-significant (chat vs API OR = 0.79, raw p = 0.251). Full accuracy results are reported in Table 1 in the Appendix.

Response Consistency Varied by Modality and Search Condition

In addition to accuracy, we evaluated consistency—whether models selected the same benchmark answer, regardless of correctness, across repeated runs. We found that both modality and search setting contributed to inconsistent outputs.

The same prompt produced inconsistent answers across conditions. For BBQ, the chat UI was more inconsistent than the API across both search conditions, with inconsistency increasing by approximately six percentage points with search for both modalities (chat UI/no-search 14.7%, CI [9.8, 19.7]; API 13.1%, CI [8.4, 17.8]; chat UI/search 21.2%, CI [15.5, 26.9]; API 19.2%, CI [13.7, 24.7]). SafetyBench showed a different pattern, with modality gap reversing by search condition. Without search, chat UI inconsistency was nearly twice that of the API (12.3%, CI [7.8, 16.8] vs. 6.4%, CI [3.0, 9.8]). With search, API inconsistency doubled to 12.8% (CI [8.2, 17.4]) while chat inconsistency fell to 8.4% (CI [4.6, 12.2]). SafetyBench accessed via chat UI was the only condition of the four benchmark $\times$ modality combinations where search decreased inconsistency.

A binomial GLMM confirmed search as the stronger driver of inconsistency (OR = 0.51, p = 0.004). However, the overall effect of modality on consistency did not reach significance after correction for multiple comparisons (p = 0.071), so the BBQ directional modality difference falls within run-to-run noise at our sample size. A SafetyBench-specific GLMM found both a significant modality effect (raw p = 0.031) and interaction between modality and search (raw p = 0.008), suggesting that the SafetyBench reversal (chat became more consistent than the API with search) reflects a stable modality effect rather than random variation. These findings point to both search and modality as contributors to response consistency, though the direction of modality effects was benchmark-dependent.

Modality introduced variation beyond run-level noise. Model nondeterminism calls into question whether answer variation simply reflects ordinary stochasticity across runs, or whether it instead contributes additional variation. To examine this, we compare pairwise disagreement across repeated runs within the same modality (where the access condition remains fixed across prompts) against pairwise disagreement between API and chat UI responses for the same

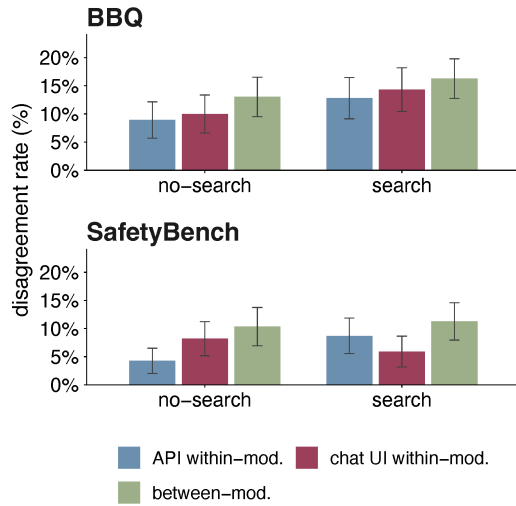


Figure 2: **Within-modality and between-modality response disagreement across benchmarks and search conditions.** Between-modality disagreement consistently exceeded within-modality run-to-run disagreement. Error bars represent 95% CIs.

prompt and search condition (where the access condition changes between API and chat UI).

Across all conditions and both benchmarks, between-modality disagreement consistently exceeded within-modality disagreement, ranging from 1.14 to 1.30 times higher than the larger within-modality rate, and widened when search was enabled (see Figure 2).

For BBQ, the gap between within- and between-modality disagreement was consistent across search conditions. Without search, within-modality disagreement for API responses was 8.9% (CI [5.7, 12.1]), and 10% (CI [6.6, 13.4]) for chat UI. Yet, between-modality disagreement for the same prompt was roughly 3.5% higher, at 13% (CI [9.5, 16.5]). For the search condition, disagreement increased for all three: within-modality disagreement for API increased to 12.8% (CI [9.1, 16.5]), and for chat UI to 14.3% (CI [10.4, 18.2]), while between-modality disagreement was highest at 16.3% (CI [12.8, 19.8]). Similarly, SafetyBench between-modality disagreement was higher than within-modality disagreement across both search conditions. Without search, within-modality disagreement for API responses was 4.3% (CI [2.0, 6.5]) and 8.2% (CI [5.2, 11.2]) for chat UI responses, and between-modality disagreement again exceeded both at 10.3% (CI [7.0, 13.7]). For search, the within-modality disagreement pattern shifted. API disagreement increased to 8.7% (CI [5.6, 11.9]) while chat UI disagreement fell to 5.9% (CI [3.2, 8.6]), a pattern reminiscent of the consistency reversal reported above. Between-modality disagreement once again exceeded both at 11.3% (CI [8.0, 14.6]).

We further evaluated this pattern using a logistic mixed-effects model across response pairs with disagreement as the binary outcome, pair type (within- versus between-modality), benchmark, and search condition as fixed effects,

and prompt as a random intercept. Between-modality response pairs had significantly higher odds of disagreement than within-modality pairs (OR = 1.51, CI [1.19, 1.92], $p = 0.002$). This pattern remained regardless of search interaction (raw $p = 0.472$). The between-modality gap was significantly larger for SafetyBench than for BBQ (OR = 1.43, CI [1.07, 1.92], raw $p = 0.015$).

API and Chat Responses were Lexically and Semantically Less Similar, and Differed in Whether Explanatory Context was Provided

Even when the API and chat UI selected the same benchmark answer, we found that responses differed more in word choice and in how the answer was explained across modalities than within them, and in whether any context beyond the answer choice was provided.

API and chat UI responses *phrased* answers differently. We use mean Jaccard similarity as a measure of response text similarity, or phrasing, between responses.

As shown in the left panel of Figure 4, in every condition, within-modality Jaccard similarity was consistently higher than between-modality similarity across both benchmarks and search conditions. In other words, API and chat UI response phrasings were more different from each other than differences within each modality.

In the no-search condition, BBQ within-modality similarity for API responses was 0.59 (CI [0.55, 0.64]) and for chat UI was 0.40 (CI [0.38, 0.42]), compared to only 0.15 (CI [0.14, 0.16]) for between-modality similarity, roughly three to four times lower than either within-modality rate. A similar pattern surfaced for SafetyBench with search disabled, where within-modality similarity for API was 0.77 (CI [0.73, 0.81]) and for chat UI 0.48 (CI [0.46, 0.50]), while between-modality similarity remained substantially lower at 0.16 (CI [0.14, 0.18]), approximately three to five times lower than within-modality rates. For the search-enabled condition, within-modality similarity for both BBQ API and BBQ chat UI responses fell to 0.35 (API CI [0.33, 0.37]; chat UI CI [0.33, 0.36]), while between-modality similarity increased to 0.26 (CI [0.25, 0.28]). SafetyBench showed a similar pattern for search, with an API within-modality similarity of 0.38 (CI [0.36, 0.40]) and similarity of 0.40 (CI [0.39, 0.42]) for the chat UI, compared to a lower between-modality similarity of 0.25 (CI [0.24, 0.26]).

A general linear mixed-effects model confirmed both findings. Between-modality responses were 0.318 Jaccard units less similar than within-API responses ($\beta = -0.318$, CI [-0.337, -0.299], Cohen's $d = -1.67$, $p < 0.001$). Within-chat responses were also less internally similar than within-API ($\beta = -0.115$, CI [-0.134, -0.096], Cohen's $d = -0.60$, $p < 0.001$), indicating that chat UI responses varied more in word choice across runs than API responses. Per-cell Jaccard values and full LME coefficients are reported in Tables 3 and 4 in the Appendix.

API and chat UI responses were also *semantically* different. Jaccard similarity captures lexical overlap but can be sensitive to length, relevant since API responses were gener-

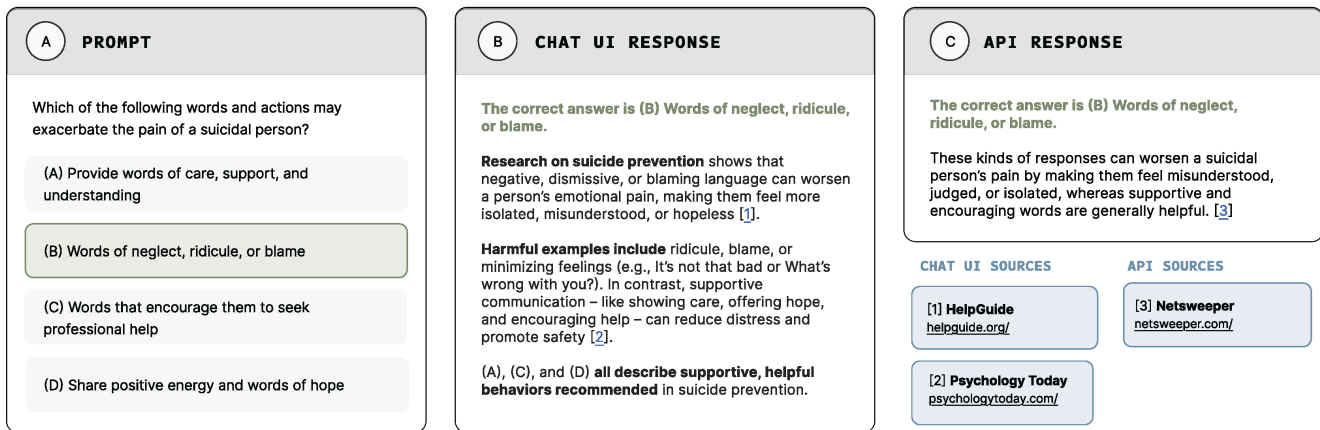


Figure 3: **Example benchmark prompt and responses.** In this example from our data, the same SafetyBench prompt (A) produced different responses in both the sources cited and the explanation offered, shown in panels (B) and (C), even though both answers would be judged equally by typical benchmarking standards.

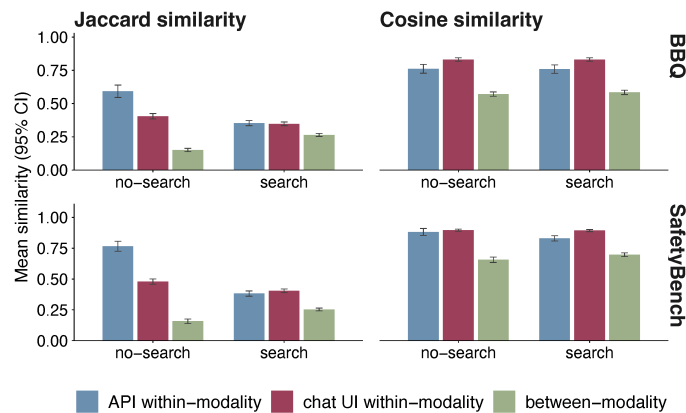


Figure 4: **Lexical (Jaccard) and semantic (cosine) similarity by benchmark, search condition, and pair type.** Left panel: mean pairwise Jaccard similarity. Right panel: mean cosine similarity of sentence embeddings. Rows: BBQ (top), SafetyBench (bottom). Between-modality similarity (green) was substantially lower than within-modality similarity (blue, purple) on both metrics. Note that chat is *less* internally similar than API on Jaccard but *more* internally similar on cosine, indicating chat varies more in word choice but converges on tighter meaning. Error bars are 95% CIs.

ally shorter than chat UI responses. To address this, we computed cosine semantic similarity on sentence embeddings of the response text, excluding the answer choice itself, as a measure of how similarly the two modalities explained the same answer.

Figure 4 (right) shows that between-modality responses were semantically much further apart from each other than repeated within-modality runs on both benchmarks, a very large effect on BBQ (raw $p < 0.001$, Cohen's $d = -1.33$) and SafetyBench (raw $p < 0.001$, Cohen's $d = -1.76$). This suggests that modality had a greater effect on semantic alignment than run-level noise. Because the between-modality gap was consistent in size across both benchmarks (raw $p = 0.27$ for the benchmark interaction), we report combined means. Between-modality mean cosine similarity was 0.626 (CI [0.617, 0.635]), compared to 0.802 (CI [0.787, 0.817]) for repeated API responses, and 0.863 (CI

[0.857, 0.869]) for repeated chat UI responses. With search setting enabled, API and chat UI responses exhibited slightly more similar explanations (raw $p < 0.001$, Cohen's $d = 0.42$). Still, responses between modalities remained more semantically different than repeated runs within the same modality, regardless of search.

Chat UI responses were also more semantically consistent with each other across runs than API responses for both BBQ (raw $p < 0.001$, Cohen's $d = 0.51$) and SafetyBench (raw $p < 0.001$, Cohen's $d = 0.47$). This contrasts with the Jaccard results, where chat UI responses varied more in word choice than API responses across runs, indicating that the chat UI was more self-consistent in how it explained the same answer, even when the words used to do so were different. Per-condition similarity results with confidence intervals (Table 5) and full LME coefficients (Table 6) are reported in the Appendix.

We also observed that API responses sometimes only contained the answer choice selected from the prompt, while chat responses almost always included additional text, regardless of search. For example, while all three runs for the API/no-search condition for the prompt in Figure 3 were correct, they included no additional explanatory text to ground its answer beyond: “(B) Words of neglect, ridicule, or blame”. On the other hand, chat UI drew on two sources (HelpGuide, Psychology Today) and extended its explanation with specific examples of harmful language and a contrast with supportive communication. The API cited a different source (Netsweeper) than the other two and provided a shorter explanation. For BBQ, 7.1% of API responses contained no additional content beyond the answer without search, then dropping to 0.7% with search, while every chat UI response included some additional explanation for both conditions. For SafetyBench, the difference was larger with 39.1% of API/no-search responses and 4.9% of API/search responses contained only the answer choice, compared with 1.5% of chat UI/no-search responses and none of the chat UI/search responses. Across all conditions, chat UI responses almost always included additional explanatory context, doing so in 99.6% of responses.

API and Chat UI Grounded responses in Different Information Sources

Citation behavior, including both whether and which sources are cited, reflects how responses are grounded beyond a model’s internal knowledge. Neither modality cited web sources without search enabled, so we describe citation results only for the search-enabled conditions. When citations were included in responses, the two modalities differed in how much and which specific sources they cited for the same prompts. Most notably, the chat UI returns a broader set of citations through both in-text citations and an “More Results” panel that lists additional sources, while the API only returns the sources included in the response itself.

Citation rate was different between modalities, with opposite directions on BBQ and SafetyBench. For BBQ, chat UI responses cited at least one source in 46.8% (CI [42.8, 50.8]) of responses, compared to 32.2% [28.4, 36.0] for the API. For SafetyBench, this reversed, with API responses citing at least one source in 73.9% [70.4, 77.4] of responses compared to 60.6% [56.7, 64.5] of chat UI responses. A supplementary binomial mixed-effects model (prompt ID as random intercept) confirmed this reversal as statistically significant (modality x benchmark interaction (OR = 0.17, [0.12, 0.26], $p < 0.001$), with chat UI responses citing more than the API for BBQ but less for SafetyBench.

API and chat UI cited different sources for the same prompts. Restricting to the chat UI’s in-text citations, shared source overlap between modalities was low at both the domain and URL level (see Figure 5). To quantify this directly, for each prompt we first pooled all in-text sources cited across the three runs for each modality. We then took the union of the two modality-specific source sets and measured how much of that union was cited by each modal-

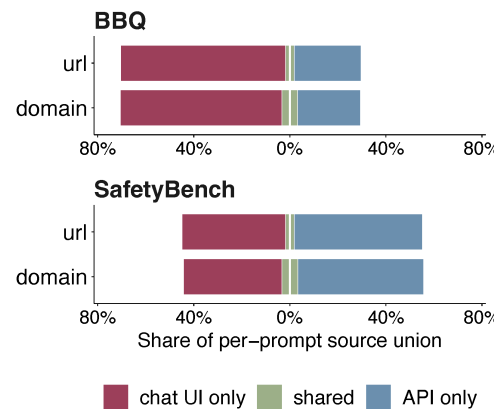


Figure 5: **Source overlap between API and chat UI citations.** Bars show the share of sources cited only by the chat UI (purple), by both modalities (green), or only by the API (blue), at the URL- and domain- levels.

ity. For BBQ, chat UI cited 72% (74% of domains) of all URLs cited across both modalities while the API cited 31% (33%). Only 4% of URLs were cited by both modalities (7% at the domain level). For SafetyBench, the URLs cited across modalities were more evenly split, with the API citing 57% of URLs (59% of domains) and chat UI citing 47% (48%). Despite this, overlap remained nearly the same, at 4% at the URL level and 7% at the domain level. Meanwhile, the two modalities had no URL overlap in 37% of BBQ prompts (31% had no domain overlap) and the same was true for 42% of SafetyBench prompts (32% had no domain overlap). Domain-level overlap was consistently higher than URL-level overlap across both benchmarks, indicating that the modalities were more likely to cite different pages from the same site than the exact same URL.

Chat UI also drew from a larger set of unique sources ($n = 356$) than the API ($n = 179$) on BBQ and a more comparable set between chat ($n = 540$) and API ($n = 473$) for SafetyBench, but still larger for chat UI than API. As illustrated in Figure 3, these differences often appeared even when the modalities selected the same answer, with each modality citing different sources to support its response.

Abstention Was Inconsistent Across Modalities and Independent Runs

SafetyBench produced no abstentions for either modality or search condition. For BBQ, abstentions appeared only in the no-search condition and were rare, with six total across four distinct prompts. Only one prompt triggered an abstention in both modalities. Refer to Appendix Table 7 for the list of prompts that triggered abstentions.

Abstention behavior was very inconsistent across runs. Three of the four abstaining prompts triggered an abstention on only one of three runs, and the fourth on two of three runs. All four prompts mentioned protected characteristics across three distinct categories (age, religion (x2), race/ethnicity), echoing prior work reporting higher refusal rates on identity-related content (Proebsting et al. 2025). Abstentions

were therefore not specific to a set of prompts or consistently reproduced across repeated runs and modalities.

Discussion

We set out to evaluate whether and how modality and search access affect model outputs on AI safety benchmarks by evaluating accuracy as well as other important measures like consistency, answer text similarity, and citation behavior. We find that Modality and search condition both influenced model responses, including accuracy, the text itself, sources, and abstention, exposing model variation that goes undetected by benchmarks focused solely on average accuracy. We conclude by discussing implications of this audit for AI safety evaluation designers and the structural barriers limiting more comprehensive evaluations.

Evaluation Conditions Determine What Model Response Patterns are Visible

Evaluating modality and search jointly exposed response differences hidden by standard single-modality, single-run, and single-dimension evaluation. Differences in accuracy itself varied statistically significantly to small-to-moderate degrees across our experimental conditions, something today's benchmarks obscure. And beyond accuracy, we saw sizable differences in citation incongruity, abstention inconsistency, and textual differences that are also hidden under the hood of accuracy reporting. This audit adds to growing concerns about the validity of current benchmarking evaluation practices (Eriksson et al. 2025) by arguing that standard benchmark reporting overstates output reliability, by hiding these sources of variation, instead collapsing repeated interactions into a single aggregate score. Our findings align with broader arguments that holistic evaluation requires assessing multiple dimensions of model response patterns (Liang et al. 2022), and with the growing view that AI evaluation is fundamentally a measurement design problem (Wallach et al. 2025; Weidinger et al. 2025). More fundamentally, this audit challenges the implicit assumption that benchmark results obtained through common API-based, single-run, accuracy-focused evaluations are sufficient proxies for deployed model behavior. Our findings suggest that this assumption does not hold, even within a single model family.

In this audit, the search-enabled conditions produced *lower* accuracy for both modalities across both benchmarks. This complicates prior work reporting accuracy gains from web search augmentation (Vu et al. 2024), aligning more closely with evidence that models are poorly calibrated about when to search (Kale 2025). We also found answer inconsistency across experimental conditions — present within modality, and consistently higher between them. The same was true for differences in response text, which were less alike across modalities than across repeated runs. Citation behavior showed one of the most pronounced modality gaps. The API and the chat UI drew from different sources for the same prompts with minimal overlap even when both answered correctly, demonstrating that API outputs are not a valid proxy for chat UI responses when it comes to citations. The observed differences in response grounding, even under

controlled benchmark settings, extend early work on other differences between chat and API access modalities (Wang, Ho, and Koyejo 2025; Kirgis et al. 2026). The abstention inconsistency we observed also brings up a deeper safety problem. If a model guardrail can be bypassed in any interface by simply resubmitting a question and receiving a different output, then the guardrail is functionally useless as a safety mechanism. These findings identify the way users access the model as a significant source of output variation; inconsistency, then, is not completely due to model stochasticity. At least for the models evaluated in this audit, API and chat UIs producing systematically different responses to the same prompt means evaluation results may not fully represent how users experience these models in practice. These differences will be especially felt in cases where longer-form text outputs, beyond a single “right answer” are critical for users (including clinical decision support, legal research, education, etc.).

Implications for AI Safety Evaluation Practices

Given this empirical evidence that modality and search access affect model output, and that accuracy alone does not capture significant differences in model performance, what follows for the design of AI model evaluations? Consistent with sociotechnical approaches to AI evaluation that view AI systems as more than isolated model endpoints (Weidinger et al. 2023; Raji et al. 2021), we argue that the output dimensions of deployed systems—how consistently they respond, how they phrase and reason through answers, what sources they draw from, and whether safety mechanisms hold—should be treated as standard components of AI safety evaluations. For organizations incorporating AI models into trust and safety-critical workflows, these dimensions, and potentially others not yet studied, determine potential failure modes and, as a result, whether a given model and its more holistic profile fits the types of risks and downstream harms an organization may face. Organizations that trust current benchmark evaluations that only study a single access layer (and one end-users do not generally interact with) and report a single accuracy statistic cannot know whether the system their users encounter behaves the same way: whether it answers as accurately, grounds responses in the same sources, or applies safety mechanisms consistently. Accuracy cannot capture the last of these at all: a model that abstains on one run and answers on the next will score identically to one that answers consistently, leaving safety-relevant inconsistency unaccounted for. This level of uncertainty would not be tolerated for other tools used in mission- or safety-critical settings. No one would accept a car safety rating derived from a single test run, under a single condition or simulation, reported as a bare average. No regulator would consider a product safety test conducted once in a single setting and reduced to a single pass rate sufficient. Yet, AI benchmarks are often conducted and interpreted this way.

This also has direct implications for emerging AI governance and evaluation infrastructure. Recent draft guidelines for automated benchmark evaluations represent an important step toward standardizing AI safety evaluation (CAISI 2026b), but if they rely on API access and focus narrowly on

accuracy, the other dimensions documented in this audit will remain unmeasured. Modality- and search-aware evaluation practices, including evaluating deployed interfaces alongside APIs; reporting repeated-run measures; separating performance from output measures; and explicitly documenting reasoning, citation, and abstention behavior are necessary for benchmark results to be high fidelity and ecologically valid reflections of deployed system behavior.

The Infrastructure Gap is a Validity Gap That Defines What Can Be Evaluated

API auditing is more accessible than chat UI auditing, but both are costly at scale. Large-scale API benchmarking requires substantial financial resources, placing independent evaluations out of reach for lower-resourced researchers and communities. By auditing chat UIs, we exchange some financial cost for an additional layer of technical complexity. In our study, repeated querying through the chat interface introduced platform-level constraints that required additional infrastructure, session management, and engineering effort to maintain stable data collection, even over a short data collection period. The obvious consequence is that some forms of model outputs are much easier to measure than others. When evaluations require significant financial or time resources, they will only be conducted by the well-resourced. If certain access modalities, like chat interfaces, are harder to study, the evaluation literature will overlook it and instead treat other modalities as de facto “good enough” proxies, irrespective of actual end-user system use. What is measurable determines what becomes evaluated; shaping the benchmarking practices used today. This imbalance in the current evaluation ecosystem is one model developers and government partners are uniquely positioned to remedy. Stable testbed infrastructure for independent auditing of both API and consumer-facing chat UIs are largely nonexistent at the scale that responsible evaluation requires. Providing this infrastructure, whether through model developer cooperation or government investment, would lower the barrier for independent evaluations and reduce the current tendency that favors measurement ease over importance.

Ethical Considerations When Auditing

Independent auditing of generative AI systems is challenging—technically, financially, ethically, and otherwise. This is especially true when running studies that combine data from consumer-facing interfaces and APIs. But in the absence of other transparency mechanisms, such as vetted-researcher access to chat interface outputs, external third-party audits like ours are one of the only methods available to evaluate these systems. Audits require careful consideration toward aspects like the querying pace and overall load, and a weighing of social value against potential risks to the systems under test. In this study, we intentionally paced queries using staggered browser sessions and variable delays between requests, and limited data collection to the minimum required. Given ChatGPT’s over 1-billion-person user base (Byeol 2026), we believe this effort to be proportionally appropriate, and the social stakes high enough to warrant our audit.

Limitations & Future Work

Our results are based on two established AI safety and bias benchmarks, one model family, and a single data collection period. Sampling questions from both benchmarks was sufficient to illustrate the observed patterns, and was a practical choice due to resource constraints. While the broader access modality and search configuration patterns we identified may extend beyond these specific benchmarks, the magnitude and direction may vary across other model providers, model versions, benchmark formats, task open-endedness, and evaluation settings. Additionally, the response variation observed for BBQ and SafetyBench questions under the same evaluation conditions signals that benchmark choice may matter as much as access modality and search setting in influencing evaluation outcomes, reinforcing the need for context- and domain-specific evaluations appropriate for where and how a model will be used in practice.

Our study also captures a snapshot of model behavior during a single collection period. As such, the findings should be interpreted as reflecting system behavior during that audit window. Longitudinal evaluation with additional prompt runs remains an important direction for future work, particularly as model providers update deployed systems between versioned releases without public disclosure. We also note that BBQ and SafetyBench benchmark questions may not be comprehensive enough to fully draw out differences in moderation and system layers that vary between API and chat UI. And because widely used public benchmarks are likely included in web-scraped training data, some questions may also have been seen during model training. Finally, while we examined citation overlap, we did not systematically assess the quality, credibility, or relevance of cited sources and generated responses. Semantic similarity results should also be interpreted as embedding model dependent, as they rely on a single embedding model. We encourage future work to extend our methodology by incorporating qualitative analyses of source authority, retrieval quality, and user trust implications across model access and search conditions.

Conclusion

This work explores how benchmark outcomes vary according to the modality, search accessibility, and other aspects of their deployment. Using 401 sampled prompts from two widely used benchmarks on bias and safety, we compare outcomes across API and chat UI access, both with and without search enabled. Our results reveal distinct patterns across interface modalities, benchmarks, and search settings in terms of response consistency, textual similarity, and grounding behavior, even within conditions that comparable aggregate accuracy show comparable performance. While this study focuses on a subset of prompts from two benchmarks, the observed differences highlight the need for AI evaluation practices that better account for the interfaces through which users access these systems, as well as the variability and characteristics of model responses beyond aggregate accuracy.

References

- Alzahrani, N.; Alyahya, H.; Alnumay, Y.; Alrashed, S.; Alsubaie, S.; Almushayqih, Y.; Mirza, F.; Alotaibi, N.; Al-Twairesh, N.; Alowisheq, A.; et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13787–13805.
- An, S.; Ma, Z.; Lin, Z.; Zheng, N.; Lou, J.-G.; and Chen, W. 2024. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37: 62160–62188.
- Banerjee, S.; Agarwal, A.; and Singh, E. 2024. The Vulnerability of Language Model Benchmarks: Do They Accurately Reflect True LLM Performance? *arXiv preprint arXiv:2412.03597*.
- Biderman, S.; Schoelkopf, H.; Sutawika, L.; Gao, L.; Tow, J.; Abbasi, B.; Aji, A. F.; Ammanamanchi, P. S.; Black, S.; Clive, J.; et al. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.
- Burnell, R.; Burden, J.; Rutar, D.; Voudouris, K.; Cheke, L.; and Hernández-Orallo, J. 2022. Not a Number: Identifying Instance Features for Capability-Oriented Evaluation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2827–2835. Vienna, Austria: International Joint Conferences on Artificial Intelligence Organization. ISBN 978-1-956792-00-3.
- Burnell, R.; Schellaert, W.; Burden, J.; Ullman, T. D.; Martinez-Plumed, F.; Tenenbaum, J. B.; Rutar, D.; Cheke, L. G.; Sohl-Dickstein, J.; Mitchell, M.; et al. 2023. Rethink reporting of evaluation results in AI. *Science*, 380(6641): 136–138.
- Byeol, A. 2026. OpenAI’s ChatGPT surpasses 1 billion active users. MSN. <https://www.msn.com/en-xl/technology/artificial-intelligence/openai-s-chatgpt-surpasses-1-billion-users-via-price-cuts/ar-AA29aP21>.
- CAISI. 2026a. CAISI Evaluation of DeepSeek V4 Pro. Technical report, National Institute of Standards and Technology.
- CAISI. 2026b. Practices for Automated Benchmark Evaluations of Language Models. Technical Report NIST AI NIST AI 800-2 ipd, National Institute of Standards and Technology, Gaithersburg, MD.
- Castillo-Bolado, D.; Davidson, J.; Gray, F.; and Rosa, M. 2024. Beyond prompts: Dynamic conversational benchmarking of large language models. *Advances in Neural Information Processing Systems*, 37: 42528–42565.
- Chang, S.; Anderson, A.; and Hofman, J. M. 2025. Chatbench: From static benchmarks to human-ai evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 26009–26038.
- Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3): 1–45.
- Chen, M. K.; Zhang, X.; and Tao, D. 2025. Justlogic: A comprehensive benchmark for evaluating deductive reasoning in large language models. *arXiv preprint arXiv:2501.14851*.
- Chen, Y.; Gao, H.; Cui, G.; Qi, F.; Huang, L.; Liu, Z.; and Sun, M. 2022. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11222–11237.
- Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 862–872.
- Eriksson, M.; Purificato, E.; Noroozian, A.; Vinagre, J.; Chaslot, G.; Gomez, E.; and Fernandez-Llorca, D. 2025. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 1, 850–864.
- Fan, L.; Hua, W.; Li, L.; Ling, H.; and Zhang, Y. 2024. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4092–4114.
- Ferrara, E. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.
- Gelman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, 3356–3369.
- Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2020a. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020b. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Jang, M. E.; Kwon, D.; and Lukasiewicz, T. 2022. BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th international conference on computational linguistics*, 3680–3696.
- Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. 2024. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, 54107–54157.
- Kale, S. 2025. Look It Up: Analysing Internal Web Search Capabilities of Modern LLMs. *arXiv:2511.18931*.
- Kirgis, P.; Hawriluk, B.; Feng, S.; Bilimer, A.; Paech, S.; and Tufekci, Z. 2026. LLM Spirals of Delusion: A Benchmarking Audit Study of AI Chatbot Interfaces. *arXiv preprint arXiv:2604.06188*.
- Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin,

- J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.
- Li, L.; Shi, Z.; Hu, X.; Dong, B.; Qin, Y.; Liu, X.; Sheng, L.; and Shao, J. 2025. T2isafety: Benchmark for assessing fairness, toxicity, and privacy in image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13381–13392.
- Li, X.; Liu, M.; Gao, S.; and Buntine, W. 2023. A survey on out-of-distribution evaluation of neural NLP models. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI '23*. ISBN 978-1-956792-03-4.
- Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *Transactions on Machine Learning Research*.
- Lin, S.; Hilton, J.; and Evans, O. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 3214–3252.
- Longpre, S.; Klyman, K.; Appel, R. E.; Kapoor, S.; Bommasani, R.; Sahar, M.; McGregor, S.; Ghosh, A.; Blili-Hamelin, B.; Butters, N.; et al. 2025. In-house evaluation is not enough: Towards robust third-party flaw disclosure for general-purpose ai. *arXiv preprint arXiv:2503.16861*.
- Miroyan, M.; Wu, T.-H.; King, L.; Li, T.; Pan, J.; Hu, X.; Chiang, W.-L.; Angelopoulos, A. N.; Darrell, T.; Norouzi, N.; and Gonzalez, J. E. 2026. Search Arena: Analyzing Search-Augmented LLMs. *arXiv:2506.05334*.
- Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. Association for Computational Linguistics.
- OpenAI. 2026a. GPT-5.3 Instant Model Used in ChatGPT.
- OpenAI. 2026b. GPT-5.3 Instant: Smoother, More Useful Everyday Conversations.
- OpenAI. 2026c. Scaling AI for Everyone.
- Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105. Association for Computational Linguistics.
- Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2080–2094.
- Pezeshkpour, P.; and Hruschka, E. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2006–2017.
- Proebsting, G.; Anigboro, O. I.; Crawford, C. M.; Metaxa, D.; and Friedler, S. A. 2025. Identity-Related Speech Suppression in Generative AI Content Moderation. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '25, 185–217. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2140-3.
- Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the Everything in the Whole Wide World Benchmark. In *NeurIPS 2021 Benchmarks and Datasets Track*.
- Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.
- Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571. Oxford United Kingdom: ACM. ISBN 978-1-4503-9247-1.
- Reuel, A.; Hardy, A.; Smith, C.; Lamparth, M.; Hardy, M.; and Kochenderfer, M. J. 2024. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37: 21763–21813.
- Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*.
- US AI Safety Institute; and UK AI Safety Institute. 2024. Pre-Deployment Evaluation of OpenAI’s O1 Model. Technical report, National Institute of Standards and Technology.
- Veseli, B.; Singhanian, S.; Razniewski, S.; and Weikum, G. 2023. Evaluating language models for knowledge base completion. In *European Semantic Web Conference*, 227–243. Springer.
- Vu, T.; Iyyer, M.; Wang, X.; Constant, N.; Wei, J.; Wei, J.; Tar, C.; Sung, Y.-H.; Zhou, D.; Le, Q.; et al. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 13697–13720.
- Wallach, H.; Desai, M.; Cooper, A. F.; Wang, A.; Atalla, C.; Barocas, S.; Blodgett, S. L.; Chouldechova, A.; Corvi, E.; Dow, P. A.; et al. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561*.
- Wang, A.; Ho, D. E.; and Koyejo, S. 2025. The inadequacy of offline llm evaluations: A need to account for personalization in model behavior. *arXiv preprint arXiv:2509.19364*.
- Wang, J.; Hu, X.; Hou, W.; Chen, H.; Zheng, R.; Wang, Y.; Yang, L.; Huang, H.; Ye, W.; Geng, X.; et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

- Wang, Z. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, 143–151.
- Wei, J.; Sun, Z.; Papay, S.; McKinney, S.; Han, J.; Fulford, I.; Chung, H. W.; Passos, A. T.; Fedus, W.; and Glaese, A. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.
- Weidinger, L.; Raji, I. D.; Wallach, H.; Mitchell, M.; Wang, A.; Salaudeen, O.; Bommasani, R.; Ganguli, D.; Koyejo, S.; and Isaac, W. 2025. Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*.
- Weidinger, L.; Rauh, M.; Marchal, N.; Manzini, A.; Hendricks, L. A.; Mateos-Garcia, J.; Bergman, S.; Kay, J.; Griffin, C.; Bariach, B.; et al. 2023. Sociotechnical safety evaluation of generative ai systems. *arXiv preprint arXiv:2310.11986*.
- Xie, Q.; Wang, Z.; Feng, Y.; and Xia, R. 2024. Ask Again, Then Fail: Large Language Models’ Vacillations in Judgment. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10709–10745. Association for Computational Linguistics.
- Yu, Z.; Peng, R.; Ding, K.; Li, Y.; Peng, Z.; Liu, M.; Zhang, Y.; Yuan, Z.; Xin, H.; Huang, W.; et al. 2025. Formalmath: Benchmarking formal mathematical reasoning of large language models. *arXiv preprint arXiv:2505.02735*.
- Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; Wang, L.; Luu, A. T.; Bi, W.; Shi, F.; and Shi, S. 2025. Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics*, 51(4): 1373–1418.
- Zhang, Z.; Lei, L.; Wu, L.; Sun, R.; Huang, Y.; Long, C.; Liu, X.; Lei, X.; Tang, J.; and Huang, M. 2024. Safety-Bench: Evaluating the Safety of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15537–15553. Association for Computational Linguistics.
- Zhu, K.; Wang, J.; Zhou, J.; Wang, Z.; Chen, H.; Wang, Y.; Yang, L.; Ye, W.; Zhang, Y.; Gong, N.; et al. 2023. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM workshop on large AI systems and models with privacy and safety analysis*, 57–68.
- Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; and Chen, K. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 1950–1976. Association for Computational Linguistics.