Introduction to the Student Projects

IE686 Large Language Models and Agents

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Student Projects

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Goals**
  - Gain practical experience with developing multi-agent LLM applications for a specific problem
  - Learn about problem domain
    - What domain-specific problems 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 with data/benchmarks/LLMs of your own choosing
  - 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Teams of four to five students
  - realize a project using LLMs as agents
  - 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 + 3 minutes discussion

- Final mark for the course
  - 70 % written final report about the project
  - 30 % project presentation

Exam Registration

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- There is no “real” exam at the end of this course
- You still need to formally register with the Prüfungsamt separate of your course registration
- The registration period started on **23.10.2024** and continues until **06.11.2024**
- Make sure to **register soon**, so you don’t forget!

Course Schedule

UNIVERSITY OF MANNHEIM
Data and Web Science Group

You are here

Day	Topic
24.10	**Project:** Introduction to Student Projects
Monday, October 28th, 23:59 Submission of Project Outlines	
31.10	**Project:** Project Coaching
07.11	**Project:** Project Coaching
14.11	**Project:** Project Coaching
21.11	**Project:** Project Coaching
28.11	**Project:** Project Coaching
Sunday, December 1st, 23:59 Submission of Project Report	
05.12	**Project:** Presentation of Project Results
Thursday, December 5th, 23:59 Submission of Presentation Slides	

Team Formation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Team formation is already done based on your preferences
  - You should have received a mail with your team number and topic area
  - It was difficult to satisfy everyone's wishes at once, so receiving the top priority topic area was often not possible
  - I had to get a bit creative but nearly everyone has gotten one of their preferences

- Now it is time to define your own project in your assigned area!

- Meet with your team after the session to organize your work!
  - Decide project specifics
  - Organize writing of project outline

Topic 1: Question Answering

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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: Text to SQL

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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 6: Code-driven Developer

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **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

Where to find Datasets/Benchmarks?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Agent Benchmarks
  - AgentBench
  - WebShop
  - WebArena
  - Meta RAG KDD Cup
  - SmartPlay
  - EvalPlus
- General Task Benchmarks
  - Papers with Code Datasets (filter by task)
  - Huggingface Datasets

Even More Benchmarks

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Method	Evaluation	Model Types	Abilities/Domain	Data Source
Benchmark	MMLU [364]	Base/Fine-tuned/Specialized	General	Human exam/practice
	BIG-bench [70]	Base/Fine-tuned/Specialized	General	Human annotation
	HELM [520]	Base/Fine-tuned/Specialized	General	Benchmark collection
	Open LLM Leaderboard [707]	Base/Fine-tuned/Specialized	General	Benchmark collection
	AGIEval [708]	Base/Fine-tuned/Specialized	General	Human exam/practice
	MMCU [709]	Base/Fine-tuned/Specialized	General	Human exam/practice
	M3KE [710]	Base/Fine-tuned/Specialized	General	Human exam/practice
	C-Eval [711]	Base/Fine-tuned/Specialized	General	Human exam/practice
	Xiezhi [712]	Base/Fine-tuned/Specialized	General	Human exam/practice
	OpenCompass [713]	Base/Fine-tuned/Specialized	General	Benchmark collection
	Chain-of-Thought Hub [714]	Base/Fine-tuned	General	Benchmark collection
	KoLA [715]	Base/Fine-tuned	Knowledge utilization	Web
	ARB [716]	Fine-tuned	Complex reasoning	Human exam/practice
	APIBench [717]	Base/Fine-tuned	Tool manipulation	Web
	APIBank [718]	Fine-tuned	Tool manipulation	Synthesis
	ToolAlpaca [719]	Base/Fine-tuned	Tool manipulation	Synthesis
	T-Bench [720]	Fine-tuned	Tool manipulation	Synthesis
	ToolBench [721]	Fine-tuned	Tool manipulation	Synthesis
	BOLAA [722]	Base/Fine-tuned	Environment interaction	Benchmark collection
	AgentBench [723]	Base/Fine-tuned	Environment interaction	Human annotation/Synthesis
	HaluEval [602]	Base/Fine-tuned	Human alignment	Human annotation/Synthesis
	PromptBench [724]	Base/Fine-tuned	Robustness	Benchmark collection
	HumanEval [105]	Base/Fine-tuned/Specialized	Code synthesis	Human annotation
	MultiMedQA [356]	Specialized	Healthcare	Benchmark collection
	FLUE [725]	Specialized	Finance	Benchmark collection
	LegalBench [726]	Specialized	Legal	Human annotation
Human	Chatbot Arena [727]	Base/Fine-tuned/Specialized	Human Alignment	Human annotation
	SciBench [728]	Fine-tuned	Complex reasoning	Human exam/practice
Model	AlpacaEval [729]	Fine-tuned	Instruction following	Synthesis
	MT-bench [727]	Fine-tuned	Human alignment	Human annotation
	TrustGPT [730]	Base/Fine-tuned	Human alignment	Benchmark collection
	LMExamQA [731]	Base/Fine-tuned	Knowledge utilization	Synthesis
	ChatEval [732]	Base/Fine-tuned	Knowledge utilization	Benchmark collection

Zhao et al.: A Survey of Large Language Models. 2024. arXiv:2303.18223

Where to Find Additional Information

UNIVERSITY OF MANNHEIM
Data and Web Science Group

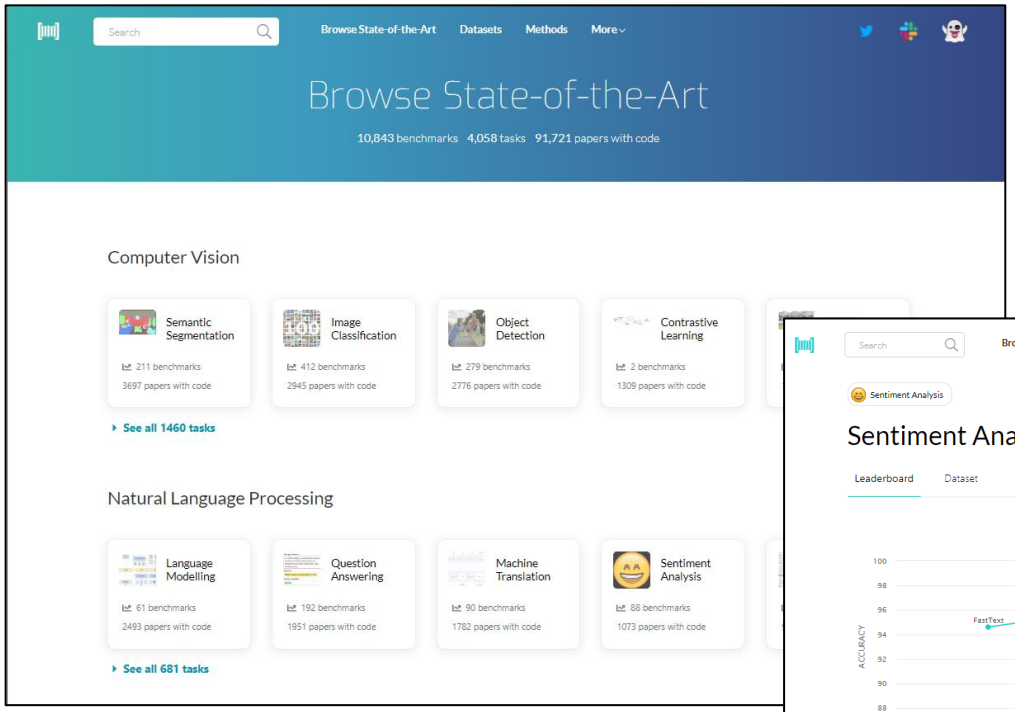
- Check out the LLMs/solutions to your problem that other people have tried.
  - by looking at leaderboards of the relevant benchmarks
  - by investigating the state-of-the-art for your your task on Papers with Code
  - or search for relevant scientific papers using Google Scholar
- Use them for inspiration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

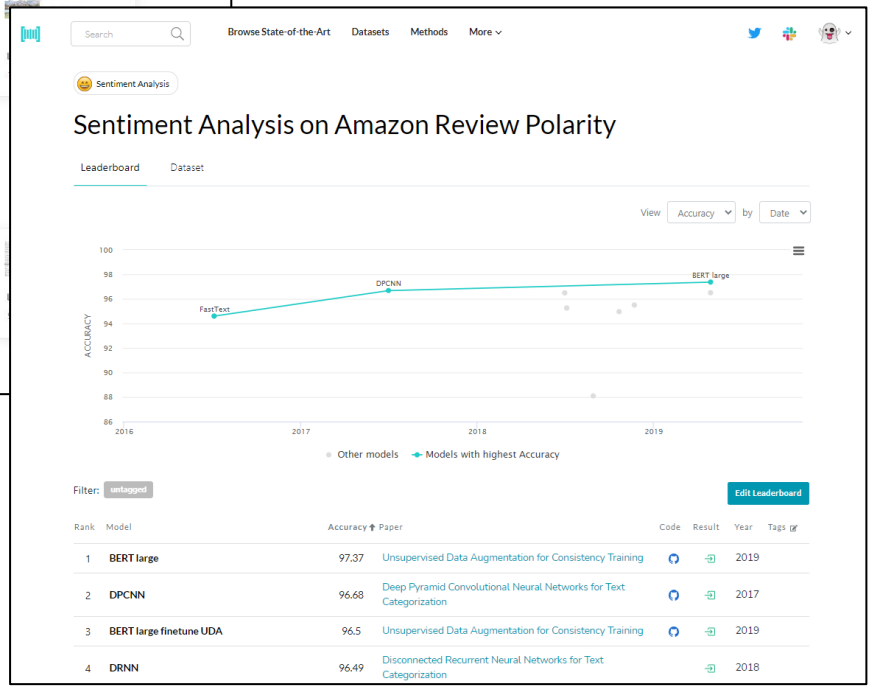
Google

Papers With Code

State-of-the-Art for Specific Tasks

UNIVERSITY OF MANNHEIM
Data and Web Science Group





https://paperswithcode.com/sota

Suggestions for APIs/Tools

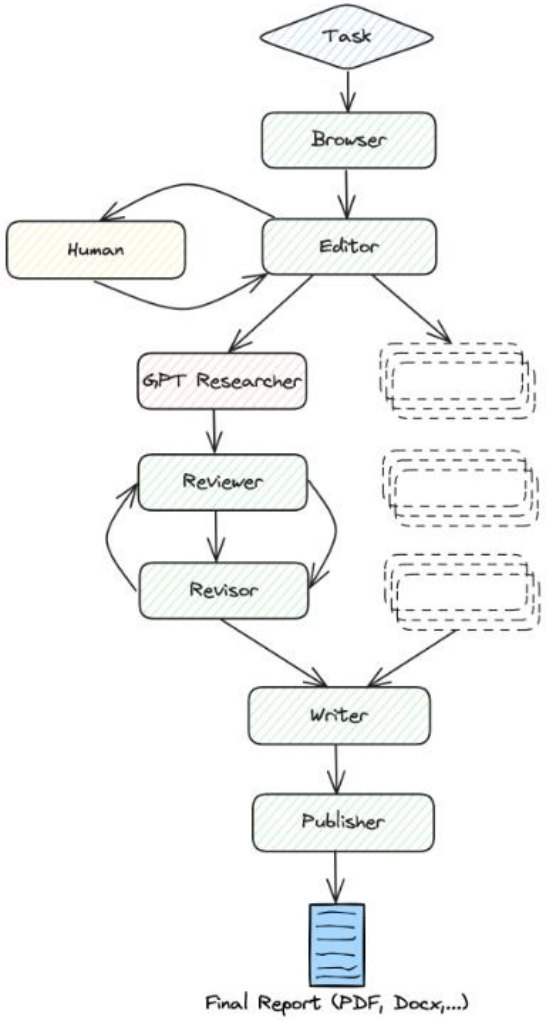
UNIVERSITY OF MANNHEIM
Data and Web Science Group

- For finding relevant APIs to query: https://apislist.com/
- Tavily search tool: https://tavily.com/
  - 1000 free requests per month (*5 team members = 😊)
  - 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 LLM inference API: https://groq.com/
  - Let's you query some powerful open-source models for free (Llama3.1-70B)
  - May run into rate limit or unavailability due to traffic
- OpenAI: gpt4o-mini (or similar Google/Anthropic/Mistral offering)
  - Quite powerful and successfully used in team project for agentic workflow
  - Comparably cheap

GPT-Researcher

UNIVERSITY OF MANNHEIM
Data and Web Science Group

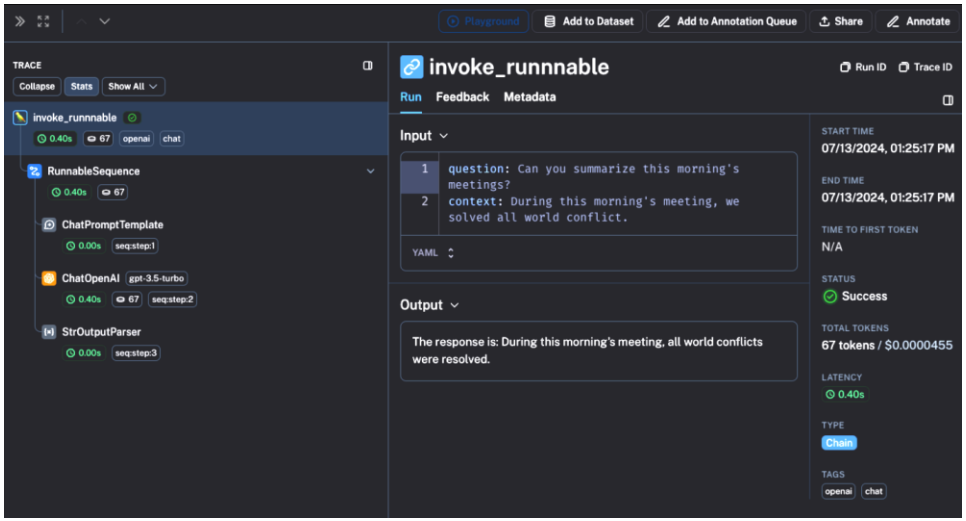
- Multi-agent system for online research
- Uses “Plan-and-Solve” prompting to divide task into subtasks…
- Which are carried out by multiple agents in parallel using web crawling as a tool.
- Written in LangGraph
- Can be used as inspiration

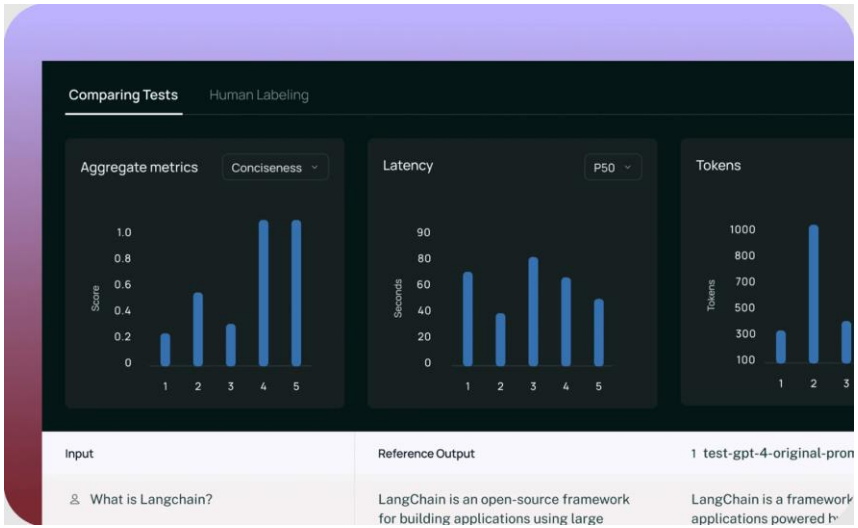


https://docs.gptr.dev/blog/building-gpt-researcher
https://github.com/assafelovic/gpt-researcher/tree/master/multi_agents

Tracking your Experiments

UNIVERSITY OF MANNHEIM
Data and Web Science Group





- 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 Outlines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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 **Monday, October 28th, 23:59**
- Send by eMail to Ralph

- Feedback about your project outlines if required:
  Thursday, 31.10.2024, lecture time (15:15-17:00)
  - I will inform you Wednesday, 30.10.2024 with some feedback via mail and let you know if you need to attend the session on Thursday, 31.10.2024
  - Then the projects start!

Project Outlines

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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)

Coaching Sessions

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- I will give you tips and answer questions concerning your project.
- At the time of the lecture (Thursdays 15:15-17:00)
- **Registration via email** is mandatory if you want coaching!
  - Via mail to Ralph
  - Until Monday night (23:59) of the respective week!
  - **Including the questions that you would like to discuss (be as clear and specific as possible so I can prepare)**
- I will assign you a time slot afterwards and inform you about the slot via email

Some Project Management Hints

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- **Define concrete milestones**: When should what be finished?
  - e.g. 08.11.24 Data exploration done and first simple baseline implemented
  - e.g. 15.11.24 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

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

Project Presentation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Present the project results to the other students
  - 12 minutes presentation + 3 minutes discussion
  - During lecture slot
  - Everyone must attend as its part of your grade
- Send your presentation in **PDF format**
  - Via mail to Ralph and Prof. Bizer
  - Until **Thursday, December 5th, 23:59**

Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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, December 1st, 23:59
- send by email to Ralph and Prof. Bizer

Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Outline for project report:
  - Application area and goals (0.5 pages)
  - Profile (structure and size) of your data (minimum 1 page)
  - Approaches to solving the problem with LLM Agents
    - 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, a discussion of the results, and a comparison to state-of-the-art results (together: minimum 2 pages)

Project Report

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- 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

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Christopher Potts

- How to evaluate your model?
  - https://www.youtube.com/watch?v=TxTblROT9lY
- 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

Deadlines - Overview

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Project outline until **Monday, October 28th, 23:59**
- Coaching Sessions
  - Every week
  - Registration for coaching must be done by Monday 23:59 of that week the latest!
- Project report until **Sunday, December 1st, 23:59**
- Project presentation as PDF until **Thursday, December 5th, 23:59**

Questions?

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Thank you

UNIVERSITY OF MANNHEIM
Data and Web Science Group