Assignment, WASP Software Engineering Course Module 2025

Laura van Weesep (AstraZeneca AB & Uppsala University)

July 2025

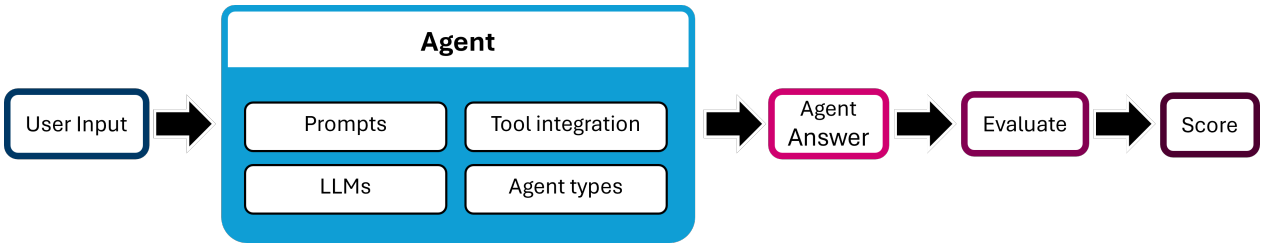

Figure 1: Schematic representation of an agentic workflow

1 Introduction

My PhD project focuses on the development of trustworthy research collaborators for drug discovery. Drug discovery is a demanding and interdisciplinary area of research [5], that has increasingly integrated artificial intelligence (AI) over the past several decades. The field has rapidly progressed and is more and more integrated in the entire drug discovery and development pipeline [2]. The trajectory of the field points towards the next step being trustworthy research collaborators that no longer rely on humans to specify each step and manually transfer the output from one tool to the other, but think along, automate workflows, and ask for input when needed. In my project, we try to use LLM-based agents with cheminformatic tools are beginning to bring this to reality [1, 3, 4].

These systems should be trustworthy and robust. Trustworthiness in this context means minimizing hallucinations, maintaining high predictive stability, and providing clear, interpretable reasoning along with transparent uncertainty estimates. They should also recognize when their own knowledge is insufficient or ambiguous and proactively seek human input when needed or request experimental wet-lab validation.

A robust research collaborator must also demonstrate adaptability to the evolving research landscapes. This involves integrating different underlying LLM models and tools, and easily accommodating improvements in methodology over time. By emphasizing robustness and modularity, these systems can remain effective and relevant as both technology and scientific understanding advance.

Agent design consists of many aspects (Figure 1). Currently, my investigation is primarily directed toward two aspects. Firstly, I am exploring the modularity of agentic systems, studying how discrete functions can be combined and recombined to build flexible, evolvable AI collaborators (as illustrated in figure 1). Secondly, I am examining how large language models interpret and respond to chemistry questions, looking at the reasoning and the embedding space involved.

2 Lecture principles

Two points that stuck from the lectures are **testing** and the **human component** in adaptation of models. In regards to testing I acknowledge that early testing is crucial and it was re-emphasized to me that also diversity of test cases is crucial. In my research this is implemented by testing components before integrating them in a system. Also I will think more about how to test my system, with increasing complexity. I can for instance ask trivial questions that require the agent to call a single tool, which they should be able to do without problems, but to think of more cases I will add multiple requests at the same time, novel research questions and also edge cases. These should not only show what the model can do, but also how it responds to tasks it was not made for. This will involve malignant use and usage outside of the scope of the agents, which it would no longer be able to preform reliably on.

The second point that really stuck is that you should think about the human that will use your solution. In my case it will be experimentalists in the wet-lab. This means that performance is important, yet reliability and usability are prioritized over peak performance. On top of this, instead of only evaluating the outcome of the agent, I will now also evaluate how user friendly it is, how well humans can follow the reasoning and how this relates to the needs of chemists.

3 Guest Lectures

From the guest lectures, the most important takeaway for me was how to **implement requirements engineering** in a more practical way. In my research I started implementing this technique by first identifying the stakeholders and I discovered that this is already quite challenging. This is where the second takeaway comes in: **manage complexity**. In a dream scenario we would like to directly design an agent capable of doing everything, yet, in a PhD and also in industry, output is already expected within a certain scope. Additionally, by starting simple, but taking into account possibilities to extend, it is easier to iterate over the prototype. In my research I apply this by narrowing down the initial scope, but incorporating the broader scope in the test set so we can monitor progress.

4 Machine Learning Scientists versus Software Engineers

Differences between machine learning scientist and software engineers The book brings forward several differences between software engineers and machine learning scientists. Software engineers traditionally focus on the actual real world application and create solutions that fulfill users need, whilst for most machine learning scientist focus is placed on data and model performance in terms of accuracy, precision or another numeric non-human based metric. Software engineering in my view seems to work based on constraints and creating solutions that fit within those, whereas machine learning to me seems to work based on data and works bottom up to generate predictions based on that data. What machine learning scientist can learn from software engineers is that you should be aware that solutions will be used by humans and that the real life application brings constraints beyond performance metrics like accuracy. I think machine learning scientists can be inspired to take into account practical constrains once applying the machine learning model. However, I think that there should also be a selection of scientists driving the field forward. These scientists could innovate based on the constraints that more application focused machine learning scientists run into. These could be improving the latency, improving explainability. In this way I see a feedback loop that inspires machine learning scientists. Also, thinking about practical implication, helps motivate certain choices, or at least report limitations and manage expectations further down the application process. They don't need to deal with everything themselves, but should think along. What are the constraints? What are the consequences of using this model? How was this data collected and what biases does it bring? Another important aspect in this, is also to think about how performance in a training setting translates to real-life and is this fair. Questions to ask in this regard in my opinion, are :"How noisy is real data?" and "Is my training data actually representative?" In my opinion, there should be another party answering these questions and it is then up to the machine learning scientist to process that information and translate it into machine learning relevant components. They should become part of an iterative process, but I don't think machine learning and engineering should happen simultaneously. Software engineers on the other hand, could benefit from getting inspiration from machine learning scientists that drive the field forward and come up with new ways to satisfy the constraints. Machine learning based models could be an additional tool that software engineers could deploy as part of their system and help them think about how to push solutions forward to achieve higher performance overall.

In regards to roles, I think that software engineers and machine learning scientists both should increase awareness of the other components, but I also see room for educating people that work at the interface. These people could help translate requirements to machine learning relevant metrics, such as "high speed" to "low latency". Also informing people that machine learning comes with limitations and what those limitations are would require new expertise in my opinion. These would not be people that are building the model, but that are aware of the choices made and the consequences of that.

5 Paper Analysis

To select 2 papers of the CAIN conference, I used an iterative process. Initially I selected papers based on the title and then to further narrow it down I read the abstract. I noticed that some titles were slightly ambiguous, as many concepts such as "safety" can be applied at different levels of a framework, as well as

refer to the application domain and the model itself. The two papers of last years additional that seemed relevant are the ones that I selected.

5.1 Paper 1: RAGProbe: Breaking RAG Pipelines with Evaluation Scenarios

I selected this paper because evaluating the outputs of LLMs is a challenge that I am currently dealing with in my PhD project.

5.1.1 Core Concept

In this paper they try to address the challenge of evaluating retrieval augmented generation (RAG) systems. Currently this is often done manually, but this limits the scalability and benchmarks that are developed based on this approach are not tailored to specific domains. Ideally, this process would be automated, which is what they do in this paper. They describe how to generate question and answer pairs and found that by asking multiple questions at the same time for instance, the failure rate of RAG systems increase, shedding a new light on what developers should prioritize. This relates to software engineering as this paper improves testing and evaluation of RAG systems. The paper describes a tool to create question and answer pairs based on the specific scenario that the user wants to evaluate. Within this approach the LLM-as-a-judge system is used to evaluate, not just the final answer, but also the weaknesses in context comprehension, formatting and completeness. In this way the paper puts forward an automated way to setup evaluation scenarios that give more insight into what element to improve exactly and also expose limitations more elaborately, ultimately helping developers prioritize and gain insight into the weaknesses.

5.1.2 Relation to Research and Adaptation

In hindsight, the paper does not answer the questions I had about verifying whether "novel" answers could also be correct, gathering information from documents and combining it in ways not anticipated by the researcher could lead to new insights, but this is not addressed in this paper. Nevertheless, the paper brought forward some interesting views to me. What I took from the paper is that by breaking down the evaluation into different elements it is easier to understand the weaknesses of the system. In my PhD project I am using the LLM to call cheminformatics tools and if I could pinpoint whether it is the question itself, something wrong with the code or the LLM not understanding the specific context, this would make improving the system more straightforward, especially as I am moving to more complex multi-agent systems.

5.1.3 Integration Into a Larger AI-intensive Project

To illustrate the use of this paper, let's take a fictional project. The idea is that for the clinical trials you have a lot of patient data and patient reports and you want to build a RAG system to evaluate the efficacy of a specific drug that you tested. This system will now be used for all upcoming drugs that you want to test in patients. It should guarantee the privacy of patients, it should be reliable and you prioritize accuracy over speed and it should also be able to use sources such as the internet, but also internal databases and scientific sources to try to explain symptoms patients are experiencing and write a summary on weather this is related to the specific drug. Since the decisions based on the outcomes are high stakes, accuracy is of utmost importance and tracing back information should be simple. For this you build a multi-agent system. Core elements include a RAG looking through patient reports, another RAG looking at patients biometric data and an agent managing tools that can lookup information in databases, on the internet. On top of this you would also need a coordinating agent and an agent that manages the information coming in from all the different components of the system. The paper would suggest to break down the evaluation of the RAG system into parts. The paper presents a way to automate the evaluation. In this case, it would make sense to use a multi question scenario, as you ask about side effects, the potential causality of them and whether they commonly occur across patients. This also means that for some questions, answers will be in different chunks of one source document and others will be divided over documents. The systems does not work as expected intially and you decide to evaluate the RAG system. By making use of the given scenario, you could now identify where the RAG system struggles. The context does not seem to be relevant for the questions you are trying to answer and the system seems to not be aware of what patient which data is from. Based on

this observation you would add an embedding of the document itself to each chunk of text. This solves the problem of forgetting patient context. My WASP project would fit into the tool-coordinating agent part. I would work on building a system that takes information about the adverse effects on the patients and based on that fetches information from the other sources to decide whether there is a correlation to the drug or not. Also positive effects will be evaluated. How the tools communicate is something I will have to do prompt engineering for. Again this paper provides guidelines on how to evaluate this subsystem. I could break down the evaluation into reliability, relevance, formatting and overall performance for instance. Nevertheless, in contrast to a RAG system, I don't directly have access to the full information. Inspired by the paper however, I could generate automatic question answer pairs. How to do this is part of my PhD project. Currently, we do this manually, but coming up with variations of questions would increase the dataset size.

5.2 Paper 2: Themes of Building LLM-based Application for Production

This paper I selected because it covers the main theme of my PhD, using foundation models (mainly LLMs) to build a drug development research collaborator. I currently work on using LLMs to communicate between agents of my multi-agent system and as backbones for some of my agents.

5.2.1 Core Concept

To adress the lack of an overview of key considerations for developing LLM-based system, in this paper they generated transcripts of video's related to LLM-based applications and analyse themes occuring in these. The core idea of this is that in this way people working towards integrating their products in production. The main topics that they came across for LLM-based systems is RAG systems, model capabilities and enhancement techniques. Also ethical challenges and risks were frequently discussed. This relates to software engineering by highlighting approaches to make LLM-based systems safe and what risks to consider. They also elaborate on how LLMs are different from more traditional ML systems, pointing out the LLMs are trained on extensive general purpose data for instance Also proposing RAG systems and LLM-based systems as part of software solutions is something to consider. In the paper they elaborate on what this entails and how the model element of LLM-based software can be enhanced through fine tuning or prompt engineering.

5.2.2 Relation to Research and Adaptation

This paper relates to my research a lot. First of all, it made me think about the differences to more traditional ML. The fact that the data it was trained on is so vast and it is also not clear what the data the model was trained on contains exactly is something I deal with regularly as we are trying to get these models to perform on chemistry tasks. Based on the themes highlighted by this article, I understand that if the project is to be successful for industry later on, factors to consider are also socio-technical and that communication with/between different stakeholders, such as chemists, cheminformatitions and managers has a high influence on how the product is going to be used and adopted in practice. Based on this, for my project I adjusted the requirements and decided to include stakeholders earlier on. The themes identified inspire me to think about practical considerations such as latency as well and reconsider the benchmarks to use for my project as they might not be relevant for my use case. It also opened my eyes to adverse attacks that can happen and that I should safeguard agains leaking API keys or resource consuming attacks.

5.2.3 Integration Into a Larger AI-intensive Project

To illustrate the use of this paper, let's go back to the previous example mentioned under paper 1. To quickly summarize: a multi-agent system with a RAG agents, a summarizing agent, a lookup agent and tool-based agents that all communicate through LLMs and/or use LLMs at their core. This paper could help to productionalise this system in multiple ways. For the RAG system components, the article points out different parameters to take into consideration, such as the chunking size. Also the article points out that RAG systems are a dynamic solution suited if you information source changes over time or per application. In this example it would change based on which clinical trial is being analyses. Moving on, the paper goes into problems that could be encountered in multi agent projects, such as getting tuck in loops. Adding safeguards to set a maximum number of steps is a way to mitigate this for production. An additional thing to think

about is the memory usage, and in practice only the most relevant information should be retained. To get users to actually use the system, this article provides some pointers for the user interface. In practice this translates to knowing your audience and what would work best for them. What kind of multimodalities do they require and how technical should the interface be. This also relates to the latency. In a production case, it should flow. Similarly, the tool should fit the purpose. If you don't need an expensive model, it is a waste to use one. To improve upon the underlying AI system itself, the article points out two main directions. First of all, models could be fine-tuned, but data quality is undershadowed at the moment. In practice you would need to focus on data quality first over optimizing fine tuning techniques according to the authors. Alternatively, prompt engineering could be used, but they are sensitive to small modifications. This is where my PhD project comes in. We are investigating how to avoid fine tuning and achieve performance through prompt engineering instead. Lastly, for an AI intensive project, evaluation is often key. They point out that human in the loop evaluations are now seen as the golden standard, but that this is not without a cost. This comes with biases and is highly resource intensive. Looking for alternative ways to do this, is something that should be part of an AI project and not just an afterthought.

6 Declaration of Originality

Through my employer I have access to NotebookLM. This could service allows you to upload papers for instance and has an interactive chat through which you can ask questions about the documents. I used this service to discuss core concepts of the RAGProbe paper as I struggled to understand some of the nuances in the paper. It did not really aid me much though, so in the end, I based my answers on the notes I had made of the paper myself. Based on this I summarized the core concept for this assignment.

References

[1] Gao, B., Huang, Y., Liu, Y., Xie, W., Ma, W.Y., Zhang, Y.Q., Lan, Y.: Pharmagents: Building a virtual pharma with large language model agents. arXiv preprint arXiv:2503.22164 (2025)

[2] Garg, A.P., Jabborova, D., Patil, A.: Artificial intelligence in biopharmaceutical: Revolutionizing drug discovery amidst industry challenges. Biomedical Journal of Scientific & Technical Research (2025)

[3] Inoue, Y., Song, T., Fu, T.: Drugagent: Explainable drug repurposing agent with large language model-based reasoning. arXiv preprint arXiv:2408.13378 (2024)

[4] Kim, H., Jang, Y., Ahn, S.: Mt-mol: Multi agent system with tool-based reasoning for molecular optimization. arXiv preprint arXiv:2505.20820 (2025)

[5] Szymański, P., Markowicz, M., Mikiciuk-Olasik, E.: Adaptation of high-throughput screening in drug discovery—toxicological screening tests. International journal of molecular sciences **13**(1), 427–452 (2011)