AIAA 4051

Introduction to Natural Language Processing

2026 Spring

CHECK ON CANVAS FOR UPDATES DURING THE SEMESTER!!!!!!!

General Information

- Time: Monday, Wednesday 12:00pm-1:20pm
- In-person lecture: E1-101
- Website: https://chenrf1121.github.io/hkustgz-nlp/
- Instructor: Sihong Xie (sihongxie@hkust-gz.edu.cn)
- Teaching assistants
    - Yazheng Liu (yliu533@connect.hkust-gz.edu.cn)
    - Rufeng Chen (rchen514@connect.hkust-gz.edu.cn)
    - Yue Chang (ychang500@connect.hkust-gz.edu.cn)

Course Logistics

Computer and Projects:

- Projects: must be coded in PyTorch (a Python deep learning package). No other programming languages (R, Matlab, Java, C++) or deep learning frameworks (Tensorflow) will be accepted.
- In this course, each student will be provided with a personal account on the Diandong platform, which grants access to a fixed amount of computing power. Detailed video tutorials will be provided to guide students on how to effectively use the platform.

Academic Integrity:

- Acceptable use of any Artificial Intelligence (AI) generated or aided material in your work is governed by the same rules as using non-AI material: it must be in accordance with the class, department, and University policies on Integrity, and you must cite or otherwise provide attribution. For more details, please refer to the HKUST-GZ Academic Quality Policy Framework, which outlines the University's academic integrity guidelines.

Assessments:

Assessment Task	Contribution to Overall Course grade (%)
In-class Quiz	24%
Homework	5%

Individual coding project	10%
Final team research project	35%
Final closed-book written exam	26%

1. **In-class Quiz:**

   Each class will have a closed-book, in-class, written quiz in the last 5 minutes of the class. During the quiz, students **are not permitted** to use any electronic devices. **Bring your own paper (1 piece of A4 should be enough) and pens**.

2. **Final closed-book written exam:**

   Cover everything from the lectures. Question types: single/multi-choice, short answer, proving. No calculation is needed.

3. **Final team research project:**

   The team research project grade including 2% team registration, poster presentations, and final reports.For the team projects, teams must consist of 2-3 members, and individual submissions will not be accepted. If you are having trouble finding a teammate, the instructor and TAs will assign you to a team. TA will interview each team, who will have 10 minutes to answer questions from the TA regarding their codes and experimental results.

Final Grade Descriptors:

Grades	Short Description	The corresponding 100-point system
A+	Excellent Performance	95-100
A	Excellent Performance	90-94
A-	Excellent Performance	85-89
B+	Good Performance	80-84
B	Good Performance	75-79
B-	Good Performance	70-74
C+	Satisfactory Performance	67-69
C	Satisfactory Performance	64-66
C-	Satisfactory Performance	61-63
D	Marginal Pass	55-60
F	Fail	0-54

Late policy:

- You will have a total quota of **7 days** for submissions and you can use them freely for individual assignments of your choice. Due dates will be stated on all assignments. If we erroneously set conflicting dates across Canvas, and the assignment document, please inform us. Until any error is corrected the earliest date applies. Students are expected to be able to submit work correctly online and to back up their data. Therefore, “forgetting to click submit”, “computer crashes”, etc, are not acceptable lateness excuses. Note that online sites’ clocks may not match yours perfectly, so don’t wait until the last moment to submit. Note that other than these late days, we will not be making exceptions and extending deadlines except for health reasons, so please try to be frugal with your late days and use them only if necessary. Assignments that are late beyond the allowed late days will lose **25%** of the total

score per late day (**any period less than 24 hours is counted as one late day**).

Grading Rubrics:

- For Homework, individual coding projects, and the Final closed-book written exam, we will provide the score for each step, the grading will be based on the steps completed.
- **Final team research project—Poster presentation (10% graded by TAs and instructor)**

Student	Introduction to the Topic	Organization and Logic	Clarity of Presentation	Fluency of Presentation	Overall

Score Rubrics: **1** (Poor) | **2** (Below Average) | **3** (Borderline) | **4** (Up Average) | **5** (Good) | **6** (Very Good) | **7** (Excellent)

- **Final team research project—Final report (10% graded by TAs and instructor)**

Content Quality (60 points)	Fundamental Concept Understanding (25 points)	Clearly explains key NLP concepts with well-structured definitions and real-world examples. (22-25 points)	Covers key NLP topics but lacks depth or fails to provide concrete examples. (17-21 points)	Mentions concepts but with vague explanations or missing critical details. (10-16 points)	Displays little to no understanding of core NLP concepts or applications. (0-9 points)
	Score				
	Review and Analyze Existing Research (20 points)	Provides a comprehensive literature review with relevant and up-to-date papers, critically analyzing methodologies and findings in NLP. (18-20 points)	Covers related works but lacks deep comparison or critical evaluation. (14-17 points)	Lists some related works without much discussion or relevance. (8-13 points)	Little to no research cited, or references are outdated/irrelevant. (0-7 points)
	Score				
	Practical Application & Critical Analysis (15 points)	Demonstrates strong analytical skills by applying NLP concepts to real-world scenarios with insightful evaluation and innovative solutions. (13-15 points)	Provides relevant applications and some critical analysis but lacks depth, originality, or a strong connection to practical challenges. (10-12 points)	Mentions basic applications but offers superficial analysis, with limited critical thinking or weak real-world relevance. (6-9 points)	Shows little to no application of concepts, with unclear, incorrect, or missing critical analysis.(0-5 points)
	Score				
Expression and Format (40 points)	Language Expression (15 points)	Clear and fluent language expression, easy to understand. (13-15 points)	Generally good expression, but some parts are not fluent. (10-12 points)	Unclear expression, affects understanding.(6-9 points)	Severely unclear expression. (0-5 points)
	Score				
	Format Adherence	Strictly follows format requirements, well-structured. (13-15	Basically follows format requirements, with minor deviations.	Insufficient format adherence, noticeable deviations. (6-9	Fails to follow format requirements. (0-5 points)

	(15 points)	points)	(10-12 points)	points)	
	Score				
	References (10 points)	Appropriately selected references, correct citation format. (9-10 points)	Generally appropriate references, minor issues in format or selection. (7-8 points)	Inadequate references or incorrect citation format. (4-6 points)	Severely lacking references or improper format. (0-3 points)
	Score				

- **Final research project—Code quality (15% graded by TAs and instructor)**

Code Quality (100 points)	Specifications (50 points)	The code works and produces the correct results and displays them correctly. This code meets all of the specifications (40-50 points)	The code works and produces the correct results and displays them correctly. It also meets most of the other specifications. (30-39 points)	The program produces correct results but does not display them correctly. (15-29 points)	The code is producing incorrect results. (0-14 points)
	Score				
	Readability (30 points)	The code is exceptionally well organized and very easy to follow. (25-30 points)	The code is fairly easy to read. (20-24 points)	The code is readable only by someone who knows what it is supposed to be doing. (10-19 points)	The code is poorly organized and very difficult to read. (0-9 points)
	Score				
	Efficiency (20 points)	The code is extremely efficient without sacrificing readability and understanding. (15-20 points)	The code is fairly efficient without sacrificing readability and understanding. (10-14 points)	The code is brute force and unnecessarily long. (6-9 points)	The code is huge and appears to be patched together.(0-5 points)
	Score				

Schedule of Topics:

	Date	Topics	Assignments, dues, and important notes.
1	2026-01-26	Course logistics; basic math and AI review	
2	2026-01-28	Language modeling and N-Grams	
3	2026-02-02	Tokenization and embeddings	
4	2026-02-04	Hidden Markov Model I	
5	2026-02-09	Hidden Markov Model II	Winter Break (2026-02-11 to 2026-02-23)
6	2026-02-25	Syntactic trees	

7	2026-02-28	Syntactic parsing	Make up day HW release
8	2026-03-02	Recurrent neural networks	
9	2026-03-04	Machine translation	
10	2026-03-09	Attention and Transformers I	
11	2026-03-11	Attention and Transformers II	HW due Individual Project release Final research project release
12	2026-03-16	Mixture of experts	
13	2026-03-18	Learning: Pretraining	Final research project topic selection and team registration
14	2026-03-23	Supervised Fine-Tuning	
15	2026-03-25	Lora	HW grade release
16	2026-03-30	RLHF	Individual Project due
17	2026-04-01	PPO	Qingming Festival (2026-04-04 to 2026-04-06)
18	2026-04-08	DPO	
19	2026-04-13	Training data synthetization	
20	2026-04-15	RAG	Individual Project grade release
21	2026-04-20	Scaling laws	
22	2026-04-22	Inference: KV cache	
23	2026-04-27	Model compression	
24	2026-04-29	Diffusion model I	Labor Day (2026-05-01 to 2025-05-05)
25	2026-05-06	Diffusion model II	
26	2026-05-11	Poster presentation	
	2026-05-18		Final research project code and report due
	2026-05-19		Team interview

	TBD		Final exam