Evaluating Generative AI and Agentic Systems: Methodologies for Systematic Assessment and Continuous Improvement

Arian Pasquali

Orq.ai, Amsterdam, The Netherlands

Abstract

As large language models and AI agents define the current era of enterprise AI adoption, organizations face a critical challenge: how to systematically evaluate and monitor these systems at scale while maintaining meaningful human oversight. This keynote draws from real-world industry practice to review evaluation methodologies for generative AI and agentic systems, with particular emphasis on the gap between academic benchmarks and production realities. We begin by examining why traditional evaluation approaches fall short in enterprise settings, where outputs are diverse, context-dependent, domain-specific, and subject to business constraints, and introduce Evaluation-Driven Development as a practical framework for continuous, iterative improvement of AI systems in production. A central theme is the operationalization of LLM-as-a-Judge approaches: how to align automated judges with human judgment, avoid common calibration pitfalls, and scale evaluation coverage through panels of judges that improve reliability and reduce individual model bias. We then turn to agentic systems specifically, covering agent simulation as a method for stress-testing agent behavior before deployment, and trace intelligence, the practice of extracting structured meaning from agent execution traces to surface usage patterns, failure modes, and improvement signals at scale. Throughout, we emphasize practical methods for closing the loop between evaluation signals and system improvements, including synthetic data generation, domain-specific metrics, and data segmentation strategies that enable rigorous development lifecycles in industry settings.

Short Bio

Arian Pasquali is an Applied AI Researcher at `Orq.ai`, an AI Engineering & Evaluation Platform focused on Generative AI and AI agents for enterprises, based in Amsterdam. He specializes in NLP and LLMs, with focus on evaluation and annotation workflows applied to AI agents. During the last years he has been bridging the gap between research and practical applications across government, education, and enterprise domains. His current research interests focus on designing processes for the systematic evaluation of large language models and agentic architectures. He has published research in top-tier conferences on keyword extraction, temporal narrative generation, and interactive systems for information retrieval. Arian is also an active contributor to the academic community, where he has served as a reviewer and chair for major international conferences, including ECIR and ECML-PKDD. His contributions have been recognized with several awards, including the World Summit Award 2019 for Government and Citizen Engagement, as well as paper awards at ECIR.

In: R Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story 26 Workshop, Delft (The Netherlands), 29-March-2026

arianpasquali@gmail.com (A. Pasquali)
https://github.com/arianpasquali (A. Pasquali)
0000-0002-3487-9397 (A. Pasquali)