

國立台灣科技大學 114學年 第2學期 課程大綱

Spring 2026 NTUST Course Outline

授課教師：陳郁堂

Instructor:Chen, Yie-Tarng

課程名稱：大型語言模型與應用

Course Title : Large Language Models and Applications

2026/6/22

課程代號： ET5347701 Course Code 學分數： 3 Credits	必選修：選修/半學年 Required/Elective: Elective/Half Yr. 先修課程： Prerequisites
節次教室： F2(IB-408) F3(IB-408) R4(IB-408) Time/Location	
專業核心能力： Core Professional Competencies	
1. 運用數學、科學及工程知識的能力。 2. 設計與執行實驗，以及分析與解釋數據的能力。 5. 培養學生具備執行跨領域專案整合、領導及溝通之能力。 6. 發掘、分析、應用研究成果及因應複雜且整合性工程問題的能力。	
課程網址： Course Website	
課程宗旨： Course Objectives	This course provides an in-depth introduction to Large Language Models (LLMs) and generative models, covering their theoretical foundations, technical architectures, model fine-tuning, and practical development with real-world applications. The curriculum explores the internal mechanisms of LLMs in detail, including language pretraining, transfer learning, and other core techniques, and examines how fine-tuning methods can be applied to adapt models to specific tasks. In addition to language models, the course also introduces generative models—particularly text-to-image generation—and discusses their training methodologies, enabling students to understand state-of-the-art technologies while developing hands-on practical skills.
課程大綱： Outline of Lectures	

1. LLM Fundamentals (optional)
 - Pytorch: Essential programming skills
 - Neural Networks: Introduction to concepts
2. The LLM Architecture
 - Transformer Overview: Encoder-decoder vs. decoder-only models
 - Tokenization: Converting text to tokens
 - Attention Mechanisms: Self-attention and scaled dot-product attention
 - Text Generation Techniques: Greedy decoding, beam search, top-k sampling, nucleus sampling
3. Building an Instruction Dataset
 - Synthetic Data Generation
 - Dataset Improvement Techniques: Evol-Instruct, regex filtering
 - Prompt Templates
4. Pre-training Models
 - Data Pipeline and vocabulary setup
 - Causal vs. Masked Language Modeling
5. Supervised Fine-Tuning
 - Full fine-tuning, LoRA, QLoRA
 - Axolotl and DeepSpeed
6. Preference Alignment
 - RLHF and preference datasets
 - PPO-based optimization
 - GRPO optimization
7. LLM Inference and Serving
 - Latency Decomposition: Prefill vs Decode
 - KV Cache Management and Memory Optimization
 - Continuous batching and scheduling
 - Long-context serving and million-token context
 - vLLM and SGLang architectures
 - GPU memory hierarchy and KV offloading
 - Quantization Methods: GPTQ, AWQ, GGUF
 - Speculative Decoding for fast inference
8. Evaluation and Benchmarking
 - LLM-as-a-Judge evaluation frameworks
 - Benchmark datasets: MMLU, HumanEval
 - RAG evaluation techniques: RAGAS
 - Automated evaluation pipelines and metrics
9. Running LLMs
 - API usage vs local execution
 - Prompt Engineering
10. Retrieval Augmented Generation (RAG)
 - Orchestrators and Retrievers
 - Knowledge integration workflows
11. LLM Agentic Framework
 - Agent-based orchestration and LangGraph
12. LLM Application Development
 - LangChain
 - SMOagents
13. New Trends
 - Mixture of Experts (MoE)
 - Multimodal Models

授課方式： 講授 Lecture：100%
 Method of Instruction 分組討論 Group discussion：0%
 案例研討 Case study：0%
 操做練習 Practical exercises：0%
 講授 Lecture：%

教科書： Class notes
 Textbooks

參考書目：
 References

修課須知：
Notice

評量方式：
Grading

- Term Exam: 45%
- Final Project 25%
- Paper Presentation: 10%
- Homework and Programs: 20%

備註說明：
Notes

Students should be familiar with Python and PyTorch and have basic deep learning courses