

The Hong Kong University of Science and Technology
UG Course Syllabus (Fall 2026-27)

Course Title: Introduction to Generative Artificial Intelligence

Course Code: COMP3213

No. of Credit: 3 credits

Prerequisite: COMP2211

Name: Chaojian Li

Email: chaojian@ust.hk

Office Hours: Thursday 1:30 PM - 2:30 PM, Room 3518

Course Description

This course introduces the principles and practice of Generative AI across multiple modalities, including text, image, and multimodal systems. It covers major paradigms of generative modeling, such as large language models as well as other generative approaches (e.g., diffusion models), with an emphasis on high-level understanding and practical applications. Students will learn how generative AI models are developed, trained, and applied in real-world scenarios. In addition to core concepts, the course covers practical aspects of deploying generative AI systems, including performance considerations, model compression, and system-level design on modern AI hardware. Through hands-on assignments and case studies, students will gain experience in using, analyzing, and improving generative AI applications. Topics also include prompting and interaction with AI systems, evaluation of model outputs, and responsible use of generative technologies.

Intended Learning Outcomes (ILOs)

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

1. Explain the fundamental concepts of generative AI across different modalities, including large language models and other generative modeling paradigms.
2. Describe how generative AI models are trained and applied in real-world scenarios.
3. Apply generative AI tools (e.g., prompting, code assistants) to solve practical computational problems.
4. Analyze and evaluate the performance of AI models using basic profiling and comparison techniques.
5. Recognize the limitations, risks, and responsible use considerations of generative AI systems.

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:

Assessments are aligned with the course ILOs. Assignment due dates are stated by teaching week; the final examination will be scheduled within the University examination period.

Assessment Task	Contribution to Overall Course grade (%)	Due date
Mid-Term Test	30%	Week 6 *
Assignment 1: Profiling Deep Neural Networks	5%	End of Week 8 *
Assignment 2: Model Compression Techniques	10%	End of Week 11 *
Assignment 3: AI-Assisted System Optimization	15%	End of Week 13 *
Final Examination	40%	7-19 December 2026 (University examination period)

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

Mapping of Course ILOs to Assessment Tasks

Assessed Task	Mapped ILOs	Explanation
Mid-Term Test	ILO 1, ILO 2, ILO 5	Assesses understanding of generative AI concepts, model development and application, and responsible-use considerations.
Assignment 1	ILO 1, ILO 3, ILO 4, ILO 5	Assesses the ability to build a PyTorch evaluation pipeline, profile deployment metrics, compare CNN and Transformer models, and comply with task-specific AI-use rules.
Assignment 2	ILO 1, ILO 3, ILO 4, ILO 5	Assesses implementation of quantization and pruning and analysis of accuracy, runtime, and resource trade-offs under a no-AI-tools condition.
Assignment 3	ILO 3, ILO 4, ILO 5	Assesses AI-assisted optimization of a CUDA sparse matrix multiplication kernel, comparison with cuSPARSE baselines, and critical analysis of AI-suggested optimizations.
Final Examination	ILO 1, ILO 2, ILO 4, ILO 5	Assesses integrated understanding, application, evaluation, and responsible analysis of generative AI models and systems.

Grading Rubrics

Each assessment is graded against the published criteria below. Percentages in the rubric tables are weights within the corresponding assessment and sum to 100% of that task's mark.

Examinations: Mid-Term Test (30%) and Final Examination (40%)

Examination questions use item-level marking schemes. Partial credit is awarded where a response demonstrates correct intermediate reasoning. AI tools are not allowed.

Criterion	Weight	Full-credit standard
Conceptual and technical correctness	40%	Accurately explains relevant generative AI concepts, methods, assumptions, and system behavior.
Application and problem solving	30%	Selects and applies appropriate methods to unfamiliar computational or system scenarios.
Analysis and justification	20%	Interprets evidence, compares alternatives, and supports conclusions with coherent technical reasoning.
Clarity and completeness	10%	Presents a complete, well-structured response using precise terminology and necessary working.

Assignment 1: Profiling Deep Neural Networks (5%)

Students build a PyTorch evaluation pipeline for pre-trained ResNet-50 and ViT-S models on an ImageNet subset, measure deployment metrics, and report on model roles and empirical performance. AI tools are allowed for pipeline development but not for profiling experiments.

Criterion	Weight	Full-credit standard
Pipeline correctness and reproducibility	30%	Implements a reliable evaluation pipeline with correct model/data handling and reproducible execution.
Profiling methodology and measurements	25%	Measures accuracy, runtime, GPU utilization, and other stated metrics correctly without AI-generated profiling experiments.
Empirical comparison and analysis	25%	Interprets results rigorously and explains meaningful performance differences between the models.
Survey, explanation, and report quality	20%	Connects CNNs and Transformers to modern GenAI systems and communicates methods, evidence, and conclusions clearly.

Assignment 2: Model Compression Techniques (10%)

Students apply post-training INT8 quantization and unstructured magnitude pruning to the Assignment 1 models, evaluate performance trade-offs, and report on compression in GenAI systems. AI tool use is not allowed.

Criterion	Weight	Full-credit standard
Compression implementation	30%	Correctly implements INT8 quantization and magnitude pruning with appropriate configurations and validation.
Experimental design and reproducibility	25%	Uses controlled, repeatable experiments and records sufficient settings and measurements.
Trade-off analysis	30%	Analyzes accuracy, runtime, memory/resource, and compression trade-offs with evidence and technically sound conclusions.
Survey and report quality	15%	Explains the role of compression in GenAI systems and presents methods and results clearly and concisely.

Assignment 3: AI-Assisted System Optimization (15%)

Students use generative AI tools to optimize a CUDA sparse matrix multiplication kernel and evaluate it against cuSPARSE baselines. AI use is unrestricted, explicitly encouraged, and part of the assessment; detailed interaction logs are required.

Criterion	Weight	Full-credit standard
Performance against cuSPARSE	80%	Performance is evaluated against cuSPARSE baselines under the specified benchmark conditions using the published performance-evaluation scheme.
Report	20%	Provides detailed AI-agent interaction logs and analyzes how different AI-suggested optimizations contribute to performance improvements, with clear methods, results, and conclusions.

Final Grade Descriptors:

Final grades reflect the overall level of achievement across the course ILOs:

Grades	Short Description	Elaboration on subject grading description
A	Excellent Performance	Demonstrates comprehensive and precise understanding of generative AI models and systems; produces robust implementations and rigorous experiments; analyzes trade-offs insightfully; and applies AI tools responsibly and transparently.
B	Good Performance	Demonstrates strong understanding of the main concepts and competent implementation, profiling, and optimization skills; analyzes evidence well with only minor gaps; and follows assessment-specific AI-use requirements.
C	Satisfactory Performance	Demonstrates adequate understanding of core concepts and can complete standard implementation and evaluation tasks; provides generally valid analysis, although depth, reliability, or clarity may be uneven.
D	Marginal Pass	Demonstrates threshold knowledge and limited ability to implement or evaluate generative AI systems; analysis is basic and may contain notable gaps, but minimum course learning outcomes are met.
F	Fail	Demonstrates insufficient understanding of core concepts or cannot reliably implement, evaluate, or reason about the systems studied; does not meet the minimum course learning outcomes or assessment requirements.

Course AI Policy

AI tool use is assessment-specific. For Assignment 1, AI tools are allowed for pipeline development but may not be used for the profiling experiments. For Assignment 2, AI tools are not allowed. For Assignment 3, AI tool use is unrestricted, explicitly encouraged, and evaluated; students must submit detailed interaction logs. AI tools are not allowed in the mid-term test or final examination. Students remain responsible for the correctness and integrity of all submitted work.

Communication and Feedback

Assessment marks for individual assessed tasks will be communicated via Canvas within two weeks of submission. Assignment feedback will be provided through Canvas using the published grading criteria; no separate written feedback report will be issued. Students who have questions about feedback or marks should consult the instructor within five working days after the feedback is received.

Resubmission Policy

Resubmission or reassessment is not permitted, except where an exception is approved under University policy for documented extenuating circumstances.

Required Texts and Materials

No textbook is required. Required materials include lecture slides, assigned readings, software and model documentation, datasets, starter code, and other resources provided through Canvas. Students will use PyTorch and CUDA-based tools and the computing resources specified for the course.

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.

Additional Resources

Official PyTorch and NVIDIA CUDA documentation, selected research papers, and model documentation announced through Canvas.