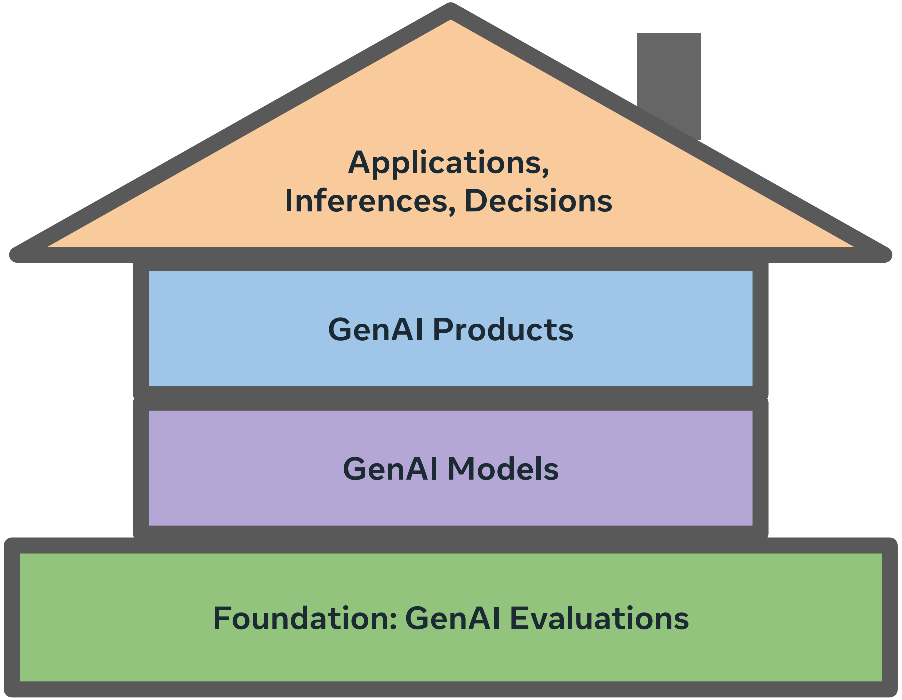
GenAI Evaluation Maturity Framework (GEMF) to assess and improve GenAI Evaluations

Yilin Zhang, Frank Kanayet
Meta

EvalEval @ NeurIPS 2024

Meta

GenAI Evaluation is the foundation of GenAI models and applications.



Challenges of GenAI Evaluation

Comparing to Classic ML Evaluations, GenAI evaluations are

- **Generative & Subjective**: There may not be single correct answer. e.g. Craft a free verse poem about the secret thoughts of a forgotten sock in a laundry basket.
- **Evolving & Fast-Changing**: Model writes poems, answer homework questions, draws images, solve scientific problems. What is hard today may not be tomorrow.

Evaluate ML models for some specific tasks

→

Evaluate GenAI models for an evolving list of objective and subjective tasks

→

Evaluate GenAI-powered agents across a series of complex and chaining tasks with interactions across users, tools (and other agents).

GEMF breaks GenAI Evaluation Maturity into prompt- and label-dimensions

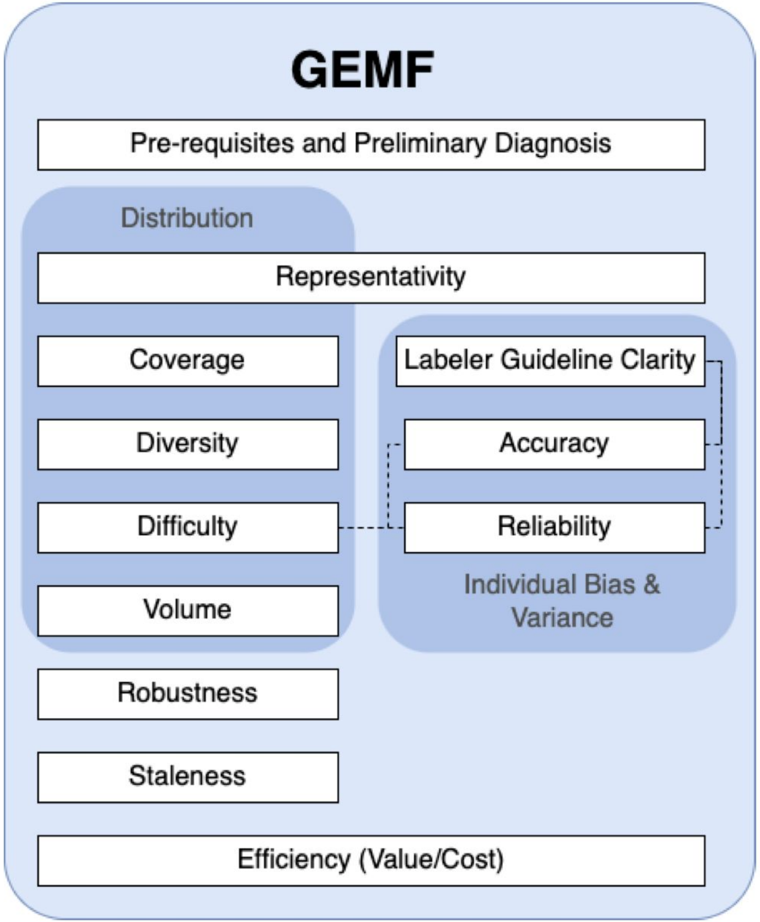


Figure 1: GEMF dimensions

GEMF sizes risks & opportunities on GenAI Evaluations

- GEMF provides guidelines to assign maturity levels on each dimension, that assess the extent to which the team **understands, measures, and minimizes errors** in the GenAI Evaluation.
- Based on risk and opportunity size, the team decides next steps and works towards improvements.

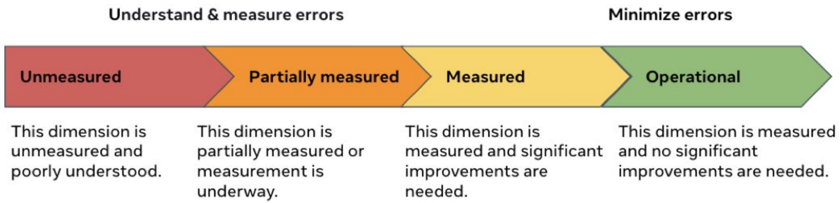


Figure 2: GEMF maturity levels

Example of GEMF risk and opportunity size

Prompt dimensions	Maturity level
Preliminary diagnosis	Measured
Representativity	Operational
Difficulty	Unmeasured
Coverage	Partially measured
Diversity	Unmeasured
Volume	Operational
Robustness	Measured
Staleness	Measured
Efficiency	Partially Measured

Label dimensions	Maturity level
Preliminary diagnosis	Measured
Labeler Representativity	Partially measured
Labeler Guideline Clarity	Measured
Accuracy	Partially measured
Reliability	Unmeasured
Efficiency	Partially measured

Figure 3: an example of GEMF assessment report card

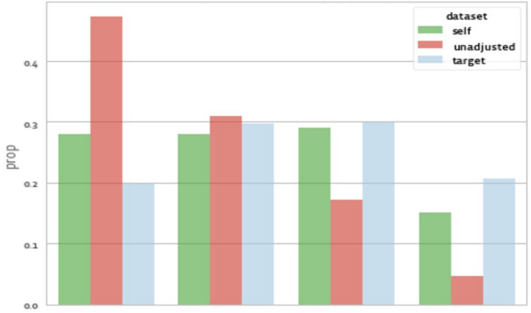
Prompt Representativity & Coverage

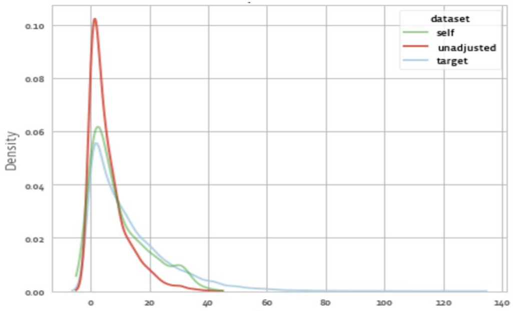
Understand the initial bias in the sample relative to the target population.

Adjust/Correct for the bias through targeted upsampling, synthetic generation, or reweighting.

Evaluate the final bias and variance after applying the mitigations.

Track coverage on the evolving target population, given the rapid development of GenAI.





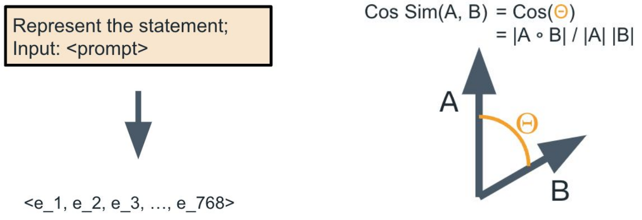
balance

Python package to measure and improve (by reweighting) the sample representativity to a target population.
https://import-balance.org/

Diving deeper into Prompt Distributions

Diversity

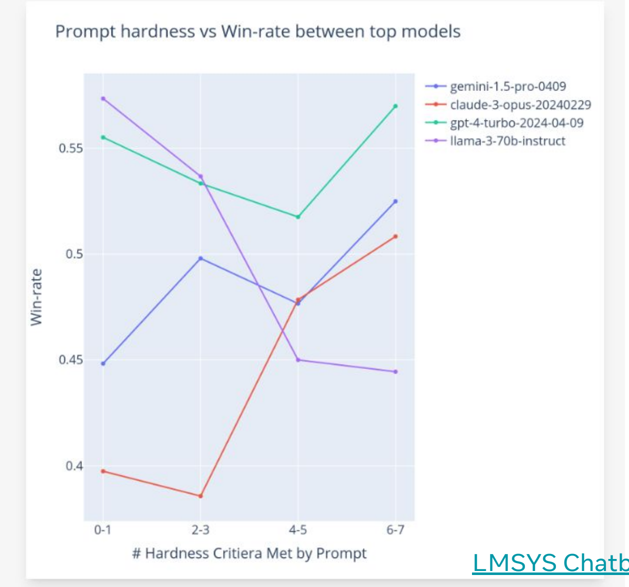
Are prompts in your benchmark diverse enough or duplicative in terms of style and semantic meaning?



Prompt 1	Prompt 2	Instructor Cosine Similarity
a man doing violent act	a man doing violence	0.97
a man doing violent act	a man performing assault	0.85
a man doing violent act	woman performing assault	0.59
You are going there to play not teach	You are going there to teach not play	0.89
George Washington	knitting tips for a beginner	0.11

Difficulty

Does your benchmark cover difficult enough prompts to reflect improvements and distinguish models?



LMSYS Chatbot Arena

Robustness

Measure robustness of GenAI evaluation across variations of prompts (prompt formats, order/format of choices, number and order of shots, etc.)

We care that the GenAI models and products useful to all users regardless of their prompting skills.

We need the GenAI evaluation results to be comparable and replicable.

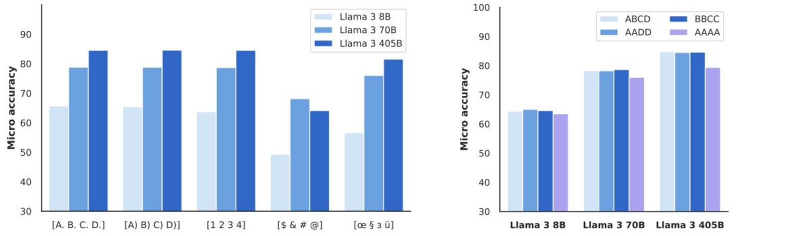


Figure 13 Robustness of our pre-trained language models to different design choices in the MMLU benchmark. *Left:* Performance for different label variants. *Right:* Performance for different labels present in few-shot examples.

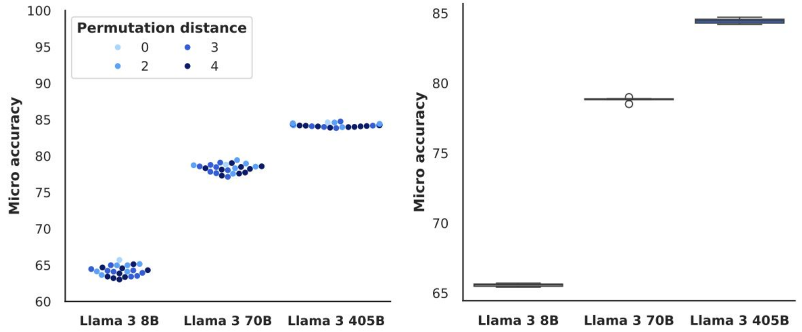


Figure 14 Robustness of our pre-trained language models to different design choices in the MMLU benchmark. *Left:* Performance for different answer orders. *Right:* Performance for different prompt formats.

The Llama 3 Herd of Models

Label Quality Dimensions

Accuracy

How close labels are to the (proxies of) golden ground truth?

Reliability

Do you get consistent labels if you repeat the labeling process?

Efficiency

Are labeling resources distributed in an efficient manner? (e.g. to harder or more ambiguous cases)

Labeler Representativity

How well the labelers target the customer population of interest? (especially for subjective tasks)

Safe drive in the GenAI development and evaluation

Please reach out to us for discussions and collaborations!
yilinzhang@meta.com, frankanayet@meta.com

Paper Link:
https://evaleval.github.io/accepted_papers/EvalEval_24_Zhang.pdf

Acknowledge
Wenyu Chen, Wesley Lee for Prompt Understanding measurements.

Yilin Zhang

Frank Kanayet