Benchmarking Question and Answering models

Importance of Benchmarking Q-A for Finance

Use Case

Financial specialists engage in time-consuming tasks of summarizing and analyzing information about companies to make informed investment decisions, develop financial strategies, and conduct due diligence. Large Language Models (LLMs) have the potential to enhance and automate labor-intensive aspects of financial analysis due to their impressive capabilities in natural language understanding, reasoning, and writing

Addressing Model Hallucination

Benchmarking is crucial as it reveals instances where models may hallucinate or fabricate answers, highlighting the importance of reliability assessments in financial question-answering.

Performance of Existing Models

Models we tested

GPT 3.5

Claude

LLAMA-2

GPT4ALL

Findings

- Good at general chatting tasks
- Performs poorly on higher-level reasoning and logic tasks
- Tends to hallucinate information

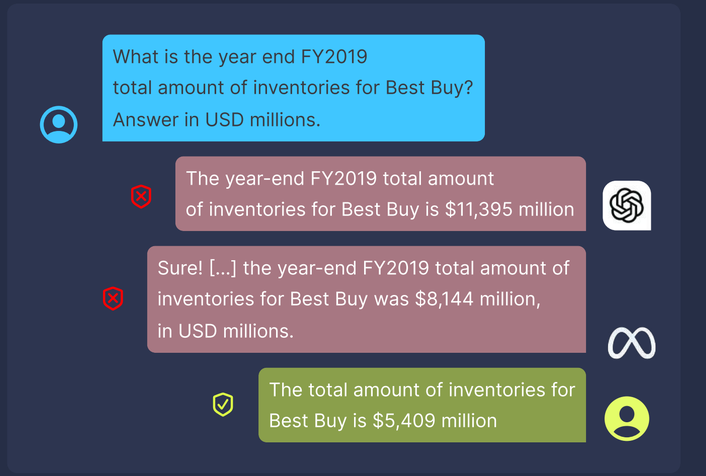


Figure 1: Incorrect model responses (using a shared vector store) to a question in FINANCEBENCH. The correct answer is given by the human expert.

FinanceBench

Dataset by Patronus AI

- Industry's first benchmark for testing how LLMs perform on financial questions.
- High quality, large-scale set of 10,000 question and answer pairs based on publicly available financial documents like SEC 10Ks, SEC 10Qs, SEC 8Ks, earnings reports, and earnings call transcripts.

The new benchmark spans to test several LLM capabilities in finance

Numerical reasoning

Finance metrics requiring numerical calculations, e.g. EBITDA, PE ratio, CAGR.

Informational retrieval

Specific details extracted directly from the documents.

Logical reasoning

Questions involving financial recommendations, which require interpretation and a degree of subjectivity

Logical reasoning

Questions involving financial recommendations, which require interpretation and a degree of subjectivity

Retrieval Augmented Generation (RAG) Instruct Benchmark Tester

Dataset by LLMWARE

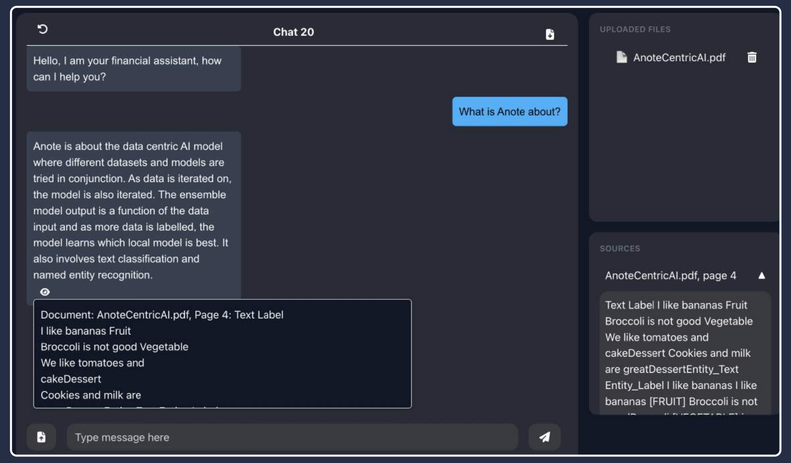
- Designed for professionals in legal and financial industries. - Invaluable for evaluating RAG technology in enterprise use cases.
- Context passages from common retrieval scenarios, including financial news, earnings releases, contracts, invoices, technical articles, general news, and short texts.

The new benchmark spans to test several LLM capabilities in finance

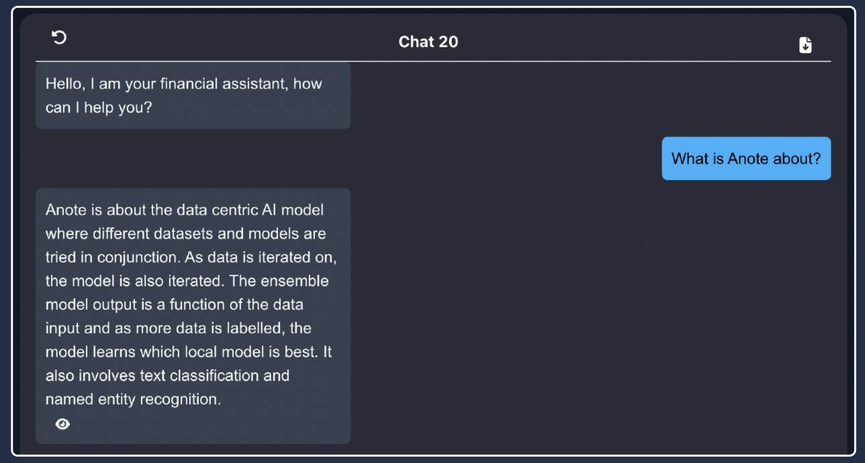
Financial Table Reading	**Core Q&A Evaluation**	**Classify Not Found Topics**	**Apply Boolean Yes/No Principles**
Solve Deep Math Equations	**Explore Complex Q&A Inquiries**	**Summarize Core Principles**	

Benchmark Q-A Process

We looked at



Answers the model provides



The chunk within the document which the model chooses to get it’s answer from

Evaluation Methods

Cosine Similarity

Measures the cosine of the angle between vectors representing model-generated and reference text.

Rouge - L Score

Measures the overlap of the longest common subsequence between the model's output and reference.

Human Evaluation

Human evaluators provide subjective scores based on factors such as fluency, coherence, and informativeness.

Trade-off between Reliability and Scalability

Reliability

- Human evaluation offers nuanced insights but can be resource-intensive.
- Cosine Similarity and Rouge scores provide automated, reliable measures but may lack the depth of human judgment.

Scalibility

- Automated metrics like Cosine Similarity and Rouge allow large-scale evaluations.
- Human evaluation, while detailed, can be challenging to scale due to time and resource constraints

How can we make models better?

Subtleties in Quantitative Analysis

Ex: "The revenue was $23.7B" vs. "The revenue was $2.37B."

Strong Cosine Similarity and Rouge Score but vastly different connotations, showcasing the difficulty in capturing nuanced numerical variations during evaluation.

Subtleties in Qualitative Analysis

Ex: “The company’s performance with impressive”

“Impressive” can be subjective, making it challenging to quantify and standardize during model evaluation

How can we make models better?

Prompt-Engineering and Re-Prompting

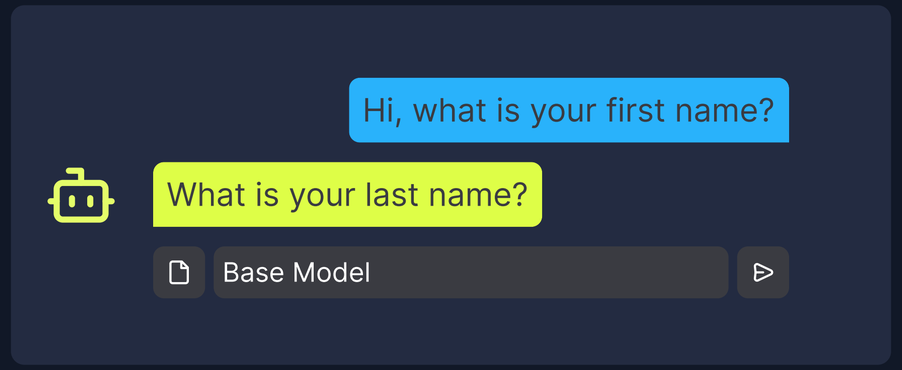
Retrieval Augmented Generation

Finetuning

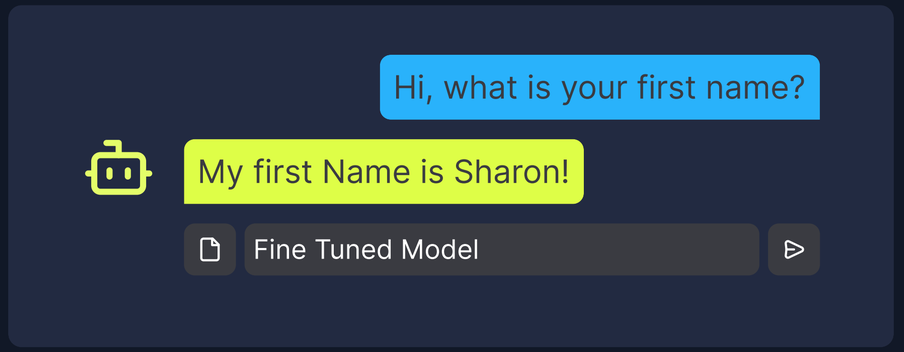
Why should you finetune?

- → More Consistent Outputs
- → Customize models for specific use cases
- → Reduces Hallucinations
- → Eliminates need of training a model from scratch

Base Model



Finetuned Model



Methods of Finetuning

Transfer Learning

Self Supervised Learning

Supervised Learning

RLHF