ChatGPT

Research Plan for Reviewing AI Model Evaluation Practices

1. **Define scope and objectives.** Clarify that the review will cover both text-only and multimodal AI models, and that you want to assess how benchmarks are designed, their limitations, and their real-world relevance. Formulate key questions (e.g. "What are the major evaluation benchmarks and metrics used today? What known flaws do they have? Do they generalize to real tasks or scientific use?"). Noting recent studies, you should aim to include critiques of benchmarks (biases, documentation gaps, data contamination, gaming, etc.) [1] [2]. This step guides your search and helps keep the review focused on "good vs bad", contamination, and real-world effectiveness.

2. **Literature survey.** Conduct a systematic search of recent literature on AI/LLM evaluation. Start with comprehensive surveys and meta-reviews such as Chang *et al.* (2023) and Eriksson *et al.* (2025), which cover evaluation tasks, metrics, and known issues [2] [1]. Key sources include LLM evaluation surveys, papers on benchmark design (e.g. HELM, BigBench, MMLU, etc.), and articles on evaluation pitfalls. Also look at papers on data contamination (e.g. Wang *et al.* 2024 [3]) and benchmarking in specific domains. Assemble a bibliography of dozens of relevant studies (the Chang *et al.* survey even maintains a GitHub list of works [2]). This survey will reveal common themes and point to concrete examples (e.g. we cited that benchmarks often suffer "imperfect gold standard" labels [4], and that most critique is gathered in meta-analyses [1]).

3. **Categorize evaluation flaws.** From the literature, identify and group the main weaknesses of current benchmarks. For example:

4. **Data/Dataset Issues:** Investigate biases or spurious cues in benchmark data. Prior work notes that careless dataset construction lets models exploit unintended signals (e.g. a medical image model learned to detect a chest drain instead of disease [5]). Also note documentation gaps – many benchmarks lack clear provenance or difficulty rubrics [5] [6]. Consider challenges like unreliable or "noisy" annotation: as Hu *et al.* (2024) point out, many tasks have an "imperfect gold standard" (misassigned labels or flawed reference answers) which undermines validity [4].

5. **Metric Limitations:** Evaluate whether the chosen metrics truly measure desired capabilities. Common metrics (accuracy, F1, BLEU, ROUGE, etc.) have known statistical and conceptual issues [4]. For instance, Hu *et al.* note that benchmarks rarely report confidence intervals or uncertainty, so we don't know if a high score is statistically significant [7]. Check if metrics align with scientific standards (e.g. do they account for multiple correct answers, or capture novel reasoning?).

6. **Benchmark Scope and Diversity:** Examine which domains and modalities are covered. Many benchmarks focus narrowly on English text tasks, neglecting multimodal inputs (images, audio, video) and non-English languages [6]. Safety/ethics benchmarks often rely on simplistic proxies (e.g. crowdworker opinions) and miss broader context [8] [9]. Critics argue benchmarks are mostly static "one-time" tests (e.g. MCQs) that do not capture dynamic, interactive use-cases [9]. Summarize calls for more holistic evaluations (longitudinal studies, multi-turn interactions, etc. [9]).

7. **Gaming and Overfitting:** Consider how models may game benchmarks. For example, "sandbagging" (intentional underperformance) has been observed when developers hide dangerous capabilities [10]. Also watch for leaderboard chasing: corporate and community incentives push teams to optimize scores rather than robustness [1] [11]. Note any documented cases where repeated runs or different random seeds produce wide score variance [12] – failure to report this suggests scores may not reflect true performance.

8. **Contamination and Data Leakage:** Focus especially on "benchmark data contamination," where models have seen the test data during training [3] [13]. Summarize survey findings (e.g. Wang *et al.* 2024 define BDC and warn it undermines trust in evaluations [3]). Collect examples of contamination: for instance, Narayanan & Kapoor (2023) showed GPT-4 solved all old Codeforces problems but **none** of the new problems released after its training cutoff [14]. This strongly implies the model memorized the earlier problems. Such case studies help illustrate why contamination is a serious flaw.

9. **Empirical checks (if feasible).** If resources allow, perform a small-scale analysis to illustrate one or more issues. For example, you might test an open-source LLM on a known benchmark both with and without data contamination. One practical experiment: pick a dataset with a clear publication date, query an LLM (or multiple models) on pre- vs post-cutoff splits to check for memorization (as above [14]). Alternatively, compare benchmark scores to human or gold-standard performance on certain tasks (to reveal metric gaps). However, given a 2-week timeline, the focus should be on secondary analysis: leverage existing reported results rather than running large evaluations from scratch. For instance, you could survey leaderboards or PaperswithCode to identify inconsistencies or missing uncertainty reporting [15] [12].

10. **Assess relevance to real-world/scientific use.** Evaluate how well benchmarks reflect practical or scientific tasks. Look for domain-specific benchmarks: e.g. CORE-Bench (Siegel *et al.* 2024) is a new benchmark of 270 tasks drawn from scientific papers, meant to test reproducibility (a core scientific task) [16]. Its low scores (best agent ~21% on hard tasks [16]) suggest current models struggle with real research tasks. Include such examples to argue that standard NLP or vision benchmarks are often poor proxies for real-world utility. Also cite critiques pointing out that famous benchmarks (like bar exams, ethics quizzes) can emphasize the wrong skills [8]. This will help you argue for more appropriate evaluation design for science and industry.

11. **Synthesize findings and write the review.** Organize the paper around the themes above. For example, sections might include *Benchmark Design & Scope*, *Data and Contamination Issues*, *Metric and Evaluation Limitations*, *Case Studies of Flaws*, and *Recommendations*. In each section, use the gathered literature to make strong, cited points. For instance, note that Eriksson *et al.* (2025) compile "nine reasons to be cautious with benchmarks," including data biases, contamination, misaligned incentives, etc. [1]. Illustrate points with concrete studies (e.g. "as shown in medical imaging and code evaluation cases [5] [14]"). Conclude by highlighting open problems (like building benchmarks that reflect real scientific workflows) and suggesting improvements (e.g. dynamic evaluation, uncertainty-aware metrics, better documentation [9] [4]).

12. **Plan timeline.** With a 2-week deadline, schedule roughly one week for intensive reading and note-taking, and one week for writing and editing. Use existing benchmark results and documented studies as much as possible to avoid lengthy experiments. Keep writing concise and well-cited. By systematically following the above steps, you will produce a thorough review of the state of AI evaluation practices, grounded in the latest research.

Each of these steps is supported by recent literature. For example, comprehensive surveys like Chang *et al.* (2023) and Eriksson *et al.* (2025) provide overviews of evaluation methods and pitfalls [2] [1]. Studies on data contamination and metric failures (e.g. Wang *et al.* 2024 [3]; Hu *et al.* 2024 [4]; Narayanan & Kapoor 2023 [14]) give concrete examples of the issues. By combining insights from these sources, your review can critically assess current benchmarks and recommend how to make evaluations more robust, transparent, and relevant to real-world science and use cases [9] [16].

Sources: Recent surveys and critiques of AI benchmarks [1] [2]; studies on data contamination [3] [14]; research on benchmark design flaws and evaluation metrics [5] [4] [9]; domain-specific benchmarks like CORE-Bench [16].

---

[1] [5] [6] [8] [9] [10] [11] [12] [13] [14] [15] Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation
https://arxiv.org/html/2502.06559v1

[2] [2307.03109] A Survey on Evaluation of Large Language Models
https://arxiv.org/abs/2307.03109

[3] Benchmark Data Contamination of Large Language Models: A Survey
https://arxiv.org/html/2406.04244v1

[4] [7] Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions11footnote 1This research was funded by supported by National Key R&D Program of China (No. 2021YFF0901400)
https://arxiv.org/html/2404.09135v1

[16] [2409.11363] CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark
https://arxiv.org/abs/2409.11363