The Hong Kong University of Science and Technology

UG Course Syllabus (Spring 2025-26)

[Course Title] Large-Scale Machine Learning for Foundation Models

[Course Code] COMP4551

[No. of Credits] 3 Credits

[Prerequisites] COMP 3211 AND COMP 3511

Name: Binhang Yuan

Email: biyuan@ust.hk

Office Hours: 3517 Tuesday 3:00 PM

Course Description

In recent years, foundation models have fundamentally revolutionized the state-of-the-art of artificial intelligence. Thus, the computation in the training or inference of the foundation model could be one of the most important workflows running on top of modern computer systems. This course unravels the secrets of the efficient deployment of such workflows from the system perspective. Specifically, we will i) explain how a modern machine learning system (i.e., PyTorch) works; ii) understand the performance bottleneck of machine learning computation over modern hardware (e.g., Nvidia GPUs); iii) discuss four main parallel strategies in foundation model training (data-, pipeline-, tensor model-, optimizer- parallelism, etc.); iv) real-world deployment of foundation model including efficient inference and fine-tuning.

Intended Learning Outcomes (ILOs)

By the end of this course, students should be able to:

1. Illustrate basic concepts and principles in foundation models.
2. Describe the main parallel training strategies for distributed training.
3. Understand the system bottleneck of computations relevant to foundation models.
4. Be able to deploy some foundation model workflows for real-world applications.

Assessment and Grading

This course will be assessed using criterion-referencing and grades will not be assigned using a curve. Detailed rubrics for each assignment are provided below, outlining the criteria used for evaluation.

Assessments:

[List specific assessed tasks, exams, quizzes, their weightage, and due dates; perhaps, add a summary table as below, to precede the details for each assessment.]

Assessment Task	Contribution to Overall Course grade (%)	Due date
Homework-1	5%	22/02/2026 *
Homework-2	5%	12/03/2026 *
Midterm examination	30%	24/03/2026
Homework-3	5%	10/04/2026
Homework-4	5%	06/05/2026
Final examination	50%	TBD

* Assessment marks for individual assessed tasks will be released within two weeks of the due date.

Mapping of Course ILOs to Assessment Tasks

[add to/delete table as appropriate]

Assessed Task	Mapped ILOs	Explanation
Homeworks	ILO1, ILO2, ILO3. ILO4	The Homework will include relevant questions to evaluate the student's ability and skill listed in ILO1, ILO2, ILO3, ILO4.
mid-term examination	ILO1. ILO2, ILO3	Midterm examination will include relevant questions to evaluate the student's ability and skill listed in ILO1, ILO2, ILO3.
Final examination	ILO1. ILO2, ILO3	Final examination will include relevant questions to evaluate the student's ability and skill listed in ILO1, ILO2, ILO3.

Grading Rubrics

The homework will include a gold standard solution for reference. Grading rubrics will be based on the comparison of the answer and the gold standard solution.

Final Grade Descriptors:

Grades	Short Description	Elaboration on subject grading description
A	Excellent Performance	Demonstrates comprehensive mastery of foundational concepts and advanced techniques in large-scale ML systems for foundation models. Expertly solves complex deployment and optimization problems with exceptional creativity and precision. Actively collaborates and exceeds course expectations through innovative approaches and real-world application.
B	Good Performance	Exhibits strong knowledge and understanding of critical principles and techniques of large-scale ML systems for

		foundation models. Effectively solves deployment and optimization problems with clear analytical skills. Demonstrates a proactive learning attitude and engages positively in teamwork.
C	Satisfactory Performance	Shows adequate comprehension of fundamental principles and practices related to large-scale ML systems for foundation models. Capably addresses standard problems in deployment and optimization, demonstrating reasonable analysis and effort toward practical applications and defined learning outcomes.
D	Marginal Pass	Possesses basic understanding of essential concepts and minimal competence in solving deployment problems related to large-scale ML systems for foundation models. Demonstrates limited analytical abilities but shows potential for improvement and professional growth with further development.
F	Fail	Fails to demonstrate a sufficient grasp of foundational concepts, system optimizations, and techniques for deploying large-scale ML systems for foundation models. Lacks necessary analytical and practical deployment skills. Displays minimal effort and does not meet fundamental learning outcomes for professional or academic progression.

Course AI Policy

The students are encouraged to use AI tools to support their learning, including completing the programming assignment in the homework.

Communication and Feedback

Assessment marks for individual assessed tasks will be communicated via Canvas within two weeks of submission. Feedback on assignments will include the score and the corresponding referred solution. Students who have further questions about the feedback, including marks, should consult the instructor within five working days after the feedback is received.

Resubmission Policy

Resubmission introduces no penalty for homework.

Required Texts and Materials

No required textbook, the slides will be self-explained.

Academic Integrity

Students are expected to adhere to the university's academic integrity policy. Students are expected to uphold HKUST's Academic Honor Code and to maintain the highest standards of academic integrity. The University has zero tolerance of academic misconduct. Please refer to Academic Integrity | HKUST – Academic Registry for the University's definition of plagiarism and ways to avoid cheating and plagiarism.

[Optional] Additional Resources

Not applicable.