Lecture 4

Resources and Evaluation

2026-04-28

Sergey Ignatenko, PhD

LINKÖPING UNIVERSITY

Resources and Evaluation

- Finding and curating AI resources for research
- Exploring open-source AI community: finding models with Hugging Face
- Designing evaluations that measure what matters
- Evaluation of machine translation: BLEU
- Metrics and methods for evaluating AI systems
- Grounding AI in human preferences
- Evaluation of modern AI workflows (RAG, tools, agents)
- “LLM-as-a-Judge” paradigm
- Reporting, transparency, and reproducibility in AI-supported research

Material is available at: **https://www.itn.liu.se/~siaih22/6fitn80.html**

Finding and curating AI resources for research

Researchers across disciplines increasingly rely on AI resources such as **models, datasets, and analytical tools** to support scientific workflows

- Literature analysis and summarization
- Data classification and pattern detection
- Natural language processing for text corpora
- Image and signal analysis

Key challenge:
selecting appropriate and reliable AI resources for specific research tasks

Finding relevant AI models

Pretrained models that perform tasks such as:

- text summarization
- named entity recognition
- image classification
- speech transcription

Example platform: Hugging Face 

Considerations when selecting models:

- task suitability
- model documentation
- training data transparency
- licensing and usage restrictions

Finding datasets relevant to research questions

Domain-specific datasets:

- public dataset repositories
- institutional research data archives
- government open-data portals
- scientific publications

Important evaluation criteria:

- relevance to research question
- data quality and completeness
- ethical and legal constraints

AI tools supporting research workflows

AI-enabled tools can assist with:

- literature review and document analysis
- data analysis and visualization
- coding and data processing
- experiment documentation

Common frameworks enabling such tools include:

- PyTorch
- TensorFlow
- BertViz

Curating reliable AI resources

Key practices:

- review documentation and model limitations
- verify outputs with independent methods
- track data sources and tool versions
- document how AI tools were used in the research process

This supports transparency, reproducibility, and scientific rigor.

When AI is used as a research tool, the critical task is not only using AI effectively but also selecting, evaluating, and documenting the AI resources that support the research workflow

Exploring open-source AI community: finding models with Hugging Face 🤗

Open platforms accelerate research transparency and experimentation. The most important platform today is Hugging Face.

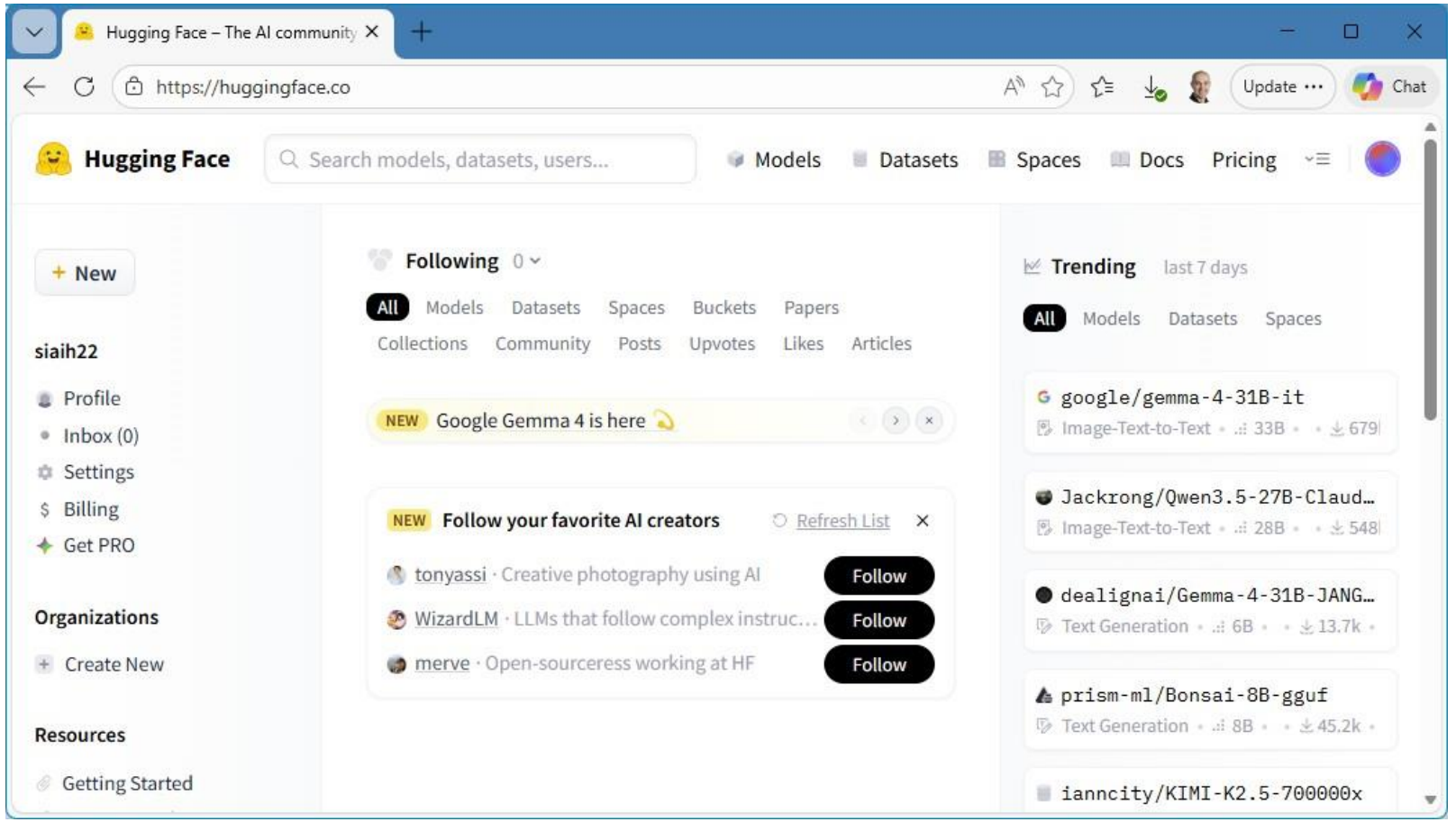
Hugging Face ~ “GitHub of AI”

How to find specialized models
- search by task
- filter by domain
- filter by model properties

Always read the model card

Use spaces (interactive demos) to test before adoption

Evaluate community signals



Open-source platforms such as Hugging Face allow researchers to rapidly discover specialized models, but effective selection requires task-based filtering, careful reading of model cards, and small-scale validation before use.

Designing evaluations that measure what matters

Why evaluation design matters

AI systems used in research must be evaluated with metrics and experimental designs that reflect the real research objective, not only technical performance

Poor evaluation can lead to:
- misleading conclusions
- overestimated system capabilities
- failure in real-world applications

Align evaluation with the research question

Evaluation should measure what the system is intended to accomplish.
- Information retrieval → precision, recall
- Text summarization → factual consistency
- Scientific reasoning → accuracy and explainability

Use multiple evaluation methods. Single metrics rarely capture the full quality of AI outputs

- Automatic metrics (accuracy, BLEU, ROUGE)
- Human evaluation
- Task-based performance tests

Consider dataset bias and benchmark limitations. Benchmarks may not represent real-world conditions

Risks include:
- dataset bias
- narrow task definitions
- models learning benchmark artifacts rather than general capabilities

This can lead to inflated performance claims.

Evaluate robustness and generalization. AI systems should be tested beyond the training distribution

Important evaluation strategies:
- out-of-distribution testing
- adversarial examples
- cross-domain evaluation

This helps reveal failure modes and limitations

Document evaluation procedures

Transparent evaluation requires
- clear reporting of datasets and metrics
- reproducible evaluation pipelines
- description of limitations and uncertainties

Documentation supports scientific reproducibility

Effective evaluation focuses not only on how well a model performs on benchmarks, but on whether the evaluation reflects the real scientific task and its constraints

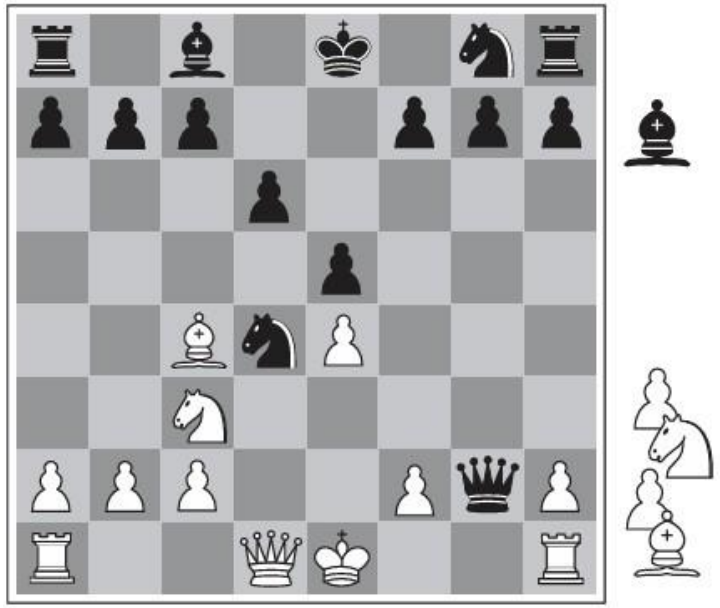
Sourses: AIMA 6.3.1(5.4.1); Bommasani, On the opportunities and risks of foundation models 2108.07258v3

Evaluation functions: example

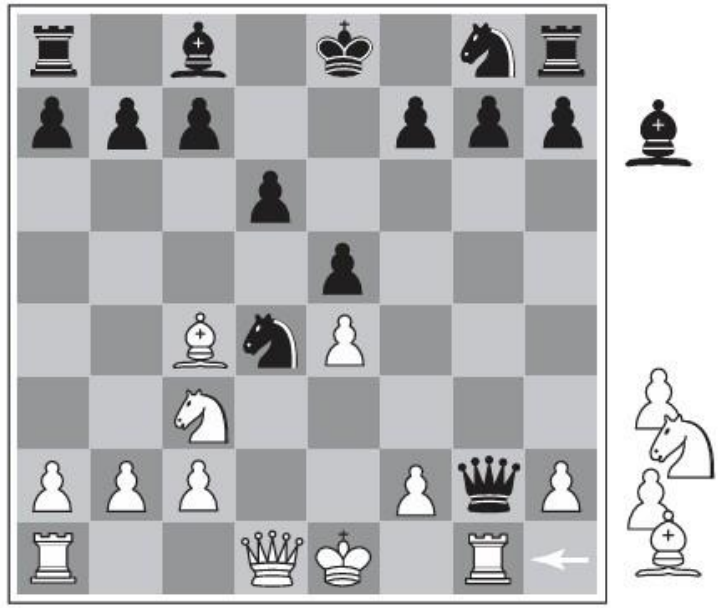
- Should order the terminal states in the true utility function
- Computation must not take too long
- For nonterminal states, the evaluation function should be strongly correlated with the actual chances of winning
- Weighted linear evaluation function

$$\text{EVAL}(s) = \sum_{i=1}^{N} w_i f_i(s)$$

where each w_i is a weight and each f_i is a feature of the position



(a) White to move



(b) White to move

Source: AIMA 6.3.1

Evaluation of machine translation: BLEU

Machine translation systems can generate many valid translations for the same sentence, so evaluation is not as simple as exact matching

BLEU (Bilingual Evaluation Understudy) became the classic automatic metric because it enables:
- fast comparison of models
- large-scale benchmarking
- reproducible evaluation across experiments

Core idea: higher overlap with high-quality references → better score

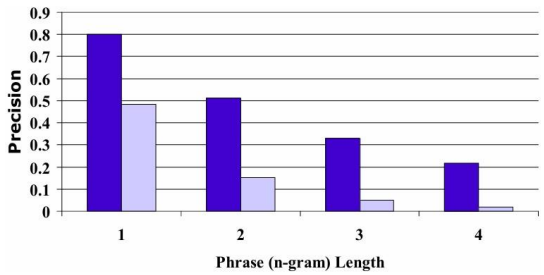
$$\text{BLEU}=\text{BP}\exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

BP – brevity penalty

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1-\frac{r}{c}\right) & \text{if } c \le r \end{cases}$$

c – candidate length
r – reference length

N – n-gram length
$w_n = 1/N$ – weights
p_n – modified precision



Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.
$p_1 = 17/18$
$p_2 = 10/17$

Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.
$p_1 = 8/14$
$p_2 = 1/13$

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

BLEU is a fast and historically important metric for machine translation evaluation, measuring the proximity of machine output to professional human references based on n-gram overlap and brevity penalty

Sources: Papineni etal, BLEU: a method for automatic evaluation of machine translation, ACL (2002)

Metrics and methods for evaluating AI systems

Why evaluation methods matters

Evaluating AI systems requires carefully selected metrics and methodologies to ensure that results reflect true system performance and usefulness for the intended task

Automatic evaluation metrics

Advantages:
- Fast and scalable
- Reproducible
- Suitable for benchmarking large models

Common examples:
- Accuracy, precision, recall
- BLEU and ROUGE for text generation
- F1-score for classification tasks

Limitations:
- may fail to capture semantic meaning, factual correctness, or usefulness

Human evaluation

Typical evaluation dimensions:
- factual accuracy
- coherence and readability
- usefulness for the task
- domain relevance

Advantages:
- captures nuanced judgments that automatic metrics miss

Limitations:
- time-consuming
- expensive
- potential variability between evaluators

Hybrid evaluation approaches

Examples:
- automatic scoring followed by human validation
- models acting as preliminary evaluators, with human oversight

Benefits:
- improved scalability
- better reliability than automatic metrics alone

Emerging approach:
- LLM-as-a-judge evaluation pipelines

Best practices for evaluation
- use multiple complementary metrics
- combine automatic and human evaluation
- report evaluation limitations and uncertainties

No single metric can capture the full quality of AI outputs. Reliable evaluation typically requires a combination of automatic metrics, human judgment, and hybrid methods

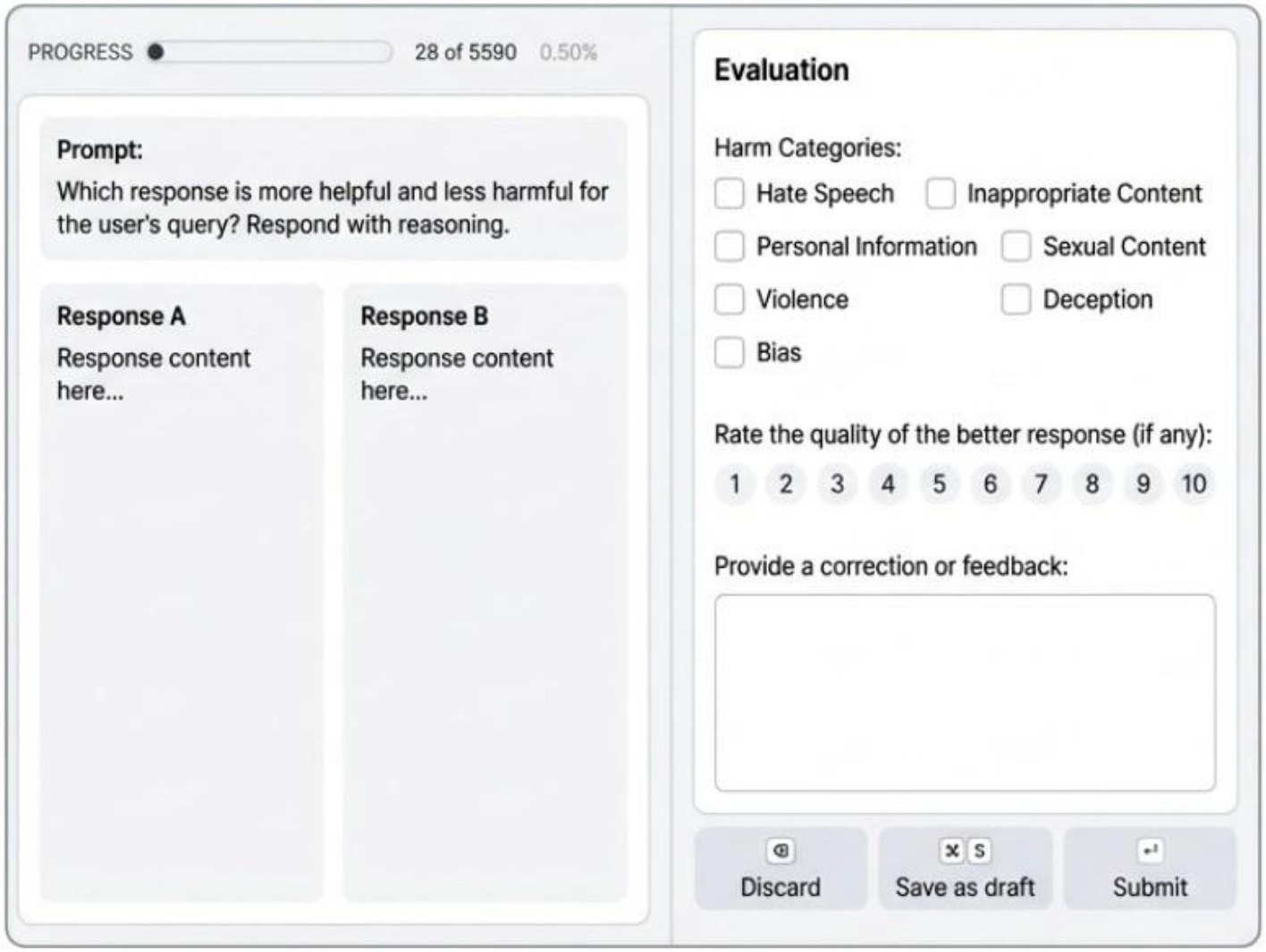
Grounding AI in human preferences

Because quality is subjective, we rely on human feedback to establish baseline truths and train preference models

Absolute scoring: Rating a single response on Likert scale. Vulnerable to human calibration errors.

Pairwise ranking: Presenting two model outputs and asking which is better. Yields highly reliable ratings.

Bottleneck: extremely resource-intensive, requiring diverse, unbiased human annotators to scale.



Evaluation of modern AI workflows (RAG, tools, agents)

Modern AI applications increasingly consist of multi-component workflows rather than standalone models
- RAG
- tool-using AI systems
- autonomous or semi-autonomous agents

Evaluation must therefore consider system-level performance, not only model outputs.

Evaluating RAG

RAG systems combine document retrieval with language model generation
Evaluation should assess both components:
- Retrieval quality
  - precision and recall of retrieved documents
  - relevance of retrieved sources
- Generation quality
  - factual accuracy
  - correct citation of retrieved information
  - consistency with source documents

Failures often arise from retrieval errors or incorrect grounding.

Evaluating tool-using AI systems

Many modern systems integrate external tools such as:
- search engines
- databases
- code execution environments

Evaluation should measure:
- correctness of tool selection
- reliability of tool outputs
- robustness of system orchestration

The focus shifts from single responses to task completion.

Evaluating AI agents

AI agents perform multi-step reasoning and decision-making.
Evaluation must consider:
- task completion success rate
- efficiency and number of steps taken
- robustness under changing conditions
- safety and unintended behaviors

Agent evaluation often requires interactive or simulation-based testing.

System-level evaluation challenges

Key challenges include:
- complex interactions between components
- error propagation across workflow stages
- difficulty defining clear metrics for complex tasks

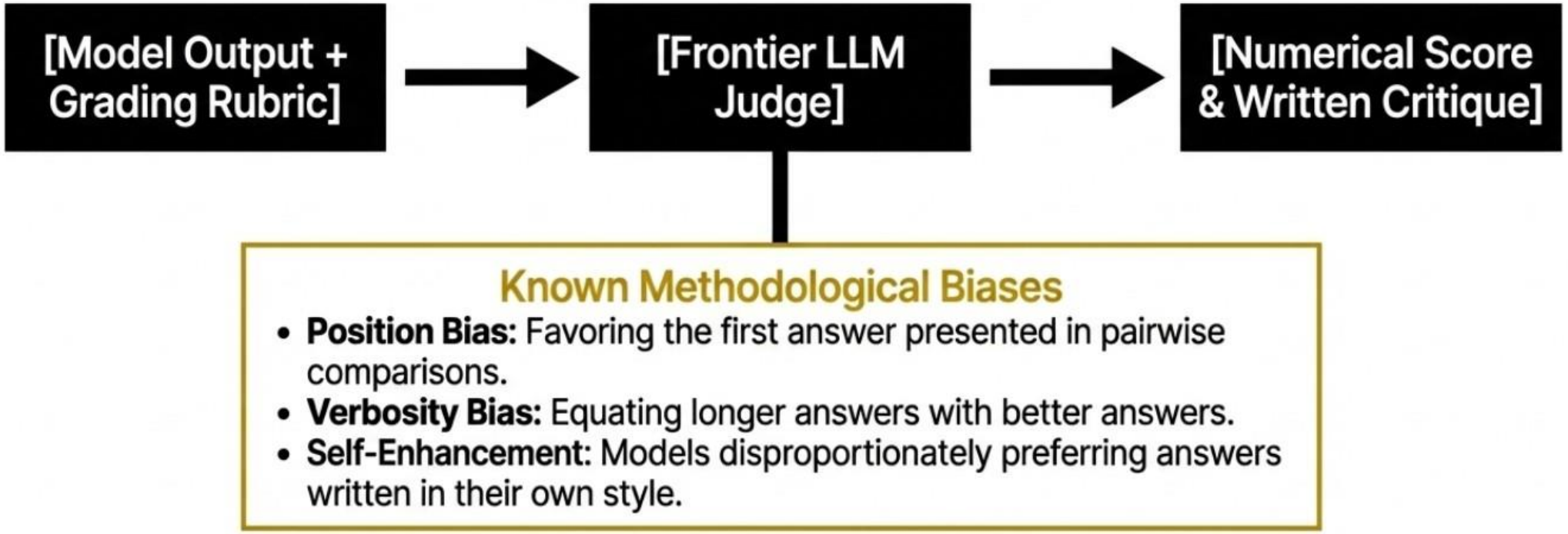
Researchers increasingly rely on task-based and end-to-end evaluations

Evaluating modern AI systems requires moving from model-centric evaluation to workflow-level evaluation, measuring how well the entire system performs complex real-world tasks.

“LLM-as-a-Judge” paradigm

Using frontier models (e.g. GPT-5) to systematically evaluate the outputs of smaller or fine-tuned models based on strict rubric

- Advantages: infinite scalability, consistent application of criteria, rapid iteration



LLM-as-a-Judge paradigm enables scalable evaluation of open-ended AI outputs, but its reliability depends critically on prompt design, calibration, and human validation

Reporting, transparency, and reproducibility in AI-supported research

Scientific research requires that results can be verified, replicated, and critically evaluated by others

When AI tools are used in research, transparency becomes essential to:
- understand how results were produced
- identify potential sources of bias or error
- allow other researchers to reproduce findings

Reporting AI use in research

Researchers should clearly document:
- which AI models or tools were used
- model versions and configurations
- prompts, parameters, or workflows
- datasets used for training or evaluation

Transparent reporting allows others to understand and reproduce the research process

Documentation of models and data

Responsible research includes structured documentation of:
- model capabilities and limitations
- training data characteristics
- intended use cases
- potential risks or biases

Examples of documentation frameworks:
- model cards for machine learning models
- datasheets describing datasets

Ensuring reproducibility

Key practices include:
- sharing code and experimental pipelines
- versioning models and datasets
- documenting preprocessing and evaluation methods
- using reproducible computational environments

These practices support verification and cumulative scientific progress.

Transparency in AI-supported workflows

When AI assists with research tasks (analysis, writing, coding), researchers should:
- disclose the extent of AI assistance
- verify AI-generated outputs
- maintain human oversight and responsibility

This helps maintain scientific integrity and trust.

Transparent reporting and reproducible workflows are essential to ensure that AI-supported research remains verifiable, trustworthy, and scientifically rigorous.

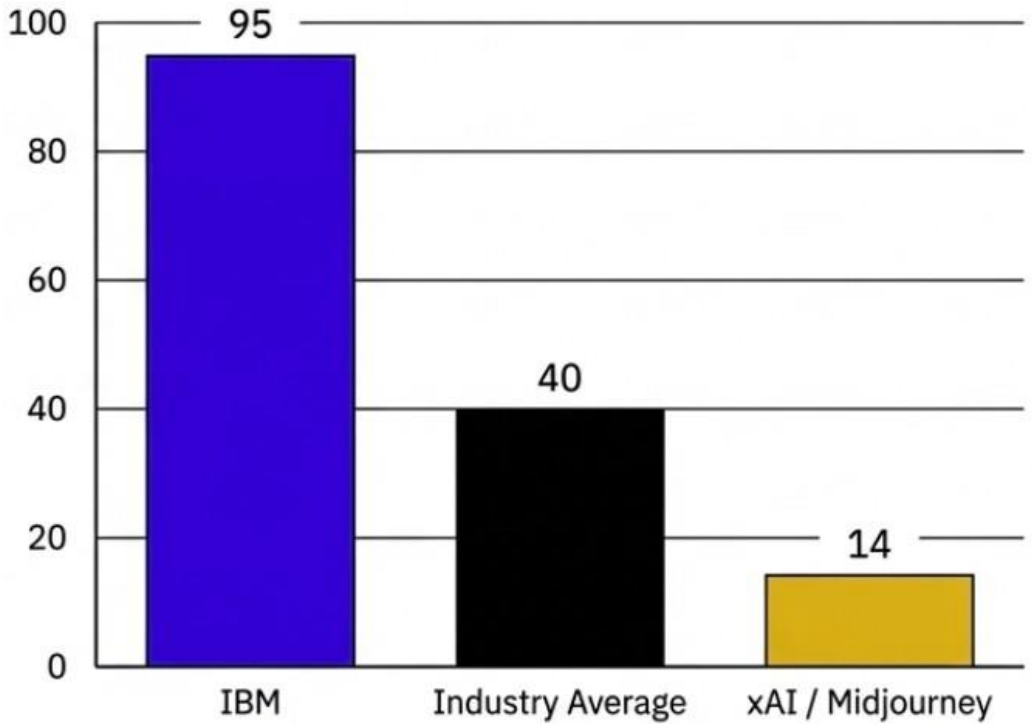
Source: Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary?, Acc. in Research 33(2) (2026)
Model Cards for Model Reporting, arXiv: 1810.03993
Improving Reproducibility in Machine Learning Research, JMLR 22, 1-20 (2021)
COPE position - Authorship and AI

Transparency imperative

"*Foundation model developers are among the world's most important companies… yet transparency is declining*"

2025 Foundation Model Transparency Index

3 domains of transparency
- Upstream: data acquisition, human labor practices, compute allocation
- Model: architecture, exact parameter counts, evaluation reproducibility
- Downstream: usage statistics, port-deployment monitoring, acceptable use policies



Seminar 1
2026-05-04 (Monday) K023 13.15