Who Am I?

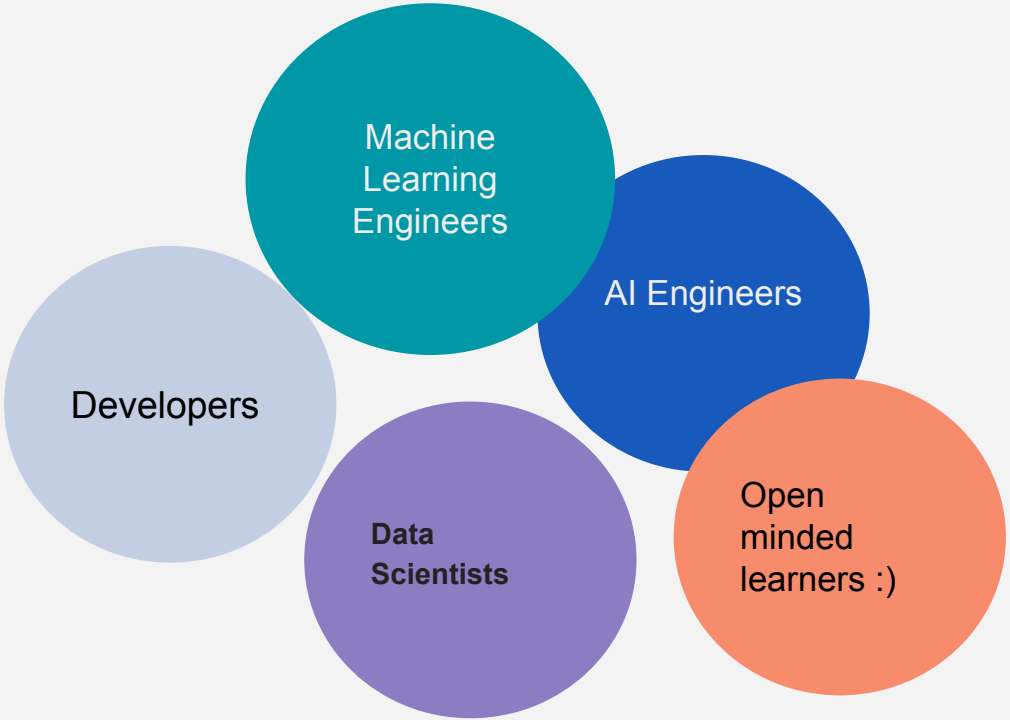
Paulo Dichone

Software, Cloud, AI Engineer
and Instructor

What Is This Course About?

- DeepSeek R1 models
  - Master DeepSeek R1 local deployment
  - Understand model optimization
  - Build practical applications/use-cases
  - ***How to really think about future reasoning models***

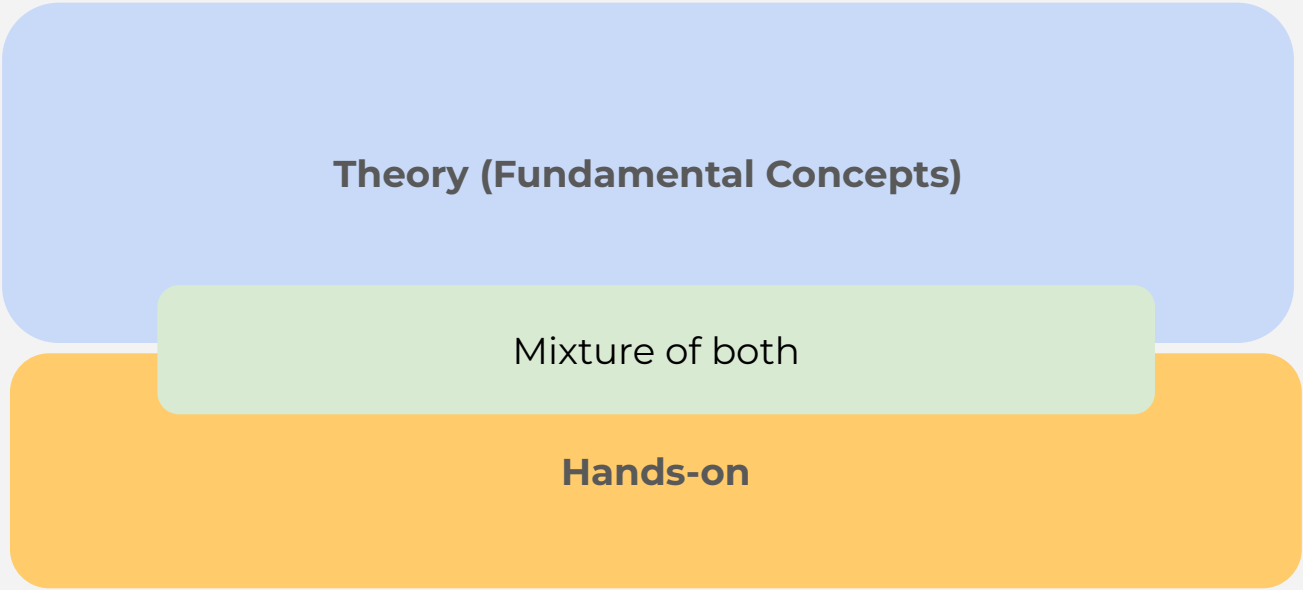
Who Is This Course For



Course Prerequisites

1. Know programming (highly *preferred... at least the basics)*
    a. *We will be using Python*
2. Basics of AI, ML, LLM, AI Agents
3. *This is not a programming course*
4. Willingness to learn :)

Course Structure



Development Environment setup

- Python
- VS Code (or any other code editor)
- OpenAI Account and an OpenAI API Key

Set up Ollama on Win and MacOS

Follow instructions here:
https://medium.com/@sridevi17j/step-by-step-guide-setting-up-and-running-ollama-in-windows-macos-linux-a00f21164bf3

Dev Environment Setup

Python (Win, Mac, Linux)

https://kinsta.com/knowledgebase/install-python/

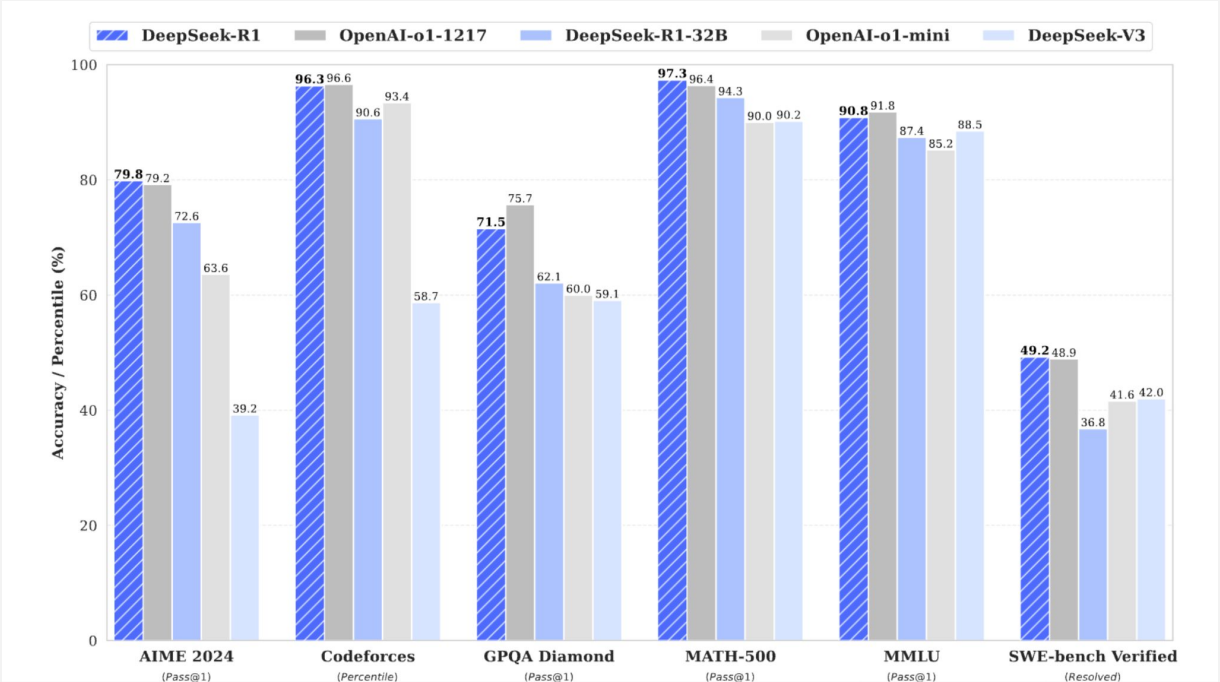
DeepSeek Fundamentals

- Introduction to DeepSeek R1
  - Deep dive into DeepSeek Universe
- Benefits
- Key concepts

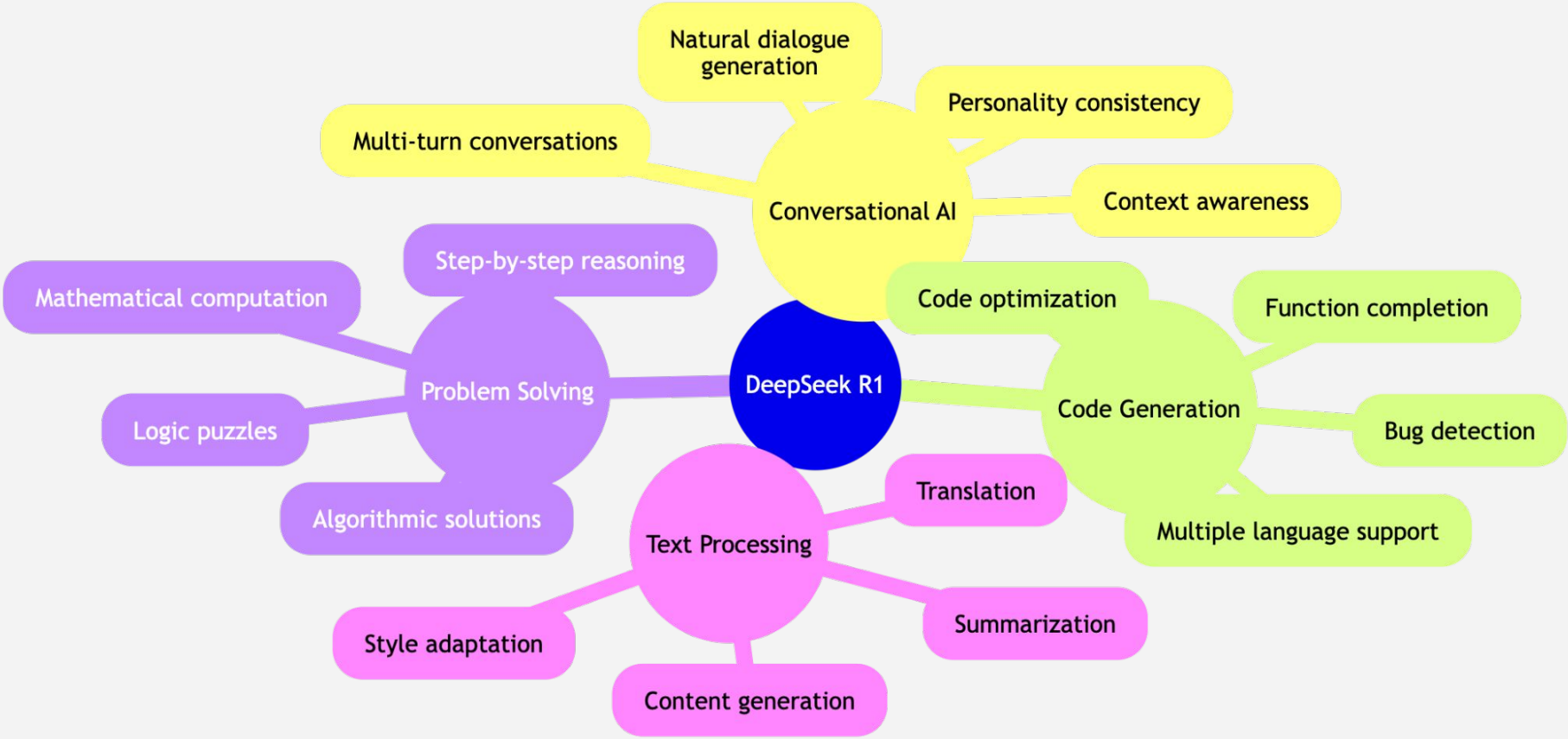
DeepSeek - Overview

A Chinese company that builds AI models (Open-source models)

DeepSeek R1 - an open-source model with **Reasoning** capabilities **comparable (outperformed)** to OpenAI o1 models.



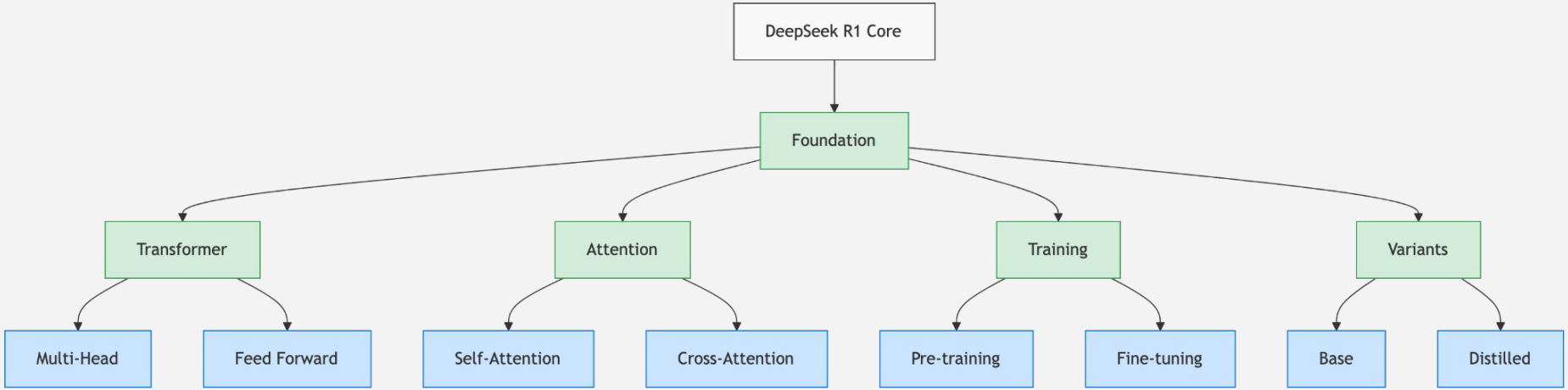
R1 Capabilities



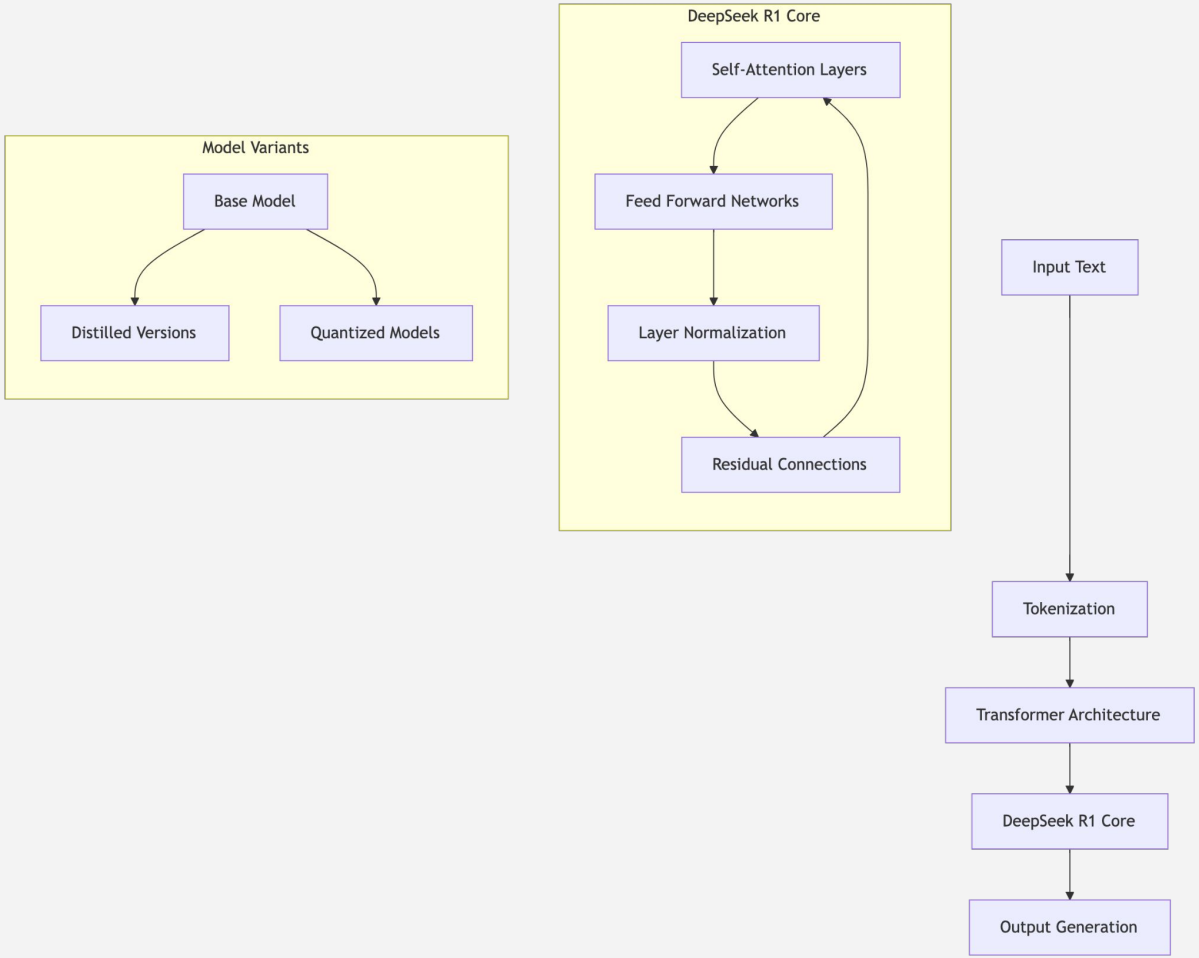
Foundation - Core Architecture

- Based on transformer architecture
- Implements advanced attention mechanisms
- Utilizes state-of-the-art training techniques
- Supports multiple model sizes through distillation

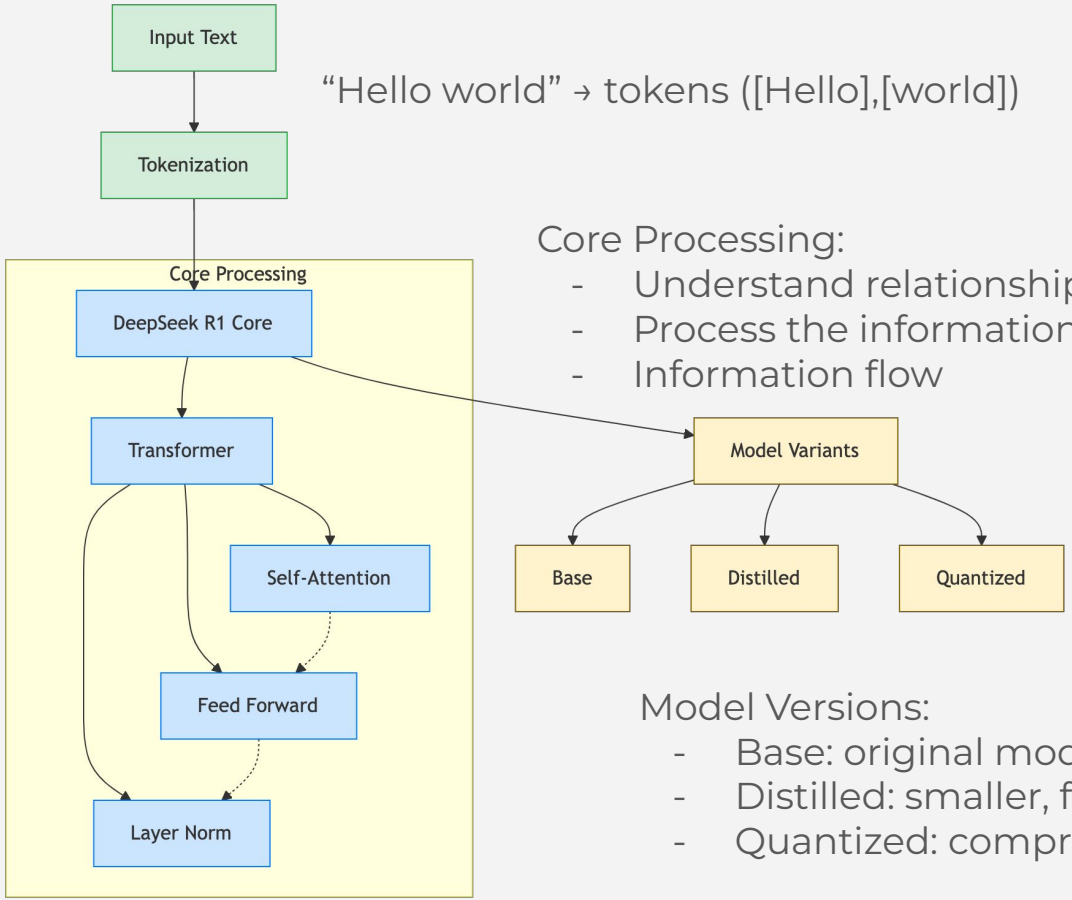
Foundation - Core Architecture



Foundation - Core Architecture



Key components



“Hello world” → tokens ([Hello],[world])

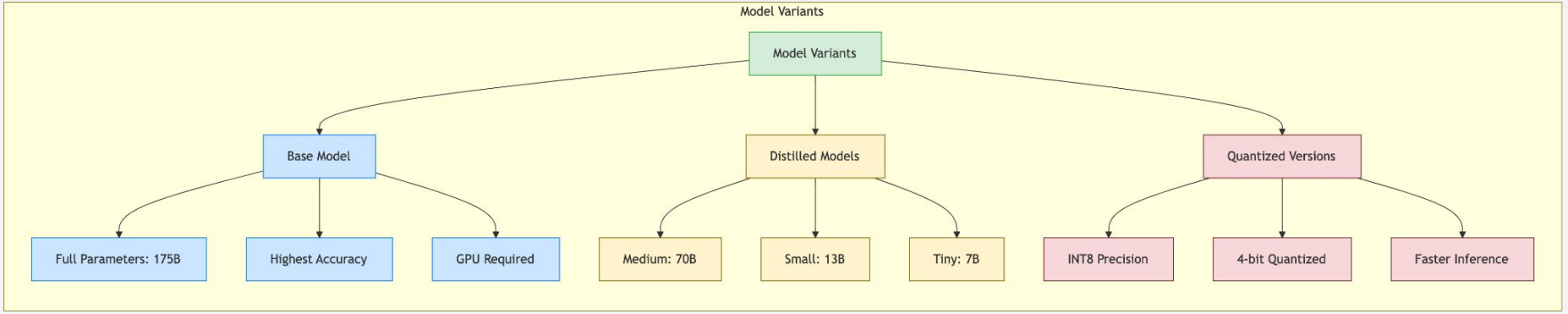
Core Processing:

- Understand relationship between words
- Process the information
- Information flow

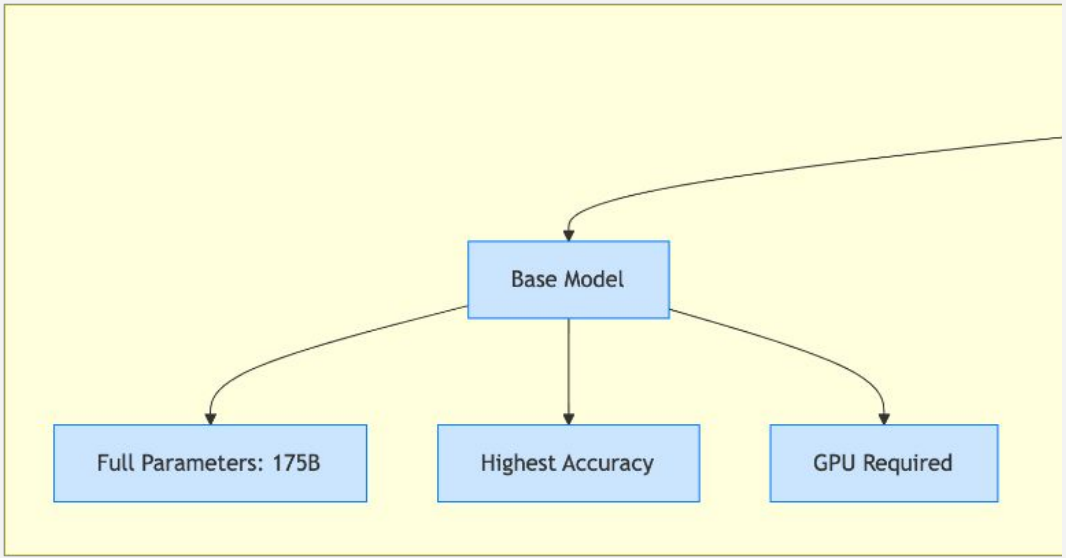
Model Versions:

- Base: original model
- Distilled: smaller, faster version
- Quantized: compressed version

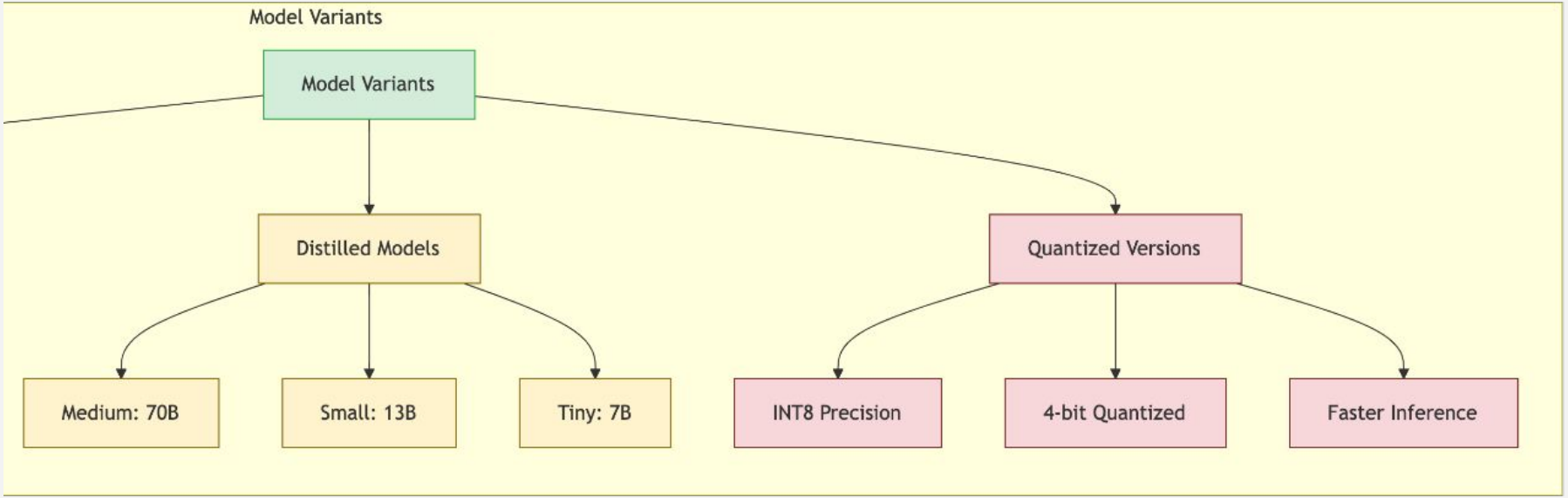
Modal Variants



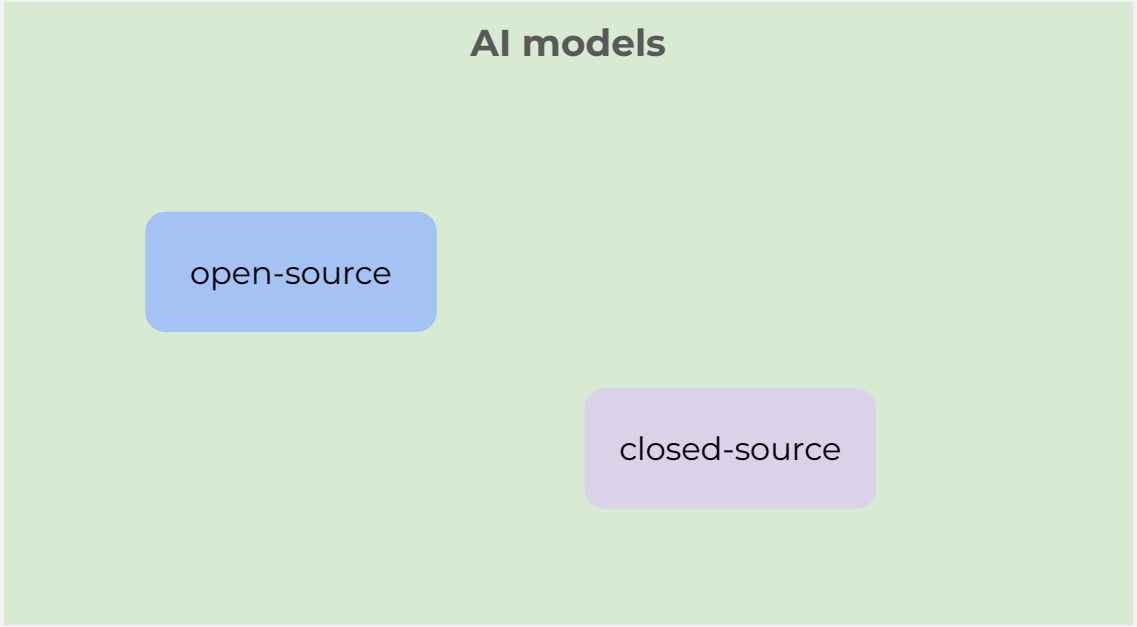
Modal Variants



Modal Variants



Open vs. Closed Source LLMs



Open-source LLMs

open-source

Transparency

Cost Control

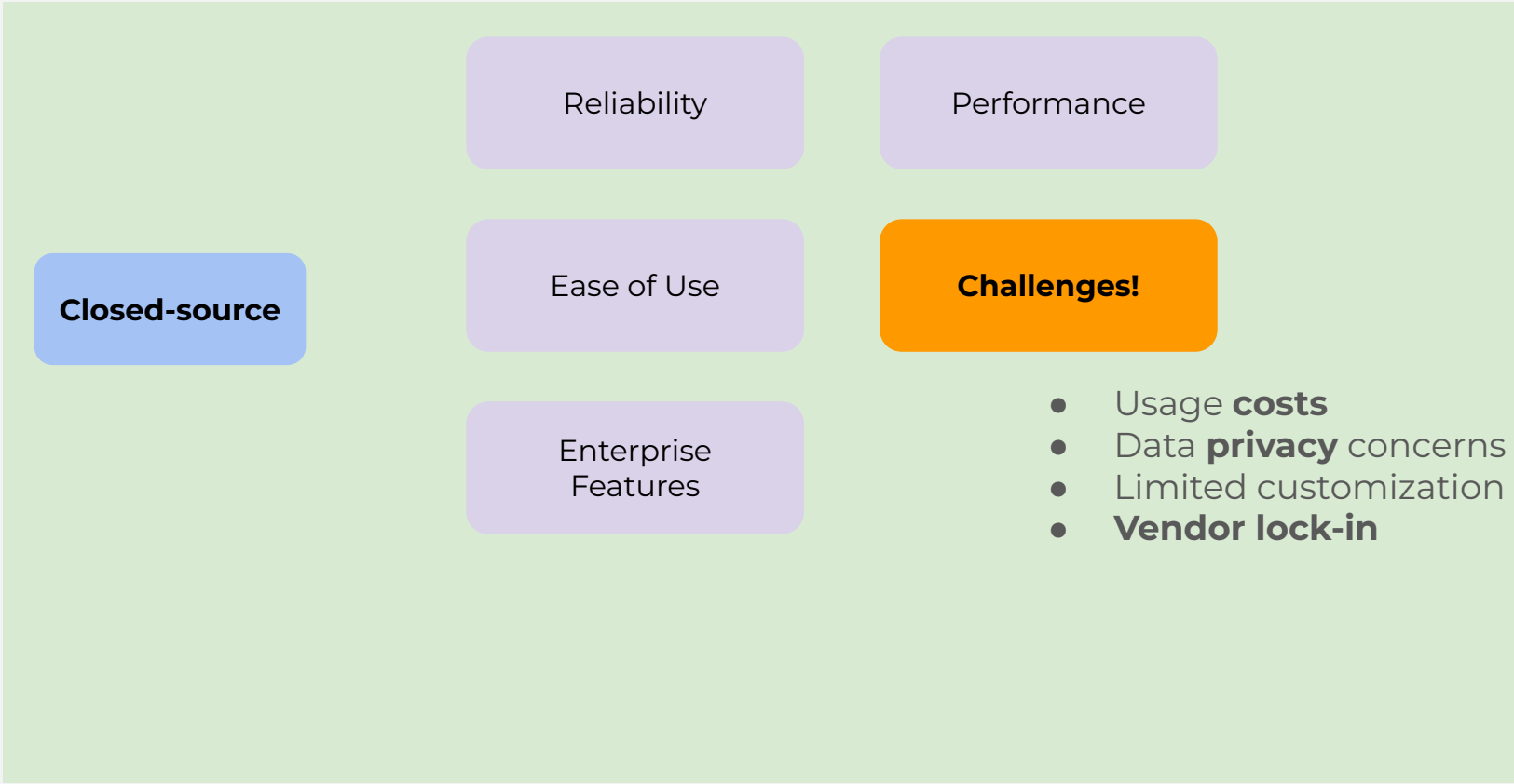
Customization flexibility

Challenges!

Community collaboration

- Initial setup complexity
- Hardware requirements
- Maintenance responsibility
- Performance optimization needs

Closed-source LLMs



Summary

1. DeepSeek R1 Foundations
    a. Transformer architecture
    b. Capabilities
    c. Model Variants
2. Open vs. closed-source Models

DeepSeek R1 Deep Dive

- DeepSeek Chatbox
- LM Studio
- Using Chatbox (Free)
- API and SDK
- Local Installation

Initial installation and Setup

1. Using DeepSeek website (chatbox)
2. Downloading R1 model locally through LM Studio
3. Downloading R1 via Ollama

DeepSeek R1 Hardware Requirements

DeepSeek R1 Hardware Requirements

Minimum (1.5B Model)

CPU: 8-core CPU
RAM: 16GB RAM
Storage: 20GB SSD
GPU: Optional (4GB VRAM)

Notes: Suitable for testing and development
Performance: Slower response times, basic capabilities

Recommended (7B Model)

CPU: 12-core CPU
RAM: 32GB RAM
Storage: 50GB SSD
GPU: 8GB VRAM GPU

Notes: Good for most use cases
Performance: Balanced performance and capabilities

Optimal (Full Model)

CPU: 16+ core CPU
RAM: 64GB RAM
Storage: 100GB NVMe SSD
GPU: 16GB+ VRAM GPU

Notes: Best for production use
Performance: Fast responses, full capabilities

Initial installation and Setup

1. Using DeepSeek website (chatbox)
2. Downloading R1 model locally through LM Studio
3. Downloading R1 via Ollama
    a. Downloading ChatBox to use R1 model

Development & Integration

Local API

- Free
- Private
- Fast inference

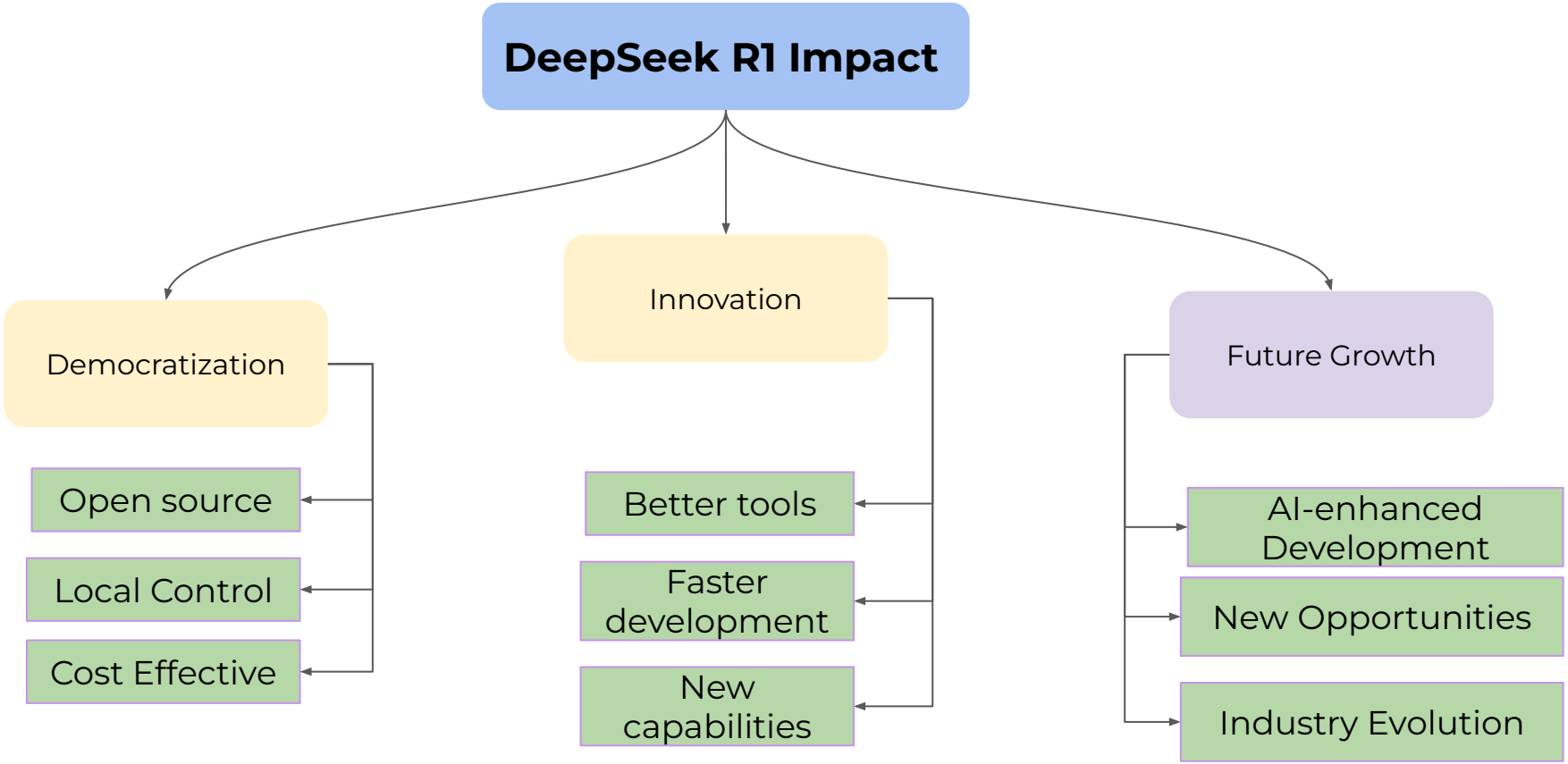
DeepSeek API

- Cost
- Not private
- Latency

Hands-on - Building R1-based Use-cases

Final Thoughts

Final Thoughts



The Goal!

- We are entering an era of AI (already here)
- These tools/models are here to:
    - **Enhance developers/users capabilities**

Congratulations!

You made it to the end!

- Next steps...

Course Summary

- DeepSeek R1
  - What is it?
  - How it works
  - Set up and Installation
    - Local installation
  - Build some Use-cases
- Hands-on

Wrap up - Where to Go From Here?

- Keep learning
  - Extend the projects we worked on in this course
  - Design and implement your own agents
- https://api-docs.deepseek.com/

Thank you!