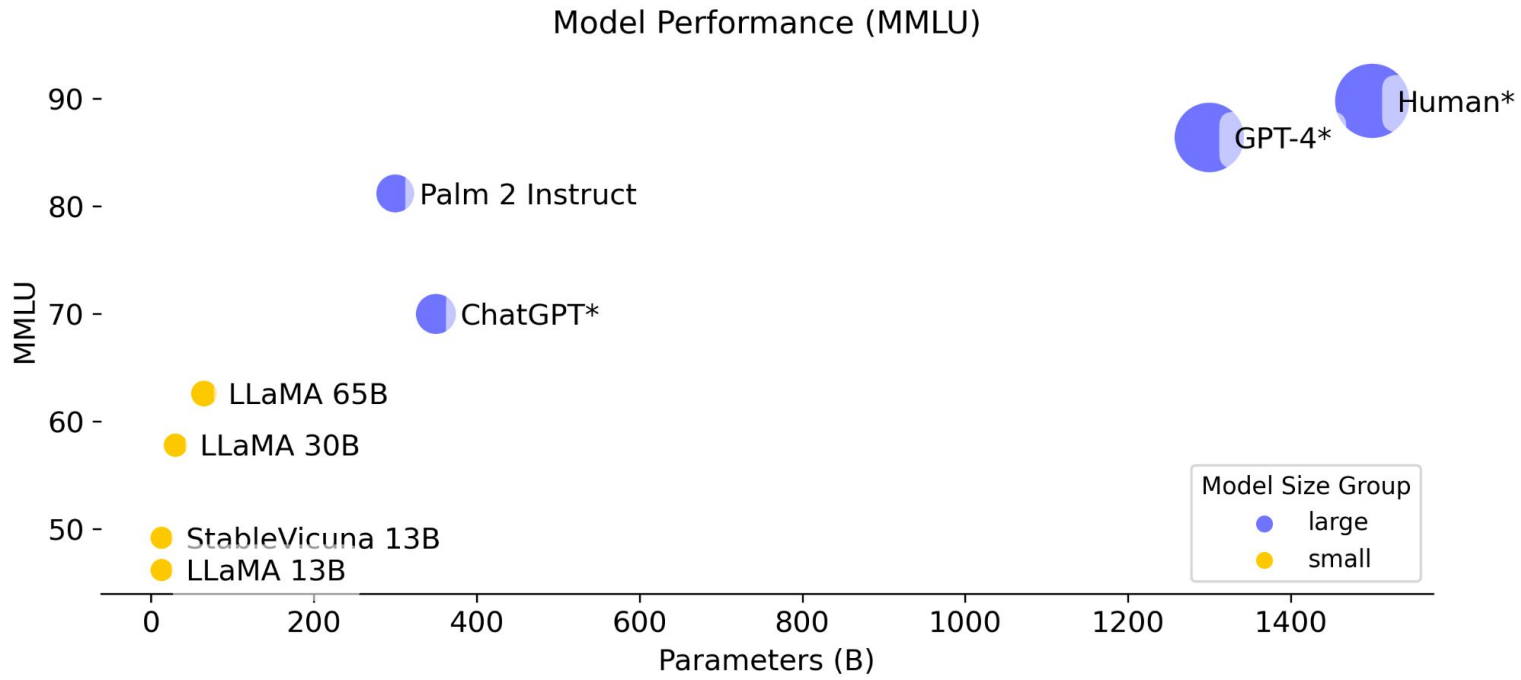
Fine-Tuning Large Language Models with LORA

Text2Json

LLM size still increase?



*Exact model size is unknown. | Data from InstructEval GitHub.

Source: Understanding Size Tradeoffs with Generative Models

GPU Performance fp32 2020 – 2023.11

LLM size (3 year -> x10)

- 2020 - GPT3 (175B)
- 2023 - Gemini Ultra (1760B)

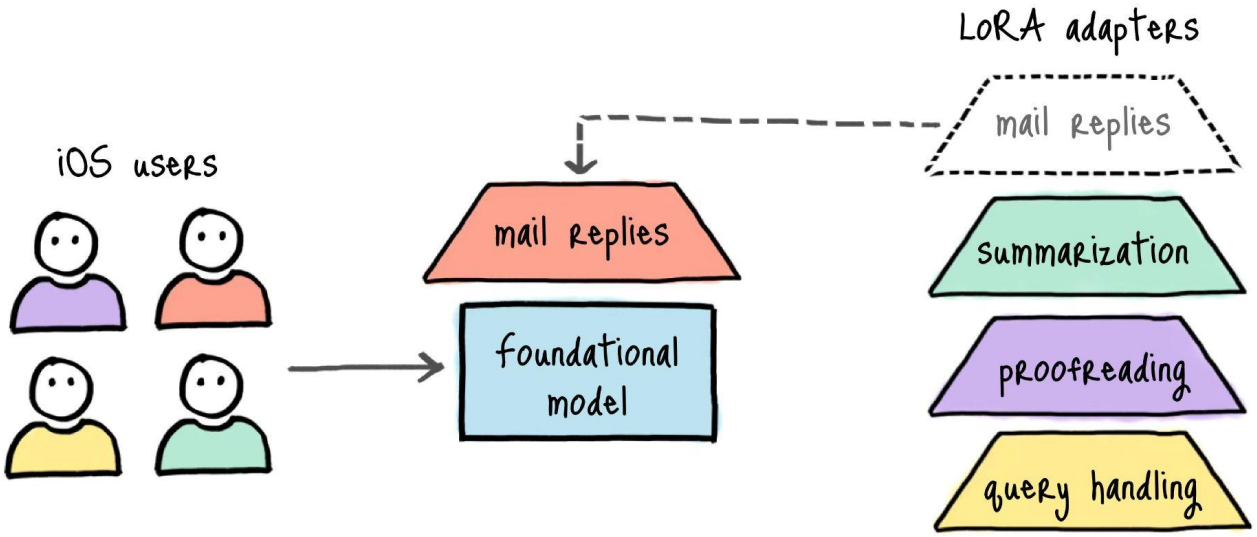
GPU (3 year -> x6.5)

- 2020 - A100 (TF32 152 TFLOPS - Mem 80GB)
- 2023 - H200 (TF32 989 TFLOPS - Mem 141GB)

Source: GPU Performance Rankings

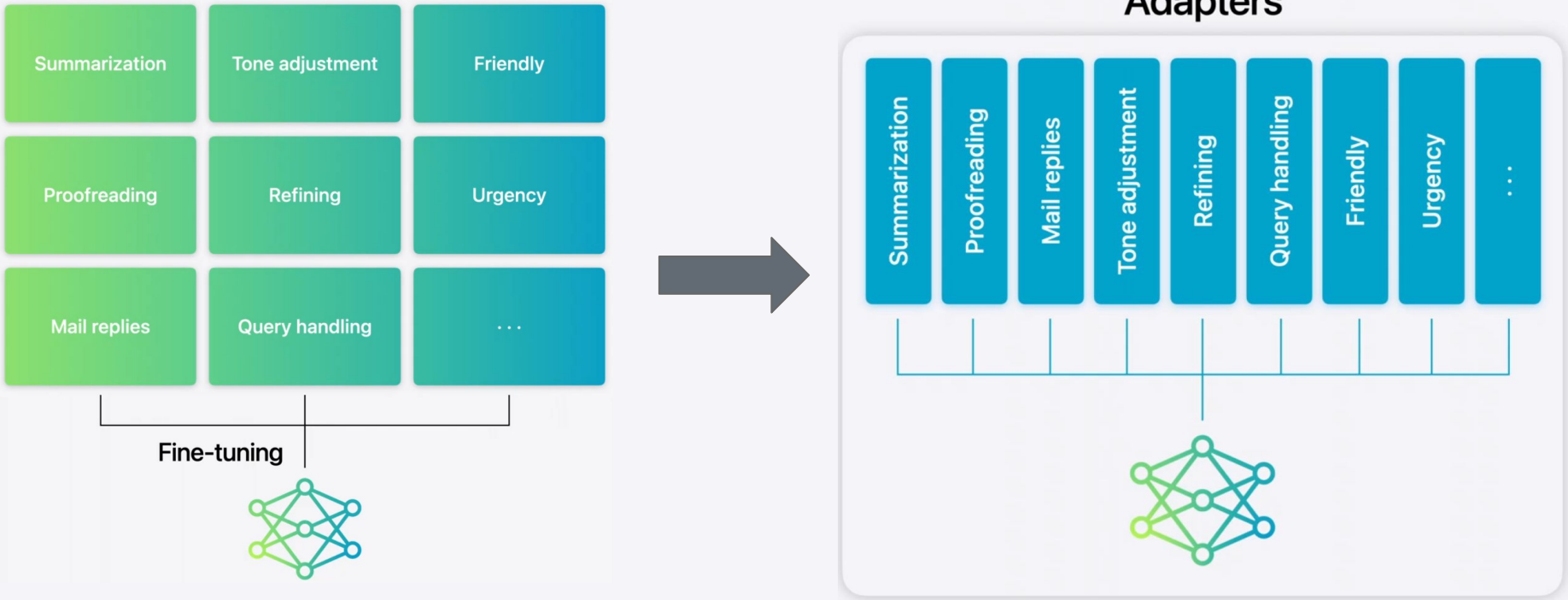
H200 SXM	141GB HBM3e - 2023.11		67	TFLOPS
NVIDIA H100 SXM5	80GB HBM3 - 2022.03		66.91	TFLOPS
NVIDIA H100 SXM5 64 GB	64GB HBM3 - 2023.03		66.91	TFLOPS
NVIDIA H100 SXM5 80 GB	80GB HBM3 - 2023.03	~126,000$	66.91	TFLOPS
NVIDIA H100 SXM5 96 GB	96GB HBM3 - 2023.03		62.08	TFLOPS
NVIDIA H100 PCIe 96 GB	96GB HBM3 - 2023.03		62.08	TFLOPS
NVIDIA H800 SXM5	80GB HBM3 - 2023.03		59.3	TFLOPS
NVIDIA H100 CNX	80GB HBM2e - 2023.03		53.84	TFLOPS
NVIDIA H100 PCIe	80GB HBM2e - 2022.03		51.22	TFLOPS
NVIDIA H100 PCIe 80 GB	80GB HBM2e - 2023.03		51.22	TFLOPS
NVIDIA H800 PCIe 80 GB	80GB HBM2e - 2023.03		51.22	TFLOPS
NVIDIA A100 PCIe	40GB HBM2e - 2020.06		19.49	TFLOPS
NVIDIA A100 SXM4 40 GB	40GB HBM2e - 2020.05		19.49	TFLOPS
NVIDIA A100 SXM4 80 GB	80GB HBM2e - 2020.11	~41,000$ - 52,609$	19.49	TFLOPS
NVIDIA A800 PCIe 40 GB	40GB HBM2e - 2022.11		19.49	TFLOPS
NVIDIA A800 PCIe 80 GB	80GB HBM2e - 2022.11		19.49	TFLOPS
NVIDIA A800 SXM4 80 GB	80GB HBM2e - 2022.08		19.49	TFLOPS

Use Case: Apple Intelligence (Adapter)

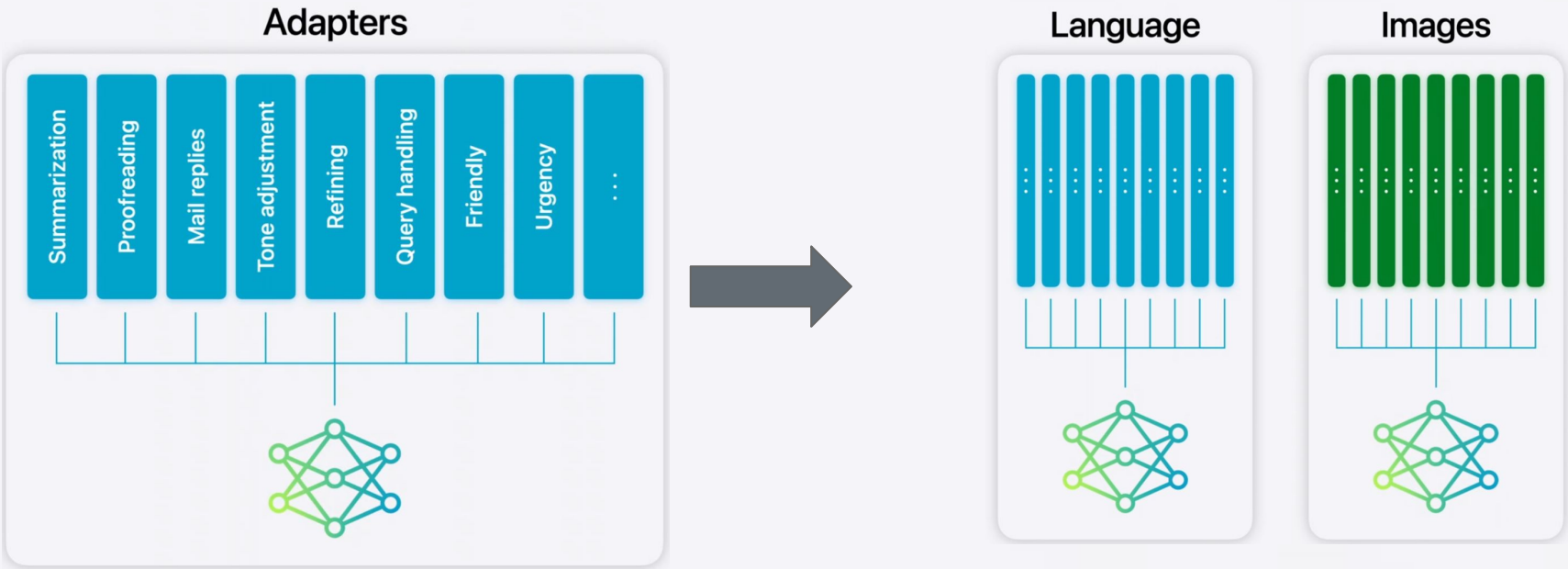


Source: X: LoRa adapters

Use Case: Apple Intelligence (Adapter)



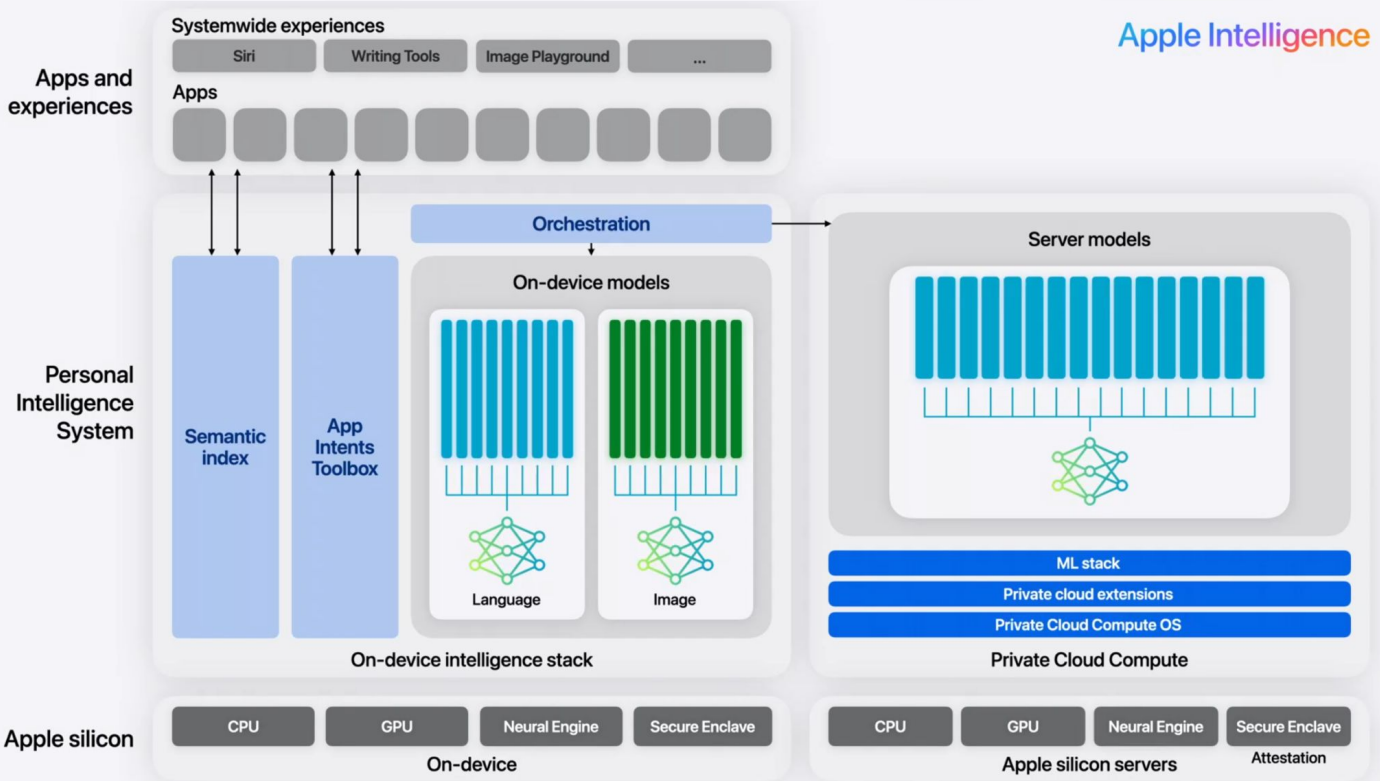
Use Case: Apple Intelligence (Adapter)



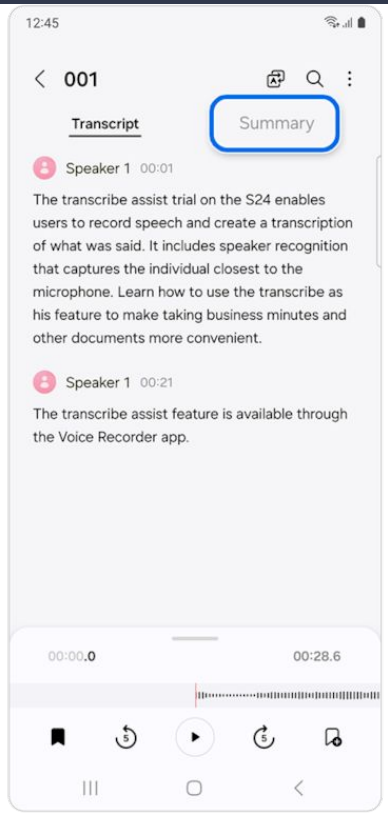
Compressing Model
Quantization from 16bit -> 4bit
Still Maintain model quality

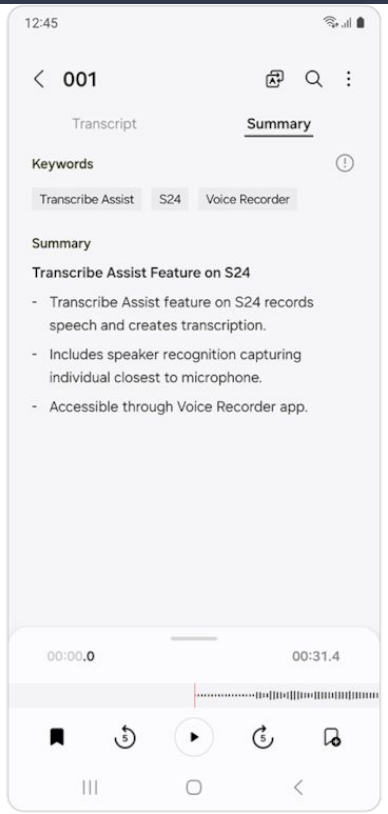
Optimize shortest time to process prompt and response
(Speculative decoding, Context pruning, Group query attention)

Use Case: Apple Intelligence (Architecture)



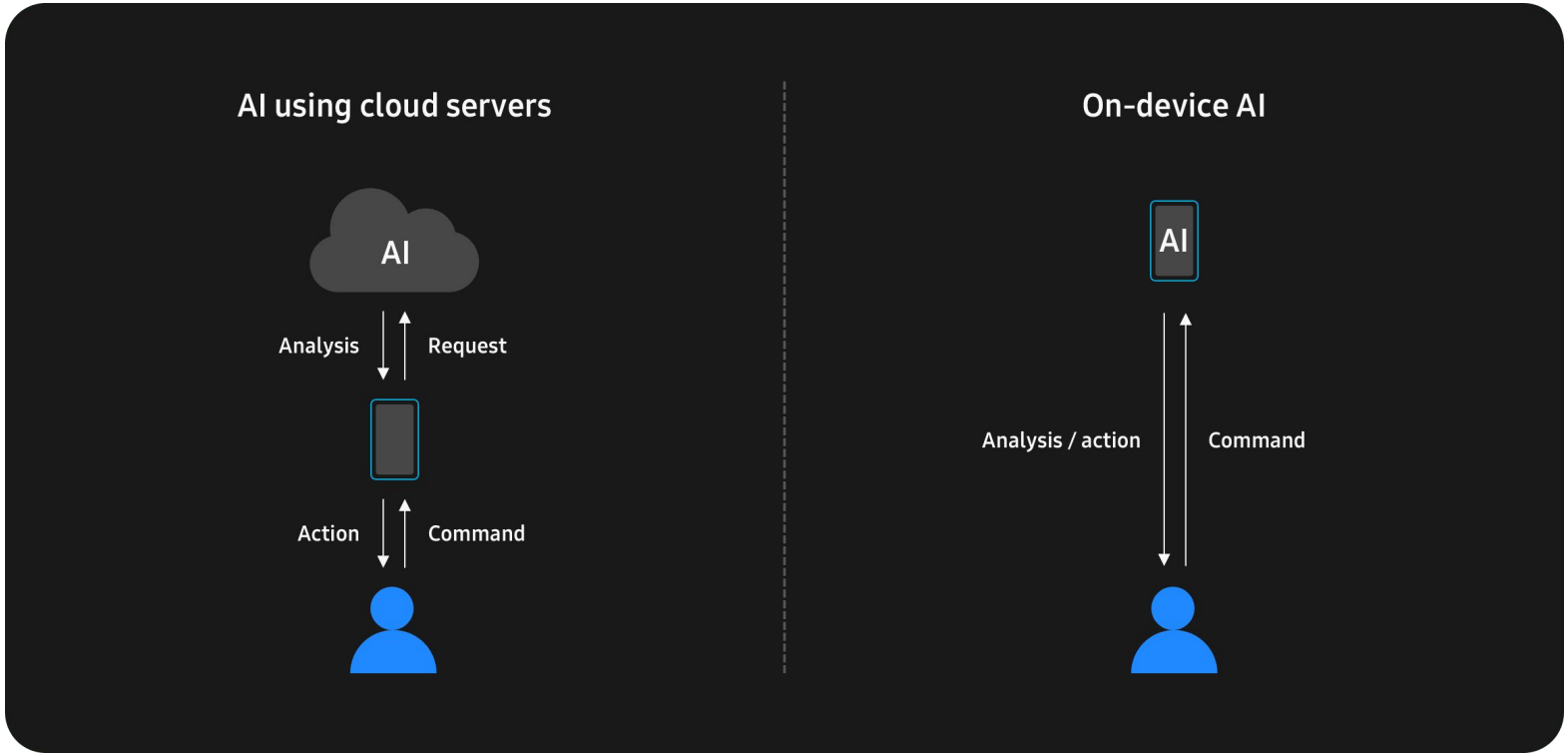
Use Case: Samsung AI





Source: Transcript Assist

AI on Cloud/Device



Source: Transcript Assist

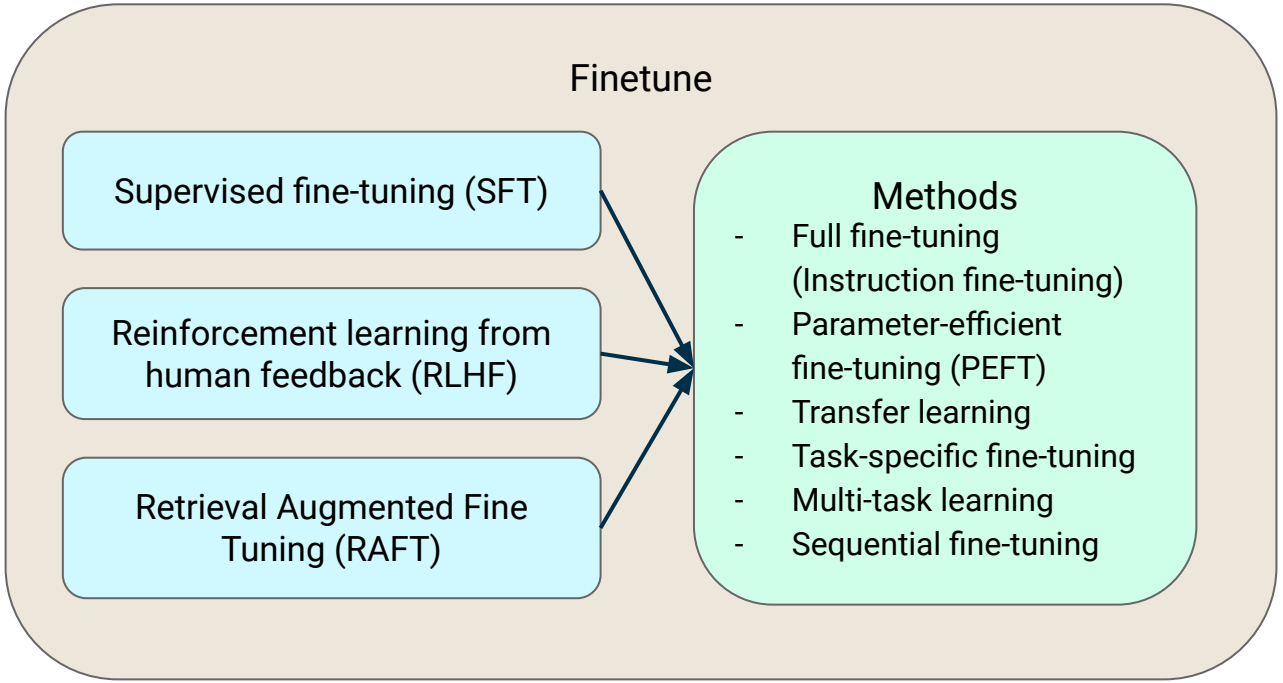
Improve LLM model performance

Prompt Engineer

Speculative decoding

Context pruning

Group query attention

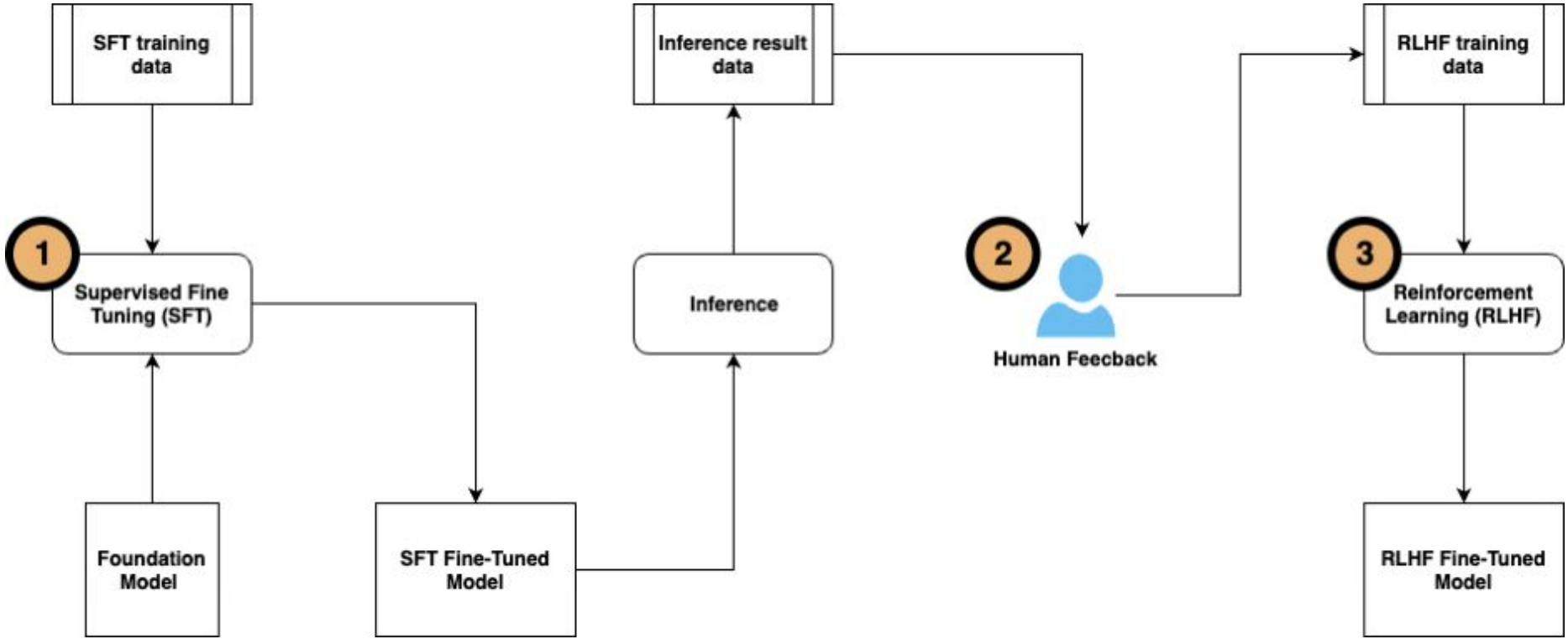


Auccracy

Latency

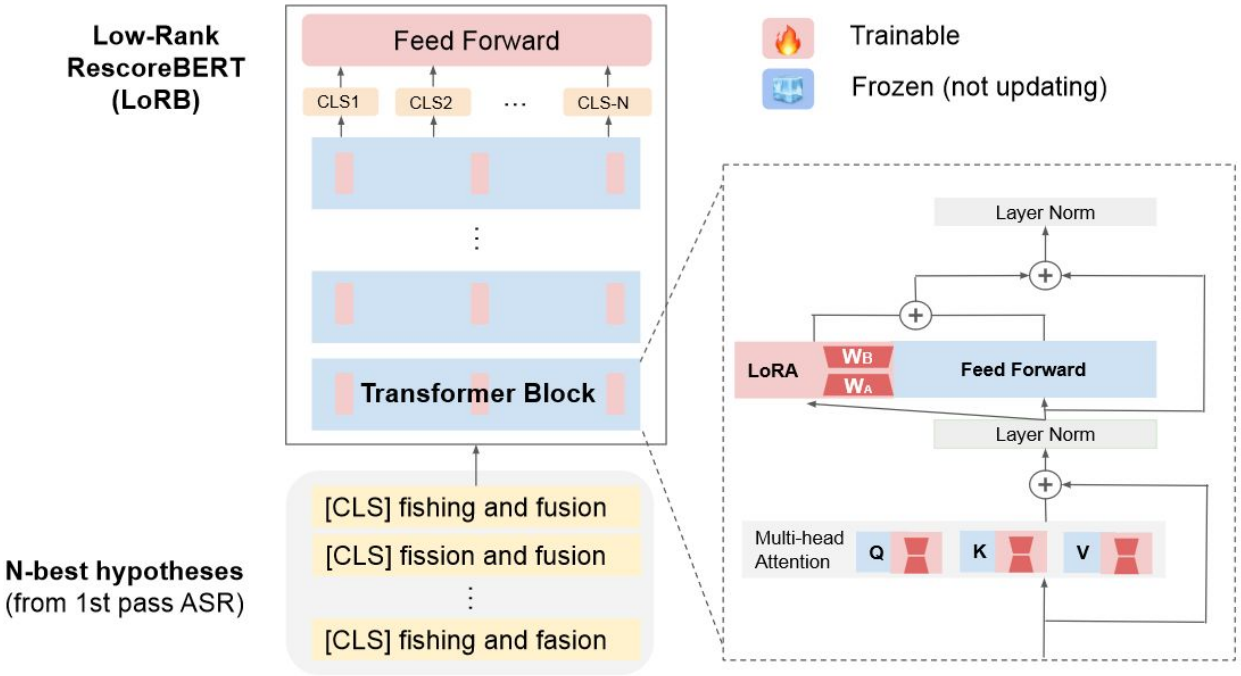
Source: Fine-tuning large language models (LLMs) in 2024

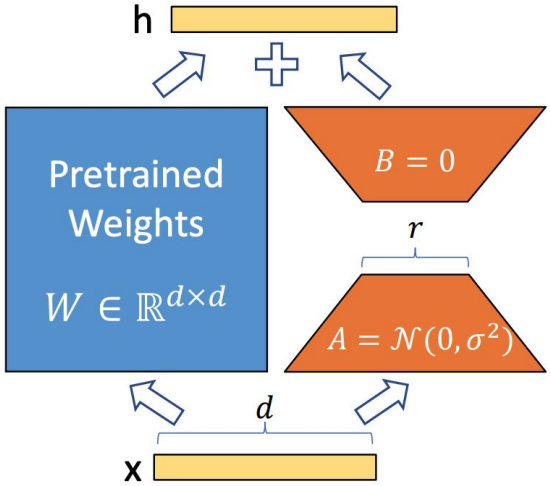
Improve bot response with supervised fine-tuning and reinforcement learning



Source: Improve LLM performance with human and AI feedback

LoRa (Low-Rank Adaptation)

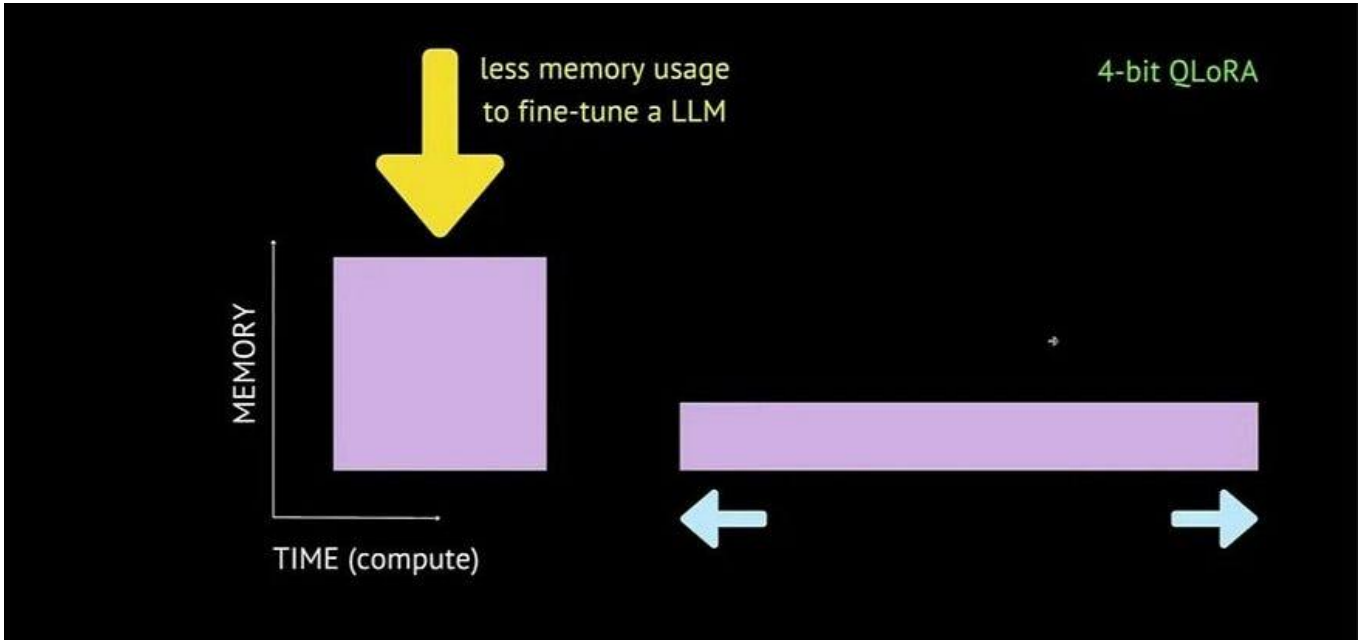




$$h = W_0 x + \Delta W x = W_0 x + BAx$$

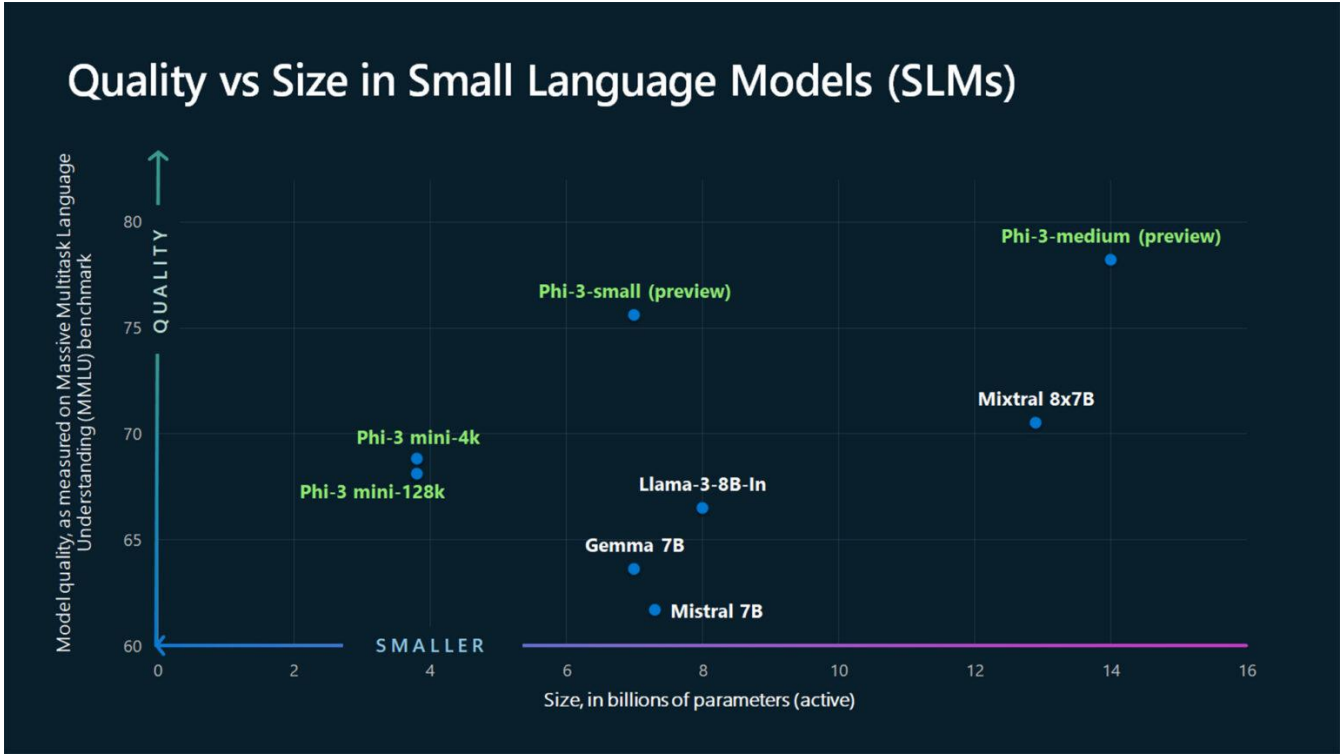
Source: [Paper - LORA: LOW-RANK ADAPTATION OF LARGE LAN- GUAGE MODELS]

Quantization



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
-1	-0.8667	-0.7334	-0.6001	-0.4668	-0.3335	-0.2002	-0.0669	0.0664	0.1997	0.333	0.4663	0.5996	0.7329	0.8662	1

Finetune Phi-3-mini-4k-instruct



Phi-3

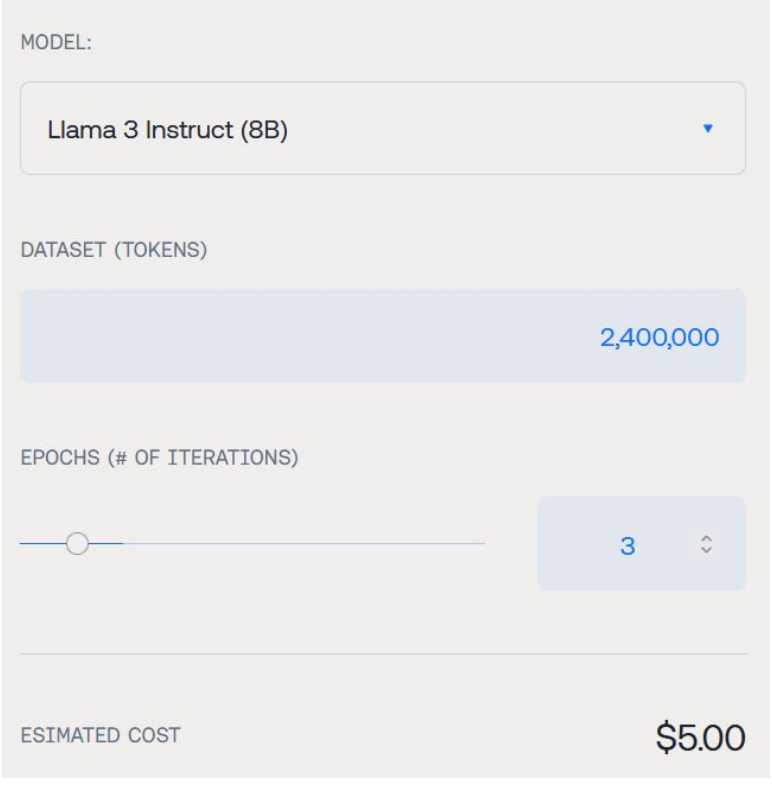
- Developed by microsoft
- Using HighQuality data
- Best smaller model
- Publish: 2023/04/23
- Latest update: 2024/07/01

Benchmarks	Original	June 2024 Update
Instruction Extra Hard	5.7	6.0
Instruction Hard	4.9	5.1
Instructions Challenge	24.6	42.3
JSON Structure Output	11.5	52.3
XML Structure Output	14.4	49.8
GPQA	23.7	30.6
MMLU	68.8	70.9
Average	**21.9**	**36.7**

Source: Tiny but mighty: The Phi-3 small language models with big potential

Show Code

Finetune Cost



Before vs After fine-tuning

Dataset (Style)	Field	Before (Acc)	After (Acc)
Formal	Name	0.872	**0.984**
	Age	0.959	**0.984**
	Job	0.612	**0.995**
Informal	Name	0.899	**0.977**
	Age	0.989	**0.998**
	Job	0.686	**0.1.0**
Novel	Name	0.882	**0.949**
	Age	0.950	**0.959**
	Job	0.551	**0.988**

Before vs After fine-tuning (Test dataset)

Dataset (Style)	Field	Before (Acc)	After (Acc)
Formal	Name	0.870	**0.965**
	Age	0.965	**0.975**
	Job	0.685	**1.0**
Informal	Name	0.895	**0.960**
	Age	0.980	**0.995**
	Job	0.750	**1.0**
Novel	Name	0.875	**0.925**
	Age	0.935	**0.930**
	Job	0.550	**0.955**

Thank You
For joining