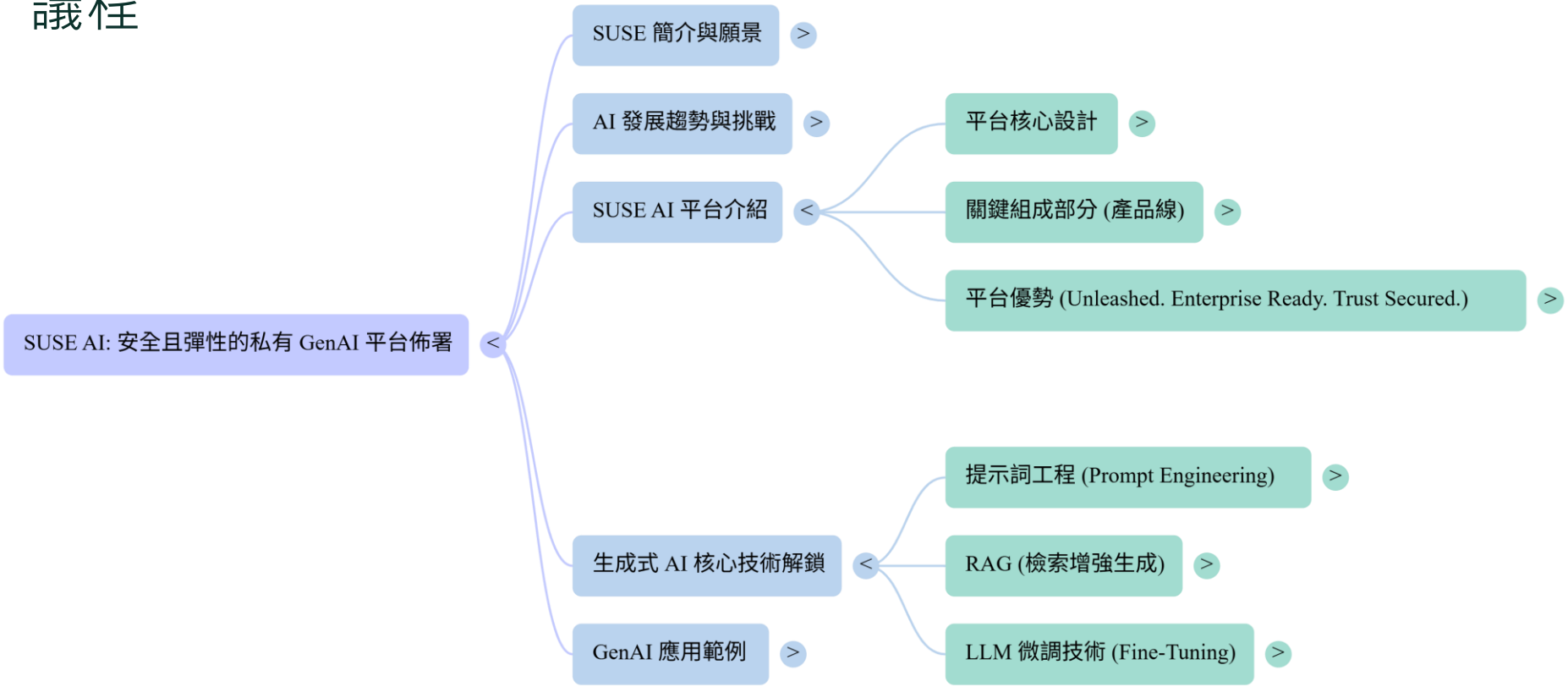
安全且彈性的私有 GenAI 平台佈署

SUSE Taiwan
Vincent Lu (盧木賢)
VLU@suse.com
0919-921-257

議程


SUSE AI: 安全且彈性的私有 GenAI 平台佈署
SUSE 簡介與願景
AI 發展趨勢與挑戰
SUSE AI 平台介紹
平台核心設計
關鍵組成部分 (產品線)
平台優勢 (Unleashed. Enterprise Ready. Trust Secured.)
生成式 AI 核心技術解鎖
提示詞工程 (Prompt Engineering)
RAG (檢索增強生成)
LLM 微調技術 (Fine-Tuning)
GenAI 應用範例

SUSE at a glance

1992
Founded in Nuremberg

HQ
Luxembourg

40
Offices

2,600
Employees

$700m+
Revenue

10,000+
Enterprise Customers

90%+
Of the world's leading companies rely on SUSE [6]

10/10
of the largest **Automotive** companies

13/15
of the largest **Pharmaceutical** companies

14/15
of the largest **Aerospace** companies

13/15
of the largest **FinServ** companies

Common Criteria SLSA

Top Supply Chain Security Certifications [3]

Developer contributions measured every day →

Publicly ranked alongside the largest technology companies

Consistently

Top 5	Top 8	Top 12
OSSRANK	CLOUD NATIVE COMPUTING FOUNDATION	OSCI

Recognised leader in Container Management & Virtualization → [2]

Gartner IDC OMDIA

United Nations Global Compact

17% emissions reduction since 2022 [4]

World-class ecosystem of partners
5 years in a row [5]

THE CHANNEL CO. CRN PARTNER PROGRAM GUIDE WINNER 2024

SUSE

[1] SUSE Company Data [2] Gartner Magic Quadrant + Omdia [3] Common Criteria, SLSA [4] The UN Global Compact [5] suse.com/partners [6] SUSE Customer Data

Our **vision** is to bring

the infinite potential of open source to the enterprise.

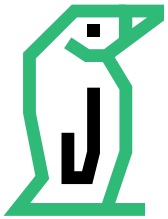
Our **mission** is to transform

open source innovations into enterprise solutions to give customers choice.

SUSE


Choice

For you to differentiate, you need freedom of choice



Linux

SUSE Linux

安全可靠的 Linux 及管理平台

Cloud Native

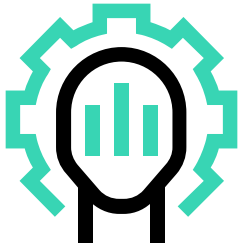
SUSE Rancher Prime, Security

下載超過1億次, 最多人使用的 K8S 管理平台與 K8S 發行版: RKE

Edge

SUSE Edge

輕量K8S 管理平台: K3S 與安全的 OS: SLE Micro



AI

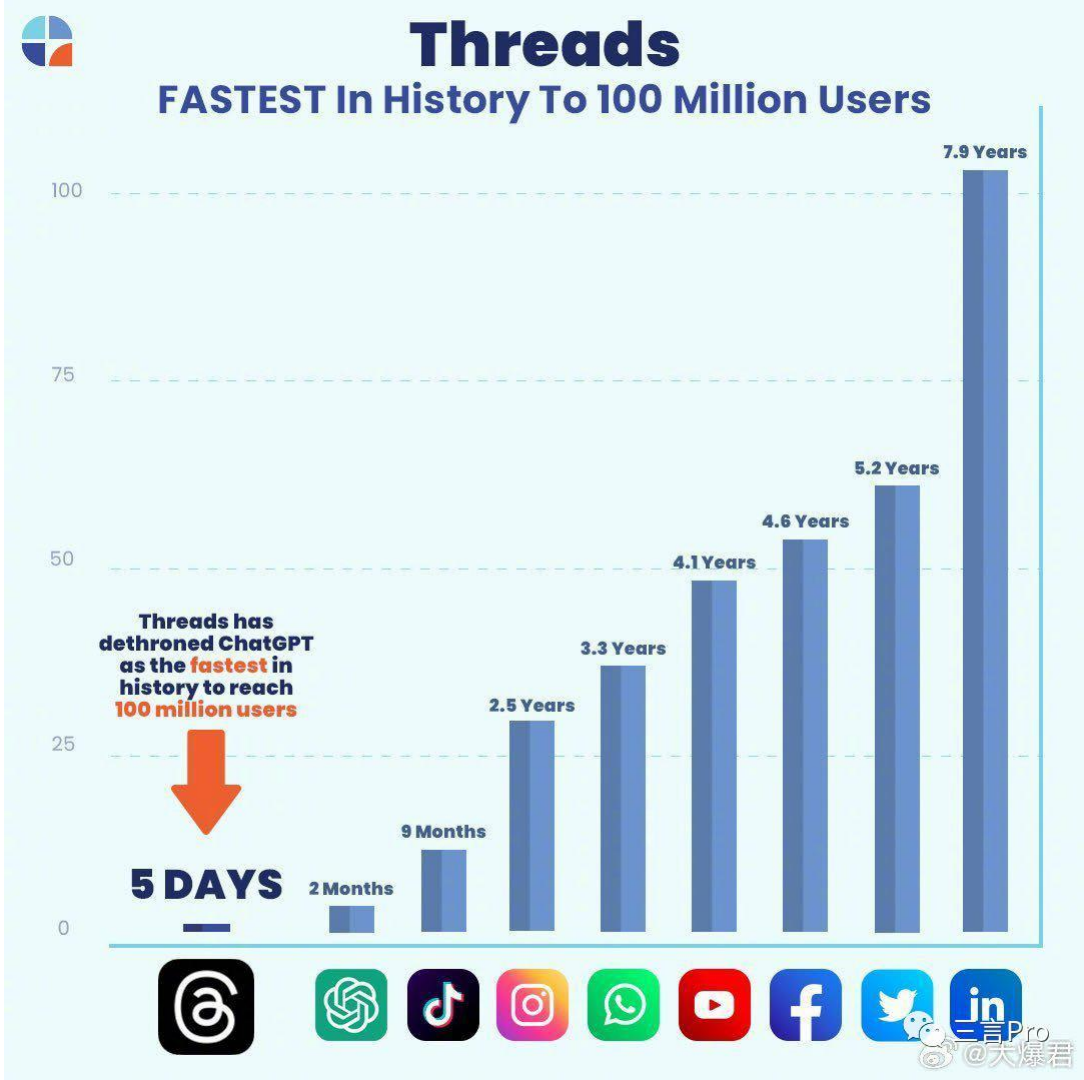
SUSE AI

開放式的 Private AI platform, 廣納 Open LLM

SUSE

從過去到現在

多久時間達到
1億個用戶數



SUSE

AI exploding, but...

of companies plan to invest more in AI over the next 3 years [1]

believe their investments have reached maturity [1]

only 44% of POCs in January 2025 made it into production [2]

Compliance challenges (regulations and fines)

Lack of trust (ethics and security)

Inconsistency (models, tools, apps)

Inability to scale (costs and access)

Complex infrastructure (new and changing)

SUSE

SUSE

Introducing SUSE AI

使用AI主要面臨的挑戰

資訊安全威脅	法令遵循 與 合規管理
推陳出新 瞬變的工具	超前沿技術 (科技, 軟體)

挑戰 #1: 數據完整性與資訊安全威脅

AI workloads present new security challenges:

- → Data poisoning attacks
- → Personal and IP data leaks
- → Misuse of model outputs
- → Model misbehaviour outside the network

SUSE AI platforms addresses these challenges.

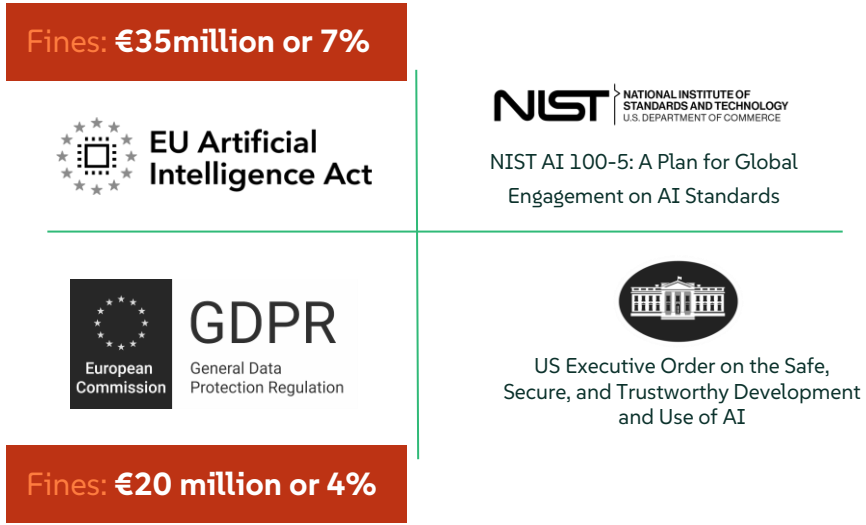
個人資料、資料合法性、資料正確性

偏見、種族、倫理、資訊安全保護

挑戰 #2: 法令遵循與合規管理

The regulatory landscape is evolving and complex and fines are costly.

SUSE provides the flexibility and security framework to meet changing demands, securely and predictably.



SUSE AI Transparency. Accountability. Security. Safety.

挑戰 #3: 推陳出新(瞬變)的 Open Source Landscape

Models & LLMs

LLaMA by Meta · nVLM · MISTRAL AI_ · Grok · 文心一言 · NOMIC

Vector Databases

Milvus · Pinecone · Chroma · qdrant · LanceDB

Synthetic training data

OpenSynthetics · gretel · datacebo Copulas · SYNTHEA · Sd

Configuration & tooling

Langflow · Tasking.AI · Open WebUI · composio · PrivateGPT · talkd.ai

Libraries & Frameworks

LangChain · LlamaIndex · Kubeflow · Ollama · LLaMA-Factory (Easy and Efficient LLM Fine-Tuning) · OpenCV · Keras · TensorFlow · PyTorch

Open Enterprise AI Infrastructure

SUSE AI

SUSE

SUSE AI
Designed to meet
these challenges

AI is the next strategic workload

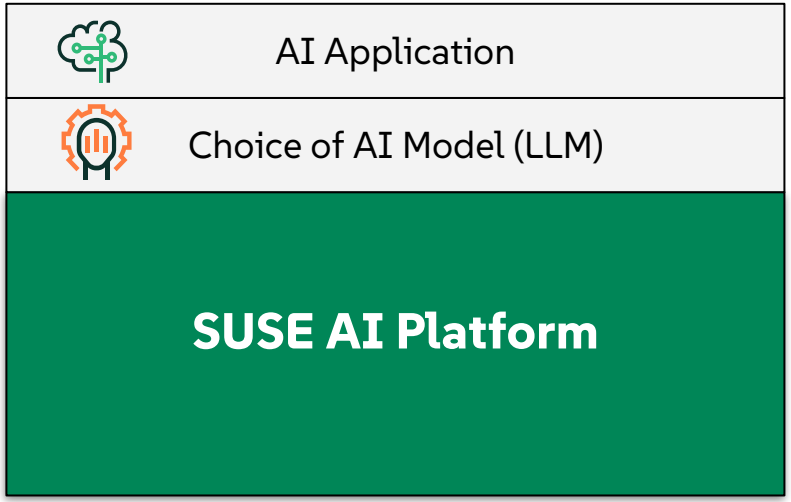
SUSE AI

A fully integrated, cloud native **platform** with sanitized components that help create secure and trusted AI applications.

Provides both **zero trust security** for your AI workloads and **observability** to give actionable insights for AI workloads.

SUSE

Enterprise-ready, cloud native **platform** to deploy and run any GenAI Workload



SUSE AI

Open/ Interoperable Container Management for AI

- Simple, consistent operations from data center to Edge
- Trusted, secure delivery
- Complete lifecycle management
- Industry leading AI, container security, storage, VM management, and Linux OS
- Access to broad ecosystem of open source technologies
- Distributed training and model serving
- Private GenAI experience

SUSE

Everything you need to deploy, run, and manage all of your containerized AI workloads

AI Security (Guardrails) Add-on	AI/ML Tools and Services **SUSE** AI	AI Observability Add-on
Full-lifecycle cluster management **SUSE** Rancher Prime	Virtual machine management **SUSE** Virtualization	Distributed storage **SUSE** Storage
Full-lifecycle container security **SUSE** Security	Operating system management **SUSE** OS anagement	Infrastructure & Application Monitoring **SUSE** Observability
Certified OS **SUSE** Linux Enterprise		

SUSE AI

AI Unleashed. Enterprise Ready. Trust Secured.

Security and Trust	• Zero trust security through SUSE Security • Insights into token usage and GPU performance for cost optimization and performance • Propel implementation of guardrails for ethical and bias free generative outputs
Freedom to Choose	• Bridge to run any LLM on the platform • Validated popular open source AI components and libraries jumpstarting GenAI workloads • Deploy in the cloud, hybrid, on premises, or air gapped
Extensible	• Designed with a modular architecture that fosters faster innovation • Adaptable to new business demands (eg: agentic workflows; guardrails technology) • Cloud native design that supports scalability

Flexible operating architecture

Industry-leading modularity

Real-time, real world protection

AI Libraries

Supported AI for Enterprise

Fully supported AI tools, libraries, and prototyping applications.

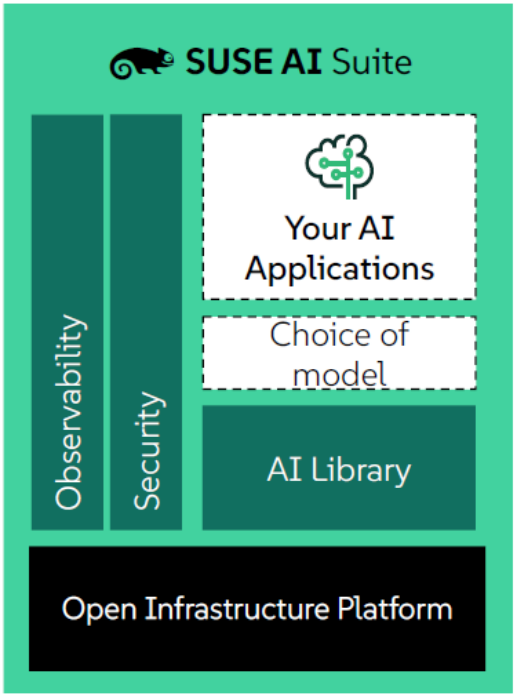
Data → Develop → Deliver

Nov 2024 GA

- Milvus — Vector Database
- Ollama — LLM Server
- Open WebUI — Prototyping

2025 Roadmap

- pgvector
- TensorFlow
- Keras
- LangChain
- LLaMA-Factory (Easy and Efficient LLM Fine-Tuning)
- PyTorch



LLM Observability
Security Templates

Use Cases Supported

- Chatbot
- Inference
- Image Management
- Fine tuning
- Training

SUSE

SUSE unlocks industry leading GenAI models,

We do not compete with them, you have choice!

LLaMA by Meta

NOMIC

MISTRAL AI_

文心一言

Grok

nVLM

Hugging Face

Falcon

Benefits of SUSE AI

Choice

Run any model

Modular and extensible to integrate with your infrastructure

No ecosystem lock-in

Total cost control & visibility

Deployment options include cloud, hybrid, on premises.

Security

Built with Secure Supply Chain; Inherits certifications

Zero-trust security

Full observability

Data sovereignty

Deployable as an air-gapped solution.

Trust

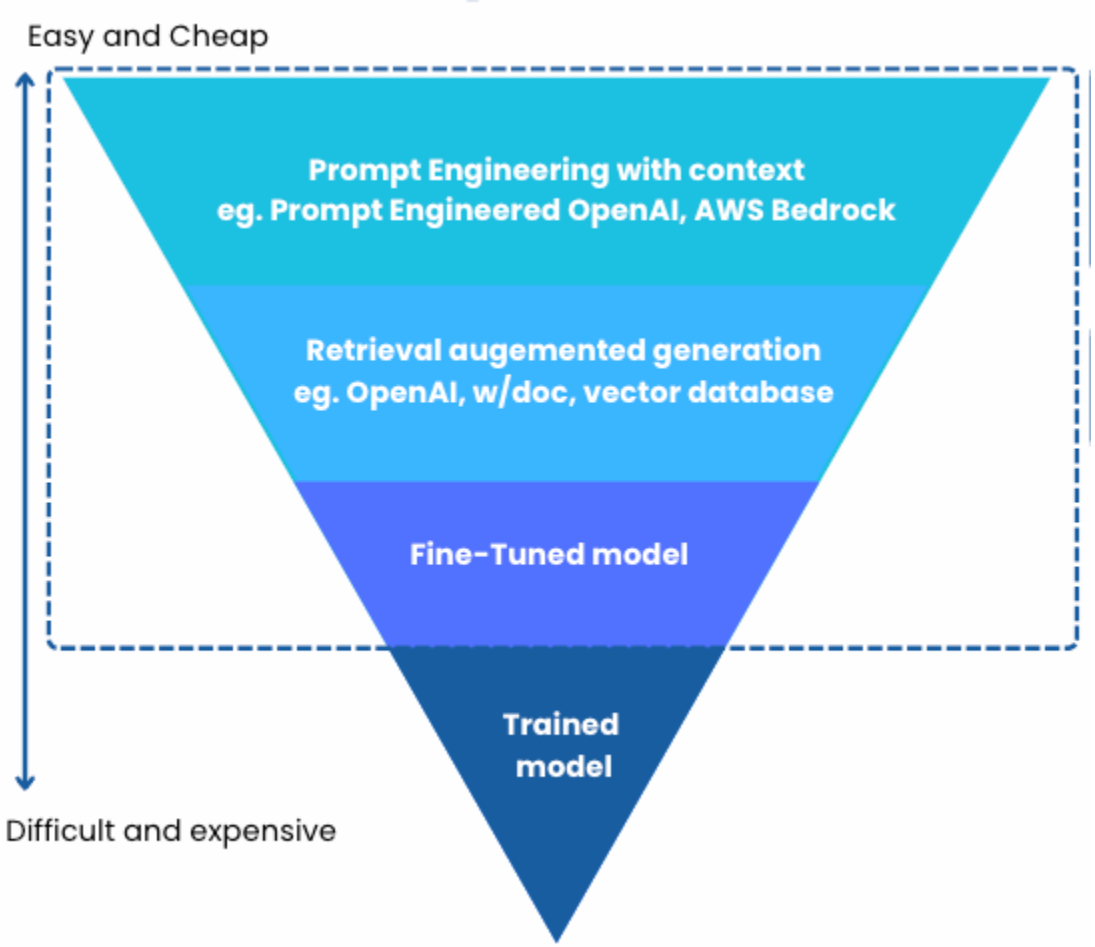
Data stays within enterprise control

Control of usage and activity

Greater confidence in the generated output

解鎖生成式AI潛力

- 提示詞工程
- RAG
- 微調技術

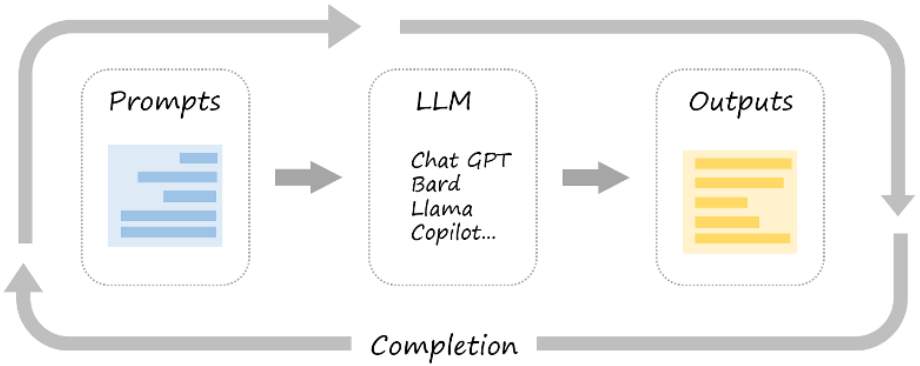


Prompt (提示詞) engineering (講清楚說明白, 好的問題)

設計提示詞 引導模型 調整和改進

跟AI溝通: 足夠的資訊, 簡單明瞭

- 任務 Task : 指定一個角色, 清楚的任務
- 背景資訊 Context : 條件, 限制, 對象
- 參考資訊 Reference
- 重新檢查評估 Evaluate
- 修改調整 Iterate



https://www.youtube.com/watch?v=wf6EwC-H4Sw

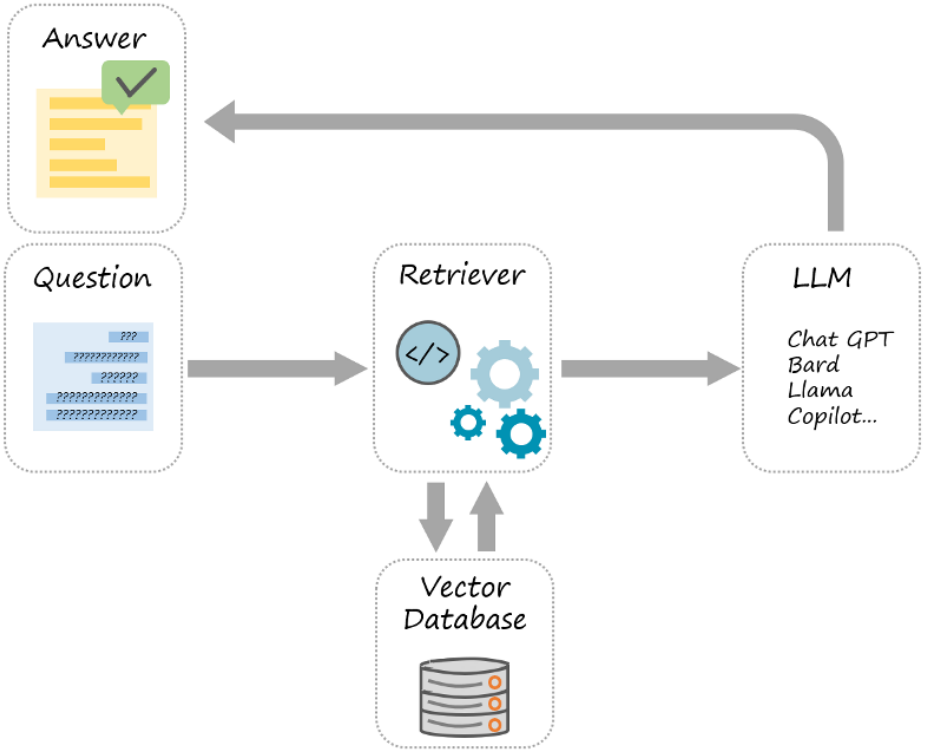
RAG(Retrieval-Augmented Generation) 檢索增強生成

檢索階段: 將用戶的查詢或問題轉向至向量資料庫進行檢索。

融合階段:將檢索到的相關資訊與原始查詢進行合併，組成上下文或背景知識，以便 LLM 能夠生成更精確的回答。

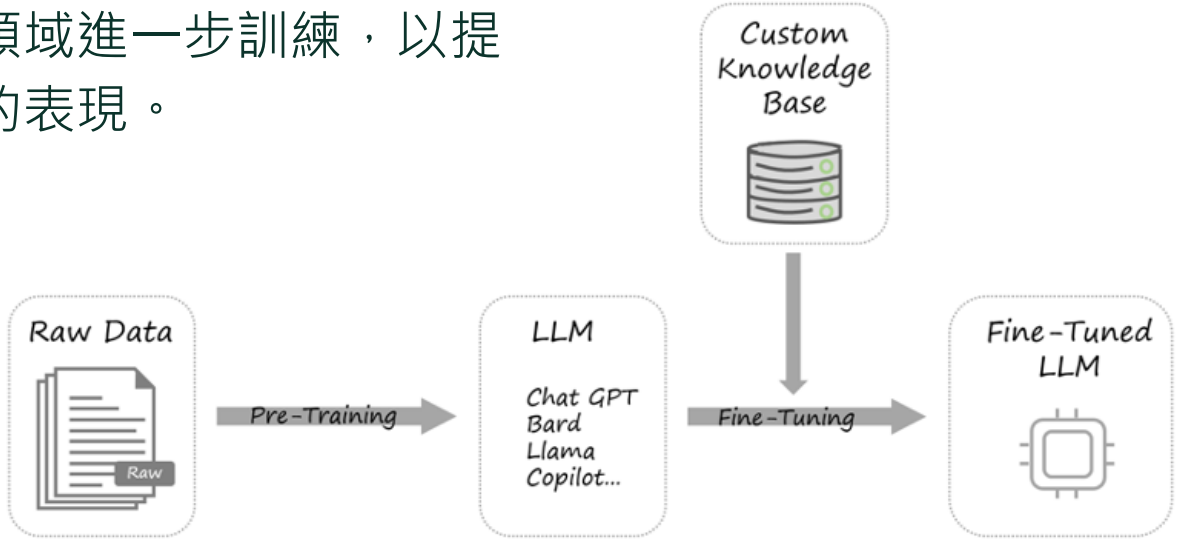
生成階段: LLM 使用這些合併的上下文和原始查詢生成最終的回答

優化和整合: 後續處理，確保回應內容與使用者的需求高度相關，並修正任何潛在的錯誤或不一致。



LLM 微調技術(**Fine-Tuning**)

Fine-tuning 是指在既有的 **Trained Model**，針對特定任務或特定領域進一步訓練，以提升模型在特定情境中的表現。

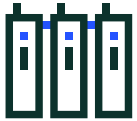


Example areas that people are applying GenAI

生成式 AI 技術正改變我們的工作模式並提升效率，
對於個人: 從學習、自動生成文本、圖像和音樂創作...等應用廣泛。
對於企業：包括客服、業務流程、行銷、法律、教育、人力資源、資料分析、文案生成... 相關等應用，
透過生程式 AI 這項技術能顯著提升工作效率和創造力，使日常工作更具創新性。

Customer Support

An enterprise wants to enhance customer support teams with GenAI agents by product.



IT Support

A telco wants to increase productivity among IT support staff

Marketing

A Retail company wants to tailor content to different customer segments based on past interactions and preferences.

Document Automation

A Consulting firm needs to analyze large sets of documentation and create summaries for employees.

Thanks !