B3 System - B3rry

Application of Large language Models (LLM)

Tao Shan
MDSAI, University of Waterloo

https://www.linkedin.com/in/taoshan88/

Agenda

1 Use Case Description

2 Data & Assumptions

3 Solutions & Values

4 Project Timeline

Use Case Description

Problem

Better detailed, context-aware responses and AI-generated reports

Need

User Satisfication, Enhance Decision-Making, Increase productivity, Competitive Advantage

Opportunity

Global Smart Manufactoring Market Size - USD XXX.XX million USD in XXXX, CAGR of XX.XX%

Solution

Utilize Data, Prompt Engineering, Fine Tuning, RLHF, RAG, Pre-training. Result with an accurate LLM.

Data

Assumptions:

- Based on basic description about type of data and sizes

Sensor Data

- RAG
- Fine Tuning

Maintenance Logs

- Fine Tuning

Production Data

- RAG
- Fine Tuning
- Pre-Train Model

Operational Metrics

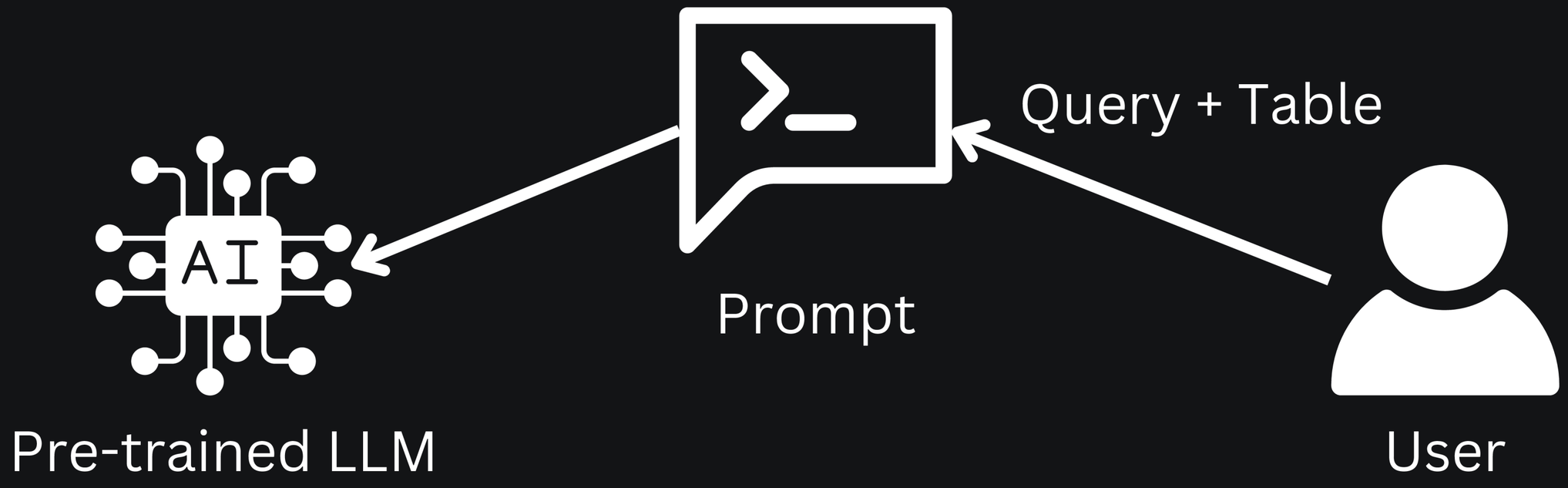
- RAG
- Fine Tuning

Chatbot Interaction Data

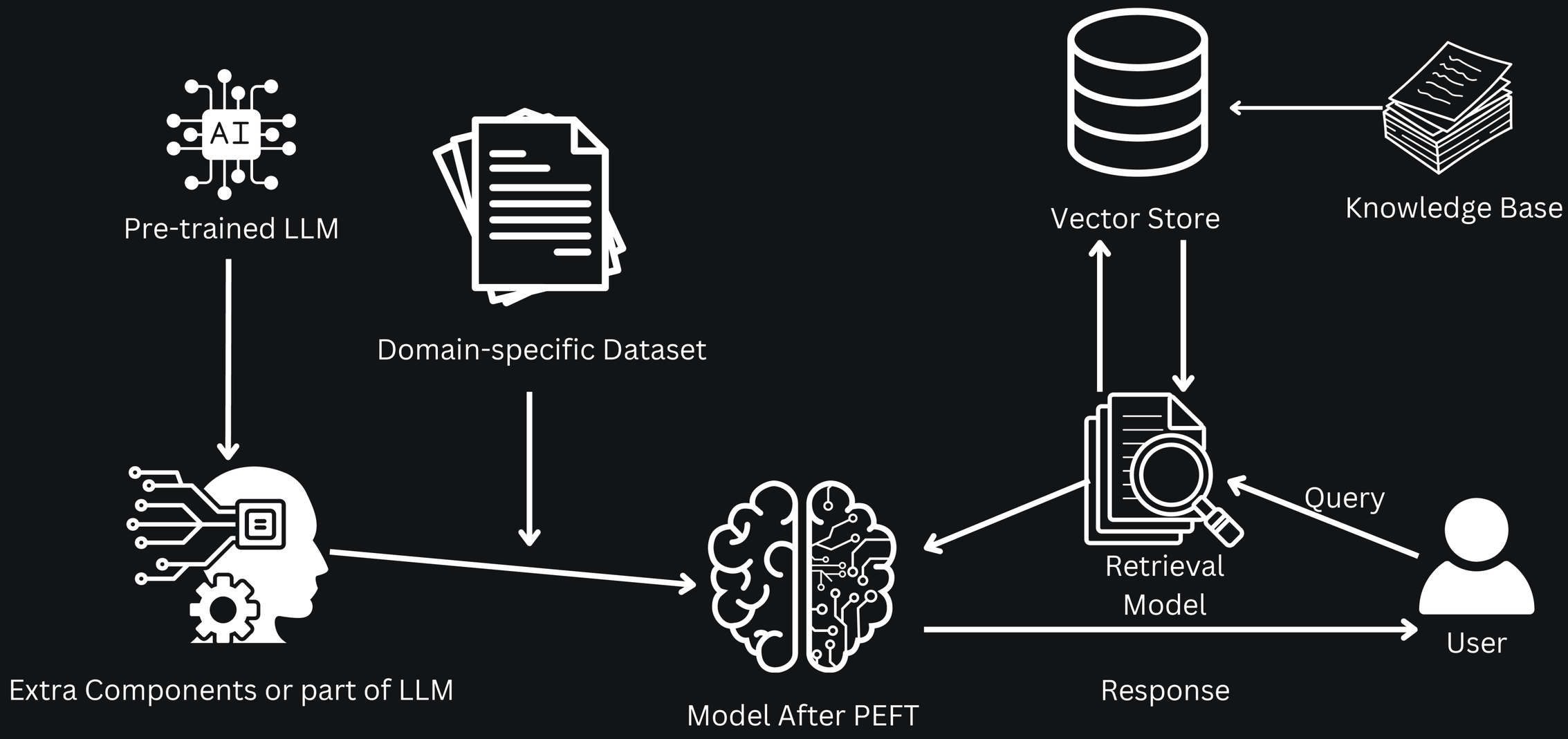
- Promopt Engineering
- Fine Tuning
- RLHF

Use these data in different parts of LLM system

Basic LLM



Advanced LLM



01 Prompt Engineering

- In-context learning (zero shot prompting)
- One and few shots prompting
- Chain of Thought (CoT) prompting
- ReAct (Reasoning and Acting)

02 Fine Tuning

- Adapt pre-trained model to our task
- Full Fine Tuning
- PEFT (Parameter Efficient Fine-Tuning)
  - LoRA
  - Adapters
  - Soft Prompt Language Model prompting
- RLHF (Reinforcement Learning from Human Feedback)

03 RLHF

Combines reinforcement learning with human input to improve an AI model's performance and align it with human preferences

04 RAG

- Framework for LLM access data not see during training
  - Retriever - vector store, train model from external data most relate to query
  - Expand previous prompt from knowledge base

05 Evaluation

- Metrics - BLEU, ROUGE, METEOR, F1 scores
- Human Rating
- LLMs Evaluating LLMs
  - Langchain's QAGenerationChain

06 Model Efficiency

- Save computational resource
- Quantization - from 32-bit to 16-bit
- Pruning - remove unnecessary weights
- Efficient Model

07 Impact

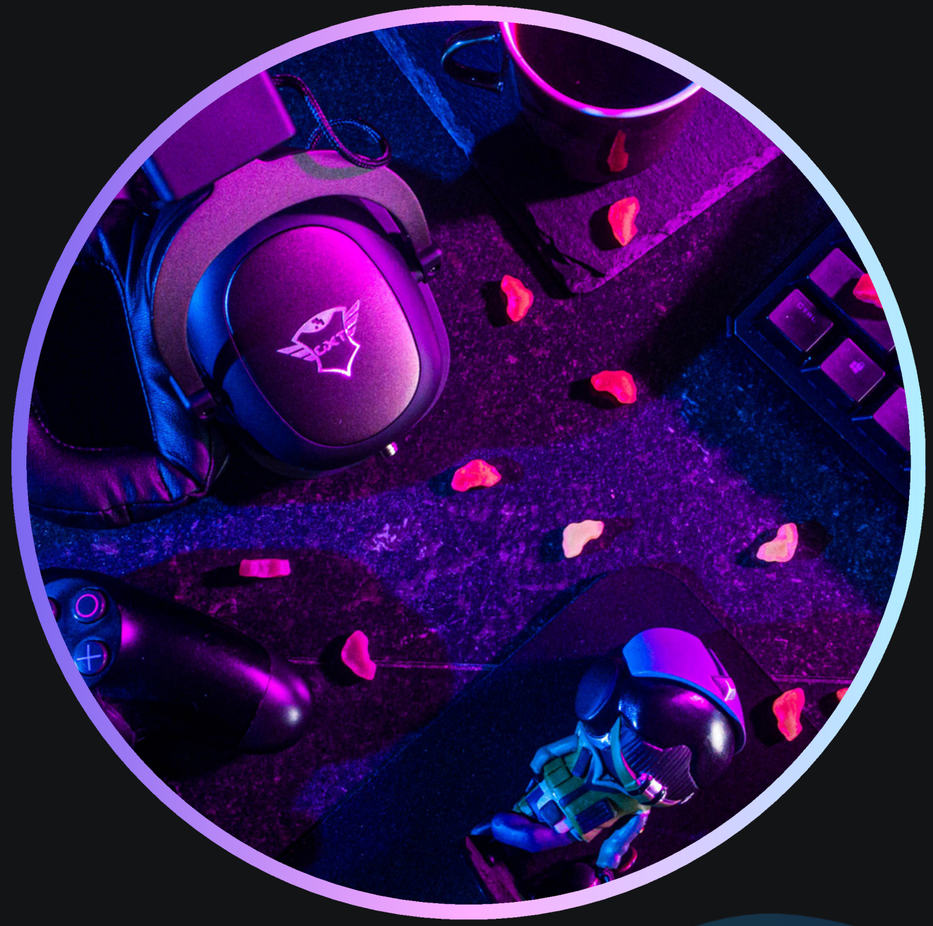
- Better Accuracy and Relevance
- Better Handling Complex, Multi-step tasks
- Reduce Bias and increase Safety
- Less Computational Cost

Time and Effort

	Prompt Engineering	Fine Tuning	RLHF	Optimization/Deployment
Training duration	Not required	Minutes to hours	Minutes to hours similar to fine-tuning	Minutes to hours
Objective	Increase task performance	Increase task performance	Increase alignment with human preferences	Increase inference performance
Cost	Low	Medium	High	Low

Time and Effort

- Use existing pre-trained model and parameters
  - Pay by usage/ free open source
    - ChatGPT, Gemini
    - Lilama 3, BLOOM, Mistral
- Pre-training new LLM model
  - High data and comoputational requirement - depend on task complexity
  - High Initial Cost vs. Ongoing Savings



Timeline - Milestones

Time	Milestone	Tasks
Week 1-2	Use Case Understanding	Understand current status, familar with current chatbot system
Week 3-4	Design Pipeline	Literature Review, Options of Fine Tuning, RAG, vectorDB
Week 5-10	Design proof of concepts	Utilize all methods to improve the model system
Week 11-14	Model Testing & Optimization	More Prompt Engineering, Validation, Optimize the model better
Week 14-16	Add value	Deploy the model, Prepare documentation and report, Github Repo

Thank You

Any Questions!

Appendix

- Smart manufactoring Market Sizes: https://www.marketsandmarkets.com/Market-Reports/smart-manufacturing-market-105448439.html
- Fine Tuning
  - instruction fine-tuning: https://arxiv.org/pdf/2210.11416
  - PEFT: https://arxiv.org/pdf/2303.15647
  - LoRA: https://arxiv.org/pdf/2106.09685
  - Prompt Tuning with soft prompt: https://arxiv.org/pdf/2104.08691
- Evaluation
  - GLUE: https://openreview.net/pdf?id=rJ4km2R5t7
  - SuperGLUE: https://super.gluebenchmark.com/
  - ROUGE: https://aclanthology.org/W04-1013.pdf
  - LLM evaluate LLM: https://arxiv.org/abs/2306.05685
- Reinforcement Learning from Human-Feedback (RLHF)
  - https://arxiv.org/pdf/2203.02155
  - https://arxiv.org/pdf/2009.01325

Appendix

- Additional Prompting
  - Chain-of-Thought: https://arxiv.org/pdf/2201.11903
  - ReAct: Synergizing Reasoning and Acting: https://arxiv.org/abs/2210.03629
- Langchain: https://github.com/langchain-ai/langchain