

# Project Topic Presentation

## IE686 Large Language Models and Agents



# General Information

- Teams of **5** students work together on an applied project to solve a problem with LLM agents
  - Teams write a 12-page report about their project and present their results during a presentation at the end of the semester
  - 3 ECTS (70% written project report, 30% presentation of results)
- How to find a topic?
  - We will propose a set of **topic areas** today
  - You prepare a **ranked list of your top three preferences** and send them to Ralph via mail
  - I will match you to the topic areas based on your preferences
  - If you already have a team/students you want to work with, please send **only one** mail as a (sub-)team

# Project Proposal

- We propose a set of topic areas that should be covered
- You define your project in one of these areas
- You are free in choosing which APIs/Benchmarks/Models you wish to use
- To make sure projects remain feasible in the allotted time, each team will prepare a project proposal outlining what they intend to do
- More information about the proposals next week!

# Course Schedule

Day	Topic
12.09	<b>Lecture:</b> Introduction to Language Models
19.09	<b>Lecture:</b> Instruction Tuning and RLHF
26.09	<b>Lecture:</b> Prompt Engineering and Efficient Adaptation
02.10	<b>Exercise:</b> Introduction to LangChain
10.10	<b>Lecture:</b> LLM Agents and Tool Use
17.10	<b>Exercise:</b> Introduction to LangGraph
24.10	<b>Project:</b> Introduction to Student Projects
31.10	<b>Project:</b> Project Coaching
07.11	<b>Project:</b> Project Coaching
14.11	<b>Project:</b> Project Coaching
21.11	<b>Project:</b> Project Coaching
28.11	<b>Project:</b> Project Coaching
05.12	<b>Project:</b> Presentation of Project Results

# Topic 1: Question Answering

- **Task:** Build a (Multi-)Agent QA system.
  - Create an agent-based environment for QA challenges
  - Make use of RAG, search engines, etc.
  - Identify problems with each approach
- **Dataset/APIs:**
  - Use existing benchmarks and relevant metrics
  - <https://www.aicrowd.com/challenges/meta-comprehensive-rag-benchmark-kdd-cup-2024>
- **Evaluation:**
  - Relevant evaluation metrics in used benchmarks
  - Compare different setups

# Topic 2: Multi-Agent Gaming

- **Task:** Build a Multi-Agent gaming application.
  - Have teams of agents cooperate to solve a game or play against each other
  - Seminar student in last semester tried playing Dungeons and Dragons with Agents with good success
  - Can be explorative or you use an existing agent benchmark
- **Dataset/APIs:**
  - <https://github.com/THUDM/AgentBench>
  - <https://github.com/microsoft/SmartPlay>
- **Evaluation:**
  - Existing evaluation when using agent benchmark
  - Otherwise depends on game, could be win rate, ELO ranking, ...

# Topic 3: Online Shopping Assistant

- **Task:** Build a (Multi-)Agent system that supports a user in making shopping decisions, e.g. for a new TV.
  - Search and present relevant products
  - Based on the users wishes
  - Present in structured format or directly perform the transaction
- **Dataset/APIs:**
  - Use existing benchmarks and relevant metrics
  - <https://github.com/THUDM/AgentBench>
  - <https://webarena.dev/>
  - <https://webshop-pnlp.github.io/>
- **Evaluation:**
  - Relevant evaluation metrics in used benchmarks



# Topic 4: Job Hunting Agents

- **Task:** Build a (Multi-)Agent application to search for job postings relevant to a user.
  - start with a user query
  - search for relevant postings
  - extract relevant facts about the jobs
  - present to user in a structured, easy-to-browse format
- **Dataset/APIs:**
  - select 2 job providers/APIs to work with
  - <https://apislist.com/category/28/jobs>
- **Evaluation:**
  - Explorative Topic, likely no relevant benchmarks exist
  - Human-based evaluation by the group members



# Topic 5: Browser-Interaction

- **Task:** Build a (Multi-)Agent system that can perform various tasks on websites.
  - Site search/link following/extraction of relevant info
  - Identify problems with each approach
- **Dataset/APIs:**
  - Use existing benchmarks and relevant metrics
  - <https://github.com/THUDM/AgentBench>
  - <https://webarena.dev/>
- **Evaluation:**
  - Relevant evaluation metrics in used benchmarks
- **Additional references:**
  - <https://arxiv.org/abs/2401.01614>
  - <https://arxiv.org/abs/2401.13919v4>
  - <https://openreview.net/pdf?id=9JQtrumvg8>

# Topic 6: Text to SQL

- **Task:** Build a (Multi-)Agent system that converts natural language queries to SQL.
  - Convert query
  - Query database
  - Refine based on result/errors
- **Dataset/APIs:**
  - <https://paperswithcode.com/task/text-to-sql>
  - <https://github.com/THUDM/AgentBench>
- **Evaluation:**
  - Relevant evaluation metrics in used benchmarks

# Topic 7: Test-first Developer

- **Task:** (Multi-)Agent system for software engineering. Start with set of seed tests written against a well-defined interface.
  - Generate code solutions based on tests
  - Run tests on solutions
  - If more than one solution remains, rank them by user preference
- **Dataset/APIs/Language:**
  - <https://evalplus.github.io/>
  - Python or Java
- **Evaluation:**
  - Relevant evaluation metrics in used benchmarks

# Topic 8: Code-driven Developer

- **Task:** (Multi-)Agent system for software engineering. Start with informal description of functionality.
  - Generate code solutions based on descriptions
  - Generate and run tests on solutions
  - Identify potential discrepancies in behavior of code solutions
  - Recommend final solution
- **Dataset/APIs/Language:**
  - <https://evalplus.github.io/>
  - Python or Java
- **Evaluation:**
  - Relevant evaluation metrics in used benchmarks

# Time to Choose!

- Decide for a **ranked list of three topic areas** by yourself or with your team if you already have someone to work with
- Send mail with your preferences to Ralph
  - until **Sunday 20<sup>th</sup> October (end-of-day)**
  - If you already have a team, please **only send a single mail** with your teams preferences
- I will assign you and present the teams in next weeks session
- More information about the project work, coachings, project proposal and final report in next weeks session!