LLM Pipelines: Seamless Integration on Embedded Devices

tinyAI Virtual Forum on Generative AI and Foundation Models on the Edge 2024

Enzo Ruedas

Equal contribution with:

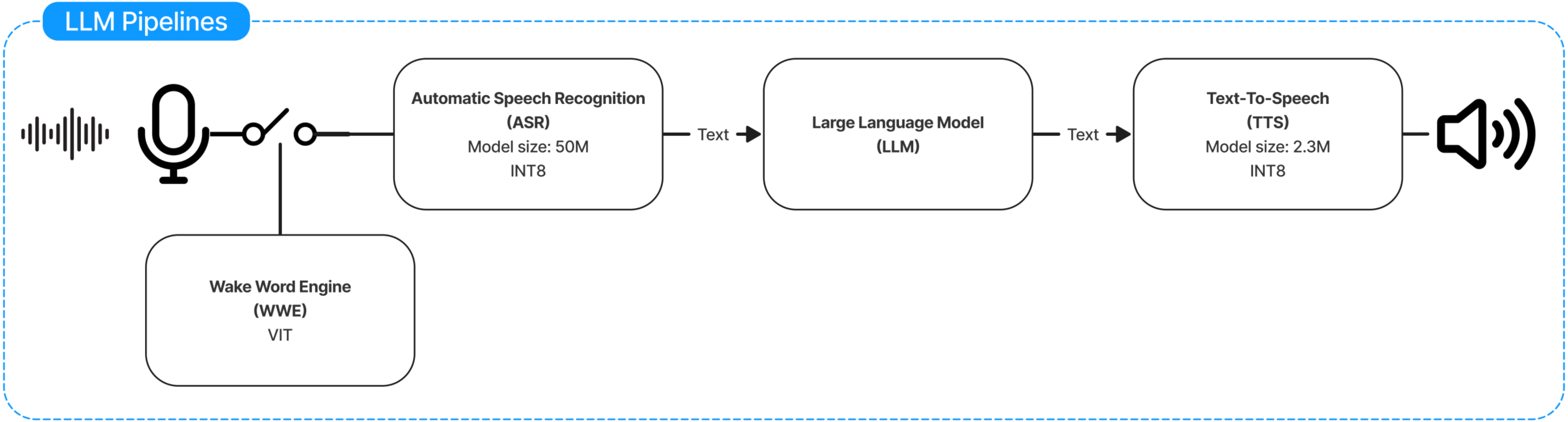
- Tess Boivin
- Baptiste Pouthier
- Laurent Pilati

Voice & Audio Team, NXP

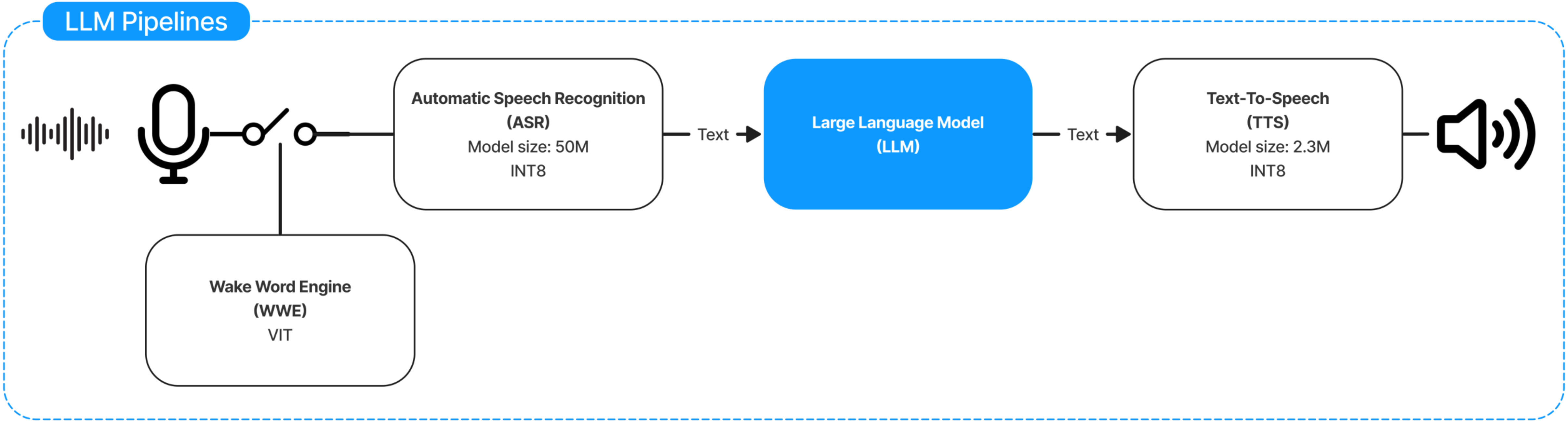
01

LLM Pipelines

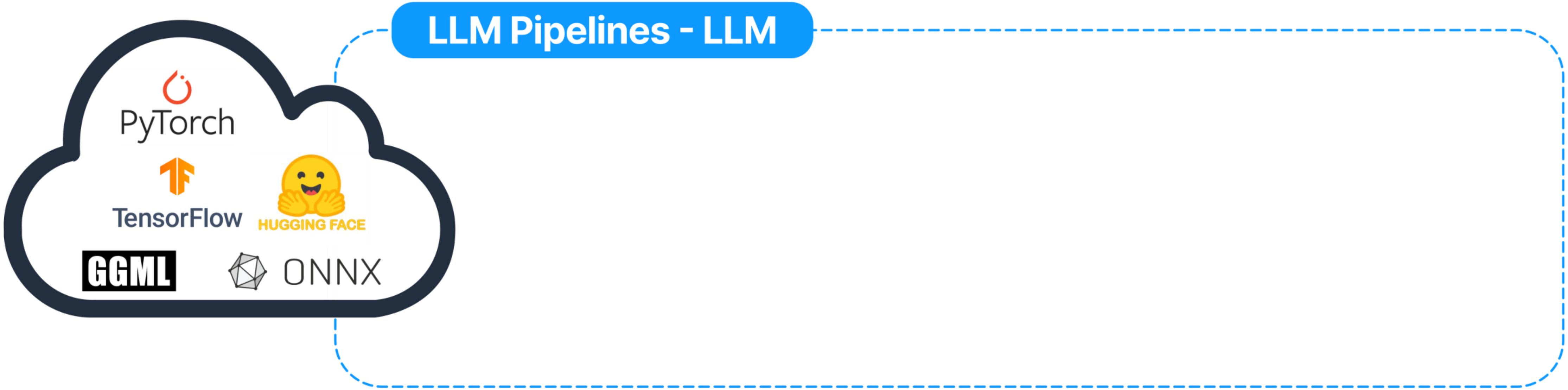
What is the LLM Pipelines project?



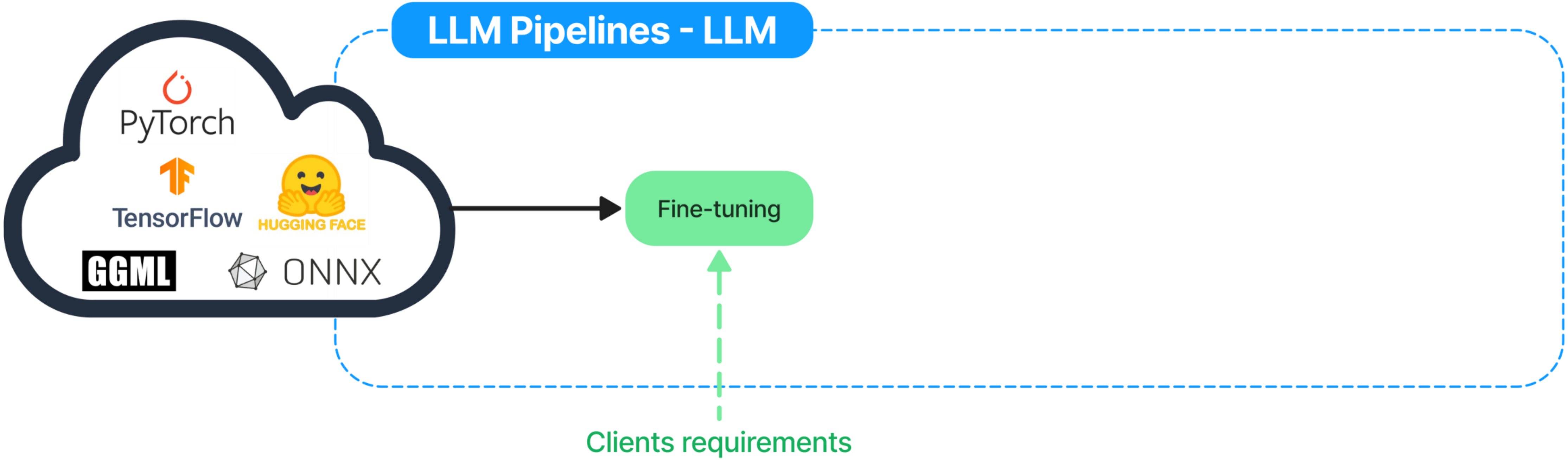
What is the LLM Pipelines project?



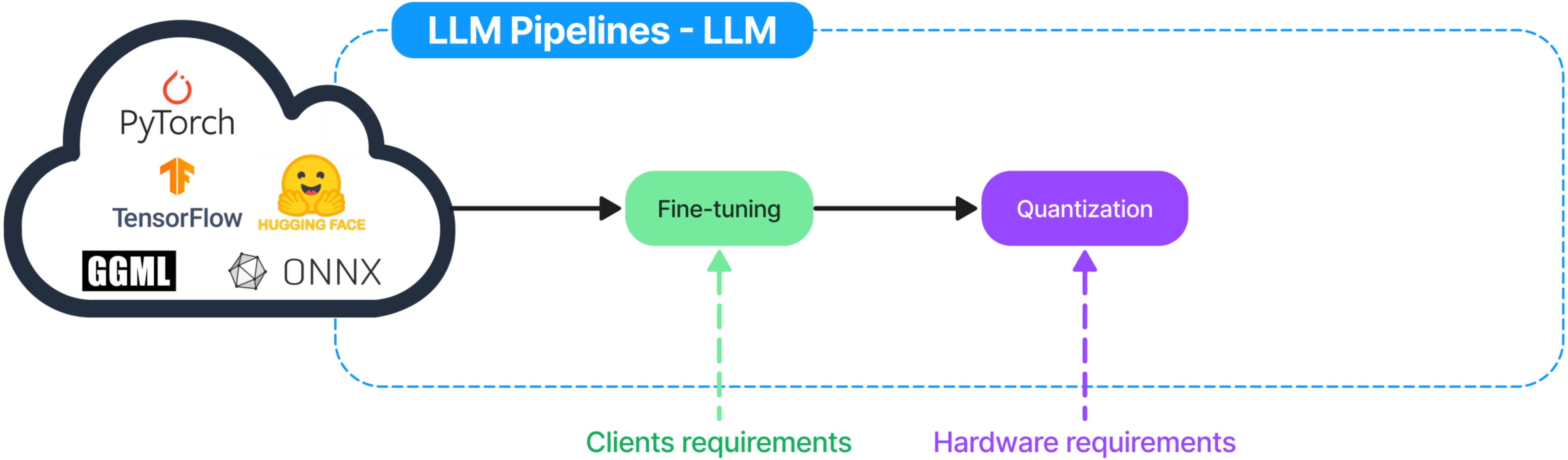
What is the LLM Pipelines project?



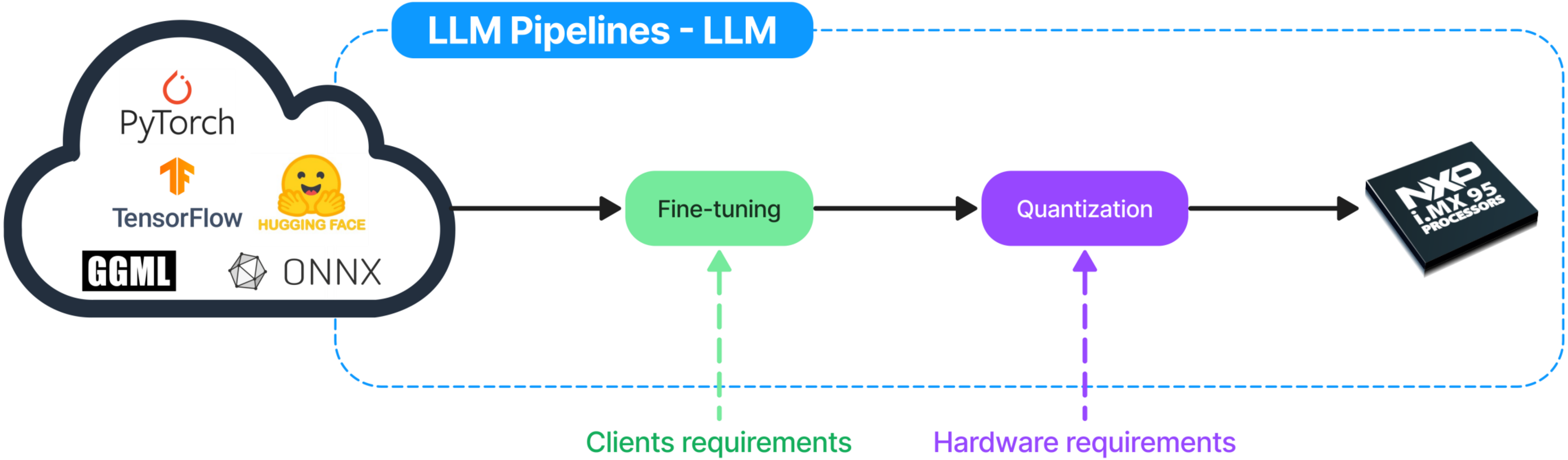
What is the LLM Pipelines project?



What is the LLM Pipelines project?



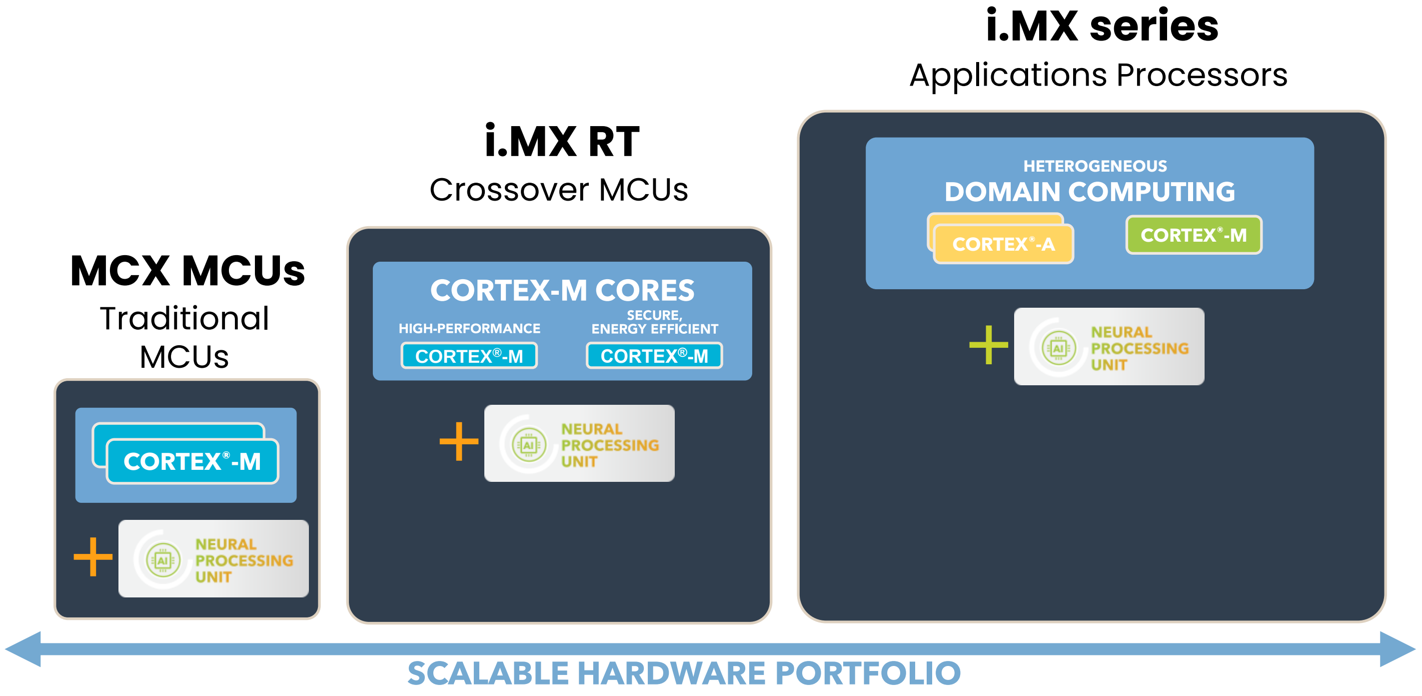
What is the LLM Pipelines project?



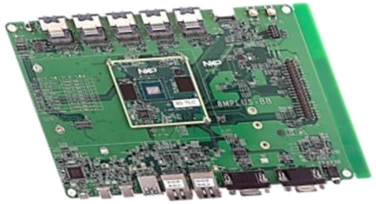
NXP MPU platforms: i.MX 8M Plus, i.MX 93, i.MX 95

Neural Processing Unit (NPU) scales Machine Learning Solutions from MCX MCUs to i.MX 9 Applications Processors

NXP eIQ Neutron **Neural Processing Unit (NPU)** hardware scales from 32 to >2k operations/cycle, for use across the range of MCUs and MPUs



NXP MPU platforms:



i.MX 8M Plus
- 4x Arm® Cortex® - A53 (1.8Ghz)
- NPU (2.3 TOPS)

i.MX 93
- 2x Arm® Cortex® - A55 (1.7Ghz)
- NPU (0.5 TOPS)

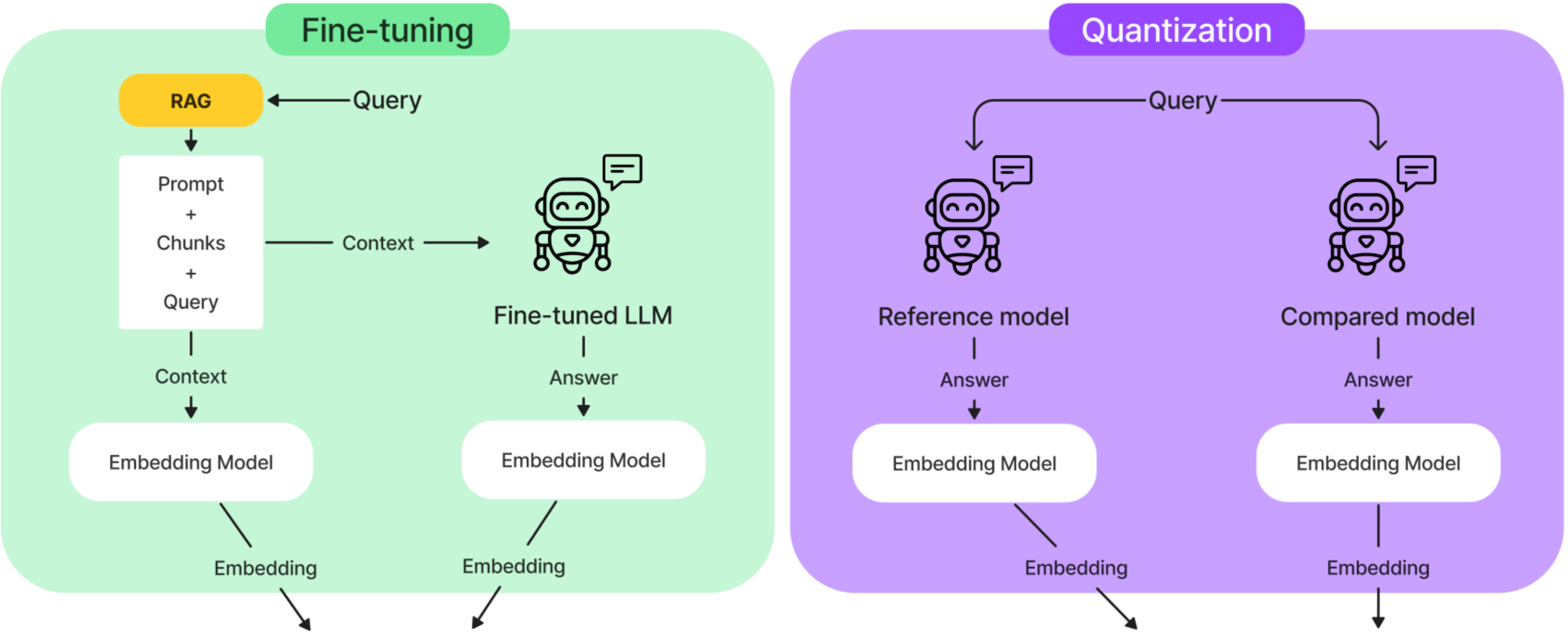
i.MX 95
- 6x Arm® Cortex® - A55 (1.7Ghz)
- eIQ Neutron NPU (2 TOPS)

02

Metrics

Our metrics

Similarity:



$$Similarity = \cos(\overrightarrow{context}, \overrightarrow{model\_answer})$$

Measure the model ability to use the given context in formulating its answer.

$$Similarity = \cos(\overrightarrow{reference\_answer}, \overrightarrow{compared\_answer})$$

Measure the quality loss of the model after quantization

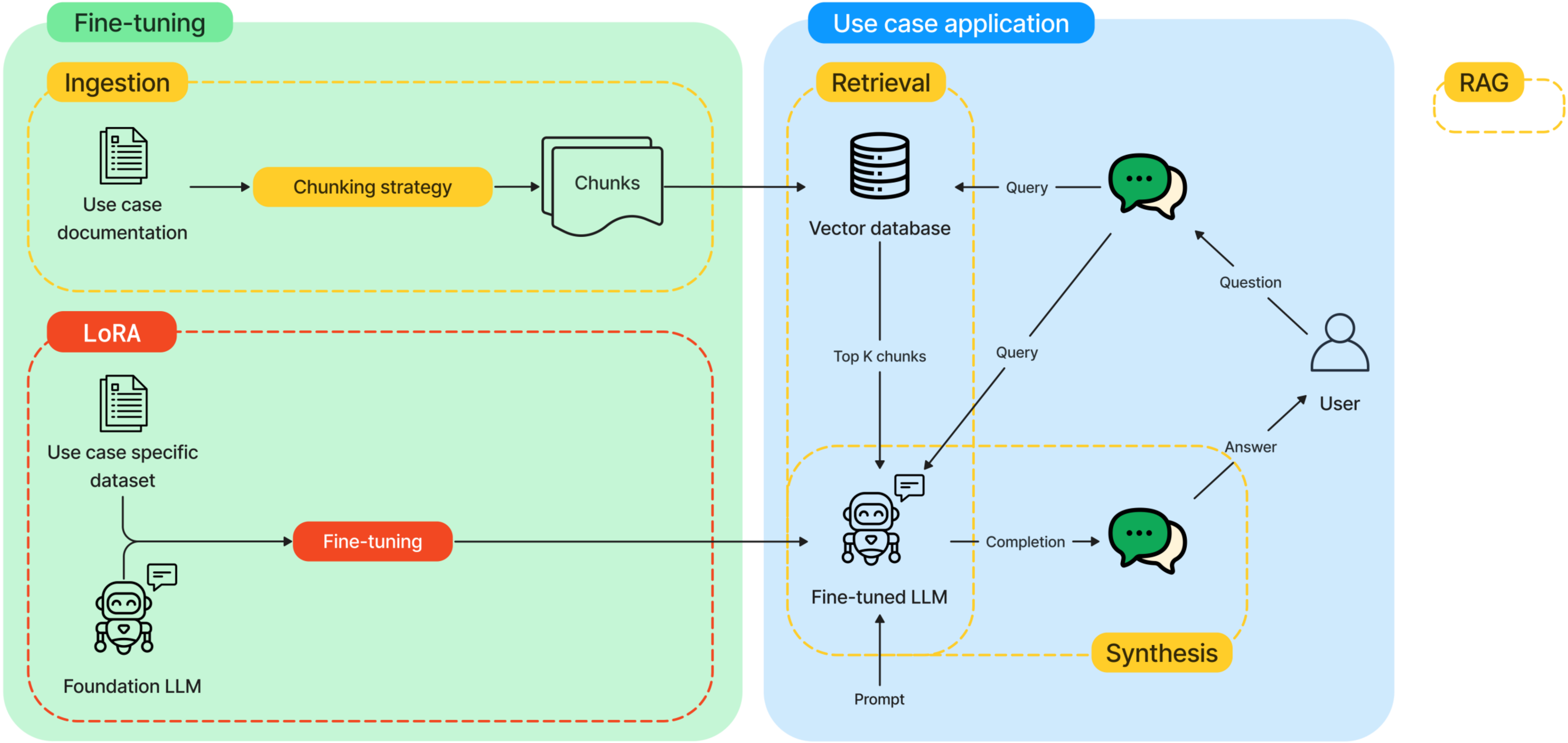
Time To First Token (TTFT): The time in seconds before the model produces its first answer token.

Token/s: Average number of token generated per second during generation time.

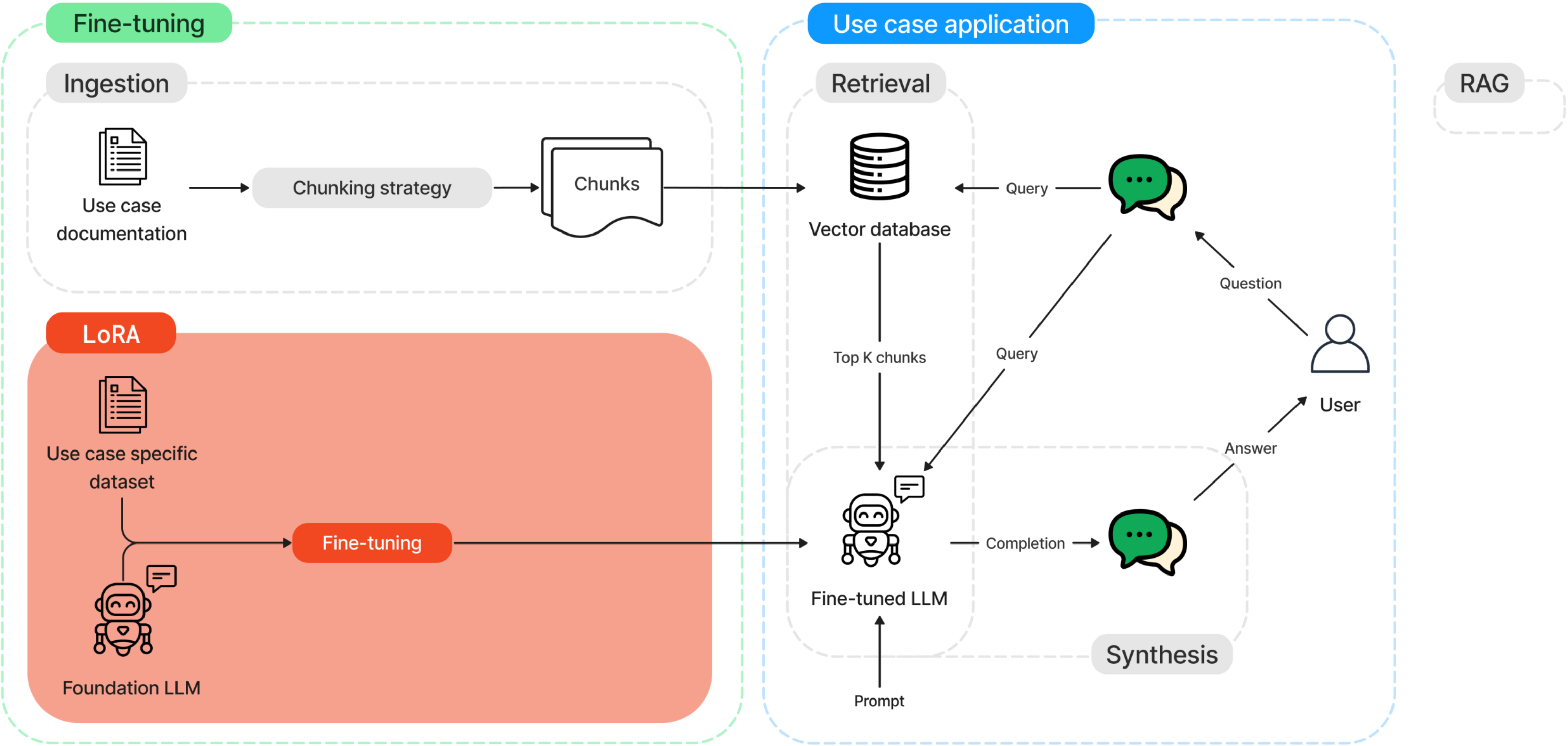
03

Fine-tuning

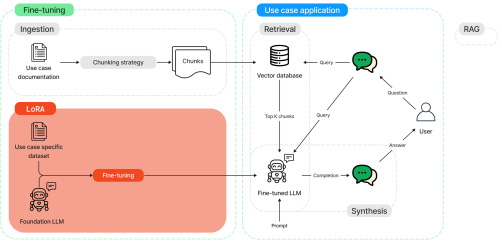
Fine-tuning strategies



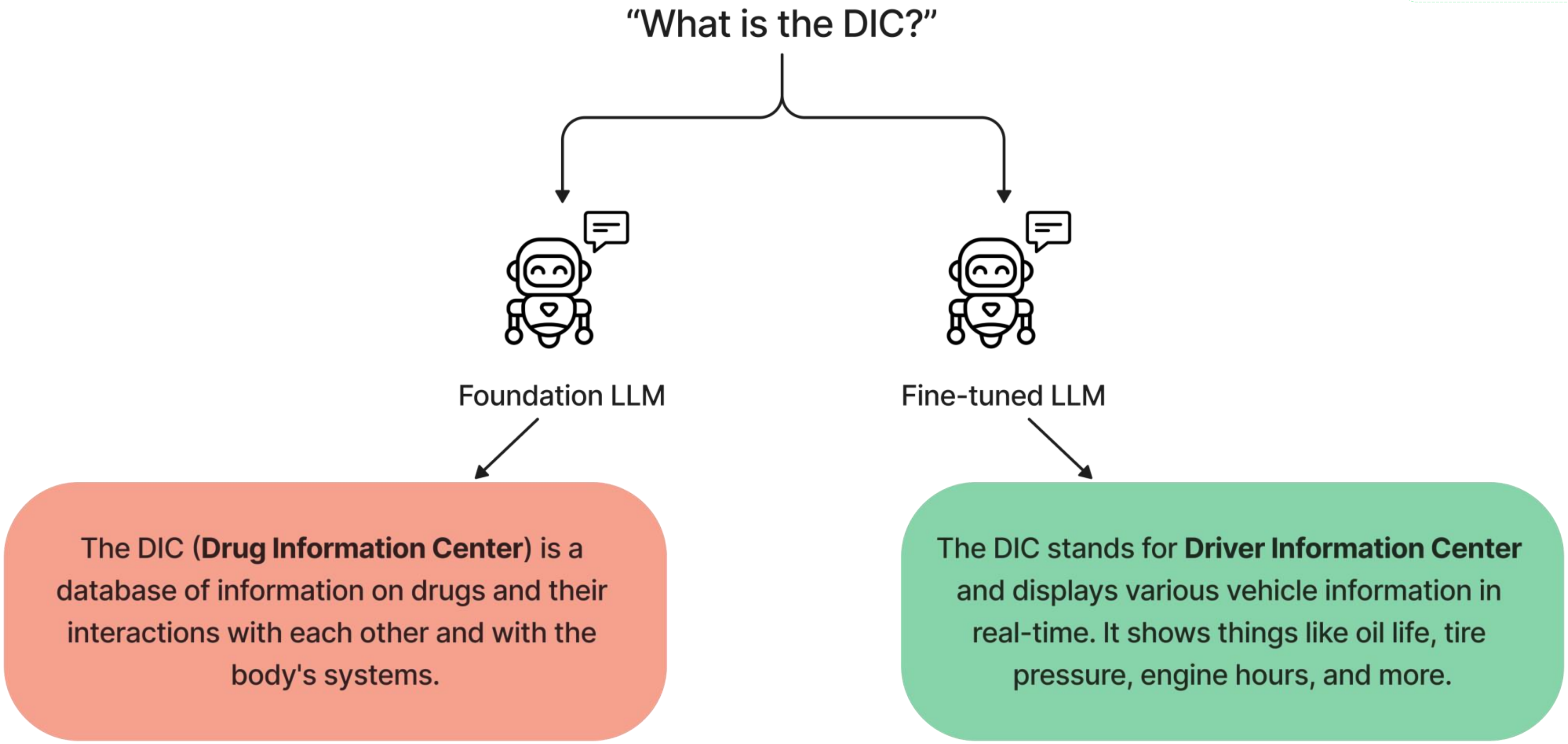
The Low Rank Adaptation (LoRA) for fine-tuning



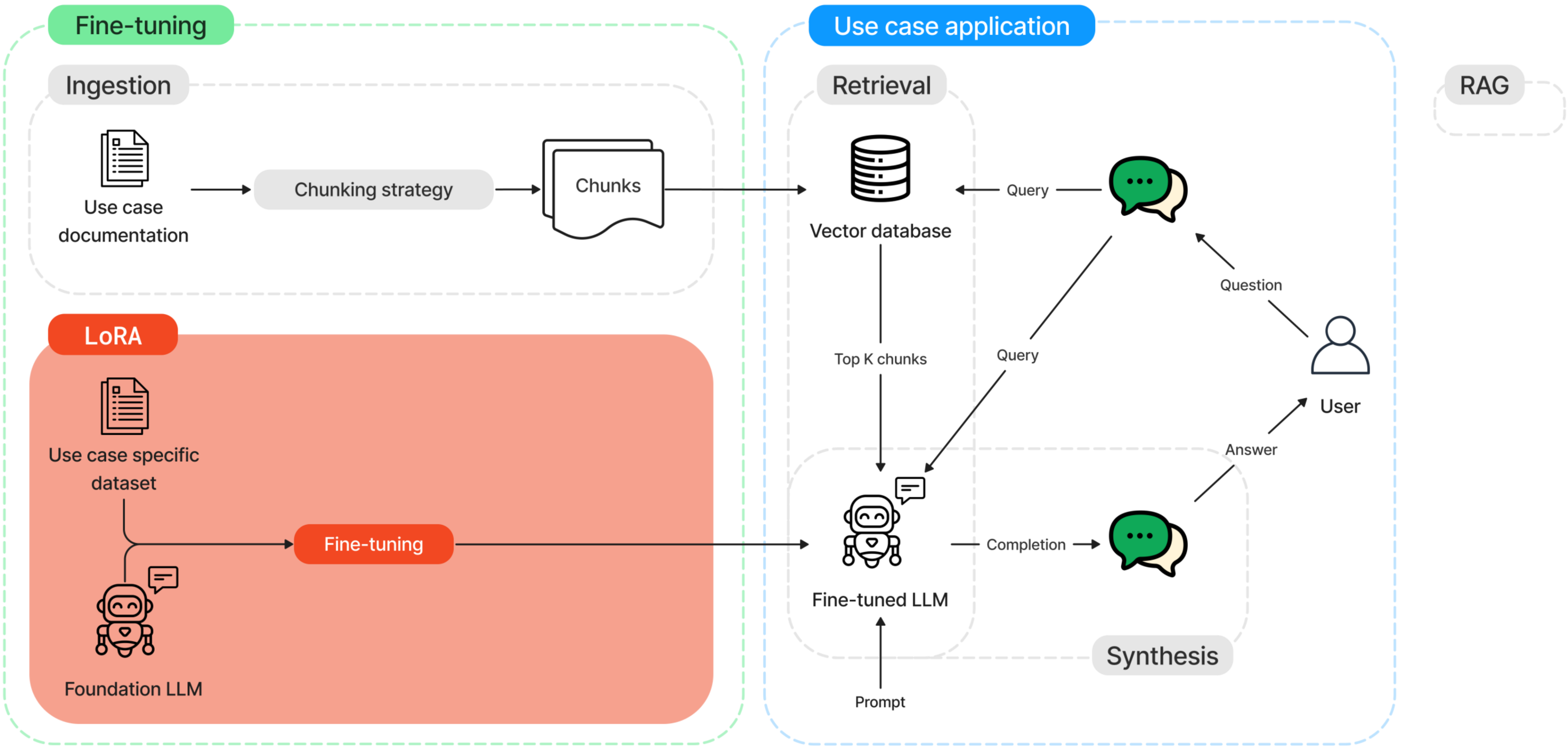
The Low Rank Adaptation (LoRA) for fine-tuning



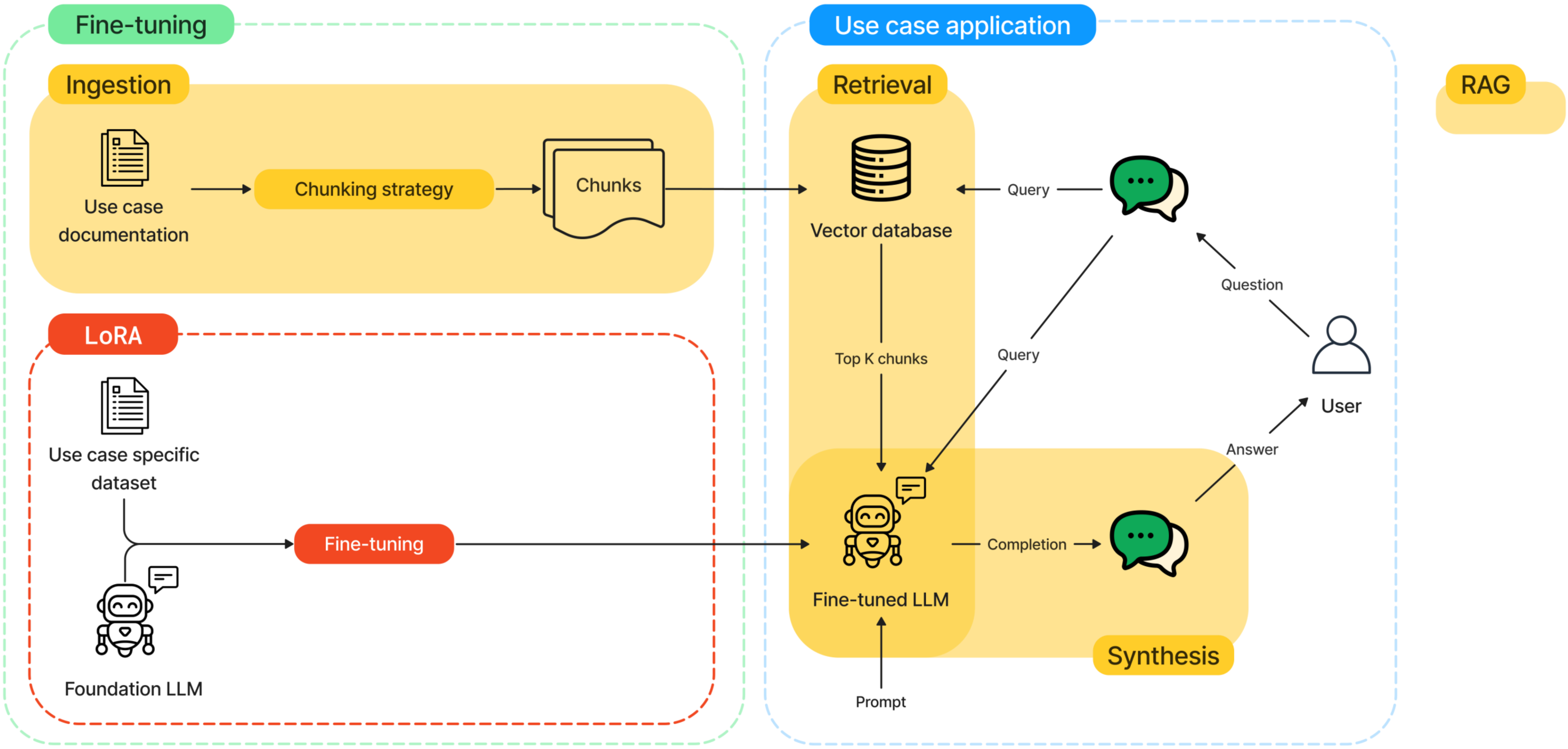
Results - automotive use case:



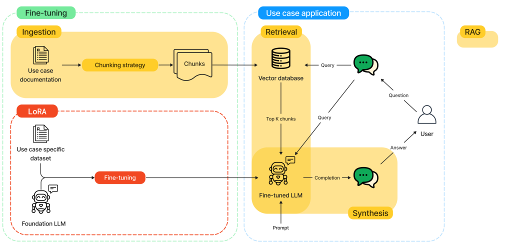
The Low Rank Adaptation (LoRA) for fine-tuning



The Retrieval-Augmented Generation (RAG) for fine-tuning



The Retrieval-Augmented Generation (RAG) for fine-tuning



Results on i.MX 95 – TinyLlama 1B - automotive use case:

- Out of domain question:

```
User: What is the weather today?
Please ask something about the car?
```

Similarity	0.236
TTFT	∅
Time to detect	0.08 s

- In domain question:
  - Without RAG:

```
User: What is the contact number for assistance?
The contact number for assistance is not mentioned in the given text.
```

Hallucination

Similarity	0.079
TTFT	1.5 s
Token/s	5.1

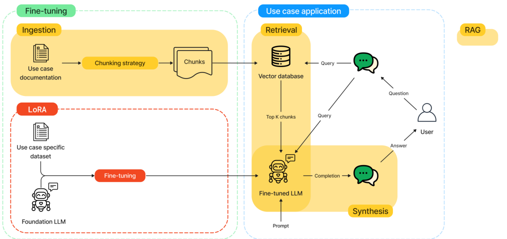
  - With RAG:

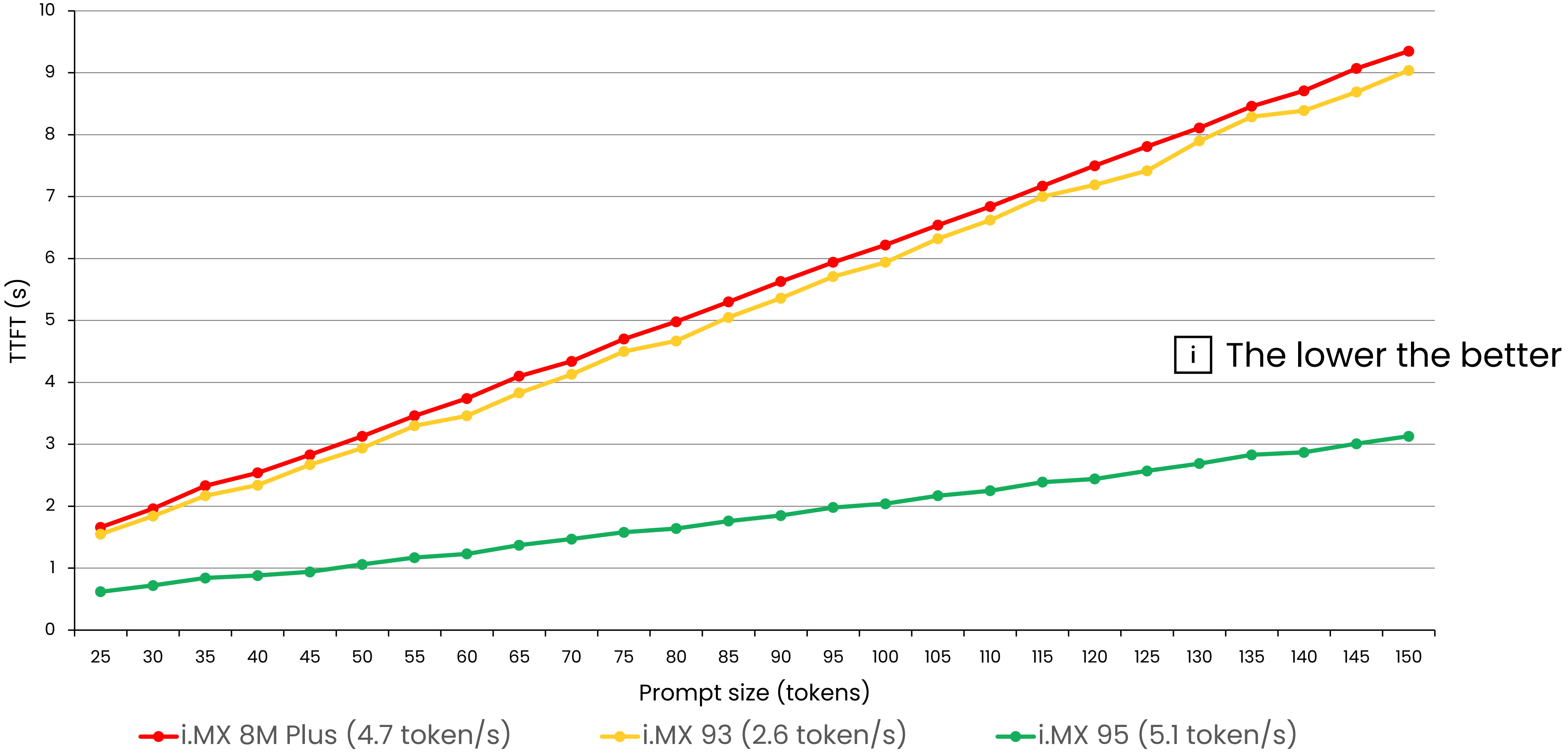
```
User: What is the contact number for assistance?
The contact number for assistance is 1-888-881-3302.
```

Real value

Similarity	0.818
TTFT	3.9 s
Token/s	5.1

The Retrieval-Augmented Generation (RAG) for fine-tuning

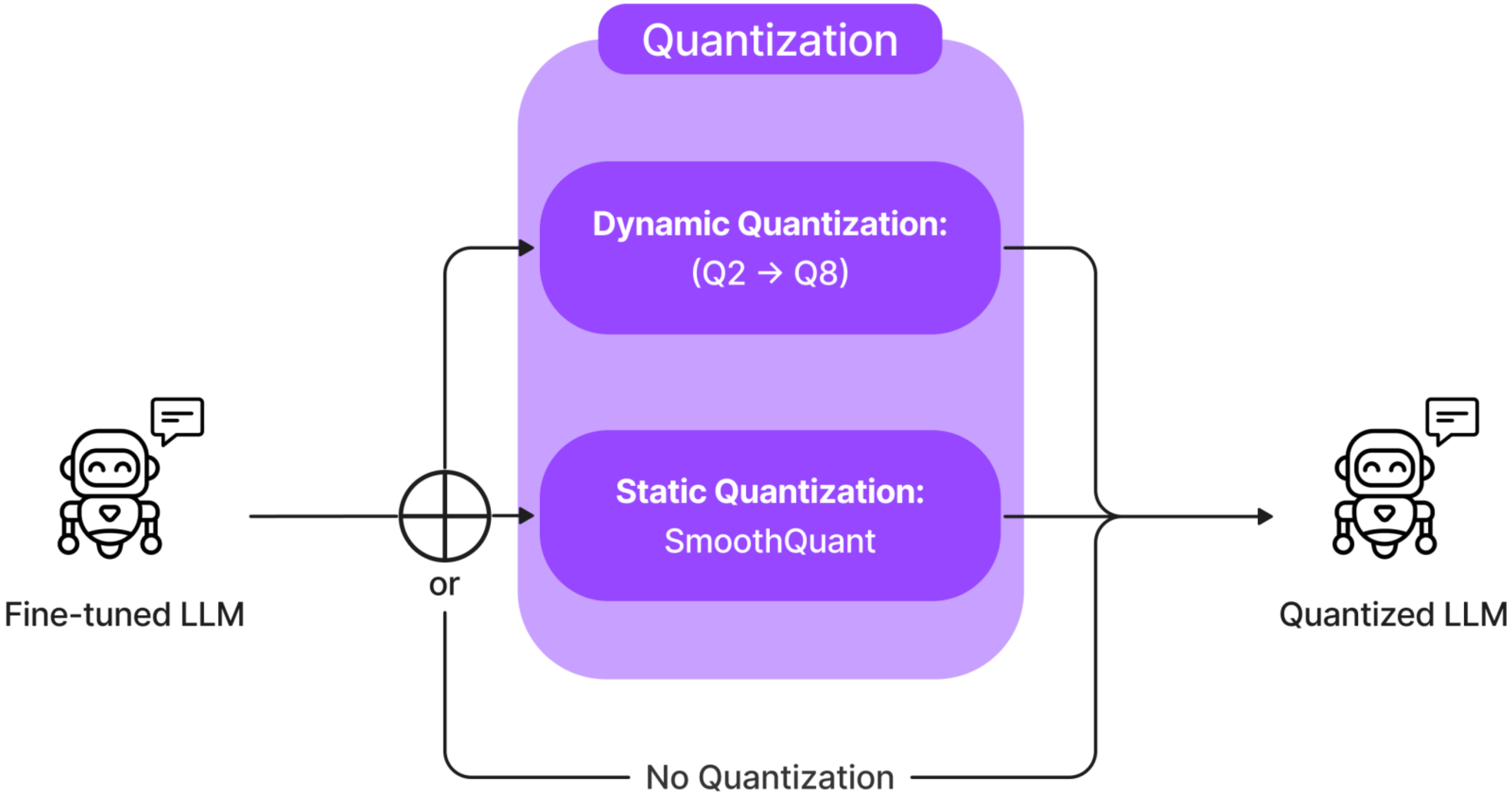




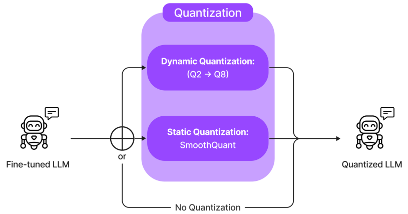
04

Quantization

Quantization strategies

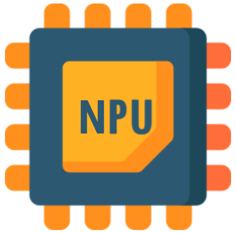


About Quantization



Reduce the model memory footprint

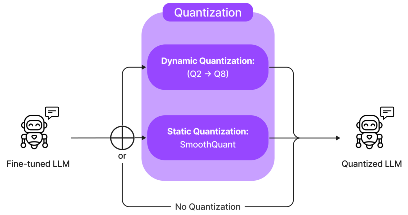
Accelerate the inference computation on CPU & NPU



Necessary for NPU enablement with integer operations

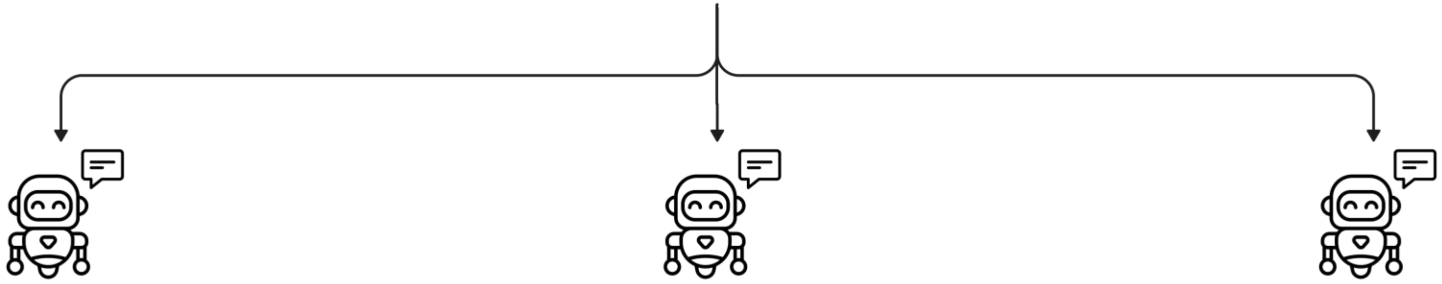
Minimize the loss of response quality after quantization

Quantization strategies



Results on i.MX 95 – TinyLlama 1B - automotive use case:

"Can I play my music on the car?"



LLM
Without Quantization (float16/32)

Yes, you can play your music on the car using a Bluetooth speaker or a USB drive.

20 tokens

Similarity	1
TTFT	12.6 s
Token/s	2.5

Quantized LLM
Dynamic Quantization (int8)

You can play your favorite music on the car using a Bluetooth speaker or a USB drive.

19 tokens

Similarity	0.887
TTFT	3.8 s
Token/s	5.6

Quantized LLM
Static Quantization SmoothQuant (int8)

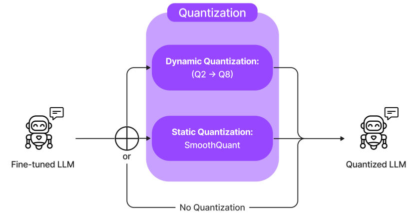
To play your music through the car, you need to connect your smartphone or tablet to the car's Bluetooth system.

27 tokens

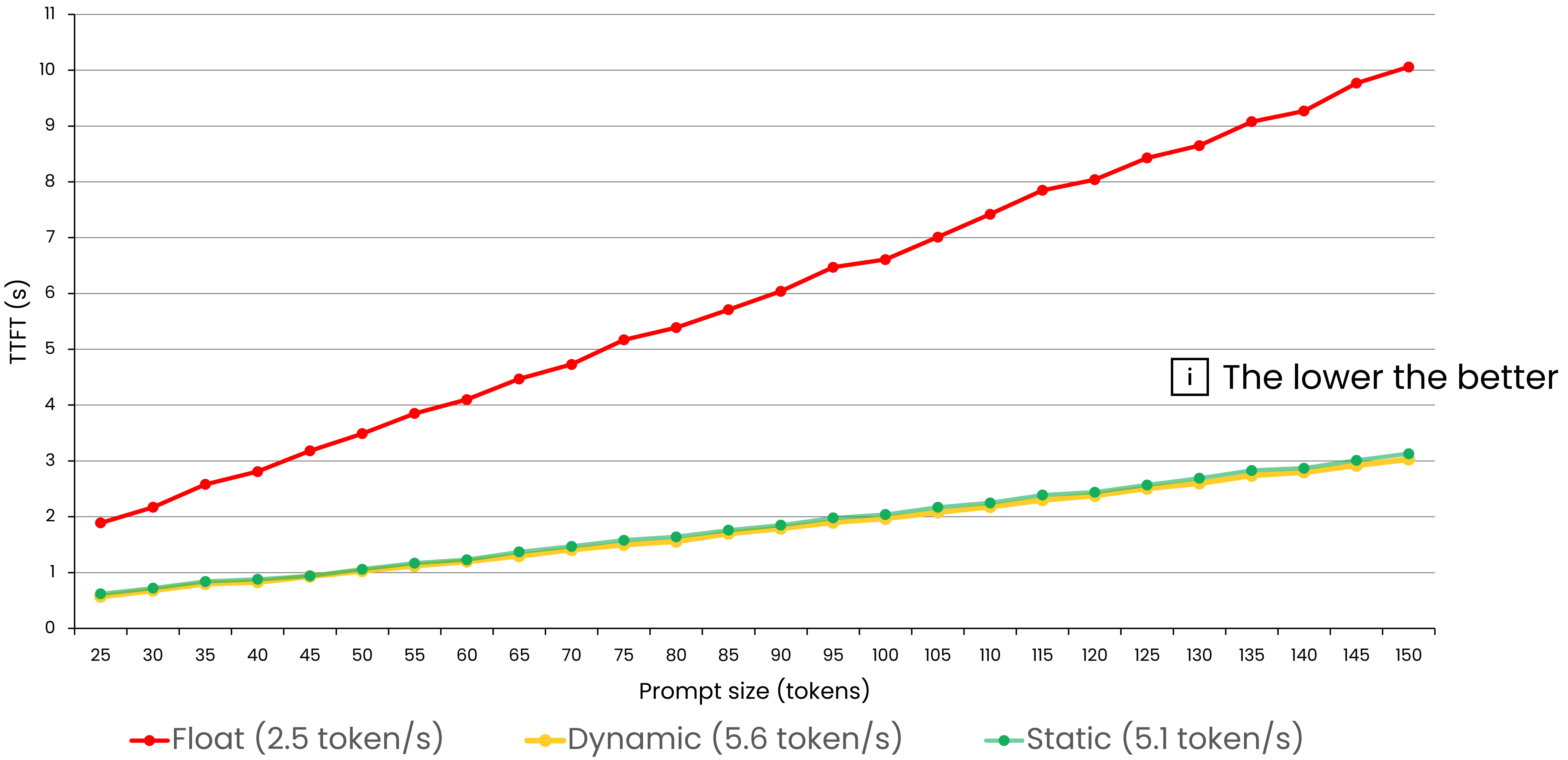
Similarity	0.804
TTFT	3.9 s
Token/s	5.1

Quantization strategies

Results on i.MX 95

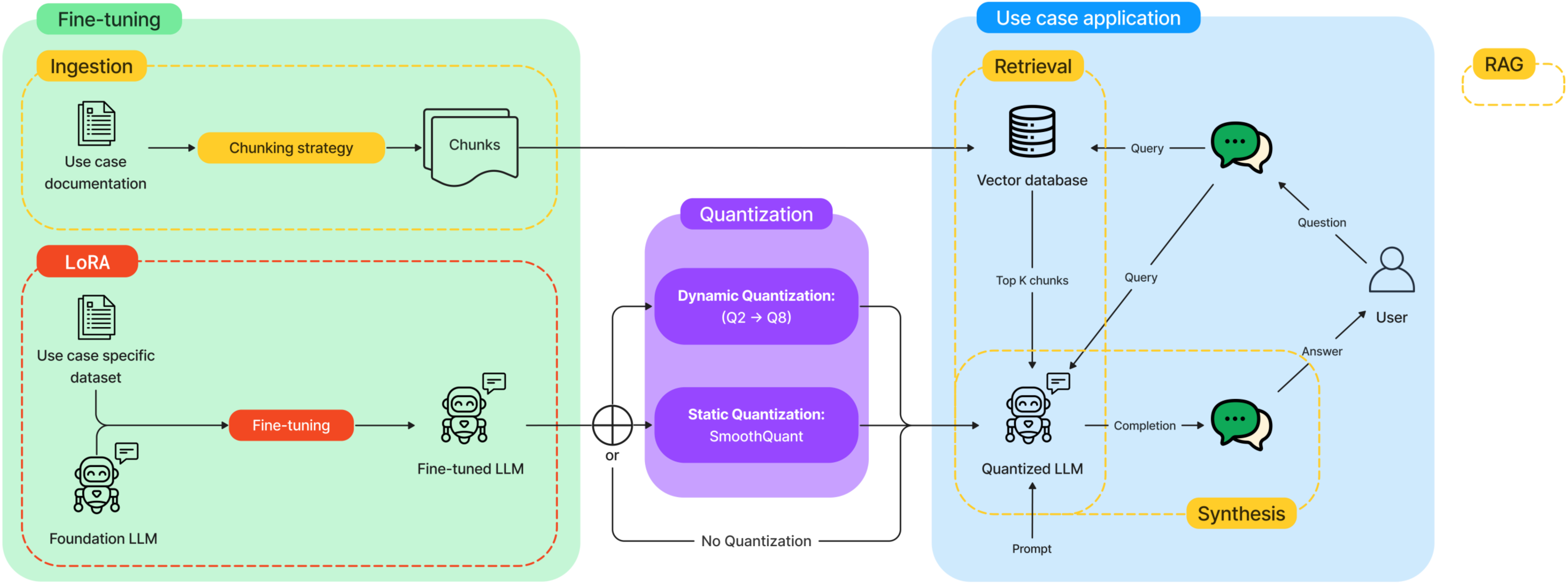


Time To First Token (TTFT) vs. Prompt size



05

Final architecture


Fine-tuning
Ingestion
Use case documentation
Chunking strategy
Chunks
LoRA
Use case specific dataset
Fine-tuning
Fine-tuned LLM
Foundation LLM
or
Quantization
Dynamic Quantization: (Q2 → Q8)
Static Quantization: SmoothQuant
No Quantization
Use case application
Retrieval
Vector database
Query
Top K chunks
Query
Question
User
Answer
Quantized LLM
Completion
Synthesis
Prompt
RAG

nxp.com