13 SEPTEMBER 2025

AI INDUSTRIAL summit 2025

global AI community

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

AI

global AI community

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

global AI community

Microsoft

Xebia

ICB | SOFTWARE INNOVATION
A Kongsberg Digital company

Baringa

Microsoft® MVP Most Valuable Professional

epam

SOFIA TECHPARK

AI CLOUD AGENCY

Kinkao_Sofia
THAIFOOD + HOMEMADE

13 SEPTEMBER 2025

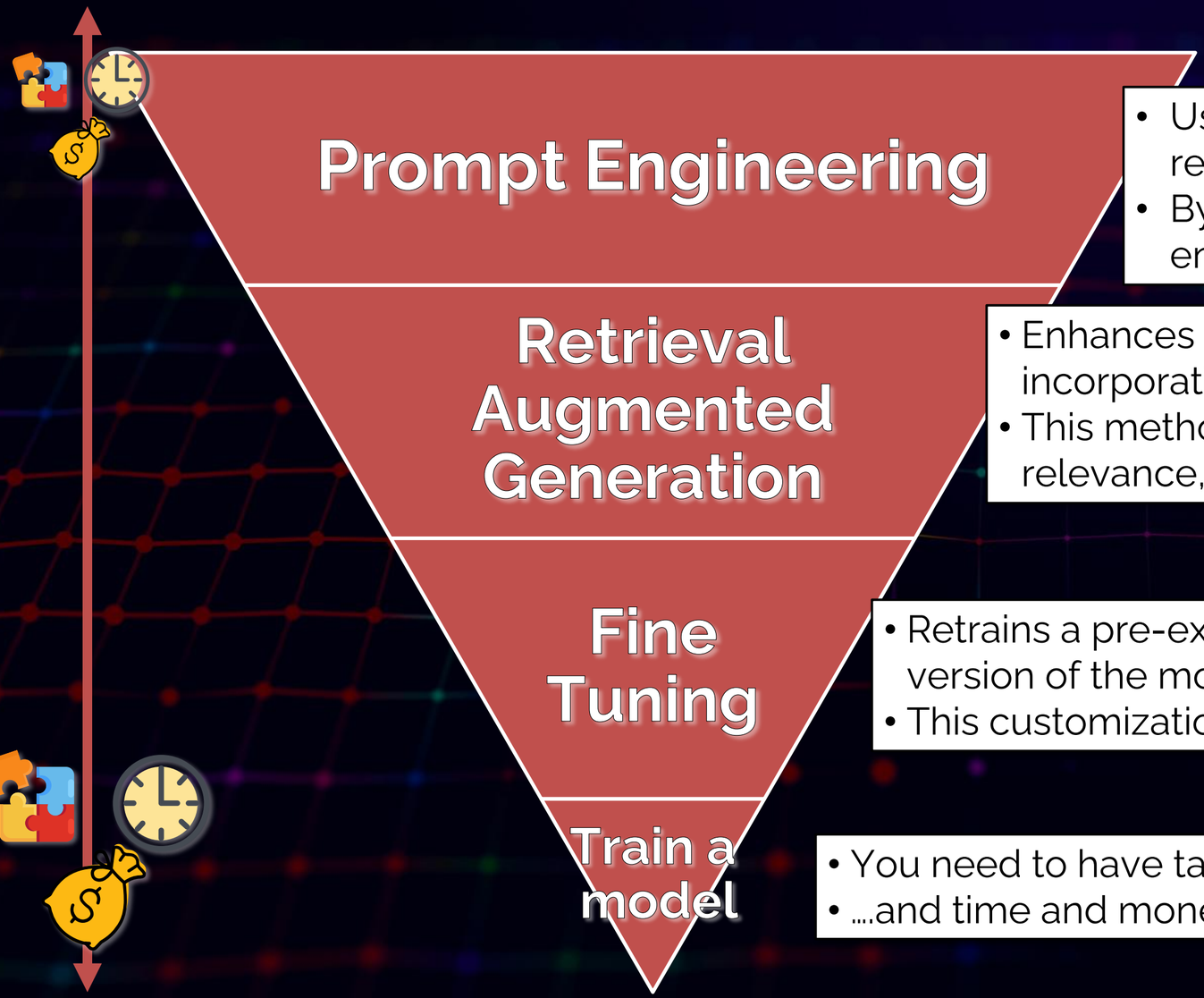
we are your local AI crew building cool stuff, sharing ideas, and making tech event

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



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

- 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.

- 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.

- 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

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

+

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

User input

Prompt engineering

LLM

Output

PROMPT ENGINEERING AND RAG

13 SEPTEMBER 2025

Basic Prompt Engineering

Retrieval / RAG

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

Tone and Style

Example Responses

Intent Mapping

Weather lookup

Past Conversations

...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 & RAG are about adding your data to the prompt!!!

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

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

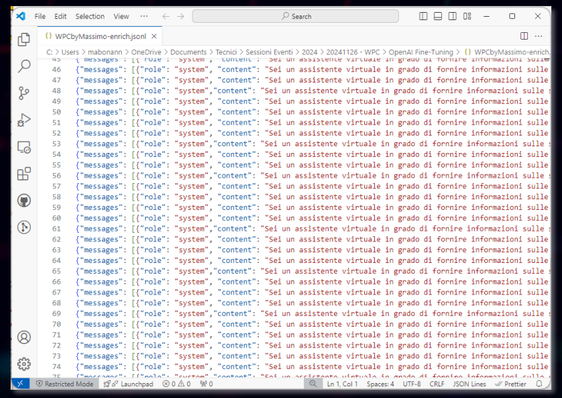
Data

Prepare 100s+ examples

Format as chat completion

Define training & validation sets

Upload data to the service



Training

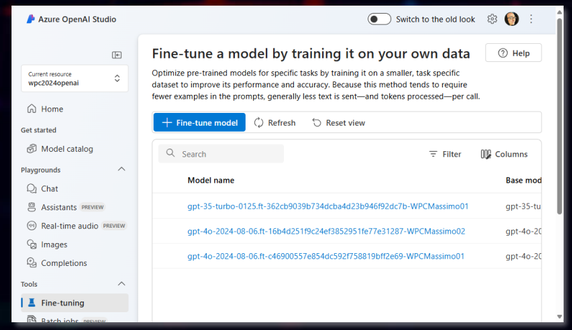
Select base model

Choose the training method and type

Ser hyperparameters

Specify data

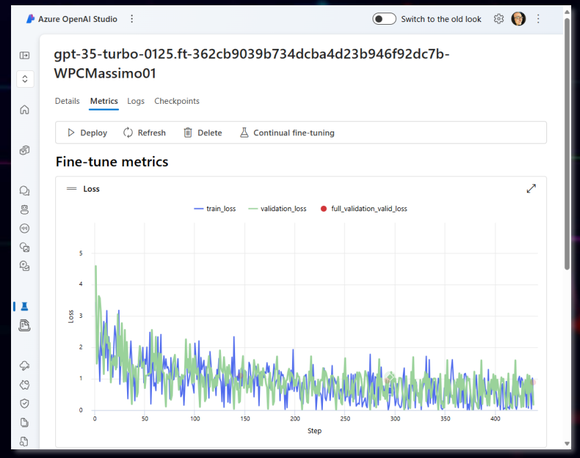
Executes on Shared Compute



Evaluation

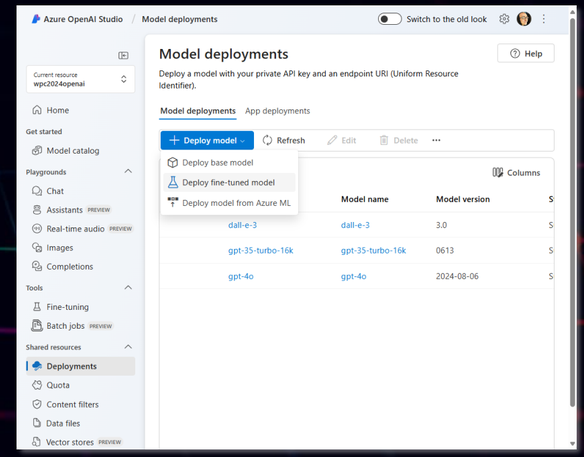
Training & validation loss

Training & validation accuracy



Deployment

Model endpoint for inferencing with AOAI



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

	Standard	Global (preview)
Where training occurs	In the current Azure OpenAI resource's region	In global training infrastructure, outside your current region
Data residency	Yes – data stays within the selected Azure region	No – data may be processed in other regions
Cost per token	Standard pricing	Lower cost per token (more affordable)
Performance	Dependent on regional compute capacity	May have faster availability due to more capacity, but still in preview
Limitations	Must have capacity in your selected region	May not be available in all regions during preview

FINE-TUNE PARAMETERS

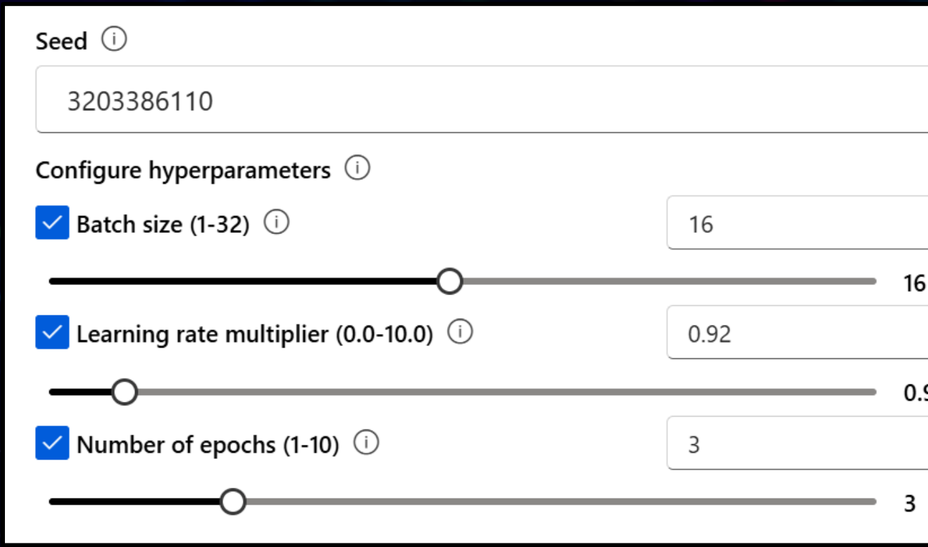
13 SEPTEMBER 2025

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).



The parameters depend on the type of training chosen.

WHAT APPROACH....

13 SEPTEMBER 2025

PE

Steer model with a few examples

PE RAG

Simple & quick implementation

RAG

Improve model relevancy

RAG

Up to date information

RAG

Factual grounding

FT

Optimize for specific tasks

FT

Instructions won't fit in a prompt

Lower costs

PE RAG FT

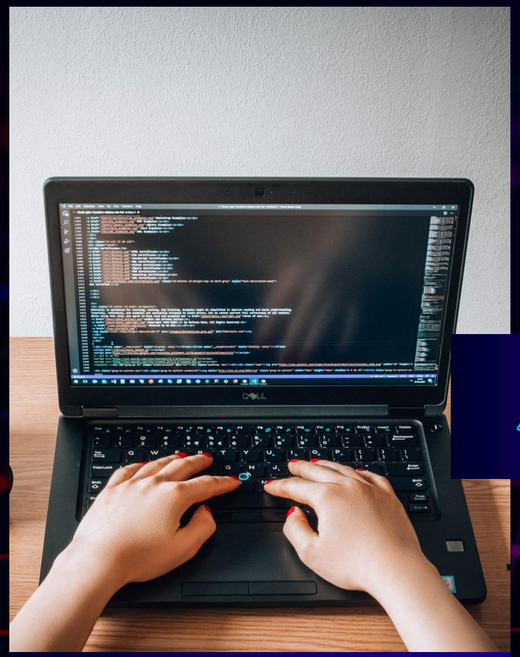
Complex, novel data or domains

Prompt Engineering PE

RAG RAG

Fine-Tuning FT

13 SEPTEMBER 2025



DEMO

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

global AI community

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

COST MANAGEMENT - TRAINING

13 SEPTEMBER 2025

Supervised / Preference Fine-Tuning

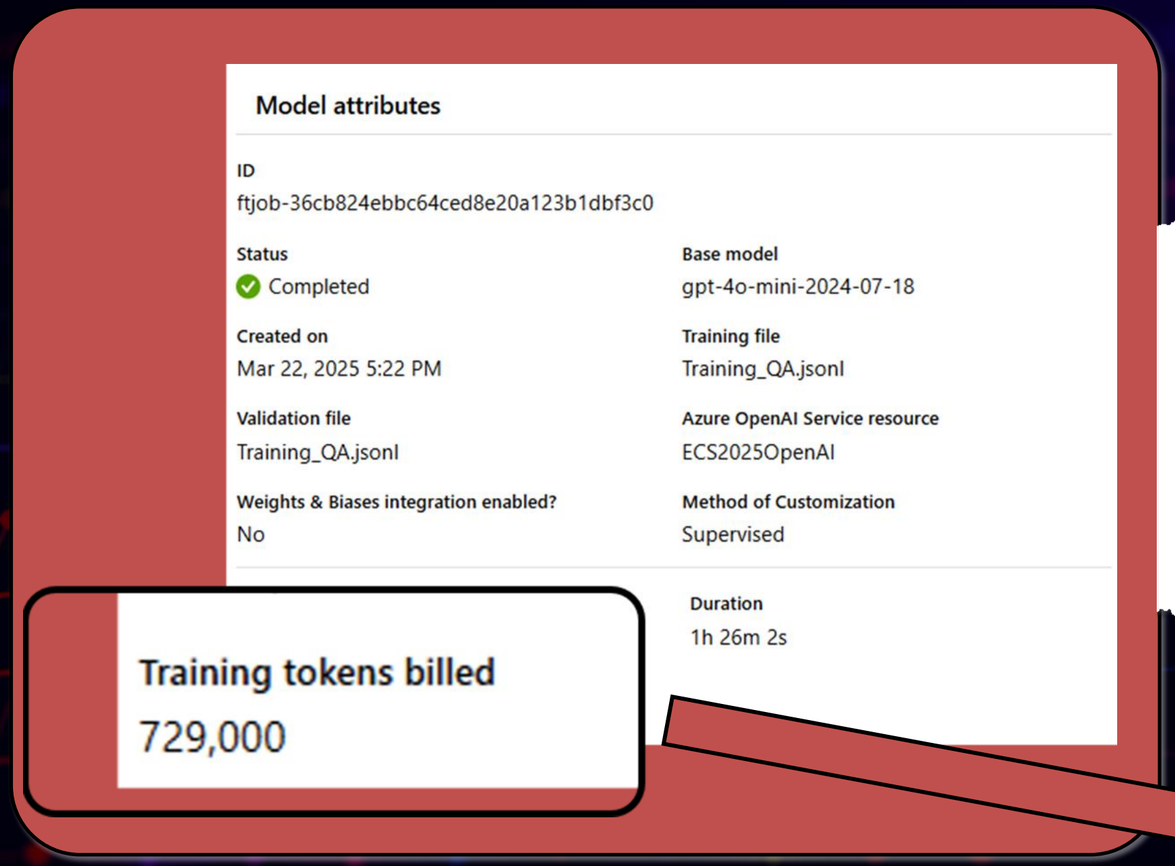
- Cost = Tokens in training file × Epochs × Price per token
- Smaller/newer models → lower token cost
- Global training = cheaper than regional if data residency not required
- Not billed for queue time, failed jobs, canceled jobs before training start

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



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

	Pro(s)	Cons(s)
RAG	Cost effective Dynamic, Up-to-date Domain Flexibility	Vector Db Dependency Relies on Data Quality Introduces Latency
Fine-Tuning	Domain specialization Improved Accuracy	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 | 04 October 2025

global AI community

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