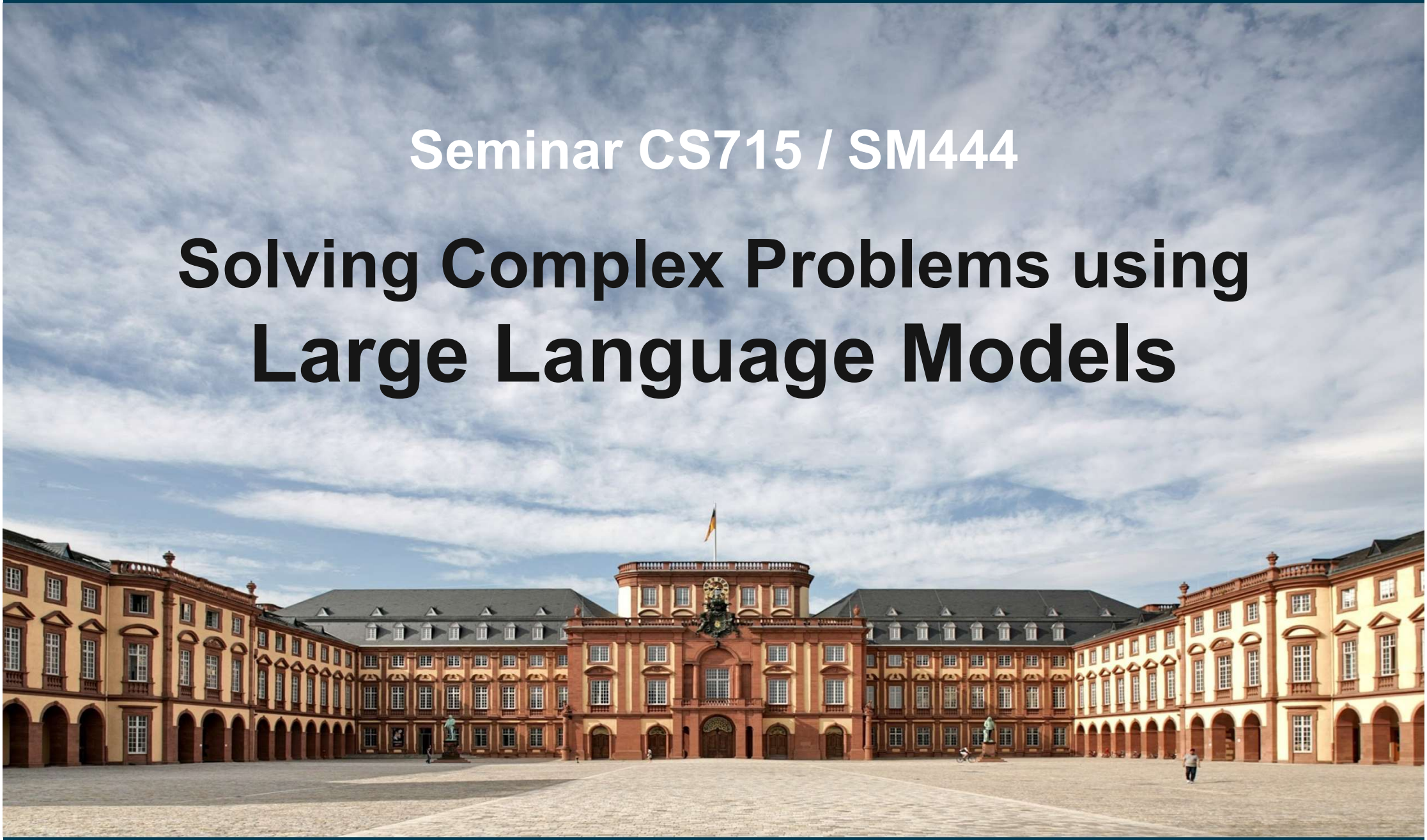
UNIVERSITY OF MANNHEIM

Seminar CS715 / SM444

Solving Complex Problems using Large Language Models



Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web-based Systems
  - Web Data Integration
  - Deployment of Data Web Technologies
- Room: B6 - B1.15
- eMail: christian.bizer@uni-mannheim.de

Hello

- **M.Sc. Wi-Inf Ralph Peeters**
- Graduate Research Associate
- Research Interests
  - Sustainable LLM-Agents
  - Entity Matching using Deep Learning
  - Data Integration
- Office: B6, 26 - C 1.04
- eMail: ralph.peeters@uni-mannheim.de

Hello

- **M.Sc. Aaron Steiner**
  - Graduate Research Associate
  - Research Interests
    - Entity Matching using LLMs
    - Data Integration
  - Office: B6, 26 - C 1.04
  - eMail: aaron.steiner@uni-mannheim.de

You and Your Experience

- A Short Round of Introductions
  - What are you studying?
  - Which DWS courses did you attend?
  - What kind of experience do you have with
    - Large Language Models (LLMs) and
    - LLM-based agents or workflows?
- Participants
  - Arnold, Carina
  - Ebner, Raphael
  - Elagin, Ksenia
  - Emsbach, Anna
  - Knüttel, Fernando
  - Minghao Lei
  - Nguyen, Anh-Nhat
  - Raheel, Esha
  - Schönweiß, Lennart
  - Schwarz, Luca
  - Wackermann, Cora

Agenda of Today‘s Kickoff Meeting

1. Seminar organization
2. Introduction to LLM-based Agents
3. Topic Assignment
4. How to structure your paper / presentation?
5. Your Questions

1. Seminar Organization

Learning Goals

- Writing a seminar thesis as an exercise for your master thesis
- Searching and citing scientific papers / journal articles
- Understanding and presenting state-of-the-art scientific literature
- Designing experiments and present experimental results
- How to structure your thesis and presentation
- How to write a scientific paper using LaTeX

Schedule

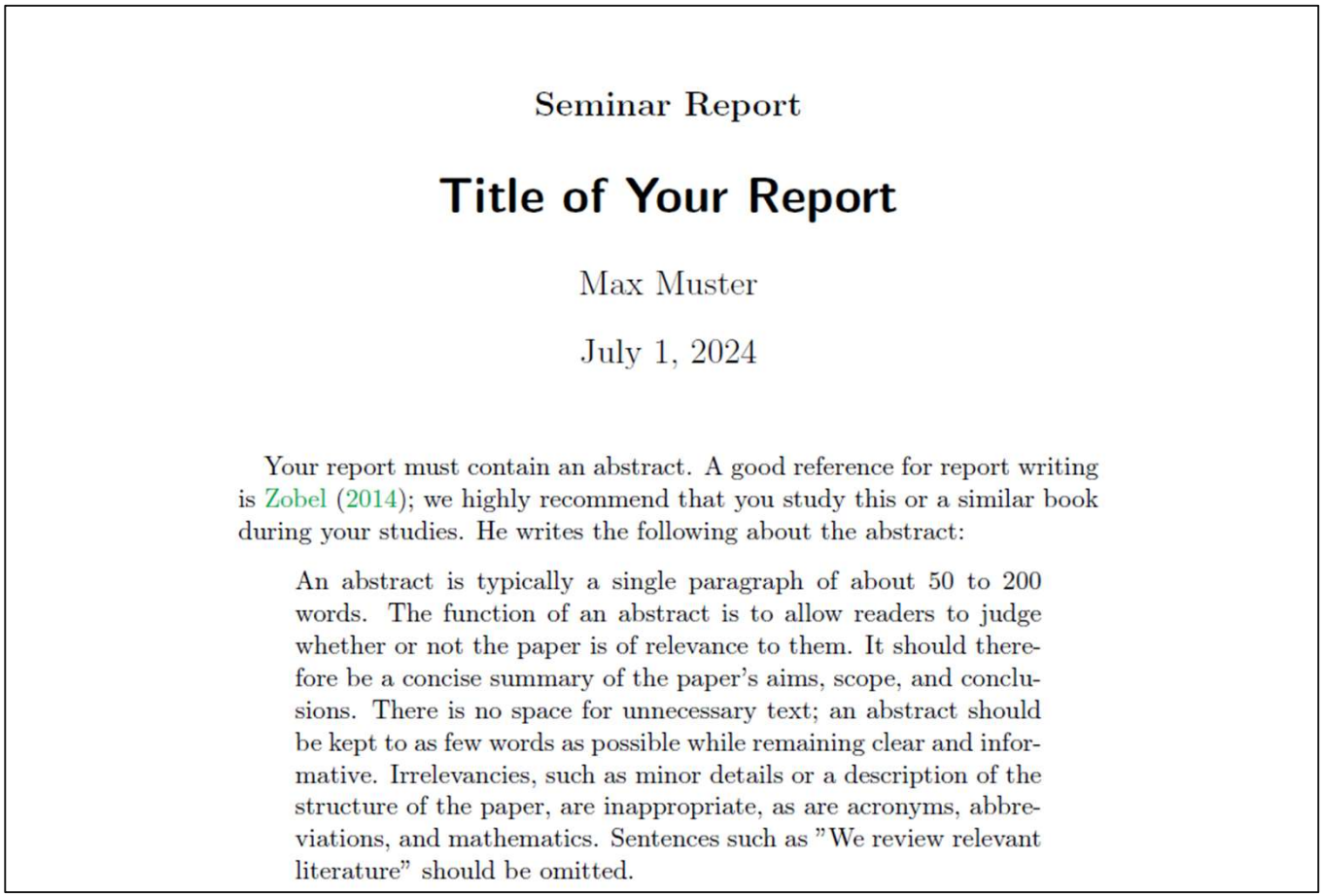
Date	Session
Tuesday, 25.02.2025 (15:30-17:00)	Kick-off meeting and topic/mentor assignment
	Read papers about your topic Search for additional literature Design experimental setup Select methods/design experiments, prepare presentation outline
Until 20.03.2025	Meet with your mentor to discuss outline and/or experimental setup
	Prepare draft of your presentation
Until 15.4.2025	Send draft presentation to your mentor
	Finalize your presentation
Monday, 28.04.2025 (10:00-12:00) (14:00-16:00)	Presentation and discussion of your topic (30 % of your final grade)
	Write seminar thesis
Friday, 04.07.2025	Submission of your seminar thesis (70 % of your final grade)

Formal Requirements

- Presentation
  - 10 minutes + 7 minutes discussion
  - should be 100% understandable for all participants
- Written report (paper)
  - 12-15 pages single column
    - including abstract and appendixes
    - not including bibliography
    - not including the page about LLM usage
    - every additional page reduces your grade by 0.3
  - written in English
  - use seminar report template from DWS templates
- Final grade
  - 70% written report
  - 30% presentation

Which template to use?

DWS Seminar Report Template



Seminar Report

Title of Your Report

Max Muster

July 1, 2024

Your report must contain an abstract. A good reference for report writing is Zobel (2014); we highly recommend that you study this or a similar book during your studies. He writes the following about the abstract:

> An abstract is typically a single paragraph of about 50 to 200 words. The function of an abstract is to allow readers to judge whether or not the paper is of relevance to them. It should therefore be a concise summary of the paper's aims, scope, and conclusions. There is no space for unnecessary text; an abstract should be kept to as few words as possible while remaining clear and informative. Irrelevancies, such as minor details or a description of the structure of the paper, are inappropriate, as are acronyms, abbreviations, and mathematics. Sentences such as "We review relevant literature" should be omitted.

https://www.uni-mannheim.de/media/Einrichtungen/dws/Files_Teaching/Theses/dws-templates.zip

Statement About the Tools that You Used

Your report **must include an extra page** about

1. which generative AI tools you used
   - ChatGPT, OpenAI API, DeepL, Deep Researcher, StudyTexter, ….
2. for which purposes
   - structuring your paper
   - summarizing related work
   - writing text for specific chapters
   - improving English grammar and formulations
   - designing experimental setup
   - writing code
   - writing prompts
   - generating training data
   - summarize log files
   - perform error analysis
   - ….
3. How useful was each tool for this?

https://www.uni-mannheim.de/infos-fuer/forschende-und-lehrende/lehren/ihre-lehre-im-fokus/

Example of AI Tools Declaration (Part of DWS Template)

Ehrenwörtliche Erklärung

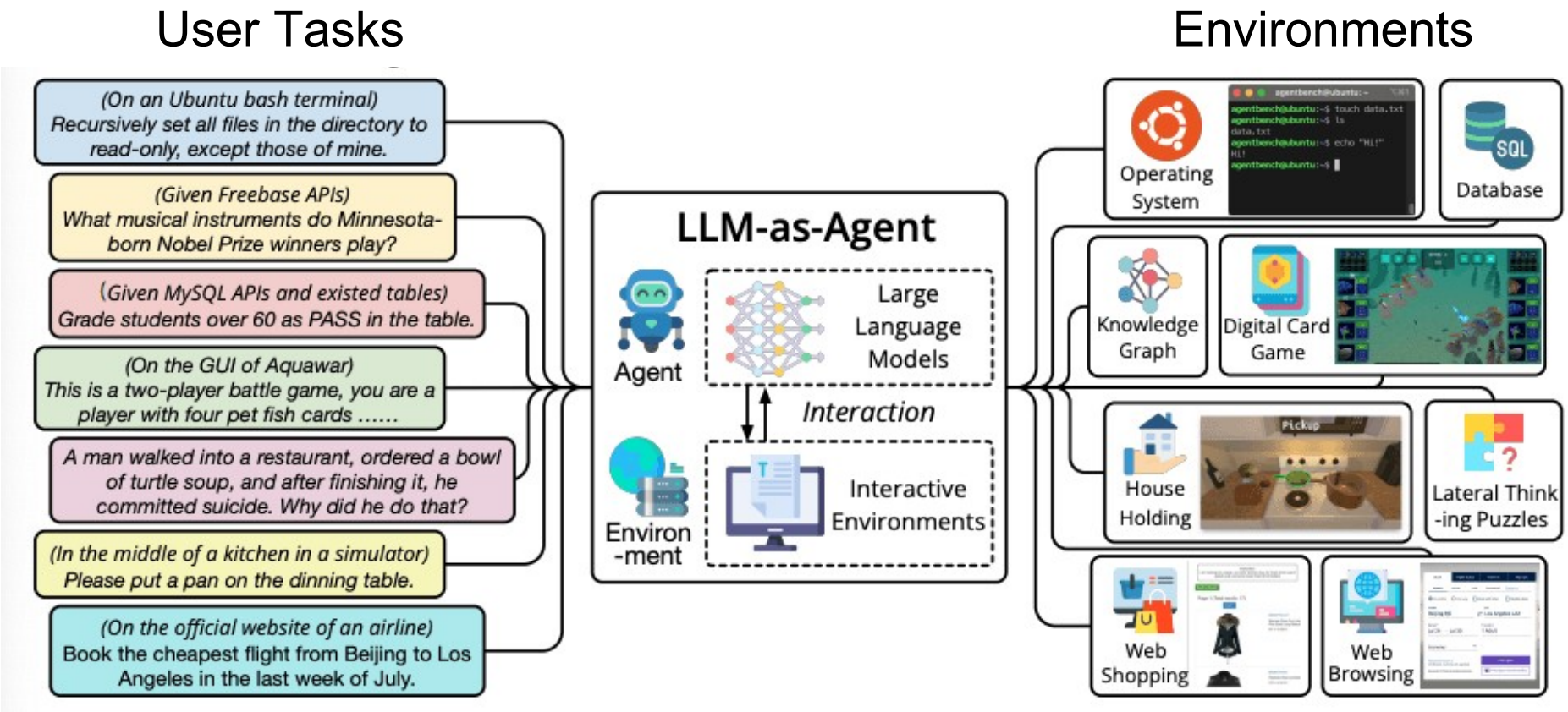
Ich versichere, dass ich die beiliegende Bachelor-, Master-, Seminar-, oder Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

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

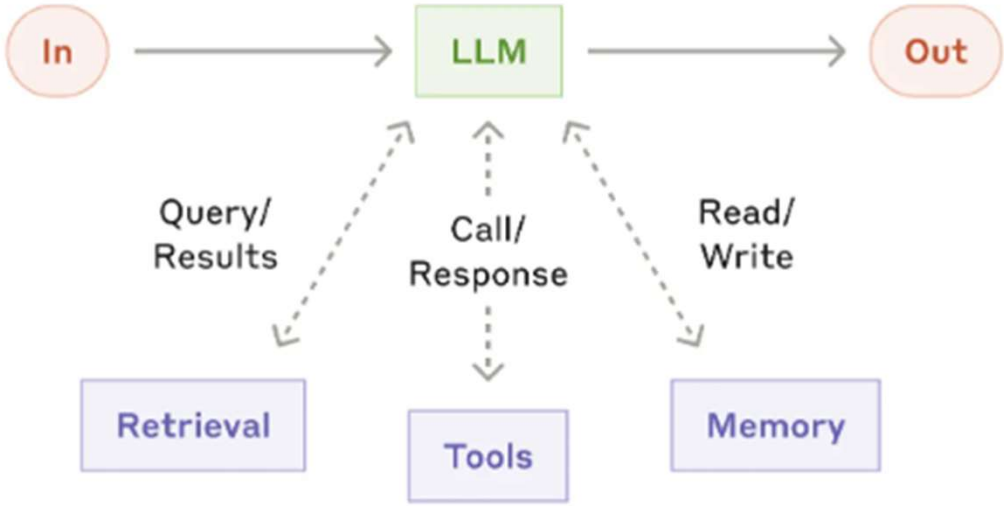
2. Introduction to LLM-based Agents

LLM-based agents autonomously interact with an environment to solve user tasks.



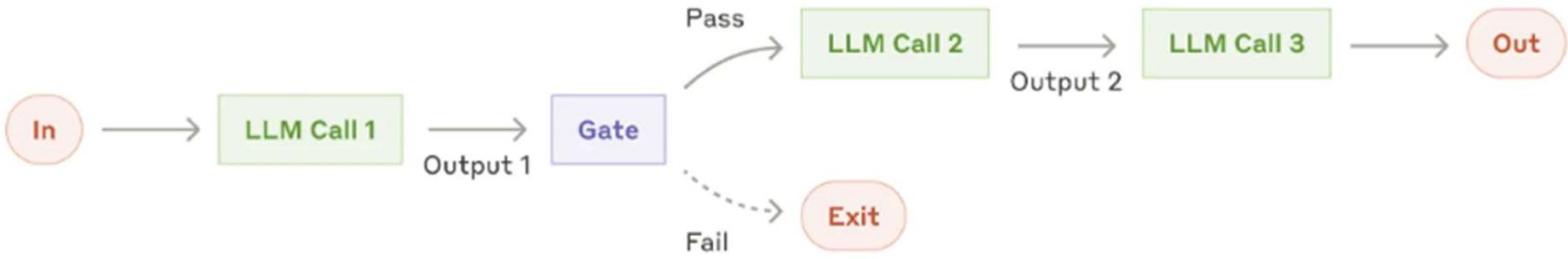
Tool Usage

- LLM-based applications increasingly **use external tools**
  - search the Web or document repositories
  - query databases and ERP systems
  - running code in execution environment
  - move the mouse or type using the keyboards
  - memorizing results of previous interactions



LLM Workflows

- LLM-based applications increasingly **perform workflows**
- prompt chaining workflow:



- example where prompt chaining is useful:
  - generating marketing text, then translating it into a different language
- the workflows are hard-coded
  - using frameworks like LangGraph

https://www.anthropic.com/research/building-effective-agents

Workflow: Orchestrator-Workers

- examples where orchestrator-workers is useful:
  - search tasks that involve gathering and analyzing information from multiple sources for possible relevant information, e.g. GPT Researcher
  - coding tasks that make complex changes to multiple files
- still hard-coded workflow

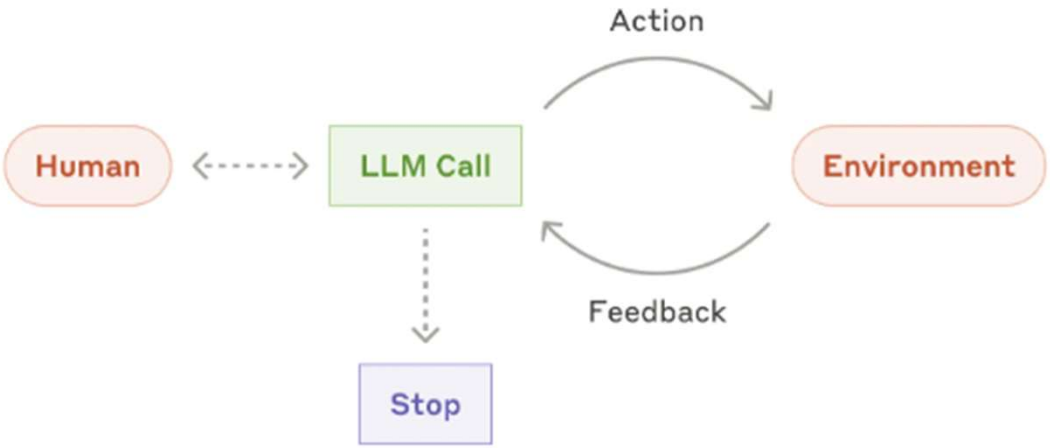
Workflow: Evaluator-Optimizer

- examples where evaluator-optimizer is useful:
  - writing code
  - text-to-SQL translation
- still hard-coded workflow

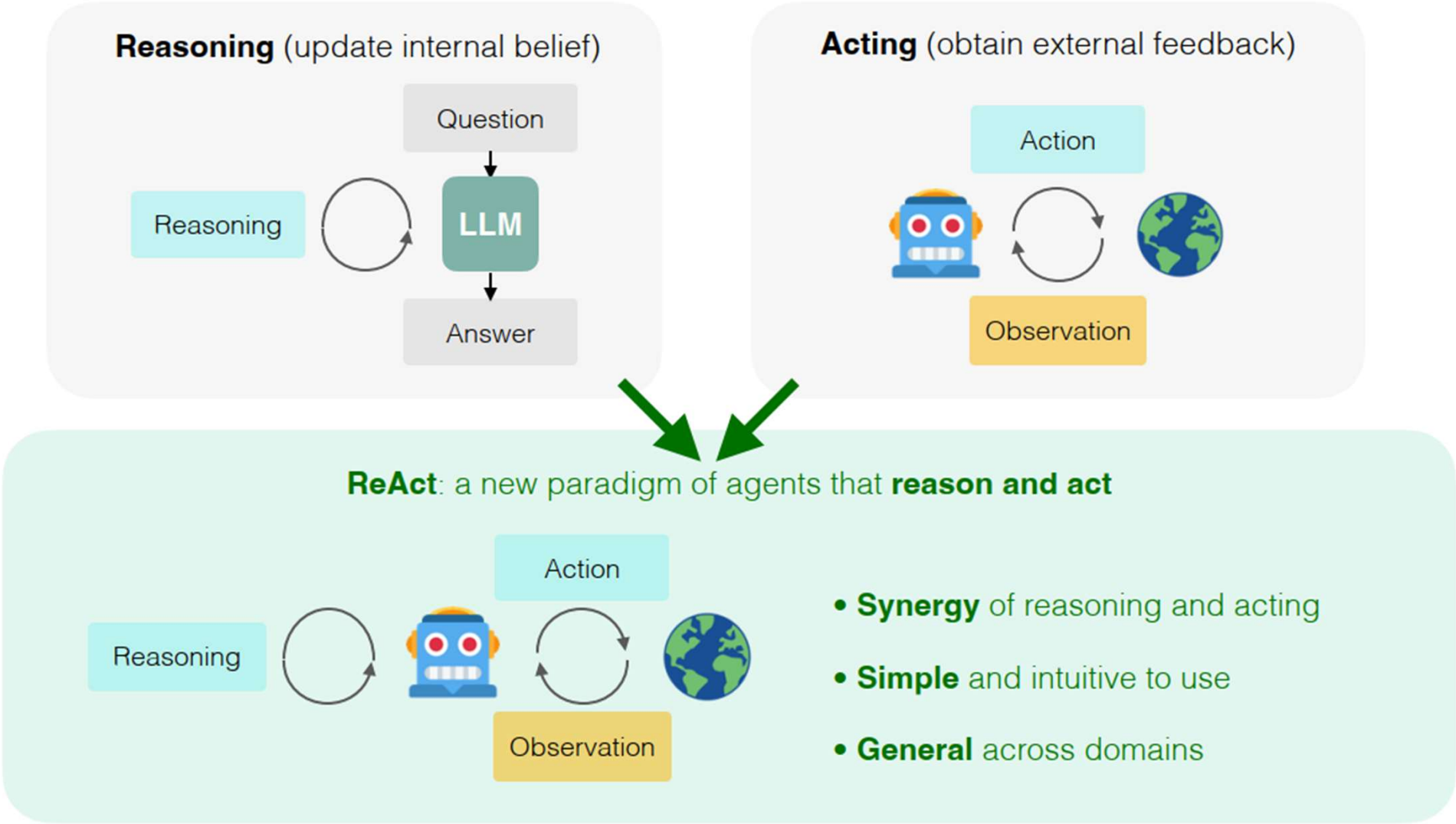
Pan, et al.: Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies, 2024

LLM-based Agents

- given a task, agents plan and operate autonomously
  - no fixed workflow, but flexible series of actions planned by agent
- agent consists of a loop iterating between LLM and action calls
- in each iteration
  1. the agent observes the environment
  2. reasons about the task given the environment and its previous actions
  3. choose an action to perform
  4. the agent's runtime environment executes the action

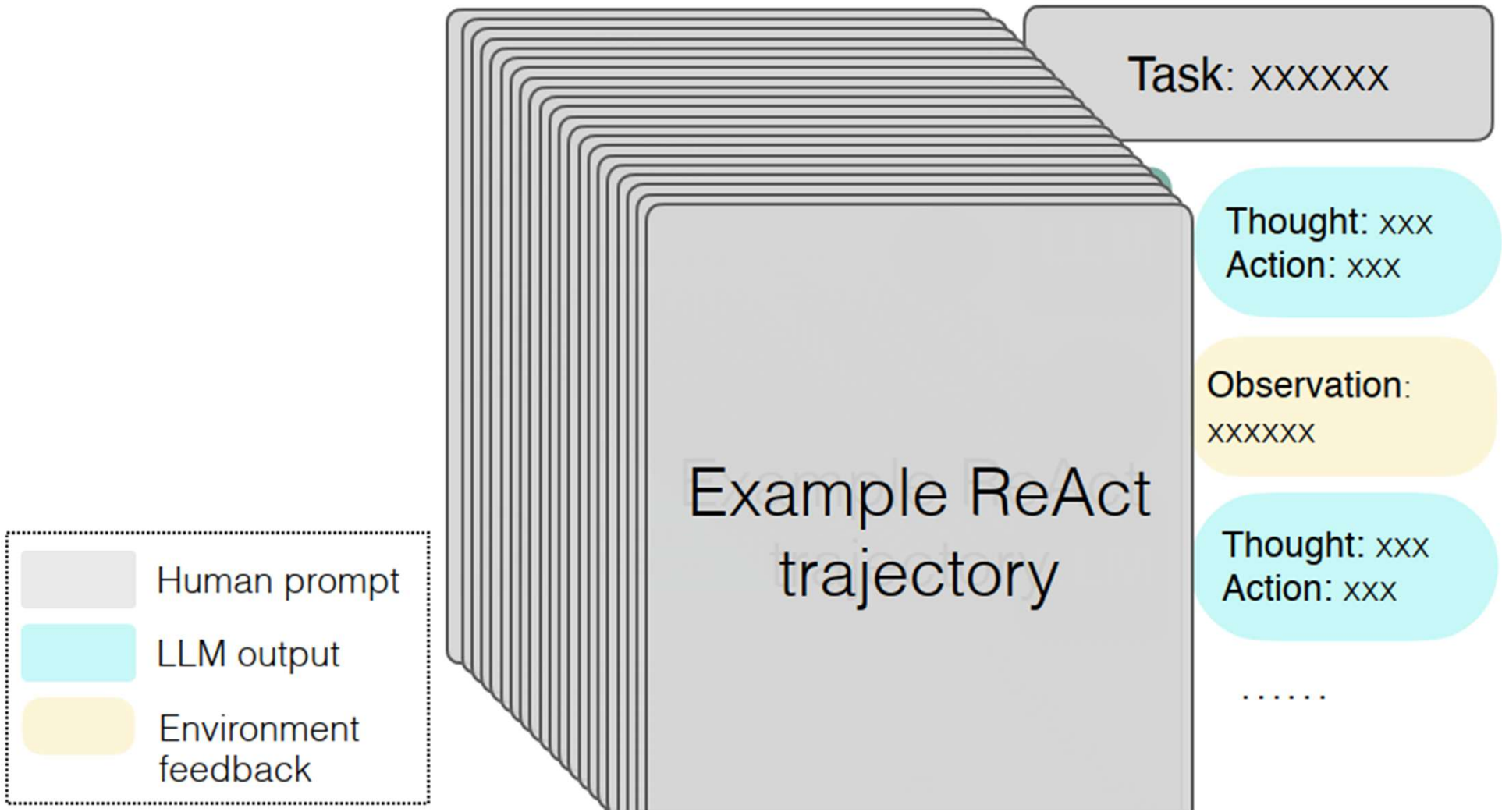


The ReAct Paradigm



Yao, S., et al., 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*.

ReAct is Simple and Intuitive to Use



Example ReAct Prompt

System prompt defines agent

You are an agent that answers questions by using two actions:
- search[query]: Google search the query. You can also use it to calculate math.
- finish[answer]: return the answer

Your generation should have the following format:
Thought: to analyze the observation and inform the action.
Action: your action

User task

Question: If I have seven trillion dollars today, can I buy Apple, Nvidia, and Microsoft? If not, how much more money do I need?

ReAct Trajectory

LLM

(GPT-4)

Thought: I need to find the current market capitalization of Apple, Nvidia, and Microsoft to understand how much it would cost to buy all of them.
Action: search["current market capitalization of Apple, Nvidia, and Microsoft"]

Observation: "Nvidia's market cap stood at $1.832 trillion as of about 10 a.m. ET Wednesday, while Alphabet's was at $1.814 trillion, bringing Nvidia behind just Microsoft (MSFT) at $3.015 trillion and Apple (AAPL) at $2.836 trillion."

(Mar 2024)

Thought: Now that I have the market cap for all companies, let's add them together and see if $7 trillion is enough.
Action: search["1.832 trillion + 3.015 trillion + 2.836 trillion"]

Observation: 7683000000000

Thought: The amount of money needed in addition to the $7 trillion is $683 billion.
Action: finish["You need an additional $683 billion to be able to buy Apple, Nvidia, and Microsoft."]

Observation Space and Action Space

Action

- call external **APIs** for extra information that is missing from the model weights (often hard to change after pre-training): **Generating multimodal outputs; Embodied Action; Learning tools; Using tools; Making tools; .....**

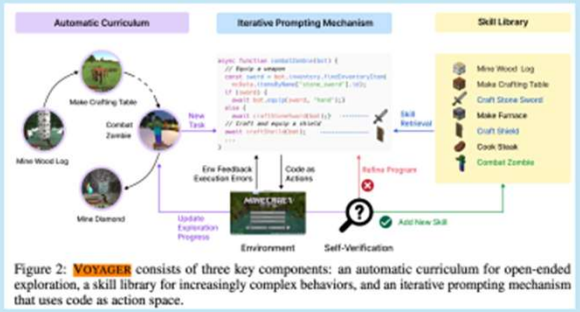
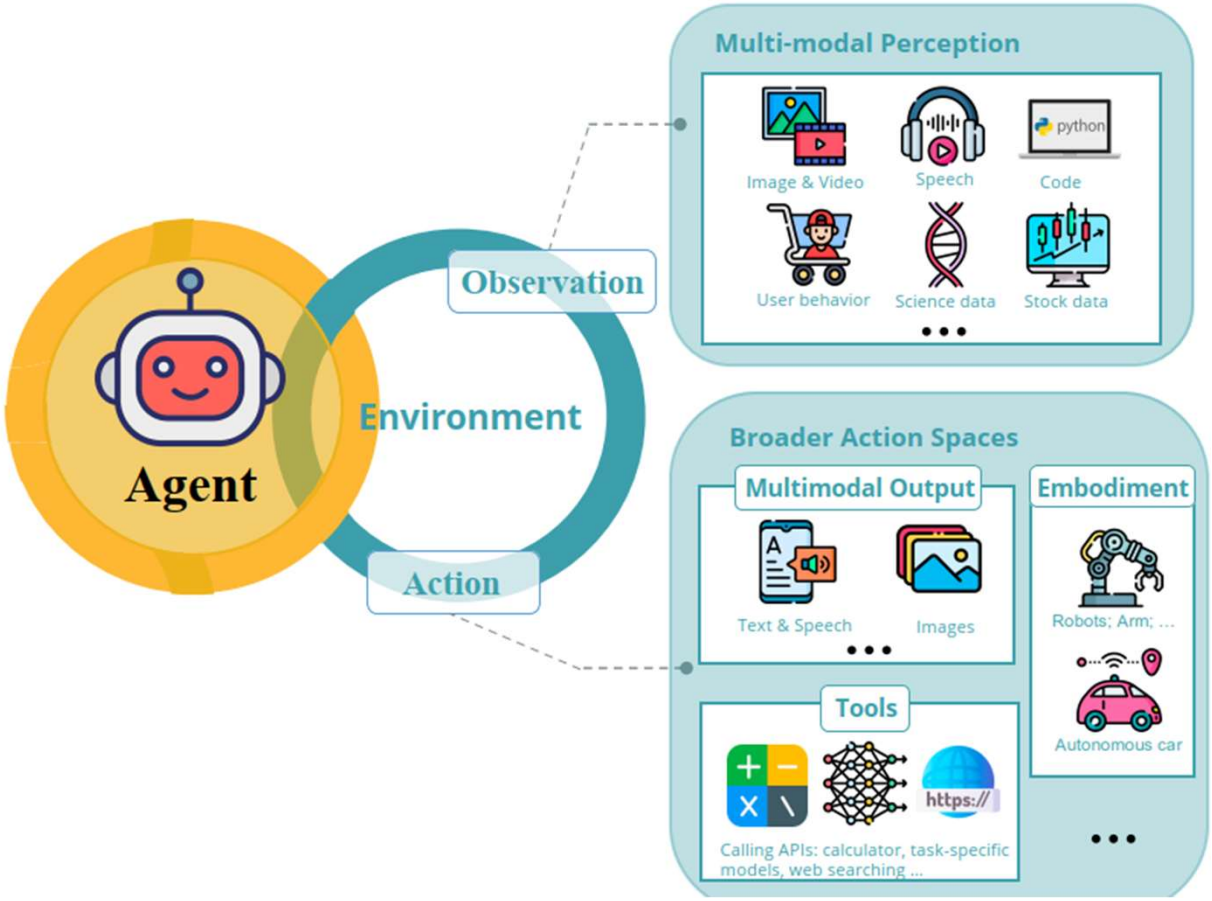
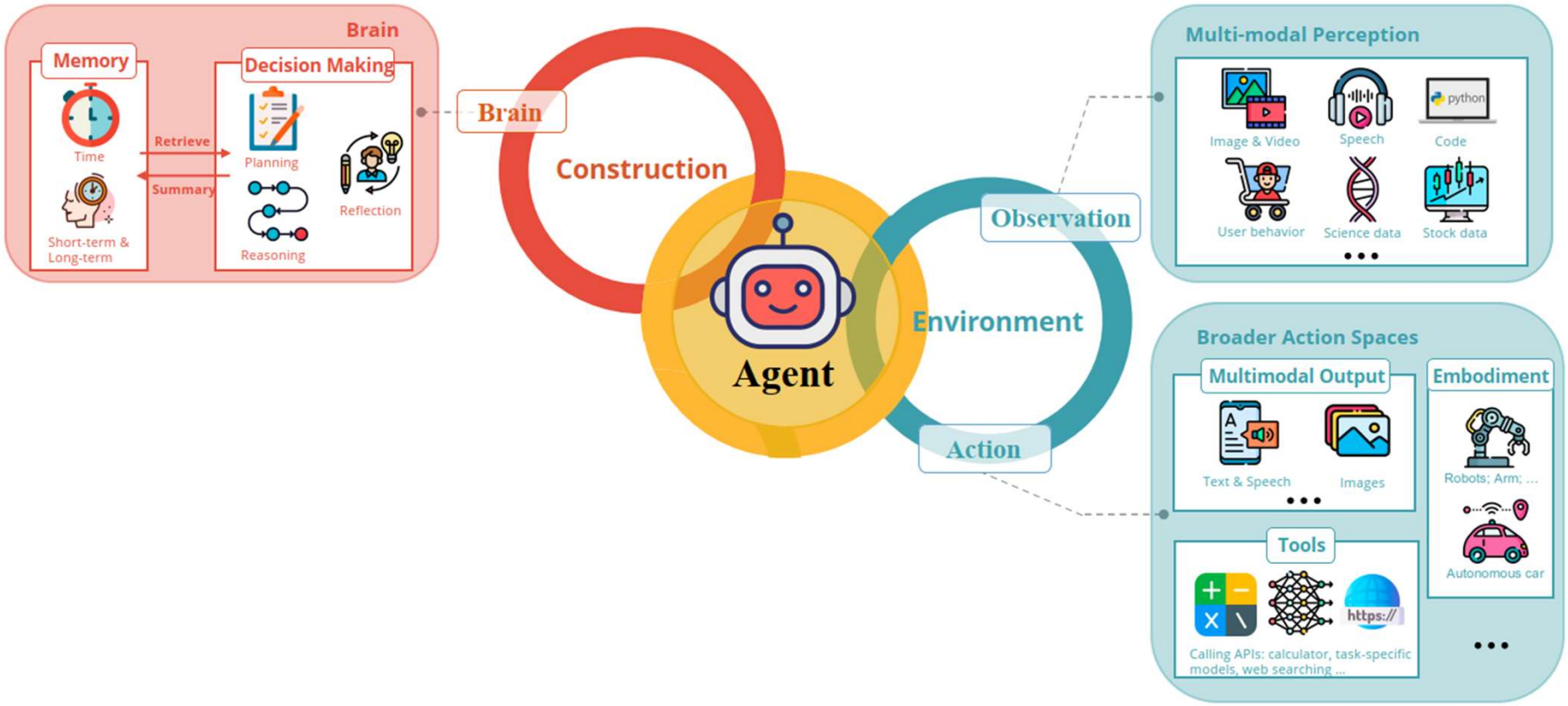


Figure 2: VOYAGER consists of three key components: an automatic curriculum for open-ended exploration, a skill library for increasingly complex behaviors, and an iterative prompting mechanism that uses code as action space.

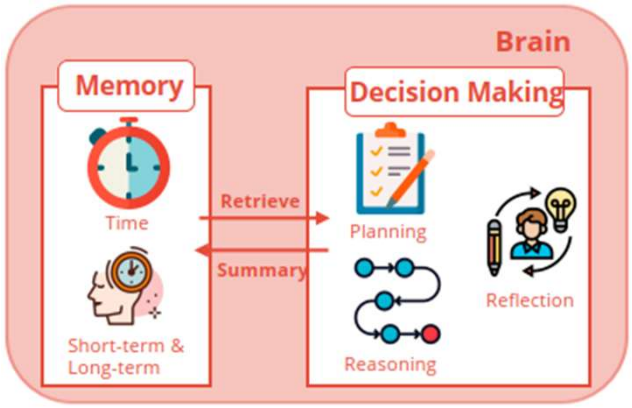


The “Brain”: Decision Making and Memory



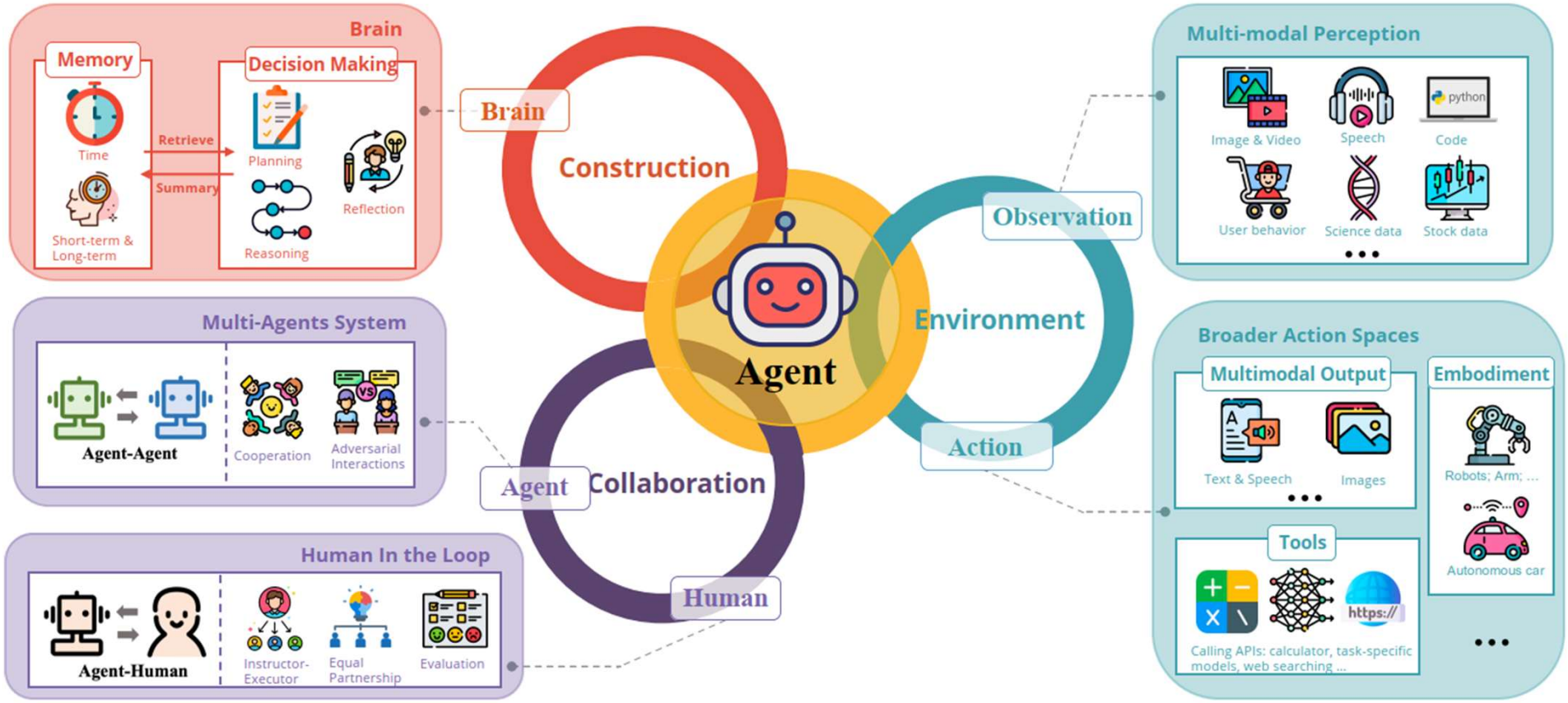
The “Brain”

- **Memory**: stores sequences of agent’s past observations, thoughts and actions
  - simple case: agent’s trajectory
    - past observations and actions
  - more sophisticated memory
    - long-term and short-term memory
    - long-term memory is abstract



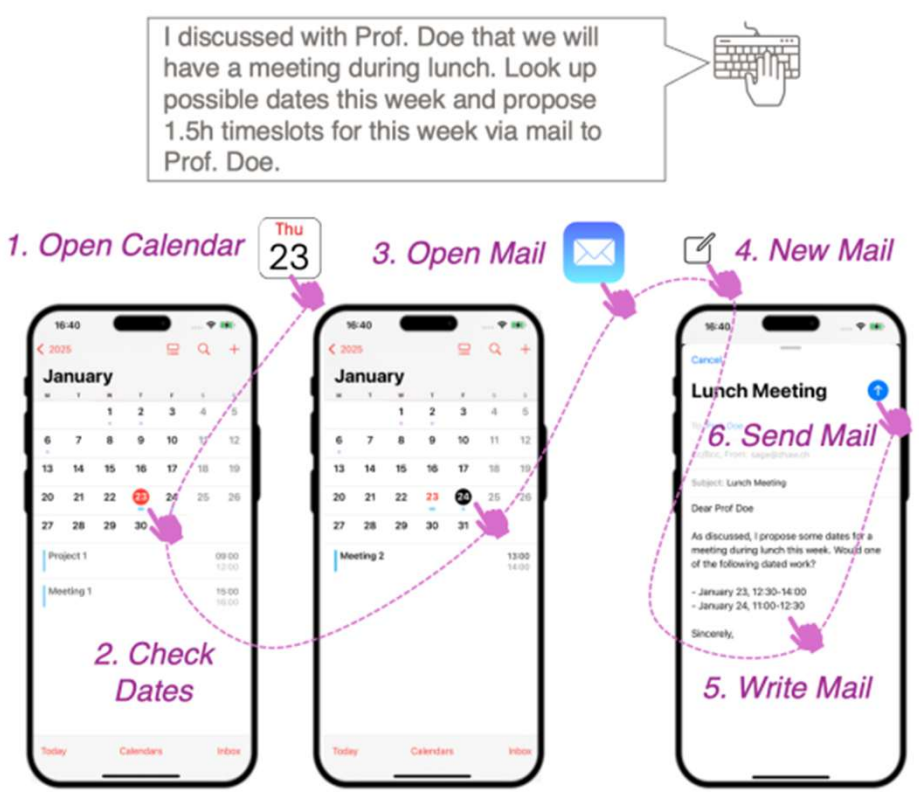
- **Decision Making Process**:
  - **Planning**: Decomposition into subgoals – Break down large tasks into smaller, manageable subgoals, enabling efficient handling of complex tasks
  - **Reasoning**: Self-criticism and self-reflection over past actions, learn from mistakes and refine for future steps

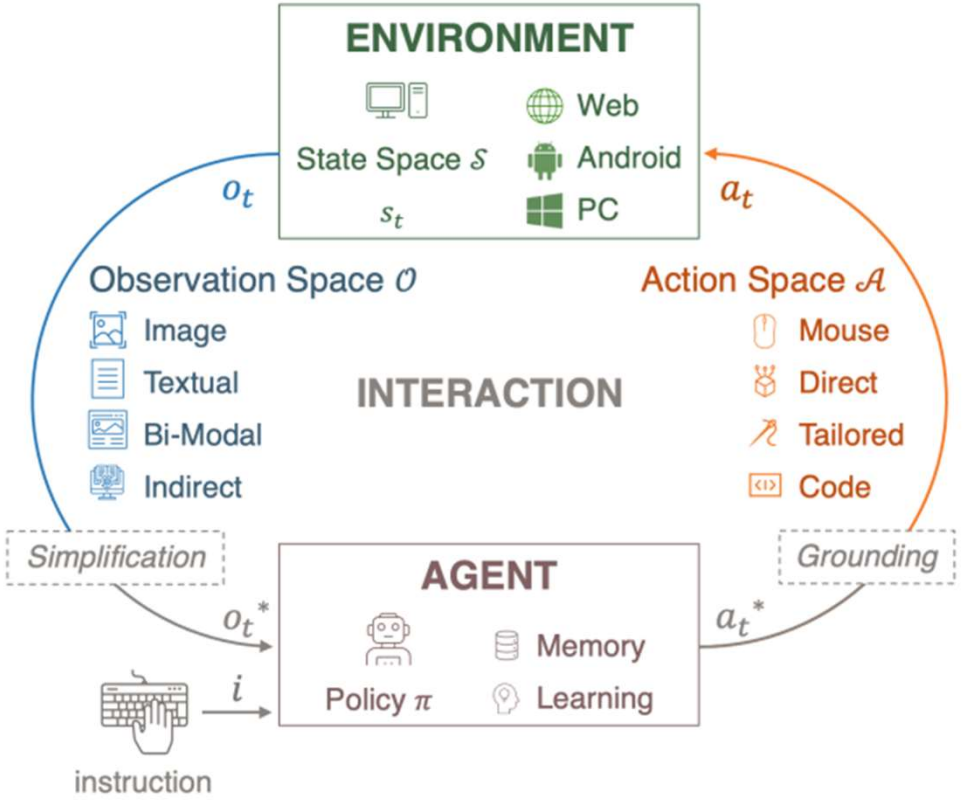
Collaboration



- multiple agents might interact with each other to solve problems in fully autonomous systems
- human-in-the-loop in cooperative systems

Computer Use and Web Agents

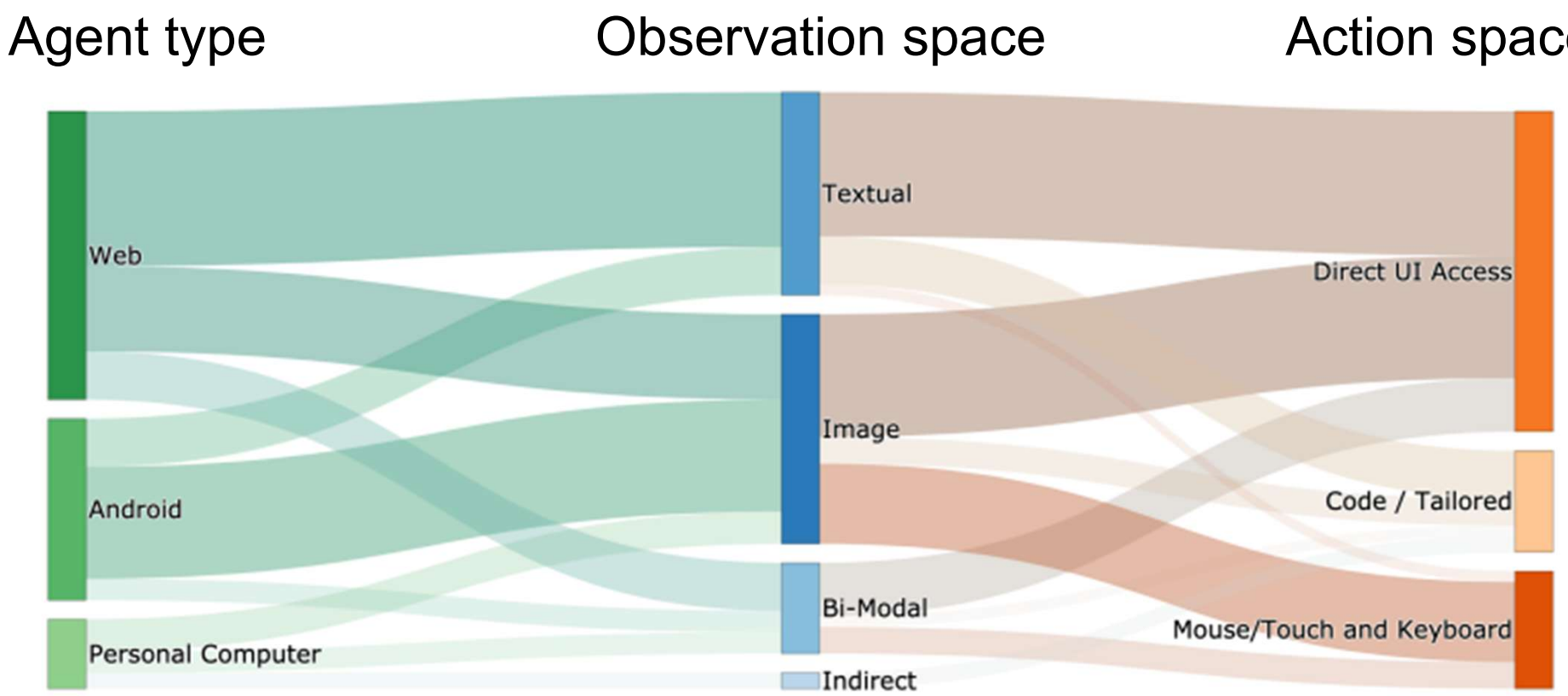




Sager et al.: AI Agents for Computer Use: A Review of Instruction-based Computer Control, GUI Automation, and Operator Assistants. arXiv:2501.16150, 2025

Survey of Computer Use and Web Agents

- recent survey covering 86 papers



Sager et al.: AI Agents for Computer Use: A Review of Instruction-based Computer Control, GUI Automation, and Operator Assistants. arXiv:2501.16150, 2025

2. Seminar Topics and Topic Assignment

- The seminar features literature as well as experimental topics.

- The goal of the **literature topics** is to describe and compare the state of the art methods/approaches concerning the respective topic.

- The goal of the **experimental topics** is to verify methods from literature by applying them to tasks beyond the tasks used in the respective papers.

1. Generalist Agents and Agent Benchmarks

- Literature topic
- Student: Carina Arnold
- Mentor: Christian Bizer

Some papers as starting point

- Wang, et al: A Survey on Large Language Model based Autonomous Agents. arXiv:2308.11432, 2023.
- Sager et al.: AI Agents for Computer Use: A Review of Instruction-based Computer Control, GUI Automation, and Operator Assistants. arXiv:2501.16150, 2025.
- Fourney et al.: Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. arXiv:2411.04468, 2024.
- Xu et al.: TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. arXiv:2412.14161, 2024.

2. Web Agents and Web Agent Benchmarks

- Literature topic
- Student: Lennart Schönweiß
- Mentor: Christian Bizer

Some papers as starting point

- De Chezelles et al.: The BrowserGym Ecosystem for Web Agent Research. arXiv:2412.05467, 2024.
- Zhou et al.: WebArena: A Realistic Web Environment for Building Autonomous Agents. arXiv:2307.13854, 2024.
- Sager et al.: AI Agents for Computer Use: A Review of Instruction-based Computer Control, GUI Automation, and Operator Assistants. arXiv:2501.16150, 2025.

3. Agents for Computer Use

- Experimental topic
- Student: Nguyen, Anh-Nhat
- Mentor: Aaron Steiner

Some papers as starting point

- DeepLearning.AI & Anthropic: Building Toward Computer Use with Anthropic – Lesson 1: Introduction. [Online Course]. https://learn.deeplearning.ai/courses/building-toward-computer-use-with-anthropic/lesson/1/introduction
- Sager et al.: AI Agents for Computer Use: A Review of Instruction-based Computer Control, GUI Automation, and Operator Assistants. arXiv:2501.16150, 2025.
- Anthropic Documentation: Build with Claude for Computer Use. Available at: https://docs.anthropic.com/en/docs/build-with-claude/computer-use
- FranceDot: ACU – Agents for Computer Use GitHub Repository. Available at: https://github.com/francedot/acu

4. Safety of LLM Agents

- Literature topic
- Student: Anna Emsbach
- Mentor: Christian Bizer

Some papers as starting point

- Gan et al.: Navigating the Risks: A Survey of Security, Privacy, and Ethics Threats in LLM-Based Agents. arXiv:2411.09523, 2024.
- Zhang et al.: Agent-SavetyBench: Evaluating the Safety of LLM Agents. arXiv:2412.14470, 2024.
- Levy et al.: ST-WebAgentBench: A Benchmark for Evaluating Safety and Trustworthiness in Web Agents. arXiv:2410.06703, 2024.

5. Energy Efficiency of LLMs and LLM Agents

- Literature topic
- Student: Esha Raheel
- Mentor: Ralph Peeters

Some papers as starting point

- Wu et al.: Addressing the Sustainable AI Trilemma: A Case Study on LLM Agents and RAG, arXiv:2501.08262, 2025.
- Zhou et al.: A Survey on Efficient Inference for Large Language Models, arXiv:2404.14294, 2024.
- Argerich et al.: Measuring and Improving the Energy Efficiency of Large Language Models Inference, IEEE Access, vol. 12, pp. 80194–80207, 2024.

6. Evaluating the Planning Capabilities of LLM Agents

- Experimental topic
- Student: Cora Wackermann
- Mentor: Aaron Steiner

Some papers as starting point

- Huang et al.: Understanding the Planning of LLM Agents: A survey. arXiv:2402.02716, 2024.
- Fourney et al.: Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. arXiv:2411.04468, 2024.

7. LLMs as Evaluators for LLM Agents

- Experimental topic
- Student: Raphael Ebner
- Mentor: Aaron Steiner

Some papers as starting point

- Xu et al.: TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. arXiv:2412.14161, 2024.
- Fourney et al.: Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks. arXiv:2411.04468, 2024.
- Zhang et al.: LLMEval: A Preliminary Study on How to Evaluate Large Language Models. arXiv:2312.07398v2, 2023.

8. Vision LLMs and their Evaluation

- Literature topic
- Student: Fernando Knüttel
- Mentor: Ralph Peeters

Some papers as starting point

- J. Zhang et al.: Vision-Language Models for Vision Tasks: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5625–5644, 2024.
- J. Huang et al.: A Survey on Evaluation of Multimodal Large Language Models, arXiv:2408.15769, 2024.
- Z. Li et al.: Benchmark Evaluations, Applications, and Challenges of Large Vision Language Models: A Survey, arXiv:2501.02189, 2025.

9. Information Extraction from Web Pages using LLMs

- Experimental topic
- Student: Minghao Lei
- Mentor: Ralph Peeters

Some papers as starting point

- Brinkmann, et al.: ExtractGPT: Exploring the Potential of Large Language Models for Product Attribute Value Extraction, Information Integration and Web Intelligence, Springer Nature Switzerland, pp. 38–52, 2025.
- Zou et al.: EIVEN: Efficient Implicit Attribute Value Extraction using Multimodal LLM, arXiv:2404.08886, 2024.
- Zhang et al.: Vision-Language Models for Vision Tasks: A Survey, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5625–5644, 2024.

10. WebRAG for Entity Matching

- Experimental topic
- Student: Ksenia Elagin
- Mentor: Aaron Steiner

Some papers as starting point

- N. Barlaug et al.: Neural Networks for Entity Matching: A Survey, ACM Transactions on Knowledge Discovery from Data, vol. 15, no. 3, p. 52:1–52:37, 2021.
- Y. Gao et al.: Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv:2312.10997, 2024.
- W. Xie et al.: WeKnow-RAG: An Adaptive Approach for Retrieval-Augmented Generation Integrating Web Search and Knowledge Graphs, arXiv:2408.07611, 2024.

11. From Supervised to Reinforcement Fine-tuning of LLMs

- Literature topic
- Student: Luca Schwarz
- Mentor: Ralph Peeters

Some papers as starting point

- S. Minaee et al.: Large Language Models: A Survey, arXiv:2402.06196, 2024.
- S. Wang et al.: Reinforcement Learning Enhanced LLMs: A Survey, arXiv: arXiv:2412.10400, 2024.
- DeepSeek-AI: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv:2501.12948, 2025.

3. How to Structure Your Paper / Presentation

Goals of Literature and Experimental Papers

- Goals of Literature Papers
  1. describe the problem / task
  2. describe several existing methods/systems for handling the task,
  3. compare the methods/systems and their evaluation using a systematic set of comparison criteria

- Goals of Experimental Papers
  1. describe state of the art concerning your problem area
  2. summarize the evaluation tasks and results from the papers
  3. design experimental setup to evaluate technique on different task
  4. compare your results to the results from the paper

How to Structure Your Literature Paper?

1. Introduction and Problem Statement
   - Which problem/task is addressed? Why is the problem important?
   - Structure of your paper
2. Description of Existing Approaches
   - Overview of existing methods and features used by the methods
   - Detailed description of selected methods (likely two)
   - Comparison of the selected methods using a set of comparison criteria
3. Evaluation
   - Comparison and critical discussion of the evaluation tasks, metrics
   - Comparison of the evaluation results using a set of comparison criteria
4. Conclusion
   - What did the comparison of the methods and evaluation results show?
   - Can something be concluded for future work?
5. Bibliography

How to Structure Your Experimental Paper?

1. Introduction and Problem Statement
   - Which problem is addressed? What is the overall approach for addressing it?
   - Overview of the existing methods/papers and their evaluation (3 pages+)
   - Structure of your paper
2. Description of Your Experimental Design
   - How to you select examples for which challenges?
   - Which method/language model combinations do you test?
3. Presentation of Experimental Results
   - Present the results of your experiments (tables containing values and deltas).
   - Present the results of your error analysis (types of errors, frequency of these types)
4. Conclusion
   - What did the experiments and the error analysis show?
   - How to your results compare to the experiments presented in the papers?
5. Bibliography

Learn from Examples

- Read survey articles and previous experimental papers and identify the structure from the previous slides
  - Why can this paragraph be found at that position?
  - What is the purpose of some section / subsection?
- Some relevant surveys
  1. Wang, et al: A Survey on Large Language Model based Autonomous Agents. arXiv:2308.11432, 2023.
  2. Sager et al.: AI Agents for Computer Use. arXiv:2501.16150, 2025.
  3. Mialon, et al.: Augmented Language Models: a Survey. arXiv:2302.0784
  4. Zhao, et al.: A survey of Large Language Models. arXiv:2303.18223
- Textbook on how to write a thesis
  - Zobel: Writing for Computer Science, 3rd Edition, Springer 2014.

Citing Different Types of Publications

1. Journal article, conference and workshop paper
   - Good to cite, current research results
2. Survey articles
   - Good to cite as overviews for specific topics, but prefer individual papers as reference for specific systems
3. Books (sometimes cited)
   - Textbooks
   - Collections of articles/papers => Cite specific paper in book
4. Websites
   - better not cited, exceptions are, e.g., documents like W3C Specifications
   - Do not cite Wikipedia, ever!
   - Use footnotes to refer to project pages, download pages, or technical documentation
5. Slide sets (especially from our lectures)
   - Never cite!

How to Find Relevant Publications?

1. Start with gathering relevant papers from the **surveys**
   1. Wang, et al: A Survey on Large Language Model based Autonomous Agents. arXiv:2308.11432. 2023.
   2. Sager et al.: AI Agents for Computer Use. arXiv:2501.16150, 2025.
   3. Mialon, et al.: Augmented Language Models: a Survey. arXiv:2302.0784

2. **Exploit references:** Given a relevant document x
   - Follow references in the past: papers y that x has cited
   - Follow references in the future: papers y that cited x (**„cited by"** functionality in Google scholar)

3. **Use Google Scholar or Semantic Scholar**
   - we use it a lot ourselves

4. Questions?