Anote

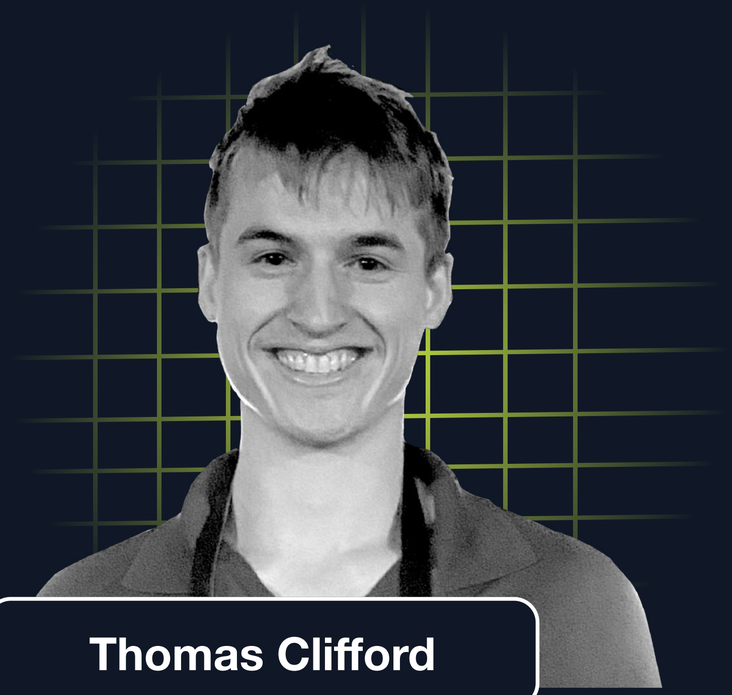
Improve AI models with human feedback.

Meet the Founders of Anote

Natan Vidra

Data Scientist at Deloitte Applied AI

Electrical & Computer Engineering | Computer Science @ Cornell University



Thomas Clifford

Software Engineer at Google

Financial Engineering | Computer Science @ Washington University

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:

Manually Labeling

Label data in a spreadsheet, build internal tooling, or send to manual data annotators

Problem?

TIME CONSUMING

ITERATIVE RELABELING

Labelbox · appen · amazon mechanical turk · scale

Document Processing

AI model handles specific document types (Ex: Receipts, invoices, K-1s, W-2, W-9)

Problem?

CAN'T DO NEW DOCUMENT TYPES

LIMITED FUNCTIONALITY

VERYFI · UiPath · Abbey · impira

Gen AI APIs

new LLM technology enables enterprises to call General Purpose Models via API

Problem?

NOT DOMAIN SPECIFIC

NOT ACCURATE

ChatGPT · Claude · Perplexity · Bard

In-house Foundation Model

Enterprises want to train AI models on their own data for domain specific insights

Problem?

REQUIRES EXPERTISE

COSTLY

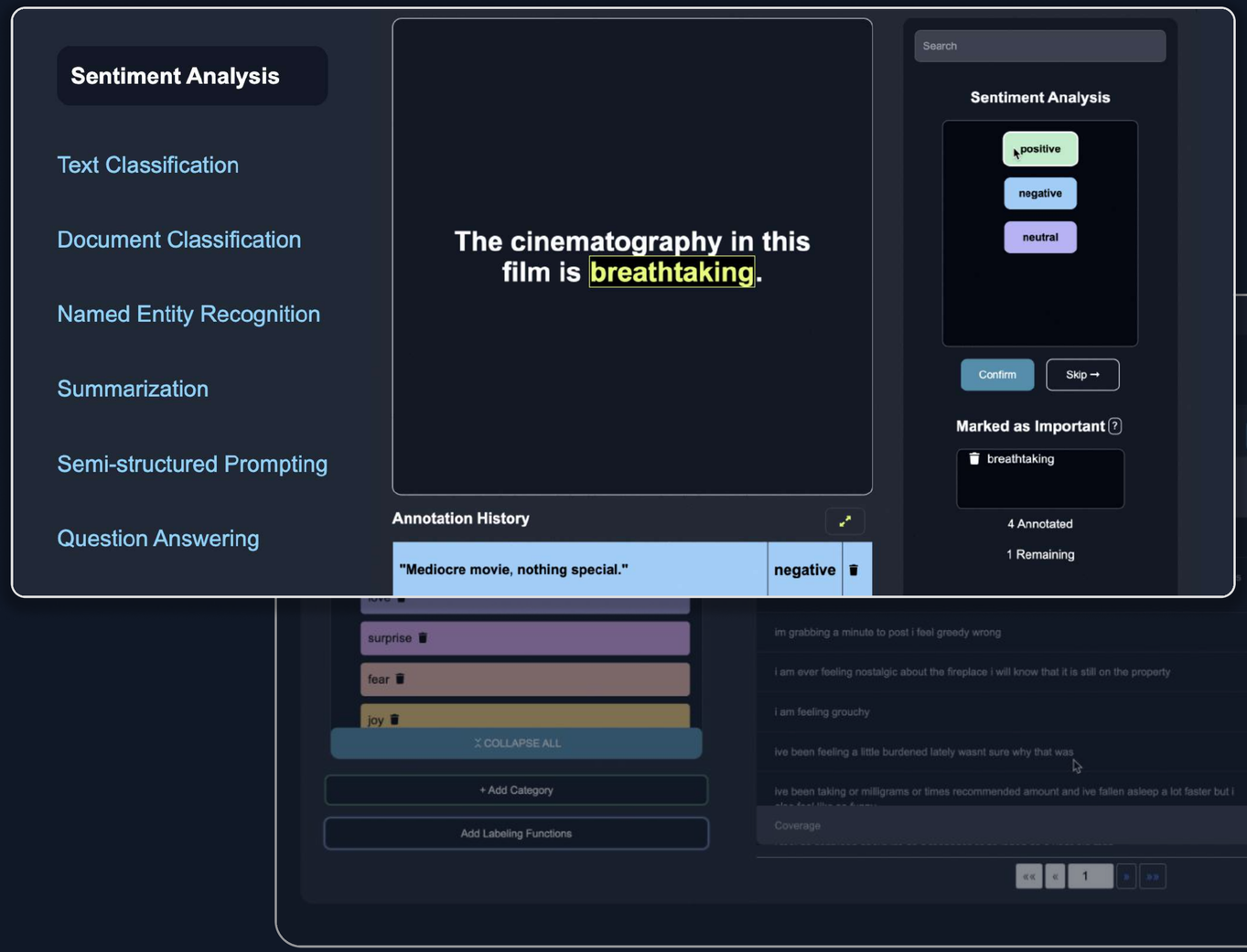
accenture · BCG · Deloitte. · slalom

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

Anote

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

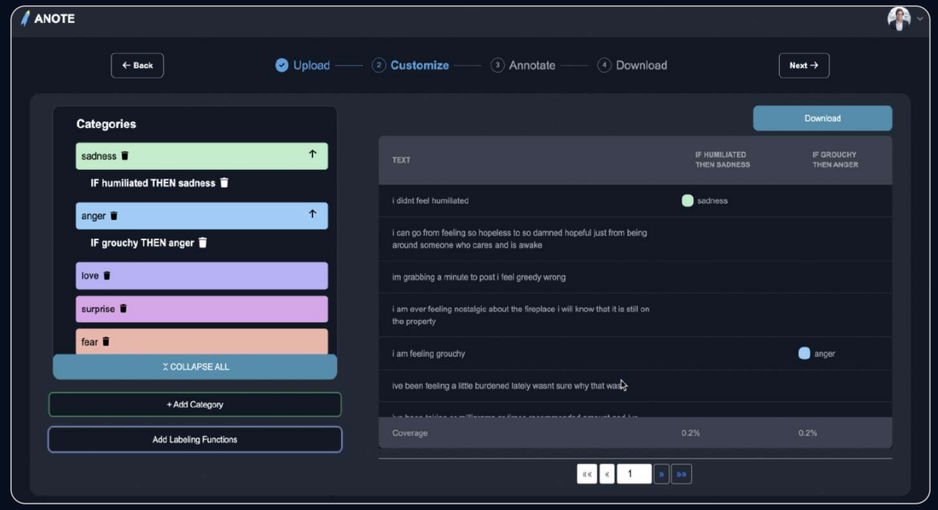


Upload

Create a new text based dataset.

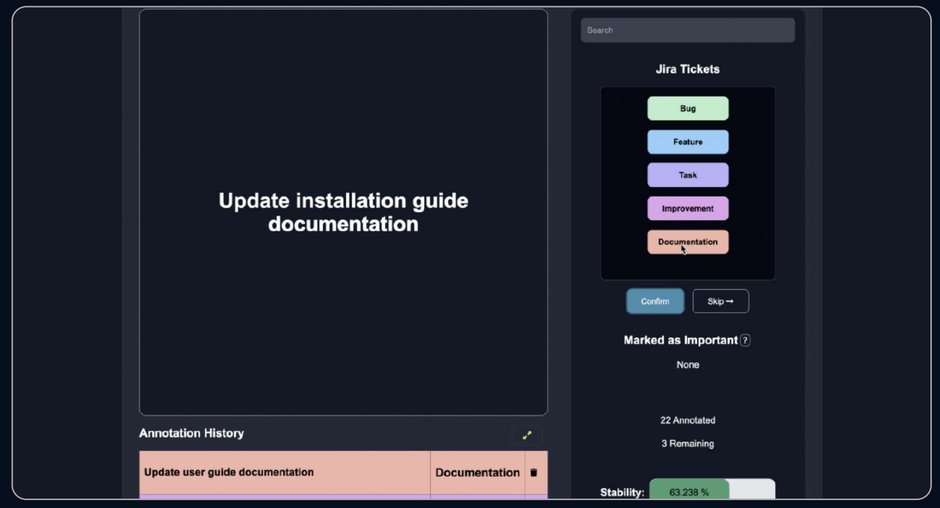
Customize

Add the categories, entities or questions you care about



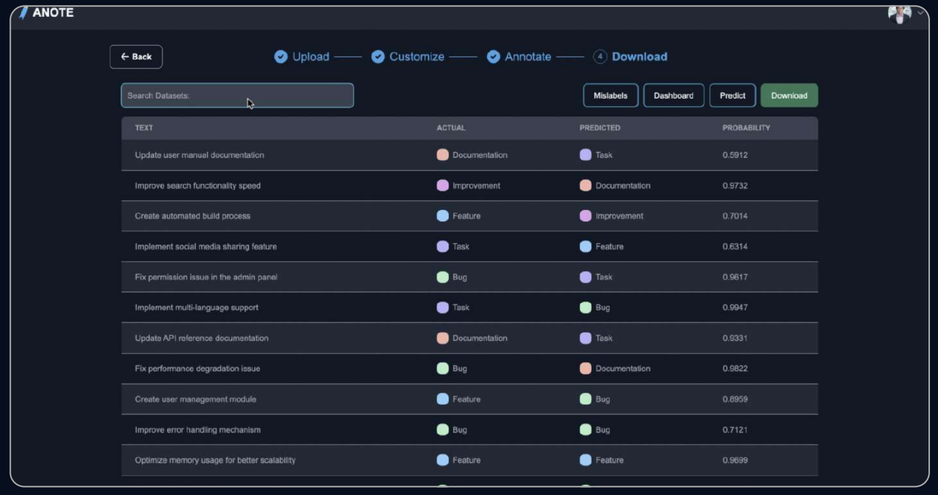
Annotate

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

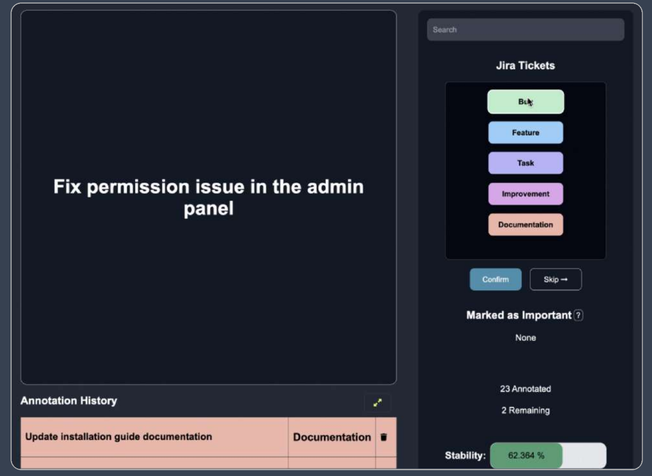


Download

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

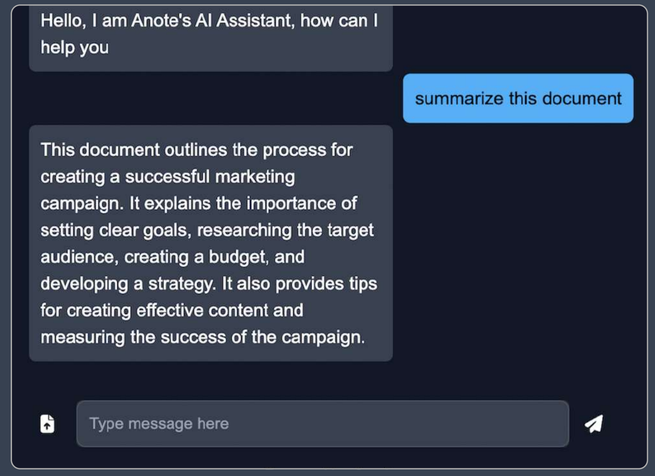


Product and Tech



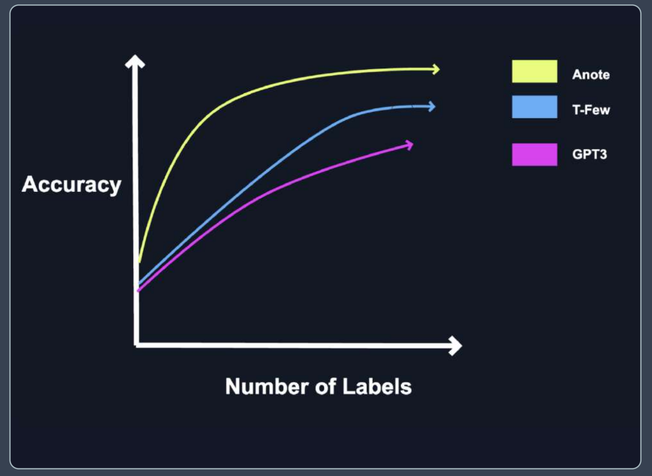
Label Text Data

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



Build Your Own Private GPT

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



Make Predictions via API

Run your Fine Tuned LLMs locally with our Model Inference API

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

	Before	After Anote
Data labeling	**Manually Labeling** their data themselves in a spreadsheet	State of the art few shot learning to make high quality predictions with a **few labeled samples.**
	Tedious, Time Consuming, Costly	Less time, less expensive, higher accuracy
	Manual Iterative Relabeling	Rapid flexibility for changing business requirements
Document Processing	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	Higher accuracy for raw unstructured documents
	Sub-optimal analytics for critical business decisions	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

Findings

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

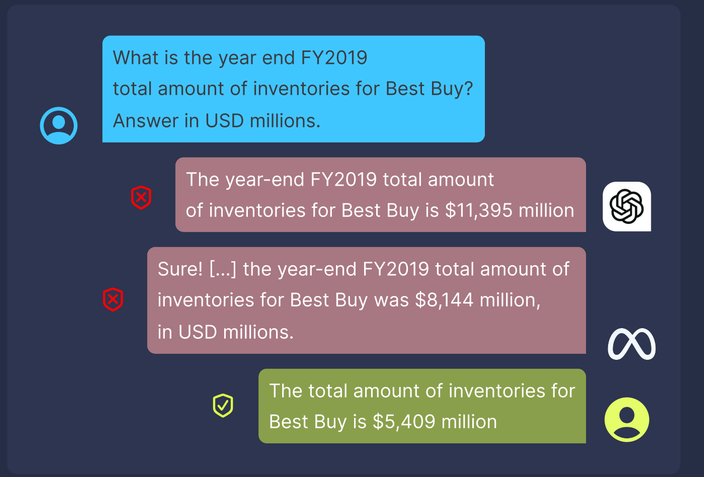


Figure 1: Incorrect model responses (using a shared vector store) to a question in FINANCEBENCH. The correct answer is given by the human expert.

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.

financebench_id string · *lengths* 21 100%	doc_name string · *lengths* 14 17 36.7%	doc_link string · *lengths* 157 167 3.3%	doc_period int64 2.02k 2%	question_type string · *classes* metrics-ge… 33.3%	question string · *lengths* 319 374 6.7%	answer string · *lengths* 1 62 53.3%	evidence_text string · *lengths* 11k 12.2k 0.7%	page_number string · *lengths* 5 7 18%
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, excep…	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

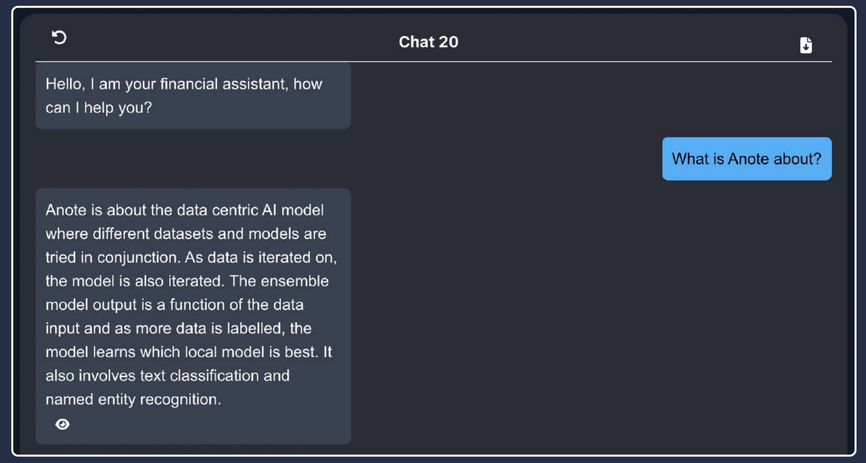
Dataset by LLMWARE

- 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.

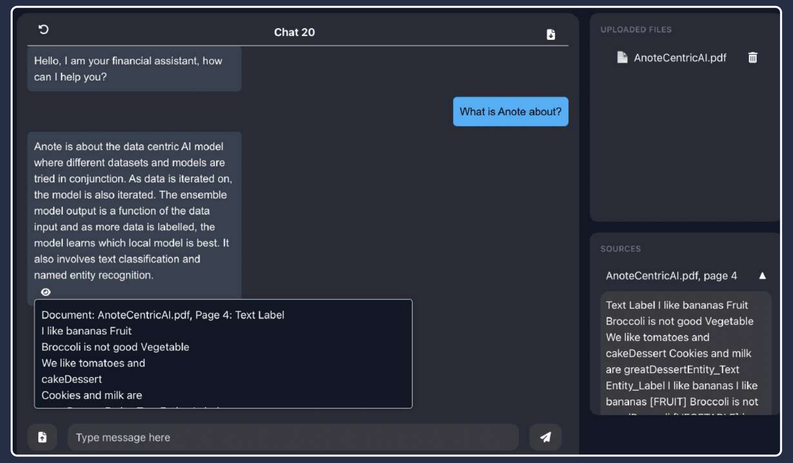
query string · *lengths* 18↔30 13%	**answer** string · *lengths* 2↔83 86.5%	**context** string · *lengths* 284↔537 23%
What is the total amount of the invoice?	$22,500.00	Services Vendor Inc. 100 Elm Street Pleasantville, NY TO Alpha Inc. 5900 1st Street Los Angeles, CA Description Front End…
What is 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 Marlington, 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

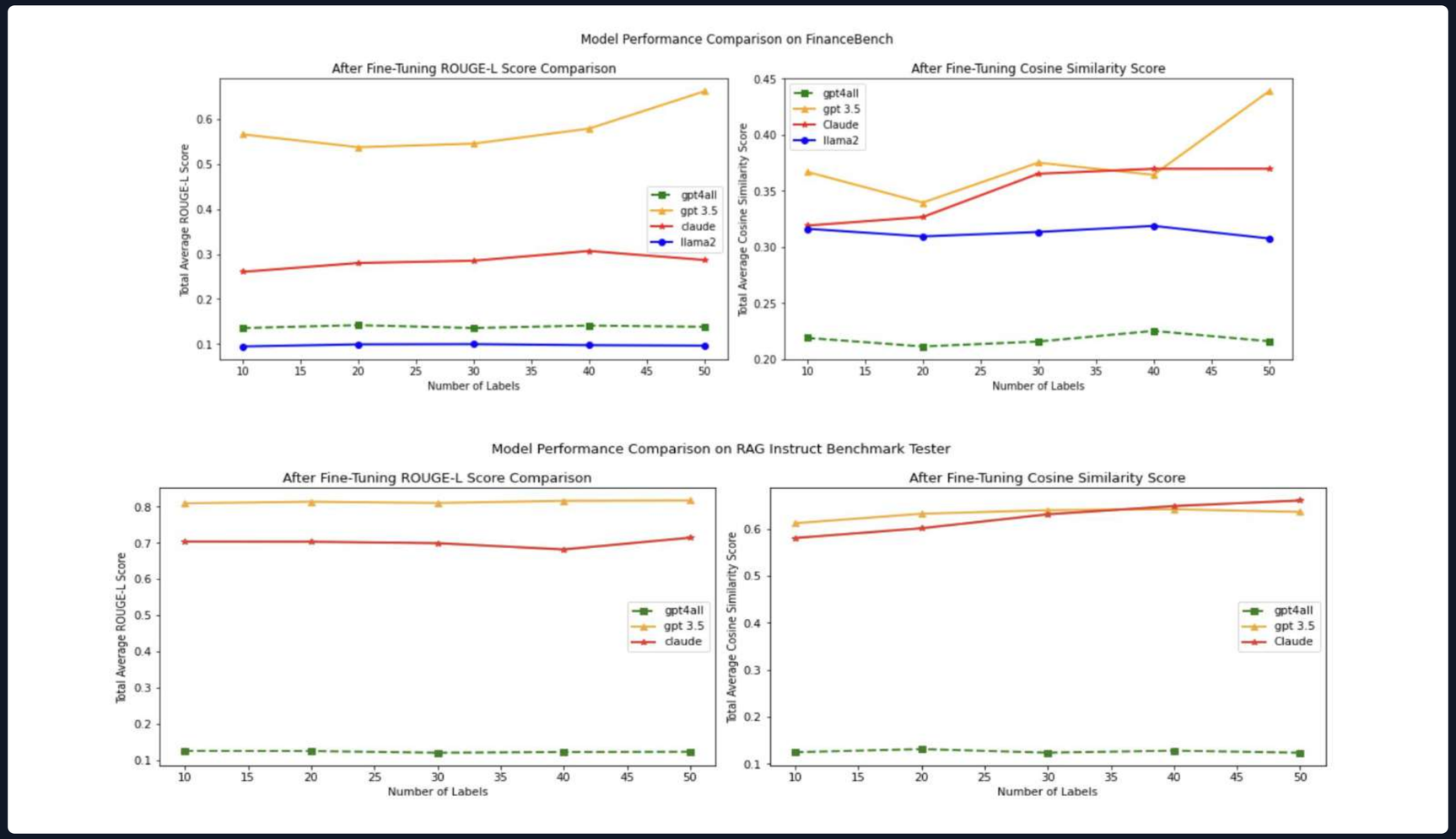
Reliability

- 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.

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

Model Evaluation



How can we make models better?

Prompt-Engineering and Re-Prompting

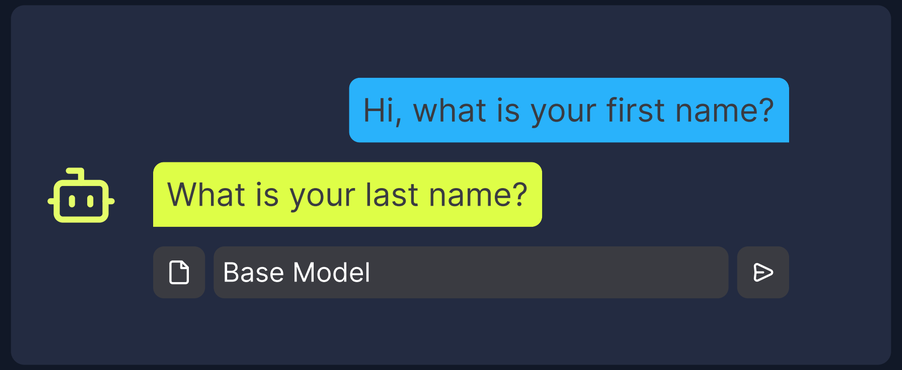
Retrieval Augmented Generation

Finetuning

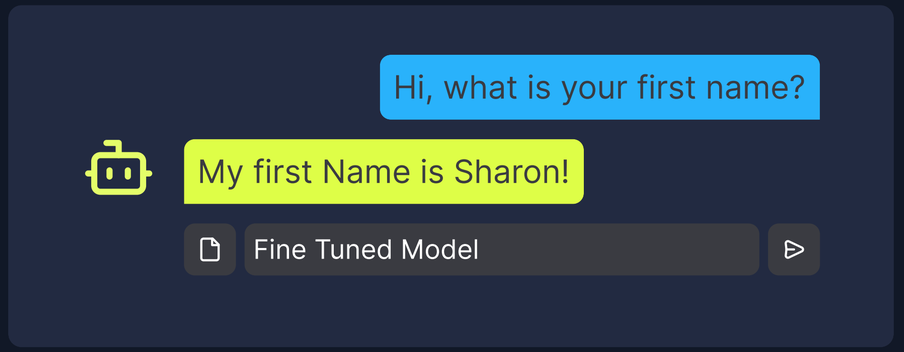
Why should you finetune?

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

Base Model



Finetuned Model



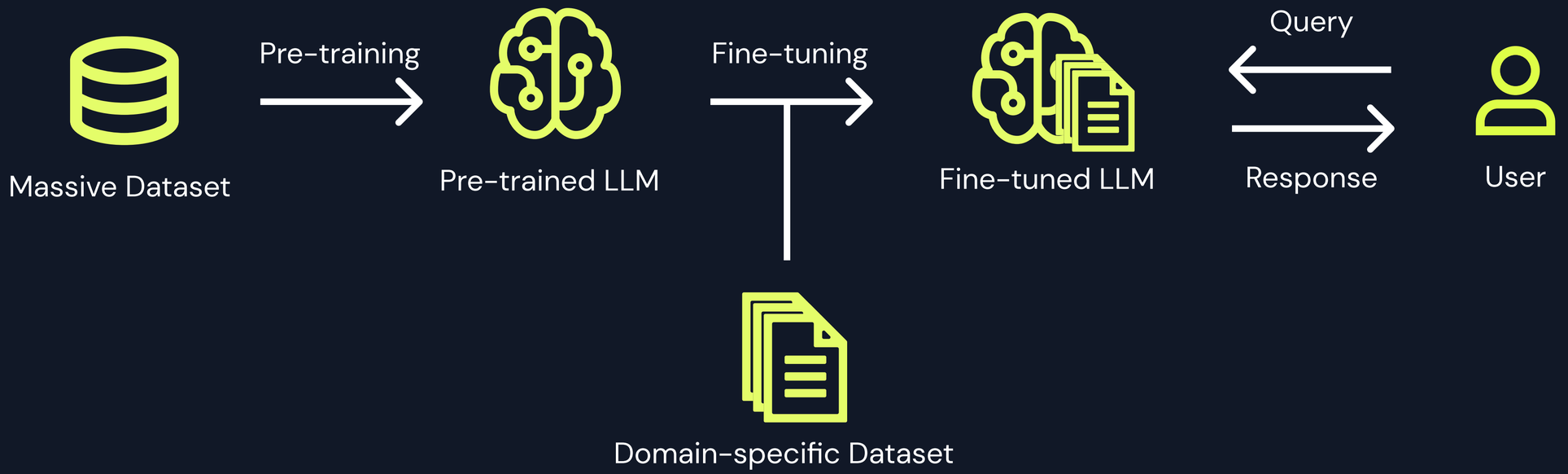
Methods of Finetuning

Self Supervised Learning

Supervised Learning

Reinforcement Learning with Human Feedback

Fine Tuning Architechture



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

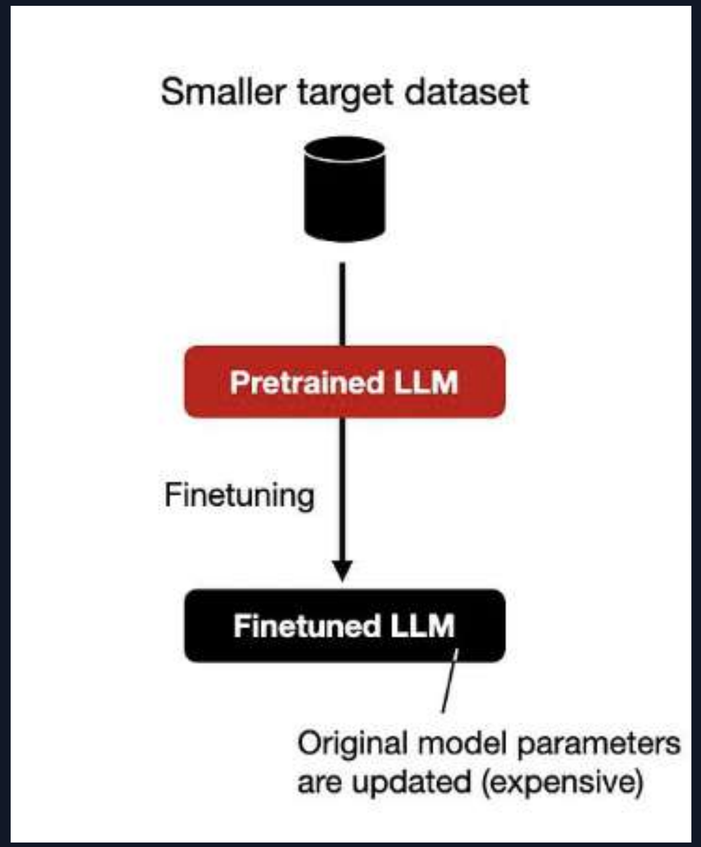
Parameter Efficient Fine Tuning

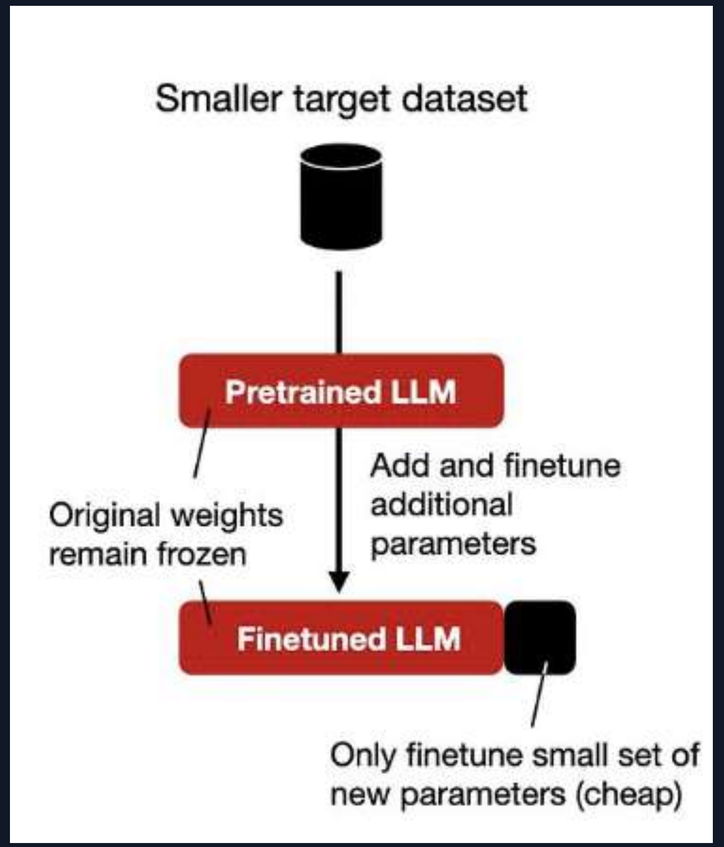
Freeze all the weights of the base LLM

Augment the model with additional parameters and finetune those

Less computationally expensive

Methods for Parameter Fine Tuning





Low Rank Adaptation (LoRA) for PEFT

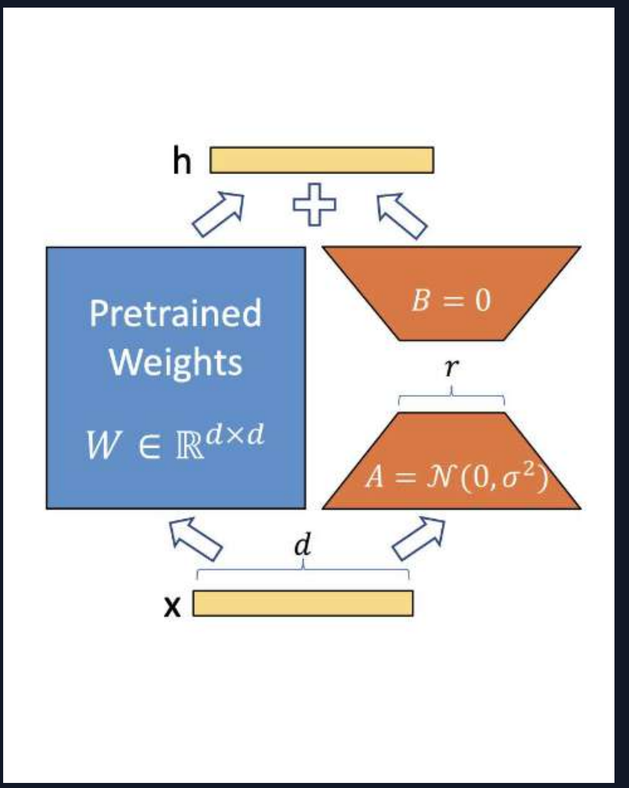
Reduces number of trainable parameters

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



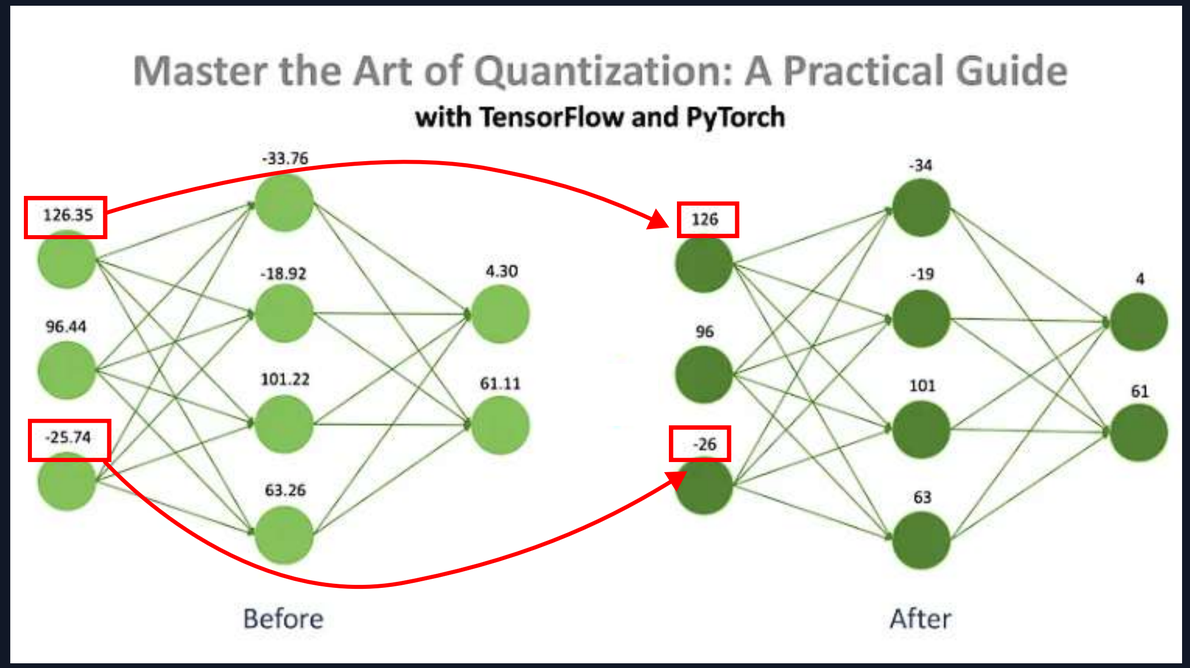
QLoRA for Quantized LLMS

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.

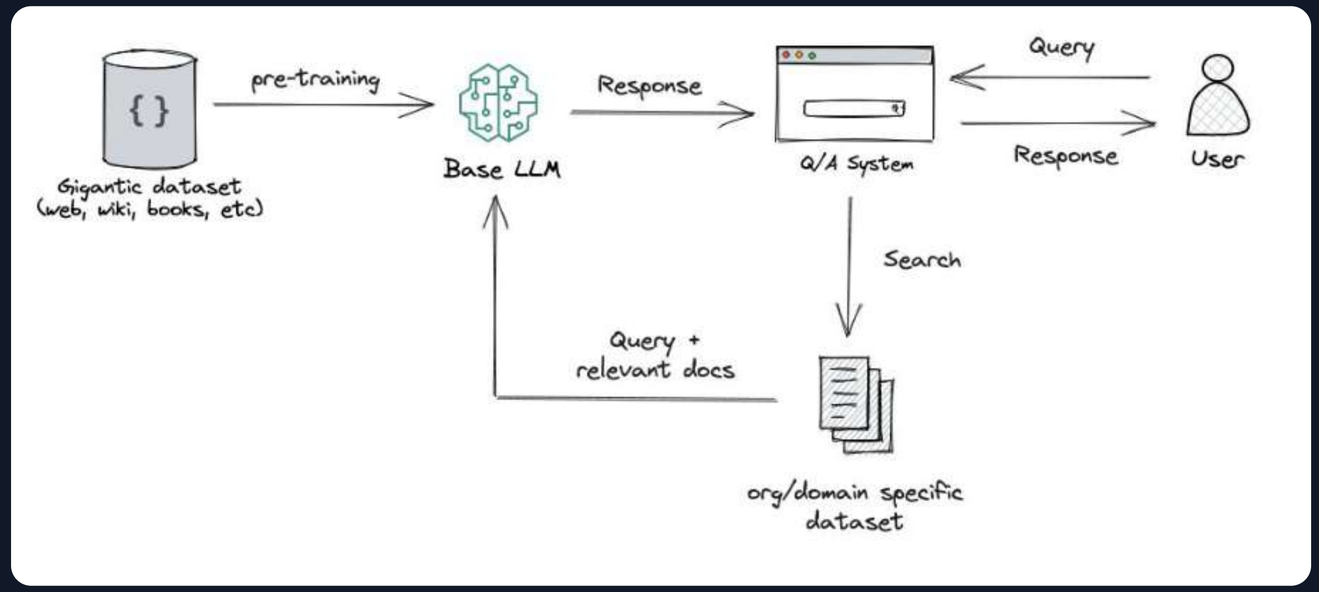
Retrieving Context for Q-A

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**