



13 SEPTEMBER 2025



AI INDUSTRIAL summit 2025



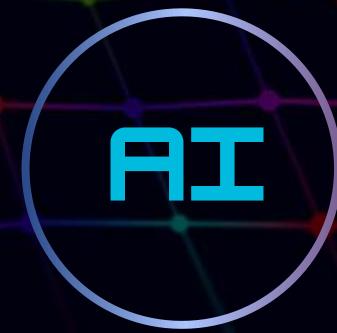
we are your local AI crew building cool stuff, sharing ideas, and making tech event



13 SEPTEMBER 2025

PERSONALIZE YOUR AI: MODEL FINE-TUNE WITH YOUR DATA IN AZURE OPENAI

Massimo Bonanni - Senior Technical Trainer @ Microsoft



we are your local AI crew building cool stuff, sharing ideas, and making tech event



<epam>



ICB SOFTWARE
INNOVATION
A Kongsberg Digital company



13 SEPTEMBER 2025

we are your local AI crew building cool stuff, sharing ideas, and making tech event

Big shoutout to our amazing sponsors thanks for believing in the power of AI and helping us bring this event to life in Sofia!

ABOUT SPONSORS



THE WAY TO USE YOUR DATA IN LLM

13 SEPTEMBER 2025

Prompt Engineering

- Uses prompts to improve the accuracy and relevancy of responses from natural language processing models.
- By optimizing prompts, the model's performance is enhanced

Retrieval Augmented Generation

- Enhances performance by fetching data from external sources and incorporating it into a prompt.
- This method allows to create customized solutions, maintain data relevance, and control costs.

Fine Tuning

- Retrains a pre-existing LLM using example data to create a custom version of the model.
- This customization optimizes model based on the provided examples.

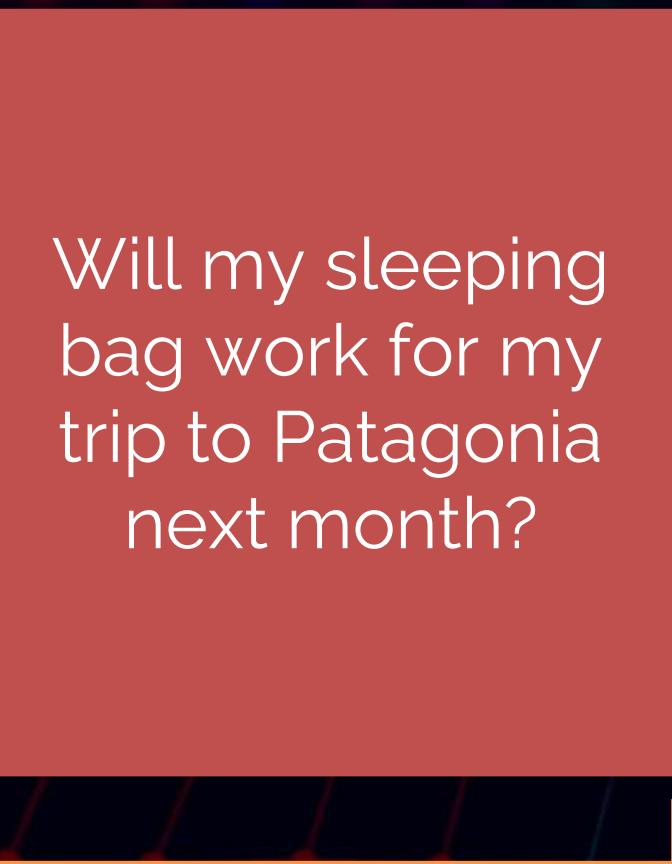
Train a model

- You need to have talent, data and GPU's to train a new model...
-and time and money also!!!!



PROMPT ENGINEERING AND RAG

13 SEPTEMBER 2025



User input

Basic Prompt
Engineering

Tone and
Style

Example
Responses

Intent
Mapping

Retrieval / RAG

Weather
lookup

Personalization

..and more!

Yes, your Elite Eco sleeping bag is rated to 21.6F, which is below the average low temperature in Patagonia in September



Prompt engineering



LLM



Output



PROMPT ENGINEERING AND RAG

13 SEPTEMBER 2025

Basic Prompt
Engineering

Retrieval / RAG

Tone and
Style

Weather
lookup

Will my sleeping
bag work for my
trip to Patagonia
next month?

Yes, your Elite Eco
sleeping bag is
rated to 21.6F,
which is below the
average low
temperature in
Patagonia in
September

**Prompt engineering & RAG are about
adding your data to the prompt!!!**

Mapping

..and more!



User input



Prompt engineering



LLM



Output



FINE-TUNING

13 SEPTEMBER 2025

Basic Prompt
Engineering

Retrieval / RAG

Tone and
Style

Weather
lookup

Will my sleeping
bag work for my
trip to Patagonia
next month?

Yes, your Elite Eco
sleeping bag is
rated to 21.6F,
which is below the
average low
temperature in
Patagonia in
September

**Fine tuning is about adding your data to
the LLM itself**



User input



Prompt engineering

Mapping



Fine-
Tuned
LLM



Output



WHAT IS FINE-TUNING?

13 SEPTEMBER 2025

Fine-tuning refers to customizing a pre-trained LLM with additional training on a specific task or new dataset for enhanced performance and accuracy

LLM
Base
Model

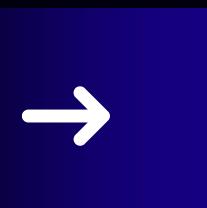


Training Data Set

Fine-Tuning



Fine-
Tuned
LLM



THE FINE-TUNE WORKFLOW

13 SEPTEMBER 2025



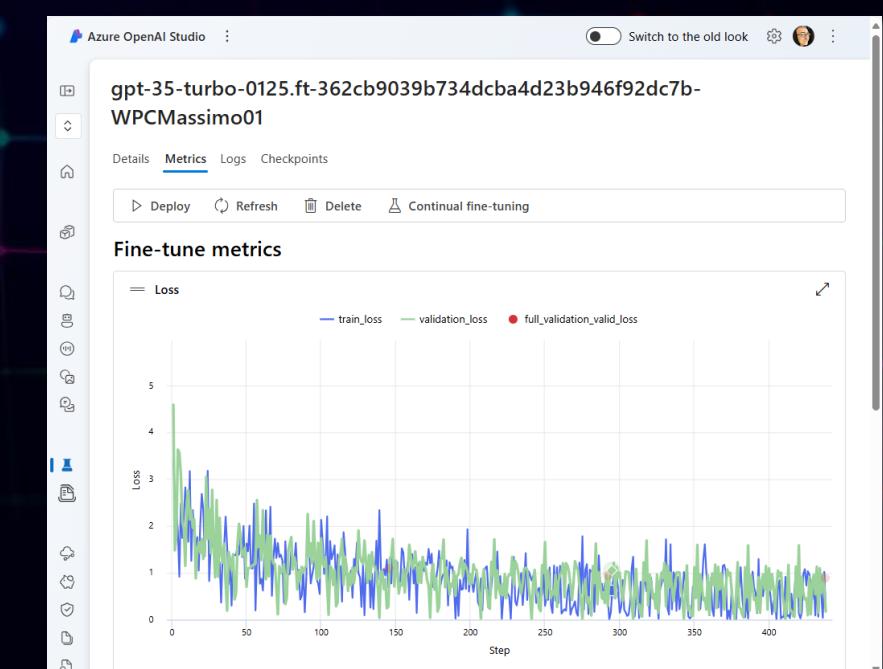
- Prepare 100s+ examples
- Format as chat completion
- Define training & validation sets
- Upload data to the service

- Select base model
- Choose the training method and type
- Set hyperparameters
- Specify data
- Executes on Shared Compute

Training & validation loss

Training & validation accuracy

Model endpoint for inferencing with AOAI



The screenshot shows the Azure OpenAI Studio interface. The top navigation bar includes 'Azure OpenAI Studio' and 'Model deployments'. A 'Switch to the old look' toggle is available, along with user profile and settings icons.

The left sidebar contains navigation links: Home, Get started, Model catalog, Playgrounds (Chat, Assistants PREVIEW, Real-time audio PREVIEW, Images, Completions), Tools (Fine-tuning, Batch jobs PREVIEW), and Shared resources (Deployments, Quota, Content filters, Data files, Model cards).

The main content area is titled 'Model deployments'. It displays a message: 'Deploy a model with your private API key and an endpoint URI (Uniform Resource Identifier).'. Below this are two tabs: 'Model deployments' (selected) and 'App deployments'.

A toolbar at the top of the deployment list includes: '+ Deploy model' (dropdown menu), Refresh, Edit, Delete, and a three-dot menu.

The deployment list table has columns: Model name, Model version, and Status (S). The table shows three entries:

	Model name	Model version	S
	dall-e-3	3.0	S
	gpt-35-turbo-16k	0613	S
	gpt-4o	2024-08-06	S

The screenshot shows the Azure OpenAI Studio interface. On the left, there's a sidebar with navigation links: Home, Get started, Model catalog, Playgrounds (Chat, Assistants, Real-time audio, Images, Completions), Tools (Fine-tuning), and a dropdown for 'Current resource' set to 'wpc2024openai'. The main area has a title 'Fine-tune a model by training it on your own data' with a sub-instruction: 'Optimize pre-trained models for specific tasks by training it on a smaller, task specific dataset to improve its performance and accuracy. Because this method tends to require fewer examples in the prompts, generally less text is sent—and tokens processed—per call.' Below this is a search bar and a table titled 'Fine-tune model' with columns 'Model name' and 'Base model'. The table contains three rows of data.

Model name	Base model
gpt-35-turbo-0125.ft-362cb9039b734dcba4d23b946f92dc7b-WPCMassimo01	gpt-35-tu
gpt-4o-2024-08-06.ft-16b4d2519c24ef3852951fe77e31287-WPCMassimo02	gpt-4o-20
gpt-4o-2024-08-06.ft-c46900557e854dc592f75881bfff2e69-WPCMassimo01	gpt-4o-20



PREPARE TRAINING AND VALIDATION DATA

13 SEPTEMBER 2025

01010
10101
01010

Your training and validation datasets should contain input and output samples that reflect the model's expected performance.

01010
10101
01010

Ensure the data is formatted as a JSON Lines (JSONL) document.

01010
10101
01010

Providing a larger number of training examples is beneficial; aim for hundreds or thousands for best results.

01010
10101
01010

Generally, expanding the dataset size can improve model quality, but be cautious as low-quality examples can negatively impact performance.

```
{
  "messages": [
    {
      "role": "system",
      "content": "You are an helpful assistant that help people to know"
    },
    {
      "role": "user",
      "content": "What does the Apparecchio tableware set include?"
    },
    {
      "role": "assistant",
      "content": "The Apparecchio tableware set includes stylish plates"
    }
  ]
}

{
  "messages": [
    {
      "role": "system",
      "content": "You are an helpful assistant that help people to know"
    },
    {
      "role": "user",
      "content": "What materials are the Apparecchio plates and bowls"
    },
    {
      "role": "assistant",
      "content": "The Apparecchio plates and bowls are made of durable"
    }
  ]
}
```



TRAINING METHODS

13 SEPTEMBER 2025

Supervised Fine Tuning (SFT)

is a method used to improve the performance of a large language model by training it further on a labeled dataset.

Direct Preference Optimization (Preview)

is a method for aligning large language models with human preferences.

Reinforcement Fine Tuning (Preview)

is a training method for improving **reasoning** models by using reward signals instead of just labeled data



TRAINING TYPES

13 SEPTEMBER 2025

Where training occurs

Data residency

Cost per token

Performance

Limitations



Standard

In the current Azure OpenAI resource's region

Yes – data stays within the selected Azure region

Standard pricing

Dependent on regional compute capacity

Must have capacity in your selected region



Global (preview)

In global training infrastructure, outside your current region

No – data may be processed in other regions

Lower cost per token (more affordable)

May have faster availability due to more capacity, but still in preview

May not be available in all regions during preview



FINE-TUNE PARAMETERS

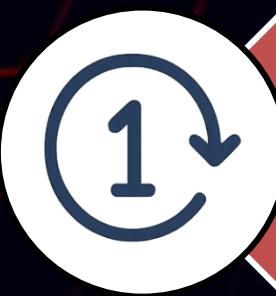
13 SEPTEMBER 2025



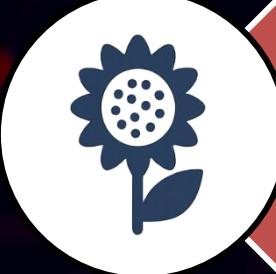
Batch Size: Number of training examples processed at once. Large batches = more stable updates.



Learning Rate: Speed of learning. Higher values = faster, but riskier.



Epochs: How many times the data is used during training.



Seed: Makes the training repeatable (same input = same result).

Seed ⓘ
3203386110

Configure hyperparameters ⓘ

Batch size (1-32) ⓘ 16

Learning rate multiplier (0.0-10.0) ⓘ 0.92

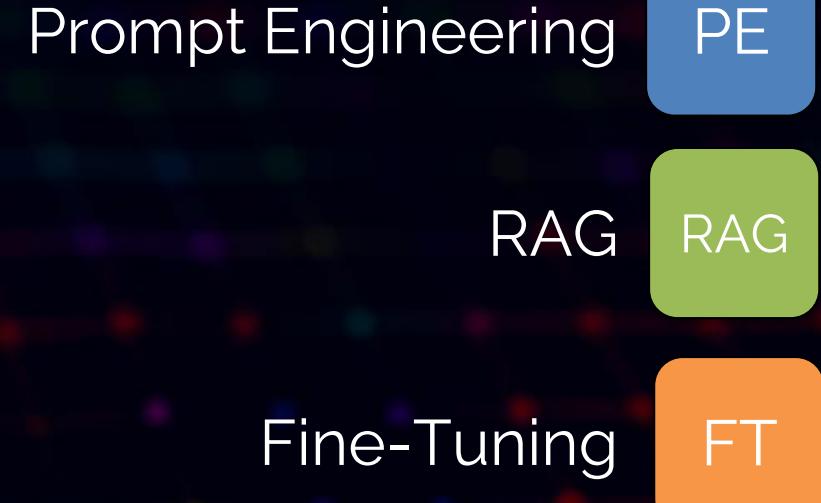
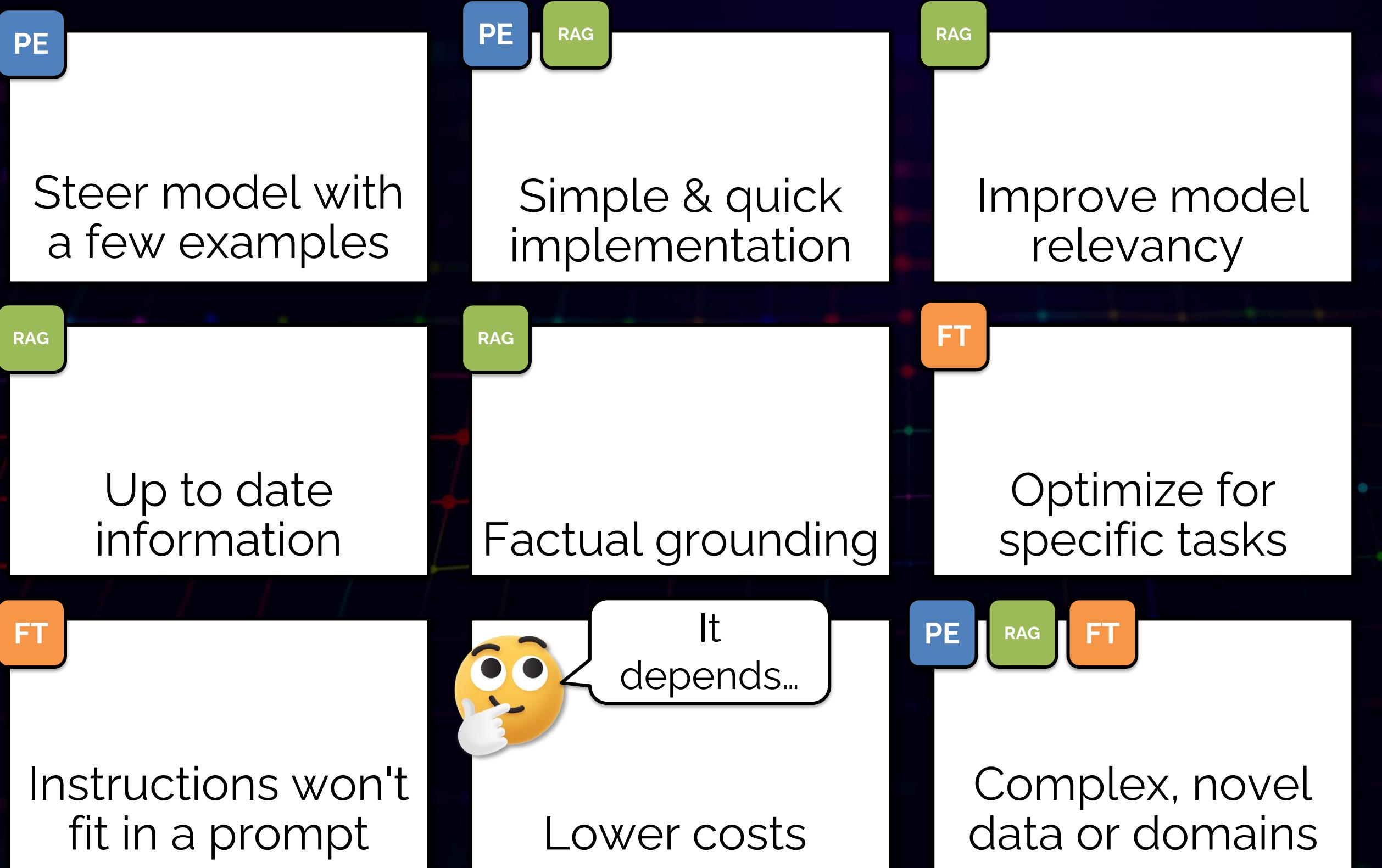
Number of epochs (1-10) ⓘ 3

The parameters depend on the type of training chosen.



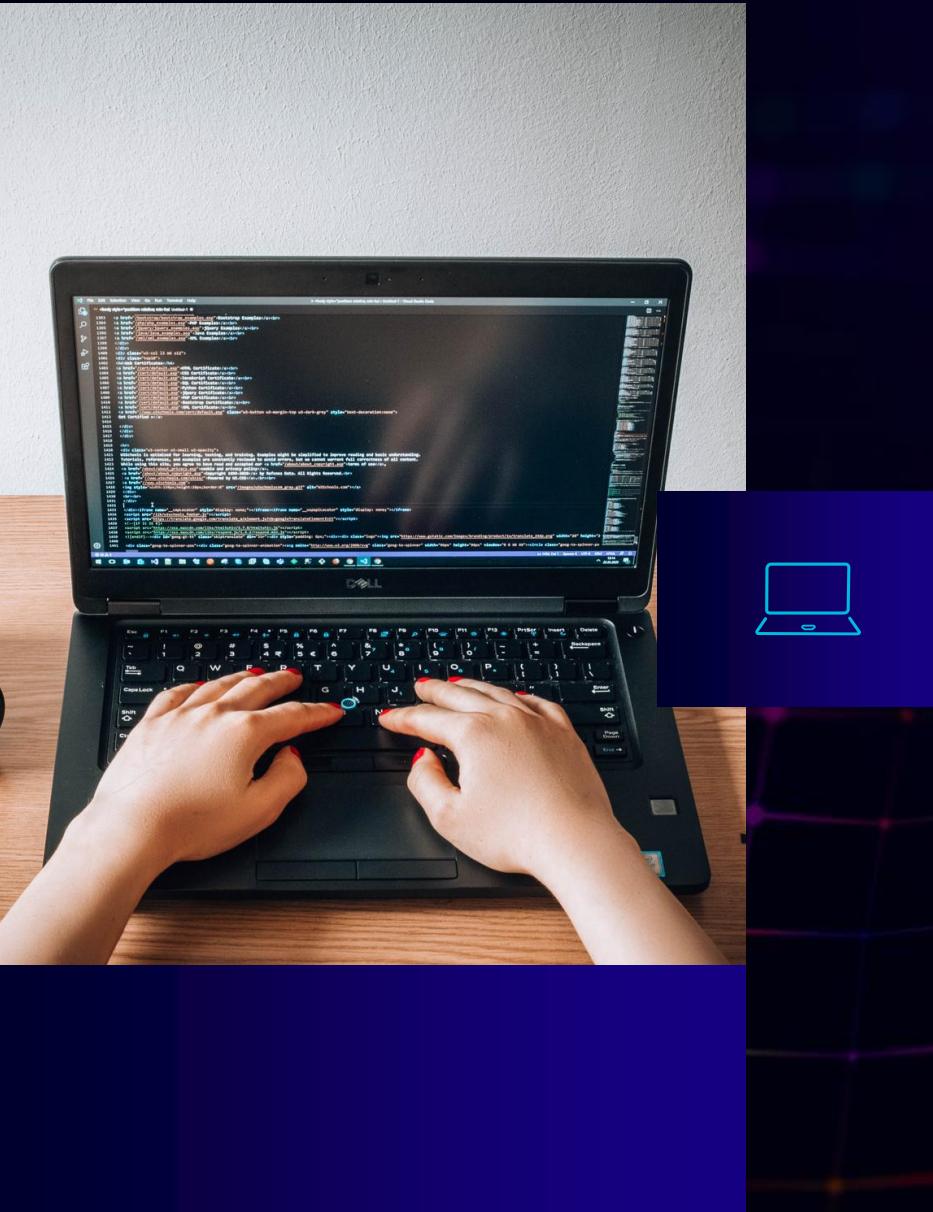
WHAT APPROACH....

13 SEPTEMBER 2025





13 SEPTEMBER 2025



DEMO
M.O.D.A.
MODERN OUTSTANDING
DESIGN ASSEMBLED



we are your local AI crew building cool stuff, sharing ideas, and making tech event



Supervised / Preference Fine-Tuning

- Cost = Tokens in training file × Epochs × Price per token
- Smaller/newer models → lower token cost
- Global training = cheaper than regional if data residency not required
- Not billed for queue time, failed jobs, canceled jobs before training start

Reinforcement Fine-Tuning

- Cost = Training time (hrs) × Hourly rate (+ grader token costs if used)
- Example rate: \$100/hr (o4-mini)
- Max per job = \$5,000 cap, job paused at cap for review/resume



M.O.D.A. TRAINING COST

13 SEPTEMBER 2025

Model attributes	
ID	ftjob-36cb824ebbc64ced8e20a123b1dbf3c0
Status	Completed
Created on	Mar 22, 2025 5:22 PM
Validation file	Training_QA.jsonl
Weights & Biases integration enabled?	No
Base model	gpt-4o-mini-2024-07-18
Training file	Training_QA.jsonl
Azure OpenAI Service resource	ECS2025OpenAI
Method of Customization	Supervised
Duration	1h 26m 2s

Training tokens billed

729,000

GPT-4o-mini

Regional

Input: €0.144/1M tokens

Cached Input: €0.079/1M tokens

Output: €0.58/1M tokens

Training: €2.9/1M tokens

Hosting: €1.5/hour

€ 2.115€



COST MANAGEMENT - HOSTING

13 SEPTEMBER 2025

Deployment Type	Token Rate	Hourly Rate
Standard	Same as base model	€1.5/hour
Global Standard	Same as base model	€1.5/hour
Regional Provisioned Throughput	None	PTU/hour
Developer Tier	Same as Global Standard	none

- Neither data residency nor availability guarantees.
- Developer Tier is designed for model candidate evaluation and proof of concepts.
- Deployments are removed automatically after 24 hours regardless of usage but may be redeployed as needed.



PROS & CONS

13 SEPTEMBER 2025



RAG

Fine-Tuning

Pro(s)

Cost effective
Dynamic, Up-to-date
Domain Flexibility

Cons(s)

Vector Db Dependency
Relies on Data Quality
Introduces Latency

Higher Costs
Static Knowledge



13 SEPTEMBER 2025

THANK YOU!

Thank you for exploring the session. Feel free to ask questions!
See you next event for

DATA
SATURDAYS



we are your local AI crew building cool stuff, sharing ideas, and making tech even



Massimo Bonanni

Sr Technical Trainer @ Microsoft

massimo.bonanni@microsoft.com

@massimobonanni



aka.ms/maxlinkedin



REFERENCES

13 SEPTEMBER 2025

- [Customize a model with Azure OpenAI in Azure AI Foundry Models - Microsoft Learn](#)
- [Fine-tuning and distillation with Azure AI Foundry - Build 2025 \(session\)](#)
- [Microsoft Build 2025 Book of News](#)
- [Fine-tune models with Azure AI Foundry - Azure AI Foundry | Microsoft Learn](#)
- [massimobonanni/MODA-FineTuning](#)
- [Azure-Samples/AIFoundry-Customization-Datasets](#)



we are your local AI crew building cool stuff, sharing ideas, and making tech event