AI Evaluations using Gemini APIs

Problem Statement

Despite Gemini's advanced multimodal capabilities, its adoption for robust AI evaluation is bottlenecked by a critical tooling gap. Developers face inconsistent integration, fragmented documentation, and a lack of practical guides across the MLOps ecosystem. This creates significant friction, hampering the ability to reliably benchmark, debug, and productionize Gemini-powered applications compared to the seamless workflows available for other major LLMs.

Project Goal

To elevate Gemini to a first-class citizen within the AI evaluation ecosystem. This project's goal was to bridge the existing integration gap by engineering native support in key frameworks, authoring centralized, high-quality documentation, and developing practical, code-first tutorials. The ultimate aim is to eliminate adoption barriers and provide developers with a seamless, production-ready workflow for evaluating sophisticated LLM applications with Gemini.

Summary of Pull Requests

The following table provides a high-level overview of contributions, summarizing their focus and strategic importance within the broader ecosystem.

Project	Repository	Contribution Focus (Inferred from PR Title)	Strategic Impact	Pull Request (Status)
TruLens	truera/trulens	Adding Google Gemini provider for evaluations.	Extends TruLens's "LLM-as-a-Judge" capabilities to leverage Gemini's models.	Merged

TruLens	truera/trulens	Tutorial for multimodal (Text, Audio, Video) evaluation with Gemini.	Positions TruLens as a key evaluation framework for evaluating multimodal AI applications.	Merged
LlamaIndex	run-llama/llama_index	Added a Developer Guide for Gemini Tool-Use	Created comprehensive documentation and practical examples for integrating Google Gemini's function calling feature.	Merged
Comet Opik	comet-ml/opik	Documentation for tracing Gemini LLM calls via Vertex AI.	Provides crucial observability for enterprise-grade Gemini deployments on Google Cloud, linking development to production monitoring.	Merged
Arize Phoenix	Arize-ai/phoenix	Adding support for Gemini API libraries in the evaluation suite.	Broadens Phoenix's model-agnostic evaluation capabilities, reinforcing its position as a comprehensive observability platform.	Merged
LiteLLM	BerriAI/litellm	Fixing Gemini API key environment variable support.	Addresses a foundational requirement for secure, standardized authentication, enabling seamless integration of Gemini into multi-cloud, multi-provider AI stacks.	Merged
W&B Weave	wandb/weave	Documentation for multi-turn conversation	Tackles the complex challenge of evaluating conversational AI, using	Merged

		evaluation with Gemini.	Gemini to score nuanced metrics like coherence and context retention.	
W&B Weave	wandb/weave	feat(docs): Add tutorial for evaluating LlamaIndex RAG agents with Gemini	Tutorial demonstrating how to build and evaluate a multi-tool LlamaIndex RAG agent, powered by Google's Gemini models, using W&B Weave	Open
Evaluations-with-GeminiAPI	sahusiddharth/Evaluations-with-GeminiAPI	Tutorials for LlamaIndex ReAct agent and tool-calling evaluation.	Serves as a practical blueprint and proof-of-concept for applying advanced evaluation patterns to Gemini-powered agentic systems.	Merged
Ragas	exploding gradients/ragas	Refactor/update gemini to genai sdk	Updates the InstructorLLM wrapper and instructor_llm_factory to migrate from the legacy Gemini provider (instructor.from_gemini) to the new Google GenAI SDK (google-genai)	Open
Ragas	exploding gradients/ragas	Add LlamaIndex FunctionAgent evaluation examples with Gemini and new Ragas	Introduces new evaluation examples for LlamaIndex's FunctionAgent, showcasing how to use Ragas with and without LLM-as-a-judge for agentic evaluations	Open

An Analysis of Contributions to the Gemini Evaluation Ecosystem

TruLens - Adding Google Gemini Provider for Evaluations (PR #2153) - MERGED

What Library Does: TruLens is an evaluation and tracking platform for LLM experiments that provides "LLM-as-a-Judge" capabilities, helping developers build reliable LLM applications by monitoring, testing, and debugging them to minimize risks like toxicity and bias.

What This PR Accomplished: Added native Google Gemini provider support to TruLens, enabling direct integration with Gemini models for LLM evaluations without requiring abstraction layers like LiteLLM.

Impact on Your Project Goal: This PR directly eliminates integration friction by giving Gemini native first-class status in TruLens, putting it on par with OpenAI and Anthropic. It extends TruLens's evaluation capabilities to leverage Gemini's advanced reasoning for judging LLM outputs, addressing the core problem of inconsistent integration that bottlenecked Gemini adoption in evaluation workflows.

TruLens - Tutorial for Multimodal (Text, Audio, Video) Evaluation with Gemini - MERGED

What Library Does: TruLens is an evaluation and tracking platform for LLM experiments that provides "LLM-as-a-Judge" capabilities, helping developers build reliable LLM applications by monitoring, testing, and debugging them to minimize risks like toxicity and bias.

What This PR Accomplished: Created a comprehensive tutorial demonstrating how to use TruLens for evaluating multimodal AI applications that process text, audio, and video inputs using Gemini models, showcasing Gemini's native multimodal capabilities within the evaluation framework.

Impact on Your Project Goal: This tutorial positions TruLens as a key evaluation framework for multimodal AI applications and fills critical documentation gaps by providing practical, code-first guidance. It showcases Gemini's unique multimodal strengths, establishing it as the preferred choice for advanced evaluation scenarios beyond simple text-based assessments, directly addressing the "lack of practical guides" problem in your project statement.

LlamaIndex - Developer Guide for Gemini Tool-Use - MERGED

What Library Does: LlamaIndex is a data framework for LLM applications that helps connect custom data sources to large language models, specializing in building RAG (Retrieval-Augmented Generation) applications and agentic workflows.

What This PR Accomplished: Created comprehensive documentation and practical examples for integrating Google Gemini's function calling feature with LlamaIndex, demonstrating how to build agents that can use tools, perform calculations, query documents, and manage stateful conversations using Gemini's native function calling capabilities.

Impact on Your Project Goal: This contribution provides the "centralized, high-quality documentation" that was missing for Gemini's advanced capabilities in LlamaIndex. It enables developers to build sophisticated AI agents and RAG applications using Gemini, directly addressing the "fragmented documentation" barrier and establishing production-ready workflows for tool-use scenarios that are critical for enterprise AI applications.

Comet Opik - Documentation for Tracing Gemini LLM Calls via Vertex AI - MERGED

What Library Does: Opik is an open-source platform that helps debug, evaluate, and monitor LLM applications, RAG systems, and agentic workflows with comprehensive tracing, automated evaluations, and production-ready dashboards.

What This PR Accomplished: Added documentation for tracing Gemini LLM calls through Vertex AI integration, enabling developers to monitor and observe Gemini model performance in production environments with comprehensive logging and cost tracking capabilities.

Impact on Your Project Goal: This contribution provides crucial observability for enterprise-grade Gemini deployments on Google Cloud, linking development to production monitoring. It addresses the "inconsistent integration" problem by establishing standardized tracing workflows for Gemini, enabling reliable debugging and performance monitoring that's essential for production AI applications - a key requirement for making Gemini a first-class citizen in enterprise evaluation pipelines.

Arize Phoenix - Adding Support for Gemini API Libraries in the Evaluation Suite - MERGED

What Library Does: Phoenix is an open-source AI observability platform designed for experimentation, evaluation, and troubleshooting. It provides tracing for LLM applications using OpenTelemetry-based instrumentation and evaluation capabilities to benchmark application performance.

What This PR Accomplished: Added support for Gemini API libraries in Phoenix's evaluation suite, enabling developers to use Gemini models for LLM evaluations through both the Google GenAI SDK and Vertex AI. The GoogleGenAIModel provides access to Google's Gemini models through the Google GenAI SDK, providing a unified interface for both the Developer API and VertexAI with multimodal support for text, image, and other content types.

Impact on Your Project Goal: This broadens Phoenix's model-agnostic evaluation capabilities and reinforces its position as a comprehensive observability platform. It eliminates the tooling gap by providing native Gemini support, enabling seamless evaluation workflows and giving developers access to Gemini's multimodal capabilities for sophisticated application monitoring and debugging - directly addressing the "inconsistent integration" barrier you identified.

LiteLLM - Fixing Gemini API Key Environment Variable Support - MERGED

What Library Does: LiteLLM is a library that provides seamless multi-LLM integration, allowing developers to use any LLM (Anthropic, OpenAI, Vertex AI, Bedrock, etc.) through a unified interface, eliminating integration headaches from the explosion of different LLM providers and their unique APIs.

What This PR Accomplished: Fixed Gemini API key environment variable support in LiteLLM, ensuring proper recognition and handling of GEMINI_API_KEY or GOOGLE_API_KEY environment variables for secure, standardized authentication across different deployment environments.

Impact on Your Project Goal: This addresses a foundational requirement for secure, standardized authentication that enables seamless integration of Gemini into multi-cloud, multi-provider AI stacks. By fixing this core authentication issue, it removes a critical barrier that was preventing reliable Gemini usage across the seven evaluation frameworks that rely

on LiteLLM, directly tackling the "inconsistent integration" problem central to your project's mission.

W&B Weave - Documentation for Multi-turn Conversation Evaluation with Gemini - MERGED

What Library Does: W&B Weave is a framework for tracking, experimenting with, evaluating, deploying, and improving LLM-based applications, designed for flexibility and scalability to support every stage of LLM application development workflow including tracing, monitoring, and evaluation.

What This PR Accomplished: Created documentation demonstrating how to evaluate multi-turn conversational AI applications using Gemini models within W&B Weave, focusing on complex metrics like coherence, context retention, and conversational flow across multiple dialogue turns.

Impact on Your Project Goal: This tackles the complex challenge of evaluating conversational AI by using Gemini to score nuanced metrics like coherence and context retention. It provides practical guidance for one of the most challenging evaluation scenarios, directly addressing the "lack of practical guides across the MLOps ecosystem" and establishing Gemini as capable of handling sophisticated conversational evaluation tasks that are critical for production chatbot and assistant applications.

W&B Weave - Tutorial for Evaluating LlamaIndex RAG Agents with Gemini - OPEN

What This PR Accomplished: Created a tutorial demonstrating how to build and evaluate a multi-tool LlamaIndex RAG agent powered by Google's Gemini models using W&B Weave's evaluation framework, showcasing the integration between LlamaIndex's agentic capabilities and Weave's monitoring tools.

Impact on Your Project Goal: This tutorial bridges two major frameworks (LlamaIndex and W&B Weave) through Gemini, creating a comprehensive workflow for building, monitoring, and evaluating RAG agents. It provides practical, code-first guidance that eliminates integration friction between popular MLOps tools, directly addressing the "fragmented documentation" problem and establishing seamless workflows for production-ready agentic applications.

Open Evaluations-with-GeminiAPI - Tutorials for LlamaIndex ReAct Agent and Tool-calling Evaluation - MERGED

What Library Does: This is a standalone repository that serves as a comprehensive tutorial collection demonstrating evaluation patterns for Gemini-powered agentic systems, specifically focusing on LlamaIndex ReAct agents and Function agents.

What This PR Accomplished: Created practical tutorials showing how to evaluate LlamaIndex ReAct agents and tool-calling Function agents using Gemini models, providing blueprint implementations for advanced evaluation patterns in agentic systems with step-by-step code examples and evaluation methodologies.

Impact on Your Project Goal: This serves as a practical blueprint and proof-of-concept for applying advanced evaluation patterns to Gemini-powered agentic systems. It directly addresses the "lack of practical guides" by providing comprehensive, code-first tutorials that developers can use as reference implementations, establishing Gemini as a viable choice for sophisticated agent evaluation workflows and eliminating adoption barriers through concrete examples.

Ragas - Refactor/Update Gemini to GenAI SDK - OPEN

What Library Does: Ragas is an open-source framework designed to supercharge LLM application evaluations, providing metrics and evaluation methodologies for RAG (Retrieval-Augmented Generation) systems and other LLM applications.

What This PR Accomplished: Updated the InstructorLLM wrapper and instructor_llm_factory to migrate from the legacy Gemini provider (instructor.from_gemini) to the new Google GenAI SDK (google-genai), ensuring compatibility with the latest Gemini API infrastructure and removing dependencies on deprecated SDKs.

Impact on Your Project Goal: This migration ensures Ragas maintains compatibility with the latest Gemini API infrastructure, preventing future integration breakdowns. It addresses the "inconsistent integration" problem by standardizing on the official Google GenAI SDK, ensuring long-term stability and reliability of Gemini evaluations within Ragas - a critical foundation for maintaining Gemini as a first-class citizen in the evaluation ecosystem.

Ragas - Add LlamaIndex FunctionAgent Evaluation Examples with Gemini and New Ragas - OPEN

What This PR Accomplished: Introduced new evaluation examples for LlamaIndex's FunctionAgent, showcasing how to use Ragas with and without LLM-as-a-judge for agentic evaluations powered by Gemini models, demonstrating how to evaluate different agent types using both pre-built Ragas metrics and custom evaluation metrics.

Impact on Your Project Goal: This creates a comprehensive evaluation blueprint for sophisticated agentic systems, establishing Gemini as the evaluation engine for complex agent workflows. It provides practical implementation examples that bridge LlamaIndex's agent capabilities with Ragas's evaluation framework, directly addressing the "lack of practical guides" and creating production-ready workflows for evaluating the next generation of AI applications that use tools and reasoning.