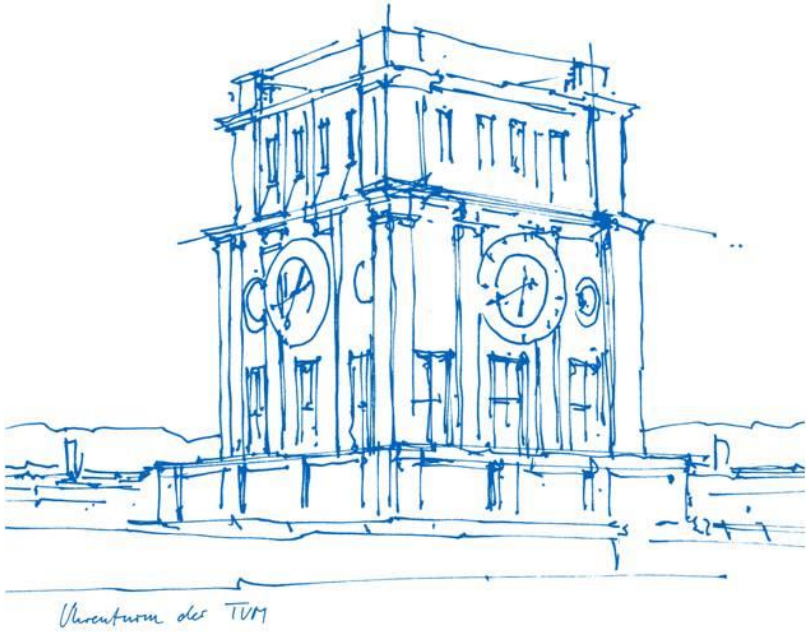
Seminar (IN0014)

Security of Large Language Models (LLMs)

Pre-meeting



Mark Huasong Meng (Dr)
mark.meng@tum.de
3 Feb 2024

Self-Introduction

- Mark MENG Huasong from Singapore

- Post-doc researcher under the Chair of Software Engineering & AI at Heilbronn Campus
- Education
  - B.Eng.(Hons), Computer Science, Nanyang Technological University, 2014
  - M.Comp., Infocomm Security, National University of Singapore, 2016
  - Ph.D., Computer Science, National University of Singapore, 2024
- Work Experience
  - Been a software engineer/analyst in the cyber-security R&D industry (2014-2019)

About this seminar module

This seminar module aims to give you a chance to explore the landscape of diverse security issues in the world of Large language models (LLMs).

You are encouraged to take this module if:

- You are interested in security or trustworthy AI.
- You want to gain some working experience with LLMs.
- You are willing to (ethically) explore the vulnerability and weakness of mainstream LLMs.
- You want to harangue your knowledge about LLM hallucination/jailbreaking with your friends one day.

Prerequisite

- Students who want to enroll in this module are recommended to have gained background knowledge about Machine Learning (ML) and Artificial Intelligence (AI).
- Students are strongly recommended to gain some working experience with LLMs and/or generative AI foundation models (FMs) (e.g., code, image, etc.) prior to attending this module.
  - Real-world observations are important!
  - Start using/playing with LLMs/FMs now!
    - Let them write for you, let them code for you, let them draw for you

Course Logistics – Topics to be covered

The broad sense of security topics for LLMs include but are not limited to:

Attack, Jailbreak, **Backdoor**
Testing, Verification, Hallucination,
Defense, Privacy Leakage, Privacy Assessment,
and many more…
Reference: https://github.com/corca-ai/awesome-llm-security

This module will begin with a few weeks of teaching, then enter into a no-class mode for the project.

Course Logistics - Requirements

Students are required to write a report in the form of review/survey papers up to 8 pages to demonstrate their understanding of a specific research question, participate in peer review, and present their report at the end of the module.

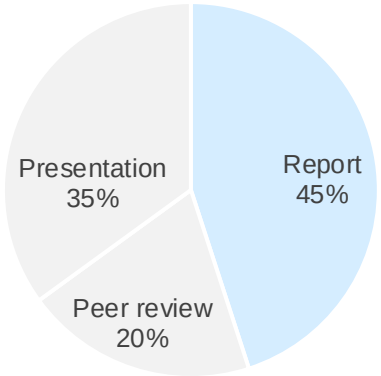
The project will be either individual or by a group of two, depending on the enrollment situation.

Continuous Assessment:
Report: to comprehensively read a research paper, gain an in-depth understanding of the research question, and learn scientific writing and typesetting (e.g., Latex).
Peer review: to criticize/appreciate from the scientific perspective, and to expand your knowledge about the other topics.
Presentation: to give an academic talk.

Course Logistics - Grading

Continuous commitment and paper submission (45%)

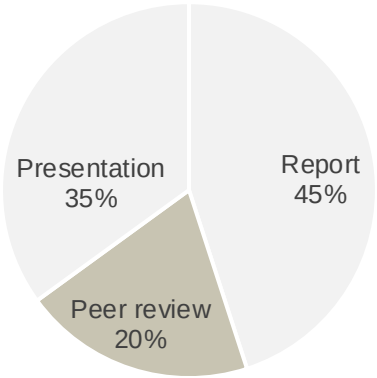
- This seminar requests students to put persistent effort to some extent to comprehensively understand the research question(s), although it is not designed as a research-intensive course and it does not aim to bring too much burden to the students.
- Grading criteria as listed below:
  - An in-depth understanding of the assigned topic and paper
  - Scientific writing skills demonstrated in the paper
  - (Last but not least) In compliance with the paper submission guidelines, your report must not exceed 8 pages (reference & appendix not counted) and must be in the given template.



Course Logistics - Grading

Peer review participation (20%)

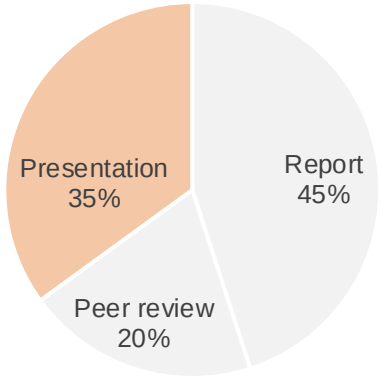
- As students would put most efforts into a specific area, this peer review process aims to involve all students in appreciating and criticizing their peers' submissions, and meanwhile, gain some exposure to diverse topics of LLM security.
- The grading of this part will be based on:
  - The quality of review comments (in terms of appreciation and critique)
  - Timeliness of the peer review contribution, i.e., your review comments must be submitted before the deadline (to be announced)



Course Logistics - Grading

Presentation in a mock conference (35%)

- The presentation at the end of the seminar aims to evaluate student's understanding of the chosen topic, and to train the students' skills in scientific communication. Students are requested to present a scientific research work to people without specific expertise.
- The grading of this part will be based on:
  - Presentation skills (speech, quality of slides, visualization, or other technique to support delivery, etc)
  - Q&A participation



Q&A