



Improve AI models with
human feedback.



Meet the Founders of Anote

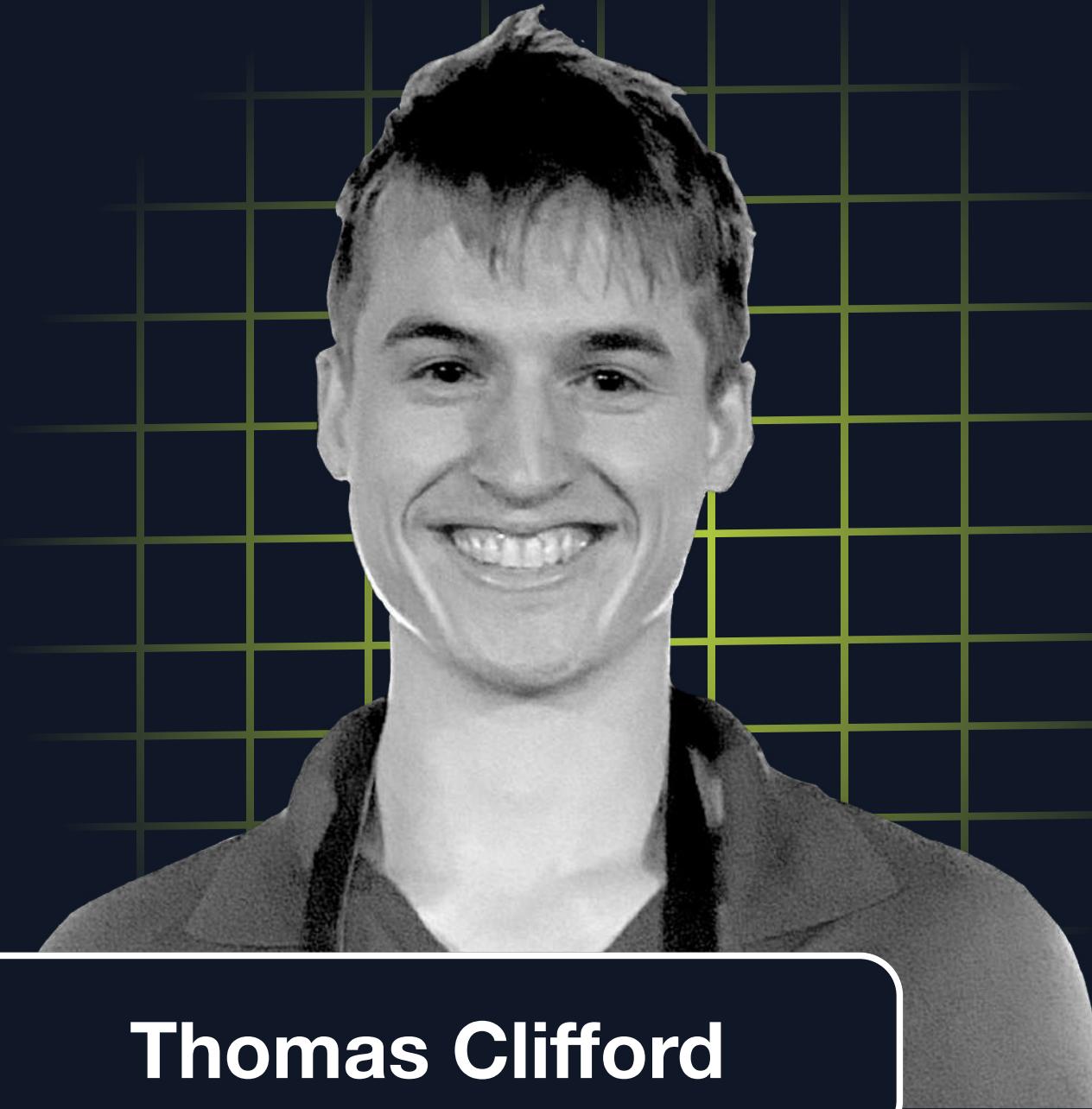


Natan Vidra

Data Scientist at Deloitte Applied AI

Electrical & Computer Engineering | Computer Science @
Cornell University

Deloitte.



Thomas Clifford

Software Engineer at Google

Financial Engineering | Computer Science @
Washington University

Google



The Problem

Enterprises are looking to **classify text**,
extract entities and **answer questions** on
unstructured text data for analytics purposes

Use Cases



Classification

Navy | government

Millions of classified documents that are bucketed to +200 CUI **categories**. 500 annotators go through these documents by hand to manually bucket this data

Navy released the CUI challenge **in search for an automated AI solution**, as the process of marking CUI documents today is entirely manual



Extract Entities

Morningstar | finance

Millions of financial documents where +1200 financial analysts work to extract entities of a taxonomy of +1,500 categories for financial products.

There are some automated solutions for basic document types, but for **raw unstructured documents the process is entirely manual**.



Answer Questions

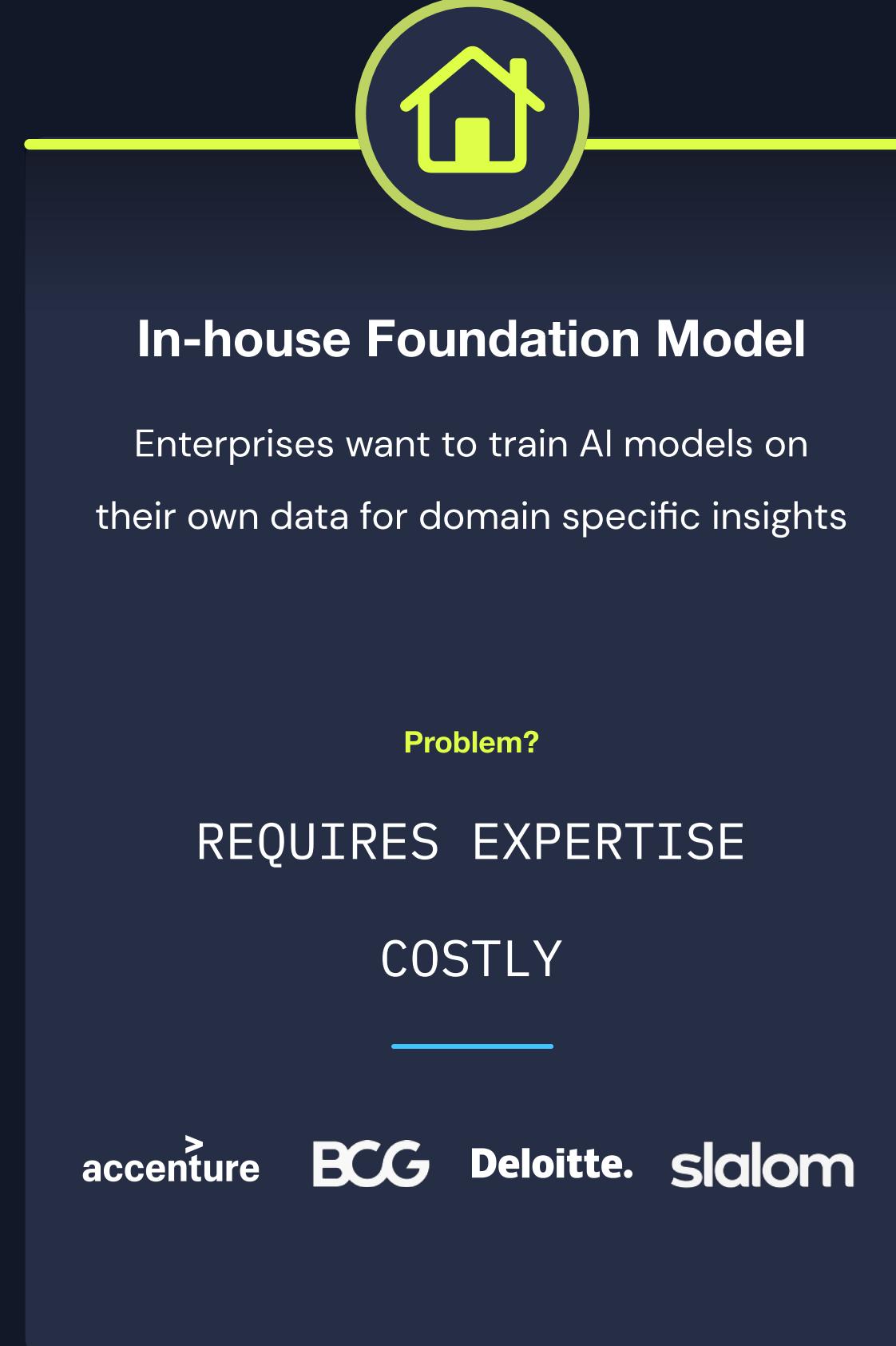
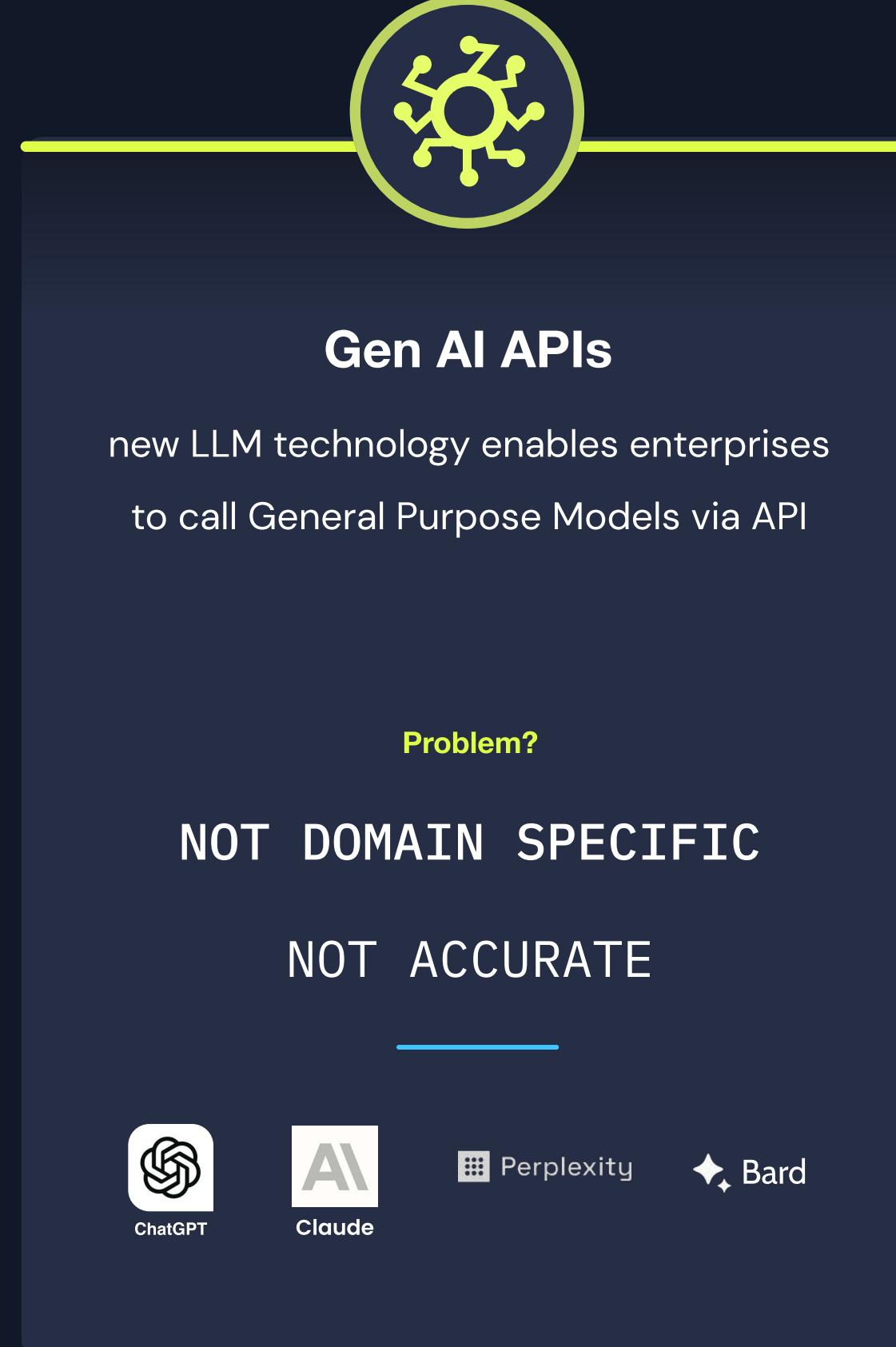
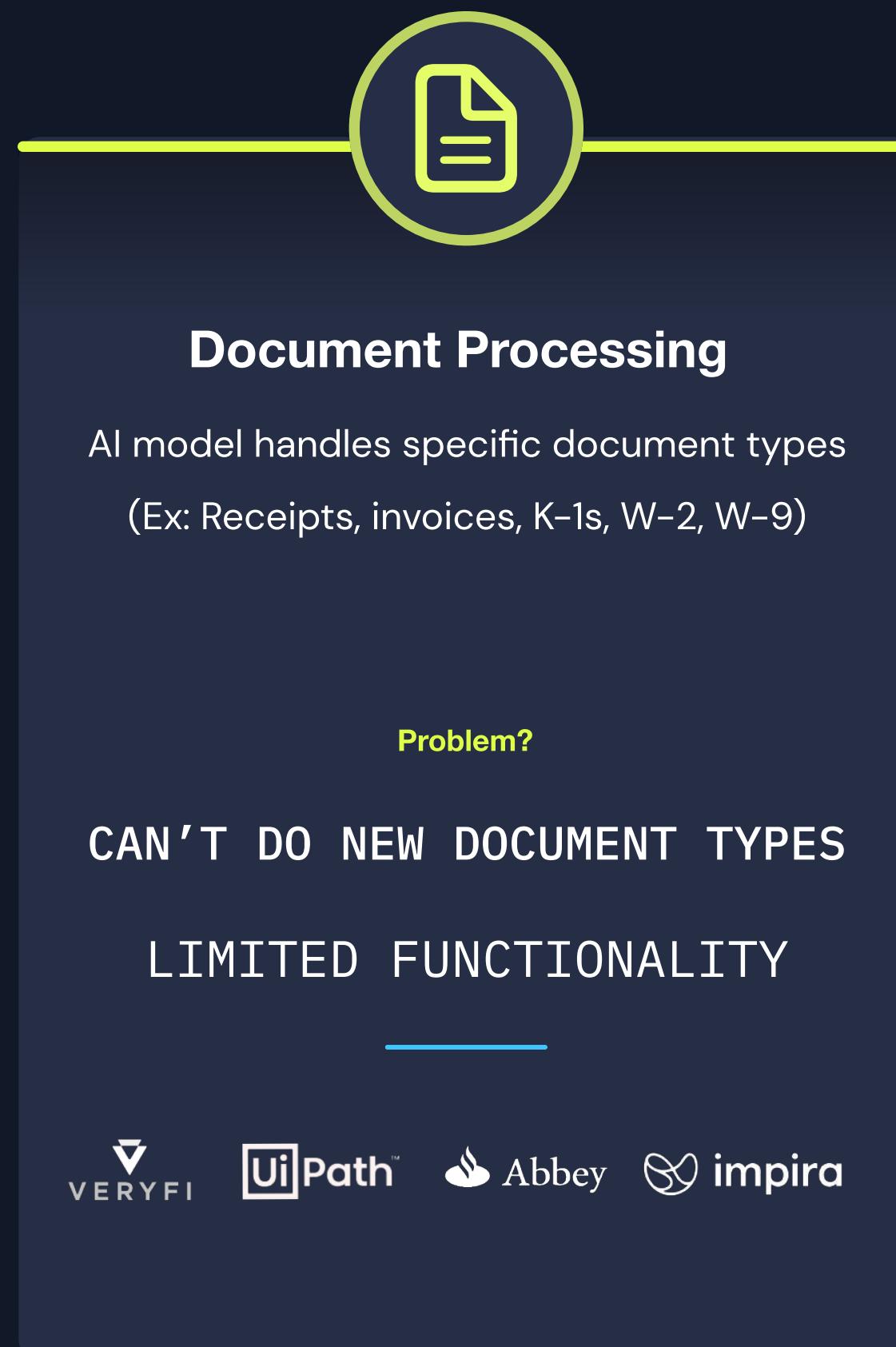
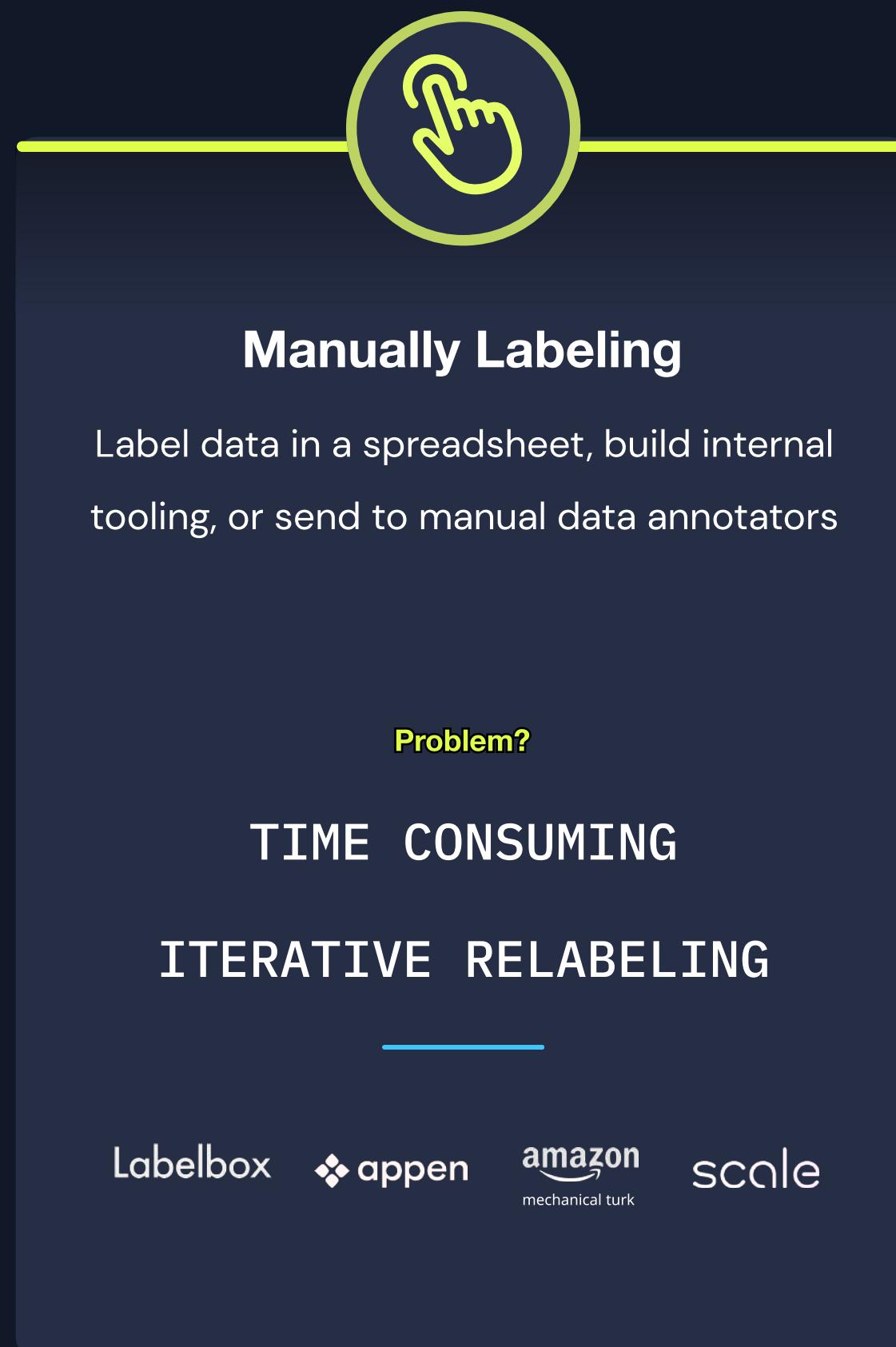
Weill Cornell | medical

Millions of medical charts are verbose and need summarization. 200 medical experts are paid to write tailored summaries for each of these medical charts.

This is **very tedious, and takes 15 minutes average** for 4th year medical students to write a summary for each specific medical chart.

What's in the Market

Currently there are a few approaches to solve this problem:



As a result, many enterprises have been looking for solutions to **fine tune or re-prompt large language models** on their own data, for tailored, accurate and domain specific results.

Our Solution



State of the art few shot learning to make high quality predictions with a few labeled samples.

The screenshot displays the Anote platform's interface for sentiment analysis. On the left, a sidebar lists various NLP tasks: Sentiment Analysis, Text Classification, Document Classification, Named Entity Recognition, Summarization, Semi-structured Prompting, and Question Answering. The main area shows a text snippet: "The cinematography in this film is **breathtaking**". Below this is an "Annotation History" section. A recent annotation for the text "Mediocre movie, nothing special." is shown as "negative". The history also includes annotations for "surprise", "fear", and "joy". At the bottom of the history section are buttons for "+ Add Category" and "Add Labeling Functions". To the right of the history is a search bar and a "Sentiment Analysis" panel with buttons for "positive" (highlighted with a cursor), "negative", and "neutral". Further right are buttons for "Confirm", "Skip →", and "Marked as Important" (with a checkbox for "breathtaking"). The status bar indicates "4 Annotated" and "1 Remaining". The bottom right corner shows a coverage metric of "0.2%" and navigation controls.

Upload

Create a new text based dataset.

The 'Create Dataset' screen displays six cards:

- Upload Unstructured**: Supports PDFs, DOCX, TXTs, and other file types.
- Upload Structured**: Supports labeled, tabular data in a CSV.
- Connect To Datasets**: Supports connections like Twitter and Reddit.
- Scrape from Website**: Can scrape HTML data from websites like Wikipedia.
- Select From Dataset Hub**: Connects to datasets on the Hugging Face hub.
- Load My Datasets**: Loads existing datasets or tutorial projects.

Customize

Add the categories, entities or questions you care about

The 'Customize' screen shows a sidebar with navigation steps: Upload, Customize, Annotate, Download. The main area has tabs for TEXT, Mislabeled, Dashboard, Predict, and Download. It lists categories and their rules:

- sadness ↑
IF humiliated THEN sadness ↑
- anger ↑
IF grouchy THEN anger ↑
- love ↑
- surprise ↑
- fear ↑

Below is a text analysis section with examples:

- i didnt feel humiliated → sadness
- i can go from feeling so hopeless to so damned hopeful just from being around someone who cares and is awake
- im grabbing a minute to post i feel greedy wrong
- i am ever feeling nostalgic about the fireplace i will know that it is still on the property
- i am feeling grouchy → anger
- ive been feeling a little burdened lately wasnt sure why that was

At the bottom, there's a coverage chart and a page navigation bar.

Annotate

As you annotate a few edge cases, the model actively learns to predict the rest.

The 'Annotate' screen shows a Jira Tickets interface with a sidebar:

- Update installation guide documentation**
- Jira Tickets**: Categories: Bug, Feature, Task, Improvement, Documentation (selected).
- Marked as Important**: None
- Annotation History**: 22 Annotated, 3 Remaining
- Stability**: 63.238 %

Download

Download the resulting labels as a CSV. Export fine tuned model as an API endpoint

The 'Download' screen shows a table of predicted labels:

TEXT	ACTUAL	PREDICTED	PROBABILITY
Update user manual documentation	Documentation	Task	0.5912
Improve search functionality speed	Improvement	Documentation	0.9732
Create automated build process	Feature	Improvement	0.7014
Implement social media sharing feature	Task	Feature	0.6314
Fix permission issue in the admin panel	Bug	Task	0.9617
Implement multi-language support	Task	Bug	0.9947
Updates API reference documentation	Documentation	Task	0.9331
Fix performance degradation issue	Bug	Documentation	0.9822
Create user management module	Feature	Bug	0.6959
Improve error handling mechanism	Bug	Bug	0.7121
Optimize memory usage for better scalability	Feature	Feature	0.9699

Buttons include Mislabel, Dashboard, Predict, and Download.

Product and Tech

Fix permission issue in the admin panel

The screenshot shows a Jira Tickets interface with a sidebar for classification: Block, Feature, Task, Improvement, and Documentation. A message at the top says "Marked as Important" with a "None" option. Below it, "23 Annotated" and "2 Remaining" are shown. A progress bar indicates "Stability: 62.364 %". At the bottom, there are buttons for "Update installation guide documentation" and "Documentation".

Label Text Data

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

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

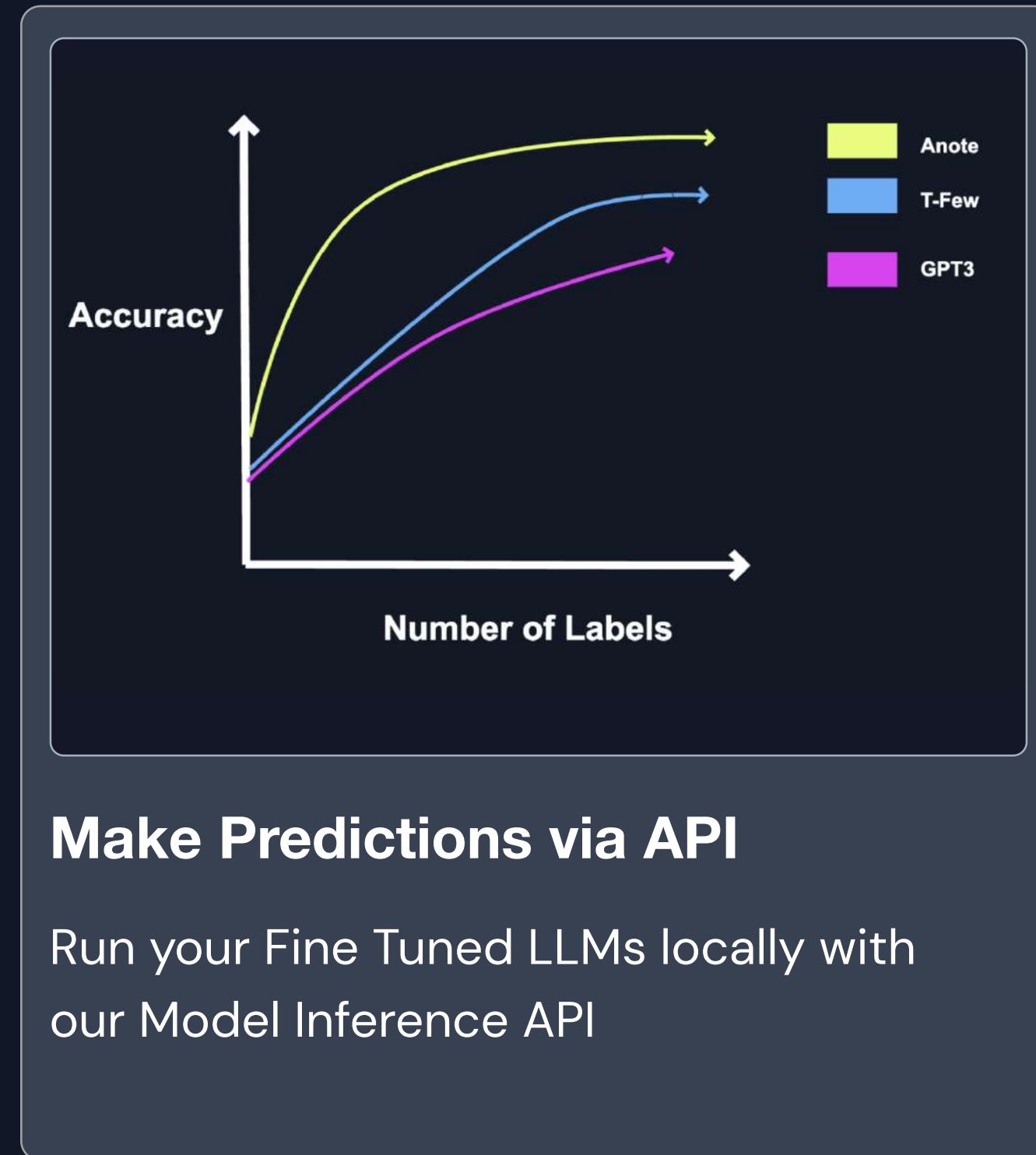
summarize this document

A message box displays a summary of a marketing campaign document. The summary includes: "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."

Type message here

Build Your Own Private GPT

Chat with your documents with LLMs while keeping your data private/secure



Differentiators



Active Learning

More accurate and tailored category predictions, entities and answers to questions, because AI models actively learns and rapidly improve from subject matter experts



Comprehensive Capabilities

Sentiment analysis, text classification, document classification, named entity recognition, semi-structured prompting, question answering, summarization etc.



Private Version

On premise enterprise-grade solution using Llama2 and GPT4All to leverage Gen AI and LLMs on your unstructured documents while keeping your data local, private and secure.

Why it Matters

Data labeling

Document Processing

Before

Manually Labeling their data themselves in a spreadsheet

Tedious, Time Consuming, Costly

Manual Iterative Relabeling

After

State of the art few shot learning to make high quality predictions with a few labeled samples.

Less time, less expensive, higher accuracy

Rapid flexibility for changing business requirements



Given a raw unstructured documents, such as a 10-k or earnings call transcript, you can't get answers to the questions right if trying to extract info, where accuracy really matters.

After a few interventions, we go from 10 questions right, to 15 questions right, to 20 questions right, to enable insights that were otherwise impossible to obtain.

Not accurate and largely manual extraction

Sub-optimal analytics for critical business decisions

Higher accuracy for raw unstructured documents

New insights that otherwise were not obtainable

Current Status

\$50k

in revenue

15

pilot projects

+30m

client rows
labeled

Next Steps

We are looking to achieve \$1M in revenue and obtain 15 enterprise clients, focusing on data teams within the finance industry.

>>>



Enhancing LLM Performance To Answer Questions and Extract Information More Accurately

[**Link to the Paper**](#)

The Abstract

Large Language Models (LLMs) generate responses to questions; however, their effectiveness is often hindered by sub-optimal quality of answers and occasional failures to provide accurate responses to questions.

To address these challenges:

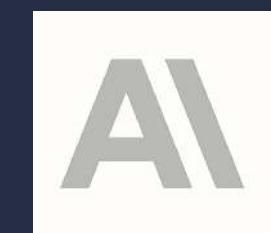
- Fine-tuning process is employed, involving feedback and examples to refine models.
- Combine fine-tuning the LLM with a process known as Retrieval Augmented Generation (RAG)

Performance of Existing Models

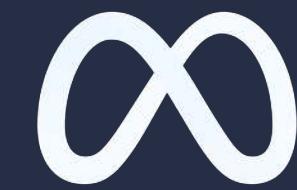
Models we tested



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

Findings

A screenshot of a user interface showing a question and four model responses. The question is: "What is the year end FY2019 total amount of inventories for Best Buy? Answer in USD millions." The first response (blue box) is correct: "The year-end FY2019 total amount of inventories for Best Buy is \$11,395 million". The second response (red box) is incorrect: "Sure! [...] the year-end FY2019 total amount of inventories for Best Buy was \$8,144 million, in USD millions.". The third response (green box) is incorrect: "The total amount of inventories for Best Buy is \$5,409 million". The fourth response (yellow box) is correct: "The total amount of inventories for Best Buy is \$5,409 million". Each response includes a small icon of a person with a checkmark or an X.

What is the year end FY2019 total amount of inventories for Best Buy?
Answer in USD millions.

✖ The year-end FY2019 total amount of inventories for Best Buy is \$11,395 million

✖ Sure! [...] the year-end FY2019 total amount of inventories for Best Buy was \$8,144 million, in USD millions.

✔ The total amount of inventories for Best Buy is \$5,409 million

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

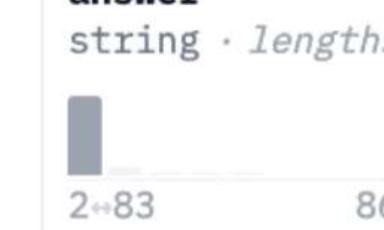
- Industry's first benchmark for testing how LLMs perform on financial questions.
- High quality, large-scale set of 10,000 question and answer pairs based on publicly available financial documents like SEC 10Ks, SEC 10Qs, SEC 8Ks, earnings reports, and earnings call transcripts.

financebench_id	doc_name	doc_link	doc_period	question_type	question	answer	evidence_text	page_number
21	100%	14~17	36.7%	157~167	3.3%	2.02k	2%	5~7
financebench_id_03029	3M_2018_10K	https://investors.3m.com/financials/sec-filings/content/0001558370-19-000470/0001558370-19...	2,018	metrics-generated	What is the FY2018 capital...	\$1577.00	Table of Contents 3M Company and Subsidiaries Consolidated Statement of Cash Flow s Years ended...	60
financebench_id_04672	3M_2018_10K	https://investors.3m.com/financials/sec-filings/content/0001558370-19-000470/0001558370-19...	2,018	metrics-generated	Assume that you are a public...	\$8.70	Table of Contents 3M Company and Subsidiaries Consolidated Balance Shee t At December 31 Decembe...	58
financebench_id_00499	3M_2022_10K	https://investors.3m.com/financials/sec-filings/content/0000066740-23-000014/0000066740-23...	2,022	domain-relevant	Is 3M a capital-intensive busines...	No, the company is managing its CAPEX...	3M Company and Subsidiaries Consolidated Statement of Income Years ended December 31 (Millions, except...	48,50,52
financebench_id_01226	3M_2022_10K	https://investors.3m.com/financials/sec-filings/content/0000066740-23-000014/0000066740-23...	2,022	domain-relevant	What drove operating margin...	Operating Margin for 3M in FY2022 has...	SG&A, measured as a percent of sales, increased in 2022 when compared to the same period last year...	27
financebench_id_01865	3M_2022_10K	https://investors.3m.com/financials/sec-filings/content/0000066740-23-000014/0000066740-23...	2,022	novel-generated	If we exclude the impact of M&A,...	The consumer segment shrunk by 0.9%...	Worldwide Sales Change By Business Segment Organic sales Acquisitions Divestitures Translation Total...	25
financebench_id_00807	3M_2023Q2_10Q	https://investors.3m.com/financials/sec-filings/content/0000066740-23-000058/0000066740-23...	2,023	domain-relevant	Does 3M have a reasonably health...	No. The quick ratio for 3M was 0.96 by...	3M Company and Subsidiaries Consolidated Balance Sheet (Unaudited) (Dollars in millions, except per...	5
financebench_id_00941	3M_2023Q2_10Q	https://investors.3m.com/financials/sec-filings/content/0000066740-23-000058/0000066740-23...	2,023	domain-relevant	Which debt securities are...	Following debt securities...	Title of each class Trading Symbol(s) Name of each exchange on which registered Common Stock, Par...	1

RAG Instruct Benchmark Tester

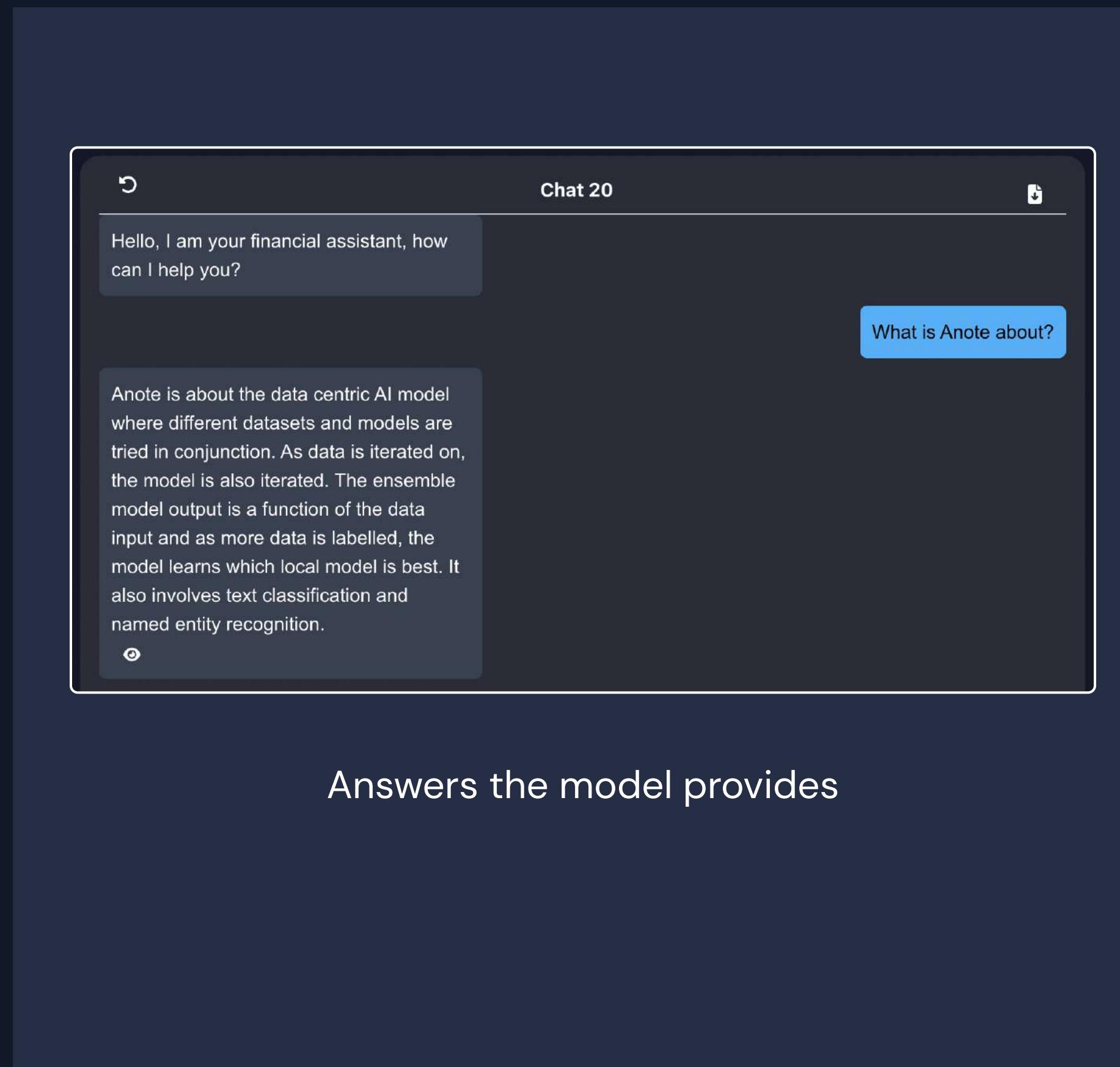
Dataset by LLMWARE

- 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.

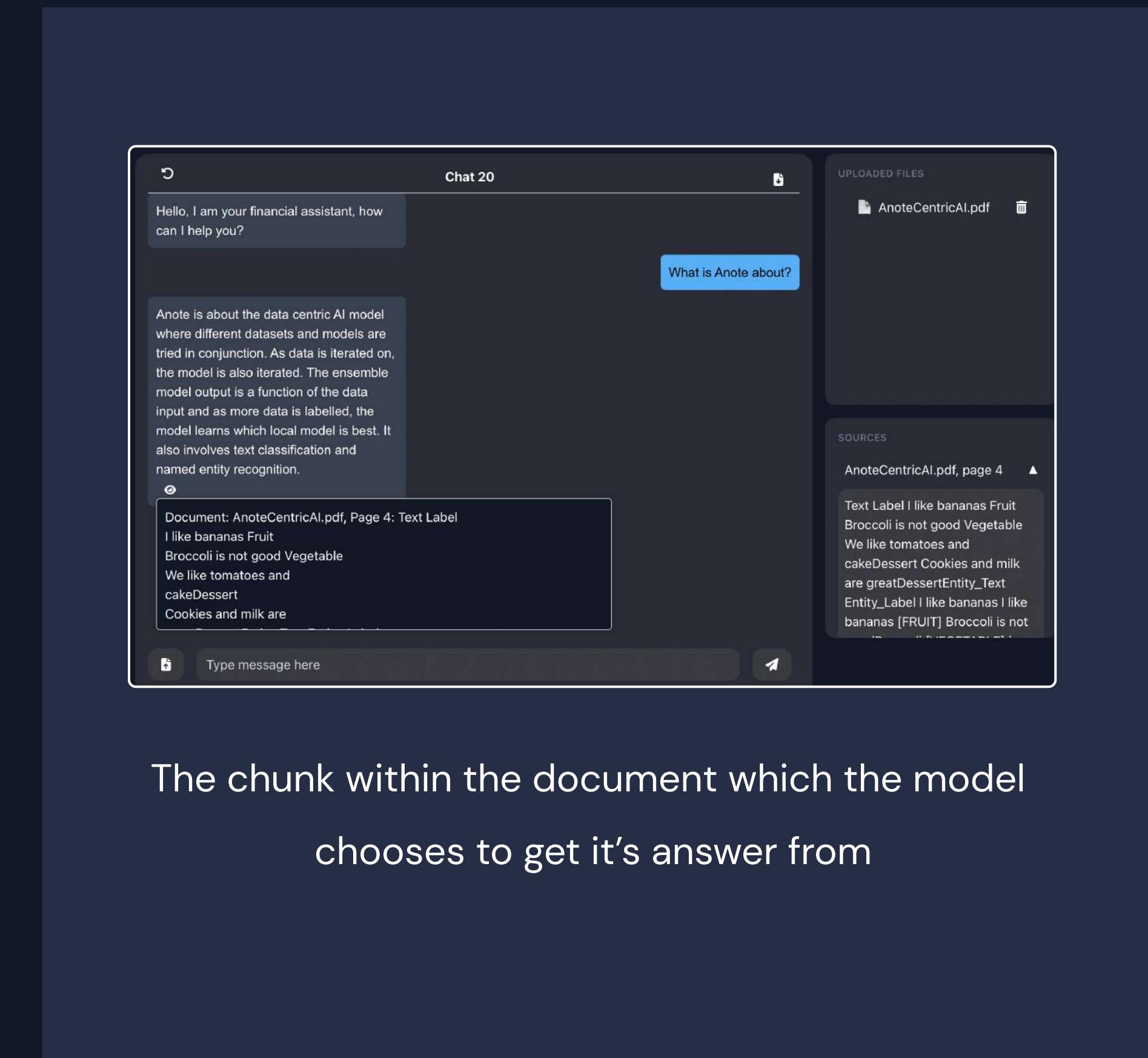
query string · lengths	answer string · lengths	context string · lengths
		
What is the total amount of the invoice?	\$22,500.00	Services Vendor Inc. 100 Elm Street Pleasantville, NY TO Alpha Inc. 5900 1st Street Los Angeles, CA Description Front End...
What is the invoice number?	#0001	Services Vendor Inc. 100 Elm Street Pleasantville, NY TO Alpha Inc. 5900 1st Street Los Angeles, CA Description Front End...
What is a list of the items being purchased?	•Front End Engineering Service; •Back End Engineering Service; •Quality Assurance Manager	Services Vendor Inc. 100 Elm Street Pleasantville, NY TO Alpha Inc. 5900 1st Street Los Angeles, CA Description Front End...
What is the name of the contact for question?	Bia Hermes	Services Vendor Inc. 100 Elm Street Pleasantville, NY TO Alpha Inc. 5900 1st Street Los Angeles, CA Description Front End...
What is the PO number?	#1000	Services Vendor Inc. 100 Elm Street Pleasantville, NY TO Alpha Inc. 5900 1st Street Los Angeles, CA Description Front End...
When is payment due?	within 30 days of 01/01/2022	Services Vendor Inc. 100 Elm Street Pleasantville, NY TO Alpha Inc. 5900 1st Street Los Angeles, CA Description Front End...
When is payment due?	May 3, 2022	Invoice DATE March 3, 2022 INVOICE NO 234 ABC Company 8675 River Run Road Marlinton, CT 09392 INVOICE TO Delta Inc. 872...

Benchmark Q-A Process

In addition to privacy and efficiency requirements 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

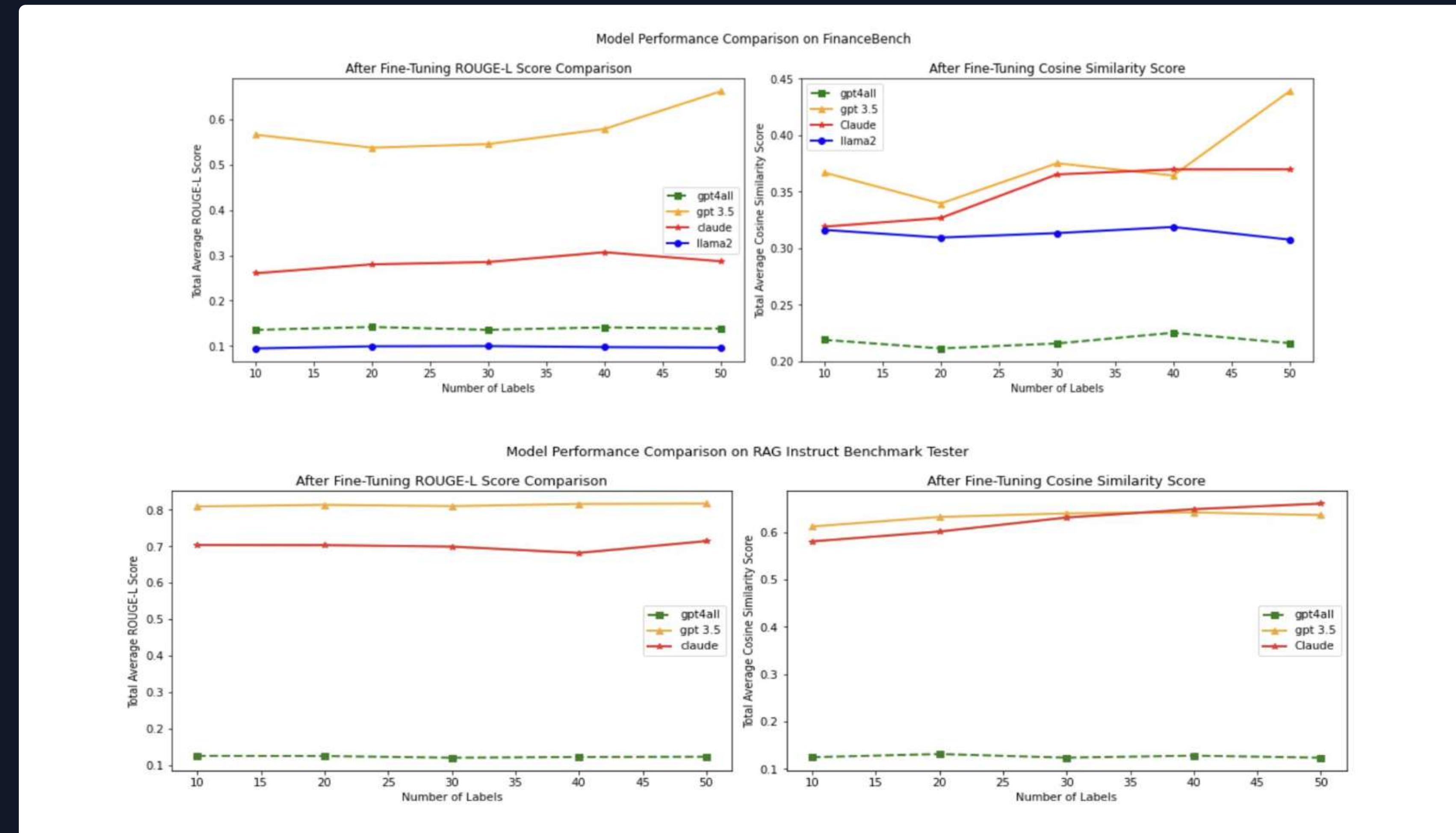
Reliability

- 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.

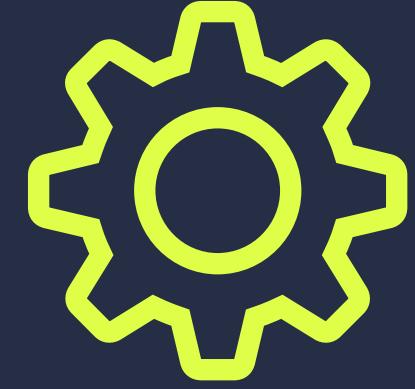
Scalability

- 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

Model Evaluation



How can we make models better?



Prompt-Engineering and
Re-Prompting



Retrieval Augmented
Generation

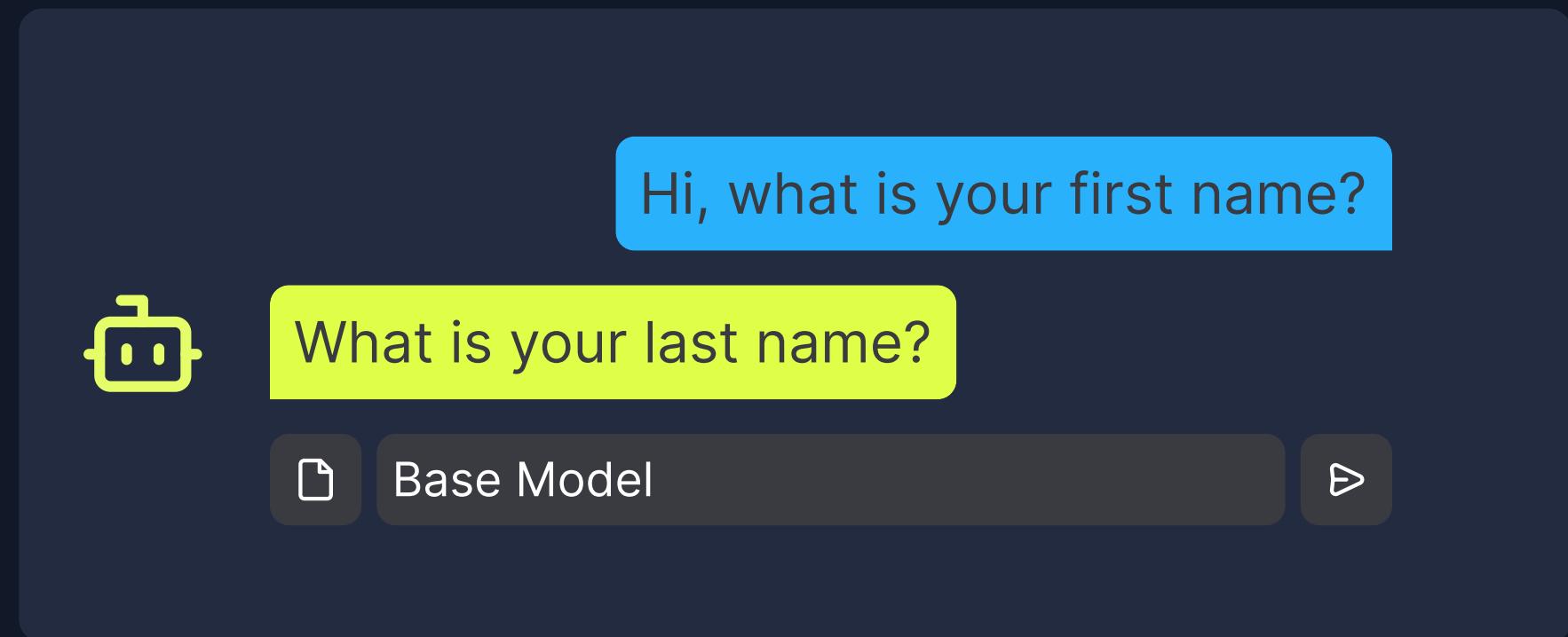


Finetuning

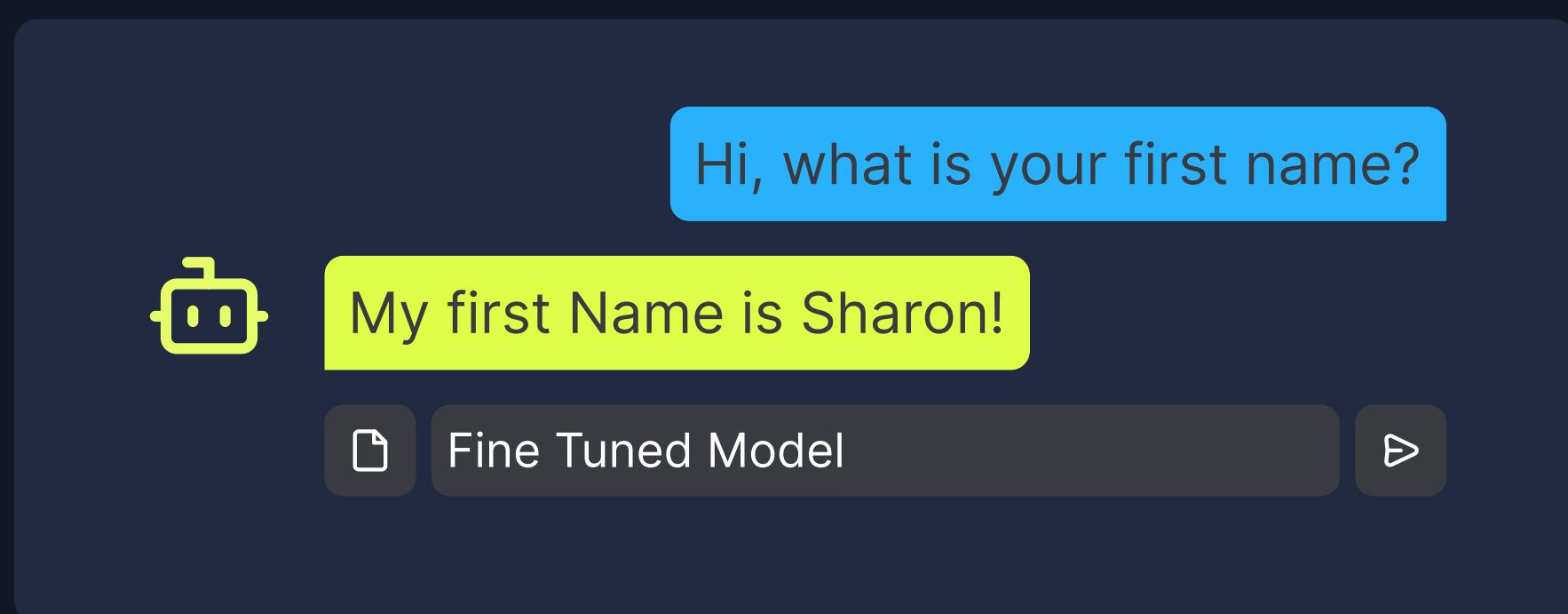
Why should you finetune?

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

Base Model



Finetuned Model



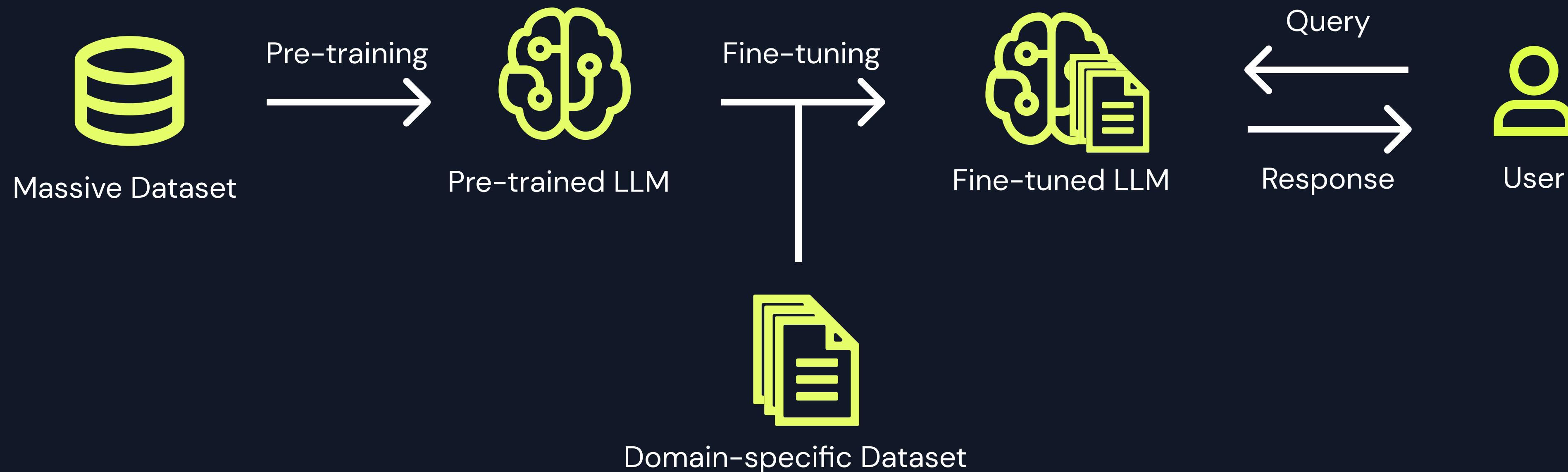
Methods of Finetuning

Self
Supervised
Learning

Supervised
Learning

Reinforcement
Learning with
Human
Feedback

Fine Tuning Architecture



Methods for Fine Tuning

Full Fine Tuning

Adjusts all parameters of the LLM using task-specific data.

Computationally expensive

Transfer Learning

Freeze all parameters except for the head of the neural network

Only finetune the layers that translate to the output layer

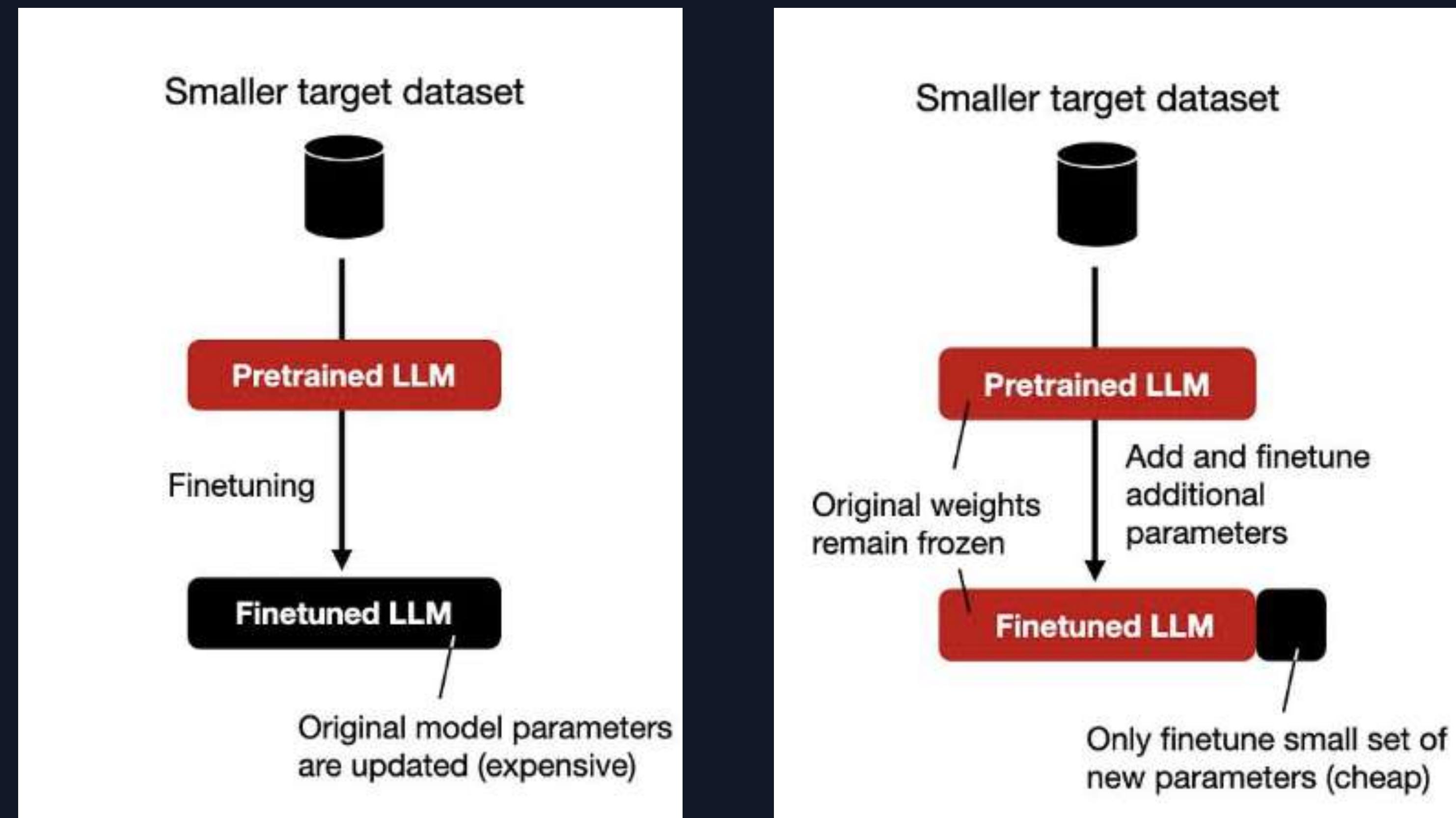
Parameter Efficient Fine Tuning

Freeze all the weights of the base LLM

Augment the model with additional parameters and finetune those

Less computationally expensive

Methods for Parameter Fine Tuning



Low Rank Adaptation (LoRA) for PEFT

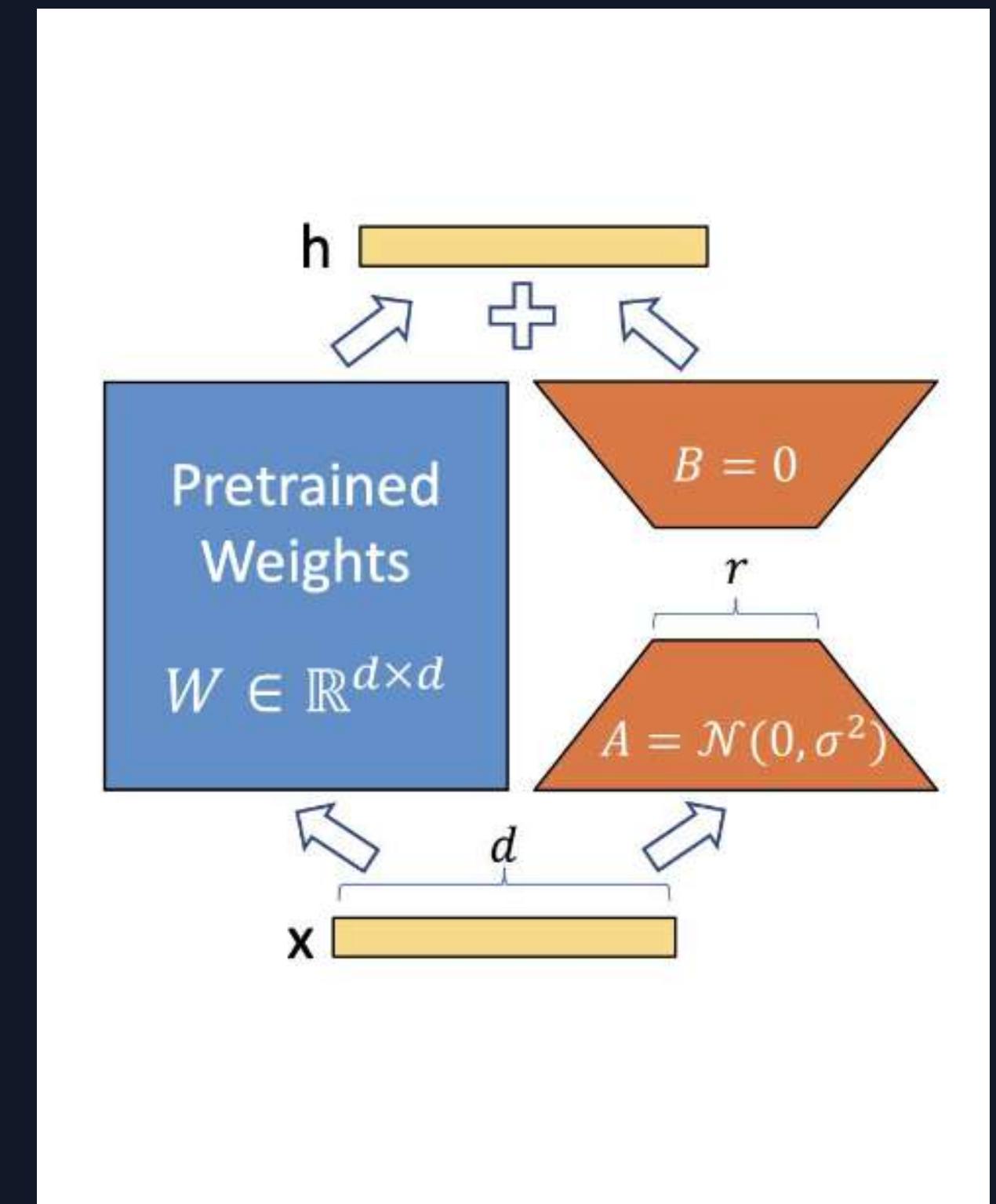
Reduces number of trainable parameters

Identifies crucial parameters for the task at hand and finetunes those

During fine-tuning, only the parameters in low-rank matrices are updated

Less chance of overfitting since only a few parameters are updated

Reduces computational and memory requirements needed to fine-tune



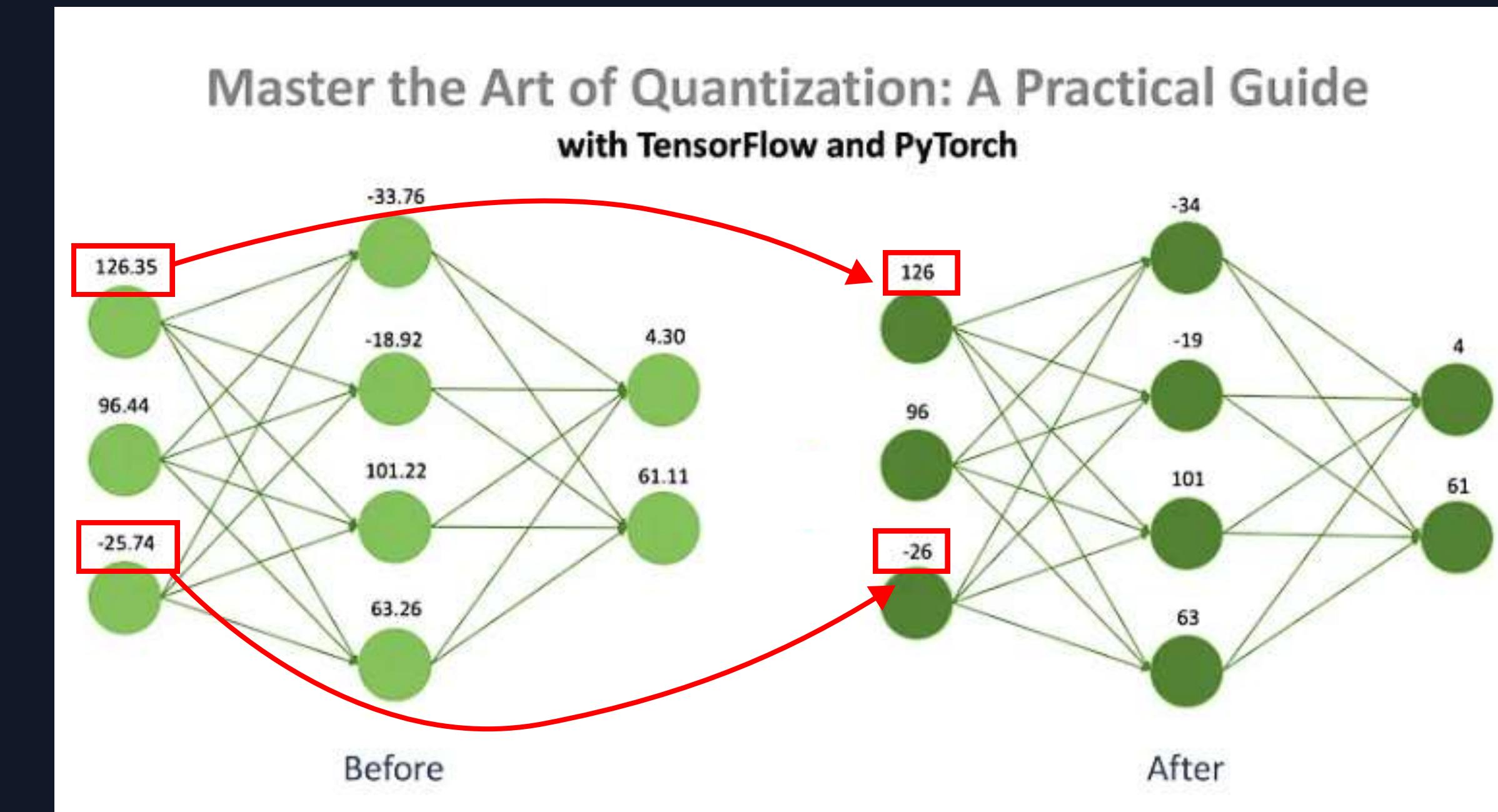
QLoRA for Quantized LLMS

Quantization: process of reducing the numerical precision of a model's tensors to make it faster and more compact

QLoRA combines quantization & low-rank adaptation

Model parameters are first quantized (usually to 4 bit precision) and then go through LoRA

Made fine tuning a lot more accessible



Fine Tuning Results on Finance-Bench

Question: What is the FY2018 capital expenditure amount (in USD millions) for 3M? Give a response to the question by relying on the details shown in the cash flow statement.

Correct Answer: \$1577M

LLAMA-2 with RAG:

The FY2018 capital expenditure amount for 3M is not directly disclosed in the cash flow statement. Therefore, I don't know the answer to this question

Fine Tuned LLAMA-2:

Based on the information provided in the 2018 10-K filing, the FY2018 capital expenditure amount for 3M is \$1577M.

Retrieving Context for Q-A

RAG usually finds relevant context via vector search of the embeddings

Retrieval algorithm does not always find the place where the correct answer is

How to improve Retrieval?

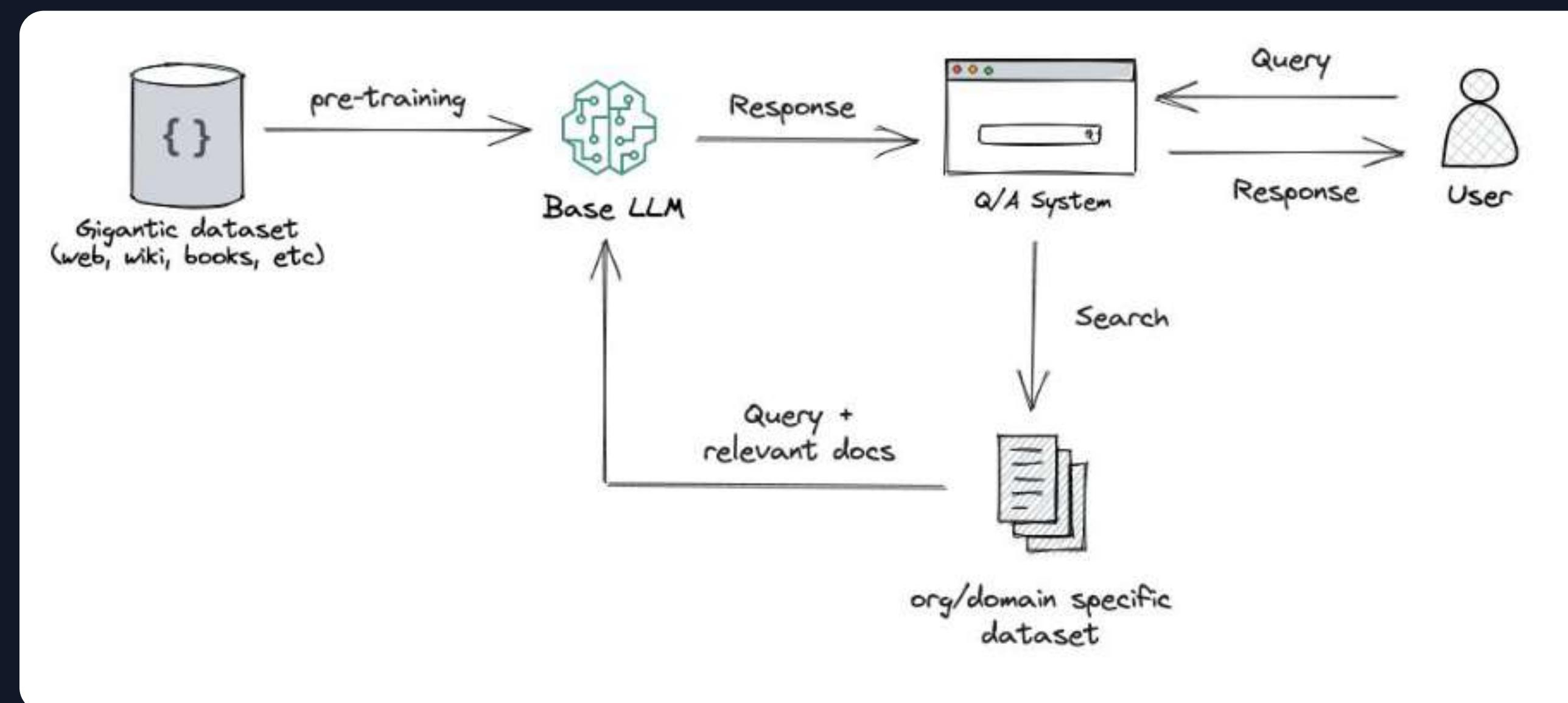
Query Expansion

Prompt Engineering

Re-ranking Algorithm

Chunking Parameters

Metadata Filtering



Thank you!

Join our **Slack Channel**

