Anote

A CASE STUDY

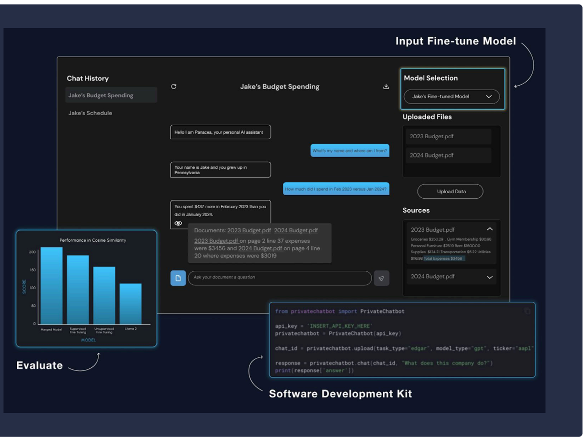
Anote’s Platform Excels in Rigorous Humane Intelligence and NIST Red Team Evaluation

BACKGROUND ON ARIA

At Anote, our mission is to create AI systems that are not only powerful but also secure, ethical, and trustworthy. These core values were rigorously tested in a red team evaluation led by the National Institute of Standards and Technology (NIST) ARIA program in partnership with Humane Intelligence. The evaluation provided a platform to demonstrate Anote’s capabilities and resilience in addressing the complex challenges of modern AI governance.

EVENT OVERVIEW

The red team evaluation, hosted in Arlington, Virginia during the CAMLIS conference, represented an opportunity to assess the security, robustness, and ethical compliance of cutting-edge AI systems. Over 30 expert testers, selected from a competitive pool of 500 applicants, conducted systematic attacks on participating AI platforms. Their goal was to identify model vulnerabilities, assess compliance risks, and test the resilience of AI-generated outputs under adverse conditions.



Hello I am Panacea, your personal AI assistant

What’s my name and where am I from?

Your name is Jake and you grew up in Pennsylvania

How much did I spend in Feb 2023 versus Jan 2024?

You spent $437 more in February 2023 than you did in January 2024.

Documents: 2023 Budget.pdf 2024 Budget.pdf

2023 Budget.pdf on page 2 line 37 expenses were $3456 and 2024 Budget.pdf on page 4 line 20 where expenses were $3019

Groceries $250.29 . Gym Membership $80.98 Personal Furniture $76.19 Rent $1600.00 Supplies $124.21 Transportation $5.22 Utilities $116.98 Total Expenses $3456

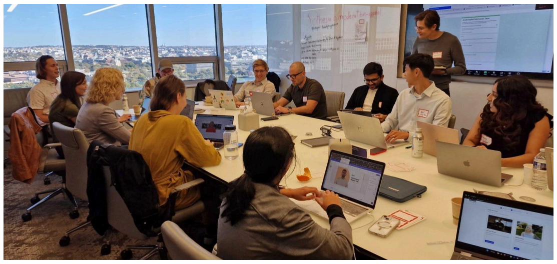
```
from privatechatbot import PrivateChatbot

api_key = 'INSERT_API_KEY_HERE'
privatechatbot = PrivateChatbot(api_key)

chat_id = privatechatbot.upload(task_type="edgar", model_type="gpt", ticker="aapl"

response = privatechatbot.chat(chat_id, "What does this company do?")
print(response['answer'])
```

EVENT HIGHLIGHTS



More than 50 certified red teamers from Camlis subjected Anote’s system to advanced challenges, including prompt injection attacks, adversarial input manipulations, and data poisoning techniques. Over 3,000 interactions during the evaluation highlighted the platform’s reliability, with moderation systems combining AI-driven automation and human oversight to ensure trustworthy outputs. Despite these tests, Anote’s platform remained resilient and trustworthy, mitigating threats.

PARTNERING ORGANIZATIONS

The evaluation was a collaborative effort involving leading organizations in the AI ecosystem. Alongside Anote, systems from Meta, Robust Intelligence, and Synthesia underwent rigorous testing, offering a comparative lens on the state of AI resilience. Anote partnered with institutions, including NIST, the Cybersecurity and Infrastructure Security Agency (CISA), and Humane Intelligence, to advance best practices in AI governance.

APPROACH

Core to Anote's success is its human-centered AI methodology, a unique framework that combines advanced AI capabilities with human expertise to tackle domain-specific challenges. This approach is built on three foundational pillars:

1 Integrating Generative AI with Human Expertise: Anote leverages the computational power of generative AI while integrating the nuanced insights of human users. This hybrid methodology ensures solutions are contextually accurate.

2 Active Learning from Users: Through iterative learning cycles, Anote refines its models based on user feedback, especially in edge cases. This continuous improvement mechanism drives enhanced performance

3 Domain-Customized Solutions: Anote tailors its outputs to align with the unique requirements, delivering results that are precise, relevant, and effective.

EVALUATION

The evaluation aimed to test Anote's ability to detect and mitigate malicious activities while ensuring adherence to ethical standards. Throughout the process, Anote's AI systems successfully identified and blocked all unauthorized attempts, providing real-time feedback to users and maintaining transparency. Ethical compliance was another critical benchmark. Anote achieved a 95% success rate in preemptively flagging harmful or misleading outputs.

RESULTS

- Resilience Against Malicious Input: Even when faced with heavily redacted documents and adversarial prompts, Anote's proprietary tagging and entity extraction algorithms maintained high accuracy.
- Dynamic Risk Mitigation: Attempts to bypass privacy-preserving summarization features were countered in real time, demonstrating robust risk management capabilities.
- Ethical Adherence: When tested with fabricated patient records designed to elicit biased outputs, Anote's system adhered strictly to compliance standards.

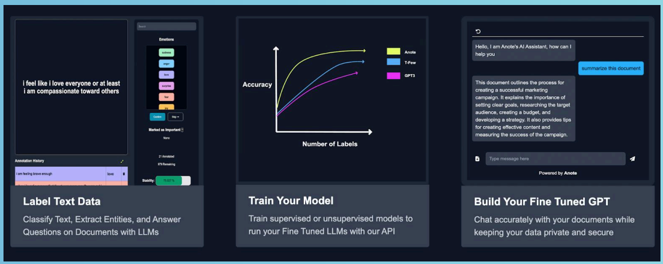
Data Annotation: Subject matter experts labeled datasets to ensure high-quality training inputs. These structured annotations were used for supervised fine-tuning.

Model Training: Anote performed supervised fine-tuning on over 2,400 curated question-answer pairs, which was complemented by Reinforcement Learning from Human Feedback (RLHF) for post-training. This combination improved the models' reliability and domain-specific accuracy.

Chatbot Integration: The fine-tuned model was integrated into Anote's chatbot, enabling users to interact with a more reliable, domain-specific, and accurate LLM. The integration of Retrieval-Augmented Generation (RAG) with domain-specific fine-tuning of LLMs enabled accurate model outputs while ensuring data privacy.

SUMMARY

This evaluation reaffirms Anote's position as a leader in ethical AI innovation. By adhering to NIST's risk management framework and collaborating with organizations like Humane Intelligence, Anote is setting new standards for secure and responsible AI applications. Anote's participation in the red team evaluation underscores our commitment to building secure, ethical, and high-performing AI systems. Our platform empowers users across industries to leverage the power of large language models responsibly.



i feel like i love everyone or at least i am compassionate toward others

i am feeling brave enough

Label Text Data

Classify Text, Extract Entities, and Answer Questions on Documents with LLMs

Train Your Model

Train supervised or unsupervised models to run your Fine Tuned LLMs with our API

Hello, I am Anote's AI Assistant, how can I help you

summarize this document

This document outlines the process for creating a successful marketing campaign. It explains the importance of setting clear goals, researching the target audience, creating a budget, and developing a strategy. It also provides tips for creating effective content and measuring the success of the campaign.

Build Your Fine Tuned GPT

Chat accurately with your documents while keeping your data private and secure

https://anote.ai/ https://www.linkedin.com/company/anote-ai/ vidranatan@gmail.com https://www.youtube.com/@anote-ai