The Chinese University of Hong Kong
Department of Linguistics and Modern Languages
First Term, 2026-27

Course Code: LING3401 **English Title:** Linguistics and Information Technology (IT) **Chinese Title:** 語言學與資訊科技
Course description: This course bridges linguistics and modern information technology by introducing computational linguistics and focusing on current developments in the Language Sciences, Natural Language Processing (NLP), Large Language Models (LLMs), and the Brain Sciences. Students will explore core principles of computational linguistics and brain sciences, understand their relationships and applications in linguistic analysis, gain practical skills in text mining and corpus analysis, and learn to work with large language models. The curriculum emphasizes hands-on experience with LLM tools through interactive tutorials and guided projects, enabling students to analyze linguistic data, apply tools to solve real-world problems, and evaluate the capabilities and limitations of language models. Target students include linguistics majors without math or programming experience, senior college students interested in the field, and anyone seeking to understand the intersection of Language Sciences and AI.

Learning outcomes

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

- Understand the core concepts of computational linguistics, NLP, and LLMs
- Compare LLMs and the human brain to understand the principles of language models and human cognition.
- Gain insights into the development of LLMs and their practical applications
- Apply NLP and corpus analysis techniques for linguistic analysis
- Develop basic skills in using LLMs and NLP tools to solve practical problems
- Appreciate how these tools contribute to language sciences, both in theory and in practice

Course syllabus

Topic	Contents/fundamental concepts
Introduction to Computational Linguistics	Overview of computational linguistics, NLP, and language models and their applications.
Introduction to NLP and LLMs	Principles of NLP and LLM techniques
Part-of-Speech Tagging	Part-of-speech tagging and related analyses
Semantic Role Assignment and Linguistic Structure Parsing	Word embeddings and vector semantics Linguistic structure annotation
Sentiment Analysis and Text Classification	Spam detection, topic categorization, and sentiment analysis
Corpus Linguistics	Introduction to linguistic corpora and corpus analysis.
NLP and LLM application	Real-world application of NLP and LLM

Course components, Teaching modes, Learning activities

Teaching Modes and Learning Activities

On-site face-to-face (hybrid or online modes may be available in extreme conditions)	Percentage of time 100%
Lectures	70% (2 hours/week)
Interactive tutorial	30% (0.75 hour/week)
or Laboratory activities	30% (0.75 hour/week)
or Discussion of case	30% (0.75 hour/week)
Out-of-classroom	Percentage of time 100%
Self study	50% (1 hours/week)
Project work	50% (1 hour/week)

Expected time allocation of each of the below learning activities during the course (per week):

Lecture **(hr)** **in /out class**	**Interactive tutorial** **(hr)** **in /out class**	**Lab** **(hr)** **in /out class**	**Discussion of case** **(hr)** **in class**	**Reading** **(hr)** **out class**	**Projects** **(hr)** **out class**	**Web-based teaching** **(hr)** **in /out class**	**Other** **(hr)** **in /out class**
2	1		0.2	1	0.2		
M	M		M	O	M		

M: Mandatory activity in the course
O: Optional activity
NA: Not applicable

Assessment type and rubrics

Assessment type	**Description**	**Percentage**
Mid-term and Final Exams	Mid-term and final exams will assess the didactic information presented in the lectures. Mid-term: 15% Final: 20%	35%
Final Project Paper	Design and conduct a survey (e.g., aptitude of LLM) and write a report or conduct a literature review on a topic in Computational Linguistics, NLP, or LLM and the brain that interests you.	35%
Experimental report	Write a testing report for each one. Describe the tasks or analyses you performed, the tools you used, and the methods you followed to obtain the results from those tools.	15%
Research presentation	Students are grouped to present a paper or application on a specific technique that interests them (e.g., how to use an NLP or LLM-related app or tool).	15%

Required and recommended readings

This is the list of recommended readings, and more readings will be announced in class. All readings are posted on BlackBoard (http://blackboard.cuhk.edu.hk).

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Week 14: Optional (critical perspective)

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. Communications Biology, 5, Article 134. https://doi.org/10.1038/s42003-022-03036-1

Week 10: Recommended

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., & Jain, A. (2024). Structured information extraction from scientific text with large language models. Nature Communications, 15, Article 1418. https://doi.org/10.1038/s41467-024-45563-x

Week 11: Optional

Glass, L., Dickinson, M., Brew, C., & Meurers, D. (2024). Language and computers (2nd ed.). Textbooks in Language Sciences 14. Language Science Press. https://doi.org/10.5281/zenodo.12730906

Weeks 1, 3, 5, 6, 7, 12, 13, 14: Required

Gougherty, A. V., & Clipp, H. L. (2024). Testing the reliability of an AI-based large language model to extract ecological information from the scientific literature. npj Biodiversity, 3, Article 13. https://doi.org/10.1038/s44185-024-00043-9

Week 11: Recommended

Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., & Birch, A. (2022). Survey of low-resource machine translation. Computational Linguistics, 48(3), 673–732. https://doi.org/10.1162/coli_a_00446

Week 12: Optional (Introduction only)

Hammond, M. (2020). Python for linguists. Cambridge University Press. https://doi.org/10.1017/9781108642408

Week 3: Recommended (tutorial preparation only)

Hugging Face. (n.d.). What is an agent? In Agents course, Unit 1: Introduction to agents. Retrieved [date], from https://huggingface.co/learn/agents-course/unit1/what-are-agents

Week 13: Recommended (practical supplementary resource)

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453–458. https://doi.org/10.1038/nature17637

Week 7: Optional

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248, 1–38. https://doi.org/10.1145/3571730

Week 9: Optional

Jurafsky, D., & Martin, J. H. (2026). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models (3rd ed.) [Online manuscript, released January 6, 2026]. https://web.stanford.edu/~jurafsky/slp3/

Weeks 1, 3, 5-9, 12: Required

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2, Article 4. https://doi.org/10.3389/neuro.06.004.2008

Week 10: Optional (advanced)

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://doi.org/10.54675/EWZM9535

Week 14: Recommended (also usable in Week 13)

Sharwood Smith, M. (2017). Introducing language and cognition: A map of the mind. Cambridge University Press. https://doi.org/10.1017/9781316591505

Week 2: Required

Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024). Language in brains, minds, and machines. Annual Review of Neuroscience, 47, 277–301. https://doi.org/10.1146/annurev-neuro-120623-101142

Week 10: Required; Week 2: Optional

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research [Comment]. Nature, 614(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

Week 11: Required

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998–6008). https://arxiv.org/abs/1706.03762

Week 8: Optional (advanced)

Supplementary Readings and Materials

Alammar, J. (2018, June 27). The illustrated Transformer [Blog post]. https://jalammar.github.io/illustrated-transformer/

Week 8 companion to Vaswani et al. (2017). Note: originally published 2018; subsequently updated (July 2020 revision to the positional-encoding figures; 2025 addition of a free companion short course).

Alammar, J., & Grootendorst, M. (2024). Looking inside large language models. In Hands-on large language models: Language understanding and generation (Ch. 3). O'Reilly Media.

Optional alternative to the blog post above. Alammar describes this chapter as the updated and expanded version of "The Illustrated Transformer," covering post-2017 developments such as multi-query attention and RoPE positional embeddings. Recommended if you want students to see the modern architecture rather than only the 2017 original.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://arxiv.org/abs/2005.14165

Sinclair, S., & Rockwell, G. (2016). Voyant Tools [Web application]. https://voyant-tools.org/

Feedback for evaluation

Students are welcome to provide comments and feedback by emailing them to the instructor's email address or by speaking with the instructor.

Grade Descriptors

Grade	Overall Course
A	Demonstrates exceptional understanding of key concepts in linguistics and information technology, including the ability to: clearly explain foundational computational linguistics concepts, such as tokenization and syntactic parsing; critically analyze the advantages and limitations of NLP tools; effectively apply corpus analysis techniques using tools; evaluate ethical considerations related to NLP, such as bias and fairness in LLMs; conduct a comprehensive and well-structured final project showcasing in-depth knowledge and application of learned methods.
A-	Shows a strong understanding of key concepts with minor weaknesses in one area, such as the ability to describe or apply corpus analysis methods or evaluate the limitations of computational tools.
B	Demonstrates a good understanding of the subject with weaknesses in no more than two major areas. A student may: show solid knowledge of computational linguistics and NLP concepts; provide acceptable but less detailed analysis of linguistic data or ethical issues.
C	Demonstrates an understanding of the course material with noticeable weaknesses in several areas, such as incomplete descriptions of NLP concepts or limited application of tools in assignments.
D	Demonstrates minimal understanding of the course material with significant weaknesses in most

	key components.
F	Fails to demonstrate sufficient understanding of the core concepts, with critical gaps in knowledge and application across the course content.

Course Schedule

Class/ week	Date	Topics and requirements	Tutorial
Part 1. How humans and computers deal with language			
Week 1	Sep 09	Why Language + AI? (Course overview)	*AI Language Challenge: test several language tasks with an LLM and identify successes and failures*
Week 2	Sep 16	Language, Cognition, and the Brain	*Human vs AI prediction: measure and calculate predictability and perplexity*
Week 3	Sep 23	Turning Language into Data	*Set up the Python environment and explore a corpus: tokens, types, annotations, etc.*
Week 4	Sep 30	No class (make up in Dec 07)	
Week 5	Oct 07	From Frequency to Prediction	*Build a tiny language model*
Part 2. From simple models to LLMs			
Week 6	Oct 14	Machine Learning and Model Comparison	*Text classification and detection*
Week 7	Oct 21	Meaning in Humans and Machines	*Human versus model semantic space: understanding and visualization*
Week 8	Oct 28	From Simple Language Models to Large Language Models	*Simple models vs. LLMs in different tasks*
Week 9	Nov 04	Using and Evaluating LLMs	*Mid-term (1.5 hours)*
Week 10	Nov 11	Do Human Brains and Language Models Represent Language Similarly?	*Tutorial: "Which model looks more brain-like?"* *Student Presentations*
Part 3. Applying LLMs in research and practice			
Week 11	Nov 18	LLMs for Language-Science Research	*Tutorial: AI research assistant* *Student Presentations*
Week 12	Nov 25	AI for Speech, Multilingual Research and Translation	*Student Presentations*
Week 13	Dec 02	AI for Language Tasks in Study, Work and Daily Life	*Demo: AI agents* *Student Presentations*
Week 14	Dec 07	Responsible AI, Integration and Future Pathways	*Final Exam (1.5 hours)*

Contact Details for Teacher(s) and Teaching Assistant(s)

Professor/Lecturer/Instructor:	Prof.
Name:	FENG Gangyi
Office Location:	G09 KKB (or 401 4/F HYS)
Office Hours:	Friday 16:00-17:00 (or by appt) English, Cantonese (native), Mandarin (native)
Telephone:	3943-3190

Email:	g.feng@cuhk.edu.hk
Teaching Venue:	ERB 712
Class/Tutorial Time:	Wed 14:30-16:15 (Lecture), 16:30 - 17:15 (Tutorial)
Website:	https://neurolanglab.github.io/people.html
Other information:	Google Scholar: Gangyi Feng (冯刚毅)

Teaching Assistant/Tutor:	TA
Name:	Qin Kangning
Office Hours and Location:	Mon 14:00-16:00，Zoom PMI: 234 852 0487
Telephone:	6584 7652
Email:	1155275319@link.cuhk.edu.hk
Teaching Venue:	ERB 712
Other information:	

Details of course website

All course announcements will be posted on Blackboard Ultra (https://blackboard.cuhk.edu.hk/ultra/).

Academic honesty and plagiarism

Attention is drawn to University policy and regulations on honesty in academic work, and to the disciplinary guidelines and procedures applicable to breaches of such policy and regulations. Details may be found at http://www.cuhk.edu.hk/policy/academichonesty/.

With each assignment, students will be required to submit a **signed declaration** that they are aware of these policies, regulations, guidelines and procedures.

In the case of group projects, all members of the group should be asked to sign the declaration, each of whom is responsible and liable to disciplinary actions, irrespective of whether he/she has signed the declaration and whether he/she has contributed, directly or indirectly, to the problematic contents.
For assignments in the form of a computer-generated document that is principally text-based and submitted via VeriGuide, the statement, in the form of a receipt, will be issued by the system upon students' uploading of the soft copy of the assignment.
Students are fully aware that their work may be investigated by AI content detection software to determine originality.
Students are fully aware of the AI approach(es) adopted in the course. In the case where some AI tools are allowed, students have made proper acknowledgment and citations as suggested by the course teacher.

Assignments without a properly signed declaration will not be graded by teachers.

Only the final version of the assignment should be submitted via VeriGuide.

The submission of a piece of work, or a part of a piece of work, for more than one purpose (e.g. to satisfy the requirements in two different courses) without declaration to this effect shall be regarded as having committed undeclared multiple submissions. It is common and acceptable to reuse a turn of phrase or a sentence or two from one's own work; but wholesale reuse is problematic. In any case, agreement from the course teacher(s) concerned should be obtained prior to the submission of the piece of work.

The copyright of the teaching materials, including lecture notes, assignments and examination questions, etc., produced by staff members/ teachers of The Chinese University of Hong Kong (CUHK) belongs to CUHK. Students may download the teaching materials produced by the staff members/ teachers from the Learning Management Systems, e.g. Blackboard, adopted by CUHK for their own educational use, but shall

not distribute/ share/ copy the materials to a third-party without seeking prior permission from the staff members/ teachers concerned.

Use of AI tools for grading and provision of feedback on student work
AI tools supported by the university may be used to grade and provide feedback on some assignments and assessment tasks. Their use will depend on the nature of the assignments and assessments.

Use of AI tools in teaching, learning, and assessments
The use of AI tools is prohibited for mid-term and final exams.
The use of AI tools is allowed with explicit acknowledgment and proper citation for assignments.

i. An example of acknowledgment: 'I acknowledge the use of (name of AI tool – e.g., ChatGPT (https://chat.openai.com/)) to (specify the support, e.g., plan my essay, generate ideas for the content, ask for examples of data collection instruments, get the dates of historical events, etc.).'

ii. An example citation. OpenAI. (2023). ChatGPT (Mar 20 version). https://chat.openai.com/chat

The type of AI tools permitted:
The type can be implemented in various forms. These AI tools involve conducting related surveys and research, such as data collection, analysis, and model building, which is permitted.

The extent of their use:
Activities related to research that do not directly involve generating research paper content, assignments, or manuscripts are permitted. Permitted AI-related research activities include testing LLMs, comparing ChatGPT with humans, using machine learning or neural network algorithms to analyze data, AI-assisted programming, developing computational models, and more.

Guideline on sharing lecture recordings
The make-up class and any online lecture will be recorded and shared on the university's online learning platform (e.g., Blackboard Ultra).

Late submission policy
For the course assignments (i.e., Final Project Paper, Experimental Report, Research Presentation), late submissions are allowed upon proper application with justification and a valid official document (e.g., a medical certificate). The application must be submitted to the instructor 14 days before the assignment deadline. No penalty is assessed in approved cases.

Make-up test arrangements
Make-up tests for the midterm and final exams are usually not allowed. In very exceptional cases, students seeking permission to be absent from an assessment and requesting a make-up arrangement must submit a written application to the course teacher 14 days before the scheduled test date, along with appropriate supporting documentation.
A 10% deduction from the full mark of the test will be applied, regardless of the reason for the absence. Only one make-up test will be scheduled for the entire class, to be held within 5 working days or 7 calendar days after the originally scheduled examination date.