

Quantix



Why we're here today

**LLMs are a black box,
and hallucinations
always happen in
production**

THE ISSUE

A

LLMs are a black box

Companies aren't ready for the randomness generated via LLMS

B

Testing is hard

Testing via edge cases, and evaluations are often overlooked when moving fast

C

LLM integration is limited

Testing softwares usually have bad first party integrations with your LLM model

What is the best method for testing LLMs?

Performance Evaluation



Assesses the efficiency and effectiveness of the LLM in real-world scenarios.



Robustness Testing

Ensures the LLM can handle diverse and unexpected inputs.

Ethical Considerations



Evaluates the LLM's adherence to ethical standards and bias mitigation.



User Feedback Integration

Incorporates user insights to refine and improve the LLM.

We simulate test cases for LLMs to make them more predictable and safe in production

THE SOLUTION

PRODUCT BENEFITS

1.

Test out your LLMs on guard rails and performance.

2.

Generate synthetic CSV data for training LLMs

3.

Connect your LLMs via different APIs. I.e. Email, Phone, or Direct URL



r/OpenAI · 1y ago

Ok-Process-2187

LLMs switching roles

Question

I have a simple scenario. LLM plays the role of a sales prospect. Human plays the role of a sales person.

But during the chat the LLM will sometimes switch roles and start replying as the sales person.



r/LLMDevs · 2mo ago

artur5092619

LLM guardrails missing threats and killing our latency. Any better approaches?

Discussion

We're running into a tradeoff with our GenAI deployment. Current guardrails catch some prompt injection and data leaks but miss a lot of edge cases. Worse, they're adding 300ms+ latency which is tanking user experience.

Anyone found runtime safety solutions that actually work at scale without destroying performance? Ideally, we are looking for sub-100ms. Built some custom rules but maintaining them is becoming a nightmare as new attack vectors emerge.

Looking for real deployment experiences, not vendor pitches. What's your stack looking like for production LLM safety?

↑ 21 ↓ 0 Share



r/LocalLLaMA · 1mo ago

DanizarQQQ

I've been trying to make a real production service that uses LLM and it turned into a pure agony. Here are some of my "experiences".

Other

Hello everyone. I hope this won't be an off topic, but I want to share my experience in creating real production service. Like a real deal, that will earn money.

For this service I've been using ChatGPT-5 and Claude Haiku 4.5 but I think this could be suitable for other LLMs too.

The idea was as simple as rock. Make an assistant bot that will communicate with people and make a scheduled appointments to the doctor.

Well in a short time I've implemented everything. The vector database that will inject doctor specific knowledge to the conversation at the right time. Multiple tools that will work with doctors data, and couple other integrations. I've extensively made very detailed system prompt, and each tool call returns instructive results. Each tools' parameters' descriptions were written in very detailed way. After testing for a week we finally deployed on production and started to receive conversations from real people.

And then real life had showed a lot of annoying and downright frustrating caveats of these LLMs.

The first frustrating thing is that LLMs makes an assumptions without calling required tool, which deceives people. It happened like this:

Test out your LLMs on guard rails and performance.

Manual testing often misses the tone of actual conversation that happens in production

General features

- Simulate real customer interaction using custom LLMs
- Test LLM responses on guard rails and real interaction just by a click
- Feel confident about your LLM's performance with mathematically-grounded feedback

Generate synthetic CSV data for training LLMs

The biggest challenge in training large language models (LLMs) isn't their architecture—it's finding high-quality, diverse, and unbiased data in a vast and noisy digital landscape. This is true whether you're building an LLM from scratch or fine-tuning a pre-trained one, as you'll need to use high-quality data compiled from multiple sources.

This article overviews LLM training, the need for public web data, and the major public data sources for highly-performant LLMs.

<https://oxylabs.io/blog/llm-training-data>

Tailored CSV data is usually hard to find without proper data cleaning.

General features

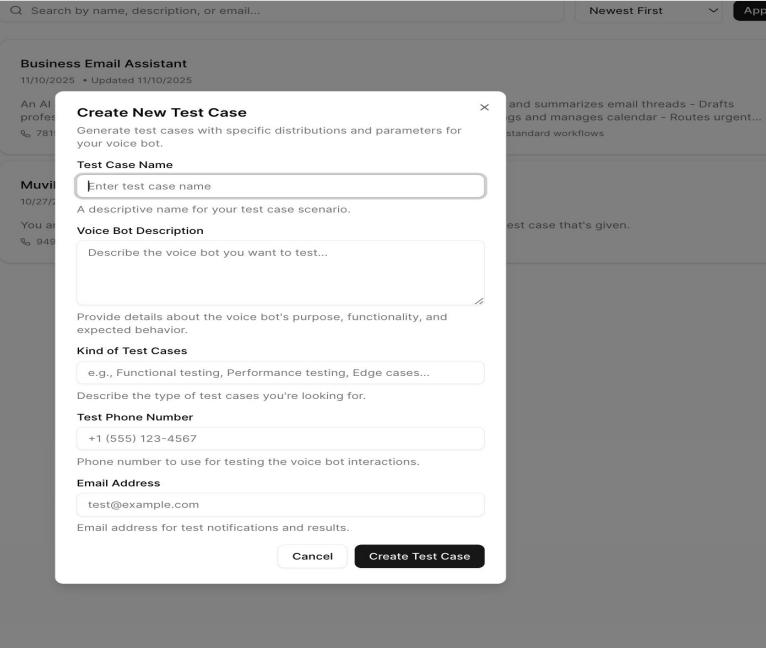
- Give shape of the data and a prompt to get accurate synthetic data for training



Connect your LLMs via different APIs. i.e. Email, Phone, or Direct URL

A simple UI to directly connect to your Email agent, phone agent or API endpoint instead of having to use multiple testing softwares to do the same task

HOW IT WORKS



The screenshot shows a software application window titled "Business Email Assistant". At the top, there's a search bar and a dropdown menu set to "Newest First". A modal dialog box is open, titled "Create New Test Case". The dialog contains fields for "Test Case Name" (with placeholder "Enter test case name"), "Voice Bot Description" (with placeholder "Describe the voice bot you want to test..."), "Kind of Test Cases" (with placeholder "e.g., Functional testing, Performance testing, Edge cases..."), "Test Phone Number" (with placeholder "+1 (555) 123-4567"), and "Email Address" (with placeholder "test@example.com"). At the bottom of the dialog are two buttons: "Cancel" and "Create Test Case". In the background, a list of email assistants is visible, including "Business Email Assistant" (11/10/2025), "An AI professional" (11/10/2025), and "Muvit" (10/27/2025).

1

Create a test suite

Prompt your agent role, test cases to test on, and type of test cases

HOW IT WORKS

2

Generate multiple test cases

Watch us generate real test cases simulating like real customers using custom LLMs

The screenshot shows a web-based application interface for managing test cases. At the top right are buttons for 'Generate Sub-Tests' (with a crossed-out icon) and '+ Add Sub-Test'. Below the title 'Business Email Assistant' and a back arrow, there's a section titled 'Test Case Details' with an 'Edit' button. The details are listed in a table-like format:

Kind of Test Cases	Test Phone Number
Happy path scenarios, successful order completions, standard workflows	7819759065
Email Address	Created
cco206@nyu.edu	11/10/2025

Description: An AI agent that helps executives manage business email communications: - Reads and summarizes email threads - Drafts professional responses - Handles contract termination requests - Schedules meetings and manages calendar - Routes urgent matters to appropriate team members - Maintains professional corporate tone - Understands business context (contracts, deadlines, legal terms)

HOW IT WORKS

- The contract renewal term was agreed to be for 12 months

LLM Judge Summary
Auto verdict: Succeeded · Task confidence: 95.0% · Safety: 100.0% · Faithfulness: 100.0%
The agent effectively summarized the email thread, highlighting key points, decisions made, and outstanding action items with clear owners and deadlines. It then offered to draft a professional reply and delivered a well-structured, courteous email summarizing the contract renewal details and next steps, including deadlines. The tone was professional, the content was aligned with the user's request, and there were no safety issues or hallucinations.

Human Label
Override or confirm the automatic judgment for this run.

Mark Correct **Mark Incorrect**

Notes (optional)
Add any comments about why this run was correct or incorrect...

Run History **Refresh**
No previous runs recorded yet for this sub-test.

3

Analyze your agent

After simulating the test case on your agent get valuable analysis

Summarize the competitive landscape



Cekura

Feature

- Voice agent testing

Strengths

- Goes beyond just conversation, captures tone, intent and other variables

Weaknesses

- Only expands to voice agents

Braintrust

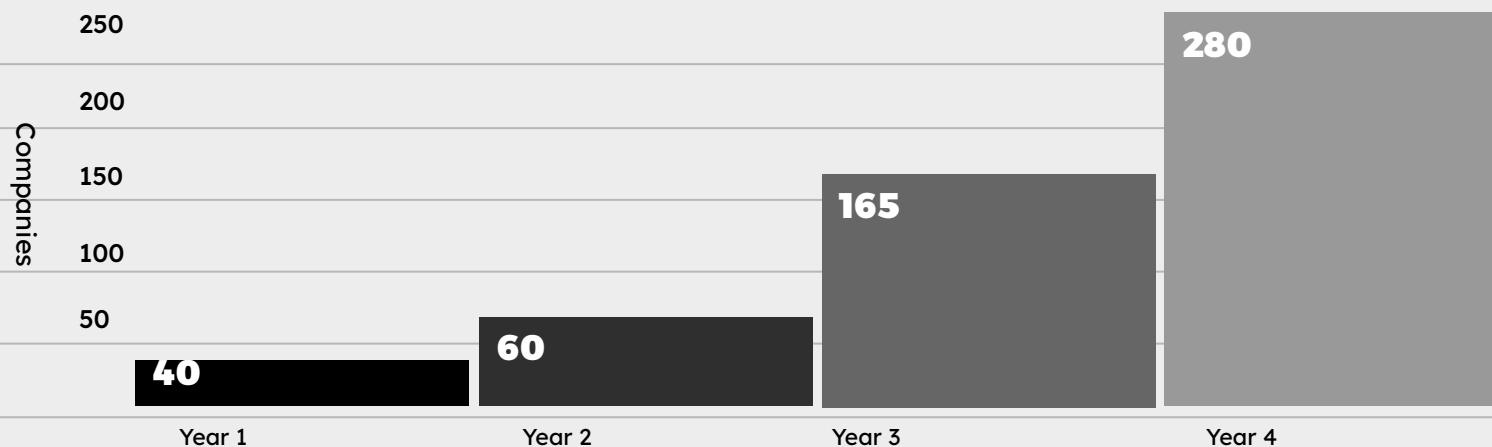
- Testing, evals, and iterations for LLMs

- Amazing insights to test out agents

- Only expands to api layer and becomes complicated when your initial workflow doesn't start with braintrust

FINANCIALS

Hybrid SaaS & Usage Based



MEET THE TEAM



Chuma Oyigbo

CEO

Satvik Seeram

CTO

Weilin Cheng

CMO

Feifan Yang

CPO

**THANK
YOU!**