BenchNavigator: Standardizing Benchmark Discovery and Selection for LLM Evaluation

Anonymous ACL submission

Abstract

Evaluating the effectiveness of large language models (LLMs) requires selecting appropriate benchmarks for the intended use case. Due to rapid growth in benchmarks, identifying and selecting the most appropriate one is difficult. Thousands of candidates now exist with heterogeneous metadata, variable documentation quality, and inconsistent terminology. Practitioners must assemble information across papers, repositories, and dataset cards, which increases cognitive load and slows decisions. Prior efforts improved single benchmark documentation and quality assessment, but they do not provide a uniform way to view and compare benchmarks for selection. In this paper, we start by surveying practitioners to analyze multi-source benchmark metadata and characterize what information needs to be unified for effective discovery. We identify the critical fields users seek, document the costs of fragmented presentation, and highlight systematic inconsistencies across sources. We introduce **BenchNavigator**, a prototype that normalizes heterogeneous metadata into a coherent representation, preserves provenance, and exposes an interface designed around practitioner priorities. Our results provide an empirical basis for harmonizing the presentation of benchmark metadata and show that a standardized view is feasible without imposing new burdens on benchmark producers.

1 Introduction

Large language models are increasingly used as components in products and workflows, for tasks such as summarization, question answering, coding assistance, customer support, and decision support. Because these systems can produce fluent but incorrect outputs, and because their behavior can shift across domains, prompts, and deployment settings, practitioners working in the field of AI need systematic ways to understand what a model can and cannot do before relying on it.

Evaluation is the mechanism that turns model choice into an evidence-based decision. In practice, evaluation supports at least four recurring needs: selecting between model candidates, verifying that a model meets requirements for a particular use case, tracking regressions as prompts and models change, and identifying risks such as unreliability or unsafe behavior. Benchmarks play a central role in this process because they provide repeatable tests: a benchmark packages a dataset (or set of tasks), a protocol, and metrics so that different models and settings can be compared under the same conditions. However, benchmarks are only useful when they are appropriate for the decision at hand. A benchmark that is misaligned with a target use case can create false confidence, for example, by overemphasizing generic skills while missing domain constraints, languages, formats, or safety requirements that matter in deployment. Poorly documented or poorly implemented benchmarks can also waste time and compute, and can lead teams to draw conclusions that are not reproducible. In short, the risk is not just no evaluation, it is evaluation that looks rigorous but answers the wrong question.

The rapid growth in the number of LLM benchmarks has created a practical problem for evaluation workflows. Practitioners struggle to discover and compare which benchmark is appropriate for a given use. For a single benchmark, finding the primary paper and a runnable implementation can require sorting across papers, repositories, and dataset cards. The same artifact is often described with different schemas, fields, and terms across sources. This fragmentation turns a selection task into a time consuming search across incompatible representations.

Recent work addresses related aspects of the evaluation ecosystem. BenchmarkCards (Sokol et al., 2024) propose structured documentation for individual benchmarks. BenchHub (Kim

et al., 2025) aggregates domain specific evaluations. Quality frameworks such as BetterBench (Reuel et al., 2024) analyze benchmarks along multiple dimensions. Studies report practitioner needs (Hardy et al., 2025) and evaluation challenges (Chang et al., 2024). These advances improve single artifacts and quality assessment, but they do not solve the presentation gap: even with clear documentation, users still face too many options that are inconsistently exposed.

At the same time, recent research has questioned whether benchmarks remain a reliable paradigm for evaluating modern AI systems. Several studies highlight problems such as benchmark saturation, distribution shifts, and the risk that models optimize for benchmark performance without reflecting real world capabilities. For example, Burnell et al. (2023) discuss structural limitations of benchmarks and argue that benchmark scores alone cannot capture the complexity of real deployments (Burnell et al., 2023). Similar concerns are raised in broader analyses of AI evaluation practices, which emphasize reproducibility, robustness, and contextual validity as ongoing challenges (Eriksson et al., 2025). Our work does not resolve this debate or claim that benchmarks are the right evaluation paradigm. Instead, we address a practical problem: how practitioners can find and compare benchmarks efficiently despite growing fragmentation.

It is also important to note that not all practitioners rely on public benchmarks. In many industry settings, teams develop internal or task specific benchmarks that reflect proprietary data and product constraints. BenchNavigator is not intended to replace those internal practices. Rather, it is designed for the subset of practitioners who consult public benchmarks and need help identifying which ones may be relevant for their use case.

Unlike prior work that focuses solely on understanding practitioner needs or evaluating benchmark quality, we specifically address the standardization challenge. We investigate not just *what* information practitioners need, but *how* that information must be presented consistently to enable efficient discovery and comparison. Crucially, in our framing, standardization *encompasses* (i) schema alignment and controlled vocabularies, (ii) normalization and reconciliation of conflicting or missing fields, and (iii) a uniform *presentation layer* for decision support. Presentation is thus the visible tip of a deeper alignment process.

Our focus on standardization reveals unique challenges: reconciling conflicting metadata across sources, handling missing information systematically, and establishing common vocabularies for comparison. Accordingly, we investigate three research questions: [**RQ1:**] What are the current practices, workflows, and challenges that practitioners face during the discovery and selection of LLM benchmarks? [**RQ2:**] What informational attributes and quality indicators do practitioners prioritize when evaluating and selecting LLM benchmarks for their specific use cases? [**RQ3:**] Can the metadata categories identified by practitioners be unified into a single interface from heterogeneous sources, and what design challenges emerge?

Our contributions are threefold. We characterize benchmark metadata fragmentation through practitioner input and multi-source analysis, derive empirical requirements for standardized metadata presentation that supports selection, and demonstrate feasibility through a functional prototype that unifies heterogeneous representations while maintaining provenance and exposing reasoning paths. Together, these results suggest design implications for discovery interfaces that complement, rather than replace, existing benchmark documentation and quality assessment.

The prototype source code is not publicly available at this stage due to institutional constraints during double-blind review. To support assessment of the design and functionality, we provide detailed screenshots of all interface components in the Appendix A and describe the full user workflow below.

2 Related Work

The benchmark ecosystem has evolved rapidly, with efforts addressing quality, documentation, and practitioner needs. Figure 1 shows this growth from three popular sources: ArXiv, GitHub, and Hugging Face. This growth has been matched by universal calls for systematic evaluation from government and standards bodies, including the U.S. National Institute of Standards and Technology (NIST), which emphasizes benchmark-driven assessment in its AI Risk Management Framework and ongoing initiatives for generative AI evaluation (AI, 2023).

However, the *standardization* problem still persists: how to present benchmarks uniformly regardless of source? We position our work relative to these complementary research directions.

Evaluation Suite Development. Early critiques revealed that single benchmarks often fail to capture model capabilities (Raji et al., 2021). The community developed comprehensive evaluation frameworks like HELM (Liang et al., 2022), Dynabench (Kiela et al., 2021), and Robustness Gym (Goel et al., 2021). While these suites broaden coverage, they also worsen the standardization problem by increasing the number of evaluation options. Our work addresses how to present these diverse options through a unified lens.

Understanding Practitioner Needs. Hardy et al. (Hardy et al., 2025) interviewed practitioners and found that benchmarks rarely inform deployment decisions due to a lack of real-world relevance. Chang et al. (Chang et al., 2024) surveyed evaluation challenges across "what," "where," and "how" dimensions. We build on these findings but focus on a different problem: how to present benchmark information consistently so practitioners can actually access what they need. For example, knowing that practitioners need contamination information is valuable; we investigate how to standardize contamination reporting across sources that document it differently.

Quality and Contamination Assessment. The community has documented numerous benchmark quality issues, for example, prompt sensitivity and option ordering affect can affect measurements (Mizrahi et al., 2024; Alzahrani et al., 2024). Data contamination inflates scores artificially (Zhou et al., 2023; Xu et al., 2024). BetterBench formalized quality assessment across multiple dimensions (Reuel et al., 2024). These efforts evaluate benchmark quality but do not address how quality information is scattered across incompatible formats. We incorporate quality dimensions into our standardization framework, ensuring practitioners can compare quality indicators regardless of the source.

Documentation Standards and Completeness BenchmarkCards provide templates for structured documentation (Sokol et al., 2024). BenchHub aggregates benchmarks with domain classifications (Kim et al., 2025). These efforts improve individual benchmark documentation but assume practitioners will adopt new standards. Our approach is complementary but distinct: rather than requiring benchmarks to adopt standard documentation, we unify existing heterogeneous representations post-hoc. This pragmatic approach works with the existing ecosystem. Empirical evidence confirms that voluntary documentation standards alone do not ensure completeness. Liang et al. (Liang et al., 2024) systematically analyzed over 32,000 AI model cards and found that most practitioners leave critical fields unfilled. Bracamonte et al. (Bracamonte et al., 2023) found that non-experts perceived full model cards as less understandable than shorter versions, suggesting a tension between completeness and usability. These findings motivate our post-hoc aggregation approach: rather than relying on producers to fill all fields, BenchNavigator aggregates what exists and makes gaps explicit.

2.1 The Standardization Gap

Prior work has improved benchmark quality assessment, documentation templates, and understanding of practitioner needs. However, no research has specifically addressed the cognitive burden imposed by fragmented benchmark presentations. Studies document what information practitioners need, but not how to present it uniformly. Standards propose ideal documentation, but cannot retroactively fix thousands of existing benchmarks. Quality assessments evaluate individual benchmarks but not how to compare quality across different reporting formats.

Our work fills this gap by focusing on standardization for discovery and comparison. Unlike efforts to assess benchmarks or understand needs, we address the practical problem of helping practitioners navigate existing fragmentation. We investigate dimensions unique to standardization: reconciling conflicting metadata, handling missing fields systematically, establishing comparison vocabularies.

3 Qualitative Analysis

Our paper combines two complementary studies: a survey examining practitioner selection practices and priorities, and an analysis of benchmark metadata characterizing the information ecosystem practitioners navigate. We then confirm that the findings can be operationalized through a prototype recommendation system.

3.1 Semi-structured Interviews

As part of preliminary work, prior research conducted ten semi-structured interviews with practitioners working in the field of AI (Anonymous,

2024).[1] The purpose was to understand how practitioners currently approach benchmark selection and what challenges they encounter. The interviews explored strategies for finding benchmarks, information considered essential for decision-making, and barriers that complicate comparison across options.

The preliminary interviews surfaced recurring pain points that later informed the survey design. Participants described frustration with missing metadata, inconsistent reporting, and unclear applicability of benchmarks. One practitioner noted, *"You can find dozens of benchmarks, but you never know what environment they expect or how much compute they need."* Another emphasized the lack of transparency around contamination: "*We stopped using one dataset after realizing it was probably in the training mix, but no one tells you that upfront.*" Concerns also extended to reproducibility. As one participant explained, "*The paper says the benchmark tests reasoning, but there are no details about the prompt format or evaluation script - you can't reproduce results without guessing.*"

These accounts illustrate why practitioners often rely on ad-hoc heuristics rather than systematic comparison and highlight these gaps our broader survey aimed to quantify. Interview themes directly informed both the survey instrument and the prototype design. We translated recurring pain points into survey items to measure prevalence (e.g., missing fields, inconsistent reporting, setup friction, contamination concerns), and we mapped those same themes to BenchNavigator features (e.g., cards, provenance trails, filters).

3.2 Practitioner Survey

We designed a survey instrument (see Appendix B) to investigate current benchmark selection practices, priorities, and challenges. The survey addresses three core questions: how practitioners currently discover and select benchmarks, what constraints shape their decisions, and what information they need but cannot currently access. IRB approval was obtained prior to recruitment. Participants were recruited via email through university networks and industry contacts, and all participants provided informed consent before beginning the survey. We surveyed **23** practitioners from **five** countries, drawn primarily from **academia** with additional responses from **industry** and non-profit. The average ML/AI experience was **4.6 years**. We report qualitative patterns given the small number, emphasizing robust themes over exact percentages. The instrument covers demographics, current practices, constraints, priorities, and trust in evaluation methods (see Appendix D for section-by-section design rationale).

3.3 RQ1: How Practitioners Currently Use Benchmarks

Typically, practitioners use benchmarks at 2 points in their research: early scoping, when they are mapping the problem and shortlisting models, and mid-project, when head-to-head comparisons determine a direction of the right choice for their project. Discovery is fragmented (see Appendix C). People triangulate across papers, Hugging Face, and GitHub, with informal recommendations filling the gaps; community leaderboards and vendor materials play a smaller role.

There are two recurring themes that were often mentioned by survey participants. First, *choice overload*: respondents describe the ecosystem as saturated, making it hard to separate canonical tests from near-duplicates. Second, *operational friction*: setup and execution break more often than they should, especially when implementations lag behind library updates or assume implicit environment details. Trust is further complicated by selective reporting, respondents notice vendors highlighting favorable results while leaving out less flattering tests. Together, these dynamics push teams toward a small working set of familiar benchmarks and ad-hoc internal checks rather than systematic exploration.

3.4 RQ2: What Practitioners Value When Selecting Benchmarks

Across responses, the selection logic is pragmatic and layered:

Core scientific signals. The quality of benchmarks construct validity (i.e., measuring the intended capability) and reliability are non-negotiable: a benchmark must measure what it claims and yield stable results. Annotation quality matters when humans are in the loop. Contamination checks, difficulty calibration, and basic statistical hygiene are valued, but respondents do not expect every benchmark to

[1] We reference this prior interview study because it informed the motivation and survey design for the present work. The study remains anonymized under double-blind review.

excel along every dimension; they do expect disclosures that let them judge fit.

Operational fitness. Ease beats elegance. Teams prefer benchmarks that are ready to run in common harnesses, come with copy-paste commands, and document seeds, versions, and prompts. Clear runtime and cost expectations help with planning; containerized or API-first options reduce integration risk. In practice, latency, memory footprint, and budget are weighed alongside accuracy.

External signals and coverage. Community recognition functions as a shortcut: cited, widely used benchmarks with active maintenance inspire confidence. For coverage, domain relevance is the gatekeeper; human-evaluation traces and language coverage are strong positives. Respondents want to know *what a benchmark actually tests* and *why that matters* for their application, not just that it is popular. To validate that empirically-derived requirements can be implemented, we developed a prototype recommendation system. The system operationalizes findings from the survey and metadata analysis into a functional tool for benchmark discovery.

Freshness and reliability expectations. Benchmarks are expected to evolve. Respondents favor regular updates to mitigate contamination and gaming. As a rule of thumb, small score swings are acceptable; what matters is transparent variance reporting and stable protocols. Typical comfort zones for test size cluster in the low-thousands, but teams trade size for feasibility if setup is smooth and provenance is clear.

3.5 Summary: What Needs Standardizing

The survey surfaces a consistent ask: *standardize the view*, not just the data. Practitioners want a uniform presentation of (1) scientific quality signals (validity, reliability, annotation method, basic stats), (2) operational requirements (cost, latency, memory, harness support), (3) coverage (domain, languages, human evaluation, robustness notes), (4) maintenance status, and (5) contamination and provenance disclosures. The presentation should be runnable, comparable, and explainable by design.

4 Benchmark Metadata Analysis

To characterize the information ecosystem practitioners navigate, we analyzed benchmark metadata from multiple sources. We extracted structured metadata including dataset identifiers, task categories, modalities, licenses, size, language coverage, and documentation completeness. We restricted the first release to text-centric benchmarks to keep scope focused; multimodal sources are planned for future iterations.

Analysis focused on metadata consistency and coverage. We examined what proportion of benchmarks document key attributes practitioners need for selection decisions: contamination status, computational requirements, evaluation metrics, domain applicability, and quality indicators. We categorized documentation completeness by identifying common missing fields across datasets. We analyzed task and domain categorization to assess whether current taxonomies support discovery, examining whether benchmarks with similar evaluation purposes use consistent terminology.

This metadata analysis serves two purposes. First, it reveals systematic gaps between the information practitioners need and the information actually available. If survey respondents indicate contamination status is critical, but few benchmarks document it, this mismatch represents a barrier to informed selection. Second, it establishes what information a recommendation system could leverage. Complete, consistent metadata enables better recommendations than sparse, inconsistent metadata.

4.1 Data Acquisition and Curation

We scraped the Hugging Face Hub for benchmarks and dataset, restricting to NLP tasks and tags to maintain focus on text modalities (audio, image, and video will be added later). For each dataset, we parsed the dataset card's YAML metadata (license, language, task categories, size, etc.) and retrieved the same fields in JSON format, along with outbound links (homepage, arXiv/DOI, GitHub, citations) for record linkage. We queried arXiv using a boolean query targeting evaluation and dataset terminology for language models (see Appendix E for the full query). We seeded our list with items from recent benchmark surveys and manually added prominent benchmarks that lack the word *benchmark* in their title. As detailed in Appendix H, benchmark metadata remains highly uneven across sources, with important fields missing, inconsistently named, or reported at different levels of granularity.

5 Implications for Design of BenchNavigator

Our survey findings and metadata analysis reveal specific requirements for a standardized benchmark discovery system. Practitioners need a unified presentation layer that aggregates fragmented metadata, provides transparent filtering and comparison, and explains recommendation rationales without imposing arbitrary rankings.

5.1 RQ3: Designing a Standardized Discovery Interface

Standardization requirements from empirical findings. The survey identified five critical metadata categories that must be presented consistently: (1) scientific quality signals including construct validity, reliability, and annotation quality; (2) operational requirements such as computational cost, runtime, and integration complexity; (3) coverage dimensions including domain relevance, language support, and human evaluation traces; (4) maintenance status and update frequency; and (5) contamination disclosures and provenance information. Practitioners emphasized that these attributes must be comparable across benchmarks regardless of source, even when underlying documentation uses different schemas or terminology.

Design principles for unified presentation. Based on practitioner priorities, we established four design principles for BenchNavigator. First, the system must aggregate heterogeneous metadata into a standardized schema while maintaining provenance. Second, it must support both exploratory search and targeted filtering across practitioner-identified dimensions. Third, recommendation rationales must be transparent and explainable through explicit relationship paths rather than black-box scoring. Fourth, the interface must acknowledge context dependency by avoiding fixed rankings that declare one benchmark universally superior to another.

Addressing information overload through structured guidance. Consider a practitioner beginning benchmark exploration, interested in hallucination measurement, whether to develop new benchmarks or evaluate existing models. A search yields over 30 relevant papers. Which should they select? How should they proceed? Standard Retrieval Augmented Generation (RAG) approaches face a dilemma: should the system return all papers, provide comparative analysis across all results, or limit output to the top 3, 5, or 10 papers?

BenchNavigator addresses this through structured guidance rather than raw retrieval. The system encodes benchmark relationships, understanding what each measures, how they differ, and which suits specific use cases. When queried about hallucination benchmarks, users receive curated explanations of primary approaches, key distinctions, and targeted recommendations, not an unfiltered list of 30+ papers. Technically, we could return all retrieved papers, but practitioners prefer filtered results that they can evaluate systematically rather than sifting through dozens of sources. Critically, we avoid imposing fixed rankings that declare one benchmark superior to another and are based on users' needs.

5.2 Interface Design and Features

BenchNavigator provides a standardized, explainable view rather than a black-box recommender. Users can query in natural language or compose boolean searches; results can be faceted by domain, task, modality, size, license, languages, and risk categories (labeled *AI Atlas risk category*). Our UI exposes these standardized fields through search and faceted filters that reflect practitioner priorities (Figure 3). Lightweight operational readiness signals (paper/GitHub/Hugging Face card availability; metrics/validation presence; community stars) appear alongside scientific descriptors. Each benchmark is summarized in a compact, standardized card surfacing the most decision-relevant attributes at a glance (Figure 5).

Following the filter interface shown in Figure 3, we present the second stage of interaction-the benchmark table view. After applying filters, users are shown a standardized list of matching benchmarks with aligned metadata fields (domain, task, modality, language, and risk indicators). This view supports efficient discovery and helps practitioners quickly identify relevant benchmarks for closer examination.

A multi-item comparison view exposes aligned columns (name, overview, domains, tasks, modality, size, languages, license, stars, risks, metrics, baselines/validation flags, and paper/GitHub/Hugging Face presence). Each result can be expanded to show a provenance trail and recommendation rationale via explicit provenance-linked rationale paths (e.g., domain tag match, task match, and source links). The comparison view

aligns these attributes column-wise to make trade-offs immediately visible (Figure 6).

BenchNavigator supports advanced boolean search, facet filters across practitioner-identified dimensions, column visibility controls, bookmarks, multi-item comparison, and shareable filter state export (see Appendix F for full specifications).

The goal is not to replace reading the paper, but to get practitioners to a defensible short-list quickly. Each benchmark result displays metadata identified as critical through survey responses: year of creation, dataset size, evaluation focus, and documentation completeness. Quality indicators flagged in the literature, contamination concerns, and known limitations appear prominently.

5.3 Validation Through Prototype Implementation

This prototype serves as an existence proof that survey findings translate to implementable design requirements. If practitioners prioritize specific metadata fields, those fields can be extracted and displayed. The system's 14 filters and 8 toggleable columns map directly to survey-identified priorities (see Appendix G for the full mapping between survey results and interface features).

Our prototype demonstrates feasibility without claiming to solve the general problem of universal benchmark assistance. This focused framing allows us to address a concrete technical challenge, standardizing heterogeneous benchmark metadata, while illustrating how such standardization enables the broader goal of accessible, trustworthy benchmark guidance.

6 Limitations

This work characterizes benchmark selection practices through practitioner surveys and metadata analysis, then validates feasibility through a prototype system. Several limitations constrain our findings and their generalizability.

Survey Scope Our survey captures self-reported practices and priorities, which may differ from actual selection behavior in naturalistic settings. Response accuracy depends on practitioners accurately recalling past decisions. Generalizability depends on achieving diverse representation across domains, organizational contexts, and experience levels. Practitioners who abandoned systematic evaluation due to selection difficulties may be underrepresented in our sample.

Prototype Validation The prototype demonstrates that empirically-derived requirements can be implemented, but does not validate effectiveness or improvement over existing approaches. Findings represent a snapshot of a rapidly evolving ecosystem. As new benchmarks emerge and documentation standards change, practitioner priorities may shift. This work examines LLM evaluation specifically and may not generalize to other ML domains with different evaluation practices. We study practitioner perspectives exclusively, not benchmark creators or platform designers who could explain why documentation remains inconsistent.

Data access and quality Consistent with our prior interviews, both the interviews and our survey indicate that data quality is a primary determinant of benchmark choice; many practitioners want to inspect raw items and annotation artifacts before deciding. Our prototype does not embed full datasets or large previews due to size and hosting constraints; instead, we preserve provenance and provide canonical links so users can directly examine the underlying data

7 Discussion

Our work reveals a fundamental tension in the benchmark ecosystem: while practitioners express skepticism about leaderboards and struggle with benchmark selection, foundation model providers invest substantial resources optimizing for benchmark performance. This paradox highlights different stakeholder perspectives that shape how benchmarks are discovered, evaluated, and ultimately used.

Future iterations could incorporate more sophisticated recommendation logic, learning from usage patterns to suggest benchmarks based on natural language task descriptions. However, our current focus on standardization and transparent filtering addresses the immediate need practitioners expressed: making sense of the fragmented landscape before automating selection decisions.

7.1 Implications for Benchmark Producers

Our findings surface an important disconnect between benchmark creation and adoption. Academic benchmark creators often prioritize scientific rigor and novel evaluation capabilities, while practitioners seek operational simplicity and clear applicability. This suggests a potential new role in the ecosystem: the "benchmark hardener" who trans-

forms scientifically interesting benchmarks into production-ready evaluation tools.

Benchmark producers might reasonably push back on some practitioner demands. Comprehensive documentation, contamination analysis, and maintaining multiple versions require significant ongoing effort. However, our approach deliberately avoids imposing new requirements on producers. Instead, we aggregate existing information and make explicit what is missing, allowing users to make informed decisions even with incomplete metadata.

The survey reveals that practitioners value transparency about limitations as much as comprehensive features. Benchmark producers who clearly document scope, known issues, and computational requirements may see better adoption than those claiming broader capabilities without substantiation.

7.2 The Leaderboard Paradox

Perhaps most intriguing is the divergent importance of benchmark leaderboards across stakeholder groups. Our survey participants-primarily technical practitioners-report limited trust in leaderboards and vendor-reported scores. Yet foundation model providers continue investing heavily in benchmark optimization, suggesting these metrics serve other audiences. Future iterations could incorporate additional documentation artifacts beyond BenchmarkCards, including EvalCards (Dhar et al., 2025), and other related structured reporting formats. This would make it possible to unify metadata, evaluation details, and audit-oriented information within the same standardized interface.

We hypothesize that benchmark scores primarily influence non-technical decision makers who lack the expertise or time to conduct thorough evaluations. When executives or procurement teams select between model providers, benchmark scores offer seemingly objective comparison points alongside pricing and terms. This dynamic parallels consumer electronics marketing, where technical specifications influence purchasing decisions despite limited understanding of their practical implications.

Respondents evaluate models across heterogeneous goals and domains (e.g., fairness, drug discovery), which makes a universal ranking inappropriate. Accordingly, BenchNavigator avoids prescribing a single "best" benchmark and instead standardizes metadata and exposes filters to support context-specific selection by the user.

This disconnect suggests that standardization efforts must serve multiple audiences with different evaluation needs. Technical users require detailed metadata for rigorous assessment, while decision makers need accessible summaries that acknowledge capabilities and limitations. BenchNavigator's tiered information presentation, from high-level domain tags to detailed quality indicators, attempts to bridge these different use cases.

8 Broader Implications for LLM Evaluation

The fragmentation we document reflects deeper challenges in LLM evaluation. The rapid proliferation of benchmarks signals both the complexity of language understanding and the difficulty of capturing real-world performance through standardized tests. No single benchmark or even suite of benchmarks adequately represents deployment readiness, yet the alternative-custom evaluation for every use case-remains prohibitively expensive for many organizations.

Our standardization framework offers a pragmatic middle path: helping practitioners navigate existing options while acknowledging their limitations. By making benchmark metadata comparable and discoverable, we enable more informed selection without claiming to solve the fundamental challenge of ecological validity in LLM evaluation.The cognitive burden imposed by fragmentation extends beyond individual practitioners.

9 Conclusion

We studied a concrete but underaddressed problem: fragmented benchmark representations make discovery unnecessarily hard. Through practitioner surveys and metadata analysis, we identified what must be standardized for effective selection and documented systematic inconsistencies across sources. Practitioners need not just better benchmarks or documentation, but a standardized way to find, compare, and justify benchmark choices across heterogeneous sources. BenchNavigator demonstrates that this standardization is feasible: it normalizes heterogeneous metadata into a consistent schema, preserves provenance, and exposes a presentation layer aligned with practitioner priorities, without imposing new burdens on benchmark producers.

References

NIST AI. 2023. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pages 100–1.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, and 1 others. 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)*, pages 13787–13805.

Anonymous. 2024. Details omitted for double-blind review.

Frank Bagehorn, Kristina Brimijoin, Elizabeth M Daly, Jessica He, Michael Hind, Luis Garces-Erice, Christopher Giblin, Ioana Giurgiu, Jacquelyn Martino, Rahul Nair, and 1 others. 2025. Ai risk atlas: Taxonomy and tooling for navigating ai risks and resources. *arXiv preprint arXiv:2503.05780*.

Vanessa Bracamonte, Sebastian Pape, Sascha Löbner, and Frederic Tronnier. 2023. Effectiveness and information quality perception of an ai model card: a study among non-experts. In *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, pages 1–7. IEEE.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. 2023. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Ruchira Dhar, Danae Sanchez Villegas, Antonia Karamolegkou, Alice Schiavone, Yifei Yuan, Xinyi Chen, Jiaang Li, Stella Frank, Laura De Grazia, Monorama Swain, and 1 others. 2025. Evalcards: A framework for standardized evaluation reporting. *arXiv preprint arXiv:2511.21695*.

Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. 2025. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. *Preprint*, arXiv:2502.06559.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*.

Amelia Hardy, Anka Reuel, Kiana Jafari Meimandi, Lisa Soder, Allie Griffith, Dylan M Asmar, Sanmi Koyejo, Michael S Bernstein, and Mykel John Kochenderfer. 2025. More than marketing? on the information value of ai benchmarks for practitioners. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1032–1047.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, and 1 others. 2021. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*.

Eunsu Kim, Haneul Yoo, Guijin Son, Hitesh Patel, Amit Agarwal, and Alice Oh. 2025. Benchhub: A unified benchmark suite for holistic and customizable llm evaluation. *arXiv preprint arXiv:2506.00482*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. Systematic analysis of 32,111 ai model cards characterizes documentation practice in ai. *Nature Machine Intelligence*, 6(7):744–753.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. 2024. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813.

Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh V. Chawla. 2024. Benchmarkcards: Standardized documentation for large language model benchmarks. *ArXiv*.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, and 1 others. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

A Supplementary Figures

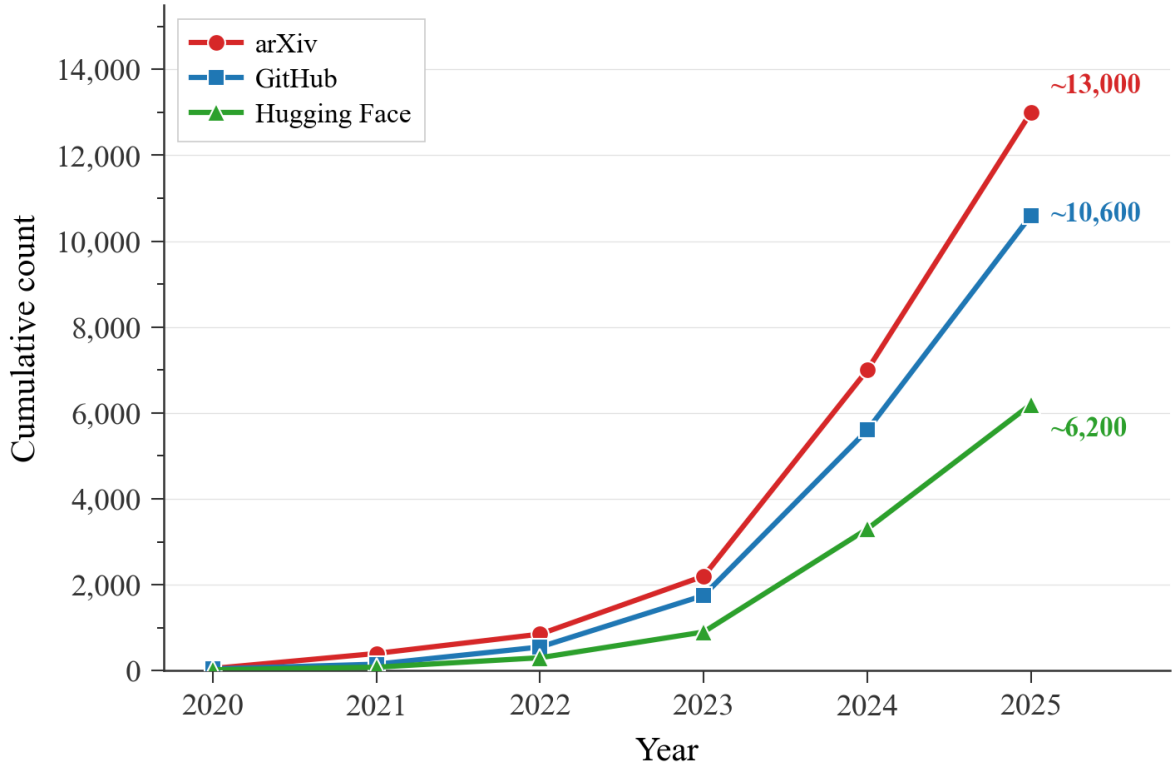

Figure 1: Cumulative growth of language model benchmarks across arXiv, GitHub, and Hugging Face in our snapshot. Growth motivates a standardized view for discovery and comparison (Data collecting until 2025. See Appendix E to see how data were collected)

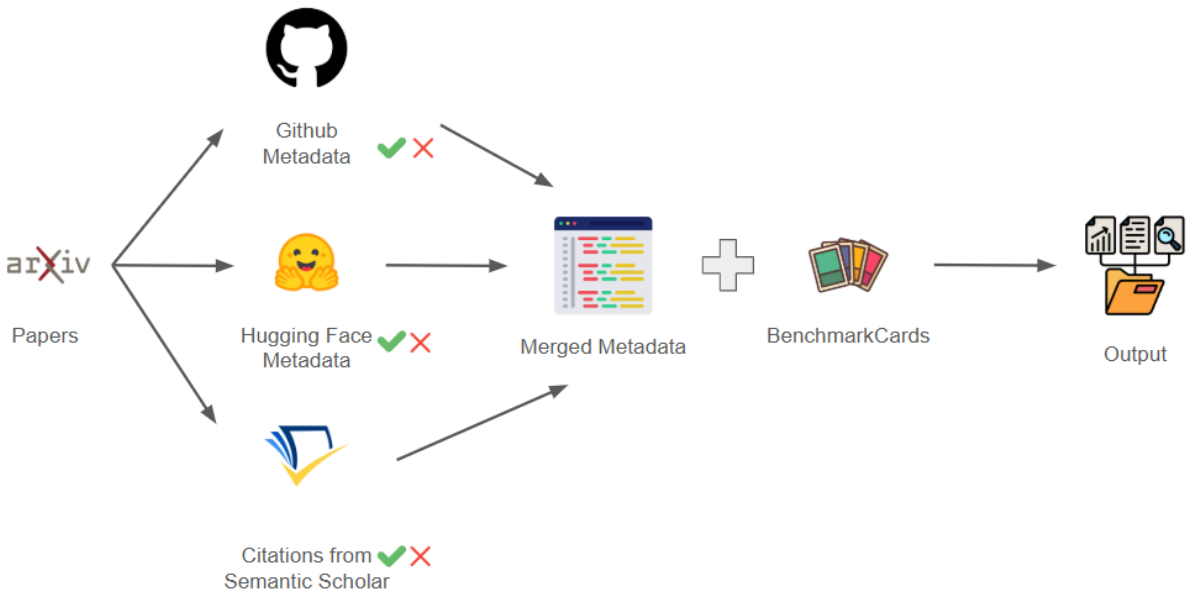

Figure 2: BenchNavigator data pipeline integrating benchmark metadata from Hugging Face, arXiv, GitHub, and BenchmarkCards.

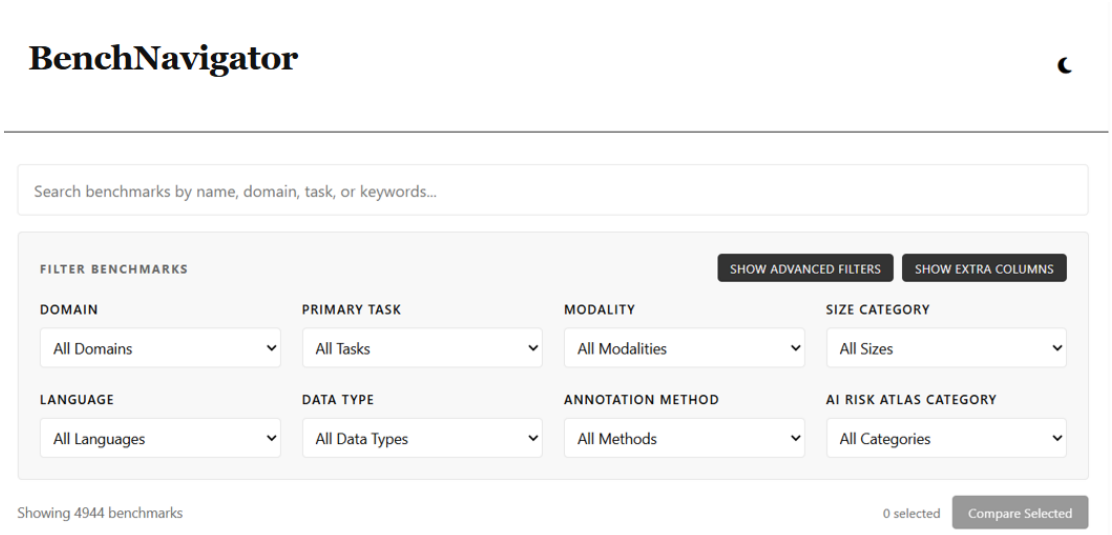

Figure 3: The user interface for our BenchNavigator prototype. It operationalizes survey findings by providing search and facet filters for key practitioner priorities, such as Domain, Primary Task, and other metadata attributes, to help users create a defensible short-list of relevant benchmarks.

COMPARE	NAME	DOMAIN	TASK	MODALITY	LANGUAGES	AI RISK ATLAS	PAPER	GITHUB	HF
☐	**TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension**	Natural Language Processing, Information Retrieval	question-answering	text	en	Fairness, Accuracy	✓	✓	✓
☐	**FigureQA**	Natural Language Processing, Computer Vision	question-answering	image	en, zh	Fairness, Accuracy	✓	✓	✓
☐	**OpenBookQA**	Natural Language Processing, Science Education	question-answering	text	en	Fairness, Accuracy	✓	✓	✓
☐	**Winogender schemas**	Natural Language Processing	question-answering	text	en	Fairness, Accuracy	✓	✓	✓
☐	**COMMONSENSE QA: A Question Answering Challenge Targeting Commonsense Knowledge**	Natural Language Processing	question-answering	text	en	Fairness, Accuracy	✓	✓	✓
☐	**CODAH (COmmonsense Dataset Adversarially-authored by Humans)**	Natural Language Processing	question-answering	tabular	en	Fairness, Robustness	✓	✓	✓
☐	**QANTA dataset (Question Answering is Not a Trivial Activity)**	Natural Language Processing, Education	question-answering	tabular	en	Fairness, Robustness	✓	✓	✓

Figure 4: Second part of the BenchNavigator interface: the benchmark table view displayed after filters are applied. The table show metadata fields such as domain, task, modality, language, and AI Risk Atlas (Bagehorn et al., 2025) categories to support transparent comparison and selection.

COMPARE	NAME	DOMAIN	TASK	MODALITY	LANGUAGES	AI RISK ATLAS	PAPER	GITHUB	HF
☐	**TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension**	Natural Language Processing, Information Retrieval	question-answering	text	en	Fairness, Accuracy	✓	✓	✓

OVERVIEW

We present TriviaQA, a challenging reading comprehension dataset containing over 650K question-answer-evidence triples. TriviaQA includes 95K question-answer pairs authored by trivia enthusiasts and independently gathered evidence documents (on average six per question) collected from Wikipedia and Web search results, designed to test complex compositional questions, substantial syntactic and lexical variability, and multi-sentence reasoning.

GOAL

To introduce TriviaQA, a new reading comprehension dataset designed to simultaneously test complex compositional questions, syntactic and lexical variability between questions and evidence, and multi-sentence reasoning, and to provide resources for training and evaluating reading-comprehension models.

DATA SOURCE

Question-answer pairs collected from 14 trivia and quiz-league websites. Evidence documents collected from two sources: (1) Web search results via Bing (top 50 results; crawled top 10 pages after filtering) and (2) Wikipedia pages identified via TAGME entity linking applied to questions. Additionally a human-annotated verified subset was created.

ANNOTATION

Evidence documents automatically gathered (distant supervision). A clean, human-annotated subset of 1,975 question-document-answer triples whose documents are certified to contain all facts required to answer the questions.

AI RISK ATLAS CATEGORIES

- Fairness
  - Data bias
- Accuracy
  - Unrepresentative data

INTENDED AUDIENCE

- Model Developers
- Machine Learning Researchers

Paper | GitHub | HuggingFace | Homepage

Figure 5: An example of a benchmark metadata card for TriviaQA within the BenchNavigator system. The card aggregates key information such as domain, task, data source, and potential AI risks in a standardized format.

Benchmark Comparison ×

ATTRIBUTE	MEDBOOKVQA	MEDDIALOG: TWO LARGE-SCALE MEDICAL DIALOGUE DATASETS
DOMAIN	Healthcare	Healthcare, Natural Language Processing
PRIMARY TASK	question-answering	question-answering
MODALITY	image	text
SIZE CATEGORY	1K	1M
LANGUAGES	en	en, zh
DATA TYPE	question-answering pairs	text (patient-doctor conversations / dialogues)
ANNOTATION METHOD	Automated extraction of medical figures paired with narrative context.	N/A
AI RISK ATLAS	• Fairness → Data bias • Accuracy → Unrepresentative data • Societal Impact → Impact on education: bypassing learning	• Privacy → Personal information in data • Fairness → Data bias
OVERVIEW	MedBookVQA is a systematic and comprehensive multimodal benchmark derived from open-access medical textbooks, consisting of 5,000 clinically relevant questions across various medical VQA task categories.	To facilitate the research and development of medical dialogue systems, we build two large-scale medical dialogue datasets: MedDialog-EN and MedDialog-CN. MedDialog-EN is an English dataset containing 0.3 million conversations between patients and doctors and 0.5 million utterances. MedDialog-CN is a Chinese dataset containing 1.1 million conversations and 4 million utterances. To our best knowledge, MedDialog-(EN,CN) are the largest medical dialogue datasets to date. The dataset is available at https://github.com/UCSD-AI4H/Medical-Dialogue-System
GOAL	To provide a comprehensive benchmark for evaluating General Medical Artificial Intelligence (GMAI) systems across diverse medical domains.	To facilitate the research and development of medical dialogue systems by providing two large-scale medical dialogue datasets (MedDialog-EN and MedDialog-CN).
DATA SOURCE	Open-access medical textbooks from DOAB (Directory of Open Access Books).	MedDialog-EN data crawled from iclinic.com and healthcaremagic.com. MedDialog-CN data crawled from haodf.com.

Figure 6: A comparison view in BenchNavigator, allowing users to evaluate multiple benchmarks side-by-side across key metadata dimensions identified through practitioner surveys.

B Survey

Section 1 of 10
BenchNavigator Survey

We are developing BenchNavigator, a recommendation tool that will help practitioners select the most appropriate benchmarks for evaluating Large Language Models (LLMs) based on their specific use cases and constraints.

What we're trying to understand:

- How practitioners currently find, select, and use LLM benchmarks in real-world settings
- What practical challenges and constraints affect benchmark selection
- Which criteria and features matter most when choosing benchmarks
- What functionality would be most valuable in an automated benchmark recommendation tool

Who should take this: People who evaluate or select LLMs (large language models) or run LLM benchmarks.

Key definitions:

- **Large language models (LLMs)** are AI systems capable of understanding and generating human language by processing vast amounts of text data.
- **Benchmark**: a combination of a dataset, evaluation metrics, and associated pre- and post-processing steps used to assess specific aspects of LLM behavior.

Time: ∼20 minutes **Privacy:** Responses are anonymized and reported in aggregate.
Contact: ANONYMOUS **IRB:** Approved, Protocol #Hidden

Privacy Notice:

- This survey does **NOT** automatically collect your email address.
- If you want to receive a gift card, you may provide an email at the end.

Section 2 of 10
Consent

By proceeding with this survey, you acknowledge that you are 18 years or older, have read the information provided, and voluntarily consent to participate in this research study.

Do you consent to participate, acknowledging that you are 18 or older, have read this information, and agree to proceed? *

- ☐ I agree and consent to participate
- ☐ I do not consent (this will end the survey)

Section 3 of 10
Background

This section collects demographic and professional context to help us interpret aggregate results. All information is kept confidential.

What is your country of residence? *

What is your gender? *

- ☐ Male
- ☐ Female
- ☐ Nonbinary
- ☐ Prefer not to say
- ☐ Other: _____

What type of organization do you work in? *

- ☐ Academia
- ☐ Industry
- ☐ Government
- ☐ Non-Governmental Organization/Non-Profit Organizations
- ☐ Military/Defense
- ☐ Other: _____

What is the name of your organization or company?
You may skip this question. If provided, your organization name will be stored separately from responses and not shown in reports. Used only for deduplication or follow-up (if you opt in).

Years of experience in ML/AI (number): *

What is your primary role? *
(Select the option that best fits. If none match, choose "Other" and specify.)

- ☐ Software Engineer / Developer
- ☐ Research Scientist / Faculty
- ☐ Domain Expert / Practitioner (e.g., healthcare, law, education)
- ☐ Machine Learning (ML) or Large Language Model (LLM) Engineer
- ☐ Data Scientist
- ☐ Product Manager (PM) or Technical Program Manager (TPM)
- ☐ Policy / Governance / Ethics Specialist
- ☐ Student (Undergrad/Grad)
- ☐ Other: _____

Have you ever created or customized a benchmark? *

- ☐ Yes
- ☐ No

Please list 1–2 specific benchmarks you have used recently (e.g., MMLU, GSM8K, SWE-bench) *

Which types of benchmarks do you typically use when evaluating LLMs? *
(Select all that apply. If none match, choose "Other" and specify.)

- ☐ General language understanding benchmarks (e.g., MMLU, HellaSwag)

- ☐ Factual accuracy and hallucination benchmarks (e.g., TruthfulQA)
- ☐ Domain knowledge benchmarks (e.g., MedQA, LegalBench)
- ☐ Reasoning benchmarks (e.g., GSM8K, MATH, ARC)
- ☐ Code generation benchmarks (e.g., HumanEval, MBPP)
- ☐ Multilingual benchmarks (e.g., FLORES, XQuAD)
- ☐ Safety/bias/toxicity benchmarks (e.g., ToxicGen, BBQ)
- ☐ Agent/tool-use benchmarks (e.g., WebShop, ToolBench)
- ☐ Question answering (open-domain, closed-book, reading comprehension)
- ☐ Instruction following benchmarks (e.g., MT-Bench)
- ☐ Table reasoning benchmarks (e.g., WikiTables QA)
- ☐ Custom/internal benchmarks
- ☐ Other: _____

Section 4 of 10
Benchmark usage & decision context

This section explores how you currently use benchmarks and what informs your selection decisions.

How often do you evaluate or compare LLMs using benchmarks? *

- ☐ In all my projects
- ☐ In more than half of my projects
- ☐ In about half of my projects
- ☐ In less than half of my projects
- ☐ In none of my projects

At what stage in your project lifecycle do you typically select benchmarks? *

- ☐ Early scoping/model exploration
- ☐ Mid-project comparison/model selection
- ☐ Pre-deployment validation/sign-off
- ☐ Post-deployment monitoring/regression checks
- ☐ Other: _____

Do you feel overwhelmed by the number of available benchmarks? *

- ☐ Yes, there are too many to evaluate
- ☐ Somewhat overwhelmed
- ☐ Neutral – neither too many nor too few
- ☐ Not at all overwhelmed

- ☐ No, the variety is helpful

Have you experienced technical failures when running benchmarks? *

- ☐ Always
- ☐ Frequently
- ☐ Sometimes
- ☐ Rarely
- ☐ Never

Do you primarily evaluate open-weight, closed-weight, or both? *

- ☐ Mostly open-weight
- ☐ Mostly closed-weight
- ☐ Both
- ☐ Not sure
- ☐ N/A

For your most recent benchmark selection, which information sources did you rely on? *
(Select all that apply)

- ☐ Academic papers (arXiv, conference proceedings)
- ☐ Hugging Face (datasets, model pages, leaderboards)
- ☐ GitHub repositories and documentation
- ☐ Community leaderboards (HELM, Open LLM Leaderboard, OpenCompass)
- ☐ Recommendations from colleagues
- ☐ Vendor documentation / blogs
- ☐ Social media and forums (Reddit, Twitter/X)
- ☐ Other: _____

How well do academic benchmark scores predict real-world performance in your use cases? *

- ☐ Completely accurate
- ☐ Very accurate
- ☐ Somewhat accurate
- ☐ Slightly accurate
- ☐ Not at all accurate

How important is it that benchmarks update regularly to prevent contamination? *

- ☐ Annually
- ☐ Quarterly

- ☐ Monthly
- ☐ Static is fine for my use cases

How concerned are you about benchmark contamination (test data appearing in training sets)? *

- ☐ Extremely concerned
- ☐ Concerned
- ☐ Moderately concerned
- ☐ Slightly concerned
- ☐ Not concerned
- ☐ N/A

Have you observed inflated scores that you suspect are due to contamination? *

- ☐ Very frequently
- ☐ Frequently
- ☐ Occasionally
- ☐ Rarely
- ☐ Never

How often do you encounter vendors selectively reporting only favorable benchmarks? *

- ☐ Always
- ☐ Often
- ☐ Sometimes
- ☐ Rarely
- ☐ Never
- ☐ N/A

When selecting models, how do these factors compare to accuracy?

	More important	**Equal**	**Less important**
Cost per token	☐	☐	☐
Inference latency	☐	☐	☐
Memory requirements	☐	☐	☐
Energy consumption	☐	☐	☐

Section 5 of 10
Typical constraints when selecting a benchmark

This section explores the practical constraints and limitations that influence your benchmark selection decisions.

Please rate how important each constraint is when you select benchmarks for your typical projects. Rate each item independently. *

	Not important	Slightly	Moderately	Important	Very imp.
Ease of integration (scripts, harnesses, API)	☐	☐	☐	☐	☐
Data privacy & compliance (GDPR, HIPAA)	☐	☐	☐	☐	☐
Monetary/compute budget	☐	☐	☐	☐	☐
Org. policies & approvals	☐	☐	☐	☐	☐
Time-to-run (setup to results)	☐	☐	☐	☐	☐
Licensing & terms of use	☐	☐	☐	☐	☐
Technical compatibility	☐	☐	☐	☐	☐
Benchmark validation & quality	☐	☐	☐	☐	☐

After section 5

Section 6 of 10
Benchmark Selection Criteria

This section explores which specific qualities and features influence your benchmark selection decisions. Please consider your general experience, not just your last benchmark selection.

Which scientific quality factors influence your benchmark selection? *
(Select all that apply)

- ☐ Construct validity: Benchmark actually measures what it claims to measure
- ☐ Reliability: Consistent results across multiple runs
- ☐ Annotation quality: Clear guidelines and high inter-annotator agreement
- ☐ Data contamination checks: Evidence that test data wasn't in training sets
- ☐ Difficulty calibration: Includes appropriately challenging test cases
- ☐ Statistical rigor: Proper confidence intervals, statistical power, sample sizes
- ☐ Other: _____

Which practical factors influence your benchmark selection? *
(Select all that apply)

- ☐ Ready-to-use implementations: Available in HELM/eval-harness/OpenCompass
- ☐ Reproducibility features: Fixed seeds, versioned prompts, documentation
- ☐ Ease of setup: Good documentation, notebooks, tutorials
- ☐ Performance estimates: Clear runtime and cost information
- ☐ Container/API support: Docker, cloud deployment options
- ☐ Other: _____

Which external signals influence your benchmark selection? *
(Select all that apply)

- ☐ Academic citations: How often the benchmark paper is cited
- ☐ Community adoption: Usage in papers and projects, stars in GitHub or Hugging Face
- ☐ Active maintenance: Recent updates, issue resolution
- ☐ Standardized metrics: Comparable across implementations
- ☐ Documentation quality: Complete dataset/model cards
- ☐ Third-party validation: Independent audits or reviews
- ☐ Other: _____

Which coverage factors influence your benchmark selection? *
(Select all that apply)

- ☐ Human evaluation data: Includes human judgments
- ☐ Domain relevance: Matches your specific use case
- ☐ Language coverage: Supports needed languages
- ☐ Safety evaluation: Includes bias/toxicity assessment
- ☐ Robustness testing: Sensitivity to prompt variations
- ☐ Other: _____

Tell us more – What is the most important criterion for you (open text) *

Section 7 of 10
Coverage Gaps & Needs

This section captures gaps you see in current benchmarks so we can align recommendations to real needs.

How much do you trust each evaluation method to provide reliable insights? *

	Don't trust	Trust a little	Neutral	Mostly trust	N/A
Human evaluation w/ IAA	☐	☐	☐	☐	☐
Automated metrics (BLEU, ROUGE)	☐	☐	☐	☐	☐
LLM-as-judge evaluation	☐	☐	☐	☐	☐
Head-to-head human (Arena-style)	☐	☐	☐	☐	☐
Adversarial/stress testing	☐	☐	☐	☐	☐

How much score variance do you consider acceptable? (e.g., "5%", "5 percent", "between 3–7%")

Which important capabilities or risks are under-tested by current benchmarks in your domain?

Which domains or data types lack good benchmarks for your work? *
(e.g., healthcare/medical, legal, education. . .)

Acceptable minimum sample size for reporting model differences on your tasks *

- ☐ <500 items
- ☐ 500–2k
- ☐ 2k–10k
- ☐ >10k
- ☐ Doesn't matter
- ☐ Other: _____

What are the main limitations of current LLM benchmarks for your work?

Section 8 of 10
BenchNavigator Features & Priorities

Imagine you had a tool, **BenchNavigator**, designed to help you quickly find and select the most relevant benchmarks for evaluating LLMs, based on your goals, constraints, and priorities. We'd like to know which features would make such a tool genuinely useful for you.

If you could add any feature or capability to BenchNavigator to make it truly valuable for you and your team, what would it be? *

Which features would you find most valuable?

	Not at all	Slightly	Moderately	Important	Very
Ranked recommendations	☐	☐	☐	☐	☐
Evidence score (docs, seeds, human eval, replication)	☐	☐	☐	☐	☐
Maintenance score (updates, issues, commits)	☐	☐	☐	☐	☐
Contamination risk & provenance indicators	☐	☐	☐	☐	☐
Compute cost/time estimator (HW/budget)	☐	☐	☐	☐	☐
Explainers: why a benchmark was recommended	☐	☐	☐	☐	☐
One-click export in JSON/Markdown	☐	☐	☐	☐	☐
Visualizations of results (graphs, interactive)	☐	☐	☐	☐	☐
Support for custom metrics	☐	☐	☐	☐	☐
Domain-specific tailoring (e.g., healthcare, law)	☐	☐	☐	☐	☐

What integrations should a recommendation tool for benchmarks support first? (select all that apply)

- ☐ Hugging Face (datasets, leaderboards, spaces)
- ☐ GitHub (repos/issues/updates)
- ☐ HELM / OpenCompass / EleutherAI eval-harness
- ☐ arXiv (paper links/citations)
- ☐ Other: _____

Section 9 of 10
Contact

This information will only be used for follow-up about this research.

If you'd like to receive the Amazon gift card, please provide an email address.

Anything else we should know?

Section 10 of 10
Thanks so much for participating!

Your feedback is invaluable and will help us make BenchNavigator more useful for the community.

C Survey Results

C.1 Participant Demographics

We surveyed 23 practitioners from 5 countries. Table 1 summarizes participant characteristics.

Table 1: Survey Participant Demographics (N=23)

Characteristic	Category	Count
Country	USA	18
	China	1
	Germany	1
	Mexico	1
	Bangladesh	1
Gender	Male	14
	Female	7
	Prefer not to say	2
Organization	Academia	19
	Industry	3
	Non-profit	1
Role	Student	13
	Research Scientist/Faculty	7
	Software Engineer	1
	Domain Expert	1
	Product Manager	1
Experience	Mean: 4.6 years	Range: 2-15
Created Benchmarks	Yes	8
	No	15

C.2 Current Practices

C.2.1 Benchmark Usage Frequency

- In all projects: 3 participants (13%)
- In more than half: 3 participants (13%)
- In about half: 3 participants (13%)
- In less than half: 12 participants (52%)
- In none: 2 participants (9%)

C.2.2 Project Lifecycle Stages

- Early scoping/model exploration: 14 participants (61%)
- Mid-project comparison/selection: 17 participants (74%)
- Pre-deployment validation: 8 participants (35%)
- Post-deployment monitoring: 7 participants (30%)

C.2.3 Information Sources Used

- Academic papers (arXiv, conferences): 21 participants (91%)
- Hugging Face: 18 participants (78%)
- GitHub repositories: 16 participants (70%)
- Recommendations from colleagues: 10 participants (43%)

- Community leaderboards: 3 participants (13%)
- Social media/forums: 5 participants (22%)
- Vendor documentation: 1 participant (4%)

C.2.4 Model Types Evaluated

- Mostly open-weight: 9 participants (39%)
- Mostly closed-weight: 3 participants (13%)
- Both: 8 participants (35%)
- Not sure/N/A: 3 participants (13%)

C.3 Challenges

C.3.1 Feeling Overwhelmed by Benchmark Quantity

- Yes, too many to evaluate: 11 participants (48%)
- Somewhat overwhelmed: 3 participants (13%)
- Neutral: 6 participants (26%)
- No, variety is helpful: 3 participants (13%)

C.3.2 Technical Failure Frequency

- Always: 3 participants (13%)
- Frequently: 5 participants (22%)
- Sometimes: 9 participants (39%)
- Rarely: 5 participants (22%)
- Never: 1 participant (4%)

C.3.3 Real-World Performance Prediction

How well do academic benchmark scores predict real-world performance:

- Very accurate: 2 participants (9%)
- Somewhat accurate: 13 participants (57%)
- Slightly accurate: 6 participants (26%)
- Not accurate: 0 participants (0%)
- N/A: 2 participants (9%)

C.3.4 Contamination Concerns

- Extremely concerned: 4 participants (17%)
- Concerned: 6 participants (26%)
- Moderately concerned: 7 participants (30%)
- Slightly concerned: 1 participant (4%)
- Not concerned: 2 participants (9%)
- N/A: 3 participants (13%)

C.3.5 Observed Contamination Impact

Have you observed inflated scores due to suspected contamination:

- Very frequently: 2 participants (9%)
- Frequently: 5 participants (22%)
- Occasionally: 6 participants (26%)
- Rarely: 4 participants (17%)
- Never: 1 participant (4%)
- N/A: 5 participants (22%)

C.3.6 Vendor Selective Reporting

How often vendors report only favorable benchmarks:

- Always: 5 participants (22%)
- Often: 7 participants (30%)
- Sometimes: 5 participants (22%)
- Rarely: 1 participant (4%)
- N/A: 5 participants (22%)

C.4 Selection Priorities

C.4.1 Scientific Quality Factors

Factors influencing benchmark selection (participants could select multiple):

- Construct validity: 17 participants (74%)
- Reliability: 16 participants (70%)
- Annotation quality: 13 participants (57%)
- Data contamination checks: 10 participants (43%)
- Difficulty calibration: 8 participants (35%)
- Statistical rigor: 7 participants (30%)

C.4.2 Practical Factors

- Ready-to-use implementations: 13 participants (57%)
- Ease of setup: 14 participants (61%)
- Reproducibility features: 13 participants (57%)
- Performance estimates: 9 participants (39%)
- Container/API support: 6 participants (26%)

C.4.3 External Signals

- Academic citations: 17 participants (74%)
- Community adoption: 16 participants (70%)
- Active maintenance: 12 participants (52%)
- Standardized metrics: 13 participants (57%)
- Documentation quality: 13 participants (57%)
- Third-party validation: 5 participants (22%)

C.4.4 Coverage Factors

- Domain relevance: 20 participants (87%)
- Human evaluation data: 17 participants (74%)
- Language coverage: 11 participants (48%)
- Safety evaluation: 6 participants (26%)
- Robustness testing: 8 participants (35%)

C.5 Constraint Importance Ratings

Table 2 shows mean importance ratings for different constraints (1=Not important at all, 5=Very important).

Table 2: Constraint Importance Ratings (Mean scores, N=23)

Constraint	Mean Rating
Benchmark validation and quality	**4.2**
Ease of integration	**3.8**
Technical compatibility	**3.9**
Time-to-run	3.6
Data privacy and compliance	3.4
Monetary/compute budget	3.5
Licensing and terms of use	3.2
Organizational policies	3.0

C.6 Trust in Evaluation Methods

Table 3 shows trust levels in different evaluation approaches (1=Don't trust at all, 5=Trust completely).

Table 3: Trust in Evaluation Methods (Mean scores, N=23)

Evaluation Method	Mean Trust
Human evaluation with inter-annotator agreement	**3.8**
Head-to-head human comparisons (Arena-style)	3.5
Adversarial/stress testing	3.3
Automated metrics (BLEU, ROUGE, etc.)	3.1
LLM-as-judge evaluation	2.9

C.7 Update Frequency Preferences

How often benchmarks should update to prevent contamination:

- Monthly: 3 participants (13%)
- Quarterly: 8 participants (35%)
- Annually: 4 participants (17%)
- Static is fine: 6 participants (26%)
- N/A: 2 participants (9%)

C.8 Acceptable Score Variance

Reported acceptable score variance (open-ended responses):

- $\leq$2%: 1 participant
- 3-5%: 9 participants
- 5%: 5 participants
- >5%: 2 participants
- Not specified: 6 participants

C.9 Sample Size Preferences

Acceptable minimum sample size for reporting:

- <500 items: 4 participants (17%)
- 500-2k items: 7 participants (30%)
- 2k-10k items: 8 participants (35%)
- >10k items: 3 participants (13%)
- Doesn't matter: 1 participant (4%)

C.10 Desired Features for BenchNavigator

Table 4 shows importance ratings for potential BenchNavigator features.

Table 4: Feature Importance for BenchNavigator (Mean ratings, N=23)

Feature	Mean Importance
Evidence score (docs, seeds, human eval)	**4.1**
Contamination risk & provenance indicators	**3.7**
Explainers: why benchmark recommended	**3.9**
One-click export in JSON/Markdown	**3.8**
Maintenance score (updates, activity)	3.6
Domain-specific tailoring	**3.8**
Ranked recommendations	3.4
Compute cost/time estimator	3.3
Visualizations of results	3.2
Support for custom metrics	3.1

C.11 Integration Preferences

Desired integrations for benchmark recommendation tool:

- Hugging Face: 20 participants (87%)
- GitHub: 18 participants (78%)
- arXiv: 16 participants (70%)
- HELM/OpenCompass/EleutherAI eval-harness: 5 participants (22%)

C.12 Commonly Used Benchmarks

Most frequently mentioned benchmarks in open responses:

- MMLU: 6 mentions
- GSM8K: 4 mentions
- SWE-bench: 3 mentions
- HellaSwag: 2 mentions
- MATH: 2 mentions
- HumanEval: 2 mentions

C.13 Identified Gaps and Limitations

C.13.1 Under-tested Capabilities

Open-ended responses identifying gaps (selected quotes):

- "Low-resource language capabilities"
- "Contextual grounding in 3D or spatial environments"
- "Fidelity of reasoning"
- "Ethics" and "Safety"
- "Generality of LLM"

C.13.2 Domains Lacking Good Benchmarks

- Healthcare/medical: 5 mentions
- Education/learning: 4 mentions
- Legal: 2 mentions
- Chemistry: 2 mentions
- Security: 2 mentions

C.13.3 Main Limitations of Current Benchmarks

Selected open-ended responses:

- "Don't fit on current GPUs - not enough memory"
- "Domain matching" issues
- "Data quality" concerns
- "Not easy to integrate to vLLM"
- "Results are hard to reproduce"
- "Hard to use" / "Not good documentation"
- "Current benchmarks are primarily text-centric and fail to capture embodied, multimodal contexts"

D Survey Design Rationale

The survey instrument progresses through five sections. Demographics and background questions establish the respondent's experience with LLM evaluation, organizational context, and typical use cases. Current practice questions examine how practitioners find benchmarks, what information sources they consult, and at what project stages selection occurs. Constraint questions probe factors influencing decisions: computational budgets, organizational policies, contamination concerns, and technical compatibility requirements. Priority questions ask respondents to rank benchmark attributes by importance for selection decisions. Trust questions investigate confidence in different evaluation approaches, from human evaluation to automated metrics to LLM-as-judge methods. Questions target specific hypotheses about selection barriers. We ask whether practitioners feel overwhelmed by the number of available benchmarks to test if the discovery problem exists at scale. We probe the frequency of encountering contamination issues to assess whether quality concerns affect selection in practice. We examine which metadata fields practitioners consider when comparing benchmarks to identify what information must be readily accessible. Open-response questions capture selection strategies and pain points not anticipated by structured items.

E ArXiv Query Details

To expand coverage beyond Hugging Face Hub listings, we queried arXiv using the following boolean query: `(benchmark OR benchmarks OR benchmarking OR evaluation OR evaluating OR dataset OR datasets OR task OR tasks OR test OR testing) AND ("language model" OR "language models" OR LLM OR LLMs OR NLP OR "natural language processing" OR "text model" OR "text models")` Because prominent benchmarks sometimes lack the word *benchmark* in their title or abstract (e.g., *Measuring Massive Multitask Language Understanding (MMLU)* or *PubMedQA: A Dataset for Biomedical Research Question Answering*), we seeded our list with items from recent benchmark surveys and manually added missing cases.

F Feature Specifications

BenchNavigator operationalizes the standardized view through: (i) advanced boolean search (`AND/OR/"exact"/-exclude`), (ii) facet filters (domain, primary task, modality, size bucket, license, and Atlas-style risk categories), (iii) column visibility controls, (iv) bookmarks and a sidecar bookmarks bar, (v) multi-item comparison with a structured table, and (vi) shareable filter state export. Users can filter benchmarks by license, data splits, region, supported libraries, instruction tuning, GitHub metrics, ArXiv availability, size, and evaluation methods, while optional columns show citations, licenses, repository statistics, organizations, metrics, baselines, and limitations. Each benchmark result displays metadata identified as critical through survey responses: year of creation, dataset size, evaluation focus, and documentation completeness. Quality indicators flagged in the literature, contamination concerns, and known

limitations appear prominently. Lightweight operational readiness signals (paper/GitHub/Hugging Face card availability; metrics/validation presence; community stars) appear alongside scientific descriptors.

G Feature-to-Survey Mapping

The system offers 14 advanced filters (license, splits, region, libraries, instruction tuning, GitHub engagement, ArXiv, size, evaluation methods) based on survey-identified priorities: 91% of practitioners consult academic papers, 78% use Hugging Face, 70% use GitHub repositories, and 70% value community adoption metrics as selection signals. Eight toggleable columns (citation, license, stars, forks, ArXiv ID, organization, metrics, baselines, limitations) address the 74% of respondents who prioritize human evaluation data and the 57% who require reproducibility features and documentation quality.

H Benchmark Code and Data Availability

To characterize how benchmark authors distribute their artifacts, we analyzed 2,150 benchmark papers submitted to arXiv in 2025. For each paper, we downloaded the PDF, extracted all URLs, classified them by hosting platform, filtered out references to generic infrastructure repositories (e.g., vLLM, LLaMA, OpenCompass), and retained only links pointing to the benchmark's own code or data. We also detected papers that *promise* a future release without providing a link. Table 5 summarizes availability status; Table 6 breaks down hosting platforms.

Table 5: Benchmark artifact availability status (N = 2,150 arXiv papers, 2025).

Status	**Count**	**%**
Link(s) found (available)	1,658	77.1%
Promised, not yet released	57	2.7%
No links at all	435	20.2%
Total	**2,150**	**100.0%**

Table 6: Hosting platform distribution (N = 2,150). Papers may appear in multiple rows.

Platform	**Papers**	**%**
GitHub	1,282	59.6%
HuggingFace	621	28.9%
Project site (`.io`)	319	14.8%
Kaggle	63	2.9%
Anonymous review (4open)	50	2.3%
Zenodo	26	1.2%
Google Drive	16	0.7%
GitLab	7	0.3%

Roughly one in five benchmark papers (20.2%) provides no public link to code or data, confirming that artifact discoverability remains a significant gap. GitHub dominates (59.6%), followed by Hugging Face (28.9%); any aggregation system must prioritize these two sources. A further 2.7% of papers promise a release without providing a link, indicating that availability is not binary and should be tracked over time.