EvalCards: A Framework for Standardized Evaluation Reporting

Anonymous ACL submission

Abstract

Evaluation has long been a central concern in NLP, and transparent reporting practices are more critical than ever in today's landscape of rapidly released open-access models. Drawing on a survey of recent work on evaluation and documentation, we identify three persistent shortcomings in current reporting practices: reproducibility, accessibility, and governance. We argue that existing standardization efforts remain insufficient and introduce Evaluation Disclosure Cards (EvalCards) as a path forward. EvalCards are designed to enhance transparency for both researchers and practitioners while providing a practical foundation to meet emerging governance requirements.

1 Introduction

Classic scientific scandals often turn on withholding of details around exactly how scientific hypotheses were evaluated. For example, the case of the Piltdown Man, where researchers selectively (mis-)reported crucial contextual details, misled evolutionary sciences for decades (Vincent, 1999). Lack of reporting standards can also lead to confusion, even when everyone acts in good faith. To illustrate, in 19th-century chemistry, the lack of agreed conventions on atomic weights left the field in chaos, with the same compounds appearing under conflicting formulas, until the Karlsruhe Congress established common standards (Ihde, 1961). These (and other similar) episodes (Goldacre, 2009) instill the same lesson: without reliable reporting conventions, even important discoveries can distort rather than advance science.

Evaluation—quantitative measurement of a model's performance on a pre-defined task or benchmark—has long been one of the central means of assessing progress in NLP (Jones, 1994; Church, 2017; Church and Hestness, 2019; Bowman and Dahl, 2021; Kiela et al., 2021; Sainz et al., 2023). Despite this, our standards for reporting evaluations have not kept pace (Bhatt et al., 2021; Belz et al., 2023, 2025a; Zhao et al., 2025a). Such a lack of standards becomes more concerning with the fast adoption of Large Language Models (LLMs) by a wide range of stakeholders, many of whom are not experts and yet heavily depend on such systems to make decisions that impact real-world outcomes (Araujo et al., 2020; Bommasani et al., 2023). As LLMs become embedded in critical domains, responsible deployment is a key consideration (Berengueres, 2024; Radanliev et al., 2024; Orr and Crawford, 2024; Tripathi and Kumar, 2025) and a major part of this includes transparency on what a model can and cannot do. Based on a survey of recent research in the field of evaluation studies, we identify three critical problems stemming from current reporting practices: the *Reproducibility Crisis*, the *Accessibility Crisis*, and the *Governance Crisis*. We discuss why current efforts at transparency (Mitchell et al., 2019; Gebru et al., 2021) need reconsideration. In light of such issues, we propose `EvalCards`): concise evaluation summaries which are (i) *easy to write*, (ii) *easy to understand*, and (iii) *hard to miss*. We present case studies of three popular models, showing how difficult it was to gather consistent evaluation details when creating sample `EvalCards` (see Figure 2), and also discuss directions of future work.

Our main argument is one for a shift in norms: evaluation reporting is not a marketing exercise but a core component of what it means to release a model responsibly. While the broader challenge of how to evaluate models remains open and complex (Laskar et al., 2024; Chang et al., 2024; Gao et al., 2025), our focus here is narrower but nonetheless critical: improving how evaluations are reported. We hope this work sparks conversation and helps move the field toward a culture of more honest and actionable evaluation disclosure practices.

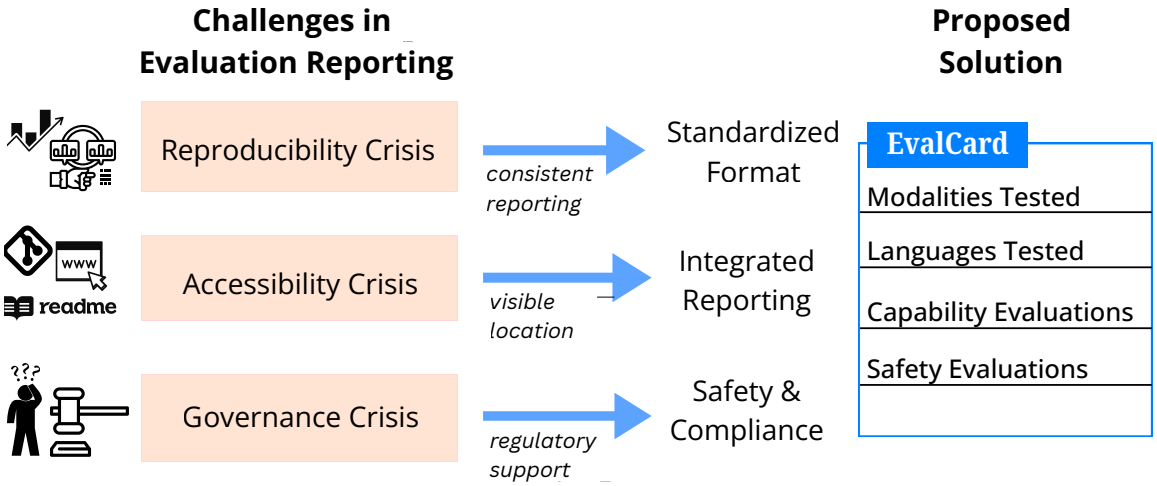

Figure 1: Challenges in evaluation reporting (Section 2) and proposed solutions via `EvalCards` (Section 4). `EvalCards` provide capability reporting to improve reproducibility, a standardized format for accessibility, and safety and compliance documentation for governance.

2 Problems with Evaluation Reporting

To ground our analysis of problems in evaluation reporting in NLP and AI, we conducted a survey of recent work with systematic keyword searches related to evaluation and reporting (e.g., evaluation, reporting, disclosure, evaluation artifacts) in ACL Anthology, DBLP, and Google Scholar. We complemented this with a reverse snowball sampling from the most recent broad-scope seminal works in NLP evaluation as seeds (Weidinger et al., 2025; Gao et al., 2025; Zhao et al., 2025a; Chang et al., 2024; Laskar et al., 2024; Biderman et al., 2024; Burnell et al., 2023; Allen et al., 2021). We then manually analyzed the works to extract recurring themes of discussion. From this, we identify three overarching crises of reproducibility, accessibility, and governance.

2.1 Reproducibility Crisis

In machine learning, the broader reproducibility crisis is well discussed (Kapoor and Narayanan, 2023) and also manifests in evaluation reporting (Bouthillier et al., 2019; Dodge et al., 2019; Belz et al., 2025a; Zhao et al., 2025a). More specifically, in the context of model evaluations, prior work repeatedly highlights that published results often cannot be trusted without careful reconstruction of undocumented choices (Belz et al., 2021; Burnell et al., 2023; Belz et al., 2023; Biderman et al., 2024; Belz et al., 2025a; Zhao et al., 2025a). We conduct some case studies on some recent LLM model releases (see Section 5 for more details) and discuss three such crucial details that are often inconsistent or missing from evaluation details:

Target Capability Model releases include benchmark names and scores, but can often fail to specify what each benchmark is intended to measure. This problem can be exacerbated for reported scores on large composite benchmarks such as SuperGLUE (Wang et al., 2019) or HELM (Bommasani et al., 2023) that aggregate tasks across domains and do not provide clarity, especially for non-experts, as to what capability is targeted.

Metric Reported scores frequently omit which evaluation metric was used, which makes it difficult to assess what a score reflects or to compare performance across models fairly (Mizrahi et al., 2024; Hu and Zhou, 2024).

Prompting Strategy Prompting, i.e, the way a query is structured and phrased for LLMs, is one of the most significant variables in model performance (Hu and Zhou, 2024; Sclar et al., 2024; Zhuo et al., 2024; Chatterjee et al., 2024), yet is often absent from reported results. Also, the absence of consistent reporting on a common baseline strategy, such as zero-shot prompting, further hinders meaningful comparison across models (Kotek et al., 2023; Ousidhoum et al., 2021).

2.2 Accessibility Crisis

A recurring theme in relevant work is the highly fragmented nature of available information on models. Documentation frameworks such as Model Cards, DataSheets, and FactSheets were introduced precisely to improve accessibility of information

about models (Mitchell et al., 2019; Gebru et al., 2021; Arnold et al., 2019; Bhardwaj et al., 2024; Luo et al., 2025). However, even today, evaluation details are dispersed across academic papers, technical appendices, GitHub READMEs, HuggingFace model cards, and blog posts, each using different terminology and presentation styles. This scattered documentation makes it difficult to locate and compare the evaluation results. This has consequences for both researchers and users.

For Researchers When evaluation details are scattered across sources, they are easily overlooked or lost altogether. Important context, such as benchmark versions, question framing, or metric details, may never reach the researchers who rely on these results (Belz et al., 2023; Belz and Thomson, 2024; Belz et al., 2025a,b).

For Users For non-technical users and decision-makers, the problem is compounded by selective reporting, where only strong benchmark results are emphasized or marketed in some sources, further reducing trust in evaluation claims (Arnold et al., 2019). As a result, users may struggle to select appropriate models, a challenge that becomes particularly critical in high-stakes or sensitive deployment contexts (Huijgens et al., 2024; Ben chaaben et al., 2025).

For evaluation to be actionable, it must be consistently visible and easily accessible. Without a standardized way to report evaluations in one place, users are left to piece together incomplete information, undermining efforts to assess a model's suitability for deployment.

2.3 Governance Crisis

AI legislations across the world today, from US to EU (Edwards, 2021; Act, 2024; Sloane and Wüllhorst, 2025; Carey, 2025) and from Singapore to China (Pande and Taeihagh, 2023; Roberts et al., 2021; Dong and Chen, 2024), are increasingly concerned with transparency (Larsson and Heintz, 2020; Agrawal, 2024) and reporting mandates (Nagendran et al., 2020; Laux et al., 2024). Without standardized evaluation reports, governance of models faces three key problems:

Risk Assessment It becomes challenging to determine model risks when the evaluation methods are not clearly reported (Hogan et al., 2021; Novelli et al., 2024; Reuel-Lamparth et al., 2024; Reuel et al., 2024a).

Algorithmic Accountability When developers can selectively report results or omit critical weaknesses, it makes it difficult for external reviewers or regulators to hold systems to consistent standards (Shah, 2018; Wieringa, 2020; Horneber and Laumer, 2023).

Compliance Washing Akin to ethics washing practices (Bietti, 2020), AI developers can satisfy regulatory requirements by disclosing something—even if that "something" is incomplete, selectively positive, or methodologically weak (Koshiyama et al., 2024). Regulatory compliance becomes a box-ticking exercise, undermining the goals of safety, accountability, and public trust (Veale and Zuiderveen Borgesius, 2021).

These crises highlight that the problem is not only how models are evaluated, but how those evaluations are reported. In the next section, we examine limitations of current standardization efforts and introduce `EvalCards` as a practical solution for improving evaluation reporting.

3 Problems with Existing Standards

Calls for transparency in AI development are not new. Over the past years, the research community has proposed frameworks aimed at improving model and dataset documentation, where notable efforts include ModelCards (Mitchell et al., 2019), DataSheets (Gebru et al., 2021), and FactSheets (Arnold et al., 2019). Below, we discuss some issues with such proposed frameworks.

Lack of evaluation focus Existing documentation frameworks rarely place evaluation at the center. Model Cards, DataSheets, and FactSheets typically treat evaluation results as only one component among many. BenchmarkCards (Sokol et al., 2024) focuses on information specific to a single benchmark only, with no reference to models. For large language models, however, evaluation is the primary means by which users, researchers, and regulators can understand capabilities and limitations. While many details about training data or internal design may be sensitive or impractical for industry to disclose, evaluation results can and should be reported clearly, since they provide the most reliable basis for responsible adoption (Allen et al., 2021; Burnell et al., 2023).

High effort for developers Most of the proposed documents, like Model Cards (Mitchell et al., 2019), are lengthy and time-consuming to produce

EvalCard: OLMO-2-1124-7B-Instruct

Modalities Evaluated

Type	Modalities
Input	Text
Output	Text

Languages Evaluated

Category	Languages
Total Number	1
List of Languages	en

Capability Evaluations

Ability	Benchmark	Metric	Zero-Shot	Alternative Prompting
Knowledge	MMLU	Exact Match	61.3 (CoT)	N.A
Knowledge	PopQA	Exact Match	N.A	23.2 (15-Shot)
Knowledge	TruthfulQA	Multiple Choice -2	N.A	56.5 (6-Shot)
Reasoning	BBH	Exact Match	N.A	51.4 (3-Shot CoT)
Reasoning	DROP	F1	N.A	60.5 (3-Shot)
Math	GSM-8K	Exact Match	N.A	85.1 (8-Shot CoT)
Math	MATH	Flex EM	N.A	32.5 (4-Shot CoT)
Instruction Following	AlpacaEval2	LC Winrate	29.1	N.A
Instruction Following	IFEval[1]	Pass@1	72.3	N.A

Safety Evaluations

Feature	Benchmark	Metric	Zero-Shot	Alternative Prompting
Safety	Tülu 3 Safety[2]	N.A	80.6	N.A

Developer Footnotes

[1] The prompt used for IFEval was "Loose".
[2] The Tülu 3 Safety dataset is a combination of 6 datasets testing different aspects of safety.

Figure 2: `EvalCard` for *OLMO-2-1124-7B-Instruct*.

since they require additional analysis like listing out all possible use-cases and users, detailed demographic factor evaluation, intersectional quantitative analyses, etc. This can be especially problematic when many models are being released at a rapid pace. Dedicating extra time to such detailed analysis may not be possible.

Limited accessibility for non-experts For decision-makers, policymakers, and many end-users, these documents are often too technical or jargon-heavy to offer real clarity (McGregor et al., 2025; Crisan et al., 2022). For example, OpenAI's system cards, like the GPT-4o card,[1] offer in-depth safety and governance information but are often very lengthy and complex. As a result, even when provided, they are underutilized by key stakeholders (Blodgett et al., 2024).

Lack of visibility While earlier works have talked about model documentation, not many have emphasized the need for visibility. Today, information about models is frequently buried in supplemental materials, obscure repositories, or separate websites—making it difficult to access and easy to overlook (McGregor et al., 2025).

[1] https://openai.com/index/gpt-4o-system-card/

4 EvalCards

To address the challenges outlined in previous sections, we propose Evaluation Disclosure Cards (`EvalCards`), a short-form standardized reporting format for model evaluations (see Figure 2). In this section, we discuss the design principles of an `EvalCard`, what it should contain, when it should be created, and where it should be available.

4.1 Design Principles of EvalCards

Any reporting format must go beyond existing documentation efforts by tackling the practical barriers identified above: focusing on evaluation, reducing the burden on developers, making results clear to a wide range of users, and ensuring that evaluation information is consistently visible wherever models are accessed. We summarize the design principles here:

Evaluation Focus Unlike broader documentation frameworks such as Model Cards or DataSheets, which include information about training data, intended use cases, and ethical considerations, `EvalCards` place evaluation at the center. The goal is not to capture every possible aspect of model development, but to provide clear and standardized details about what was evaluated, how it was evaluated, and under what conditions. By narrowing the scope to evaluation results, `EvalCards` ensure that the most critical information for understanding and comparing models is reported consistently and without distraction.

Easy to Write For transparency to become standard practice, evaluation reporting must be easy to implement. `EvalCards` are designed to capture only the essential details of model evaluation, making them quick to produce and maintain. This is especially important for smaller organizations and open-source model developers that lack the resources to run large test suites. By focusing on reporting of whatever evaluations have been run, rather than fixing what benchmarks to evaluate on and demanding completeness, `EvalCards` lower the barrier to adoption while still providing meaningful insights into a model's capabilities and risks.

Easy to Understand Transparency is meaningless if only a handful of experts can interpret it. Evaluation reports must be designed for broad accessibility, enabling not just researchers but also all those without domain expertise to grasp a model's capabilities and risks quickly. This requires clear mapping between each benchmark score and the specific capability or risk it evaluates (Liu et al., 2024b), disclosure of evaluation metrics used for each benchmark, and reporting on a common testing format for all benchmarks. For example, ensuring evaluations are consistently reported under at least one comparable prompting strategy (e.g., zero-shot settings) with optional reporting on advanced formats.

Hard to Miss As discussed, evaluation details are buried in academic papers, supplementary materials, or hidden deep within repositories. Standardized evaluation reports should be integrated directly into any landing pages where models can be accessed, whether that is on HuggingFace Hub, API dashboards, or third-party model provider repositories. By ensuring that evaluation disclosures are always visible and linked to the model itself, we would create a culture where understanding a model's capabilities and risks becomes a default part of using models, not an optional deep dive. This visibility also supports downstream accountability, making it easier for regulators, developers, and users to verify claims and make informed decisions (Raji et al., 2020).

In summary, an effective reporting format should place evaluation at its core and satisfy three principles: it must be easy to write, easy to understand, and hard to miss.

4.2 What should an EvalCard contain?

Modalities Evaluated `EvalCards` specify which input and output modalities—such as text, image, or audio—the model has been evaluated on. This upfront clarity helps users quickly assess whether a model aligns with their intended use cases, without wading through documentation. It also supports informed model selection by making functional boundaries visible from the start.

Languages Evaluated As with modalities, clearly stating which languages a model has evaluated on helps define the scope of its real-world applicability. Many models advertise multilingual "training" or "support", but such claims do not indicate whether those languages have been explicitly tested in any systematic way. Without evaluation, such claims can be misleading (Joshi et al., 2020; Blasi et al., 2022; Talat et al., 2022). This section makes the model's evaluated language capabilities transparent, allowing users to quickly ground claims of "language support". We recommend listing language ISO codes to enable easy cross-referencing across language variants (Dalby et al., 2004; Gillis-Webber and Tittel, 2020).

Capability Evaluation Clear evaluation of model capabilities requires more than just reporting benchmark scores. Today, many model reports include benchmark results without indicating what capability the benchmark is intended to measure, how well it does so, or under what prompting conditions. This undermines interpretability and limits the usefulness of the evaluation (Blodgett et al., 2024; Tedeschi et al., 2023). `EvalCards` do not prescribe specific benchmarks, but we require developers to explicitly state which core abilities (e.g., summarization, reasoning, factual recall, mathematical problem solving) were evaluated, and to indicate the benchmark chosen for each. Additionally, all reported results should be accompanied by

the metric used (e.g., exact match, accuracy, precision@1), zero-shot prompting strategy (to enable better baseline comparison across models), and any alternative prompting strategies tested (e.g., few-shot, chain-of-thought).

Safety Evaluation AI models pose well-documented risks, such as bias (Dai et al., 2024), toxicity (Luong et al., 2024), and misinformation generation (Zhang et al., 2024; Chen and Shu, 2024). These issues are often under-reported or selectively presented in current evaluation practices (Burnell et al., 2023; McGregor et al., 2025). `EvalCards` should include a dedicated section for such safety risks. As with capability evaluations, developers should specify the safety feature evaluated, the benchmarks used, the metrics applied, and both the zero-shot and any alternative prompting scores. Consistent and transparent reporting of safety evaluations enables better model comparison and supports emerging regulatory requirements by making known vulnerabilities explicit (Hogan et al., 2021; Ye et al., 2023).

Developer Footnotes In this free-text section of the `EvalCard`, model developers can choose to have relevant footnotes or include any information they think is relevant for users.

4.3 When should EvalCards be created?

`EvalCards` should be generated as part of the initial model release workflow, whether for open-source models or commercial APIs. *First,* model developers are the ones who trained the model, chose the data, designed the architecture, and tuned the objectives. Standardized evaluation reporting works best when it is done by the people who know the model inside out and at the point of release. Also, most of these evaluations align with internal testing already conducted by developers during model validation phases, and direct reporting can prevent multiple runs of the model on the same tests, leading to reduced climate impact. *Second,* most models that require evaluation today—especially large foundation models with hundreds of billions of parameters—are built by organizations with substantial compute resources. If a team can train a model with billions of parameters, it is likely to have enough compute to run a standard suite of evaluations.

By embedding `EvalCard` creation into the release pipeline, developers ensure that transparency is delivered upfront, not left to third-party auditors. Furthermore, `EvalCards` should be updated with each major version change or significant fine-tuning event, reflecting how model behavior may evolve. This keeps users informed of both improvements and potential regressions across the capability and safety dimensions.

4.4 Where should EvalCards be displayed?

Visibility is a core principle of `EvalCards`: Evaluation summaries should appear where the model appears. This includes:

- **Model Repositories**: `EvalCards` should be a standard, prominently linked file in Huggingface Hub or Github—similar to a README or license—ensuring that anyone downloading or browsing the model can immediately access it.
- **Commercial Platforms**: For closed-source or hosted models accessed through APIs or user interfaces (e.g., ChatGPT, Google Gemini), `EvalCards` should be integrated into developer dashboards, product documentation, or user-facing pages. This allows users to review evaluation and safety information before deployment or interaction, supporting informed use and regulatory compliance.
- **Research Publications**: For both open and closed-source models, `EvalCards` should be present as a section in the main text or appendix of academic papers and any technical blogs as a concise summary of evaluation results—providing a clear alternative to selective performance highlights.

By ensuring that `EvalCards` are consistently attached to the points of access and decision-making, we eliminate the common problem where evaluation details are buried in supplementary materials or scattered across inconsistent sources. This "hard to miss" philosophy helps establish a cultural norm: interacting with an NLP model always begins with understanding its core capabilities and risks.

4.5 Model Card vs EvalCard

Model Cards (Mitchell et al., 2019) are an early effort at increasing transparency. However, `EvalCards` are not designed to serve the same purpose and are complementary to Model Cards. Below, we highlight how `EvalCards` differ from

Model Cards and why they are essential in the current landscape:

- **Evaluation Focus:** Model Cards were developed at a time when most models were custom-built, and detailed information about training, design motivations, and use cases was essential. But with the rise of general-purpose foundation models, the training and use-cases are often similar. Most users need to understand capabilities and limitations *more than anything else*.

  → `EvalCards` elevate evaluation to the primary focus and provides a decision-time snapshot to help users and regulators quickly evaluate if a model is appropriate for their needs.

- **Ease of Adoption:** Model Cards focus on open-ended description of model information, including detailed analysis of ethnographic and ethical considerations. This information is useful for theorists and researchers who seek deeper knowledge of the model. However, it can serve as a barrier to adoption by overwhelming end-users with less technical expertise.

  → `EvalCards` are designed to be lightweight and easy to adopt, with structured fields that capture high-signal evaluation details with minimal overhead.

- **Visibility:** Model Cards do not specify where or how they should be displayed, and the information is often buried across multiple sources. However, this is not optimal for evaluation since visible evaluation is key to responsible AI adoption.

  → `EvalCards` are displayed where the model is accessed—on model hubs, APIs, or UIs.

As the norm shifts towards the use of off-the-shelf generative AI models, evaluation becomes the key requirement for responsible model deployment. Evaluation reporting must have its own dedicated and standardized format, which is what `EvalCards` provide. Just as developers include a `model.md` file with model documentation, we propose adding an `eval.md` for evaluation details. This separation makes both model understanding and model assessment more accessible and maintainable as new evaluations become available.

5 EvalCard: Case Studies

To implement our idea, we started with case studies of three popular LLMs. We document the issues we found for collecting details for each model's evaluation followed by the creation of an `EvalCard` for each model[2]. In Figure 2, we provide an example of an `EvalCard` for the model *OLMO-2-1124-7B-Instruct*, from the OLMo project on transparency and open research from the Allen Institute for AI (OLMo et al., 2025). Thanks to its emphasis on openness, creating an `EvalCard` for OLMo 2 was comparatively straightforward. However, we faced few key challenges:

- **Reproducibility:** The metrics and prompt details were difficult to collect, and we often had to manually cross-reference cited research papers and benchmark websites to determine what metric was used. There was also no comparable prompting setup, with each benchmark being tested using very different conditions.

- **Accessibility:** In the research paper, evaluations were scattered in multiple tables across the main paper and the appendix, which was not optimal for a quick glance evaluation of model capabilities.

- **Governance:** The model's safety performance was reported only as a single average across six datasets, without disaggregated scores. This made it impossible to determine how the model behaved on specific risks such as toxicity, bias, or hallucination. The lack of per-dataset transparency limited our ability to meaningfully assess or compare the model's safety profile.

While open-source models like OLMo 2 generally provide more information than their closed counterparts, the process of compiling that information remains time-consuming. `EvalCards` offer a structured way to consolidate key details in one place, reducing the need to spend hours navigating scattered documentation and supplementary sources. In Appendix A, we also provide two more case studies on creating `EvalCards` for models *Qwen3-4B-Base* and *Gemini Flash 2.0* and discuss challenges we faced in the process.

[2]We will release all relevant code for creating EvalCards upon publication.

6 Alternative Views

While we advocate for the adoption of EvalCards, it is important to acknowledge reasonable concerns from different stakeholders. Here, we outline two commonly raised perspectives and address them.

A Developer's View Developers may argue that evaluation is already done internally, and publishing it publicly increases workload or reputational risk, especially when performance is uneven. However, as regulatory frameworks like the EU AI Act and US or UK guidelines begin to demand transparency in model capabilities and risks (Act, 2024), structured evaluation disclosures might become a requirement, not a preference. EvalCards offer a proactive way to meet these expectations. Even seeing what has not been evaluated is useful to avoid misinterpretations and build credibility ahead of external audits or compliance checks (Raji et al., 2020; Koshiyama et al., 2024).

A Researcher's View Researchers may note that evaluation is context-sensitive, where evaluations vary by tasks (Chang et al., 2024) or user groups (Hershcovich et al., 2022a) and that standards for "good" benchmarks are still evolving (Reuel-Lamparth et al., 2024; Liu et al., 2024b; Blodgett et al., 2024). Introducing a fixed reporting format might seem premature. But EvalCards are not rigid templates. They do not enforce benchmark choices but merely require clarity about what was tested and how. This transparency helps everyone interpret results accurately, compare across models, and build on prior evaluations rather than duplicating or misapplying them.

7 Conclusion

Evaluation reporting is a critical but underprioritized part of responsible NLP and AI. Non-standardized practices create hurdles for researchers, users, and regulators, fueling reproducibility, accessibility, and governance crises. Existing documentation efforts like Model Cards (Mitchell et al., 2019) or BenchmarkCards (Sokol et al., 2024) have aimed to improved transparency, but they do not put evaluation at the center. EvalCards aim to close this gap with a format that is easy to write, easy to understand, and hard to miss. By making evaluation details visible and consistent, they turn scattered disclosures into a foundation for cumulative research, informed adoption, and accountability. Looking ahead, we highlight three directions to strengthen and extend the EvalCard framework.

Increased Transparency EvalCards can help establish a unified pipeline that links evaluation and reporting into a single transparent process by reducing ambiguity across sources (Biderman et al., 2024). In the longer term, EvalCards could link to Benchmark Cards(Sokol et al., 2024) for each mentioned benchmark, creating a connected reporting ecosystem where model, benchmark, and evaluation details are transparently linked to ensure that both the tests and the results behind model claims are easy to trace and verify.

Increased Adoption Widespread adoption will require making EvalCards easy to produce and use. Methods like automated extraction from technical reports (Liu et al., 2024a) or community contribution to scores (Zhao et al., 2025b) can help ease the burden on model developers, while usability testing can help refine design and language for diverse stakeholders (Crisan et al., 2022). This will help ensure that EvalCards are not only technically sound but also popular.

Regulatory Integration : EvalCards can play a key role in bridging technical evaluation and regulatory transparency. They can serve as a standardized reporting format for models within compliance processes, such as regulatory sandboxes under the EU AI Act (Lanamäki et al., 2025) or US AI Risk Management Framework (AI, 2023), by offering a consistent way to report model performance and limitations. Partnerships with platforms like HuggingFace or emerging standards bodies could help maintain vetted benchmark sets that align with evolving priorities. Future work in technical governance (Reuel et al., 2024a,b) can explore how EvalCards can be incorporated into regulatory compliance workflows, e.g., by establishing minimum mandatory fields tied to emerging AI regulations.

As model development accelerates, reporting practices must evolve with equal urgency. Evaluation should drive progress, not confusion—but that is only possible when what models can and cannot do is made clear. EvalCards take a small but concrete step toward that goal, embedding transparency into the model release process itself. We hope this work sparks deeper reflection and concrete action toward standardizing how we evaluate and report on models.

Limitations

This work focuses on the design and motivation of `EvalCards` as a framework for standardized evaluation reporting. While we ground our proposal in a survey of prior literature and case studies of recent model releases, our analysis has a few limitations.

First, the scope of fields covered in `EvalCards` is not necessarily complete. While we emphasize evaluation for capabilities and safety, many other aspects are becoming increasingly relevant to stakeholders today, such as: efficiency (Husom et al., 2025), latency (Fan and Tao, 2025), cost, and environmental impact (Hershcovich et al., 2022b; Samsi et al., 2023; Luccioni et al., 2025). Extending the framework to incorporate these dimensions will be important for ensuring comprehensiveness.

Second, our case studies are illustrative rather than exhaustive. We highlight recurring problems by examining a small set of model releases, but we do not conduct a large-scale quantitative audit across the full landscape of open and closed-source models. Future work could extend our survey to systematically measure reporting gaps across hundreds of releases.

Third, we do not evaluate user adoption or usability of `EvalCards`. Although we outline design principles aimed at lowering developer burden and improving accessibility for non-experts, we have not yet validated these principles through user studies or integration with model release platforms. Assessing the best design for such reports and how well they meet the needs of diverse stakeholders remains an open question for future work. While our focus is on capabilities and safety, many other aspects are relevant for specific stakeholders—for example, climate impact (Hershcovich et al., 2022b).

Finally, while we discuss alignment with emerging governance frameworks, we stop short of specifying how EvalCards would be formally incorporated into compliance or auditing processes. This work is therefore best seen as a step toward standardization rather than a fully implemented solution.

Ethical Considerations

We do not anticipate any risks in our work. We used three models for our case studies under appropriate license considerations (in Section 4 and Appendix A) and have provided model research papers or linked to source documentation as applicable for each model. We have created the `EvalCard` artifact in our paper and will share all relevant code with relevant license for future use upon publication.

References

EU Artificial Intelligence Act. 2024. The eu artificial intelligence act.

Gaurav Agrawal. 2024. Accountability, trust, and transparency in ai systems from the perspective of public policy: Elevating ethical standards. In *AI healthcare applications and security, ethical, and legal considerations*, pages 148–162. IGI Global.

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

Bibb Allen, Keith Dreyer, Robert Stibolt Jr, Sheela Agarwal, Laura Coombs, Chris Treml, Mona Elkholy, Laura Brink, and Christoph Wald. 2021. Evaluation and real-world performance monitoring of artificial intelligence models in clinical practice: try it, buy it, check it. *Journal of the American College of Radiology*, 18(11):1489–1496.

Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI & society*, 35(3):611–623.

Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, and 1 others. 2019. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Simon Mille, and Craig Thomson. 2025a. Standard quality criteria derived from current NLP evaluations for guiding evaluation design and grounding comparability and AI compliance assessments. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26685–26715, Vienna, Austria. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 91–105, Torino, Italia. ELRA and ICCL.

Anya Belz, Craig Thomson, Javier González Corbelle, and Malo Ruelle. 2025b. The 2025 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 1002–1016, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3676–3687, Toronto, Canada. Association for Computational Linguistics.

Eya Ben chaaben, Janin Koch, and Wendy E. Mackay. 2025. "should i choose a smaller model?': Understanding ml model selection and its impact on sustainability. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

J. Berengueres. 2024. How to regulate large language models for responsible ai. *IEEE Transactions on Technology and Society*, 5(2):191–197.

Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, and Christoph Becker. 2024. Machine learning data practices through a data curation lens: An evaluation framework. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1055–1067.

Shaily Bhatt, Rahul Jain, Sandipan Dandapat, and Sunayana Sitaram. 2021. A case study of efficacy and challenges in practical human-in-loop evaluation of nlp systems using checklist. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 120–130.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, and 1 others. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Elettra Bietti. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 210–219.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 39–43, Miami, Florida, USA. Association for Computational Linguistics.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Xavier Bouthillier, César Laurent, and Pascal Vincent. 2019. Unreproducible research is reproducible. In *International Conference on Machine Learning*, pages 725–734. PMLR.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

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, and 1 others. 2023. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138.

Samuel Carey. 2025. Regulating uncertainty: Governing general-purpose ai models and systemic risk. *European Journal of Risk Regulation*, pages 1–17.

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.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14550–14565, Miami, Florida, USA. Association for Computational Linguistics.

Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368.

Kenneth Ward Church. 2017. Emerging trends: I did it, i did it, i did it, but... *Natural Language Engineering*, 23(3):473–480.

Kenneth Ward Church and Joel Hestness. 2019. A survey of 25 years of evaluation. *Natural Language Engineering*, 25(6):753–767.

Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation.

In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 427–439.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447.

David Dalby, Lee Gillam, Christopher Cox, and Debbie Garside. 2004. Standards for language codes: developing ISO 639. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Huijuan Dong and Junkai Chen. 2024. Meta-regulation: An ideal alternative to the primary responsibility as the regulatory model of generative ai in china. *Computer Law & Security Review*, 54:106016.

Lilian Edwards. 2021. The eu ai act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)*, 1:25.

Xiaojing Fan and Chunliang Tao. 2025. Towards resilient and efficient llms: A comparative study of efficiency, performance, and adversarial robustness. In *Proceedings of the 2024 7th Artificial Intelligence and Cloud Computing Conference*, AICCC '24, page 429–436, New York, NY, USA. Association for Computing Machinery.

Mingqi Gao, Xinyu Hu, Xunjian Yin, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, pages 1–28.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Frances Gillis-Webber and Sabine Tittel. 2020. A framework for shared agreement of language tags beyond ISO 639. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3333–3339, Marseille, France. European Language Resources Association.

Ben Goldacre. 2009. *Bad science*. Harper Perennial.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022a. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022b. Towards climate awareness in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Neil R Hogan, Ethan Q Davidge, and Gabriela Corabian. 2021. On the ethics and practicalities of artificial intelligence, risk assessment, and race. *The Journal of the American Academy of Psychiatry and the Law*, 49(3):326–334.

David Horneber and Sven Laumer. 2023. Algorithmic accountability. *Business & information systems engineering*, 65(6):723–730.

Taojun Hu and Xiao-Hua Zhou. 2024. Unveiling llm evaluation focused on metrics: Challenges and solutions. *arXiv preprint arXiv:2404.09135*.

Hennie Huijgens, Devrim Eskiyerli, and Burcu Kör. 2024. Help! which ai should i choose?-work in progress. In *The 1st Human-Centered AI Education Practice Conference, Dublin, Ireland*.

Erik Johannes Husom, Arda Goknil, Merve Astekin, Lwin Khin Shar, Andre Kåsen, Sagar Sen, Benedikt Andreas Mithassel, and Ahmet Soylu. 2025. Sustainable llm inference for edge ai: Evaluating quantized llms for energy efficiency, output accuracy, and inference latency. *ACM Transactions on Internet of Things*.

Aaron J Ihde. 1961. The karlsruhe congress: a centennial retrospective. *Journal of Chemical Education*, 38(2):83.

Karen Spärck Jones. 1994. Towards better nlp system evaluation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9).

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Adriano Koshiyama, Emre Kazim, Philip Treleaven, Pete Rai, Lukasz Szpruch, Giles Pavey, Ghazi Ahamat, Franziska Leutner, Randy Goebel, Andrew Knight, and 1 others. 2024. Towards algorithm auditing: managing legal, ethical and technological risks of ai, ml and associated algorithms. *Royal Society Open Science*, 11(5):230859.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Arto Lanamäki, Karin Väyrynen, Fanny Vainionpää, Heidi Hietala, Erkki Tervo, Alex Moltzau, Sophie Weerts, and Demian Niemeyer. 2025. What to expect from the upcoming eu ai act sandboxes: Panel report. *Digital Society*, 4(2):42.

Stefan Larsson and Fredrik Heintz. 2020. Transparency in artificial intelligence. *Internet policy review*, 9(2):1–16.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, and 1 others. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816.

Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2024. Three pathways for standardisation and ethical disclosure by default under the european union artificial intelligence act. *Computer Law & Security Review*, 53:105957.

Jiarui Liu, Wenkai Li, Zhijing Jin, and Mona Diab. 2024a. Automatic generation of model and data cards: A step towards responsible AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1975–1997, Mexico City, Mexico. Association for Computational Linguistics.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024b. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand. Association for Computational Linguistics.

Alexandra Sasha Luccioni, Giada Pistilli, Raesetje Sefala, and Nyalleng Moorosi. 2025. Bridging the gap: Integrating ethics and environmental sustainability in ai research and practice. *arXiv preprint arXiv:2504.00797*.

Xufei Luo, Bingyi Wang, Qianling Shi, Zijun Wang, Honghao Lai, Hui Liu, Yishan Qin, Fengxian Chen, Xuping Song, Long Ge, and 1 others. 2025. Lack of methodological rigor and limited coverage of generative ai in existing ai reporting guidelines: A scoping review. *Journal of Clinical Epidemiology*, page 111903.

Tinh Luong, Thanh-Thien Le, Linh Ngo, and Thien Nguyen. 2024. Realistic evaluation of toxicity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1038–1047, Bangkok, Thailand. Association for Computational Linguistics.

Sean McGregor, Allyson Ettinger, Nick Judd, Paul Albee, Liwei Jiang, Kavel Rao, William H Smith, Shayne Longpre, Avijit Ghosh, Christopher Fiorelli, and 1 others. 2025. To err is ai: A case study informing llm flaw reporting practices. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28938–28945.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

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.

Myura Nagendran, Yang Chen, Christopher A Lovejoy, Anthony C Gordon, Matthieu Komorowski, Hugh Harvey, Eric J Topol, John PA Ioannidis, Gary S Collins, and Mahiben Maruthappu. 2020. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *bmj*, 368.

Claudio Novelli, Federico Casolari, Antonino Rotolo, Mariarosaria Taddeo, and Luciano Floridi. 2024. Taking ai risks seriously: a new assessment model for the ai act. *AI & SOCIETY*, 39(5):2493–2497.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

Will Orr and Kate Crawford. 2024. Building better datasets: Seven recommendations for responsible design from dataset creators. *Journal of Data-centric Machine Learning Research*.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

Devyani Pande and Araz Taeihagh. 2023. Navigating the governance challenges of disruptive technologies: insights from regulation of autonomous systems in singapore. *Journal of Economic Policy Reform*, 26(3):298–319.

Petar Radanliev, Omar Santos, Alistair Brandon-Jones, and Adam Joinson. 2024. Ethics and responsible ai deployment. *Frontiers in Artificial Intelligence*, 7:1377011.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44.

Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, and 1 others. 2024a. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*.

Anka Reuel, Lisa Soder, Benjamin Bucknall, and Trond Arne Undheim. 2024b. Position: Technical research and talent is needed for effective ai governance. In *Forty-first International Conference on Machine Learning*.

Anka Reuel-Lamparth, 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.

Huw Roberts, Josh Cowls, Jessica Morley, Mariarosaria Taddeo, Vincent Wang, and Luciano Floridi. 2021. *The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation*. Springer.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. 2023. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–9. IEEE.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Hetan Shah. 2018. Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170362.

Mona Sloane and Elena Wüllhorst. 2025. A systematic review of regulatory strategies and transparency mandates in ai regulation in europe, the united states, and canada. *Data & Policy*, 7:e11.

Anna Sokol, Elizabeth Daly, Michael Hind, David Piorkowski, Xiangliang Zhang, Nuno Moniz, and Nitesh Chawla. 2024. Benchmarkcards: Standardized documentation for large language model benchmarks. *arXiv preprint arXiv:2410.12974*.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

Ajay Tripathi and Vinod Kumar. 2025. Ethical practices of artificial intelligence: a management framework for responsible ai deployment in businesses. *AI and Ethics*, pages 1–12.

Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112.

John B Vincent. 1999. Piltdown man: Combining the instruction of scientific ethics and qualitative analysis. *Journal of chemical education*, 76(11):1501.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*.

Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 1–18.

Wentao Ye, Mingfeng Ou, Tianyi Li, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Haobo Wang, Junbo Zhao, and 1 others. 2023. Assessing hidden risks of llms: an empirical study on robustness, consistency, and credibility. *arXiv preprint arXiv:2305.10235*.

Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1302–1305.

Dora Zhao, Qianou Ma, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. 2025a. SPHERE: An evaluation card for human-AI systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1340–1365, Vienna, Austria. Association for Computational Linguistics.

Zhimin Zhao, Abdul Ali Bangash, Filipe Roseiro Côgo, Bram Adams, and Ahmed E. Hassan. 2025b. On the workflows and smells of leaderboard operations (lbops): An exploratory study of foundation model leaderboards. *IEEE Transactions on Software Engineering*, 51(4):929–946.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A `EvalCard` Case Studies

In this section, we provide `EvalCards` for two more models and also discuss some issues we found while trying to create these.

A.1 Model *Qwen3-4B-Base*

Qwen3-4B-Base, released on the 29th of April 2025, for illustration purposes in Figure 3. The model is released under the Apache 2.0 license and is accessible through platforms like Hugging Face, ModelScope, and Kaggle. During our attempt at creating this card, we noticed the following:

- **Delayed Evaluations**: Up until a week after the release of the model, there were only a few evaluations present in the Blog. This is problematic for responsible and informed use.
- **Ambiguous Language Information**: The blog mentions the 119 languages trained on, but does not clearly mention which languages have been evaluated. Evaluations mention some benchmarks, leaving it up to users to spend hours examining benchmarks to find what languages they contain.
- **Ambiguous Capability Evaluations**: According to the blog and HuggingFace repository, the model excels "in creative writing, role-playing, multi-turn dialogues, and instruction following" and "significantly enhanced reasoning capabilities". However, there is no information on exactly which benchmarks or datasets have been used to test these capabilities. The updated technical report provides some evaluations with vague capabilities like "General Task" and "Alignment" or conflating different capabilities like "Math & Text Reasoning" or "Agent & Coding". This is extremely problematic for end-users who are not familiar with benchmarks and can easily misinterpret phrases like "General Tasks" or "Alignment".
- **Absent Evaluation Details**: Information on evaluation metrics or prompting strategies was not available for many benchmark scores reported in the blog. Since there was no way to verify whether the prompting strategy was zero-shot or not, we had to place the available scores in the alternative prompting score column.

- **Absent Safety Evaluations**: There are no specific safety evaluations reported yet, even though the model has been released.

In our `EvalCard`, we provide the information on cases where the capability could be clearly deciphered from the technical report or our laboriously painstaking personal investigation into benchmarks mentioned. Creating `EvalCards` would also likely encourage developers to not only be more open about evaluation metrics and prompts but also be more aware about what capability they aim to test—instead of bombarding users with random benchmark scores.

A.2 Model *Gemini Flash 2.0*

We provide an example of an `EvalCard` for the model *Gemini Flash 2.0* in Figure 4, accessible via the Gemini API and described in both the official Developer Studio documentation[3] and a December 2024 blog post[4]. Although there is a basic Model Card available in the Developer Studio, several challenges made it difficult to construct a comprehensive `EvalCard` for the model.

Ambiguous multilingual support While the Developer Studio lists languages the model is "trained to work on", it does not specify whether the model's performance on those languages was ever evaluated, leaving users uncertain about where the model has been tested versus where support is simply assumed due to training.

Absent evaluation details For all benchmark scores, information on metric and prompting strategy was unavailable—which is why we place scores in the alternative prompting column. This reduces understanding of capabilities and leaves users without guidelines on which prompting methods work.

Absent safety evaluations The blog post references a review by Google's Responsibility and Safety Committee (RSC) to "identify and understand potential risks", but does not disclose any concrete safety findings, red-teaming procedures, or benchmarks used for this purpose.

[3] https://ai.google.dev/gemini-api/docs/models

[4] https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#gemini-2-0

EvalCard: Qwen3-4B-Base

Modalities Evaluated

Type	Modalities
Input	Text
Output	Text

Languages Evaluated

Category	Languages
Total Number	66
List of Languages	af, ar, az, be, bg, bn, ca, cs, cy, da, de, el, en, es, et, eu, fa, fi, fr, gu, he, hi, hr, hu, hy, id, it, ja, ka, kk, kn, ko, lt, lv, mk, ml, mr, ms, ne, nl, no, pa, pl, pt, ro, ru, sk, sl, so, sq, sr, sv, sw, ta, te, th, tl, tr, uk, ur, uz, vi, zh, zh-Hans, zh-Hant

Capability Evaluations

Ability	Benchmark	Metric	Zero-Shot	Alternative Prompting
General Knowledge	MMLU-Redux	N.A	N.A	77.3
General Knowledge	GPQA-Diamond	Accuracy	N.A	41.7
Math	MATH-500	N.A.	N.A	84.8
Math	AIME-24	Accuracy	N.A	25.0
Math	AIME-25	Accuracy	N.A	19.1
Reasoning	ZebraLogic	N.A.	N.A	35.2
Reasoning	AutoLogi	N.A.	N.A	76.3
Coding	LiveCodeBench v5	N.A	N.A	21.3
Coding	CodeForces	Rating/Percentile	N.A	842/33.7%
Instruction Following	IFEval	Accuracy	N.A	81.2
Tool Integration	BFCL v3	FC Format	N.A	57.6

Safety Evaluations

Feature	Benchmark	Metric	Zero-Shot	Alternative Prompting
N.A	N.A	N.A	N.A	N.A

Developer Footnotes

None

Figure 3: `EvalCard` for *Qwen3-4B-Base*.

EvalCard: Gemini Flash 2.0

Modalities Evaluated

Type	Modalities
Input	Text, Image, Video, Audio
Output	Text

Languages Evaluated

Category	Languages
Total Number	N.A
List of Languages	N.A

Capability Evaluations

Ability	Benchmark	Metric	Zero-Shot	Alternative Prompting
General Knowledge	MMLU-Pro	N.A	N.A	76.6
Factual Knowledge	FACTS Grounding	N.A	N.A	83.6
Code	Natural2Code	N.A	N.A	92.9
Code	Bird-SQL Dev	N.A	N.A	56.9
Code	LiveCodeBench	N.A	N.A	35.1
Math	MATH	N.A	N.A	89.7
Math	HiddenMath	N.A	N.A	63
Reasoning	GPQA(Diamond)	N.A	N.A	62.1
Long Context	MRCR(1M)	N.A	N.A	69.2
Image	MMMU	N.A	N.A	70.7
Image	Vibe-Eval(Reka)	N.A	N.A	56.3
Video	EgoSchema(Test)	N.A	N.A	71.5
Audio	CoVoST2 (21Lang)	N.A	N.A	39.2

Safety Evaluations

Feature	Benchmark	Metric	Zero-Shot	Alternative Prompting
N.A	N.A	N.A	N.A	N.A

Developer Footnotes

As part of our safety process, we've worked with our Responsibility and Safety Committee (RSC), our longstanding internal review group, to identify and understand potential risks.

Figure 4: `EvalCard` for *Gemini Flash 2.0*.