

# Introduction to the Student Projects

## IE686 Large Language Models and Agents



# Student Projects

- **Goals**

- Gain practical experience with developing (multi-) agents / LLM workflows for a specific task
- Learn about problem domain
  - What domain-specific challenges exist and make the task hard?
  - Come up with ways to try to solve them with agents

- **Expectation**

- You explore your assigned topic area and its problems by
  1. implementing a (multi-) agent system or LLM workflow and
  2. evaluating it
- You solve the problem by
  - Building on the general methods we have seen in the course
  - Exploring additional methods that may be specific to the domain
  - Comparing multiple approaches using some form of evaluation

# Procedure

- Teams of **five** students
  - realize a project using LLM agents / LLM workflows
  - write a 12-page summary of the project and the methods employed in the project
  - present the project results to the other students
    - 12 minutes presentation + 8 minutes discussion
- Final mark for the course
  - 70 % written final report about the project
  - 30 % project presentation

# Project Examination Registration

- You need to formally register with the Prüfungsamt separate of your course registration **for the project IE686!**
- The general registration period starts on **08.04.2026** and continues until **22.04.2026**
- Put a reminder in your calendars to **register** as we can only grade you with a proper examination registration via Portal2!
- If you take both IE685 and IE686 this semester, **make sure to register for both courses separately!**

# Course IE 686 Schedule

Day	Topic
<b>24.03</b>	Introduction to Student Projects
<b>26.03</b>	<b>Deadline:</b> Submit Project Topic Preferences per Team
<b>27.03</b>	Assignment of Project Topics
<b>Wednesday, April 15<sup>th</sup>, 23:59 Submission of Project Outlines</b>	
<b>15.04</b>	Project Coaching
<b>22.04</b>	Project Coaching
<b>29.04</b>	Project Coaching
<b>06.05</b>	Project Coaching
<b>13.05</b>	Project Coaching
<b>Sunday, May 17<sup>th</sup>, 23:59 Submission of Project Report</b>	
<b>19.05</b>	Presentation of Project Results
<b>20.05</b>	Presentation of Project Results
<b>Wednesday, May 20<sup>th</sup>, 23:59 Submission of Presentation Slides</b>	

# Team Formation

- Team formation is already done based on your preferences
- You have been assigned to your team in ILIAS with your team number
- Now it is time to think about the topic area you want to work on!  
(see next slides)
- Meet with your team after the session to organize your work!
  - Reach out to team mates that may not be in the session today
  - Decide on topics you want to work on as a team
  - Make a ranked list of **3 topics**
  - Fill in [this Google Form](#) as a team to send us your preferences  
(**please send only 1 form per team!**)

# Topic 1: Multi-Agent Gaming

- **Task:** Build a Multi-Agent gaming application.
  - Have teams of agents cooperate to solve a game or play against humans or each other
  - Students in last semester tried playing Codenames with agents with good success
  - Can be explorative (you choose and implement the game) 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 2: Question Answering Agents

- **Task:** Build a RAG (Multi-)Agent QA system.
  - create an agent-based environment for QA challenges
  - that uses a web search engine to derive answers
  - system should be able to explain why it chose specific pieces evidence for it's answer
- **Dataset/APIs:**
  - Use existing benchmarks and relevant metrics
  - <https://openai.com/index/browsecomp/>
  - <https://www.aicrowd.com/challenges/meta-comprehensive-rag-benchmark-kdd-cup-2024>
- **Evaluation:**
  - Relevant evaluation metrics in used benchmarks
  - Compare different setups

# Topic 3: Job Hunting Agents

- **Task:** Build a (Multi-)agent application to search for job postings relevant to a user.
  - start with a user query and a short CV
  - search for relevant postings
  - extract relevant facts about the jobs
  - present results 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
  - LLM- and human-based evaluation by the group members

# Topic 4: 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 features of recommended products in structured format
- **Dataset/APIs:**
  - Explorative or 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
  - LLM- and human-based evaluation

# Topic 5: Personal Tutor

- **Task:** Build a (Multi-)Agent system that supports a user in studying for their major
  - helps explore and explain relevant concepts
  - identifies users knowledge gaps and gives targeted practice
  - tracks level of knowledge and improvement of user
- **Dataset/APIs:**
  - search for benchmarks which may exist
  - use your favorite Uni Mannheim courses!
- **Evaluation:**
  - LLM- and human-based evaluation

# Topic 6: Personal Assistant

- **Task:** Build a (multi-)agent system that supports a user in day-to-day tasks, e.g. manage
  - Calendar
  - Mails
  - Chat
  - ...
- **Dataset/APIs:**
  - Own task set or search for existing benchmarks
  - Use your personal data 🤨
- **Evaluation:**
  - LLM- and human-based evaluation

# Topic 7: How Do Different Personas Read the News?

- **Task:** Investigate how agents having different personas consume online news about current political topics
  - Which new sources do they choose? Where do they dig deeper?
  - how do the news influence the agent's memories and beliefs?
  - What differences emerge between the personas?
- **Dataset/APIs:**
  - persona definitions covering different political standpoints will be provided
  - the agents are restricted to a set of news outlets and their archives
- **Evaluation:**
  - LLM- and human-based comparative evaluation

# Topic 8: Wikipedia Article Generation

- **Task:** Build a RAG workflow / agent system for the automatic generation of Wikipedia articles
- **Dataset/APIs:**
  - WIKIGENBENCH
  - WikiAutoGen
- **Evaluation:**
  - LLM- and human-based evaluation using existing Wikipedia articles as ground truth
  - derive rubrics from articles

# Topic 9: Skill-Mining using Public Data

- **Task:** Build a RAG system that summarizes the project experience and skills of software developers given a public github profile
  - What are the skills of the developer?
  - Which roles does she take in her projects?
  - Does she take lead roles?
- **Dataset/APIs:**
  - github developer profiles
- **Evaluation:**
  - LLM- and human-based evaluation

# Topic X: Your Own Idea

- When submitting your project topic preferences you will also have the option to **suggest your own topic of choice**
- If we think your idea fits the course and the time allotted for the project, we are happy to coach your ideas
- Otherwise we will fall back to the preference list and assign one of our proposed topics

# Project Outlines

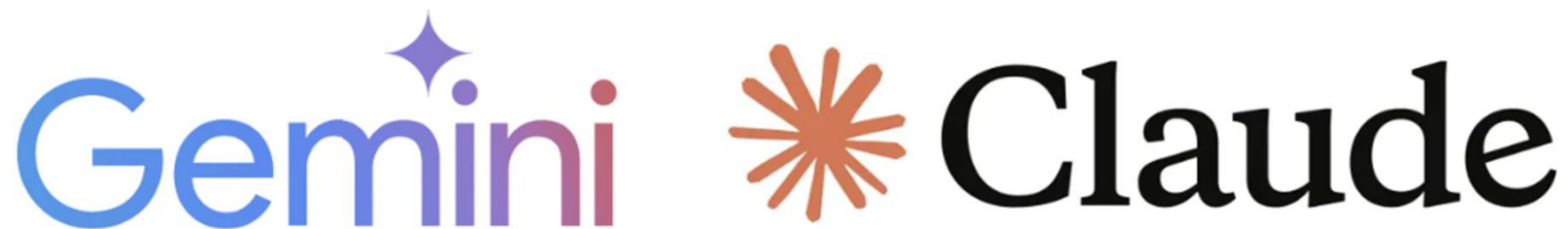
- Maximum 4 pages (sharp!) including title page
  - Using [DWS Seminar thesis layout](#) (PDF!)
  - Include a project name, your team number and name on the first page!
- Due **Wednesday, April 15th, 23:59**
- Upload to ILIAS submission (Every team member can submit, last submission will be evaluated)
- Feedback about your project outlines if required:  
Tuesday, 21.04.2026, (10:15-11:45)
  - We will inform you Monday, 20.04.2026 with some feedback via mail and let you know if you need to attend the session on Tuesday

# Project Outlines

- Answer the following questions:
  1. What is the problem you are solving?
  2. What data will you use?
    - Where will you get it?
    - How will you gather it?
  3. How will you solve the problem?
    - What LLMs do you plan to use?
    - Which methods do you plan to apply? Be as specific as you can!
    - What is your idea for a multi-agent workflow for your task?
  4. How will you measure success? (Evaluation method)

# How to Do Research for Your Outline

- Discuss your topic with an LLM of your choice.
  - Ask for existing methods / approaches
  - Ask for relevant APIs
  - Ask for existing benchmarks
  - Ask for alternative evaluation approaches



# How to Do Research for Your Outline

- Check out the solutions to your problem that other people have tried.
  - by looking at leaderboards of the relevant benchmarks
  - by investigating the state-of-the-art for your task
  - by searching for relevant scientific papers using Google Scholar
- Use them for inspiration and/or comparison



# How to Do Research for Your Outline

- Have the **proposal writing agent** from last week's exercise write a proposal for your topic
  - Compare your ideas and resources that you found to the ideas generated by the agent, merge good ideas from both
- Run your proposal through the **critique agents** to spot major shortcomings



OpenAI



Claude

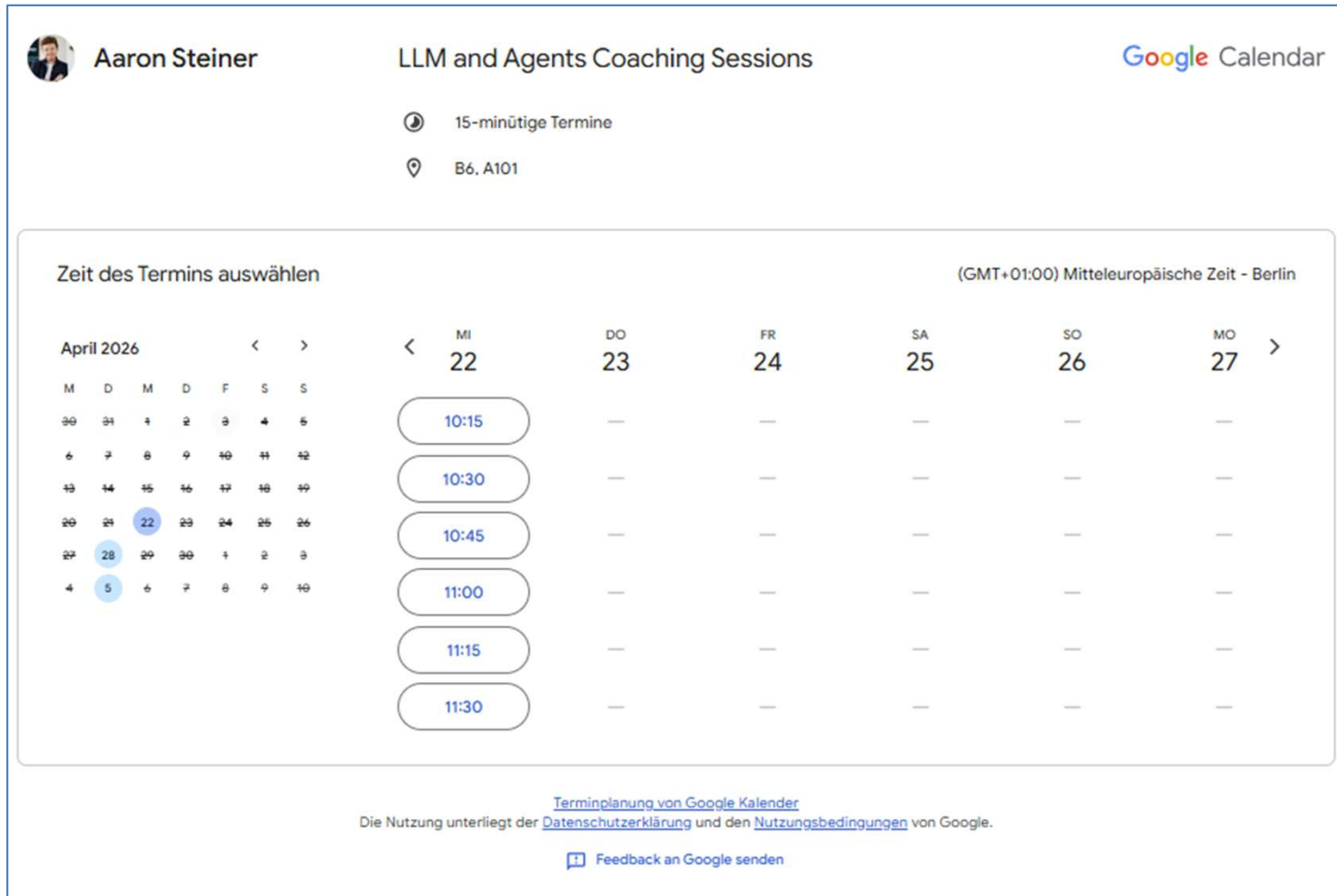


tavily

# Coaching Sessions

- We give you tips and answer questions about your project.
- At the time of the lecture (Tuesdays 10:15-11:45)
- **Registration** is mandatory if you want coaching!
- Make sure to register until Monday (23:59) of the week you want to attend the coaching session
- Each coaching session lasts for 15 minutes
  - Include your questions when booking the session, so we can prepare!
  - Most time efficient use of the session
  - We will of course also answer any question you pose directly in the session

# Booking Coaching Sessions: How-to



The screenshot shows the Google Calendar booking interface for Aaron Steiner. The title is "LLM and Agents Coaching Sessions". The duration is set to "15-minütige Termine" and the location is "B6, A101". The time selection interface is active, showing a calendar for April 2026 and a list of available times for Monday, April 22nd. The times listed are 10:15, 10:30, 10:45, 11:00, 11:15, and 11:30. The time 10:45 is currently selected. The interface also includes a "Feedback an Google senden" button and a link to Google's terms of service.

Aaron Steiner

LLM and Agents Coaching Sessions

Google Calendar

15-minütige Termine

B6, A101

Zeit des Termins auswählen (GMT+01:00) Mitteleuropäische Zeit - Berlin

MI	DO	FR	SA	SO	MO
22	23	24	25	26	27
10:15	—	—	—	—	—
10:30	—	—	—	—	—
10:45	—	—	—	—	—
11:00	—	—	—	—	—
11:15	—	—	—	—	—
11:30	—	—	—	—	—

[Terminplanung von Google Kalender](#)

Die Nutzung unterliegt der [Datenschutzerklärung](#) und den [Nutzungsbedingungen](#) von Google.

[Feedback an Google senden](#)

[https://calendar.google.com/calendar/u/0/appointments/schedules/AcZssZ2hONVF9I9A8zr4uM-jeC7FGgjOZxiKvu\\_Pf5wJW6AnmgspM3BMXvTWWDMxCncy4ocN\\_NtJSfvz](https://calendar.google.com/calendar/u/0/appointments/schedules/AcZssZ2hONVF9I9A8zr4uM-jeC7FGgjOZxiKvu_Pf5wJW6AnmgspM3BMXvTWWDMxCncy4ocN_NtJSfvz)

# Some Project Management Hints

- Organize your project in **multiple iterations**
  - every artefact will be improved over time!
- Get a **simple process running early** on to have a baseline
- **Parallelize tasks** while keeping centrally track of results
  - e.g. one central document with results plus reference to exact version of the notebooks/datasets that produced these results
  - sub-groups should explore specific ideas for a specified amount of time

# Some Project Management Hints

- **Define concrete milestones:** When should what be finished?
  - e.g. 28.04.26 Data exploration done and first simple baseline implemented
  - e.g. 30.04.26 Subgroup using decentralized agent communication adds results to central document
- **Infrastructure**
  - use shared folder for result document, versions of data, processes, slideset (e.g. MS Teams, Google Drive, GitHub)
  - use LLMs for inspiration about additional methods as well as for coding support

# Tasks within the Iterations of the Project

1. Data Exploration
2. Establish/update baseline (simple LLM agent(s) based method or non LLM baseline)
3. Try different methods for applying agents using different LLMs
  - Iteratively improve on methods...
  - solving problems as they come up
  - Track your experiments and all the things you did!
4. Perform **error analysis** in order to understand what is going on!
  - Stream the outputs, see whats going on and going wrong!
  - Analyse mistakes and reason about why system fails
  - Take a sample of errors, come up with meaningful error classes as humans, then sort the errors into the classes to get a quantitative result about frequency of errors!
  - Use your human sample as a starting point to automate the error analysis to a larger sample with LLMs!

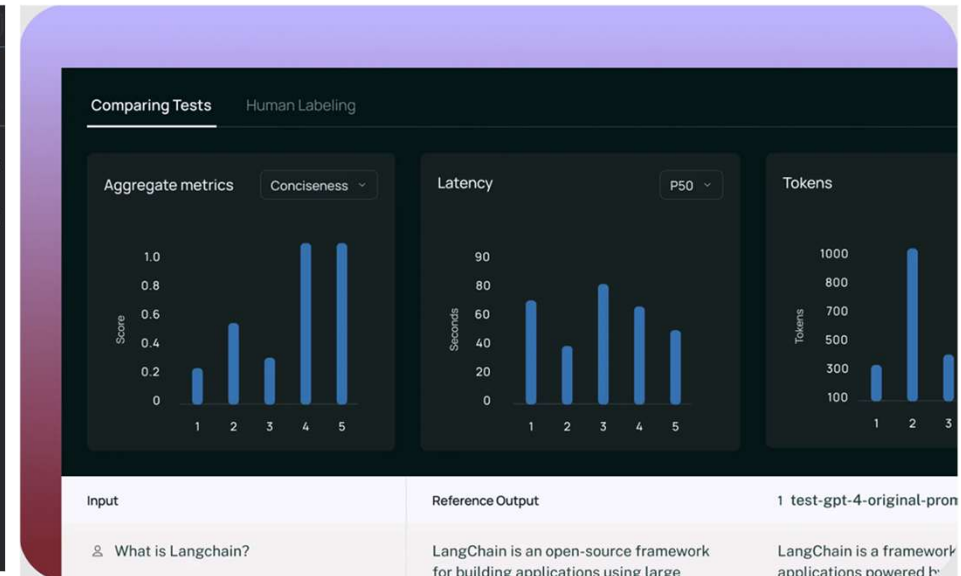
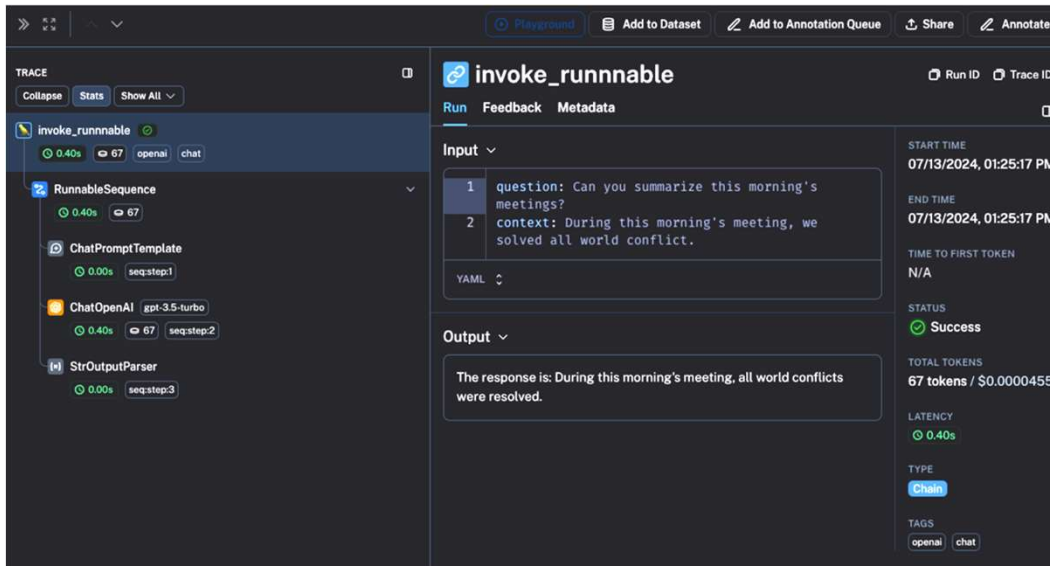
# Suggestions for APIs/Tools

- For finding relevant APIs to query: <https://apislist.com/>
- **Tavily** search tool: <https://tavily.com/>
  - 1000 free requests per month (\*5 team members = 😊 + free student tier)
  - Also possible to just use search engine
  - then find and extract relevant info from HTML
- **Unstructured** data transformation tool: <https://unstructured.io/>
  - Offer many open-source libraries to prepare data for LLMs you can use for free (<https://github.com/Unstructured-IO>)
- **Groq** inference API: <https://groq.com/>
  - Let's you query some powerful open-source models for free
  - May run into rate limit or unavailability due to traffic
- **nVidia** inference API: <https://build.nvidia.com/models>
- **Open Router**: <https://openrouter.ai/>
  - Allows you to switch between different LLMs

# OpenRouter

- This semester we can **fund your projects with up to 100\$** each!
- Each team will get an OpenRouter API key from us which you can freely use for your experiments
  - We would still advise you to start with a free API like Groq to develop a first basic working agentic system
  - Move on to a small and cheap model like gpt5-mini, gemini-flash, ... for the main experimental part
  - Once you have determined which parts may benefit from a powerful LLM, try using one (try to estimate costs first!)
- There will be **no refills!** Be careful when using your funds!

# Tracking your Experiments



- Print to console and save to logfiles, or...
- Use LangSmith
  - Commercial offering by LangChain
  - Let's you track and save what's happening during your workflow executions (traces)
  - Easy way to review what's happening and where things go wrong
  - 5000 free traces per month (\*5... 😊)

# Project Presentation

- Present the project results to the other students
  - 12 minutes presentation + 8 minutes discussion
  - Everyone must attend as its part of your grade
- Required content in slideset
  - Relevant prompts
  - Architecture overview
  - Result tables
  - Error analysis tables with error classes and frequency
- Upload your presentation in **PDF format**
  - Via your ILIAS Group
  - Until **Wednesday, May 20th, 23:59**

# Project Report

- Max. 12 pages including title/toc page and reference page
  - max. 10 pages content, no appendix
  - Each extra page and each day of late submission downgrades your mark by 0.3!
- Due **Sunday, May 17th, 23:59**
- Upload in **PDF format** via your ILIAS group

# Project Report

- Outline for project report:
  - Application area and goals (0.5 pages)
  - Profile (structure and size) of your data / tasks (minimum 1 page)
  - Approaches to solving the problem with LLM Agents / LLM Workflows
    - Describe different approaches that you tried
  - Evaluation
    - results
    - including problems that you faced with each approach and what steps you took to fix them
    - including an analysis of the errors still made (error classes and frequency), a discussion of the results, and a comparison to state-of-the-art results (together: minimum 2 pages)

# Project Report

- Requirements
  - You have to use the latex template for [DWS Seminar Theses](#)
  - please cite sources properly and use your references page
  - also submit your Python code and (a subset) of your data
  - include your names and your team number on the first page!
- Usage of additional AI Tools for non-method specific tasks needs to be declared in a separate table

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing	Throughout	+
DeepL	Translation	Throughout	+
ResearchGPT	Summarization of related work	Sec. 2.2	-
Dall-E	Image generation	Figs. 2, 3	++
GPT-4	Code generation	functions.py	+
ChatGPT	Related work hallucination	Most of bibliography	++

# Get Additional Advice from a Stanford Professor

- How to evaluate your model?
  - <https://www.youtube.com/watch?v=TxTbIROt9IY>
- How to structure your project report?
  - <https://www.youtube.com/watch?v=DZNwO-p5PGY>
- How to present the results of your project?
  - <https://www.youtube.com/watch?v=GGx7klcahZy>



**Christopher Potts**

# Deadlines - Overview

- Project outline until **Wednesday, April 15th, 23:59**
- Coaching Sessions
  - Every week
  - Registration for coaching must be done by Monday 23:59 of that week the latest!
- Project report as PDF until **Sunday, May 17th, 23:59**
- Project presentation as PDF until **Wednesday, May 20th, 23:59**

Thank you

