GAICo: Demonstrating a Unified Framework for Multi-Modal GenAI Evaluation

Pallav Koppisetti, Nitin Gupta, Kausik Lakkaraju, Biplav Srivastava

University of South Carolina
Columbia, South Carolina, USA

Abstract

The rapid evolution of Generative AI, yielding outputs across text, structured data, images, and audio, has outpaced the development of standardized evaluation tools, leading to fragmented and non-reproducible practices. GAICo (Generative AI Comparator) offers a solution: a deployed, open-source Python library that provides a unified, extensible, and reproducible framework for multi-modal GenAI evaluation. Our demonstration highlights GAICo's utility through a practical case study: evaluating and debugging composite AI Travel Assistant pipelines. We show how GAICo facilitates isolating performance issues, for instance, distinguishing orchestrator LLM planning deficiencies from specialist image model generation flaws, by consistently comparing diverse outputs against tailored references. This framework streamlines development, improves system reliability, and promotes reproducible evaluation, making it a critical tool for building safer and more effective AI. Its rapid adoption, evidenced by over 14,000 downloads, underscores its relevance and impact within the AI community.

Code — github.com/ai4society/GenAIResultsComparator

Introduction

Generative AI is transforming research and practice, producing outputs across text, images, audio, and structured data. Yet evaluation has not kept pace. Current approaches rely on ad hoc scripts, siloed metrics, and manual inspection, resulting in comparisons that are fragmented, subjective, and difficult to reproduce (Lopresti and Nagy 2021). The lack of a standardized, multi-modal framework has become a critical barrier for both researchers and practitioners.

GAICo (Generative AI Comparator) establishes a unified, extensible library for evaluating generative outputs across modalities (Figure 1). Implemented as a deployed, open-source Python package, GAICo integrates diverse metrics into a consistent interface, normalizes scores for comparability, and automates reporting and visualization, enabling practitioners to build trustworthy AI systems more quickly and reliably (Pekka et al. 2018). **GAICo acts as a crucial collaborative bridge by unifying evaluation methodologies across traditionally disparate AI disciplines such as NLP, computer vision, audio processing, and sub-discipline-specific tasks like automated planning and time series forecasting.** The demonstration video and a collection of 17 ready-to-run Python notebooks in the code repository show sample evaluation of LLM outputs in individual modalities (text, audio, image) as well as complex multi-modal cases.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

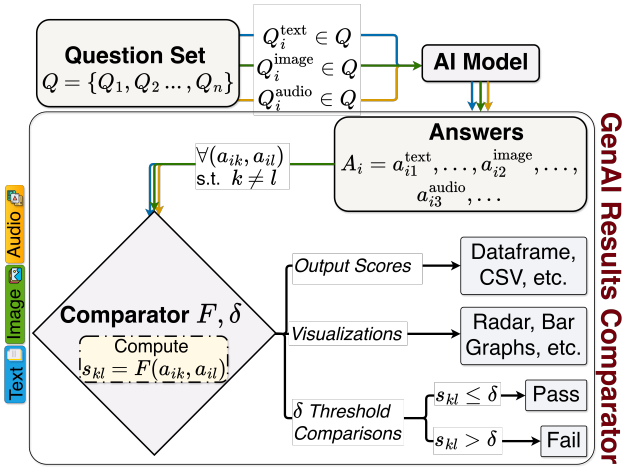

Figure 1: The multi-modal GAICo workflow. The framework processes answers from multi-modal (text, image, audio) AI models, computes pairwise similarity scores (s_{kl}), and constructs several outputs: raw data reports, visualizations, and pass/fail assessments against a threshold δ.

Demonstration on Assistant Evaluation

This demonstration will illustrate GAICo's capabilities in not just comparing, but also diagnosing performance issues with complex, multi-modal AI systems. We achieve this by evaluating three composite AI pipelines designed as travel assistants, depicted in Figure 2 (left). Each pipeline includes an orchestrator LLM that generates a structured itinerary (day-level descriptions, action sequences, and budgets), and specialist models that produce images and audio summaries from the orchestrator's prompts. Pipeline A is treated as the reference, while Pipeline B and Pipeline C are compared against it. In this demonstration, we focus on the image

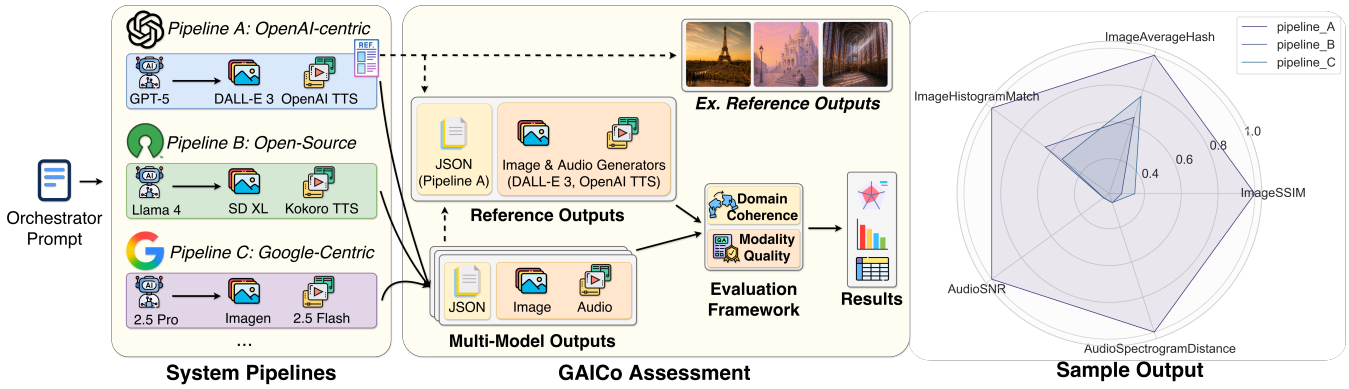

Figure 2: Demonstration of GAICo on a multi-modal travel assistant case study. (*Right*) 3 pipelines, Pipeline A (GPT-5 + DALL·E 3 + OpenAI TTS) as reference, Pipeline B (Llama 4 + Stable Diffusion XL + Kokoro TTS), and Pipeline C (Gemini 2.5 Pro + Imagen + Google TTS), generate plans, images, and audio, which are evaluated by GAICo. (*Left*) The output radar plot shows image and audio fidelity relative to references derived from each pipeline's prompts.

modality, while details on text and audio evaluation can be found in the extended GAICo paper (Gupta et al. 2025).

GAICo evaluates the fidelity of specialist image models by applying three complementary metrics: SSIM (Wang et al. 2004) capturing structural fidelity and luminance similarity, average hashing (Dufournaud, Schmid, and Horaud 2004) encoding global layout as a perceptual fingerprint, and histogram matching (Shen 2007), assessing alignment of color distributions. Together, these metrics provide a balanced assessment of visual quality.

The results highlight distinct strengths across pipelines. As shown in Figure 2 (right), Pipeline C achieves higher structural fidelity, while Pipeline B more closely matches color distributions. Such complementary patterns illustrate the value of GAICo's multi-metric approach: no single measure is sufficient to capture all dimensions of quality. GAICo also outputs standardized CSV reports, along with radar and bar plots, making differences interpretable and reproducible. By applying these metrics to a realistic travel assistant scenario, GAICo demonstrates how fragmented evaluation tasks can be unified into a transparent, extensible workflow. While this paper highlights image evaluation, GAICo generalizes seamlessly to text, structured data, and audio, enabling holistic analysis of composite AI systems.

Related Work

General-Purpose Toolkits: Hugging Face's evaluate library (Wolf et al. 2019) and scikit-learn (Pedregosa et al. 2011) provide widely used metrics for NLP and ML but are confined to specific domains. NLTK (Bird, Klein, and Loper 2009) and SpaCy (Honnibal et al. 2020) include basic overlap-based measures but lack support for structured or multimedia data.

Integrated Frameworks: Ragas (Es et al. 2024) and DeepEval (Ip and Vongthongsri 2025) offer pipelines that couple evaluation with LLM inference, often leveraging "LLM-as-a-judge" paradigms. While these frameworks provide flexibility, they introduce challenges in terms of reproducibility, cost, and dependency on external APIs.

Domain-Specific Tools: Specialized libraries exist for individual modalities: plan validators for automated planning (Howey, Long, and Fox 2004), dtaidistance for time-series (Berndt and Clifford 1994), scikit-image for image metrics (Van der Walt et al. 2014), or librosa for audio (McFee et al. 2015). These remain siloed and require developers to manually integrate the results.

GAICo addresses the limitations of prior approaches by offering a post-hoc, unified comparator. It decouples evaluation from inference, supports diverse modalities, normalizes results to a common scale, and provides visualization and reporting out of the box.

Significance and Conclusion

GAICo integrates over 15 metrics into a single, extensible framework that supports evaluation across text, structured data, images, and audio, making it broadly applicable to composite AI systems. **A collection of 17 ready-to-run Python notebooks** within the repository further lowers the barrier to adoption, providing practical entry points for diverse tasks. GAICo has already demonstrated strong community uptake, with over 14,000 downloads in its first three months (PyPI Stats 2025). Its design emphasizes openness and extensibility (to custom metrics), ensuring the framework evolves alongside advances in generative AI. Standardized CSV reports and visualizations make results replicable, interpretable, and easy to share, addressing the reproducibility issue with modern GenAI evaluation. In the travel assistant case study, GAICo exposed complementary strengths of specialist image models. Such distinctions are critical for debugging and improving AI pipelines, yet difficult to capture with isolated tools. Taken together, GAICo is both a **practical toolkit for immediate adoption and a collaborative bridge across AI subfields**. Its open-source availability, early adoption, and extensible design ensure that it is a promising step towards transparent, reproducible, and trustworthy generative AI systems.

References

Berndt, D. J.; and Clifford, J. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining*, 359–370.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Dufournaud, Y.; Schmid, C.; and Horaud, R. 2004. Image matching with scale adjustment. *Computer vision and image understanding*, 93(2): 175–194.

Es, S.; James, J.; Anke, L. E.; and Schockaert, S. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158.

Gupta, N.; Koppisetti, P.; Lakkaraju, K.; and Srivastava, B. 2025. GAICo: A Deployed and Extensible Framework for Evaluating Diverse and Multimodal Generative AI Outputs. *arXiv preprint arXiv:2508.16753*.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Howey, R.; Long, D.; and Fox, M. 2004. VAL: Automatic plan validation, continuous effects and mixed initiative planning using PDDL. In *16th IEEE International Conference on Tools with Artificial Intelligence*, 294–301. IEEE.

Ip, J.; and Vongthongsri, K. 2025. deepeval.

Lopresti, D.; and Nagy, G. 2021. Reproducibility: evaluating the Evaluations. In *International Workshop on Reproducible Research in Pattern Recognition*, 12–23. Springer.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. *SciPy*, 2015: 18–24.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pekka, A.; Bauer, W.; Bergmann, U.; Bieliková, M.; Bonefeld-Dahl, C.; Bonnet, Y.; Bouarfa, L.; et al. 2018. The European Commission's high-level expert group on artificial intelligence: Ethics guidelines for trustworthy ai. *Working Document for stakeholders' consultation. Brussels*, 1–37.

PyPI Stats. 2025. GAICo Download Statistics. https://pepy.tech/projects/gaico?timeRange=threeMonths&category=version&includeCIDownloads=true&granularity=daily&viewType=line&versions=0.3.0,0.2.0,0.1.5,0.1.4. Accessed Aug 18, 2025.

Shen, D. 2007. Image registration by local histogram matching. *Pattern Recognition*, 40(4): 1161–1172.

Van der Walt, S.; Schönberger, J. L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J. D.; Yager, N.; Gouillart, E.; and Yu, T. 2014. scikit-image: image processing in Python. *PeerJ*, 2: e453.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.