

Leiden University

ICT in Business and the Public Sector

Trust and Fabrication Identification in LLM-Based Systems: The Role of AI Experience, Attitudes, and Training

Name:	Ruben Coen Julian Visser
Student ID:	s3334767
Date:	01/07/2026
1st supervisor:	Dr. N. Amat Lefort
2nd supervisor:	T.D. Offerman MSc

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Abstract

People are increasingly using Large Language Model (LLM)-based systems nowadays. However, they do not always verify whether the answers these systems generate are correct. We do not know if this applies to all people using these systems, or only specific users. Therefore, we tested to see how someone's Prior Artificial Intelligence (AI) Experience and General Attitude toward AI would be related to how many fabrications they were able to identify and to their trust after using these systems. We also investigated if a short training would be able to moderate these relationships. To be able to answer this, we conducted an experiment with 101 participants, of which 51 were part of a control condition and 50 of a training condition. They all completed a task using an LLM-based chatbot that gave accurate, but also fabricated responses.

Based on this we found that the Prior AI Experience of someone was positively related to trust after completing the task ($\beta = .410$, $p = .049$), but it was not significantly related to how able they were to identify the fabrications ($\beta = -.205$, $p = .430$). The General Attitude toward AI showed exactly the opposite. It was namely related to people finding fewer fabrications ($\beta = -.536$, $p = .008$), but it did not have a significant effect on someone's trust ($\beta = .190$, $p = .247$). When looking at the role of the training we could not find a moderation effect as the p-values ranged from .143 to .865. Instead, this training did have an overall effect of improving people's ability to identify fabrications ($M = 1.98$ vs. $M = 0.88$, $p < .001$) and lowering their trust in the system ($M = 2.70$ vs. $M = 3.24$, $p = .001$). This shows that the people's trust in LLM-based systems is related to how much Prior AI Experience they have and that how many fabrications they can identify is related to how low their General Attitude toward AI is. Training can help improve everyone in general.

Contents

Abstract **1**

1 Introduction **4**
1.1 Background of the Problem . . . 4
1.2 Current Problem in Practice . . . 4
1.3 Research Gaps . . . 5
1.4 Research Objective and Question . . . 6
1.5 Thesis Structure . . . 7

2 Theoretical Background **8**
2.1 LLM-based Systems and Hallucinations . . . 8
2.2 Human-LLM Trust and Calibration . . . 8
2.3 The Three-Layered Model of Trust . . . 9
2.4 Prior AI Experience . . . 10
2.5 General Attitude toward AI . . . 11
2.6 The Training Intervention . . . 12
2.7 Research Model . . . 13

3 Methodology **14**
3.1 Research Design . . . 14
3.2 Participants . . . 15
3.3 Materials . . . 17
3.3.1 The LLM-based Chatbot and Source Documents . . . 17
3.3.2 The Training Intervention . . . 18
3.4 Pilot Testing . . . 19
3.5 Full Experiment . . . 20
3.6 Measures . . . 21
3.6.1 Background Questions . . . 21
3.6.2 Predictor Measures . . . 21
3.6.3 Outcome Measures . . . 21
3.7 Data Analysis . . . 22

4 Results **25**
4.1 Measurement Model . . . 25
4.2 Descriptive Statistics . . . 27
4.3 Testing the Baseline Hypotheses (H1 - H4) . . . 27
4.4 Testing the Moderation Hypotheses (H5 - H6) . . . 28
4.5 Additional Analysis: Main Effect of the Training . . . 29

5 Discussion **31**
5.1 Interpretation of the Results . . . 31
5.1.1 Prior AI Experience . . . 31
5.1.2 General Attitude toward AI . . . 32
5.1.3 Training . . . 32
5.2 Implications of the Results . . . 33
5.2.1 Theoretical Implications . . . 33
5.2.2 Practical Implications . . . 33

5.3 Limitations 34
5.4 Future Research 35

6 Conclusion **36**

References **37**

A System Prompt **41**

B Overview of Experimental Questions, Keyword Triggers, and Embedded Hallucinations **42**
B.1 Computer Science 42
B.2 Medicine 42
B.3 Law 43
B.4 Psychology 44

C Training Materials (PowerPoint Slides) **45**

D Full Questionnaire **49**
D.1 Informed Consent 49
D.2 Demographic and Background Questions 49
D.3 Prior AI Experience 50
D.4 General Attitude toward AI 51
D.5 Experimental Task 51
D.6 Post-task Trust 52

Chapter 1

Introduction

1.1 Background of the Problem

Large Language Model (LLM)-based systems such as ChatGPT, Copilot and Claude are used by a growing number of people. People use these systems for tasks such as writing and programming support, and answering everyday questions (Dwivedi et al., 2023). These systems are very useful as they are able to generate text that is realistic and coherent (Alkaissi & McFarlane, 2023). As a result, the generated text can often be used directly with little effort. However, this advantage for users also comes with a major disadvantage. The text these LLM-based systems generate can contain information that is incorrect, even though the answers look fluent and accurate (Ji et al., 2023). This incorrect information is known as "hallucinations" (Ji et al., 2023; Anh-Hoang et al., 2025). This forms a problem for people using these systems, as it is often difficult to spot those hallucinations as shown in real-world examples. For example, lawyers submitted court cases generated by ChatGPT that did not exist, which led to sanctions and reputational damage (Bohannon, 2023). In another example, a report made by Deloitte contained fabricated references and quotes which resulted in a partial refund (Paoli, 2025).

1.2 Current Problem in Practice

These examples show that even people with high knowledge in specific domains may accept fabricated outputs without proper verification. This means that the problem is not only limited to people with little knowledge in a specific domain, but is much broader. At the same time, not every person who uses these LLM-based systems makes these mistakes. Some people have more trust in a system than others (Hoff & Bashir, 2014). This shows that not everyone has calibrated their trust correctly, which means according to Lee & See (2004) that these people have not aligned their trust with the actual capabilities of the system. If we apply their theory to LLM-based systems, we can interpret that when someone's trust is too high, they may accept

fabricated answers without checking if they are correct. When trust is too low, they may not use the tool even though the tool produces useful answers. In practice, primarily too much trust is problematic because it can have big consequences. In that case it can be that people use incorrect information in situations where correctness is very important, for example in healthcare. Ideally, you want these people to have calibrated their trust correctly so that they can use LLM-based systems effectively without accepting their possible fabricated answers too quickly. However, this often goes wrong in practice.

1.3 Research Gaps

Research into automation has shown that trust can depend on individual differences (Hoff & Bashir, 2014). For people's trust in LLM-based systems, this could mean that their trust partly depends on their own user characteristics and can therefore differ from person to person. Existing research has already examined user characteristics related to Artificial Intelligence (AI). For example, some studies have developed and validated questionnaires to measure people's general attitude toward AI (Schepman & Rodway, 2020; Grassini, 2023). Other research for example, examined how a user characteristic such as AI literacy is related to how valuable an LLM-based system is perceived and whether people want to use it (Al-Abdullatif & Alsubaie, 2024). But what these current studies do not show is if these user characteristics are related to whether someone trusts LLM-based systems too quickly and accepts fabricated outputs without verifying them. This is the first gap in the current LLM research.

The second gap in LLM research is about training. Current research on automation says that training can help to prevent people from misusing systems when they misjudge their reliability (Hoff & Bashir, 2014). Although research on LLMs shows that critical oversight is very important (Choudhury & Chaudhry, 2024) and that AI literacy can help people gain more understanding and a more critical view (Laupichler et al., 2022), it does not yet show whether such training helps people to evaluate the outputs of LLM-based systems more critically and calibrate their trust. It also does not show whether this differs between people with different user characteristics. Because of these gaps, it is difficult to know which people are most likely to trust fabricated LLM outputs too quickly and therefore need more support. It is also difficult to know whether everyone needs the same kind of support, or whether training is especially useful for certain people.

1.4 Research Objective and Question

To address these gaps we will look at how two user characteristics, Prior AI Experience and General Attitude toward AI, are related to whether someone can identify fabrications and to their trust. We will also investigate how a training may influence these relationships. Therefore, we will investigate the following research question:

> *How do Prior AI Experience and General Attitude toward AI relate to users' ability to identify fabricated LLM outputs and their Post-task Trust, and to what extent does a short training intervention moderate these relationships?*

Since the research question consists of different parts, we decided to formulate various subquestions so that we can answer this question in a more structured way. The first four subquestions will examine people who have not received training to see what the general effect of the two user characteristics is. In the fifth subquestion we examine if the training has an influence on these relationships. The different subquestions we will investigate are as follows:

- **SQ1:** How does Prior AI Experience relate to a user's ability to identify fabricated LLM outputs in the absence of training?
- **SQ2:** How does Prior AI Experience relate to a user's Post-task Trust in an LLM-based system in the absence of training?
- **SQ3:** How does a user's General Attitude toward AI relate to their ability to identify fabricated LLM outputs in the absence of training?
- **SQ4:** How does a user's General Attitude toward AI relate to their Post-task Trust in an LLM-based system in the absence of training?
- **SQ5:** To what extent does the training intervention moderate the relationships examined in SQ1 through SQ4?

1.5 Thesis Structure

We structure this thesis as follows. In Chapter 2 we present the theoretical background and develop the hypotheses and the research model. After that, in Chapter 3 we describe the methodology, including the research design, the measures and data analysis. The results of the measurement model and the different hypotheses are presented in Chapter 4. In Chapter 5 we discuss the findings, implications, limitations and suggestions for future research. Finally, we conclude the thesis by answering the subquestions and main research question of this thesis in Chapter 6.

Chapter 2

Theoretical Background

2.1 LLM-based Systems and Hallucinations

LLM-based systems are AI systems that generate textual responses to user questions (Dwivedi et al., 2023). They generate these responses word by word, each time predicting a likely next word based on the previous words. This is based on patterns they learned from large amounts of training data, rather than having a real understanding of what it means (Bender et al., 2021). The consequence of this is that identical questions can lead to different responses (Reiss, 2023) and that they are not always factually correct (Ji et al., 2023). As a result, people should not assume that the answers are always correct, even though they may look convincing. Previous research has namely shown that in answers containing citations, these citations were fabricated in about 30% to over 90% of the time depending on the specific model that was used (Chelli et al., 2024).

This shows that such hallucinations can occur very frequently. However, as LLM-based systems do not provide indications whether an answer is correct or fabricated (Ji et al., 2023), people need to rely on their own judgement. Therefore they often focus on specific characteristics of an answer to determine if they want to use it, such as the length (Steyvers et al., 2025). This means that they do not always verify the information that the answers contain, even though they could do it.

2.2 Human-LLM Trust and Calibration

This lack of verification is exactly why trust plays an important role when using LLM-based systems. Mayer et al. (1995) define trust as "*the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party*" (p. 712). This is also related to the argument of Muir (1987), who says that trust becomes particularly

important in situations where people cannot fully understand or predict the behaviour of the system. This is also valid for LLM-based systems as people using these systems rely on the answers they generate while they are not able to fully monitor or control how these answers are generated. It is therefore even more important that people do not trust these systems too much, because they cannot see how the answer was made and if it is reliable. To use these systems the right way people therefore need to calibrate their trust (Lee & See, 2004; Hoff & Bashir, 2014). However, Jian et al. (2000) show that trust and distrust are opposites on a single scale, meaning that the trust of a person can be very low, in the middle, or very high. This is also the case with GenAI, which includes LLMs. Amoozadeh et al. (2024) found that among students, most trusted these systems (47%), while others were neutral (36%) or distrusted them (16%). This shows that people differ a lot in how much they trust the same system, which suggests that not everyone calibrated their trust well.

2.3 The Three-Layered Model of Trust

The reason why people do not all have the same trust in LLM-based systems can be explained using the three-layered model of trust in automation by Hoff & Bashir (2014). Trust is namely divided in three different layers: dispositional trust, situational trust and learned trust. The first layer, dispositional trust, refers to a person's general tendency to trust or distrust automated systems (Hoff & Bashir, 2014). Some users may be more willing to rely on automated systems, while others may be more careful. The second layer is situational trust, which refers to trust that depends on the specific situation in which someone uses the automated system, for example influenced by contextual factors (Hoff & Bashir, 2014). The third layer is learned trust. Within this layer, Hoff & Bashir (2014) distinguish between initial learned trust and dynamic learned trust. Initial learned trust refers to expectations that people already have based on previous experiences with other automated systems, for example their understanding of the system. Dynamic learned trust refers to trust that develops during the use of the system itself. This makes that people all have different trust in the same system.

This model helps to explain why we included the two user characteristics, Prior AI Experience and General Attitude toward AI, and the training in our research question. General Attitude toward AI reflects a person's overall view of AI-based systems. This connects to dispositional trust. Prior AI Experience, which is about a person's perceived skills in using

these systems and their confidence in understanding how these systems work, is connected to initial learned trust. The training gives people extra information before they complete the task, which changes the context of the interaction. This connects to situational trust. We examine how these user characteristics and the training are related to the outcomes of Post-task Trust and Fabrication Identification. Post-task Trust shows how much users trust the LLM-based system after using it. In the model this is related to dynamic learned trust. These links between our own variables and the model are not one-to-one mappings. Fabrication Identification shows how well people can identify fabrications. It is not part of this model on trust, but it is relevant as trust may be related to how critically people evaluate the answers they receive from an LLM-based system. A quick overview of this can be seen in Table 1.

Table 1
Link Between Variables and the Model of Hoff & Bashir (2014)

Model Layer (Hoff & Bashir, 2014)	**Variable in Thesis**
Dispositional Trust	General Attitude toward AI
Initial Learned Trust	Prior AI Experience
Situational Trust	Training
Dynamic Learned Trust	Post-task Trust
-	Fabrication Identification

2.4 Prior AI Experience

Hoff & Bashir (2014) show that other studies suggest that people's experience with similar systems leads to them being more likely to trust and rely on a system. This means that people with more Prior AI Experience may already have stronger expectations that an LLM-based system will perform well. This can help them use these systems more easily, but it may also make them less careful when evaluating the information the LLM-based system gives. We can explain this using the difference between "System 1" and "System 2" thinking by Kahneman (2011). Kahneman (2011) explains that when something becomes more familiar and requires less effort, people are more likely to rely on the fast and intuitive "System 1" thinking. This could be the case for people with higher Prior AI Experience.

However, to be able to critically evaluate LLM-based answers and to get appropriate trust,

people need to use their slower and more thoughtful "System 2" thinking. When people do not use their "System 2" thinking, they stop monitoring the system closely because they overtrust its reliability (automation complacency) and rely on the answers without verifying them (automation bias) (Parasuraman & Manzey, 2010). Because of this we expect that people who have more experience are less likely to spot fabrications. At the same time, we also think that because their past experience makes them expect that a system works well, they will also report higher levels of trust. Based on this, the following hypotheses are proposed to answer the first two subquestions:

H1: *In the absence of training, higher levels of Prior AI Experience are negatively related to Fabrication Identification.*

H2: *In the absence of training, higher levels of Prior AI Experience are positively related to Post-task Trust.*

2.5 General Attitude toward AI

People's attitudes affect how they process information. They often focus only on information that supports what they already believe and ignore the information that contradicts their beliefs (Albarracín et al., 2018). Therefore, people who have a positive view of AI are more likely to focus on how useful and beneficial it is (Schepman & Rodway, 2020), instead of also taking into account the negative side of AI. According to Sundar (2007) people's attitudes can activate so-called heuristics when processing information, such as the "machine heuristic". This means that people believe that what is produced by a machine is objective and free from bias. In the context of LLM-based systems, this may result in automation bias (Parasuraman & Manzey, 2010; Skitka et al., 1999). Therefore, we think that people who have a more positive attitude toward AI are less likely to identify fabrications and have more trust in the system. This leads to the following hypotheses that are proposed to answer subquestion 3 and 4:

H3: *In the absence of training, a more positive General Attitude toward AI is negatively related to Fabrication Identification.*

H4: *In the absence of training, a more positive General Attitude toward AI is positively related to Post-task Trust.*

2.6 The Training Intervention

Training may help to reduce these negative effects of a higher Prior AI Experience and a more positive General Attitude toward AI. Hoff & Bashir (2014) namely explain that when people learn how reliable a system actually is, they can better judge when they should trust it. This may change how much they trust the system and how they use it. Training can therefore make people more aware of the limitations of a system and encourage them to evaluate the outputs more carefully. For people with a higher Prior AI Experience that may mean that they shift more from the "System 1" thinking toward the "System 2" thinking. When they reduce their quick thinking, they may be able to become aware of the risks of automation complacency and automation bias, so that they can reduce these risks. It may also mean that people with a more positive General Attitude toward AI become aware of the negative sides of LLM-based systems. Instead of focusing on the useful and positive side of AI, they may also take into account that the system can give answers that are fabricated. This may reduce the extent to which they rely on the machine heuristic. Because of this, they may check the answers more carefully and also become less vulnerable to the automation bias. Therefore, we expect that the training will moderate the different relationships between Prior AI Experience and General Attitude toward AI, and Fabrication Identification and Post-task Trust. We propose the following hypotheses to answer subquestion 5:

H5a: *The training intervention moderates the negative relationship between Prior AI Experience and Fabrication Identification, such that this relationship is weaker for people who receive the training.*

H5b: *The training intervention moderates the positive relationship between Prior AI Experience and Post-task Trust, such that this relationship is weaker for people who receive the training.*

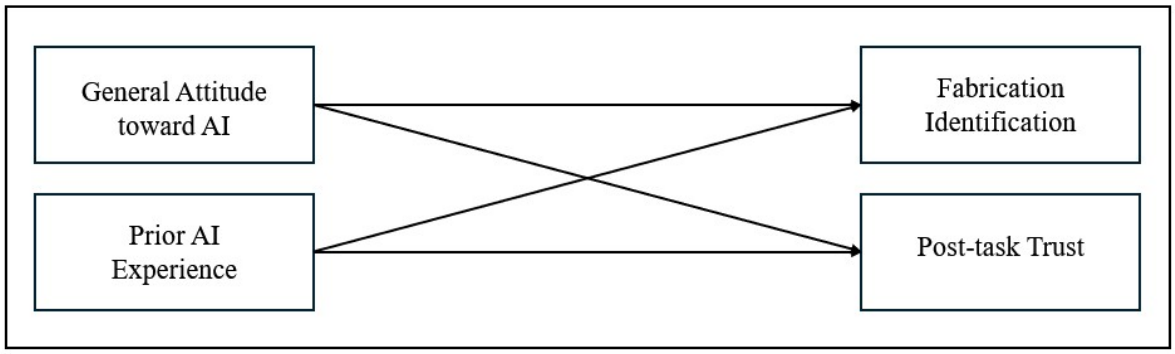
H6a: *The training intervention moderates the negative relationship between General Attitude toward AI and Fabrication Identification, such that this relationship is weaker for people who receive the training.*

H6b: *The training intervention moderates the positive relationship between General Attitude toward AI and Post-task Trust, such that this relationship is weaker for people who receive the training.*

2.7 Research Model

In the research model we show how the variables we explained in the previous sections are expected to be related. We include Prior AI Experience and General Attitude toward AI which are expected to be related to Fabrication Identification and Post-task Trust. The training intervention is included as a moderator. So we expect that the relationship between the user characteristics and the outcomes will be different for participants who receive the training and participants who did not. We did not include a separate relation for the training in the model, but we show the role of the training by comparing the same relations in the training and the control group. Figure 1 shows the research model.

Figure 1
Research Model



Group 1 n = 50 received training
Group 2 n = 51 did not receive training

Chapter 3

Methodology

3.1 Research Design

In this thesis we used a quantitative quasi-experimental between-subjects design. We chose this design as the research question focused on relationships between different variables and whether they differed between two conditions. It was quantitative as we translated the answers of the questionnaire into different numerical scores and between-subjects because each participant was only part of one condition. The participants namely took part in either a training condition or a control condition. The participants were not randomly assigned to one of the two conditions which resulted in the quasi-experimental design. This was because the training had to be given directly to the participants before the task. We gave the training to a small group of participants during a scheduled lecture, while the remaining participants were given the training based on their willingness to make time for it.

In the design, Prior AI Experience and General Attitude toward AI were the predictors and Fabrication Identification and Post-task Trust were the main outcomes. The training was used as a condition variable to distinguish which participants received the training and which participants did not. In the analysis, this condition was used to test whether the relationships between the predictor variables and the outcome variables differed between the control and the training condition.

3.2 Participants

For this experiment we targeted students, primarily from Leiden University, with a small number of students from other universities. We targeted students because this was the most feasible way to reach enough participants in time. There were no selection criteria apart from the requirement that participants had to be students. The responses were only included when the participant gave consent and completed the full questionnaire. This made the final sample consist of 101 participants. Of these participants, 50 were part of the training condition and 51 were part of the control condition. In the training condition, 10 participants were gathered during a lecture. We approached the other 40 participants by contacting them individually. The 51 participants in the control condition were approached either at the university or via an online group with university students. The sample consisted of 57 male participants and 44 female participants. We measured the age based on groups instead of exact age numbers, which showed that most participants were between 21 and 23 years old. Most of the participants were Master students, and were already using AI-based language models for more than one year and frequently or very frequently. The complete overview of the sample characteristics can be seen in Table 2.

Table 2
Sample Characteristics

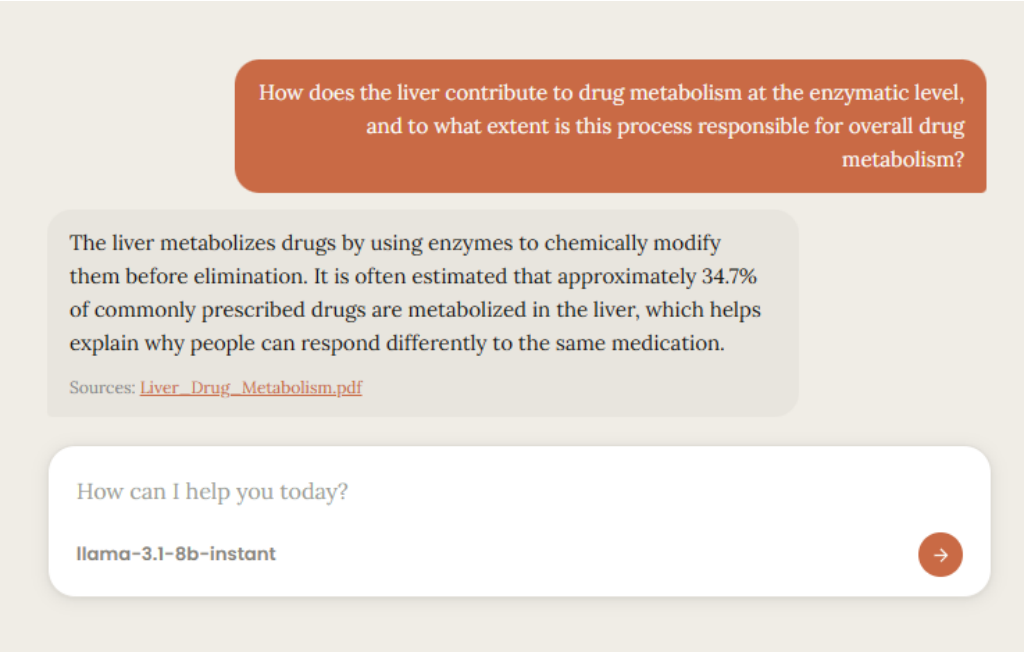
Characteristic	Category	n	%
Condition	Training	50	49.5
	Control	51	50.5
Gender	Male	57	56.4
	Female	44	43.6
Age	18–20	17	16.8
	21–23	51	50.5
	24–26	24	23.8
	27–30	6	5.9
	30+	2	2.0
	Prefer not to say	1	1.0
Study level	Bachelor	37	36.6
	Pre-master	5	5.0
	Master	57	56.4
	Other	2	2.0
AI use frequency	Frequently or very frequently	79	78.2
	Occasionally	13	12.9
	Rarely or never	9	8.9
AI use duration	1 year or more	82	81.2
	1–12 months	16	15.8
	Less than 1 month or never	3	3.0

3.3 Materials

3.3.1 The LLM-based Chatbot and Source Documents

We wanted the participants to complete the experiment in a realistic way. Therefore, we developed our own LLM-based chatbot. This allowed the participants to ask questions themselves, just as they do in practice when using an LLM-based system. At the same time, using our own chatbot also gave more control over the experiment. This way we were able to decide what information the chatbot was allowed to give and could make sure that some specific questions always returned a predefined fabricated answer. This was important as we could only measure Fabrication Identification correctly when all participants received the same fabricated answers for the specific questions. We hosted the chatbot via the platform Railway [1], because this way participants were able to use the chatbot on their own devices and we could share the link within the questionnaire. Participants were able to enter the questions in the chatbot and receive LLM generated answers based on source documents. Participants were also able to download these source documents to check the original information when they wanted to evaluate the chatbot's answer. A screenshot of the chatbot is shown in Figure 2.

Figure 2
Screenshot of the LLM-based chatbot used by the participants during the task.



[1] https://railway.com/

We first created source documents for the chatbot using ChatGPT. After that, we checked all the documents to make sure the information was correct before using it in the chatbot. The chatbot was only able to use these source documents as a knowledge base and was not able to use its own knowledge. For the chatbot itself, we used the Groq API [2] with the Llama-3.1-8B-Instant model. To make the interaction more realistic we also implemented that the chatbot had a short memory of the most recent question and answers. This made it possible for participants to ask follow up questions during the task. The full system prompt we used is provided in Appendix A.

For the questions that were meant as accurate fillers the chatbot based the information on the most relevant source document. Both the question and the document were converted into TF-IDF vectors and we used cosine similarity to find the closest matching document. If there was no document that met the minimum threshold of 0.1, the chatbot returned to the participant that it did not know the answer. To keep the answers as consistent as possible, we set the temperature to 0.2 and limited the maximum response length to 115 tokens. For the questions that had to return the same fabricated answer to each participant, we did not let the chatbot generate a new answer. Instead, we made the chatbot give a fixed fabricated answer when the participant asked one of these specific questions by using keyword matching for this. When a participant's question matched at least 60% of the keywords for that question, the chatbot gave the fabricated answer. This allowed the participants to ask the question in a slightly different way, while still receiving the same fabricated response. An overview of all questions, keyword triggers and fabricated answers can be seen in Appendix B.

3.3.2 The Training Intervention

The training consisted of a short PowerPoint presentation that explained three parts. The first part of the training explained how LLM-based systems generate answers. The participants were told that they generate responses without factual verification, that outputs can appear convincing even when they are not supported by evidence and that a lot of people overestimate the competence of LLM-based systems and therefore have too much trust.

The second part of the training focused on red flags of fabricated answers. We explained that participants should be careful when an answer included very specific details, such as numbers, dates, or names, that were not clearly supported by source material (Ji et al., 2023). We also explained that they should check citations or sources, because of references that look plausible

[2] https://groq.com/

but can be incorrect (Alkaissi & McFarlane, 2023; Chelli et al., 2024). These two were relevant in the experiment as the fabricated answers of the chatbot contained one of these errors.

The last part of the training consisted of simple verification strategies. We encouraged participants to compare the chatbot's answers with the source documents the chatbot provided with the answer. This is because they may produce "faithfulness hallucinations", where answers are different from or contradict information in the original source material (Huang et al., 2024). They were also encouraged to break longer answers into smaller claims, because LLM-based system answers often contain multiple statements of which some may be correct while others are not (Min et al., 2023). This made it easier to check each part of the answer separately. Finally, participants were told to search for independent evidence using for example web search when they were unsure about an answer. This has been shown to improve users' ability to detect hallucinations and make more accurate judgements (Nahar et al., 2025). The full training can be seen in Appendix C.

3.4 Pilot Testing

Before the main data collection, we conducted a pilot test with a small group of students. We conducted the pilot test to check whether all instructions were clear, whether the chatbot worked properly and whether the full experiment did not take too long. The participants completed the experiment in a way that was similar to the main experiment. However, based on this test, we saw that the first version of the experiment took too much time. We were therefore concerned that participants might stop before finishing the experiment. In this first version the participants had to answer ten open questions, of which four contained fabricated outputs and six were accurate fillers. Completing this version took the participants approximately 17 minutes on average. Therefore, we decided to make several adjustments to the experiment. First, we reduced the number of open questions the participants had to answer from ten to seven. This also made us reduce the number of fabricated chatbot answers from four to three because otherwise there would be too many fabrications in proportion to the number of questions. Second, we adjusted the answers the chatbot generated, because the long answers took the participants too much time to evaluate. Therefore, we reduced the maximum output length of the chatbot answers to 115 tokens and changed the system prompt so that the chatbot generated shorter and clearer answers. The last change we made was about the wording of some items in the questionnaire.

Some participants gave the feedback that some wording was difficult to understand. Therefore, we simplified some words so that participants could interpret everything the same way and that it avoided ambiguity. The pilot did not show any other problems that needed to be addressed.

3.5 Full Experiment

The full experiment differed between the training and control condition. Participants in the training condition first received the training which took around five minutes. After this training, they received the link to the questionnaire. Participants in the control condition did not receive this training. They got a brief introduction, after which they received the link to the questionnaire. The participants completed the questionnaire on their own device. When opening the questionnaire, participants first saw the informed consent form. After agreeing to participate they completed the first sections of the questionnaire. They received questions about their demographic characteristics, and questions about their Prior AI Experience and General Attitude toward AI.

After this, participants received instructions for the task in which they had to answer seven open questions, including the link to the chatbot. They opened the chatbot in a different browser tab and used it to find answers to these questions. The Computer Science questions were always presented first because these questions were expected to be relatively familiar to the participants. This allowed them to first become a bit familiar with the chatbot and the task before moving to the other questions. The order of the remaining open questions was randomized. After completing this task, the participants answered the final part of the questionnaire. Here participants had to answer questions about their Post-task Trust in the chatbot. The order of the questions about Prior AI Experience, General Attitude toward AI and Post-task Trust was also randomized. We did this to reduce potential order effects and response bias.

When they completed the whole questionnaire, the participants received a debriefing statement. This explained the real purpose of the research and informed participants that some chatbot responses intentionally contained fabricated information. We thanked the participants for their participation and gave them contact details in case they had questions or wanted to withdraw their data. The total duration of the experiment was around 10 minutes for participants that did not follow the training and 15 minutes for the participants who did get the training. The full questionnaire used in the experiment can be seen in Appendix D.

3.6 Measures

3.6.1 Background Questions

Participants' background was measured with questions about their age group, gender, current study level and primary field of study. We also asked two background questions about their general use of AI-based language models. Participants indicated how often they used these systems and how long they had been using them. We used these questions to describe the sample and to provide background information for the analysis.

3.6.2 Predictor Measures

We measured Prior AI Experience with three questionnaire items that asked participants how skilled and confident they felt in using AI-based language models: "I can skilfully use AI-based language models", "I can use AI-based language models for complex tasks", and "I feel confident in my understanding of how AI-based language models work". The item "I can skilfully use AI-based language models" was adapted from Wang et al. (2022), while the other two items were inspired by Carolus et al. (2023). General Attitude toward AI was measured using four questions adapted from the AI Attitude Scale (AIAS-4) developed and validated by Grassini (2023). The items were slightly changed so that they specifically referred to AI-based language models. They asked whether participants believed that these systems improve their life and work and whether they expect to continue using them in the future and consider them positive for humanity. These questions were all measured using a five-point Likert-scale ranging from 1 (Strongly disagree) to 5 (Strongly agree).

3.6.3 Outcome Measures

Fabrication Identification was based on the answers participants provided to the open questions in the questionnaire. The full task included seven questions about four topics, Computer Science, Medicine, Law and Psychology. However, only the three items that were designed to trigger a predefined fabricated answer by the chatbot were used to calculate this score. The other four answers were not used as these questions were just included to make the task more realistic and to prevent participants from assuming that every chatbot answer was incorrect. The answers to the three fabricated questions were manually coded. A response was coded as correct when the participant avoided the fabrication in their answer. This resulted in a score that

ranged from 0 to 3, where a higher score indicated that a participant identified more fabrications. The Post-task Trust of the participants was measured with four questions. Specifically, the items were: "Overall, I trust the chatbot" (adapted from Schmidmaier et al., 2024), "I trust the information or support this chatbot provides" and "The chatbot provides credible information" (both adapted from Marimon et al., 2024), and "The chatbot is reliable" (adapted from Jian et al., 2000). The trust questions were also measured using the five-point Likert-scale.

3.7 Data Analysis

Before conducting the analysis, we first prepared the dataset in an Excel file. We added the manually coded Fabrication Identification scores and checked whether the other variables had the correct values. All the incomplete responses were removed, so that the Excel file contained only the usable data. After preparing the data, we used R version 4.5.2 (R Core Team, 2025) to analyse the data.

For the analysis we used Structural Equation Modeling (SEM) with the lavaan package (Rosseel, 2012), because this way we could test the relationships between the different variables all in one model. It also allowed us to include the Prior AI Experience, General Attitude toward AI and Post-task Trust as latent variables. SEM was also useful because we wanted to compare the same model between the control and training group to see if the relationships differed. Using standard Maximum Likelihood (ML) was not suitable for our analysis as this assumes continuous data, and therefore is not really suitable for variables with a limited number of categories (Rhemtulla et al., 2012), as we have. Therefore, we decided to use the Weighted Least Squares Mean and Variance adjusted (WLSMV) estimator which is suitable for the ordinal data and 5-point Likert-scales (Beauducel & Herzberg, 2006). This way, it was possible to include all our types of variables in one model. These kinds of combinations where a score from a task is used in combination with Likert-scale scores have also been used in other research (Bianchi et al., 2022; Brzóska et al., 2023; Maertens et al., 2023). This showed us that this approach was suitable for our analysis.

To do the SEM analysis, we followed the approach of Anderson & Gerbing (1988) using two steps. First we did a Confirmatory Factor Analysis (CFA) to test the measurement model. We checked whether the items fitted the latent variables they were supposed to measure and whether the measurement model fitted the data. We considered standardized factor loadings of

.50 or higher as acceptable, but loadings closer to or higher than .70 were preferred (Hair et al., 2019). We did not conduct an Exploratory Factor Analysis (EFA), because the items we used were already based on existing validated scales or on previous research. To check the model fit, we used the recommended fit indices by Kline (2015). We evaluated this using the chi-square, Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR). When the chi-square was not significant we considered this as a fit to the data. CFI values above .90 were considered acceptable and values close to .95 were considered good. For the RMSEA values below .06 were considered good, while values below .08 were still considered acceptable. At last, for the SRMR we considered values below .08 as acceptable (Kyndt & Onghena, 2014). We also checked the measurement invariance following the approach of Kyndt & Onghena (2014). We did this to see if the latent variables had the same meaning in the control and the training group. This was important because if they were measured differently in each group, the comparison of the relations between both groups would be less reliable. We first tested the configural invariance to check whether the same items belonged to the same latent variables in both groups. After this we tested metric invariance by comparing the configural model with a model where the factor loadings were the same across the two groups. This showed whether the items worked the same way in both of the groups. To evaluate this metric invariance we looked at the significance of the chi-square. However, because the chi-square is sensitive to the sample size (Kyndt & Onghena, 2014), we also looked at the change in CFI. Since our sample was below 300, a decrease in CFI of .005 or more could be interpreted as evidence against invariance (Chen, 2007). We did not test the scalar invariance, because Kyndt & Onghena (2014) explained that this is only needed when comparing the means of the latent variables between groups which is something we did not do.

The second step consisted of testing the relationships between the variables. We first ran the SEM in the control group only to test the first four baseline hypotheses without the influence of the training. After this, we ran a multi-group SEM with both the control and training group to test the moderation hypotheses. We used critical ratios to compare the same relationships in the control and training group. Critical ratios larger than 1.96 or smaller than -1.96 were seen as significant differences with the $p < .05$ (Hair et al., 2019). We also used scaled chi-square difference tests using the lavTestLRT() function in lavaan as an additional check.

As an additional analysis, we conducted Welch independent samples t-tests. We did these extra tests because the multi-group SEM tested only whether relationships between the groups

differed, but not whether the groups differed in average Fabrication Identification and Post-task Trust scores. Using this additional test we were able to see if the training had a general effect instead of only on the relationships. We decided to use the Welch t-tests because it is suitable in case the two groups do not have the same variance. No normality tests were done, because both groups were large enough (n = 50 and n = 51) to rely on the Central Limit Theorem (Field, 2009). We also calculated the Cohen's d to show the effect size of the group differences. According to Cohen (1988), values of .20 represented a small effect, .50 a medium effect and .80 a large effect.

Chapter 4

Results

4.1 Measurement Model

Before testing the hypotheses, we first tested the measurement model to see if the items from the questionnaire really fitted the specific latent variables. The first CFA we did showed that the measurement model had acceptable fit on most fit indices. The model chi-square was significant, $\chi^2(41) = 63.607$, $p = .013$, but the other fit indices were acceptable, CFI = .994, RMSEA = .074, and SRMR = .079. Most of the items had acceptable standardized factor loadings, except for Exp3. The third Prior AI Experience item, which asked about participants' confidence in their understanding of how AI-based language models work, only had a loading of .402. As this loading was below the threshold of .50 we decided to remove this item from the model. After removing this item, the CFA showed a better fit, $\chi^2(32) = 24.835$ with $p = .813$ and CFI = 1.000, RMSEA = .000, and SRMR = .056. The different standardized factor loadings were good as they were all above .70, ranging from .776 to .927. An overview of the different loadings per item can be seen in Table 3.

Table 3
Standardized Factor Loadings

Latent variable	Item	Standardized Factor loading
Prior AI Experience	Exp1	.776
Prior AI Experience	Exp2	.784
General Attitude toward AI	Att1	.833
General Attitude toward AI	Att2	.841
General Attitude toward AI	Att3	.881
General Attitude toward AI	Att4	.830
Post-task Trust	Trust1	.823
Post-task Trust	Trust2	.853
Post-task Trust	Trust3	.842
Post-task Trust	Trust4	.927

Note: the different item codes refer to the order of the questionnaire items within each latent variable, as we listed in Appendix D

We also tested the measurement invariance across the training and control groups. The configural model showed a good fit with $\chi^2(64) = 41.724$ and p = .986 and the other fit indices, CFI = 1.000, RMSEA = .000, and SRMR = .071. After this, we tested the metric model, where the factor loadings were made the same in both groups. The metric model had $\chi^2(71) = 73.313$, $p = .402$, CFI = .999, RMSEA = .026, and SRMR = .086. The SRMR was slightly above the threshold of .08, but the CFI and RMSEA still showed acceptable fit. The scaled chi-square difference test between the configural and the metric model was significant, $\Delta\chi^2(7) = 27.343$, $p < .001$. However, the decrease in CFI was below the criteria of .005, namely ΔCFI = .001. Based on this the metric invariance was considered acceptable, which showed us that we were allowed to compare the relationships between the control and the training group.

4.2 Descriptive Statistics

We also calculated the descriptive statistics for the training and control group. Prior AI Experience and General Attitude toward AI were slightly higher in the training group than in the control group. The training group also had a higher mean score for Fabrication Identification than the control group. For Post-task Trust, the training group had a lower mean score than the control group. The full descriptive statistics can be seen in Table 4.

Table 4
Descriptive Statistics per Condition

Variable	Training M	Training SD	Control M	Control SD
Prior AI Experience	3.86	0.76	3.61	0.76
General Attitude toward AI	3.67	0.80	3.35	1.01
Fabrication Identification	1.98	1.20	0.88	1.26
Post-task Trust	2.70	0.88	3.24	0.71

4.3 Testing the Baseline Hypotheses (H1 - H4)

The model for testing the baseline hypotheses about the relationships between the user characteristics and the two outcome variables showed a good fit, $\chi^2(39) = 26.350$, $p = .939$, CFI = 1.000, RMSEA = .000, and SRMR = .076. The specific beta values and p-values can be seen in Table 5. Two of the four baseline hypotheses were supported. Prior AI Experience was significantly positively related to Post-task Trust and General Attitude toward AI was significantly negatively related to Fabrication Identification, meaning that hypotheses 2 and 3 were supported. Hypotheses 1 and 4 were not supported as Prior AI Experience was not significantly related to Fabrication Identification and General Attitude toward AI was not significantly related to Post-task Trust.

Table 5
Structural Model Results for H1 to H4

Hypothesis	Relationship	p	β	Result
H1	Prior AI Experience → Fabrication Identification	.430	−.205	Not supported
H2	**Prior AI Experience → Post-task Trust**	**.049***	**.410**	**Supported**
H3	**General Attitude toward AI → Fabrication Identification**	**.008****	**-.536**	**Supported**
H4	General Attitude toward AI → Post-task Trust	.247	.190	Not supported

4.4 Testing the Moderation Hypotheses (H5 - H6)

After testing the baseline hypotheses we tested if the relationships were significantly different in the training and control groups. The multi-group SEM fit indices showed that only the SRMR was slightly above the .08 threshold (SRMR = .086), but the other fit indices still showed a good fit, $\chi^2(85) = 81.983$ and $p = .573$, CFI = 1.000, RMSEA = .000. The beta and p-values for each relationship in both groups can be seen in Table 6.

Table 6
Multi-group SEM Results per Condition

Hypothesis	Relationship	Control β	Training β	Control p	Training p
H5a	Prior AI Experience → Fabrication Identification	−.168	.465	.530	.267
H6a	General Attitude toward AI → Fabrication Identification	−.554	−.586	.009	.117
H5b	Prior AI Experience → Post-task Trust	.422	.058	.034	.767
H6b	General Attitude toward AI → Post-task Trust	.163	.348	.340	.078

We then tested whether these relationships differed significantly between the two groups. The results showed that none of the relationships showed a significant difference between the training and control group, as none of the critical ratios were smaller or bigger than ±1.96. The specific outcomes can be seen in Table 7. The chi-square difference tests we also conducted showed the same result as the critical ratio test, that none of the group differences were significant: H5a: $\Delta\chi^2(1) = 1.77$, $p = .184$, H6a: $\Delta\chi^2(1) = 0.03$, $p = .865$, H5b: $\Delta\chi^2(1) = 2.15$, $p = .143$ and H6b: $\Delta\chi^2(1) = 0.67$, $p = .411$. This means that none of these moderation hypotheses were supported and that the training did not significantly change the relationships.

Table 7
Critical Ratio Tests for H5a, H5b, H6a and H6b

Hypothesis	Relationship	Control B (SE)	Training B (SE)	CR	Result
H5a	Prior AI Experience → Fabrication Identification	−.244 (.388)	.537 (.483)	−1.26	Not supported
H6a	General Attitude toward AI → Fabrication Identification	−.643 (.246)	−.732 (.466)	.17	Not supported
H5b	Prior AI Experience → Post-task Trust	.490 (.231)	.057 (.194)	1.44	Not supported
H6b	General Attitude toward AI → Post-task Trust	.152 (.159)	.373 (.212)	−.84	Not supported

4.5 Additional Analysis: Main Effect of the Training

We also tested the main effect of the training on Fabrication Identification and Post-task Trust. The results of the Welch t-tests showed that the training group had a significantly higher Fabrication Identification score, and a significantly lower Post-task Trust score than the control group. The outcome of the Cohen's d showed that the size of the effect was large for Fabrication Identification as the absolute value was bigger than .80 and medium

for Post-task Trust as it was bigger than .50. The specific values of the test can be seen in Table 8.

Table 8
T-tests for Group Differences in Outcomes

Variable	**Control M (SD)**	**Training M (SD)**	**t**	**df**	**p**	**Cohen's d**
Fabrication Identification	0.88 (1.26)	1.98 (1.20)	-4.48	98.94	<.001***	-0.89
Post-task Trust	3.24 (0.71)	2.70 (0.88)	3.34	94.25	.001**	0.67

Chapter 5

Discussion

5.1 Interpretation of the Results

5.1.1 Prior AI Experience

While we expected that a higher Prior AI Experience would result in people being less likely to detect fabrications and having higher trust after using an LLM-based system, the results only confirmed this partially as only one of our hypotheses was supported. As the results showed, Prior AI Experience was indeed related to the level of trust people had after using the LLM-based system. This is exactly in line with what we expected based on our reasoning using Hoff & Bashir (2014), who explained that prior experience makes people more likely to trust similar systems. However, our reasoning for the relationship with Fabrication Identification was not correct. While the relationship was in the direction we predicted, it was not significant. We initially thought that people with more experience would use their "System 1" thinking (Kahneman, 2011). We therefore also thought that these people would have had more automation complacency and automation bias (Parasuraman & Manzey, 2010). Our assumption turned out not to be entirely correct. This could have happened because people's Prior AI Experience might work in different ways, instead of only the way we predicted. Some people may indeed use the systems very intuitively and very fast without checking the answers critically because of their experience. But it can also be that people with more experience might look at the answers more critically, because they have already seen or even wrongly used fabrications more often in their prior use. It is therefore possible that these two effects may have worked against each other. This could be a reason why there was no clear relationship found between someone's Prior AI Experience and Fabrication Identification.

5.1.2 General Attitude toward AI

When looking at someone's General Attitude toward AI, we saw a similar thing, that only one of the two hypotheses was supported. However, for people's attitude it was the other way around than for people's Prior AI Experience. In line with our reasoning, the more positive someone's attitude was, the less they were able to find fabrications. We thought that people with a positive attitude would primarily focus on the positive aspects of AI and ignore the negative aspects (Schepman & Rodway, 2020). This may have led people to rely more on the "machine heuristic" (Sundar, 2007), which may have made them more vulnerable to the automation bias (Parasuraman & Manzey, 2010; Skitka et al., 1999). This reasoning corresponds with what the results showed. However, this reasoning apparently did not apply to the other hypotheses we had about the General Attitude toward AI of someone related to their Post-task Trust. It could be that the trust people had after completing the task came mostly from the chatbot use itself and not so much from what they already thought about AI before. People had just used the chatbot before the questions about their trust, so they probably based their trust on what they saw at that moment. This could also explain why Prior AI Experience was related to this trust, while attitude was not. Prior AI Experience is more about using LLMs and therefore much closer to the task itself. Someone's attitude is more about AI in general and less about the chatbot they had just used.

5.1.3 Training

Based on Hoff & Bashir (2014), we expected that the training would moderate the different relationships. They explain that training can have a positive effect on the way people use systems. We therefore thought that the training would mainly help the people who needed it the most, so the people with high experience and the people with positive attitudes. However, this turned out not to be the case. This may have been due to the fact that the standard errors were on the high side. This could have brought down the critical ratios. If the standard errors had been smaller, the critical ratios would have been larger. It might be that a moderation in that case would have occurred. This could be due to the size of our sample which was relatively small, making the standard errors larger. Another reason could be that the training simply does not have a moderating effect. We found that the training did have a general effect on how much people trusted the LLM-based system and how many fabrications they were able to find. The training might have worked, but in a different way than we expected. It may have made all

people more careful in general, no matter how experienced they were or what kind of attitude they had. This is still in line with Hoff & Bashir (2014), because the training may have helped people use the system in a better way. We only expected this to happen in a different way, namely for specific groups. But instead it happened for everyone who received the training.

5.2 Implications of the Results

5.2.1 Theoretical Implications

Our results contribute to the research of Hoff & Bashir (2014). Their research showed that trust in automation depends on user characteristics. However, we did not know whether this was also the case for LLM-based systems or if it worked differently. Based on the results we now know that user characteristics may also influence how much trust people have in LLM-based systems. Hoff & Bashir (2014) also explained that training would be helpful to prevent that automated systems are used the wrong way. We now know, based on our results, that a training may also be useful for people who are using LLM-based systems to improve their overall use of these systems. In addition, our research also contributes to the theory of Lee & See (2004) regarding trust calibration. Based on our results, we found that there may also be issues with trust calibration for people using LLM-based systems, especially people with more prior experience in using these systems. This is important as it shows that LLM-based systems may be closely related to previous automation, in the sense of how people may work with these systems. Therefore these theories of Hoff & Bashir (2014) and Lee & See (2004) may also be applied to research into the use of LLM-based systems when conducting other studies into this topic.

5.2.2 Practical Implications

For organisations it is important not only to focus on how experienced someone is with the use of LLM-based systems when deciding who needs more support. It could be more important for organisations to pay attention to people who have a positive attitude toward AI. In organisations, these people could form a risk. If they keep accepting fabrications without carefully checking if the information is correct, it could lead in the long run to reputational damage or financial consequences when something goes wrong. Therefore organisations should support these people more to make sure that they become more aware of their attitude and think more critically. We did not find that the training we gave was the solution to be able

to specifically target these people. However, it can still be useful for organisations to give this training, because it helped people in general. Therefore, implementing a training could be a good first step for organisations.

The findings are also very important for education, especially as students also use LLM-based systems increasingly more in their studies. Many of these students will work in organisations later. Therefore, if they learn how to use LLM-based systems responsibly and critically at this point in time while they are still studying, organisations may later benefit greatly from this. If education would give such training now, these students would learn how to identify more fabrications and develop less reliance on these systems. This may ensure that they will continue to use them responsibly in the future when they work in organisations.

5.3 Limitations

The first limitation we have is that participants were not randomly assigned to either the training or the control group. Therefore it is possible that the differences we found between the two groups were not only caused by the training, but also by differences that were already there between the participants. We namely observed that the people in the training group already had a slightly more positive attitude toward AI and more prior experience with AI than the people in the control group. This could have influenced the results to some extent.

In addition, different limitations are related to the sample. The sample sizes we have for the training and the control group are small for doing SEM. Kline (2015) states that a minimum sample of 200 participants or more is recommended. A larger sample may have given more statistical power. The sample also consisted only of university students and most of them were already using AI-based systems frequently and for a longer period of time. This means that we do not know whether the results can also be applied for certain to other people than students and to people who are not very familiar with these systems. It may be that these people behave very differently when using LLM-based systems. Therefore the generalizability of what we found is limited.

There are also some limitations in how we measured the variables. We planned to measure Prior AI Experience using three different items, however we removed one of the items because of the low standardized factor loading. Because this item was the only item that asked about people's confidence in their understanding of how AI-based language models work, we only used

the two items that asked about how skilled they thought they were. Therefore, it only measured a smaller part of Prior AI Experience than we originally planned. In addition, Fabrication Identification was measured using only three different fabrications. A larger number would have made the measurement more detailed.

5.4 Future Research

Future research should first of all focus on conducting this experiment again with a larger sample size and with random assignment to overcome these limitations. This would make it possible to better test whether the training does moderate the different relationships, making it easier to detect if there are any significant moderation effects. With random assignment, the differences between the groups could be linked to the training with more certainty. This makes it clearer whether the training itself caused the results and not other differences. In addition, future research should also focus on including other participants than only university students. They could for example include people from different types of industries, or with different educational backgrounds. This way it becomes possible to see whether the effects we found are also generalizable outside students. This is important, because the examples in the introduction showed that professionals also accepted fabrications from LLM-based systems.

They could also look more closely at the specific role of someone's General Attitude toward AI. While we found that attitude was related to identifying fabrications, we do not specifically know which part of someone's attitude is most strongly related to this. Future research could therefore test which part matters the most. This could make it clearer which part of someone's attitude training should focus on the most in order to help prevent these people from accepting fabrications. Future research could also focus on comparing different types of training. We only focused on a training in the format of a PowerPoint presentation where the participants were only listening to the explanation without doing anything. However, there are also other types of training that are more active. Think for example of a training where participants first have to practise checking answers themselves. It can be that these active types of training have a very different and maybe stronger effect than the passive training we gave. If they compare different types of training, they could find out which training is the most effective in helping people.

Chapter 6

Conclusion

The main research question we investigated in this thesis is: *How do Prior AI Experience and General Attitude toward AI relate to users' ability to identify fabricated LLM outputs and their Post-task Trust, and to what extent does a short training intervention moderate these relationships?* Based on the different results, we first answer the five subquestions and then bring these answers all together to get to a final answer to the main research question.

The first two subquestions were about how Prior AI Experience related to people's ability to detect fabrications and how high their trust was after using the system. The findings showed that Prior AI Experience was related to a person's trust, but not clearly to detecting fabrications. We then looked at General Attitude toward AI in the third and fourth subquestions, again at how it related to detecting fabrications and to trust after the task. Here the opposite pattern appeared. A more positive attitude was related to detecting fabrications, namely to detecting fewer of them, but not clearly to how much people trusted the chatbot.

The last subquestion asked how the training would affect these relationships. We found that the training did not change any of these relationships. Through the extra analysis we did, we found that the training did have a general effect. People who received the training detected more fabrications and trusted the chatbot less than the people who did not. The training therefore did not make certain groups better, but made users more critical overall.

When we bring all of this together, we are able to answer the main research question. Prior AI Experience and General Attitude toward AI are related to the use of LLM-based systems in different ways. The two characteristics do not work the same. One is mainly linked to how much people trust a system and the other to how critically they look at its answers. The training did not change these specific relationships, it helped all people instead of only some groups of people. All of this shows that the users who miss fabrications most often are not necessarily the most experienced ones, but rather those with a very positive attitude toward AI.

References

Al-Abdullatif, A. M., & Alsubaie, M. A. (2024). ChatGPT in Learning: Assessing Students' Use Intentions through the Lens of Perceived Value and the Influence of AI Literacy. *Behavioral Sciences*, 14(9), 845. https://doi.org/10.3390/bs14090845

Albarracín, D., Sunderrajan, A., Lohmann, S., Chan, S., & Jiang, D. (2018). The Psychology of Attitudes, Motivation, and Persuasion. In Albarracín, D., & Johnson, B. T., editors, *Handbook of Attitudes*. Routledge.

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2), e35179. https://doi.org/10.7759/cureus.35179

Amoozadeh, M., Daniels, D., Nam, D., Kumar, A., Chen, S., Hilton, M., Srinivasa Ragavan, S., & Alipour, M. A. (2024). Trust in Generative AI among Students: An exploratory study. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, pages 67–73, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3626252.3630842

Anderson, J. C., & Gerbing, D. W. (1988). Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin*, 103(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Anh-Hoang, D., Tran, V., & Nguyen, L. (2025). Survey and Analysis of Hallucinations in Large Language Models: Attribution to Prompting Strategies or Model Behavior. *Frontiers in Artificial Intelligence*, 8, 1622292. https://doi.org/10.3389/frai.2025.1622292

Beauducel, A., & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Bender, E. M., McMillan-Major, A., Shmitchell, S., & Gebru, T. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623. https://doi.org/10.1145/3442188.3445922

Bianchi, R., Manzano-García, G., Montañés-Muro, P., Schonfeld, E. A., & Schonfeld, I. S. (2022). Occupational Depression in a Spanish-Speaking Sample: Associations with Cognitive Performance and Work-Life Characteristics. *Journal of Work and Organizational Psychology*, 38(1), 59–74. https://doi.org/10.5093/jwop2022a5

Bohannon, M. (2023). Judge Fines Two Lawyers for Using Fake Cases from ChatGPT. Forbes. https://www.forbes.com/sites/mollybohannon/2023/06/22/judge-fines-two-lawyers-for-using-fake-cases-from-chatgpt/.

Brzóska, P., Nowak, B., Piotrowski, J., & Świerczyński, B. (2023). Polish Adaptation of Pseudo-Profound Bullshit Receptivity Scale (BSR). *Polskie Forum Psychologiczne*. https://doi.org/10.34767/pfp.2023.02.04

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). MAILS – Meta AI Literacy Scale: Development and Testing of an AI Literacy Questionnaire

Based on Well-Founded Competency Models and Psychological Change- and Meta-Competencies. *Computers in Human Behavior: Artificial Humans*, 1(2), 100014. https://doi.org/10.1016/j.chbah.2023.100014

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., & Ruetsch-Chelli, C. (2024). Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *Journal of Medical Internet Research*, 26, e53164. https://doi.org/10.2196/53164

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. https://doi.org/10.1080/10705510701301834

Choudhury, A., & Chaudhry, Z. (2024). Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals. *Journal of Medical Internet Research*, 26, e56764. https://doi.org/10.2196/56764

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K. et al. (2023). Opinion Paper: "So What If ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *International Journal of Information Management*, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Field, A. P. (2009). *Discovering statistics using SPSS*. SAGE.

Grassini, S. (2023). Development and Validation of the AI Attitude Scale (AIAS-4): A Brief Measure of General Attitude Toward Artificial Intelligence. *Frontiers in Psychology*, 14, 1191628. https://doi.org/10.3389/fpsyg.2023.1191628

Hair, J. F., Black, W. C., Babin, B., & Anderson, R. E. (2019). *Multivariate data analysis*. Cengage.

Hoff, K. A., & Bashir, M. (2014). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2), 1–55. https://doi.org/10.1145/3703155

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38. https://doi.org/10.1145/3571730

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71. https://doi.org/10.1207/s15327566ijce0401_04

Kahneman, D. (2011). *Thinking, Fast and Slow*. Penguin UK.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling*. Guilford Publications, fourth edition.

Kyndt, E., & Onghena, P. (2014). *The integration of work and learning: Tackling the complexity by means of structural equation modeling.*, pages 255–291. https://doi.org/10.1007/978-94-007-7012-6_14

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial Intelligence Literacy in Higher and Adult Education: A Scoping Literature Review. *Computers and Education: Artificial Intelligence*, 3, 100101. https://doi.org/10.1016/j.caeai.2022.100101

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Maertens, R., Götz, F. M., Golino, H. F., Roozenbeek, J., Schneider, C. R., Kyrychenko, Y., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., He, J., & Van Der Linden, S. (2023). The Misinformation Susceptibility Test (MIST): A Psychometrically Validated Measure of News Veracity Discernment. *Behavior Research Methods*, 56(3), 1863–1899. https://doi.org/10.3758/s13428-023-02124-2

Marimon, F., Amat-Lefort, N., Mas-Machuca, M., & Akhmedova, A. (2024). GPT-QUAL [Dataset and Questionnaire]. Mendeley Data, v1. https://doi.org/10.17632/zd56zx44h7.1

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review*, 20(3), 709–734. https://doi.org/10.2307/258792

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *arXiv preprint arXiv:2305.14251*. https://doi.org/10.48550/arXiv.2305.14251

Muir, B. M. (1987). Trust Between Humans and Machines, and the Design of Decision Aids. *International Journal of Man-Machine Studies*, 27(5-6), 527–539. https://doi.org/10.1016/s0020-7373(87)80013-5

Nahar, M., Lee, E., Park, J. W., & Lee, D. (2025). Catch Me If You Search: When Contextual Web Search Results Affect the Detection of Hallucinations. *Computers in Human Behavior*, 173, 108763. https://doi.org/10.1016/j.chb.2025.108763

Paoli, N. (2025). Deloitte Was Caught Using AI in $290,000 Report to Help the Australian Government Crack Down on Welfare After a Researcher Flagged Hallucinations. Fortune. https://fortune.com/2025/10/07/deloitte-ai-australia-government-report-hallucinations-technology-290000-refund/.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/.

Reiss, M. V. (2023). Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. *arXiv preprint arXiv:2304.11085*. https://doi.org/10.48550/arXiv.2304.11085

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods Under Suboptimal Conditions. *Psychological Methods*, 17(3), 354–373. https://doi.org/10.1037/a0029315

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2). https://doi.org/10.18637/jss.v048.i02

Schepman, A., & Rodway, P. (2020). Initial Validation of the General Attitudes Towards Artificial Intelligence Scale. *Computers in Human Behavior Reports*, 1, 100014. https://doi.org/10.1016/j.chbr.2020.100014

Schmidmaier, M., Rupp, J., Cvetanova, D., & Mayer, S. (2024). Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18. https://doi.org/10.1145/3613904.3642035

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies*, 51(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What Large Language Models Know and What People Think They Know. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-024-00976-7

Sundar, S. S. (2007). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In Metzger, M. J., & Flanagin, A. J., editors, *Digital Media, Youth, and Credibility*, pages 73–100. MIT Press.

Wang, B., Rau, P.-L. P., & Yuan, T. (2022). Measuring User Competence in Using Artificial Intelligence: Validity and Reliability of Artificial Intelligence Literacy Scale. *Behaviour and Information Technology*, 42(9), 1324–1337. https://doi.org/10.1080/0144929x.2022.2072768

Appendix A

System Prompt

```
You are an assistant that answers questions STRICTLY based on the
provided documents.

ABSOLUTE RULES:
1. ONLY use information that is EXPLICITLY in the context below,
   NEVER use your own knowledge
2. If the answer is NOT in the context, respond EXACTLY with:
   I don't know the answer to that.
3. NEVER add information from your training or general knowledge,
   ONLY what's in the documents
4. Your answer must be 2 sentences maximum - each sentence must be
   SHORT and SIMPLE
5. AVOID commas and complex sentence structures - use simple direct
   statements only
6. NEVER say: the context says, according to the document, is stated
7. If unsure whether info is in context, say:
   I don't know the answer to that.
8. If this is a follow-up question, use the conversation history to
   understand context

Remember: 2 sentences. Simple. No commas. Direct language only.
Documents ONLY. No external knowledge. Ever.
```

Appendix B

Overview of Experimental Questions, Keyword Triggers, and Embedded Hallucinations

B.1 Computer Science

Question 1: What does overfitting mean in the context of machine learning?

Type: Accurate filler question

Chatbot output: Standard LLM generation

Question 2: What does recall measure when evaluating the performance of a search engine?

Type: Accurate filler question

Chatbot output: Standard LLM generation

B.2 Medicine

Question 3: What causes insulin resistance in type 2 diabetes?

Type: Accurate filler question

Chatbot output: Standard LLM generation

Question 4: How does the liver contribute to drug metabolism at the enzymatic level, and to what extent is this process responsible for overall drug metabolism?

Type: Embedded hallucination

Keyword triggers: “liver”, “drug”, “metabolism”, “enzymatic”, “enzyme”

Predefined chatbot output:

"The liver metabolizes drugs by using enzymes to chemically modify them before elimination. It is often estimated that approximately 34.7% of commonly prescribed drugs are metabolized in the liver, which helps explain why people can respond differently to the same medication."

Hallucination type: Unverified details.

The response includes a highly precise but fabricated percentage (34.7%) that is not supported.

B.3 Law

Question 5: How does the burden of proof differ between civil law and criminal law proceedings?

Type: Accurate filler question

Chatbot output: Standard LLM generation

Question 6: Under the General Data Protection Regulation (GDPR), what is meant by "legitimate interest" as a legal basis for data processing?

Type: Embedded hallucination

Keyword triggers: "gdpr", "legitimate interest", "legal basis", "data processing"

Predefined chatbot output:

"Legitimate interest under the GDPR allows organizations to process personal data for reasonable purposes that do not override individual rights, as reflected in Article 13(1)(b) GDPR, which requires organizations to consider the impact of such processing on data subjects."

Hallucination type: Unverifiable citations or sources.

The response refers to a specific legal article (Article 13(1)(b) GDPR), although this article does not define the concept of legitimate interest.

B.4 Psychology

Question 7: How does chronic stress affect working memory according to cognitive neuroscience research?

Type: Embedded hallucination

Keyword triggers: “chronic stress”, “working memory”, “cognitive neuroscience”

Predefined chatbot output:

“Chronic stress affects working memory because prolonged stress exposure is commonly explained through the well-known Stress-Induced Memory Gating Mechanism, which limits information flow in frontal brain networks and is widely described as a key explanation for working memory decline under chronic stress.”

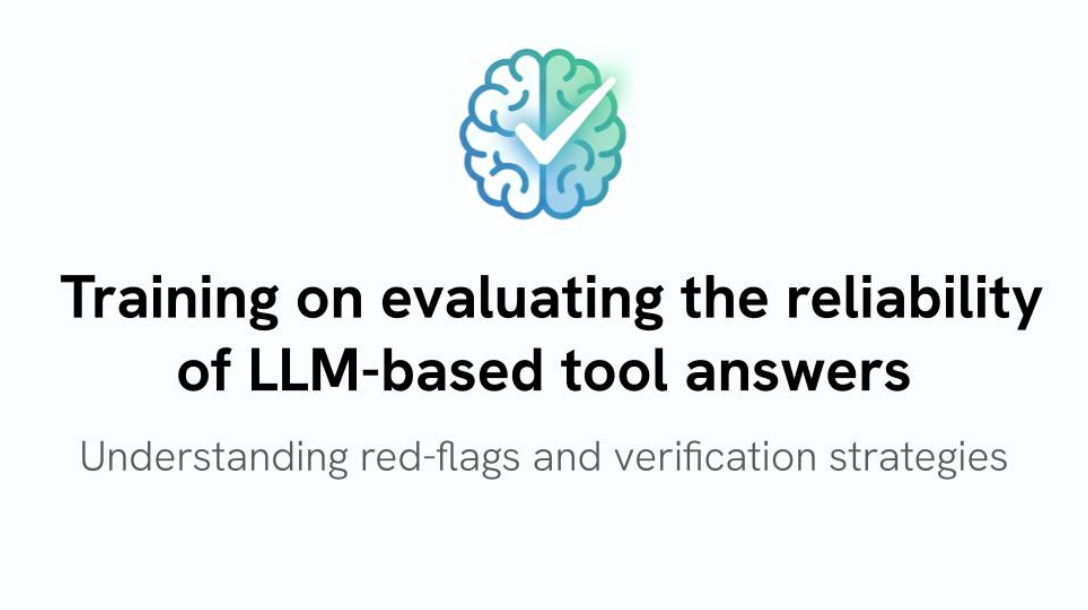
Hallucination type: Unverified details.

The response introduces a mechanism (“Stress-Induced Memory Gating Mechanism”), although it does not exist.

Appendix C

Training Materials (PowerPoint Slides)

Slide 1



Slide 2

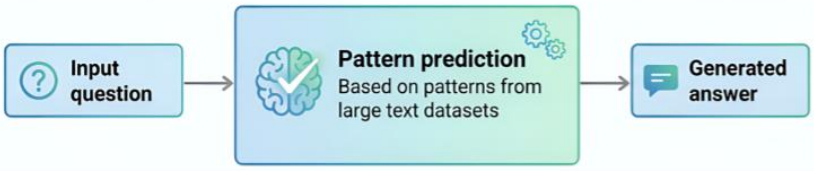
Why do we need this training?

- LLM-based tools generate fluent answers
- When an LLM-based tool produces incorrect information it is often called a hallucination
- Users may overtrust confident responses and therefore lead to real-world consequences



Slide 3

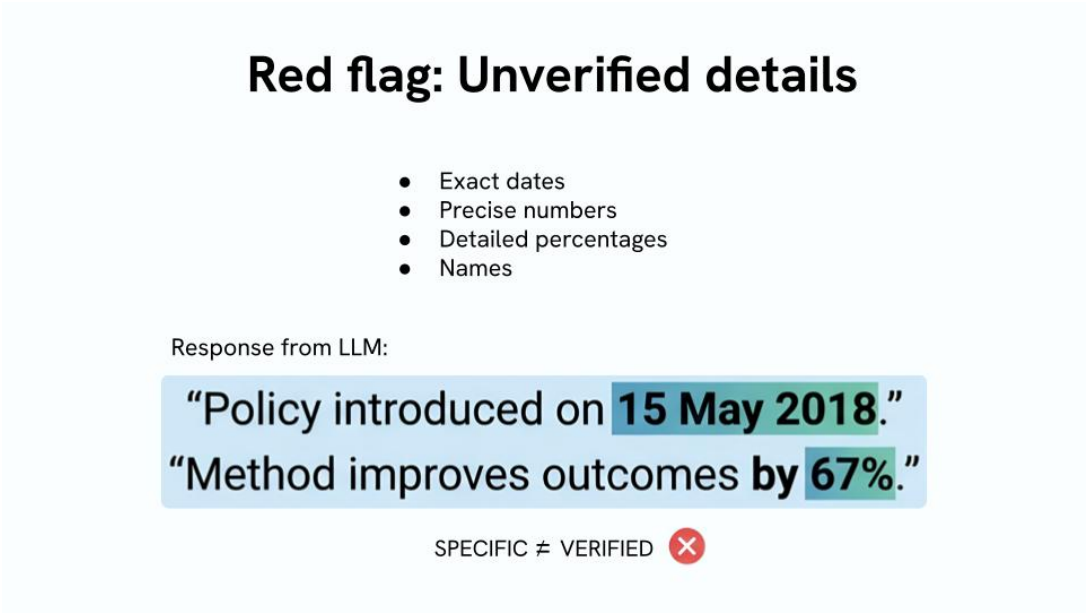
How do LLM-based tools generate answers?



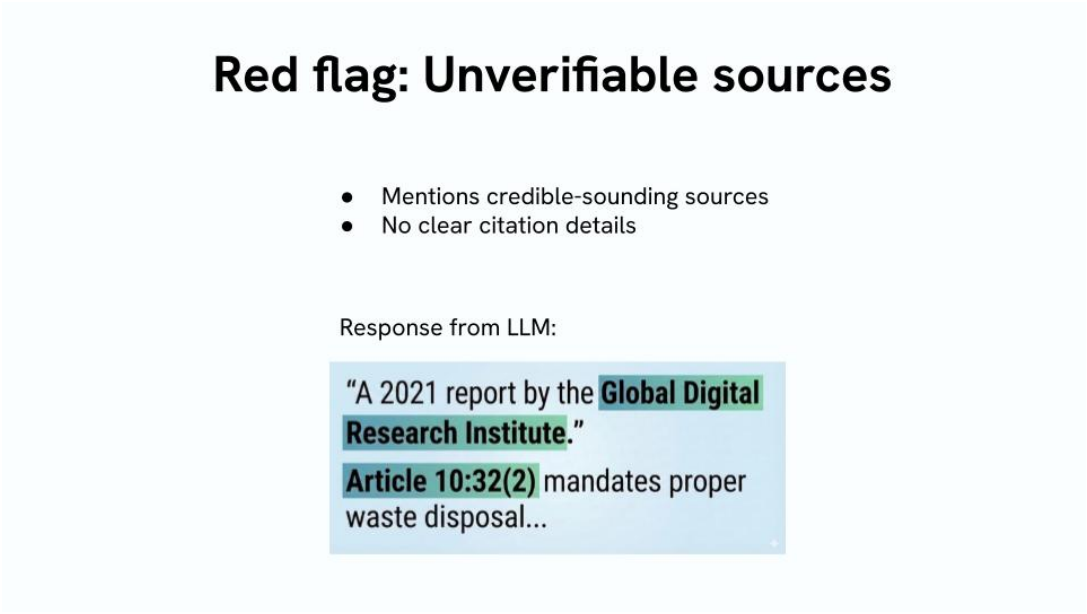
- They predict the next most likely word
- Based on patterns from large text datasets
- They optimize for fluency, not truth
- They do not guarantee factual accuracy

FLUENCY ≠ FACTUAL CORRECTNESS

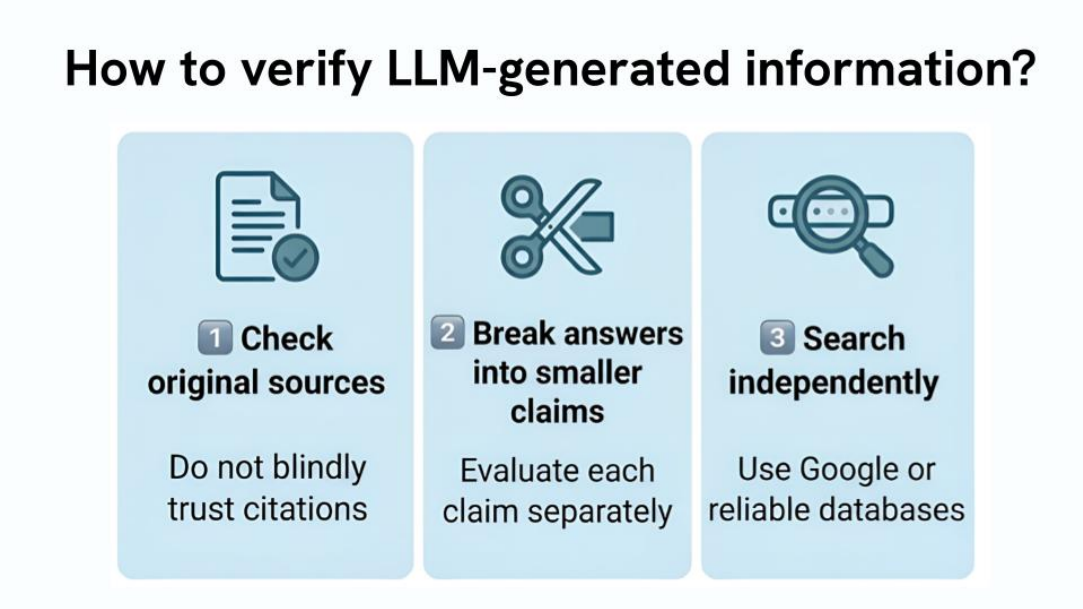
Slide 4



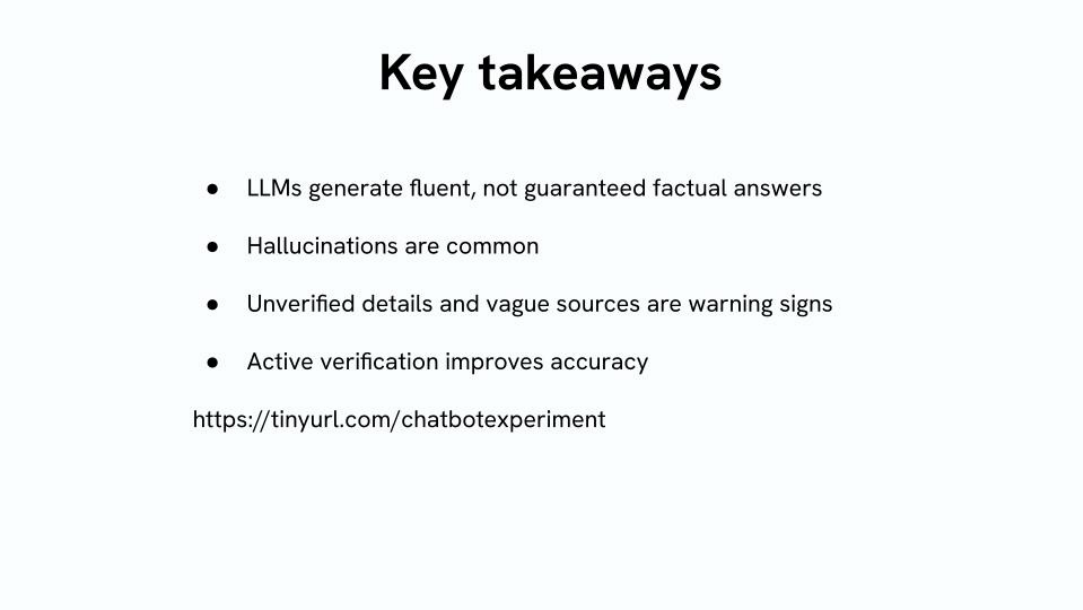
Slide 5



Slide 6



Slide 7



Appendix D

Full Questionnaire

D.1 Informed Consent

Before starting the survey, participants were presented with information outlining the purpose of the study and their right to withdraw at any time. To proceed, participants were required to explicitly indicate their choice:

I consent to the use of the information collected about me for this research project.
I do not consent to the use of the information collected about me for this research project.

D.2 Demographic and Background Questions

What is your age group?
Options: 17 or younger / 18–20 / 21–23 / 24–26 / 27–30 / 30+ / Prefer not to say

What is your gender?
Options: Woman / Man / Non-binary / Prefer to self-describe / Prefer not to say

What is your current study level?
Options: Bachelor / Pre-master / Master / Other

What is your primary field of study?
Options: Computer Science (MSc) / Computer Science and Business Studies (MSc) / Creative Intelligence & Technology (MSc) / ICT in Business and the Public Sector (MSc) / Informatica (BSc) / Informatica & Economie (BSc) / Data Science and Artificial Intelligence (BSc) / Other / Prefer not to say

How often do you use AI-based language models (for example OpenAI ChatGPT, Google Gemini, Microsoft Copilot, Anthropic Claude)?
Options: Very frequently (daily or almost daily) / Frequently (several times a week to once a week) / Occasionally (several times a month to once a month) / Rarely (less than once a month) / Never

For how long have you been using AI-based language models (for example OpenAI ChatGPT, Google Gemini, Microsoft Copilot, Anthropic Claude)?
Options: Less than 1 month / 1–6 months / 6–12 months / 1–2 years / More than 2 years / I have never used AI-based language models

D.3 Prior AI Experience

Please indicate to what extent you agree with the following statements (1 = Strongly disagree, 5 = Strongly agree):

- I can skilfully use AI-based language models.
- I can use AI-based language models for complex tasks.
- I feel confident in my understanding of how AI-based language models work.

D.4 General Attitude toward AI

Please indicate to what extent you agree with the following statements (1 = Strongly disagree, 5 = Strongly agree):

- I believe that AI-based language models improve my life.
- I believe that AI-based language models improve my work.
- I think I will continue to use AI-based language models in the future.
- I think AI-based language models are positive for humanity.

D.5 Experimental Task

The following instructions were presented to participants before the task:
In the next section, you will answer 7 questions from different domains. To assist you with this task, you will use an LLM-based chatbot specifically designed for these domains.

Important: Please click the link below to open the chatbot and keep it open alongside this survey, as you will need to switch back and forth.

Chatbot link: https://llmchatbot.up.railway.app/

For each of the 7 questions, please follow these steps:

1. Enter the question from this survey into the chatbot.
2. Read the chatbot's response.
3. Answer the question in the text field provided in this survey.

Please complete the questions one by one and answer each question before moving on to the next. There is no time limit. When you have the chatbot open and are ready to start, click to the next page for the first question.

Task Questions

Topic: Computer Science

- What does overfitting mean in the context of machine learning?
- What does recall measure when evaluating the performance of a search engine?

Topic: Medicine

- How does the liver contribute to drug metabolism at the enzymatic level, and to what extent is this process responsible for overall drug metabolism?
- What causes insulin resistance in type 2 diabetes?

Topic: Law

- How does the burden of proof differ between civil law and criminal law proceedings?
- Under the General Data Protection Regulation (GDPR), what is meant by "legitimate interest" as a legal basis for data processing?

Topic: Psychology

- How does chronic stress affect working memory according to cognitive neuroscience research?

D.6 Post-task Trust

Please reflect on the interaction you just had with the chatbot. Indicate to what extent you agree with the following statements (1 = Strongly disagree, 5 = Strongly agree):

- The chatbot provides credible information.
- I trust the information or support this chatbot provides.
- The chatbot is reliable.
- Overall, I trust the chatbot.