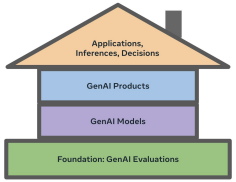
GenAI Evaluation Maturity Framework (GEMF) to assess and improve GenAI Evaluations

Yilin Zhang, Frank Kanayet

Meta

Background and Motivation

GenAI Evaluation is the foundation of GenAI models, products, applications, and decisions.



It is crucial to understand how good your GenAI evaluations are and have high quality enough GenAI evaluations.

Comparing to classic ML evaluations, GenAI evaluations are challenging due to:
- 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).

Label dimensions

Accuracy: How close are the labels to the golden ground truth?

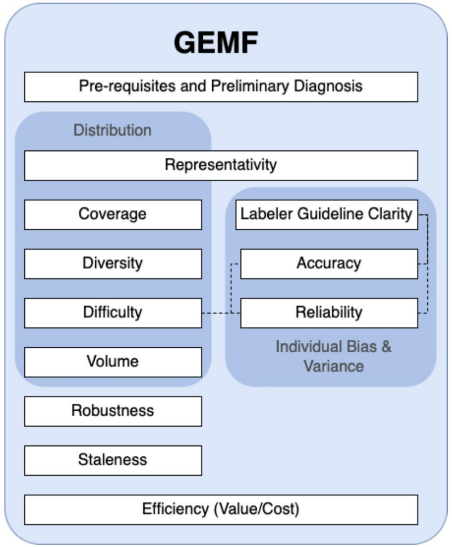
Reliability: How consistent are the labels if repeat the labeling process?

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

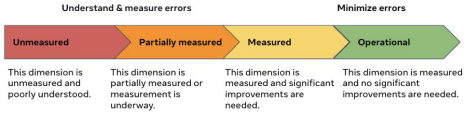
Efficiency: Are labeling resources distributed in an efficient manner? (e.g. to harder or more ambiguous cases)

GEMF dimensions

GEMF assesses the maturity of GenAI Evaluations across Prompt- and Label- dimensions.

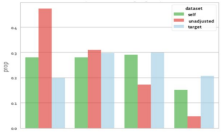


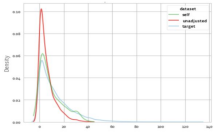
GEMF sizes risks and opportunities across four maturity levels in each dimension.



Prompt dimensions

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.



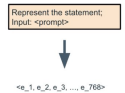


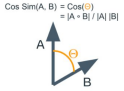
balance

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

Diversity

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

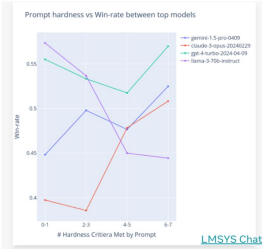




Prompt 1	Prompt 2	Instructor Cosine Similarity
a man doing violent act	a man doing violence	0.97
a man doing violent act	a man performing assault	0.85
a man doing violent act	woman performing assault	0.59
You are going there to play not teach	You are going there to teach not play	0.89
George Washington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

Paper Link: https://evaleval.github.io/accepted_papers/EvalEval_24_Zhang.pdf
Please reach out to us for discussions and collaborations!
yilinzhang@meta.com, frankanayet@meta.com

Acknowledge Wenyu Chen, Wesley Lee for Prompt Understanding measurements.


Yilin Zhang


Frank Kanayet