Course title: Retreival Augmented Generation

Lecturer: Kornai, András

Prerequisites: Some knowledge of machine learning, NLP, LLMs, python

Course description: Large Language Models (LLMs) produce grammatically, but not necessarily factually, perfect or near-perfect natural language (NL) output in response to NL prompts. In particular, they have a significant *hallucination* problem providing plausibly sounding, but quite incorrect answers to factually sounding questions, even when the correct answer is known to be present in their training data [arxiv 2309.05922,2402.02420, 2406.09155, etc]. Retrieval-augmented generation (RAG) has emerged as a popular method for mitigating the problem.

The course will look at RAG and other approaches to the general problem of bringing declarative (e.g. database) information to LLM-based question answering. We will have several guest lecturers from industry using RAG in real-life settings sharing their insights. The course will start on Week 2, and will also be accessible by Zoom. Besides registering on Neptun before Week 1, you need to send email to kornai@math.bme.hu with Subject: RAG to get the zoom link.

During the semester, you need to either (a) present a relevant paper or (b) build a baseline system and demonstrate how it is improved by RAG. Such projects can be the joint work of teams (maximum 4 participants/team), but presentations (in English) need to be done alone.

Manner of teaching: Lectures. No dedicated lab sessions.

Grading: Based on presentation and/or project work. No exam. Students outside BME/ELTE who cannot enroll in Neptun will be issued an attendance certificate upon request, but have to make their own arrangements for credit at their home institution.

Make up: The last lecture will be given over to project presentations. If substantially incomplete, they can still be completed in the exam period.

Consultation: As requested by the students.

Course materials: https://www.zs.com/insights/large-language-models-llm-and-hallucination-generative-ai
https://www.freecodecamp.org/news/retrieval-augmented-generation-rag-handbook/

Contact hours	28
Homework prep	10
Presentation/project	22