Benchmarking Q-A models

Katherine Jijo

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

Model Understanding Drives Optimization: In-depth understanding of current models obtained through benchmarking highlights areas for iterative improvement, paving the way for targeted finetuning and optimization.

Performance of Existing Models

Models we tested:

- GPT 3.5
- Claude
- LLAMA-2
- GPT4ALL

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

Dataset: FinanceBench

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:

1. Numerical reasoning: Finance metrics requiring numerical calculations, e.g. EBITDA, PE ratio, CAGR.
2. Information retrieval: Specific details extracted directly from the documents.
3. Logical reasoning: Questions involving financial recommendations, which require interpretation and a degree of subjectivity.
4. World knowledge: Basic accounting and finance questions that analysts are familiar with.

Dataset: Retrieval Augmented Generation (RAG) Instruct Benchmark Tester

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 benchmark spans to measure LLM capabilities in legal & financial services:

- 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

Evaluation Methods

1. **Cosine Similarity**
   a. Measures the cosine of the angle between vectors representing model-generated and reference text.
2. Rouge - L Score
   a. Measures the overlap of the longest common subsequence between the model's output and reference.
3. Human Evaluation
   a. 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.
- **Scalability:**
  - 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.

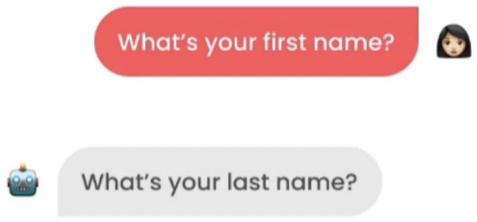
Performance

Insert overall performance graph here

How can we make models better?

1. Prompt-Engineering and Re-Prompting
2. Retrieval Augmented Generation
3. Finetuning

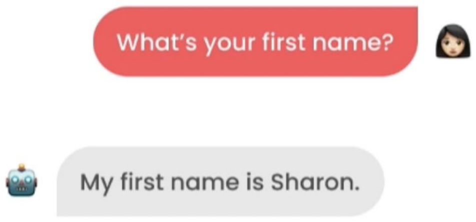
Why should you finetune?



What’s your first name?

What’s your last name?

Base Model



What’s your first name?

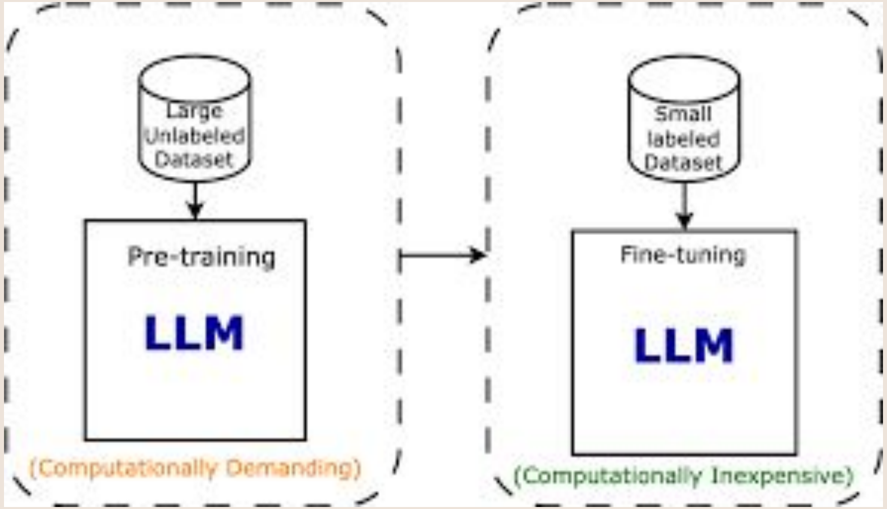
My first name is Sharon.

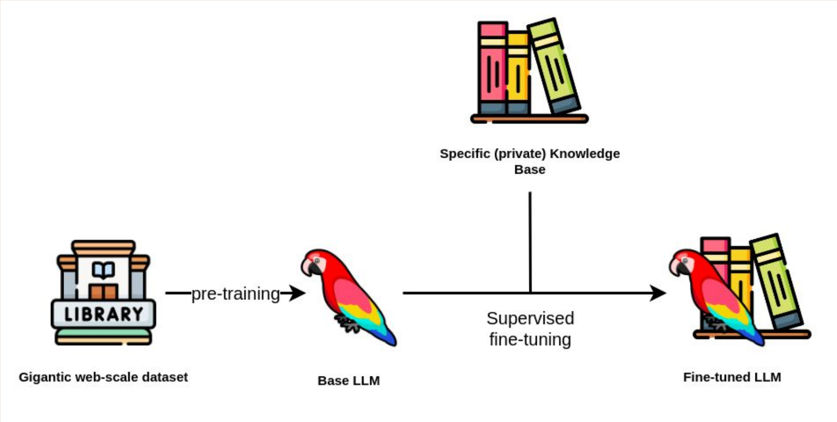
Finetuned Model

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

Methods of Finetuning

- Transfer Learning
- Self Supervised Learning
- Supervised Learning
- Reinforcement Learning with Human Feedback


Large Unlabeled Dataset
Pre-training
LLM
(Computationally Demanding)
Small labeled Dataset
Fine-tuning
LLM
(Computationally Inexpensive)


Specific (private) Knowledge Base
LIBRARY
pre-training
Supervised fine-tuning
Gigantic web-scale dataset
Base LLM
Fine-tuned LLM