01em

0.01em

0.0.01em

Navigating the Model Labyrinth: Adapting PRISMA for Systematic Selection in the Hugging Face Ecosystem

Andrew Wellman Taylor, Anwar Sleiman Haidar, Ana Luiza Ruskowski Mees, Carina Rodriguez, Fatih Karatay, Madihah Shaik, Sarah Spence, and Sarv Parteek Singh

Whiting School of Engineering EP Program
Johns Hopkins University
Baltimore, MD, USA

Abstract

The proliferation of publicly available machine learning models—especially on platforms such as Hugging Face—has democratized access to cutting-edge tools while simultaneously introducing a paradox of abundance: the more models are available, the harder it becomes to choose one. In this context, model selection is no longer merely a technical decision, but a methodological and epistemological challenge.

1 Introduction

The PRISMA framework (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) was designed to address a comparable crisis in the medical sciences [Moher et al., 2009]. When clinical studies exploded in volume, researchers struggled to synthesize results coherently and reproducibly. PRISMA introduced a rigorously structured protocol for systematically searching, selecting, evaluating, and synthesizing heterogeneous sources of evidence [Page et al., 2021]. While born in healthcare, PRISMA's intellectual scaffolding provides a compelling analogue for responsible model selection in AI—especially in open-source ecosystems where model availability far outpaces validation.

This paper explores how PRISMA can inform and transform how we evaluate and select models from Hugging Face, not as a rigid prescription but as a set of guiding principles grounded in transparency, reproducibility, and evidence-based reasoning.

2 From Clinical Trials to Transformers: Why PRISMA Matters in AI

AI research and deployment increasingly mirror the complexity of clinical domains: stakes are high, interventions are opaque, and outcomes are context-dependent. A Hugging Face model may outperform rivals on a benchmark dataset yet fail silently when applied in the wild due to unseen distribution shifts or deployment constraints.

In medicine, PRISMA helps synthesize diverse, often contradictory, findings into coherent assessments. Similarly, ML practitioners need a framework to adjudicate among models with varying licenses, training corpora, hyperparameters, evaluation criteria, and reporting quality. PRISMA's core components—predefined inclusion criteria, structured search protocols, bias assessment, evidence synthesis, and reproducibility—are transferable to model selection with minimal modification.

3 Systematic Model Discovery: Formalizing the Search Phase

Model selection often begins with vague searches like "text-classification" or "summarization" on Hugging Face. While such queries yield results, they lack intentionality. PRISMA demands that search strategies be systematic, documented, and replicable [Moher et al., 2009]. This means going beyond keyword filters to incorporate semantic search, architectural taxonomies (e.g., transformers, CNNs, diffusion models), and dependency graphs (e.g., fine-tuning lineage or pretraining base).

Federated search extensions across GitHub, arXiv, and institutional repositories align with PRISMA's recommendation to broaden the evidence base beyond a single publication venue. Further, structured metadata—model card completeness, release date, update frequency, institution of origin—can serve as proxies for credibility and maintainability [Mitchell et al., 2019].

Advanced search tactics might include clustering models by performance on related tasks or by shared fine-tuning objectives, helping uncover high-potential candidates that lack widespread exposure.

4 Inclusion and Exclusion: Making Criteria Explicit

One of PRISMA's most consequential features is that it forces researchers to declare in advance what will qualify as acceptable evidence—and why. This has a direct analogue in model selection. Inclusion criteria might include licensing constraints (e.g., only Apache 2.0 models), computational feasibility (e.g., $\leq$500MB memory footprint), support for multilingual inference, or evidence of responsible development practices (e.g., ethical considerations in the model card).

On the exclusion side, the lack of a model card, missing evaluation scripts, absence of downstream usage citations, or outdated architecture may constitute legitimate grounds. Cru-

cially, PRISMA insists that every exclusion be logged with justification—a principle that could powerfully counteract the unconscious biases that plague model selection today.
Building on these principles, a PRISMA-informed search strategy might look like this:

Ordering to a Filtered Screening	Criteria
Step 1: Define Initial Search Query (Text Search)	Explicit terms of scope, relevance
Step 2: Apply Hugging Face Ecosystem Filters	Filter 1: Task Tags
	Filter 2: Language
	Filter 3: Model Architecture
	Filter 4: License
	Filter 5: Model Size
	Filter 6: Last updated
Step 3: Secondary Filters	Filter 7: Popularity Indicators
	Filter 8: Model Card Completeness
Step 4: Optional AI-Driven Filters (Semantic Search)	Filter 9: Semantic Search (Embeddings)

This is a proposed ordering of filters that may be best for general use. The actual ordering of a search may depend on the user's priorities and thought process. For instance, some like to trust the wisdom of the crowd, using popularity indicators early on in the process of selection. However, this is proposed as one formulation that makes sense generally for most tasks: task type, then language of the model, and so on, would presumably be issues that must be decided first. Other desiderata would follow naturally. One interesting option is semantic search via "sentence-transformers" or Hugging Face's built-in embeddings-based searches. The justification for that being semantic filtering captures relevance beyond keywords alone, aligning closely with intended use cases. There are a lot of orderings possible that would clearly alter results. The initial filters should quickly narrow down results to relevant, feasible options, while subsequent filters refine quality, ethical standards, and community validation, aligning closely with PRISMA's structured, progressive approach to systematic selection.
There are many other considerations not yet mentioned that could factor into the process including (perhaps most obviously) the performance metrics reported, bias and ethical factors, documentation quality, endorsement, deployment documentation, energy and efficiency metrics, robustness to dataset shift, specific repository activity, and the community or support that exists. The following sections will dive into some of these criteria and what they mean to model selection, and why do they matter.

5 Bias, Risk, and Robustness: From Box Checking to Meaningful Assessment

Bias in machine learning has become a recognized challenge, but how it's measured—or whether it's measured at all—varies widely. PRISMA's emphasis on bias assessment offers a corrective. While model cards may mention limitations, very few quantify them or examine them across demographic strata.
A PRISMA-like model selection would include subgroup performance analysis (e.g., disaggregated accuracy across race, gender, geography), adversarial robustness tests, calibration

error estimates, and fairness audits. Moreover, just as meta-analyses distinguish between methodological bias and reporting bias, a model review might differentiate between training data imbalance and cherry-picked evaluations.
Publication bias—where only high-performing results are published—also pervades AI. Funnel plots and asymmetry tests, common in clinical meta-analysis, could be adapted to detect suspiciously uniform or inflated model scores, particularly among corporate releases.

6 Evidence Synthesis and Transitive Benchmarking

The Hugging Face ecosystem suffers from evaluation fragmentation: models are often tested on disjoint datasets or under inconsistent conditions. PRISMA provides a template for resolving this through standardized metric normalization and transitive benchmarking.
Just as medical studies compute effect sizes (e.g., relative risk), ML researchers could normalize performance gains (e.g., Δ F1 per 10M parameters, or improvement per unit of training compute). These can be pooled across datasets using meta-analytical models, allowing the emergence of global winners or performance-efficiency frontiers.
Network meta-analysis allows for indirect comparisons—if Model A beats B on Dataset X and B beats C on Dataset Y, one can infer A's likely standing relative to C.

7 Heterogeneity and Generalizability

Not all models generalize equally. PRISMA's attention to heterogeneity—variability in effect size not explained by chance—offers another layer of insight. In AI, this might translate to measuring how model performance varies across deployment conditions: hardware platforms, input noise, low-resource domains, or regional dialects.
Meta-regression could examine the impact of scale (e.g., parameter count), training set diversity, or supervision signal quality on downstream performance.
Heterogeneity is critical and touches upon a central tension in enterprise model adoption. While the PRISMA-inspired approach rightly pushes us towards standardized, transitive benchmarking to resolve evaluation fragmentation, we must be cautious not to create a new kind of "evaluation monoculture." A model that excels on a suite of standardized benchmarks—even one designed to test for generalizability—may still fail silently when faced with the unique, high-variance "long tail" of production data.
For example, a model's performance might not just vary across broad categories like "hardware platforms" or "regional dialects," but across more subtle, emergent contexts: a specific OEM's sensor data, a single user's idiosyncratic input patterns, or a sudden shift in market behavior. Therefore, a key part of the discussion isn't just
if a model generalizes, but how we build a system that gracefully handles the inevitable moments when it doesn't. This involves architecting for fallback mechanisms, continuous online monitoring to detect performance anomalies in specific subprocesses, and creating a rapid feedback loop from production systems back into the model selection and fine-tuning process. The PRISMA framework provides the rigor for initial selection, but adapting it to a live environment requires embracing this heterogeneity as an ongoing operational reality, not just a variable to be measured at a single point in time.

8 Lifecycle and Temporal Relevance

Unlike static interventions, ML models evolve. Some decay due to concept drift; others are rendered obsolete by new architectures. PRISMA's cumulative meta-analysis and time-aware evidence synthesis offer frameworks for capturing this temporal dimension.
Update frequency and GitHub activity may act as proxies for maintenance—akin to survival analysis in biomedical domains.

9 Resource Efficiency and Pareto Optimization

AI adoption is increasingly gated by deployment constraints—compute, power, latency, storage. PRISMA invites us to borrow from healthcare's cost-effectiveness analyses to evaluate models not just by accuracy but by efficiency per cost unit.
Metrics like FLOPs-per-inference, energy-per-token, or accuracy-per-MB can feed into multi-objective optimization, where the goal is not to find the best model, but the best trade-off given operational constraints.

10 Reproducibility, Documentation, and Accountability

The final virtue PRISMA imparts is a commitment to transparency. This includes pre-registration of evaluation criteria, version-controlled benchmarks, and open-source replication kits. Tools like `asreview`, model cards, datasheets for datasets, and the NeurIPS reproducibility checklist already gesture in this direction—but a PRISMA-inspired selection pipeline would require these by default.
Moreover, documenting the selection process itself—what was searched, what was excluded, and why—enables teams to audit their own decisions, replicate past studies, and defend choices to stakeholders or regulators [Pineau et al., 2020, van de Schoot et al., 2021].

11 Putting it all together: PRISMA for Hugging Face

Now that we have covered all the core components of PRISMA as they apply to Hugging Face, let us consolidate these and do a side-by-side comparison of the original PRISMA table [Moher et al., 2009] with the meaning of that feature when applied to the Hugging Face model selection problem:

Item	Checklist Item Description	Application to Hugging Face Model Selection
	TITLE	

Continued on next page...

...continued from previous page

Item	Checklist Item Description	Application to Hugging Face Model Selection
1	Identify the report as a systematic review.	**Project Titling:** Clearly title your project document to reflect its purpose, e.g., "A Systematic Selection of a Transformer Model for Sentiment Analysis of Customer Feedback."
		ABSTRACT
2	Provide a structured summary.	**Executive Summary:** Write a concise summary of the task, key selection criteria (e.g., accuracy, license, size), shortlisted models, and the final "winning" model with its benchmark performance.
		INTRODUCTION
3	Rationale	**Problem Statement:** Explain the business or research need. For instance, "Current manual methods are inefficient; we need an automated solution for analyzing customer reviews."
4	Objectives	**Define Goals:** State the explicit goal, such as "To identify and validate the best-performing open-source model from the Hugging Face Hub for our specific text classification task."
		METHODS
5	Eligibility criteria	**Inclusion/Exclusion Criteria:** Define non-negotiable requirements. *Inclusion:* PyTorch framework, MIT or Apache 2.0 license. *Exclusion:* Models requiring >16GB VRAM, models without a model card.
6	Information sources	**Define Search Space:** State the primary source is the Hugging Face Hub. Note if other sources like research papers or GitHub repositories were also consulted.
7	Search strategy	**Document Search Queries:** Record the exact filters and keywords used on the Hub (e.g., `task:text-classification`, `language:en`, `library:pytorch`, `sort:downloads`).
8	Selection process	**Screening Workflow:** Describe how models were screened. For example, "Two team members independently reviewed the top 20 models from the search results to reduce bias."
9	Data collection process	**Information Extraction:** Detail how data was gathered from model cards. For example, using a shared spreadsheet to log details for each potential model.
10	Data items	**List of Variables:** List all extracted data points: model name, architecture, license, downloads, reported metrics (e.g., F1-score), and model size.

Continued on next page...

...continued from previous page

Item	Checklist Item Description	Application to Hugging Face Model Selection
11	Risk of bias assessment	**Model Limitation Analysis:** Assess each model's potential flaws. Review the "Bias and Limitations" section of the model card and consider if the training data aligns with your use case.
12	Effect measures	**Define Your Metrics:** Specify the metrics for your own benchmark (e.g., "Primary metric: Macro F1-score. Secondary metrics: Inference latency and memory usage on our hardware.")
13	Synthesis methods	**Benchmarking Protocol:** Describe the head-to-head comparison plan. "Finalist models will be fine-tuned on our training set and evaluated on our hold-out test set."
		RESULTS
17	Study selection	**Flow Diagram:** Report the numbers for your selection funnel, ideally with a PRISMA-style flow diagram. (e.g., "150 models identified, 30 screened, 5 eligible, 3 included in final benchmark.")
18	Study characteristics	**Model Summaries:** Present a table summarizing the characteristics of the finalist models (e.g., architecture, parameters, original training data).
20	Results of individual studies	**Benchmark Results Table:** Display the performance of each finalist model on your own test data, according to the metrics you defined in the methods section.
21	Results of syntheses	**Summary of Findings:** Provide a summary statement of the benchmark results. For example, "RoBERTa-base achieved the highest F1-score of 0.92, while DistilBERT was 30% faster."
		DISCUSSION
24	Interpretation of results	**Analysis and Trade-offs:** Discuss the results in context. Explain why one model might have outperformed others and discuss the trade-offs (e.g., accuracy vs. speed).
25	Limitations	**Process Limitations:** Acknowledge the limitations of your selection process. "Our search was confined to the Hugging Face Hub; performance may differ on out-of-distribution data."
		OTHER INFORMATION
26	Registration and protocol	**Internal Documentation:** For business settings, link to the internal protocol document (e.g., on a wiki) that was created before the selection process began.

Continued on next page...

...continued from previous page

Item	Checklist Item Description	Application to Hugging Face Model Selection
27	Support & Data Availability	**Reproducibility:** State where the benchmarking code, evaluation scripts, and final results are stored (e.g., in a shared Git repository) to ensure transparency and reproducibility.

12 Limitations and Considerations

While the PRISMA framework offers valuable principles for model selection, its direct application to machine learning is not without challenges. Unlike clinical studies, machine learning evaluations often lack standardized reporting protocols. Reproducibility issues plague public papers due to data leakage, undocumented preprocessing, missing seeds, and unshared code Kapoor and Narayanan [2022]. In 2022, a study found 329 papers with errors across 17 fields "leading to overly optimistic conclusions" Kapoor and Narayanan [2022]. The rapid pace of model development also means that systematic reviews can become outdated quickly. PRISMA can bring guidelines and transparency to model selection. However, its effectiveness hinges on the machine learning community establishing standardized data formats, evaluation criteria, and documentation practices.

13 Conclusion

Adopting PRISMA in machine learning isn't about mimicking medicine; it's about recognizing that scientific legitimacy arises from process, not performance. Hugging Face, for all its openness, is still prone to opacity in model evaluation. By aligning our selection criteria with PRISMA's epistemic commitments—clarity, rigor, transparency—we convert intuitive practices into principled ones.
In doing so, we not only choose better models. We build better systems, better research pipelines, and ultimately, better trust in AI.

References

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science. *arXiv preprint arXiv:2207.07048*, 2022. doi: 10.48550/arXiv.2207.07048.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019. URL `https://arxiv.org/pdf/1810.03993`. arXiv:1810.03993.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269, 2009.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer Tetzlaff, Elie A Akl, Megan Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. doi: 10.1136/bmj.n71.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *arXiv preprint arXiv:2003.12206*, 2020. doi: 10.48550/arXiv.2003.12206.

Rens van de Schoot, Joost de Bruin, Rianne Schram, Payam Zahedi, Jolien de Boer, Freek Weijdema, Bianca Kramer, Martijn Huijts, Lars Tummers, and Daniel L Oberski. An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133, 2021. doi: 10.1038/s42256-020-00287-7.