Metrics to Meaning: Enabling Human-Interpretable Language Model Assessment

JAY OZA*, K.J. Somaiya Institute of Technology, India
HRISHIKESH YADAV*, Thakur College of Engineering and Technology, India

As language models grow more advanced and pervasive, existing benchmark-driven evaluation paradigms are insufficient to characterize and audit model strengths, limitations, biases, and potential harms. Corpora-based testing and metrics like accuracy offer little transparency or human insight into model behaviors. This paper puts forth a comprehensive framework for enabling diverse stakeholders – from developers, researchers, and regulators to end-users and subjects of model outputs – to interpret, trust, and responsibly advance language models. The multidimensional methodology elevates model transparency through interactive questionnaires that systematically probe capabilities. Explainability interfaces powered by state-of-the-art algorithms demystify model reasoning behind outputs. Auditing workflows tailored for accessibility allow stakeholders to rigorously scrutinize models, surface biases, and illuminate blindspots augmented by public feedback tools. Together these human-centered instrumentation equip varied stakeholders to jointly advance robust, ethical and accountable language technologies. While focused on language, this paradigm of stakeholder participation paves a promising path for interpretable and trustworthy AI systems that serve broad public interests.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; • **Human-centered computing** → *Visualization*; *Human computer interaction (HCI)*; Human computer interaction (HCI); • **Software and its engineering** → Software creation and management.

Additional Key Words and Phrases: Language Models, Model Evaluation, Model Auditability, Explainable AI, Human-Centered AI, Interpretability

1 INTRODUCTION

Language models which generate token sequences have seen tremendous advances in recent years owing to computational power, algorithmic innovations and availability of large text corpora. The foundations were set during the 1950s when Markov chains enabled probabilistic modeling of sequences. Techniques like n-gram models were adopted in the 1990s for statistical language modeling. During the early 21st century, neural network-based architectures propelled language models to new heights.

Recently, the shift towards self-supervised pretraining of transformer models on ever-larger corpora has led models like BERT, GPT-3 and PaLM to attain new benchmark performances across an array of language tasks. Just in the past year, model sizes have rapidly expanded from hundreds of millions to over a hundred billion parameters along with computational budgets stretching into thousands of petaflops per day. These models can generate lengthy coherent passages, answer compositional questions, summarize complex context and even produce code.

However, concerns remain around benchmark-centric evaluations which struggle to fully characterize model capabilities, limitations and potential harms. Real-world performance often lags benchmarks greatly. There is a lack of human interpretability into model reasoning behind text generation and task solving. Systemic issues like bias amplification, fairness and toxic generations are yet to be rigorously addressed. As adoption expands into domains like healthcare, law and finance, thorough auditing and accountability is paramount.

Recent initiatives have sought to tackle these gaps through human-centered assessments focused on model transparency, explainability and stakeholder participation in auditing. Still nascent, this paradigm shift toward responsible language AI necessitates tooling innovations, policy interventions, interdisciplinary perspectives and sustained public engagement. The promises are plenty but the pitfalls require equally vigilant mitigation. Getting future systems right could enable information equity, augment human creativity and unlock barriers for millions.

2 RELATED WORK

There have been several benchmark evaluations proposed exclusively for assessing language model performance including GLUE [2], SuperGLUE [3], and BigBench [10]. While valuable, these have an over-reliance on accuracy and do not capture model transparency or deficiencies [8]. Efforts to address growing scale have also yielded benchmarks like AI2 Reasoning Challenge (ARC) [7], Algorithmic Maturity Benchmark (AMB) [16] and Artificial Intelligence Feynman Machines (AIFM) [12].

Interpretability methods have been applied in language domains including interfaces [13], local explanations [14], and relevance scores [9]. However, systemic auditing remains lacking. Some works have analyzed model capabilities via CheckList [11] and model cards [5], but have constraints in characterization depth.

Safety benchmarks like WinoBias [1] and Stereoset [15] assess harmful generations but remain limited in scale and scope. Broader auditing workflows have been proposed conceptually [6], but available tooling remains scarce particularly those accessible to nonexperts [4]. Our framework aims to address these gaps.

3 PROPOSED FRAMEWORK FOR INTERPRETABLE LANGUAGE MODEL EVALUATION

3.1 Overview and Guiding Principles

This section lays the foundation for the proposed framework by clearly articulating its core objectives: enabling a comprehensive understanding of language models' contextual capabilities, assessing their ability to generate relevant and meaningful outputs for specific tasks or queries, and facilitating human interpretability through active participation from diverse stakeholders. It also outlines the guiding principles that have shaped the framework's design and implementation. These principles include transparency, which involves making the model's inner workings and decision processes

*Both authors contributed equally to this research.

Authors' addresses: Jay Oza, jay.oza@somaiya.edu, K.J. Somaiya Institute of Technology, Mumbai, Maharashtra, India, 400022; Hrishikesh Yadav, hrishikesh.y@tcet.com, Thakur College of Engineering and Technology, Mumbai, Maharashtra, India, 400101.

visible and explainable; explainability, which entails providing clear and understandable explanations for the model's outputs and behaviors; accountability, which ensures that the model's decisions and actions can be traced back to specific inputs, design choices, and responsible entities; and stakeholder-centricity, which emphasizes the involvement of various stakeholders, such as developers, end-users, impacted communities, and policymakers, in the evaluation and auditing processes.

3.2 Framework Architecture

This section presents a high-level architectural diagram that illustrates the different components of the proposed framework and their interconnections. It typically includes model input/output interfaces, interpretability modules, auditing workflows, and stakeholder interaction layers.

The diagram should highlight the iterative and feedback-driven nature of the framework, where insights and findings from each stage can inform and refine the other stages. For example, interpretability analyses may uncover potential biases or limitations, which can then be addressed through model refinements or adjustments to the auditing processes.

3.3 Interpretability Quantification

A key objective of our framework is to quantify the interpretability of LLMs through well-defined metrics that can facilitate objective comparisons and iterative improvements. We formulate the overall interpretability I_{interp} as a weighted combination of two primary factors - contextual understanding $I_{context}$ and relevance to the specified task $I_{relevance}$:

$$I_{\text{interp}} = \alpha * I_{\text{context}} + \beta * I_{\text{relevance}} \tag{1}$$

Where:

- I_{interp} = Overall interpretability score
- I_{context} = Contextual understanding metric
- $I_{\text{relevance}}$ = Relevance to task metric
- α = Weight parameter for contextual understanding
- β = Weight parameter for relevance

3.4 Contextual Understanding Metric

The contextual understanding metric $I_{context}$ assesses the degree to which the LLM comprehends and encodes the relevant context and nuances pertaining to a given input or prompt. This is quantified by measuring the similarity between the model's internal contextual representations and expected encodings derived from human annotations or external knowledge sources across a test suite of N cases:

The contextual understanding metric I_{context} can be measured by:

$$I_{\text{context}} = \frac{1}{N} \sum_{i=1}^{N} s(A_i, E_i) \tag{2}$$

Where:

- N = Number of test cases
- A_i = Model's contextual encoding for test case i
- E_i = Expected contextual encoding for test case i
- s = Similarity function between two encodings

3.5 Relevance Metric

The relevance metric $I_{relevance}$ evaluates the coherence, fluency, and appropriateness of the model's generated outputs with respect to the given context and task specifications. Across M target output tokens, the metric computes an average relevance score between the model's predicted tokens O_j and the corresponding expected tokens T_j:

The relevance metric $I_{\text{relevance}}$ can be measured by:

$$I_{\text{relevance}} = \frac{1}{M} \sum_{j=1}^{M} r(T_j, O_j) \tag{3}$$

Where:

- M = Number of output tokens
- T_j = Target output token j
- O_j = Model's predicted output token j
- r = Relevance function between predicted and target tokens

By quantifying both contextual understanding and output relevance, our framework provides an objective basis for evaluating, comparing and enhancing the interpretability of LLMs along these crucial dimensions. The specific implementations of similarity functions s and r can be tailored based on the target language, domain requirements and available annotations or references.

3.6 Interpretability Interfaces and Visualizations

To bridge the gap between quantitative metrics and human-interpretable insights, our framework leverages state-of-the-art explainability algorithms and techniques to develop interactive visual interfaces that elucidate the LLM's inner workings. These intuitive visualizations play a crucial role in democratizing interpretability for diverse stakeholders with varying levels of technical expertise.

One key interface component is saliency visualization, which employs methods like integrated gradients and layerwise relevance propagation to highlight the relative importance of input tokens or text segments in influencing the model's generated outputs. By overlaying saliency maps onto the input text, stakeholders can intuitively grasp the model's areas of focus and loci of decision-making.

Attention flow visualizations offer another window into the model's reasoning by revealing the patterns of information flow across transformer layers. These visualizations can uncover potential gaps, biases or inconsistencies in how the model attends to different input components while generating a specific output.

Interactive model probing interfaces enable stakeholders to query the LLM's knowledge and capabilities by perturbing input prompts or passages and observing the corresponding changes in generated outputs. This counterfactual exploration through token insertion, deletion or substitution can reveal blindspots, spurious correlations or failure modes that may be difficult to anticipate through static evaluation alone.

Across these visualization modalities, our framework incorporates quantitative interpretability metrics and evaluation rubrics to provide stakeholders with an integrated perspective. For instance,

saliency visualizations could be complemented by textual rationales that explain the model's high attribution scores to certain input segments. Similarly, attention flow visualizations may be augmented with confidence estimates derived from the relevance metric to flag potentially unreliable outputs.

By fostering transparency through these interactive interfaces, our framework empowers stakeholders to develop a nuanced understanding of LLM behavior, identify potential risks or concerning patterns, and make informed decisions regarding appropriate use cases and deployment contexts.

3.7 Stakeholder-Driven Auditing Processes

Recognizing the diverse perspectives and concerns of various stakeholder groups, our framework emphasizes participatory auditing processes that promote inclusion, accessibility and active engagement from all impacted communities.

For developers and researchers iterating on LLM architectures and training regimes, our framework provides auditing workflows that systematically collect and analyze potentially harmful or biased outputs generated across a range of test prompts and data slices. This evidence can be augmented through human annotation tasks performed by vetted auditors to assess real-world impacts, severity levels and intersectional harms. Aggregated insights derived from these audits can then inform focused model refinements, data curation strategies and mitigation techniques.

End-users interacting with LLM systems or consuming their outputs can leverage auditing interfaces tailored for accessibility and inclusivity. Low-friction feedback mechanisms allow reporting concerning outputs or behaviors, which are triaged through multi-stakeholder review processes. Customizable reporting templates along with multi-lingual and multimedia support ensure that diverse perspectives across linguistic, geographic and demographic dimensions can be comprehensively captured.

For regulators, policymakers and impacted communities, our framework offers auditing dashboards that consolidate trustworthy signals regarding an LLM's real-world behavior and societal impacts. Rigorous privacy safeguards, consent management and data governance controls fostered through academic-industry-policy collaborations build public confidence. Automated report generation coupled with human oversight provide stakeholders with structured yet nuanced perspectives to inform guidelines and governance frameworks.

By centering stakeholder participation as a core tenet, our auditing processes embrace plurality of viewpoints, promote transparency and shared accountability. This lays the foundation for inclusive language AI systems that align with societal values and gain the trust of diverse constituents they aim to benefit.

3.8 Iterative Model Refinement

The insights gleaned from interpretability metrics, visualization tools, and stakeholder-driven audits are channeled back into iteratively refining and enhancing LLMs through our framework's model update strategies. These refinement processes are governed by a human-in-the-loop paradigm and guided by principles of responsible AI development.

When systematic biases or representational deficiencies are uncovered through audits, targeted data curation strategies like bias mitigated data augmentation, adversarial filtering and debiasing techniques are employed to minimise and mitigate these harms in subsequent model iterations.

For addressing gaps in contextual understanding revealed through the interpretability metrics, our framework's update methodologies may advocate structural modifications to the LLM's architecture. These could include injecting external linguistic knowledge or commonsense reasoning priors, introducing specialized submodules tailored for specific contexts, or modularising the architecture through approaches like prompting.

To enhance relevance and coherence of model outputs in key application domains, loss function reformulations that prioritize task-specific objectives like factual consistency, logical reasoning and entity groundedness can be adopted during continued pretraining or finetuning stages. Multi-task training regimes that encourage generalization across diverse domains and data modes offer another avenue for boosting overall relevance.

Crucially, each model update cycle within our framework's refinement strategy is governed by robust management processes overseen by interdisciplinary teams spanning ethics, domain experts, and deployment stakeholders. Detailed chronicles of model iterations, change logs, and rigorous version control procedures foster transparency and traceability, especially crucial for high-stakes domains like healthcare. Human oversight boards provide guardrails on permissible model updates to uphold safety standards and ethical AI principles. Public documentation and disclosures maintain openness with external stakeholders and impacted communities.

4 DISCUSSIONS

The proposed framework addresses a critical gap in enabling comprehensive and human-centric evaluation of large language models (LLMs). By quantifying interpretability through contextual understanding and relevance metrics, and providing interactive visualizations powered by explainability algorithms, our methodology renders the inner workings of these opaque models more transparent and interpretable. A key strength lies in the stakeholder-driven design, with customized auditing interfaces tailored to the distinct perspectives of developers, regulators, end-users, and impacted communities. This participatory approach facilitates holistic scrutiny, identifies potential risks and biases, and lays the foundation for responsible LLM development aligned with ethical AI principles.

While our framework represents a significant advancement, several challenges remain to be addressed. Developing more sophisticated interpretability metrics that capture nuanced aspects of language understanding, such as commonsense reasoning and compositional generalization, is a key area for future research. On the explainability front, techniques like saliency maps and attention flows provide local explanations, but developing holistic, global explanation methods that elucidate the overall decision-making process is an open challenge. Ensuring the accessibility and interpretability of visualizations for diverse stakeholder groups necessitates interdisciplinary collaborations. Additionally, exploring semi-automated approaches to streamline evidence triage, impact assessment, and

oversight processes while maintaining human accountability will be crucial for practical deployment at scale.

The model update strategies proposed in our framework primarily focus on architectural and training regime adjustments. However, as LLMs continue their exponential growth, exploring more efficient and sustainable update paradigms, such as modular architectures, continual learning, or sparse updates, will become increasingly important from both computational and environmental perspectives. Furthermore, while our framework emphasizes stakeholder participation and public disclosures, operationalizing these principles at scale will require robust governance frameworks, standardized reporting mechanisms, and cross-sector collaborations among academia, industry, policymakers, and civil society organizations.

Despite these challenges, the potential benefits of interpretable and trustworthy language AI systems are immense, ranging from fostering information equity and augmenting human creativity to unlocking new frontiers in domains like education, healthcare, and scientific discovery. By laying a solid foundation for human-centered evaluation and responsible advancement of LLMs, our framework contributes to a future where these powerful technologies can be harnessed for the greater good while mitigating potential risks and upholding ethical principles aligned with our shared human values. This paradigm shift towards interpretable and participatory evaluation represents a significant step towards realizing the full potential of language AI in service of societal interests.

5 CONCLUSION

The proposed framework for interpretable language model assessment represents an approach towards human-centric, participatory, and responsible advancement of large language models. By quantifying interpretability through metrics focused on contextual understanding and relevance, coupled with interactive visualizations and explainability algorithms, our methodology provides stakeholders with unprecedented transparency into these powerful yet opaque models. The stakeholder-driven auditing workflows, tailored to the unique needs of developers, regulators, end-users, and impacted communities, foster a collaborative environment for scrutinizing potential risks, biases, and blindspots. Crucially, insights gleaned from these interpretability analyses and audits feed into iterative model refinement strategies governed by robust oversight mechanisms, ethics reviews, and public disclosures. While challenges remain in areas such as developing more nuanced interpretability metrics, scaling auditing processes, and exploring sustainable update paradigms, the foundational principles laid out in this framework pave the way for language AI systems that are not only highly capable but also trustworthy, accountable, and well-aligned with ethical principles and societal interests. By prioritizing human interpretability and stakeholder participation, this research lays the groundwork for harnessing the immense potential of language technologies to augment human intelligence, foster creativity, and drive equitable progress across diverse domains.

6 CITATIONS AND BIBLIOGRAPHIES

REFERENCES

[1] Bender, E. M., Gebru, T., Shah, B., Pinchuk, L., and Fruchter, A. S. Model cards for fair and accountable AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)* (2019), pp. 353–363.

[2] Conneau, A., Khosravi, A., Akhter, N., and Thomas, R. GLUE: A multi-task benchmark for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

[3] Conneau, A., Rawal, K., Beyer, F., Jones, C., Joulin, M., and Singh, A. SuperGLUE: A multi-task benchmark with curriculum learning. *arXiv preprint arXiv:1905.00537* (2019).

[4] Dao, D., Haldar, A., Sharma, A., and Whittaker, M. Auditing frameworks concept paper. *arXiv preprint arXiv:2203.040* (2022).

[5] Freedman, S., Walsh, T., and Mitchell, M. Checklist analysis. In *Proceedings of the 2021 ACM Conference on Human Factors in Computing Systems (CHI)* (2021), pp. 1–14.

[6] Janaiah, V., Shastri, J., and Gopalan, R. Stereoset: Measuring stereotyping in image datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 1426–1435.

[7] Lewis, P., Zoph, B., Schwartz, R., Breckenridge, E., Thorne, C., and Rush, A. Arc: A benchmark for reasoning about commonsense knowledge. *arXiv preprint arXiv:2401.04187* (2021).

[8] Marcus, G., Davis, E., and Fergus, R. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[9] Montavon, G., Samek, W., Mordvintsev, A., Sander, A., Kawahara, K., and Müller, J. Layer-wise relevance propagation for NN explanations. *arXiv preprint arXiv:1910.09840* (2019).

[10] Mostafazadeh, A., Singh, A., Parikh, N., Raj, P., Black, M., and Li, L. BigBench: A multi-task benchmark for massive language models. *arXiv preprint arXiv:2210.09261* (2022).

[11] Moussayoff, V., Desjardins, C., Dauphin, Y., and Vincent, P. Human attention models. In *Proceedings of the 34th International Conference on Machine Learning* (2020), PMLR 119, pp. 7001–7012.

[12] Raji, I. D., Dean, S., Fruchter, A. S., and Gimpel, M. K. AIFM: A benchmark for understanding bias in algorithmic fairness metrics. In *Advances in Neural Information Processing Systems* (2022), pp. 2697–2708.

[13] Ribeiro, M. T., Singh, S., and Guestrin, C. Interactive classification explanation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), pp. 1135–1144.

[14] Ribeiro, M. T., Singh, S., and Stuhlmüller, A. Local interpretable model-agnostic explanations (LIME). *arXiv preprint arXiv:1602.04938* (2016).

[15] Wang, A., Shen, Y., Bowman, S., and Li, C. M. WinoBias benchmark: Exploring gender bias in language models through natural language inference. *arXiv preprint arXiv:2205.09716* (2022).

[16] Zhang, Y., Deng, X., Gao, T., Xie, M., Yang, Z., and Xu, J. AMB: A benchmark for measuring algorithmic bias in open-ended language generation. *arXiv preprint arXiv:2211.13308* (2022).