Position: Bayesian Statistics Facilitates Stakeholder Participation in Evaluation of Generative AI

Yanan Long
ylong@uchicago.edu
yanan.long439@gmail.com
University of Chicago & Queer in AI

Abstract

The evaluation of Generative AI (GenAI) systems plays a critical role in public policy and decision-making, yet existing methods are often limited by reliance on benchmark-driven, point-estimate comparisons that fail to capture uncertainty and broader societal impacts. This paper argues for the use of Bayesian statistics as a principled framework to address these challenges. Bayesian methods enable the integration of domain expertise through prior elicitation, allow for continuous learning from new data, and provide robust uncertainty quantification via posterior inference. We demonstrate how Bayesian inference can be applied to GenAI evaluation, particularly in incorporating stakeholder perspectives to enhance fairness, transparency, and reliability. Furthermore, we discuss Bayesian workflows as an iterative process for model validation and refinement, ensuring robust assessments of GenAI systems in dynamic, real-world contexts.

ACM Reference Format:
Yanan Long. 2018. Position: Bayesian Statistics Facilitates Stakeholder Participation in Evaluation of Generative AI. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The evaluation of generative AI (GenAI) systems is crucial for public policy decision-making in contexts in which they are deployed [8]. However, exisiting methods suffer from important shortcomings. For example, NLP evaluations have been shown to be susceptible to a "benchmark culture" that prioritizes quantitative improvements in certain curated datasets over broader societal goals of language understanding, especially in real-worreal-world applicationstions ("in the wilthat generally span[6, §4.2]. Such benchmarking usually takes of the form of showing that a new system outperforms previous state-of-the-art (SOTA) on a set of chosen tasks (approximately) representing some functionalities, where the judgements are made by comparing the performances of the systems in terms of *point estimates*, occasionally also including assessment of confidence or uncertain quantification (UQ), but this is not always the case. In other words, in order to know whether a new system really is superior to pre-existing ones, we need to be able to able to draw robust and rigorous inferences. Finally, we close by discussing challenges of the proposed framework.

However, in the case of GenAI systems, this is challenging on several fronts. First, GenAI systems are inherently random and high-dimensional, which prevents traditional statistical methods from adequately being applied for robust inference. Another issue is that GenAI systems are computationally expensive for both training and prediction (also called "inference" in the GenAI literature) [1, §3]. As a result, GenAI evaluations may have to rely on a limited amount of data [12, §3.6]. Finally, there is also the problem of measurement, where researchers have to deal with the gap between observable phenomena and theoretical constructs — the former are tied to real world consequences of GenAI systems, affecting stakeholders who likely hold domain expertise, whereas the latter can often be obscure and only accessible to trained AI modellers [12]. So far, stakeholder participation has been undervalued in AI system development, leading to various kinds of harm [10].

In this work, we argue that Bayesian statistics is well-poised to tackle the above problems since it offers a *unified* workflow which (1) takes into account *context-dependent prior information* or *beliefs*, allowing the principled inclusion of domain expertise; (2) adapts *continuously* to new data; and (3) by providing posterior inference, provides a robust way of uncertainty quantification.

2 Bayesian Statistics

2.1 Bayesian Inference as Generative Modelling

In the most basic form, Bayesian statistics refers to the use of Bayes theorem — given *parameters* $\boldsymbol{\theta}$ and data y, Bayes theorem statest that

$$p(\boldsymbol{\theta}|\mathrm{y}) \propto p(\mathrm{y}|\,\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{1}$$

where $p(\boldsymbol{\theta})$ is called the *prior*, $p(\mathrm{y}|\,\boldsymbol{\theta})$ the *likelihood*, and $p(\mathrm{y}|\,\boldsymbol{\theta})$ the *posterior*. Note that in Eq. (1) the normalization constant has been omitted because it can be difficult to calculate and not strictly necessary in practice. Moreover, $\boldsymbol{\theta}$ designates all the parameters collectively may have a highly complex *hierarchical* structures, reflecting the modeler's belief in the interdependencies between the individual parameters. Employing a hierarchical structures is usually used to model partial pooling.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

As a pedagogical example, consider the following model specified by:

$$\mathrm{y}_{ij} \mid \theta_i, \sigma^2 \sim \mathcal{N}(\theta_i, \sigma^2) \tag{2a}$$

$$\theta_i \mid \mu, \tau^2 \sim \mathcal{N}(\mu, \tau^2) \tag{2b}$$

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{2c}$$

$$\tau^2 \sim \text{Inverse-Gamma}(\alpha, \beta) \tag{2d}$$

$$\sigma^2 \sim \text{Inverse-Gamma}(\gamma, \delta), \tag{2e}$$

where $\sim$ means that the random variable on the left-hand side follows the (conditional) probability distribution given on the right-hand side. The joint posterior is thus given by

$$p\left(\{\theta_i\}, \mu, \tau^2, \sigma^2 \mid \{\mathrm{y}_{ij}\}\right) \propto \left[\prod_{i=1}^{M}\prod_{j=1}^{N_i} p(\mathrm{y}_{ij} \mid \theta_i, \sigma^2)\right] \times \left[\prod_{i=1}^{M} p(\theta_i \mid \mu, \tau^2)\right] \times p(\mu) \times p(\tau^2) \times p(\sigma^2). \tag{3}$$

In the above example, the data $\mathbf{y}$ are the observed data $\{\mathrm{y}_{ij}\}$ grouped by the indices i and j, and the parameters (modeller-specified) are $\boldsymbol{\theta} = \{\sigma_0, \mu_0, \alpha, \beta, \gamma, \delta\}$ whereas the *hyperparameters* (not modeller-specified) are $\{\mu, \tau^2, \sigma^2, \{\theta_i\}\}$. This model can also be succinctly and intuitively represented by a graphical model using the *plate notation*, where a shaded node represents observed data, an unshaded node a hyperparameter, and a bare node a parameter (Fig. 1). Parameter inference is then performed by drawing (generating) samples from the joint posterior distribution, and as such, Bayesian inference is known as "generative modelling". In this way, we obtain not only the point estimate or the confidence interval of a parameter, as is the case with the classical statistical method, but also *simulations* from the posterior, enabling graphical assessment of the *distribution* of the parameters of interest and hence improved UQ. Note that our discussion on UQ is quite general and without any distributional assumption, extending the treatment given by [7].

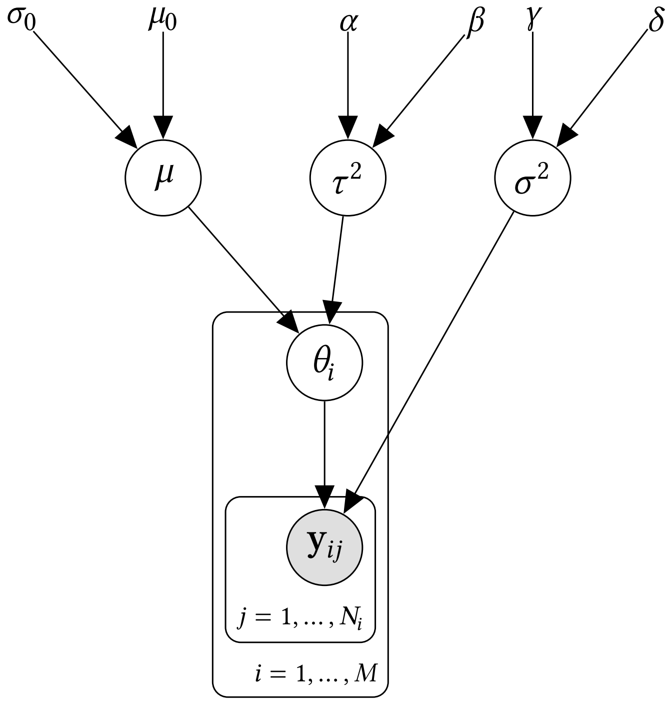

Figure 1: Plate notation for the hierarchical model given by Eq. (2). A shaded node represents observed data, an unshaded node a hyperparameter, and a bare node a parameter. Nodes contained within a square are repeated as indicated.

Before proceeding further, we note that the actual data-generating processes of typical GenAI systems cannot be expressed in simple close-formed equations but are rather to be treated as black-boxes. Even in cases where the model weights have been publicly released, they do not carry interpretable meanings comparable to parameters (e.g. regression coefficients) in typical theory-driven scientific models.

2.2 Prior Elicitation with Stakeholders

Bayesian inference naturally incorporates prior beliefs in the form of specification of the prior distribution $p(\boldsymbol{\theta})$, which can be parametric as in the example shown in the previous subsection, or *non-parametric*, such as the Gaussian process for continuous hyperparameters or the Dirichlet process for discrete hyperparameters. Such prior beliefs should ideally be based on best available information at the time of modelling. It is also important to note that the choice of the prior should generally be made in connection with the likelihood, which is also known as the *data prior* [3]. In the case of GenAI evaluation, one method is to consult important stakeholders such as minoritized communities particularly affected by GenAI systems [2, 10]. Drawing on standpoint epistemology, we contend that these communities are the relevant *domain experts* whose *lived experiences* ought to be accentuated for robust GenAI evaluation [13, 14]. The process of consulting domain experts to inform prior choices is called *prior elicitation* [9]. Broadly speaking, prior elicitation can be performed on either the *parameter space* ($\mathbf{y}_P$) or the *observable space* $\mathbf{y}_O$, or both, and is treated as Bayesian inference itself, reflecting the belief of the modeller about the domain expertise of the recruited stakeholders [9]:

$$p(\boldsymbol{\theta}|\mathbf{y}_P, \mathbf{y}_O) \propto p(\mathbf{y}_P|\boldsymbol{\theta})p(\mathbf{y}_O|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{4}$$

For GenAI systems, their black-box nature means that the focus is generally on the the observable space. To illustrate the process, consider bias evaluation in large language models (LLMs) and vision language models (VLMs) (e.g. [5, 11]). This typically involves first designing a set of *prompts* to be fed into the LLMs/VLMs, whose outputs (texts, images etc.) are then inspected by annotators who assign values to specific variables that are designed to map to concrete system behaviours. Such values may be binary (e.g. whether the output should be classified as harmful), categorical (e.g. whether the output affects given identitiy groups), ordinal (a Likert scale of a qualitative measure, e.g. satisfaction with model output) or continuous (e.g. a normalized score ranging from 0 to 100 or a probability ranging from 0 to 1). In any case, the prompts are *part of* the parameters i.e. $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_{\text{Prompt}}, \boldsymbol{\theta}_M\right]$, where $\boldsymbol{\theta}_M$ stands for the model parameters (e.g. temperature, top-k/top-p sampling) for the LLMs/VLMs that may itself be drawn from some other prior distribution. We may rewrite the posterior as

$$p(\boldsymbol{\theta}|\mathbf{y}_O, \Phi) \propto p(\mathbf{y}_O|\, \boldsymbol{\theta}_{\text{Prompt}}, \boldsymbol{\theta}_M, \Phi) p(\boldsymbol{\theta}_M|\Phi)\, p(\Phi) \tag{5}$$

where $p(\Phi)$ is the empirical distribution obtained by querying/sampling the LLMs/VLMs.

2.3 Bayesian Workflow

So far we have introduced *Bayesian inference*, which is to be distinguished from the broader process of *Bayesian workflow* of which it is a part [4]. In the workflow, there are the additional steps of model building and model validation, which is carried out iteratively (cf. Fig. 2). Bayesian inference refers to the step **fit model**, usually by means of Markov Chain Monte Carlo (MCMC). After obtaning the posterior draws, we can test if they are valid (i.e. the chains have mixed well), and adjust the model accordingly. After that, we can produce sample from the posterior predictive distribution to simulate new data and inspect their plausibility. To continue with the LLM/VLM bias evaluation above, we have:

$$p(\mathbf{y}'|\Phi) = \int p(\mathbf{y}'|\, \boldsymbol{\theta}_{\text{Prompt}}, \boldsymbol{\theta}_M, \Phi) p(\boldsymbol{\theta}|\mathbf{y}_O, \Phi)\, d\boldsymbol{\theta}, \tag{6}$$

where $\mathbf{y}'$ stands for new (observed or simulated) data. The Bayesian workflow is also well-suited for the situation where the same set of evaluations is repeated across different contexts or settings (e.g. geographic locations, GenAI versions, etc.) or over time (e.g. evaluation is done periodically). In the former case, the output variable can be grouped by the locations, version and other reasonable grouping structures, resulting in a hierarchical model, and parameter estimation can give insights into how the behaviours of the GenAI system can be modulated by those variables. In the latter case, the posterior obtained at each round of evaluation will become the new prior for the next round, and the process can continue for multiple rounds until some stopping criterion is met.

3 Discussion

Integrating the Bayesian workflow into the evaluation of GenAI systems offers a comprehensive framework that addresses the limitations of traditional evaluation methods. By incorporating prior knowledge and continuously updating beliefs with new data, Bayesian approaches provide robust uncertainty quantification, enhancing the reliability of GenAI assessments. This iterative process aligns with the dynamic nature of real-world applications, allowing for adaptive evaluations that remain relevant over time.

By way of conclusion, we now discuss some challenges of the proposed framework. Since the GenAI system is treated as a blackbox, we cannot have direct access to the GenAI prior $p(\Phi)$ and must rely on simulations. As a result, we may have to make (possibly) strong assumptions about the relationship between the output and the GenAI system. Moreover, our framework will require larger numbers of model queries than typical benchmarking.

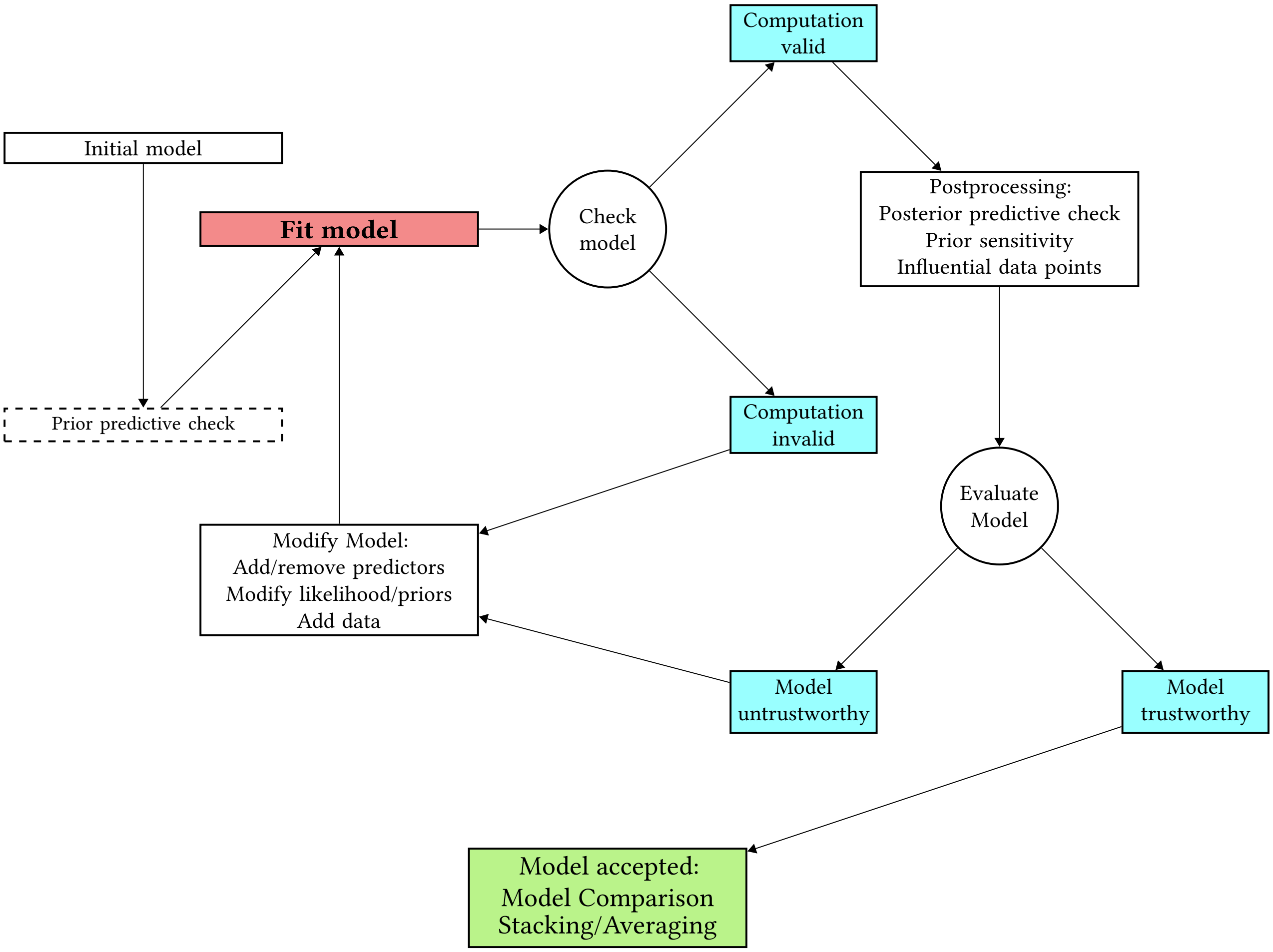

Figure 2: Illustration of the Bayesian workflow, adapted from Figure 1 in [4].

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □ *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. doi:10.1145/3442188.3445922

[2] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of Risks of Generative Text-to-Image Models *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 396–410. doi:10.1145/3600211.3604722

[3] Andrew Gelman, Daniel Simpson, and Michael Betancourt. 2017. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy* 19, 10 (Oct. 2017), 555. doi:10.3390/e19100555 Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[4] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. Bayesian Workflow. doi:10.48550/arXiv.2011.01808 arXiv:2011.01808 [stat].

[5] Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring \ensuremath<mask\ensuremath>: evaluating bias evaluation in language models, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 2209–2225. doi:10.18653/v1/2023.findings-acl.139

[6] Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13310–13325. doi:10.18653/v1/2023.emnlp-main.822

[7] Rachel Longjohn, Giri Gopalan, and Emily Casleton. 2025. Statistical Uncertainty Quantification for Aggregate Performance Metrics in Machine Learning Benchmarks. doi:10.48550/arXiv.2501.04234 arXiv:2501.04234 [stat].

[8] Laura Manduchi, Kushagra Pandey, Robert Bamler, Ryan Cotterell, Sina Däubener, Sophie Fellenz, Asja Fischer, Thomas Gärtner, Matthias Kirchler, Marius Kloft, Yingzhen Li, Christoph Lippert, Gerard de Melo, Eric Nalisnick, Björn Ommer, Rajesh Ranganath, Maja Rudolph, Karen Ullrich, Guy Van den Broeck, Julia E. Vogt, Yixin Wang, Florian Wenzel, Frank Wood, Stephan Mandt, and Vincent Fortuin. 2024. On the Challenges and Opportunities in Generative AI. doi:10.48550/arXiv.2403.00025 arXiv:2403.00025 [cs].

[9] Petrus Mikkola, Osvaldo A. Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, Paul-Christian Bürkner, and Arto Klami. 2024. Prior Knowledge Elicitation: The Past, Present, and Future. *Bayesian Analysis* 19, 4 (Dec. 2024), 1129–1161. doi:10.1214/23-BA1381

[10] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1882–1895. doi:10.1145/3593013.3594134

[11] Eddie Ungless, Bjorn Ross, and Anne Lauscher. 2023. Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image Models, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 7919–7942. doi:10.18653/v1/2023.findings-acl.502

[12] Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. 2025. Position: Evaluating Generative AI Systems is a Social Science Measurement Challenge. doi:10.48550/arXiv.2502.00561 arXiv:2502.00561 [cs].

[13] Katherine Ward. 2024. *Standpoint Phenomenology: Methodologies of Breakdown, Sign, and Wonder.* Springer Nature Switzerland, Cham. doi:10.1007/978-3-031-55456-8

[14] David Gray Widder. 2024. Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1295–1304. doi:10.1145/3630106.3658973