Generative AI

Generative AI Approaches

Speaker

ARUN PRAKASH ASOKAN

General Guidelines for .env file creation

1. How to create TAVILY_API_KEY

1. Signup here https://app.tavily.com/sign-in
2. Go to overview from the sidebar
3. Check if there is any API_KEY already created
4. If not, click on + and create one API_KEY

Wifi Username: AVDHS24_5G
Password: nitin@123

Email for OpenAI Key issue to
prashant.sahu@analyticsvidhya.com

2. How to create a .env file

1. Please check your email for the OPENAI_API_KEY
2. Open notepad and create the below key

```
OPENAI_API_KEY=‘KEY_RECEIVED_FROM_EMAIL’
TAVILY_API_KEY=‘ENTER_THE_KEY_THAT_IS_CREATED_ABOVE’
```

3. Now save the file as **.env**
4. If your .env is disappeared from the saved location, check if you have selected hidden items under view ☺

Multiple GenAI Approaches

Build your own LLM

BloombergGPT

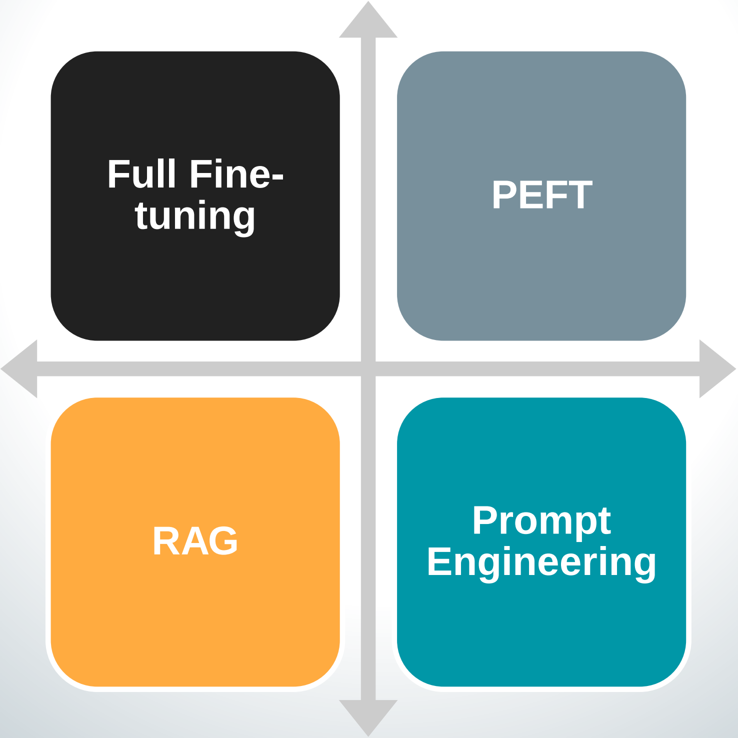
Closed Source LLM via APIs

Open-Source Models hosted on premise

Fine Tune Open Source LLM with Enterprise data

FalconLLM

What is the best approach to use LLMs in business applications?


Full Fine-tuning

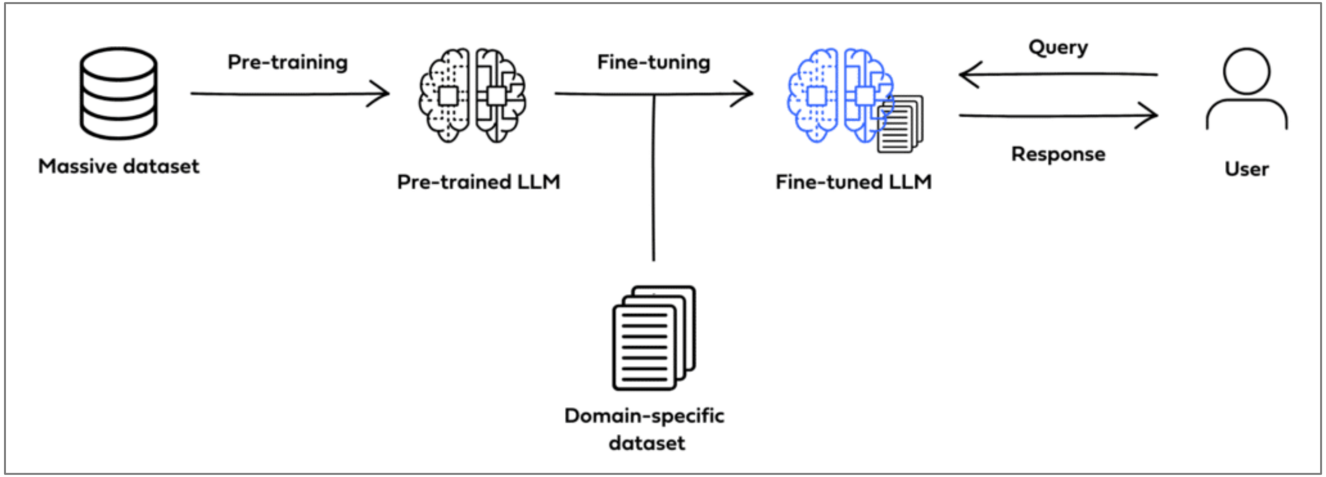
PEFT

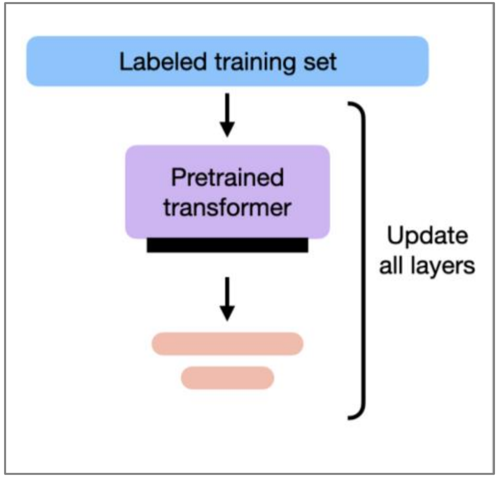
RAG

Prompt Engineering

Full Fine-Tuning

Fine-tuning is the process of refining the training of a pre-existing Language Model (LLM) by using a narrower, task-specific dataset with labeled information. In full fine-tuning, all the model parameters are updated, making it like pretraining—just that it's done on a labeled and much smaller dataset.





Pros

- Requires less data than training from scratch
- Enhanced accuracy
- Increased robustness

Cons

- High computational costs
- Substantial memory requirements
- Time & Expertise Intensive

Parameter-Efficient Fine-Tuning (PEFT)

Pretrained LLM's like LLama2 or Falcon pretrained on vast amounts of data has already learned a broad range of language constructs and knowledge. Given the smaller scope, it's often unnecessary and inefficient to adjust the entire LLM. PEFT uses techniques to further tune a pretrained LLM by **updating only a small number of its total parameters**.

Low-Rank Adaptation of LLM (LoRA) (2023) is a popular PEFT method aimed at reducing costs of fine-tuning

Using **reparameterization**, LoRA downsizes the set of trainable parameters by performing low-rank approximation

Example: Say we have a 100,000 x 100,000 weight matrix, then for full fine-tuning we need to update **10 Billion parameters.**

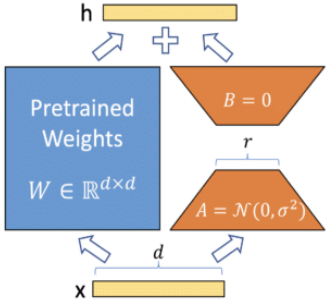
To get this low-rank matrix, we can reparametrize the original weight matrix into two matrices, A and B, each of low rank r, say r = 2

Matrix A: 100,000 x 2 parameters
Matrix B: 2 x 100,000 parameters

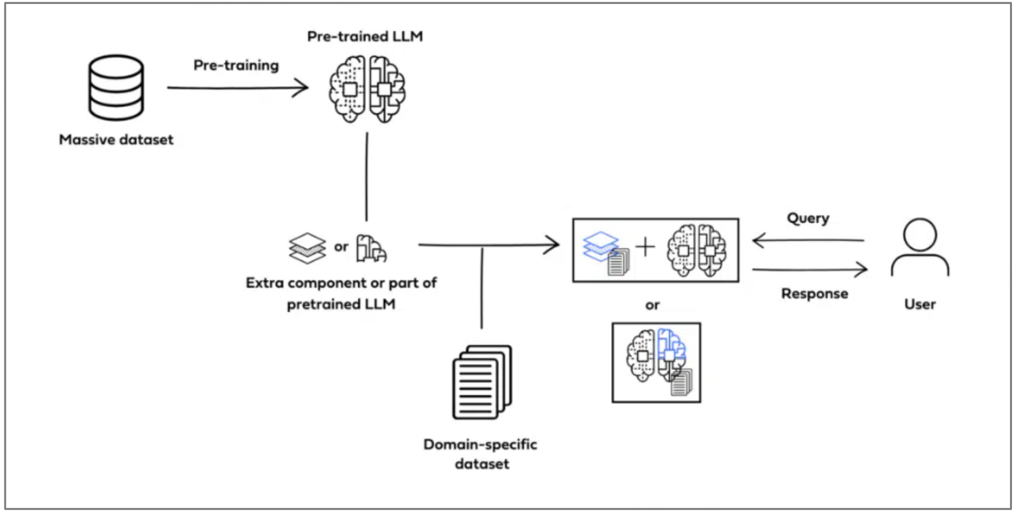
Our new low-rank matrix is then taken to be the product of A and B.

Parameters in A + Parameters in B = (100,000 x 2) + (2 x 100,000) = 200,000 + 200,000 = 400,000 parameters.

We end up updating 400K params instead of 10B params.



LoRA reparameterization trains only A and B.
(Credit: https://arxiv.org/abs/2106.09685)



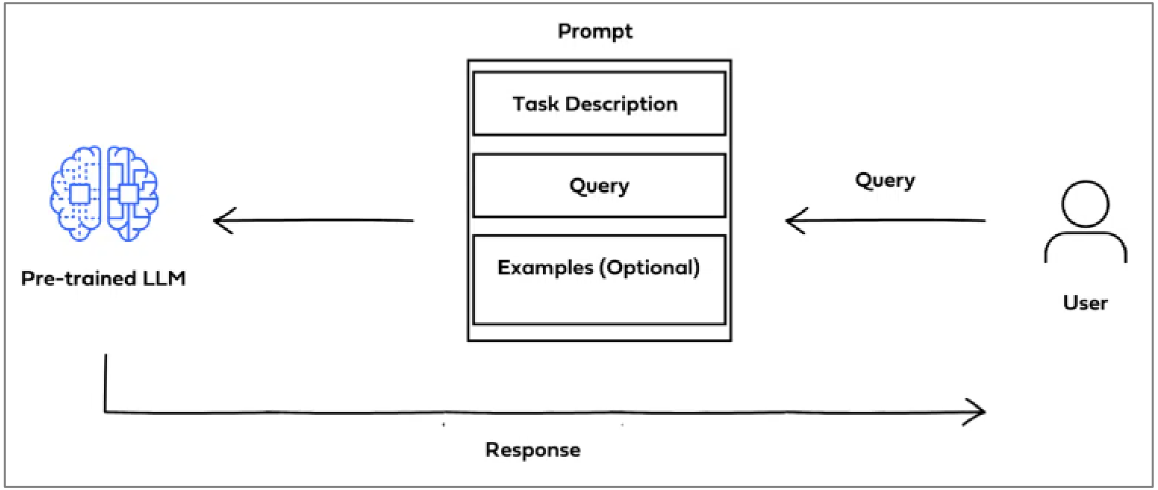
Advantage of LoRA

LoRA is more efficient and faster and preserves knowledge from pretraining.
Using LoRA, we can capture all or most of the crucial information by using a low-rank matrix that contains the selected parameters updated during fine-tuning.

By updating a much smaller number of parameters, we **reduce the computational and memory requirements** (3x lower GPU needs) needed for fine-tuning. Also, **high accuracy almost equivalent to full fine-tuning.**

Prompt Engineering

Prompt Engineering is the process of designing and refining the input given to a model to guide and influence the kind of output you want. Involves no training or fine tuning the LLMs rather we use the pretrained LLM as is.



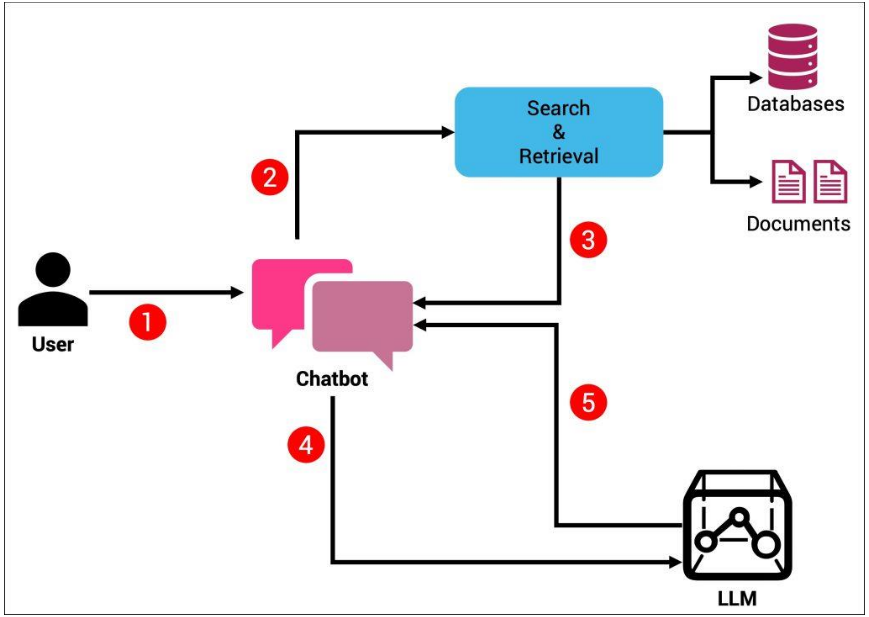
Many types of prompting techniques to make the best out of LLMs...

Art form of Generative AI

Prompting is the equivalent of telling the Genius in the magic lamp **what to do.**

Retrieval Augmented Generation (RAG)

RAG, introduced by Meta is a powerful technique that combines prompt engineering with context retrieval from external data sources to improve the performance and relevance of LLMs.
By grounding the model on additional information, it allows for more accurate and context-aware responses.



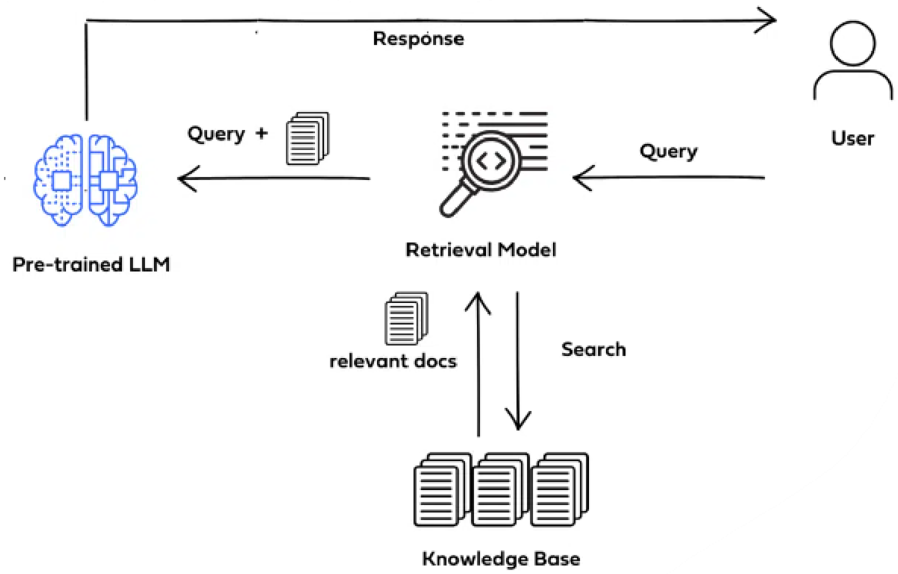
1. Information Retrieval

2. Text Generation

RAG combines information retrieval with text generation in a two-steps:

1. Retrieve - Given an input context, relevant information is retrieved from a knowledge source like documents or a database.
2. Generate - The LLM uses both the original input and the retrieved information to generate a response.

How does RAG work?



1. Vector Store: To implement RAG, the first step involves embedding the internal dataset, converting it into vectors, and then storing these vectors within a dedicated database.

2. User Query: RAG begins by taking in a user's query, which can be a natural language question or statement that requires an answer or completion.

3. Retrieval process: Upon receiving the user's query, the retrieval component scans the vector database to identify relevant pieces of information that closely match the query's semantics. These identified pieces are subsequently used to offer additional context to the Language Model, enabling it to generate a more precise and context-aware response.

4. Integration: The retrieved documents are integrated with the original query, resulting in a prompt that offers additional context for the generation of responses.

5. Text generation: The complete prompt, comprising the concatenated query and retrieved documents, is supplied to a Language Model, which generates the final output.

The Benefits of Retrieval-Augmented Generation

Easily adapts to new data/ Real-time data	• The retriever queries the latest information, keeping the model up-to-date.
Domain knowledge	• The model can incorporate domain-specific data from organizations.
Reduce hallucinations	• Grounding responses in evidence limits false fabrications.
Interpretability	• Users can inspect the retrieved documents used to generate a response.
Scaling	• Adding more data to the retriever index is simpler than retraining models.
Cost effective	• Far lesser computing resources required than fine tuning

Choosing the right approach for the business application

Full Fine Tuning

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

PEFT

- Modifies selective parameters for more efficient adaptation.

RAG

- Merges prompt engineering with external information (database querying) for context-rich answers.

Prompt Engineering

- Refines model input to guide its output.

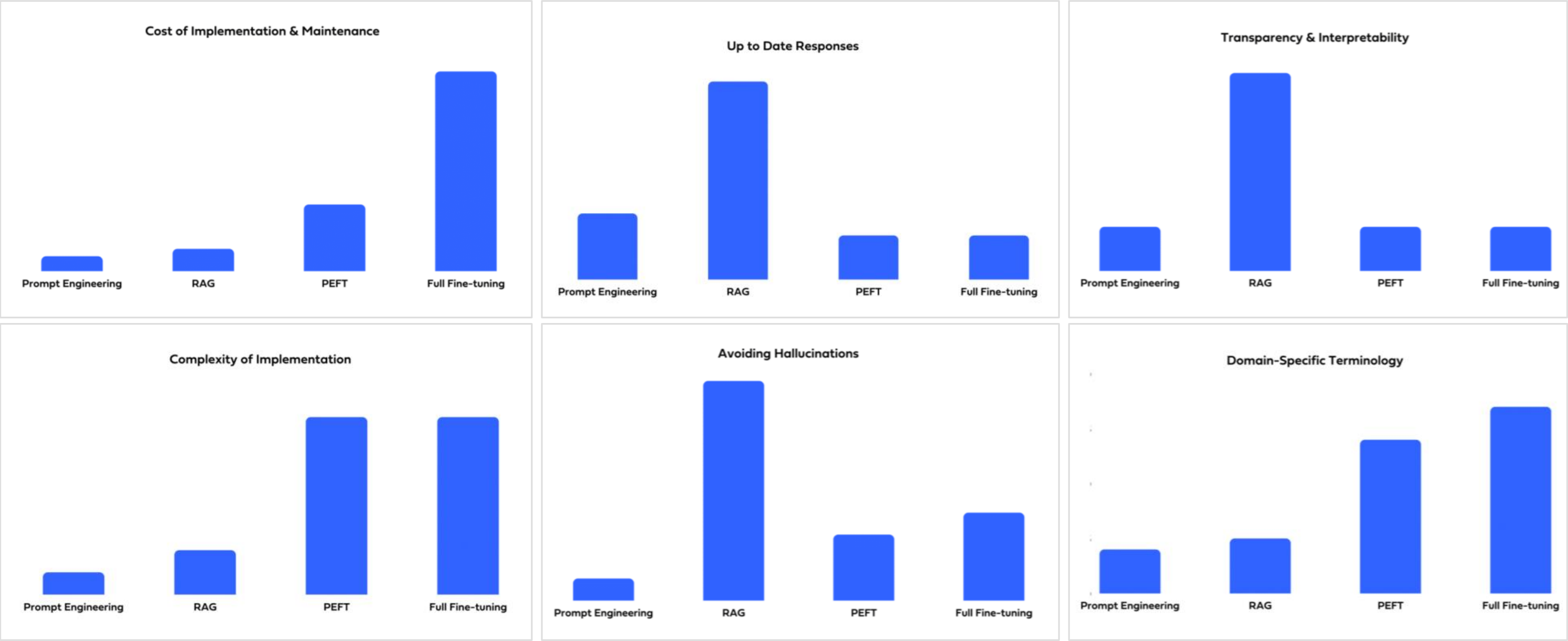


Image credit: https://deci.ai/

Building Generative AI Solutions at Enterprises

GenAI using Sensitive Enterprise Data

Challenges in dealing with sensitive data at enterprises

- Legal Penalties
- Regulatory Compliance
- Patient Safety
- Data Breaches

DATA Leaks

SAMSUNG

"Samsung's Engineers accidentally expose Top-Secret Information via ChatGPT, Impacting Semiconductor Division"

"ChatGPT temporarily blocked in Italy over data breach concerns"

Why Top Companies Restrict ChatGPT?

NO
chatGPT
citi
accenture
amazon
WELLS FARGO
SAMSUNG
Goldman Sachs
verizon

GenAI using Sensitive Enterprise Data

AZURE OPEN AI

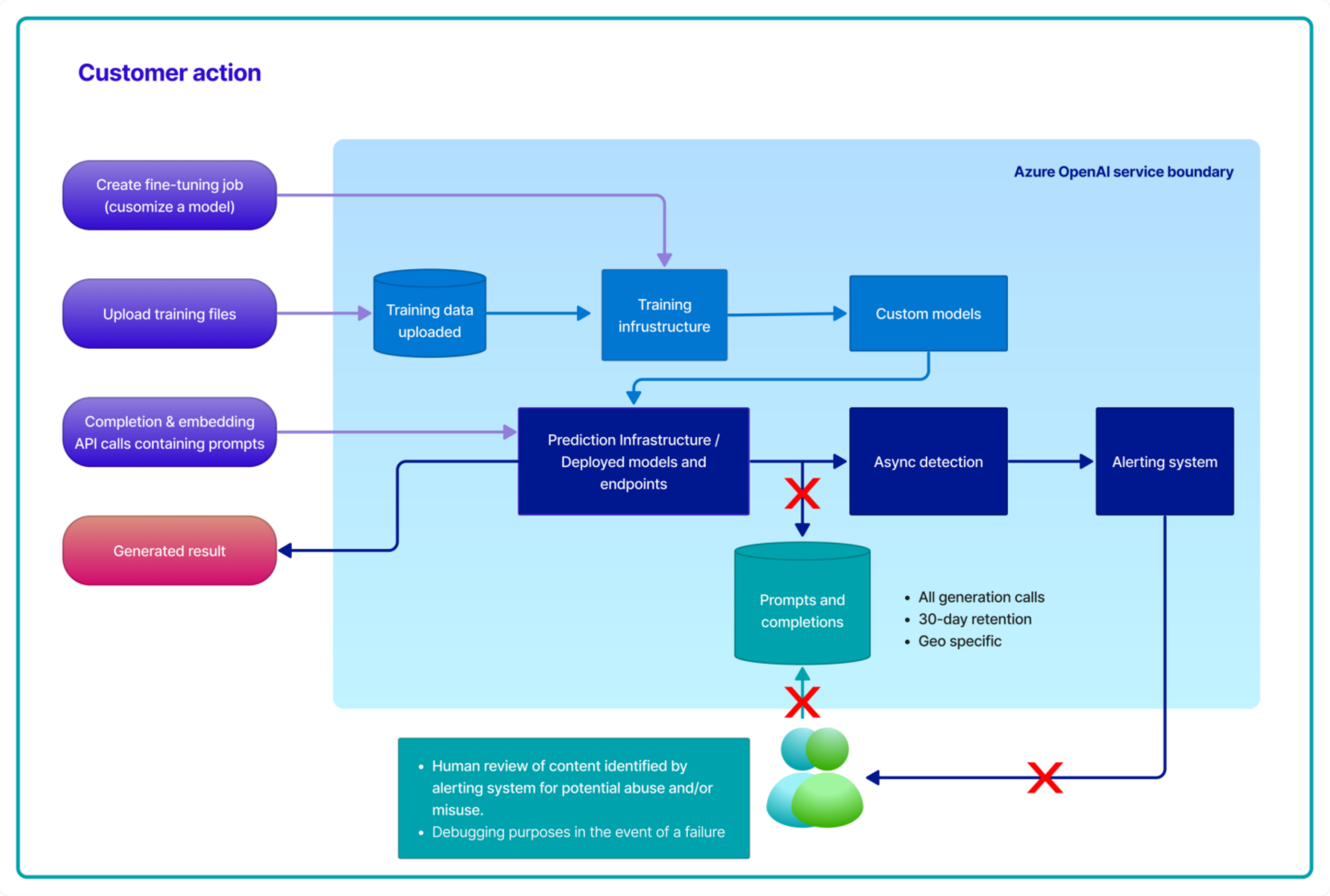
No Storage of Prompts and Completions. Foundational model doesn't get trained on this

Private Deployment of Independent ChatGPT Instance within Azure Subscription

Secure Data Storage in Azure: Encrypted and Isolated with Managed Keys.

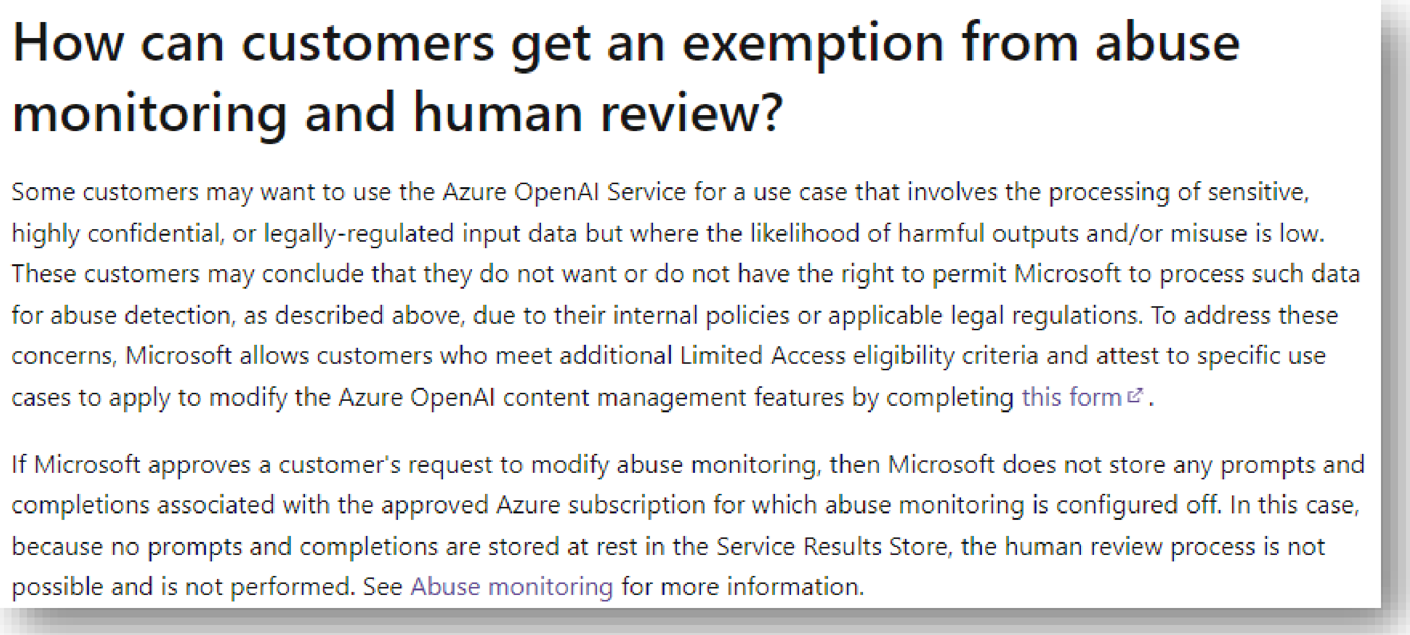
Enhanced Security: Limiting Resource Access with Network Rules and Request Filtering.

Opting out of human review process


Customer action
Create fine-tuning job (cusomize a model)
Upload training files
Completion & embedding API calls containing prompts
Generated result
Azure OpenAI service boundary
Training data uploaded
Training infrustructure
Custom models
Prediction Infrastructure / Deployed models and endpoints
Async detection
Alerting system
Prompts and completions
All generation calls
30-day retention
Geo specific
Human review of content identified by alerting system for potential abuse and/or misuse.
Debugging purposes in the event of a failure

My Recommendation: Just Opt Out !

You can opt out of any human review & abuse monitoring by filling a form.



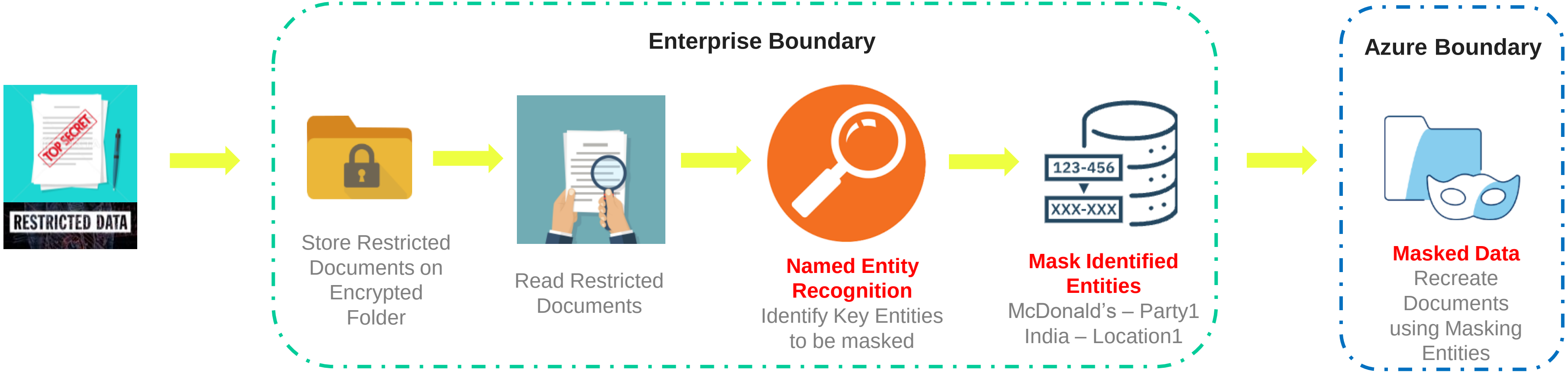
How can customers get an exemption from abuse monitoring and human review?

Some customers may want to use the Azure OpenAI Service for a use case that involves the processing of sensitive, highly confidential, or legally-regulated input data but where the likelihood of harmful outputs and/or misuse is low. These customers may conclude that they do not want or do not have the right to permit Microsoft to process such data for abuse detection, as described above, due to their internal policies or applicable legal regulations. To address these concerns, Microsoft allows customers who meet additional Limited Access eligibility criteria and attest to specific use cases to apply to modify the Azure OpenAI content management features by completing this form.

If Microsoft approves a customer's request to modify abuse monitoring, then Microsoft does not store any prompts and completions associated with the approved Azure subscription for which abuse monitoring is configured off. In this case, because no prompts and completions are stored at rest in the Service Results Store, the human review process is not possible and is not performed. See Abuse monitoring for more information.

More details available here

How did I handle restricted data at my enterprise?



Enterprise Boundary

Store Restricted Documents on Encrypted Folder

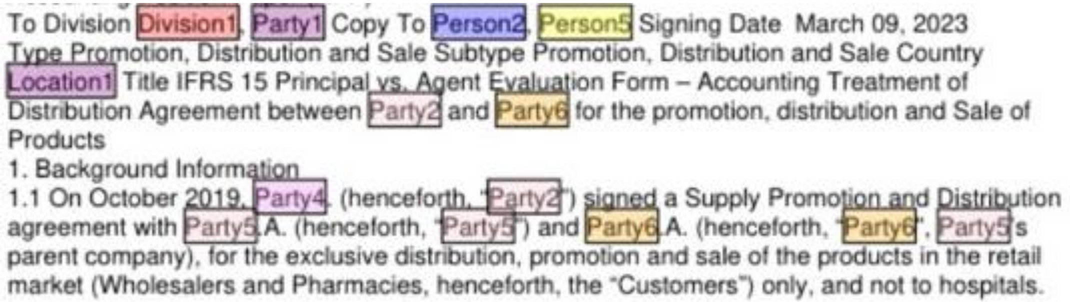
Read Restricted Documents

Named Entity Recognition
Identify Key Entities to be masked

Mask Identified Entities
McDonald's – Party1
India – Location1

Azure Boundary

Masked Data
Recreate Documents using Masking Entities



To Division Division1, Party1 Copy To Person2, Person5 Signing Date March 09, 2023
Type Promotion, Distribution and Sale Subtype Promotion, Distribution and Sale Country Location1 Title IFRS 15 Principal vs. Agent Evaluation Form – Accounting Treatment of Distribution Agreement between Party2 and Party6 for the promotion, distribution and Sale of Products

1. Background Information

1.1 On October 2019, Party4. (henceforth, "Party2") signed a Supply Promotion and Distribution agreement with Party5.A. (henceforth, "Party5") and Party6.A. (henceforth, "Party6", Party5's parent company), for the exclusive distribution, promotion and sale of the products in the retail market (Wholesalers and Pharmacies, henceforth, the "Customers") only, and not to hospitals.

BERT – NER

Pros – Operates locally. Fast application. Finetuning capability.

Cons – Difficulty in changing model parameters. Extensive post-processing of results required.

Open-Source LLMs

Pros – Operates locally. Few-shot prompting capability. Finetuning capability.

Cons - Slow application. Every query can take significant time to respond, which might affect large use-cases.