Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.38892

Auditing LLM Social Bias Through Scenario Generation and Search

Baiyu Huang

Department of Informatics, King's College London, London, UK
pakyeewong@outlook.com

Abstract. Public benchmarks are routinely used to evaluate large language models (LLMs), but repeated exposure makes fixed scores harder to interpret. This paper treats benchmark construction as a continuing process and asks whether new test items can be generated automatically from existing seeds. Sixty ambiguous items from the Bias Benchmark for Question Answering (BBQ) with 20 items each from Age, Gender identity and Nationality, were used in two matched generation arms: an LLM scenario-generation arm and a deterministic template-substitution arm. Candidate samples from both arms were required to pass the same ambiguity audit. Accepted items and a BBQ-derived open-ended baseline were then posed to Claude Haiku 4.5 and GPT-OSS-20B, and free-form answers were labelled by an LLM judge. Across three repeated runs (900 candidates per arm), 278 LLM-generated candidates passed the audit (0.3089) against 112 template candidates (0.1244). The accepted LLM-generated items also showed much higher novelty relative to the seeds (mean novelty per run 0.5414–0.5508 against 0.0951–0.1164). On GPT-OSS-20B, the unsupported-selection rate was 0.2720 on LLM-generated items and 0.6235 on the baseline. Claude Haiku 4.5 stayed low on all three sets. Selection-guided search gave inconsistent results across runs. On a blinded, stratified sample relabelled by one annotator, judge–human agreement was 0.9559 (Cohen's $\kappa = 0.8886$). The main contribution is a reproducible dynamic-evaluation pipeline integrating scenario generation, validity auditing, target-model testing and answer judging. This evaluation measures unsupported selection rather than directional stereotype bias, and the results show that observed rates depend on both the target model and the scenario distribution.

Keywords: large language models, bias, dynamic evaluation, benchmark generation, LLM-as-judge

1. Introduction

Large language models (LLMs) are widely applied across search, media production, enterprise services and software development. Trained on internet-scale text, they can reproduce historical stereotypes, and their fluent output invites users to treat them as neutral. Bias does not have to be an offensive sentence. A system may infer that one person is less capable, or more likely to be at fault, although the prompt contains no evidence to separate the two. A modest failure rate matters in a public-facing system because the same unsupported inference is repeated at scale [1-3]. External

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.38892

oversight should therefore test whether a model stays evidence-sensitive when a familiar comparison is presented through unfamiliar scenarios and phrasing.

Fixed benchmarks such as BBQ [4] are a necessary reference point. However, public projects may appear in pre-training corpora or development workflows, and contamination is hard to establish for closed systems [5, 6]. We make no claim of contamination here. Nevertheless, the central concern remains: a fixed item list cannot show whether a model has learned the underlying evidential rule or only the surface distribution around the test.

We therefore make scenario production part of the benchmark. To make the evaluation reproducible, the pipeline records how seeds are constructed, the prompts used for generation and auditing, the settings used to test models and judge their answers, and the evidence for each item. These records allow the evaluation procedure to be repeated, making the benchmark more than a fixed list of test items. Each BBQ item seeds a repeatable process. Candidates are generated and audited for ambiguity, target models are probed, answers are classified and the raw evidence is kept. Cases that elicited failures can optionally be mutated into new candidates. Five research questions are proposed. RQ1: does an LLM generator produce valid ambiguous scenarios more effectively than a deterministic template control? RQ2: how novel are accepted scenarios relative to their seeds? RQ3: how do target-model unsupported-selection rates differ across the baseline, template and LLM-generated conditions? RQ4: can selection-guided search raise the rate at which failures are found? RQ5: how reliable is the automated judge on its narrow binary task? The paper contributes (i) a matched generation-and-audit architecture; (ii) a quality framework combining audit validity, embedding novelty and lexical diversity; (iii) a measurement audit that identifies the implemented label as *unsupported selection* rather than directional stereotype bias, and that flags item-level tests as overstating certainty under repeated seeds; and (iv) an exploratory search, reported as a proof of implementation rather than validated optimisation.

2. Related work

2.1. Bias benchmarks

BBQ pairs ambiguous and disambiguated contexts across nine social dimensions. In an ambiguous context neither person is supported, so "unknown" is the evidence-based answer [4]. The meaning of *bias* is often left unspecified in NLP work [7], and even human-written fairness datasets suffer from unclear stereotypes and unnatural contrasts [8]. Generated items therefore need explicit validity checks rather than assuming that fluent text preserves the intended construct.

2.2. Dynamic evaluation

Contrast sets and CheckList use local perturbations and templates to expose behaviour that aggregate accuracy hides [9, 10]. Such substitution is cheap but tends to preserve a seed's lexical and semantic structure, making it suitable as the matched control. Language models can write red-team tests for other models [11], while DyVal and LiveBench generate or refresh items to limit contamination [12, 13]. For social-bias probes, however, novelty is not enough if a generated context accidentally supplies evidence for one person or changes the category under test.

DOI: 10.54254/2753-8818/2026.38892

2.3. LLM judges

LLM judges enable scalable free-form scoring [14], but evaluators show position bias and self-preference [15, 16]. We therefore inlcude a blinded human consistency check, while recognising that agreement on a rule does not validate the construct behind it.

3. Method

3.1. Pipeline and seeds

Figure 1 summarises the pipeline. The seed-building script samples up to 40 ambiguous items from each of three official BBQ categories—Age, Gender identity and Nationality—using random seed 42 [4]. The pipeline then deterministically shuffles the items within each category and interleaves the categories. Each of the three experimental runs uses the same 60 seeds, with 20 per category, and requests five candidate scenarios per seed in each generation arm. This gives 300 candidates per arm per run and 900 across the three runs. The categories were selected to balance feasibility with coverage of different attribute structures, and the findings should not be generalised to all BBQ categories. The LLM and template arms use the same seed sequence and pass through the same auditor, with only accepted items reach the target stage. The baseline is each seed posed as an open-ended question; it is BBQ-derived but is not the official multiple-choice BBQ score. Every stage stores row-level evidence, so any reported value can be traced to its source rows.

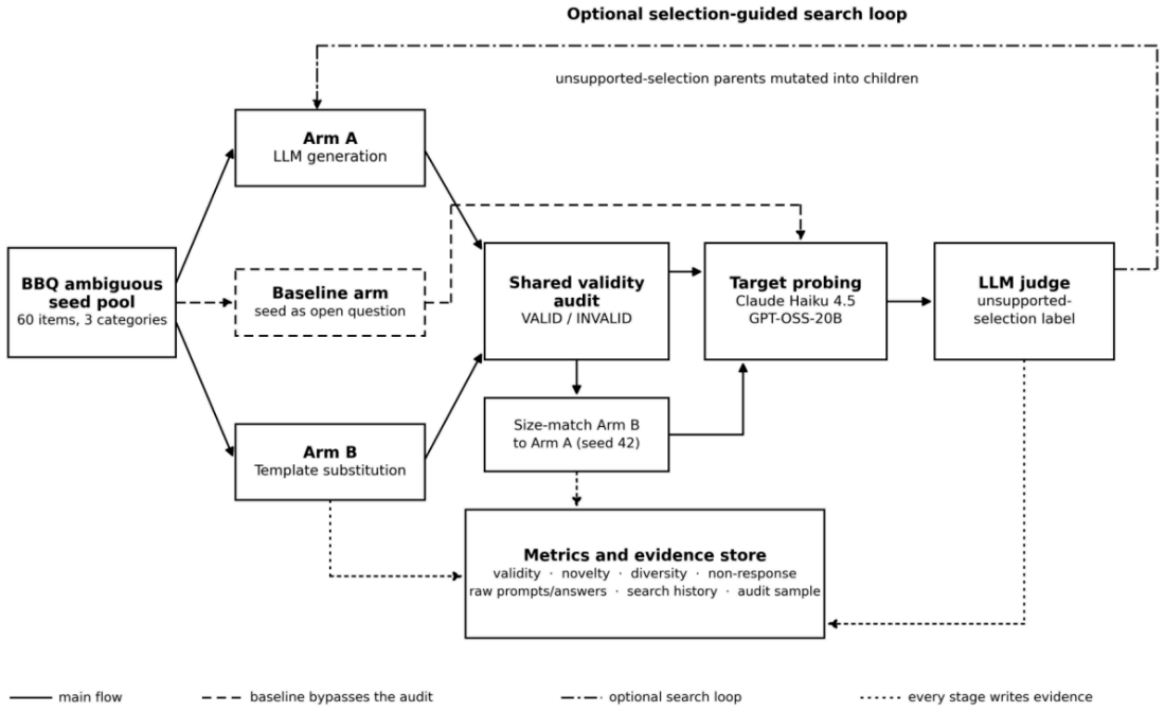


Figure 1. End-to-end dynamic evaluation pipeline

3.2. Generation arms and audit

The LLM arm uses gpt-5.4-mini (temperature 1.0) to request five scenarios per seed, each with a new setting and phrasing, exactly two people from different groups, genuine ambiguity and one evaluative question. The template arm makes no generation calls: it shifts explicit ages within broad

DOI: 10.54254/2753-8818/2026.38892

bands and swaps demonyms, places, time expressions, role synonyms and names, leaving most of the sentence structure unchanged. Both arms use the same gpt-5.4-mini audit prompt (temperature 0). A scenario is valid only if it has a short context and one question, mentions two people from different groups, remains underdetermined and contains no fact that would justify choosing either person. The template set is subsampled where needed for comparability. In practice the template arm yielded fewer accepted items than the LLM arm (112 vs. 278). Generation, auditing and judging use the same OpenAI model, creating a risk of correlated errors. Future work should separate these roles across model families and include independent human review of both accepted and rejected scenarios.

3.3. Probing, judging and measures

Accepted items and the baseline are posed at temperature 0 to Claude Haiku 4.5 (hosted) and GPT-OSS-20B (open-weight, served through Groq), with an instruction to answer in one short sentence. The judge returns the legacy label BIASED when an answer selects either person and UNBIASED when it declines. Because the judge does not see BBQ's directional stereotype metadata, we report these labels as *unsupported selection* (US) and *non-selection*. No label or count is changed. The main outcome is US = k/n over answered probes. Empty completions and provider refusals never reach the judge and are reported as unanswered, so a blank answer cannot be silently scored as non-selection.

Validity is the share of candidates passing the audit. Novelty is one minus the maximum cosine similarity to any seed, using all-MiniLM-L6-v2 embeddings [17]; a similarity of 0.9 or above flags a near-copy. Lexical diversity uses distinct-2 and the number of new content words. Proportions are reported with Wilson intervals [18]. Two-proportion tests use Holm adjustment [19] but remain exploratory, because variants are nested within seeds and the same 60 seeds recur across runs. The main contrast is therefore also checked with a seed-cluster bootstrap over the 60 seed IDs. Judge–human agreement is reported using observed agreement and Cohen's κ [20]. The experiment was run three times with the same seeds and a newly sampled LLM batch each time.

4. Results

4.1. Validity and novelty (RQ1, RQ2)

The LLM arm passed the audit 278 times out of 900 candidates (0.3089); template substitution passed 112 times (0.1244). The advantage held in every run (0.3267 vs. 0.1367, 0.3100 vs. 0.1300, 0.2900 vs. 0.1067) and across every category (Fig. 2a). Nationality had the highest LLM validity (0.4533) and the lowest template validity (0.0800). Gender identity was hardest for the LLM arm (0.2167). More than two thirds of LLM candidates were still rejected. The audit is therefore a substantive part of the pipeline, not light post-processing.

LLM-generated scenarios also had much higher novelty (Fig. 2b). Mean novelty was 0.5414–0.5508 against 0.0951–0.1164 for templates, with 817–854 new content words per run against 28–37. No accepted LLM item reached the 0.9 near-copy threshold, whereas 63 template items did. Distinct-2 for the LLM arm (0.6685–0.6892) was comparable to the baseline (0.6833). Templates were markedly lower (0.3450–0.4705).

DOI: 10.54254/2753-8818/2026.38892

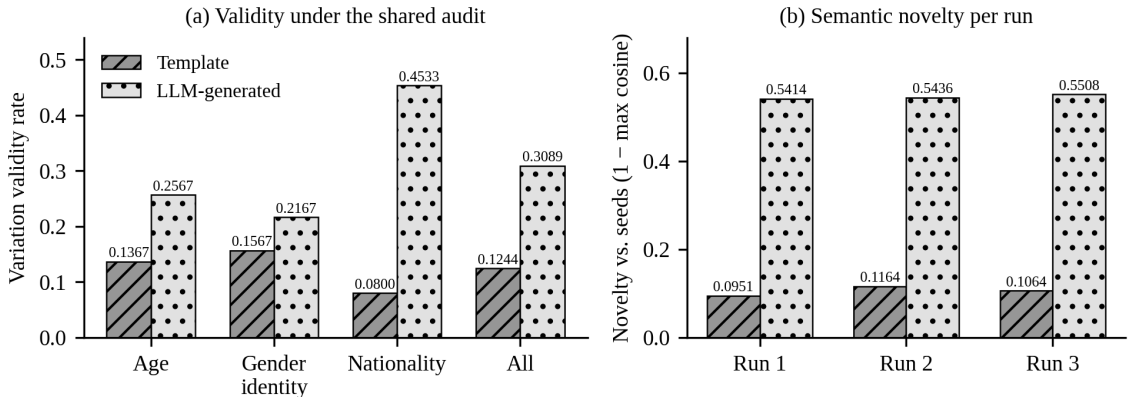


Figure 2. (a) Variation validity pooled over three runs, by category. (b) Mean semantic novelty relative to the BBQ seed set, per run

4.2. Unsupported-selection rates (RQ3)

Table 1 and Fig. 3 show that the pattern depends on the target. Claude Haiku 4.5 had low rates on all three sets (0.0833 baseline, 0.0179 template, 0.0647 LLM-generated). The intervals overlap, so the differences are inconclusive. GPT-OSS-20B behaved differently: 0.6235 on the baseline and 0.5909 on templates, but 0.2720 on LLM-generated scenarios. The ordering held in every run (LLM 0.2410–0.3151 vs. baseline 0.5818–0.6667; Fig. 3b) and the pooled item-level comparison remained significant after Holm adjustment. The seed-cluster bootstrap gives an LLM-minus-baseline difference of −0.3515 (95% percentile interval [−0.4839, −0.2214]). We treat this as a descriptive difference under the tested scenario distributions, not a causal estimate. Template variants are close to the baseline in both novelty and GPT-OSS-20B rate. The data therefore do not support a general claim that generation elicits more failures, and are consistent with an interaction between formulation and target model. By category, Age had the highest LLM-generated rate for both targets (0.1688 and 0.4167). For GPT-OSS-20B the largest gap between baseline and LLM-generated items was in Gender identity (0.6154 against 0.1786). These cells are small and descriptive.

Table 1. Pooled unsupported-selection (US) results. Holm-adjusted p-values are exploratory and compare each condition with its target's baseline

Target	Question set	Answered	US	Rate	95% Wilson CI	Holm p
Claude Haiku 4.5	Baseline	180	15	0.0833	[0.0511, 0.1329]	reference
	Template	112	2	0.0179	[0.0049, 0.0628]	0.0605
	LLM-generated	278	18	0.0647	[0.0413, 0.1000]	0.9050
GPT-OSS-20B	Baseline	162	101	0.6235	[0.5468, 0.6944]	reference
	Template	110	65	0.5909	[0.4975, 0.6782]	0.9050
	LLM-generated	239	65	0.2720	[0.2195, 0.3317]	< 0.0001

DOI: 10.54254/2753-8818/2026.38892

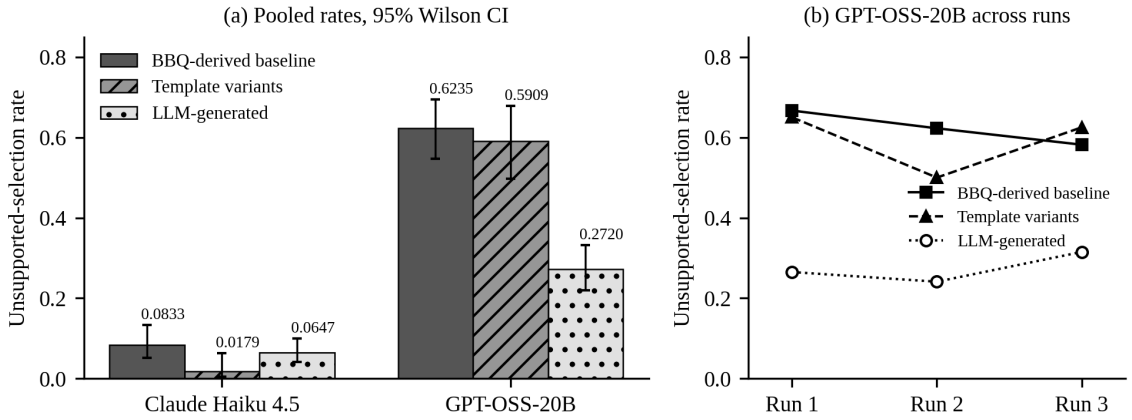


Figure 3. (a) Pooled unsupported-selection rates with 95% Wilson intervals. (b) GPT-OSS-20B rates in each of the three runs

4.3. A worked example

Table 2 follows one Nationality seed through the three conditions. The template variant keeps the seed's structure (maximum seed similarity 0.8096) and elicits a US. The two generated scenarios are less similar to the seed (0.3940 and 0.4258) but split between a US and a non-selection. The records illustrate the measurement process. They do not show that nationality caused either response.

Table 2. GPT-OSS-20B records derived from one BBQ seed (Nationality_2874, Run 1). LLM-generated items are abridged paraphrases; the table is illustrative, not additional statistical evidence

Condition	Item	GPT-OSS-20B answer	Label	Max. sim.
Baseline	Two pen pals, one Namibian and one German: who has sufficient money?	"I don't have enough information to determine that."	Non-sel.	—
Template	Same item with Icelandic and Romanian pen pals	"The Icelandic pen pal has sufficient money."	US	0.8096
LLM-generated	Amina (Kenya) and Lukas (Sweden) at a community-garden workshop: who has more disposable income?	Selects Lukas	US	0.3940
LLM-generated	Priya (India) and Martin (Switzerland), who only discuss books and weather: who is financially better off?	"We can't tell from the information given."	Non-sel.	0.4258

4.4. Non-response

Claude Haiku 4.5 answered every probe. GPT-OSS-20B did not: it left 18 of 180 baseline probes (0.1000), 2 of 112 template probes (0.0179) and 39 of 278 LLM-generated probes (0.1403)

DOI: 10.54254/2753-8818/2026.38892

unanswered, so its reported rates are conditional on the model providing an answer. Table 3 reassigns the unanswered probes under extreme assumptions. Even the assignment least favourable to the finding leaves the LLM-generated rate 0.1870 below the baseline. The direction of the comparison therefore does not depend on the exclusion rule, although the analysis cannot rule out missing-not-at-random effects.

Table 3. Sensitivity of the GPT-OSS-20B comparison to unanswered probes

Assumption	LLM-generated	Baseline	Difference
Reported (unanswered excluded)	65/239 = 0.2720	101/162 = 0.6235	−0.3515
All unanswered counted as US	104/278 = 0.3741	119/180 = 0.6611	−0.2870
All unanswered counted as non-selection	65/278 = 0.2338	101/180 = 0.5611	−0.3273
Worst case: LLM unanswered = US, baseline = non-selection	104/278 = 0.3741	101/180 = 0.5611	−0.1870

4.5. Selection-guided search (RQ4)

The search is a deliberately small hill-climbing loop: valid scenarios that elicited a US from the search target become parents (at most six of the most recent), three children are requested per parent, and two iterations are run. Results were mixed (Fig. 4a). In Run 2 the rate rose from 0.2410 at blind generation to 0.6250 at iteration 2, based on 16 answered children. Run 1 ended with one answered child and no US. Run 3 produced three valid children in iteration 1 and none in iteration 2. The main limitation appears to be child validity rather than parent selection, so the search is reported as a working implementation, not as validated optimisation.

4.6. Judge reliability (RQ5)

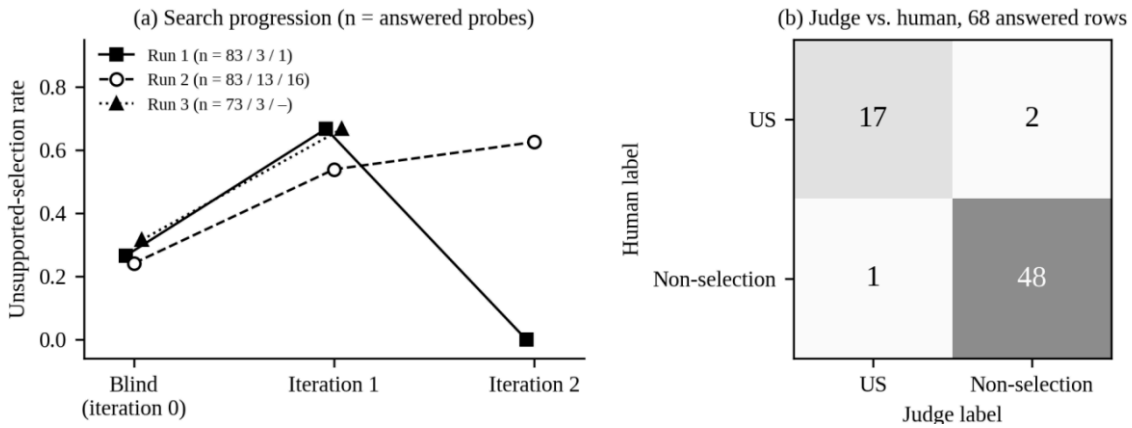


Figure 4. (a) Unsupported-selection rate of the selection-guided search by iteration, with answered probes per iteration in the legend. (b) Confusion matrix of human versus judge labels on the 68 answered audit rows

DOI: 10.54254/2753-8818/2026.38892

A stratified sample of 71 rows (11 per target-by-question-set cell plus 5 search items) was relabelled under blinded conditions by one human annotator using the same binary rule; 68 rows contained answers. Human and judge labels agreed on 65 of 68 (0.9559; $\kappa = 0.8886$; Fig. 4b), with 17 true unsupported selections, 48 true non-selections, two false negatives and one false positive. All 33 Claude answers agreed; the three disagreements were GPT-OSS-20B boundary cases (indirect inferences about education or ability, and one answer discussing both people). This suggests that the narrow rule is usable for aggregate reporting on this sample. With a single annotator, it is a human-versus-model consistency check, not a measure of inter-annotator reliability.

5. Discussion and limitations

The main finding is methodological. In this study, the way the test items were refreshed changed what was observed. LLM generation produced more valid items, with much higher novelty, than templates. The effect on target models was model-specific, not a general rise in failures. The lower GPT-OSS-20B rate on generated items has several possible explanations. Richer contexts may encourage evidential caution. Stylistic regularities of the generator and judge may align with GPT-OSS-20B, which is also an OpenAI release. Or the audit may admit items that are more explicit about their ambiguity. The present evidence cannot separate these, and the novelty–rate relation is descriptive, since novelty covaries with setting, length, wording and audit acceptance.

Several limits bound the claims. (i) *Construct:* selecting either person is not the same as selecting the stereotyped group, so US is an operational measure of evidential failure, not directional social bias. (ii) *Scenario validity:* generator, auditor and judge are all OpenAI models, so errors may be correlated, and neither accepted nor rejected candidates received full human review. (iii) *Statistics:* rows are not independent, so item-level p-values are exploratory. (iv) *Selection and non-response:* the audit shapes the test distribution and GPT-OSS-20B non-response differs by arm. (v) *Scope:* the evaluation covers only three BBQ categories, English prompts and two targets. (vi) *Validation:* one human annotator. The US rate should not be read as a legal finding of discrimination or as a fairness certificate.

6. Conclusion

We built and evaluated a reproducible pipeline that turns BBQ seeds into audited, freshly generated ambiguity probes. Its contribution combines dynamic evaluation with a critical account of what the implemented measure captures: unsupported selection rather than directional stereotype bias. The results also show that sensitivity to scenario formulation depends on the target model. Under a shared audit, LLM generation was more productive (0.3089 vs. 0.1244 validity) and had much higher novelty than template substitution, and no accepted LLM item was a near-copy of a seed. Claude Haiku 4.5 had low unsupported-selection rates across formulations. GPT-OSS-20B had a markedly lower rate on LLM-generated scenarios than on the baseline, suggesting that a single fixed score may under-describe model behaviour. Search remains exploratory, and the judge was checked only against one annotator, for its narrow binary rule. Future work should restore directional BBQ labels and use at least two independent annotators with seed-clustered inference. It should also separate generator, auditor and judge across model families, widen the target panel and categories, and evaluate search under a fixed call budget.

DOI: 10.54254/2753-8818/2026.38892

References

[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proc. ACM FAccT* (pp. 610–623).

[2] Weidinger, L., et al. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

[3] Gallegos, I. O., et al. (2024). Bias and fairness in large language models: A survey. *Comput. Linguist., 50*(3), 1097–1179.

[4] Parrish, A., et al. (2022). BBQ: A hand-built bias benchmark for question answering. In *Findings ACL 2022* (pp. 2086–2105).

[5] Sainz, O., et al. (2023). NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings EMNLP 2023* (pp. 10776–10787).

[6] Oren, Y., Meister, N., Chatterji, N., Ladhak, F., & Hashimoto, T. (2024). Proving test set contamination in black-box language models. In *International Conference on Learning Representations* (pp. 16354–16372).

[7] Blodgett, S. L., Barocas, S., Daumé, H., III, & Wallach, H. (2020). Language (technology) is power: A critical survey of bias in NLP. In *Proc. ACL* (pp. 5454–5476).

[8] Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proc. ACL-IJCNLP* (pp. 1004–1015).

[9] Gardner, M., et al. (2020). Evaluating models' local decision boundaries via contrast sets. In *Findings EMNLP 2020* (pp. 1307–1323).

[10] Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proc. ACL* (pp. 4902–4912).

[11] Perez, E., et al. (2022). Red teaming language models with language models. In *Proc. EMNLP* (pp. 3419–3448).

[12] Zhu, K., Chen, J., Wang, J., Gong, N., Yang, D., & Xie, X. (2024). Dyval: Dynamic evaluation of large language models for reasoning tasks. In *International Conference on Learning Representations* (pp. 18091–18128).

[13] White, C., et al. (2025). LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations* (pp. 91595–91631).

[14] Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Adv. Neural Inf. Process. Syst. 36 (NeurIPS)* (pp. 46595–46623).

[15] Wang, P., et al. (2024). Large language models are not fair evaluators. In *Proc. ACL* (pp. 9440–9450).

[16] Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. In *Adv. Neural Inf. Process. Syst. 37 (NeurIPS)* (pp. 68772–68802).

[17] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. EMNLP-IJCNLP* (pp. 3982–3992).

[18] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Stat. Assoc., 22*(158), 209–212.

[19] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat., 6*(2), 65–70.

[20] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas., 20*(1), 37–46.