

Extended Syllabus

Course Title	Introduction to Data-Centric AI for Language Technology	Semester	Spring 2026
Credit	3 credits	Course Number	ENGG232
Class Time	Wed / Fri 15:00 ~ 16:15	Enrollment Eligibility	Undergraduate Junior and Senior

Instructor's Photo	Name: Heo Yoon Seok (허윤석)	Homepage: haicolab.sogang.ac.kr
	E-mail: nlpysheo@sogang.ac.kr	Telephone: 02-705-8295
	Office: J825 Office Hours: Wed/Fri 16:30 ~ 18:30 (* By appointment via email in advance)	

I . Course Overview

1. Description					
This course is designed to understand, evaluate, and improve Large Language Models (LLMs) from a "Data-Centric" perspective. Instead of focusing on complex model architecture or heavy coding, students will learn how to control AI using Prompt Engineering and how to evaluate AI performance by constructing their own Benchmark Datasets. The course empowers students to leverage their domain expertise (Linguistics, Culture, Education, etc.) to discern the critical differences between human and AI capabilities and construct high-quality evaluation data to rigorously verify these limitations. The final goal is to produce a KCI-level academic journal by building a domain-specific benchmark. ● NOTE1 - Differential Grading: Undergraduate students will be evaluated separately from graduate students to ensure fairness in grading standards. ● NOTE2 - Adjustable Project Scope: The scope and expectations for the final research paper can be adjusted through consultation with the instructor. Students are encouraged to discuss their project direction early to align with their capabilities. * The finalized course syllabus will be distributed on the first day of class.					
2. Recommended Prerequisites					
- Linguistics with AI (ENG2112), Data Analysis for AI (ENG3510), or equivalent basic programming/AI courses. - Basic interest in language technology and humanities.					
3. Course Format (%)					
Lecture	Discussion	Experiment	Projects	Presentations	Other
30%	5%	15%	35%	15%	0%
4. Evaluation (%)					

Mid-term Exam	Final exam	Quizzes	Presentations	Projects	Assignments	Attendance	Other
15%	0%	10%	10%	40%	20%	5%	0%

II. Course Objectives

Knowledge:

- Understand the fundamental principles of LLMs and the importance of Data-Centric AI.
- Comprehend various **evaluation benchmarks** (MMLU, GSM8K, etc.) and evaluation metrics (ROUGE, BLEU, **LLM-as-a-judge**).
- Recognize ethical issues in AI, including bias, hallucination, and toxicity.
- Understand the methodologies of **synthetic data generation**

Skill:

- Master Prompt Engineering techniques (Few-shot, Chain-of-Thought, Tree-of-Thought).
- Construct and refine Synthetic Datasets using LLM-based generation methods.
- Execute basic Python code for data processing and API usage via Google Colab or VSCODE.

Attitude:

- Develop a critical perspective on AI technology based on humanities insights.
- Collaborate effectively in a research team to solve real-world problems.
- Adhere to ethical guidelines in data construction and AI evaluation.

III. Course Format

(* In detail)

This course combines theoretical lectures with hands-on lab sessions.

- Weeks 1-8: Focus on individual skill-building (Prompting, Evaluation, Data Generation).
- Weeks 9-16: Operated as a "Research Lab." Students form teams to build a benchmark dataset, with weekly milestones and seminars replacing traditional lectures.

[Grad students] Every team MUST submit at least one KCI-level or SCIE journal at the end of the semester.

- **NOTE1 - Differential Grading: Undergraduate students will be evaluated separately from graduate students to ensure fairness in grading standards.**
- **NOTE2 - Adjustable Project Scope: The scope and expectations for the final research paper can be adjusted through consultation with the instructor. Students are encouraged to discuss their project direction early to align with their capabilities.**

IV. Course Requirements and Grading Criteria

1. Required Tools & Costs

- Access to advanced AI models (e.g., ChatGPT Plus, Gemini Advanced) may be required (* Will be announced)
- Students should be prepared for subscription fees and potential separate API usage costs necessary for data generation and evaluation.

2. Grading Criteria

- **NOTE1 - Differential Grading: Undergraduate students will be evaluated separately from graduate students to ensure fairness in grading standards.**
- **NOTE2 - Adjustable Project Scope: The scope and expectations for the final research paper can be adjusted through consultation with the instructor. Students are encouraged to discuss their project direction early to align with their capabilities.**
- Lab sessions and assignment (20%)
 - In-class lab session: Students are required to participate in weekly lab sessions.
 - Submission: Completed Colab notebooks or progress reports must be submitted after each lab session.
- Mid-term Project: Individual Proposal (15%)
 - Format: A 5-minute video presentation pitching a unique benchmark idea.
 - Content: The proposal must explain the target domain, the specific capability gap in AI, and the planned evaluation method.
 - Peer Review: Students will watch and evaluate peers' proposals to form research teams.
 - Note: This project replaces the traditional written mid-term exam.
- Final Project: Team Benchmark Construction (40%)
 - Goal: Construct a novel benchmark dataset and evaluate current LLMs.
 - Components:
 - ✓ Dataset: A clean, domain-specific benchmark dataset (JSON format).
 - ✓ Paper: A research paper following the NeurIPS Datasets & Benchmarks track format.
 - ✓ Poster Presentation: Final presentation of the research findings.
 - ✓ Grade Incentive: Teams producing at least one KCI-level journal or SCIE paper quality result will be awarded an A+.
- Presentation: Paper seminar (10%)
- Quiz: 10%
- Attendance: 5%

V. Course Policies

- **The FA policy and all course regulations will be applied based on the university's academic policies.** Additional course-related regulations will follow the guidelines announced on the cyber campus.
- Any form of academic dishonesty will result in an F grade.
- Except for officially recognized reasons for absence (e.g., military reserve duty, family emergencies), **NO make-up quizzes or exams will be given, and no extensions for assignment submissions will be allowed.**
- AI Tools Policy: Collaboration with generative AI tools such as Co-Pilot and ChatGPT is allowed, treating them as collaborators in the problem-solving process. However, the direct solicitation of answers or copying solutions, whether from peers or external sources, is strictly prohibited. Note that indiscriminate reliance on AI tools can seriously harm learning efficiency, and this may be revealed as a significant gap between high assignment performance and low exam scores. Therefore, students are strongly encouraged to use AI tools as collaborators rather than as substitutes for their own work.

VI. Materials and References

- Weekly PowerPoint slides and lab materials will be provided on CyberCampus
- Reading materials will be announced in details on class.
- **Students are highly recommended for subscription fees of AI models(e.g. ChatGPT, Cluade, Gemini etc) and prepared for potential API usage costs**
- Students are asked to bring their own laptops on class..

VII. Course Schedule

(* Subject to change)

Week 1 (Mar 4 th / 6 th)	Learning Objectives	Orientation and AI Literacy
	Topics	Introduction to Data-Centric AI: Why Data Matters More Than Models?
	Class Work (Methods)	Lecture / Lab session
	Materials (Required Readings)	TBA
	Assignments	TBA
Week 2	Learning Objectives	LLM Basics & Prompt Engineering

(Mar 11th / 13th)	Topics	How LLMs Work: Instruction & Context.
	Class Work (Methods)	Lecture / Lab session
	Materials (Required Readings)	TBA
	Assignments	Hands-on practice: Prompt engineering
Week 3 (Mar 18th / 20th)	Learning Objectives	Advanced Reasoning & Logic Design
	Topics	Reasoning, Chain-of-Thought(CoT), Tree-of-Thought(ToT)
	Class Work (Methods)	Lecture / Lab session
	Materials (Required Readings)	TBA
	Assignments	Hands-on practice: CoT, ToT
Week 4 (Mar 25th / 27th)	Learning Objectives	Evaluation and Benchmark
	Topics	LLM Benchmarks, Linguistic Metrics, and LLM-as-a-judge
	Class Work (Methods)	Lecture / Lab session
	Materials (Required Readings)	TBA
	Assignments	- Hands-on practice: LLM-as-a-judge - Benchmark survey
Week 5 (April 1st / 3rd (No class))	Learning Objectives	The shadows of LLMs
	Topics	Bias and Ethics / Hallucination
	Class Work (Methods)	Lecture (April 1st) / No class on April 3rd(School holiday)
	Materials (Required Readings)	TBA
	Assignments	NLP Bias Survey
Week 6 (April 8th /	Learning Objectives	Synthetic Data Generation
	Topics	Self-instruct & Evol-instruct, Alignment

10th)	Class Work (Methods)	Lecture Lab session: Building a pipeline-for generating and refining QA datasets
	Materials (Required Readings)	TBA
	Assignments	Hands-on practice for synthetic data generation
Week 7 (Apr 15th/ 17th)	Learning Objectives	Synthetic Data Generation
	Topics	Case study / Data quality
	Class Work (Methods)	Lecture
	Materials (Required Readings)	TBA
	Assignments	Hands-on practice
Week 8 (Apr 22nd / 24th)	Learning Objectives	Mid-term
	Topics	
	Class Work (Methods)	
	Materials (Required Readings)	
	Assignments	A proposal for final project, video presentation, and peer review
Week 9 (Apr 29th / May 1st)	Learning Objectives	Project kick-off
	Topics	Team building and Overleaf
	Class Work (Methods)	Lecture / No class on May 1st (Labor Day)
	Materials (Required Readings)	TBA
	Assignments	TBA
Week 10	Learning Objectives	Project milestone 1

(May 6th / 8th)	Topics	Data schema and seed data construction
	Class Work (Methods)	Progress presentation / Paper seminar
	Materials (Required Readings)	TBA
	Assignments	Project
Week 11 (May 13th / 15th)	Learning Objectives	Project milestone 2
	Topics	Synthetic Data Generation / Data quality filtering
	Class Work (Methods)	Progress presentation / Paper seminar
	Materials (Required Readings)	TBA
	Assignments	Project
Week 12 (May 20th / 22th)	Learning Objectives	Project milestone 3
	Topics	Project Progress Presentation / Seminar
	Class Work (Methods)	
	Materials (Required Readings)	
Week 13 (dd/mm)	Assignments	
	Learning Objectives	Project milestone 4
	Topics	Project Progress Presentation / Seminar
	Class Work (Methods)	
	Materials (Required Readings)	
Week 14	Assignments	
	Learning Objectives	Project milestone 5

(dd/mm)	Topics	Paper Drafting
	Class Work (Methods)	
	Materials (Required Readings)	
	Assignments	
Week 15 (dd/mm)	Learning Objectives	Project milestone 6
	Topics	Peer Review and Final polish
	Class Work (Methods)	
	Materials (Required Readings)	
	Assignments	
Week 16 (dd/mm)	Learning Objectives	Project Presentation
	Topics	
	Class Work (Methods)	
	Materials (Required Readings)	
	Assignments	

VIII. Special Accommodations

Students are strongly encouraged to make active use of **office hours (OH)** for questions, feedback, and additional support throughout the semester.

In addition, recitation sessions may be scheduled during the semester if necessary. These sessions are not mandatory, but will be designed to provide opportunities for extended Q&A, clarification of lecture materials, and preliminary guidance for upcoming assignments. The specific times and details will be announced late

IX. Aid for the Challenged Students

Everyone will be welcomed, and I will try to make the class environment accessible for anyone with any type of challenge. Please do not hesitate to contact me.