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

Yilin Zhang
Meta
Menlo Park, CA
yilinzhang@meta.com

Frank Kanayet
Meta
Menlo Park, CA
frankanayet@meta.com

Abstract

We introduce a general framework to assess and improve the maturity of GenAI evaluations, across two Areas: Prompts and Labels, each with multiple dimensions. The GEMF assessment provides a report card with maturity levels across each prompt- and label- dimension, a comprehensive summary on the status of the GenAI evaluation, and suggested directions on where to improve.

1 Introduction

With the rapid growth of Generative AI models (Achiam et al. (2023), Team et al. (2023), Dubey et al. (2024)) and various applications Gozalo-Brizuela and Garrido-Merchán (2023), it is critical to develop solid evaluations that enable the development and assessment of reliable and trustworthy GenAI models and the inferences derived from them. Due to its generative manner and flexible purposes, the evaluation for GenAI models is more challenging than traditional machine learning models Kenthapadi et al. (2024).

We propose the GenAI Evaluation Maturity Framework (GEMF) to assess the maturity of GenAI evaluations. Different from platforms like HELM Liang et al. (2022) and EleutherAI Harness Gao et al. (2024) which evaluate the GenAI models, our framework evaluates the GenAI evaluations. GEMF serves two purposes: 1) To construct a GenAI evaluation, leverage GEMF to understand which dimensions matter for your use case, and use that to guide the selection or construction of benchmark datasets. 2) Given a GenAI evaluation, use GEMF to provide a comprehensive assessment, and identify gaps and opportunities to improve the evaluation. We list out GEMF prompt- and label-dimensions in Section 2, and provide more details on measurement metrics, an application example, and discussion for future work in Appendix.

2 GenAI Evaluation Maturity Framework (GEMF)

2.1 Prompt dimensions for Maturity Assessment

Prompts are input to the LLM model describing the task that LLM should perform. We consider the following dimensions to assess prompt maturity, regardless of sources (online users, existing benchmarks, human annotators, etc.) and applies for both levels (single- or multi- turn) for assessment.

Representativity: In this dimension, we assess how representative the prompt set is to the target population of possible prompts of interest, i.e. how well the prompt set distribution matches the target population distribution across key covariates. Unrepresentative prompt sets for evaluation could lead to biased evaluation metrics and gaps to in-production performance. See Appendix A.2 for methods and tools to measure and improve prompt representativity.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Difficulty: One important category in representativity is the difficulty of the task. From test theory, a good evaluation test should distinguish between different levels of ability of the test takers and a test should have a mix of item-difficulty. Measurement methods are discussed in Appendix A.2.

Coverage: This dimension considers how well the prompt set covers the evolving target population. Given the rapid development of the field, the target capabilities keep evolving as customers and developers find new applications and test the boundaries of GenAI products. Related while different from Representativity, Coverage takes into account the evolvement of the target population. Measurement metrics are included in the Appendix A.2.

Diversity: In this dimension, we consider the extent to which each individual prompt in a segment of interest adds incremental value to the evaluation instead of being duplicative in terms of style and semantic meaning. We can use an embedding-based metric to measure diversity.

Volume: For evaluation data, we assess whether we have enough prompts in each segment to derive precise aggregated metrics with confidence intervals narrow enough to make a decision.

Robustness: This dimension assesses the robustness of the GenAI evaluation across variations in prompting techniques like chain of thought or number of shots. Model assessment should report impact of these variations on performance.

Staleness: In this dimension, we consider whether the cadence of your prompt data collection is regular enough with period relevance given the rate of underlying population change and whether there are processes in place to identify when prompts need to be refreshed in terms of feasibility of cadence and data "shelf life". Staleness also considers the problem of "saturation" which happens when models perform too well on a test to be informative either because the test is part of the training data or because the model has improved significantly.

Efficiency: In this dimension, we consider if the cost of the prompt is providing value as efficiently as possible. Methods to measure and improve prompt efficiency are discussed in Appendix A.2.

2.2 Label dimensions for Maturity Assessment

Labels are decisions on how good LLM responses to the given prompt are. We consider the following dimensions to assess label maturity, regardless of sources (human annotators, online user feedback, model-based judges, heuristics, existing benchmarks, etc.).

Labeler Representativity: Different from prompt representativity, this dimension assesses how well labelers target the customer population of interest. Different sources and backgrounds of labelers could assess response quality differently, especially for more subjective tasks. This dimension does not apply to objective or factual question-answering use cases.

Reliability: This dimensions assesses the consistency of the label if you repeat the label generation process. Inconsistency could come from both the multi-reviewing process by human/model annotators, and LLM's own sampling uncertainty in response generation. We want to control inconsistency from multi-reviewing, but inconsistency from LLM response is usually not a bad thing given that creativity is one of the values of LLMs. This dimension interacts with labeler representativity. If labelers are diverse and the task is subjective, we should expect (and want) some label inconsistency determined by personal preference. For evaluations where we want to be consistently correct, like factuality, reliable measures are needed. Measurement metrics are discussed in Appendix A.3.

Accuracy: In this dimension, we assess the bias of the label. Measurement methods (for objective or subjective scenarios, when there are or no golden ground truth labels) are discussed in Appendix A.3.

Label Guideline Quality: Labeling guideline has been a key part that affects the reliability and accuracy of labels. Guidelines should clearly articulate the criteria to label (e.g. helpfulness, harmlessness, honesty), sub-dimensions to consider for each criteria (e.g. conciseness, coherence, relevancy for helpfulness), detailed task context (e.g. responses will be used to write college essay Vs. to learn on my phone about a topic I saw on social media) and instructions (and examples) to make labeling/scoring decisions. See Appendix A.3 for measurement methods.

Efficiency: In this dimension, we consider whether the cost of the label is providing value as efficiently as possible. Note that this is separate from prompt efficiency as the generation of both prompt and label can be costly. See Appendix A.3 for measurement metrics.

References

2023. Using GPT-4 for content moderation. https://openai.com/index/using-gpt-4-for-content-moderation/

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732* (2021).

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Youssef Benchekroun, Megi Dervishi, Mark Ibrahim, Jean-Baptiste Gaya, Xavier Martinet, Grégoire Mialon, Thomas Scialom, Emmanuel Dupoux, Dieuwke Hupkes, and Pascal Vincent. 2023. Worldsense: A synthetic benchmark for grounded reasoning in large language models. *arXiv preprint arXiv:2311.15930* (2023).

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 7432–7439.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036* (2018).

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161* (2019).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* (2024), 1–79.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.12608602

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. 2024. Are We Done with MMLU? *arXiv preprint arXiv:2406.04127* (2024).

Roberto Gozalo-Brizuela and Eduardo C Garrido-Merchán. 2023. A survey of Generative AI Applications. *arXiv preprint arXiv:2306.02781* (2023).

Robert M Groves and Lars Lyberg. 2010. Total survey error: Past, present, and future. *Public opinion quarterly* 74, 5 (2010), 849–879.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* (2021).

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).

Maurice George Kendall. 1948. Rank correlation methods. (1948).

Krishnaram Kenthapadi, Mehrnoosh Sameki, and Ankur Taly. 2024. Grounding and Evaluation for Large Language Models: Practical Challenges and Lessons Learned (Survey). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6523–6533.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664* (2023).

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

Don McNicol. 2005. *A primer of signal detection theory*. Psychology Press.

Viet-An Nguyen, Peibei Shi, Jagdish Ramakrishnan, Udi Weinsberg, Henry C Lin, Steve Metz, Neil Chandra, Jane Jing, and Dimitris Kalimeris. 2020. CLARA: confidence of labels and raters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2542–2552.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. 2023. Epistemic neural networks. *Advances in Neural Information Processing Systems* 36 (2023), 2795–2823.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728* (2019).

Tal Sarig, Tal Galili, and Steve Mandala. 2022. Balance: a python package for balancing biased data samples. `https://import-balance.org/`

Burr Settles. 2009. Active learning literature survey. (2009).

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937* (2018).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

Yilin Zhang, Aude Hofleitner, Hannah Furnas, Paul Chung, Wesley Lee, Xu Chen, and Ben Fifield. 2022. Introducing the Ground Truth Maturity Framework for assessing and improving ground truth data quality. `https://tinyurl.com/2cnhb7jz`

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* (2023).

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2024).

A More details on GEMF dimensions

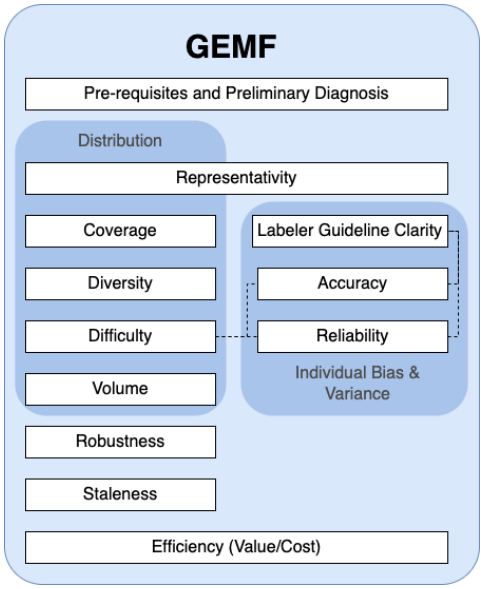

Figure 1: GEMF dimensions

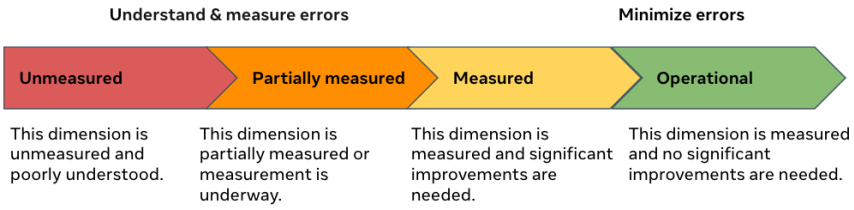

Figure 2: GEMF maturity levels

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

GEMF extends from the Ground Truth Maturity Framework (GTMF) for assessing and improving label quality for machine learning models. Zhang et al. (2022).

A.1 Pre-requisites and Preliminary Diagnosis

Pre-requisites: Before understanding the maturity of the GenAI evaluation, it is important to clearly articulate the goal (e.g. guide launch decisions or model performance improvement), scenario (e.g. helpfulness or safety), models or products (e.g. Llama3.1 Dubey et al. (2024) or some chatbot empowered by Llama3.1), for the GenAI evaluation. This is important because the prompts and labels should be measuring that theoretical construct as closely as possible, ensuring "construct validity" in the language of the Total Survey Error framework Groves and Lyberg (2010). Meanwhile, teams shall be mindful of the discrepancy between what we try to measure and what we can measure, which is more of an issue in GenAI compared to traditional ML evaluation.

Preliminary Diagnosis: Once definitions have been explicitly stated and aligned on, in this step, teams shall identify whether there is anything obviously wrong with the dataset or its collection process. This is a chance to take stock of what exactly your goals are while understanding potential shortcomings that may not be fully captured in the later steps. Much of the preliminary diagnosis step is cataloging existing understand work or past major events in the prompt and label generation/collection lifecycle. Questions on **Pipeline Error** are a chance to catalogue any known issues or biases in the data pipeline, or past major pipeline error events that may or may not be fully resolved. Similarly, drawing on answers in the definition stage, questions involving **Definition Differences** are an opportunity to explicitly state how the observed ground truth labels may or may not fully capture the target construct of interest. Besides that, questions involving **Contamination** considers whether there's any circularity between the training or finetuning or RLHF data and the evaluation data which leads to biased model performance.

A.2 More details on Prompt Dimensions

Prompt Representativity *How to measure and improve:*

- **Define the target population**. The target population could be the online usage of users or prompt set in external benchmarks, depending on the use case of the evaluation (to indicate online engagement or for chatbot Arena).
- **Define the taxonomy** with the key covariates of the target population for which it's important to achieve balance between the sample and the target population.
- **Measure and improve (by reweighting)** the prompt representativity through *balance* - a Python package for balancing biased data samples Sarig et al. (2022).

Difficulty *How to measure:* We can measure the difficulty for LLM and labelers through metrics like LLM uncertainty (e.g. with semantic sets Kuhn et al. (2023)) and rater consistency Nguyen et al. (2020), or accuracy metrics Gallegos et al. (2024), Kadavath et al. (2022). We can also use methods from Item Response Theory Baker (2001) such as item information plots, or the difficulty and discrimination parameters to characterize the extent to which prompt sets and evaluation suites capture the right difficulty levels and allow us to discriminate great models and responses from the median.

Coverage *How to measure:* After aligning on a comprehensive ideal taxonomy of capabilities for a given model, we can calculate the following metrics to measure the current state of coverage and improvement from the last assessment. Current state: We calculate coverage as the percentage of capabilities that we currently measure with evaluations that are at least "measured" in the framework:

$$\text{Coverage}(\text{taxonomy}, \text{population}) = \frac{\#\text{ measured segments }(\text{taxonomy}, \text{population})}{\#\text{ segments }(\text{taxonomy}, \text{population})}$$

Improvement from last assessment: We calculate improvement on coverage as difference on coverage metric from the last assessment, under the current population and taxonomy.

$$\Delta\text{Coverage} = \text{Coverage}_{\text{current}}(\text{taxonomy}, \text{population}) - \text{Coverage}_{\text{last}}(\text{taxonomy}, \text{population})$$

Though within the "measured" maturity level, measurement could be further improved through more granularity or more advanced metrics. We take into account the improvement of measurement

through the percentage of capabilities that we improved on the measurement.

$$\frac{\#\,\text{segments with improved measurement}(\text{taxonomy}, \text{population})}{\#\,\text{total segments}(\text{taxonomy}, \text{population})}$$

Prompt Efficiency *How to measure and improve:* Measurement for Cost per prompt includes metrics e.g. the average handling time per prompt and average cost per prompt. To save cost on prompt generation, we could leverage LLM to simulate/generate and sample and combine from external auto benchmark prompt sets. We consider Value per prompt in terms of 1) increasing the design effect in GenAI evaluation, 2) improving the other dimensions for prompt e.g. diversity, or 3) improving model performance for GenAI pre-training/finetuning. This type of value could be improved through active learning Settles (2009) and RAG methods Gao et al. (2023).

A.3 More details on Label Dimensions

Reliability *How to measure:* For the reliability of multi-review labeling, we can measure using metrics including: Kendall's τ Kendall (1948) to evaluate the reliability of ranking for information retrieval. Krippendorff's α Krippendorff (2011) to perform detailed analyses of reliability by taking agreement by chance into account. We can improve through dynamic multi-review Nguyen et al. (2020). For LLM consistency, if needed, could be controlled by LLM temperature and top-p parameters and measured by logit-based, verbalized-based, multi-review-based method (Sec 5.4.3 in Zhang et al. (2023)), or epistemic type of method Osband et al. (2023).

Accuracy *How to measure:* On factuality type of tasks where there's golden ground truth, accuracy of labels can be measured with standard methodologies and metrics such as F1 scores for categorical labels and root-mean squared error (RMSE) for continuous labels. We can also use Signal Detection Theory McNicol (2005).

Without a golden set, teams will likely need to look at the construct validity of ground truth labels or other related proxies for label accuracy. Below are several types of validity checks, which can provide signal on the accuracy of the ground truth data when there is no golden set.

- Convergent validity tests whether the constructs that theoretically should be related are actually related.
- Discriminant validity tests whether the measures that are not supposed to be related are actually unrelated.
- Predictive validity tests whether the labels are predictive of things that it conceptually should be predictive of.

Label Guideline Quality *How to measure:*

- Develop a checklist of elements that a good labeling guideline should have.
- Check if there are mechanisms to collect feedback and questions from labelers and to improve guidelines based on the feedback.
- Develop a measure of labeler adherence to guidelines by identifying the set of critical guideline elements and check if annotators correctly adhere to the elements on a sample of prompts.
- We can use an approach similar to con (2023) with an LLM to label a golden set following the same guidelines and using the number of LLM mistakes (and reasons) as a proxy for guideline lack of clarity. This method depends on having confidence in the quality of a golden set as ground truth.

Label Efficiency *How to measure:* Measurement for Cost per prompt includes metrics e.g. the average handling time per label and average cost per label. Conversion rate (proportion of usable/converted labels) and decision rate (in multi-review, how many labelers needed for a final decision) should also be considered when measuring label cost. To save cost on label generation, we could leverage LLM to automate label generation Zheng et al. (2024) and guideline generation. We consider Value per label in terms of affecting the evaluation result if we flip the value of the label, where we should flag if several influential labels are able to significantly affect the evaluation result, or improving the model performance for GenAI pre-training/finetuning. We could leverage importance sampling to more effectively optimize for the value increased from the same cost.

B GEMF assessment on a suite of benchmarks

GenAI Evaluation dataset: A suite of benchmarks with MMLU Hendrycks et al. (2020), DROP Dua et al. (2019), WorldSense Benchekroun et al. (2023), SQUAD Rajpurkar et al. (2016), CommonSenseQA Talmor et al. (2018), Social IQA Sap et al. (2019), PIQA Bisk et al. (2020), QuAC Choi et al. (2018), GSM8K Cobbe et al. (2021), Math Hendrycks et al. (2021), MBPP Austin et al. (2021), etc.

Summary of GEMF assessment on this suite of benchmarks as the GenAI evaluation dataset: Overall, a large majority of this benchmark suite has maturity between unmeasured and partially measured, indicating a lot of opportunities for improvement, and risks that should be aware of when making decisions using the evaluation results derived from this benchmark suite.

B.1 Top recommendations:

Prompt Coverage & Representativity

- **Maturity Level:** Partially Measured
- **Risks:** Academic benchmarks tend to nominally measure a given construct or capability but the definition of the benchmark might not match the definitions or needs of a given model or product developer so relying on benchmarks as the main source of evaluations will be risky. If a taxonomy of ideal capabilities exist, it is critical to understand (1) how the benchmarks map to the capabilities in the taxonomy, and (2) how to measure and improve on coverage and balance across capabilities represented by the benchmark suite.
- **Recommendations:** Shift evaluation focus from benchmark-driven to capability-driven. Prioritize aligning on a comprehensive taxonomy of capabilities, and measuring coverage and representativity by mapping prompts from all benchmarks into the capability taxonomy. Merge capability taxonomies into a centralized place.

Prompt Difficulty

- **Maturity Level:** Unmeasured
- **Risks:** Measuring difficulty of evaluations is essential to properly understand where each model stands against top competitors, especially in frontier capabilities. Currently, difficulty of benchmarks and suites is unmeasured except for some high level description in some benchmark papers.
- **Recommendations:** Develop and leverage difficulty measurement methods to establish the definition and enable the measurement. Define which evaluations create better separability between top performing models or between early and late checkpoints.

Label Accuracy and Reliability

- **Maturity Level:** Partially Measured
- **Risks:** Though some benchmarks shared their label verification process (multi-reviewing, expert-label verification, or a deterministic process), errors were found in benchmark labels e.g. MMLU Gema et al. (2024).
- **Recommendations:** Develop a standard way of reporting label verification process in academic benchmarks and build mechanisms to audit the quality of labels and curate prompts and labels from large scale datasets.

Efficiency

- **Maturity Level:** Unmeasured
- **Risks:**
  - There's significant operational cost to discover and select among a huge number of benchmark data sets. There lacks of a standardized way for prompt understanding and characterization across benchmarks.

  - Value of prompt and labels is unclear: There lacks of understanding on the ROI of the prompts/benchmarks for a strategic sampling/selection/combination among benchmarks/prompts. There lacks of on the impact of label errors in the evaluation metrics.

- **Recommendations:** Leverage centralized platforms and automated tools to scale efforts. Understand the ROI of prompts and benchmarks. Measure impact of label values on model performance.

C Future Work

We are developing tools to automate and scale GEMF assessment across benchmarks. We are also working on defining and operationalizing metrics to standardize the assessment. For example, defining more clearly how the dimensions should be measured in specific context like pre-training Vs. post-training.