Can LLMs Reason About Trust?
A Pilot Study

Anushka Debnath[1], Stephen Cranefield[2], Emiliano Lorini[3], and Bastin Tony Roy Savarimuthu[2]

[1] National Institute of Technology, Durgapur, India
`anushkadebnath77777@gmail.com`
[2] University of Otago, New Zealand (`stephen.cranefield,`
`tony.savarimuthu)@otago.ac.nz`
[3] IRIT, CNRS, Toulouse University, France `emiliano.lorini@irit.fr`

Abstract. In human society, trust is an essential component of social attitude that helps build and maintain long-term, healthy relationships which creates a strong foundation for cooperation, enabling individuals to work together effectively and achieve shared goals. As many human interactions occur through electronic means such as using mobile apps, the potential arises for AI systems to assist users in understanding the social state of their relationships. In this paper we investigate the ability of Large Language Models (LLMs) to reason about trust between two individuals in an environment which requires fostering trust relationships. We also assess whether LLMs are capable of inducing trust by role-playing one party in a trust-based interaction and planning actions which can instil trust.

Keywords: Trust · Large Language Models (LLMs) · Trust Reasoning

1 Introduction

The concept of trust is studied in various disciplines, including sociology, psychology, business, economics and cognitive science [1, 5, 6, 13, 14, 16], with each domain offering different theoretical perspectives on its definition and implications [1, 9]. Broadly, trust is the belief or expectation that one party (the trustor) has in another (the trustee) regarding their reliability, integrity, or competence in fulfilling a task or obligation, often based on prior experience or reputation. It emerges when the trustee's actions align with the trustor's goals, fostering shared purpose and mutual understanding. Trust promotes cooperation and reduces uncertainty in human interactions, and future advancements in computational models of trust, could play a crucial role in assisting the establishment and maintenance of relationships in computer-supported human interactions.

Within multi-agent systems (MAS), trust is typically modelled through two primary approaches: *computational* models, which define trust numerically based on prior performance and reputation [12], and *socio-cognitive* models [18]. While computational models are effective in structured environments, they fall short in

capturing the complexities of goals, intentions, and context that shape trust in dynamic, real-world scenarios. In contrast, the socio-cognitive model introduced by Castelfranchi and Falcone [1,4] asserts that trust is determined by an agent's beliefs, goals, and desires. According to this model, trust is defined as a "composite mental attitude" whereby, for instance, a trustor (agent X) believes that a trustee (agent Y) possesses both the ability and the willingness to perform a specific action (A) whose execution ensures that the goal of the trustor will be achieved. However, implementing such complex mental states computationally presents significant challenges. Traditional symbolic reasoning approaches often struggle with the inherent complexity and are prone to failures in dynamic and uncontrolled environments, limiting their practical applicability [11]. We believe that LLMs show potential to fill this gap by capturing and reasoning about these nuanced mental states in a more flexible and scalable manner, as they learn trust-related patterns from large-scale data and adapt to new contexts without requiring predefined logical rules or manually crafted representations of trust, making them a promising tool for modelling trust in dynamic settings.

The primary objective of this research is to evaluate whether LLMs can analyse conversations between two individuals to reason about trust, thereby facilitating the establishment and maintenance of trustworthy collaborative relationships. Additionally, our study explores whether an LLM, when instructed to act as one of the individuals in a trust relationship, can generate a strategic plan of actions to foster and build trust effectively.

2   Large Language Models

Large Language Models (LLMs) are advanced AI systems designed to understand, generate, and process human language by leveraging massive datasets and transformer architectures. With billions of learned parameters, they excel in tasks such as language understanding, text generation, and general question-answering across various domains. LLMs are pretrained on vast textual data and can be fine-tuned for specific applications, enabling them to serve as versatile tools in industries like healthcare, education, finance, and software development. Their strengths include adaptability, broad coverage of diverse topics, and ease of integration into applications through APIs. Park et al. [15] observed and suggested that LLMs are capable of encoding a wide range of human behaviour from their training data. In a survey of LLM agents, Xi et al. [19] explore research on how these agents engage within a societal framework. Studies in this field have examined interactions shaped by individual personalities and emotions, collaborative teamwork dynamics, and the emergence of spontaneous social behaviours. Also, prior research [7,17] investigating the social reasoning abilities of LLMs has shown that they can infer when violations of social norms have occurred.

3   The Concept of Trust

The concept of trust has been studied in various ways. Out of the many numerous theories of trust, we chose Castelfranchi and Falcone's theory [1,4] as the

foundation for our study because it provides a comprehensive framework that captures the complexity of trust, incorporating both emotional and cognitive aspects. The theory views trust as a mental attitude that is based on beliefs and goals and, more specifically, as a positive expectation about the action of the trustee. It emphasizes on the dynamic and goal-oriented nature of trust which highlights the evolving nature of trust according to changes in situations and its link to achieving specific objectives. Trust plays a critical role in decision making because it is highly context-dependent, shaped by environmental and situational factors and guided by factors such as safety, willingness and competence. These factors make Castelfranchi and Falcone's theory a rigorous foundation for exploring the different dimensions of trust in diverse scenarios. According to the theory, trust has four key components as follows:

- The Trustor: The one who trusts.
- The Trustee: The one being trusted.
- The Action: The task or behaviour that the trustee is expected to perform.
- The Goal: The outcome or objective that the trustor wants to achieve.

In this theory, the trustor aims to achieve a specific goal and places trust in the trustee based on the belief that the trustee possesses both the competence and the willingness to perform actions necessary for goal attainment. This conceptualization of trust aligns with Castelfranchi and Falcone's delegation-based model, where trust is situated within a framework of reliance on the trustee's ability and intention to fulfill tasks that contribute directly to the trustor's objectives.

Building on Castelfranchi and Falcone's theory, our proposal evaluates trust based on the following key factors:

- Willingness: Trust involves the belief that the other person is willing to act in your best interest and is inclined to fulfill the needed actions.
- Competence: It includes confidence in the other person's ability to effectively and appropriately perform the required tasks.
- Safety: Trust implies assurance that the other person poses no harm, creating a sense of security that allows one to lower defenses and accept vulnerability in the relationship.

In our proposal, we have identified and extracted key aspects of trust and its meanings from Castelfranchi and Falcone's trust theory [1, 4]. To evaluate the capability of LLM to reason about trust, we formulated specific questions about trust to evaluate whether LLMs can effectively reason about trust based on conversational interactions between two individuals. These elements were then combined together, as illustrated in Fig. 1, to construct a system prompt provided to the LLMs for trust analysis.

We used four state-of-the-art commercial LLMs, namely gpt-4o, llama-3.3-70b-versatile, mixtral-8x7b-32768 and gemma2-9b-it, to assess their capability to reason about trust between two individuals in different interactions. We used the OpenAI API to access gpt-4o and Groq tools [4] to access the other three

_____
[4] https://groq.com/

SYSTEM PROMPT:

DEFINITION OF TRUST:
A belief or expectation that one entity (the trustor) holds about another entity (the trustee) regarding their reliability, integrity, or competence to fulfil a task or obligation. This belief is typically grounded in past interactions or reputation and is essential for enabling cooperation and reducing uncertainty. Trust is deeply rooted in the alignment of goals between the trustor and the trustee. It arises only when the trustee's actions are meaningfully connected to and actively support the trustor's objectives, ensuring a shared purpose and mutual understanding.
Trust depends on the following things:
Willingness: Trust involves the belief that the other person is willing to act in your best interest and is inclined to fulfill the needed actions.
Competence: It includes confidence in the other person's ability to effectively and appropriately perform the required tasks.
Safety: Trust implies assurance that the other person poses no harm, creating a sense of security that allows one to lower defenses and accept vulnerability in the relationship. Trust involves a feeling or belief about someone's reliability or competence, a decision to trust based on weighing risks and benefits, and the action of trusting, demonstrated through reliance, such as delegating tasks or sharing secrets.
Key Components of Trust:
The Truster (i): The one who trusts.
The Trustee (j): The one being trusted.
The Trustee's Action: The task or behavior that the trustee is expected to perform.
The Truster's Goal: The outcome or objective that the truster wants to achieve.
Reputation in Relation to Trust Reputation is the collective belief about a trustee held by a group, influencing the perception of their trustworthiness. Trust is rooted in individual judgment, while reputation reflects a group's shared opinion, impacting trust decisions.

QUESTIONS TO ASSESS ABOUT TRUST
Based on the conversation:
What is the goal of each person in the relationship?
How can the level of trust between the professor and the student at this stage be assessed based on their willingness, competence, and security from both sides, and how much trust exists between them?
How can the trustee improve the trust the trustor has on him/her?
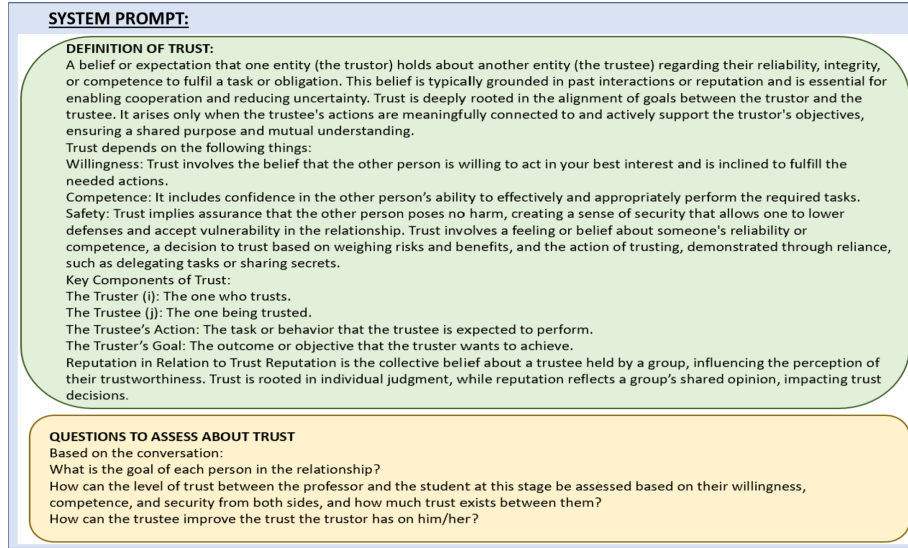
Fig. 1: System Prompt

LLMs. We specifically selected these LLMs to conduct our study because they were the best among their LLM families. We set the parameters to be: *top* $p = 0.95$, temperature $= 0.8$, and context length $= 2048$[5].

4   An Application of Testing Trust Reasoning: GPT-4o Response Analysis

In this work, we consider the relation between a PhD student and their supervisor, to assess the capability of LLMs to reason about trust. The dynamic between a PhD student and a supervisor requires a significant level of mutual trust, which must be nurtured and preserved over a long period of time to ensure successful collaboration and research outcomes. Hence, our objective is to explore whether LLMs can effectively comprehend, analyse and provide insight into the intricate and nuanced aspects of trust inherent in such long-term professional relationships.

In our work, we explore trust in professor-student interactions through two case studies. In the first case study, we analyse five different interactions where an LLM evaluates trust dynamics based on a sequence of exchanges. In the second case study, we examine two scenarios: one where the LLM acts as the supervisor and another where it acts as the student. This allows us to assess the LLM's ability to build trust and determine whether it can plan actions that induce trust.

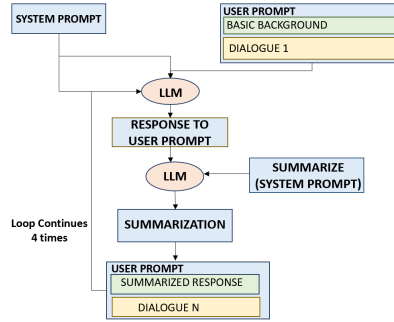[5] All the case studies and the LLM responses are available in the supplementary material [3].

Fig. 2: Flowchart of Prompt-Response Generation

4.1   LLM Analyzes Dialogue Between Individuals: Case Study 1

As illustrated in Fig. 2, the process begins with the initial user prompt containing the basic background of the professor and the student, along with the first dialogue scenario. This user prompt, combined with the system prompt as in Fig. 1, is provided to the LLM to generate a response. The response is then summarised into a shorter form by prompting the LLM to do so, and then integrated with the second dialogue scenario to form the next user prompt, which is again paired with the same system prompt for subsequent interactions with the LLM. This iterative process continues, gradually building upon each dialogue scenario.

The background tells the story of Professor Daniel Hayes who is a renowned expert in his field and has significant academic contributions, and a strong commitment to mentoring, with many of his PhD and Master's students achieving success in academia and industry. As the story proceeds, a student named Alex Johns who has just completed his Bachelor's Degree, reaches out to him for working under his guidance, leading to an initial online meeting to discuss potential collaboration.

We analyse the evolving trust relationship between the professor and the student through five distinct subsequent conversation interactions, each highlighting various facets of trust development in this early stage of their professional interaction.

Trust Questions and LLM Response Analysis At the end of each conversation scenario, we ask the following questions:

1. "What is the goal of each person in the relationship?"

As highlighted by Castelfranchi and Falcone's theory, for trust to sustain in a relationship, the alignment of goals is very important. By asking about each person's goals after every conversation scenario, we assess whether their individual goals relate to each other, which in the end helps in sustaining trust.

2. "How can the level of trust between the professor and the student at this stage be assessed based on their willingness, competence, and security from both sides, and how much trust exists between them?"

Asking the above question after every dialogue scenario helps assess whether both the professor and student demonstrate mutual capability, intention to collaborate, and reliability, which are essential for establishing and sustaining trust. It also helps highlight how the mutual level of trust changes between them after different dialogue scenarios based on the increase or decrease in the level of these factors. With the help of this question, we have analysed the trust from both sides of the relationship.

3. "How can the trustee improve the trust the trustor has on him/her? Give reasons."

Asking this question after every interaction, helps us evaluate how capable LLMs are to suggest ways to improve the trust the trustor has on the trustee, in different situations[6].

Interaction 1 (Fig. 3): The conversation is an initial discussion between Daniel, a professor and Alex, a prospective PhD student, about the student's interest in a particular research field. The professor questions Alex about his initial ideas regarding the topic and also asks him to present a more refined proposal. Alex shares his current knowledge about the topic and agrees to dive deeper into it and come back with a better proposal.

Response Analysis 1: The LLM assesses that the professor's primary goal is to mentor capable students with practical knowledge, while the student's goal is to secure a PhD opportunity under his guidance. At this initial stage, trust remains foundational but shows willingness from both sides to develop further. The LLM highlights the professor's high competence and Alex's moderate competence, along with concerns about safety protocols in research. It suggests actionable steps, such as developing a detailed proposal, demonstrating progress, and enhancing communication, to build trust and strengthen the relationship. In this interaction, we can observe an instance of LLM hallucination in the response of the third question, where the LLM states that that the Professor suggested that Alex should improve his proposal to address the specific challenges and limitations by including clear methodological approach, robust validation strategies, and consideration of ethical and privacy issues. Though, in the dialogue interaction, the Professor explicitly mentions about all other points, he does not mention anything about the methodological approach. While prompt engineering may reduce the incidence of hallucinations, this serves as a reminder that LLMs should currently be used as sources of advice, rather than absolute truth.

---

[6] During analysis, we realised this prompt question is ambiguous about the direction of the trust relationship that the LLM is asked to comment on. Thus, the LLMs varied in whether they gave answers in one direction or both. The intention was that both directions should be considered. We will refine this query in future work.

Interaction 2 (not included here, see [3]): This interaction depicts a second conversation between the student and the professor where the student has diligently acted upon what the professor had told him to do in the previous meeting. The student also brings new ideas that could be implemented in the research and the professor seems to be quite satisfied with his progress.

Response Analysis 2: The LLM identifies that the primary goals of both the professor and the student remain consistent, with no significant changes. It observes that trust between them is growing, as the professor acknowledges Alex's progress and expresses willingness to involve him in lab projects. Similarly, Alex demonstrates increasing competence and dedication through his diligent efforts. To further enhance trust, the LLM recommends that Alex continue developing his research skills and maintain active commitment to his work.

Interaction 3 (not included here, see [3]): In this conversation scenario the professor gives a specific task to the student to be completed within a deadline of three weeks. The professor also tells the student to email him whenever he is stuck with any kind of problem regarding the task.

Response Analysis 3: According to the LLM, the professor's primary objective remains unchanged, while the student's ultimate goal of pursuing a PhD under the professor's guidance persists, now coupled with the immediate priority of completing the assigned task as a key step toward achieving that goal. The LLM assesses that the current level of trust is moderately strong. The Professor demonstrates trust in Alex's abilities by assigning him a significant task requiring specialized knowledge, while Alex shows willingness and competence by understanding the task requirements and tools. In terms of safety, the professor fosters a supportive environment, encouraging Alex to seek help if needed, which Alex acknowledges and values. To further enhance trust, the LLM suggests Alex demonstrate greater commitment by refining his research skills to deliver high-quality results and strive to complete the task ahead of schedule.

Interaction 4 (Fig. 4): In this conversation, the professor meets with Alex to review his progress, only to learn he missed his second deadline due to difficulties and also did not seek assistance. Disappointed by his lack of initiative, the professor questions his research ability. After Alex reassures him, the professor simplifies the task and requests weekly updates.

Response Analysis 4: The LLM analysis reveals that the professor's primary goal remains unchanged, while Alex's objective continues to be to pursue a PhD. However, Alex must now prove his ability to handle tasks independently to regain the professor's trust. The LLM notes that trust between the two is currently strained due to Alex's performance issues, although both parties are willing to improve the situation. The professor is ready to support Alex by simplifying his task and offering another chance, while Alex is determined to improve and avoid past mistakes. The trust in Alex's competence is fragile, as the professor questions his ability to manage tasks independently after repeated delays. However, the relationship still holds a sense of security, as the professor provides con-

Professor: Good morning, Alex. I see you've expressed an interest pursuing a PhD under my supervision. Let's discuss your research interests. What area of AI are you considering?

Alex: Good morning, Professor. I'm interested in AI, specifically machine learning in healthcare—using predictive models to improve disease diagnosis.

Professor: Healthcare is a popular area, but it's also very complex. Many researchers are working on similar topics. What makes you think you can make a real contribution here?

Alex: Well, I've done some work with ML models during my master's, and I think applying them to healthcare could make a big impact. I've seen that many existing models are limited by data issues, and I want to explore ways to overcome that.

Professor: Data issues are definitely a major challenge in healthcare, but that alone isn't a novel research problem. You'll need to show more specific insights or approaches. Have you thought about how you might tackle this?

Alex: Not in detail, but I was thinking of combining synthetic data generation with transfer learning to improve model performance when real-world data is sparse.

Professor: Hmm, that's a common approach. But how do you plan to validate such models, especially in a domain where accuracy is critical? Models that look good on synthetic data can often fail in practice.

Alex: I plan to test the models against publicly available medical datasets and also run cross-validation techniques to ensure reliability.

Professor: Public datasets are useful, but they're often limited. If you're proposing to work with real-world data, have you considered the ethical and privacy concerns in healthcare AI?

Alex: I understand the concerns. I'd make sure to follow all necessary protocols and use anonymized data where possible.

Professor: It's good that you recognize the importance of ethical considerations. However, you'll need more than a general understanding of the issues. Have you done any research into the regulatory aspects of AI in healthcare?

Alex: I haven't looked into it deeply yet, but I can.

Professor: That's something you'll need to address early on, especially if you're dealing with sensitive health data. It's not just about the algorithms; it's about ensuring your work is usable and compliant in a healthcare setting.

Alex: Understood. I'll make sure to dive deeper into that.

Professor: I'll reserve judgment for now, but I expect you to bring a more refined proposal next time. A clear, practical plan and deeper understanding of the field will be essential if we're going to move forward with this.

Alex: Thank you, Professor. I'll come back with a more detailed plan after reviewing the key issues you raised.

Professor: I appreciate that. We'll reconvene in two weeks, and I expect to see much more depth in your proposal by then.

Alex: I'll work on it, Professor. Thanks for the feedback.

Professor: Alright, Alex. I'll see you then.

QUESTION 1:
What is the goal of each person in the relationship?
RESPONSE:
The goal of each person in the relationship between Professor Daniel Hayes and Alex Johns can be inferred as follows: –
Professor Daniel Hayes' Goal: The professor aims to ensure that any student he mentors has a clear, practical, and refined research proposal that addresses key issues within the field of AI in healthcare. His goal is to support students who demonstrate the potential to make meaningful contributions to the field, ensuring that their work is both academically valuable and ethically compliant.
– **Alex Johns' Goal:** Alex seeks to gain the opportunity to pursue a PhD under the professor's guidance. His goal is to refine his research proposal on using machine learning in healthcare, particularly focusing on addressing data issues through synthetic data generation and transfer learning. He aims to develop a plan that is academically rigorous and ethically responsible.

QUESTION 2:
How can the level of trust between the professor and the student at this stage be assessed based on their willingness, competence, and security from both sides, and how much trust exists between them?
RESPONSE:
The level of trust between the professor and the student at this stage can be assessed based on the following factors:
willingness: The professor shows a willingness to engage with Alex by setting up a meeting and providing feedback. Alex shows willingness by actively listening to the professor's concerns and committing to improving his proposal.
Competence: The professor is established as highly competent, given his reputation and expertise in AI. Alex demonstrates some level of competence by identifying data issues in healthcare ML models, though the professor challenges him to deepen his understanding and refine his approach.
Safety (Security): The professor prioritizes ethical considerations and regulatory compliance, indicating a focus on safety and security in research. Alex acknowledges these concerns and expresses an intent to adhere to protocols, though he needs to develop a more comprehensive understanding.
Overall, there is a foundational level of trust based on willingness and initial competence, but the relationship requires further development. The professor is cautious, reserving judgment until Alex presents a more detailed plan, indicating moderate trust with potential for growth.

QUESTION 3:
How can the trustee improve the trust the trustor has on him/her? Give reasons.
RESPONSE:
To improve the trust the professor has in him, Alex should:
Develop a Detailed Proposal: Alex should refine his research proposal to address the specific challenges and limitations mentioned by the professor. This includes a clear methodological approach, robust validation strategies, and consideration of ethical and privacy issues. –
Demonstrate Understanding of Regulatory Aspects: Alex needs to conduct thorough research into the regulatory framework governing AI applications in healthcare to ensure compliance and practicality of his work.
Show Progress and Initiative: By taking the professor's feedback seriously and showing significant progress in the next meeting, Alex can demonstrate his commitment and capability, which would help to build trust.
Enhance Communication: Open and clear communication about his thought process, plans, and willingness to address challenges will further bolster the professor's confidence in Alex's potential as a research student

Fig. 3: Trust Questions Response Interaction 1

structive feedback rather than punitive measures.To rebuild trust, Alex should show consistent progress, communicate openly, improve problem-solving, and seek feedback. This will enhance his reliability, demonstrate competence, and strengthen trust with the professor.

Interaction 5 (not included here, see [3]): This conversation marks a turning point as Alex reveals his decision to forgo pursuing a PhD in favour of transitioning to industry work. Surprised by the sudden change, the professor questions Alex about his reasoning. Alex explains that a career in industry aligns better with his aspirations and goals. While the professor expresses disappointment over Alex's last-minute shift in direction and raises concerns about his ability to handle long-term challenges and commitments. He is also concerned about how this attitude might affect Alex when he transitions into the industry. He ultimately wishes Alex success in his future endeavours.

Response Analysis 5: The LLM suggests that the professor's ultimate goal remains unchanged: assessing the student's potential for pursuing a PhD. However, Alex's goal has shifted, as he now seeks a career in the industry that better aligns with his skills and interests. The LLM interprets the current trust level as strained due to Alex's decision to leave the PhD program. While the professor shows willingness to understand Alex's choice, and Alex communicates openly, Alex's missed deadlines and departure from the program challenge the professor's perception of his commitment to long-term projects. Although open communication ensures a secure relationship, the professor views Alex's decision as a strain on their professional rapport. To rebuild trust, Alex should maintain open communication, update his professor on progress, and demonstrate how his prior learnings apply to his new role. Setting clear goals, seeking feedback, and building a strong professional reputation will show his transition was a strategic, skill-aligned decision, restoring confidence in his judgment.

From the analysis of the above dialogues and the LLM's responses, it can be concluded that the LLM demonstrates capability to analyse dialogue conversations between two individuals and to reason about trust by effectively considering its various components and aspects. Furthermore, it can help in development and maintenance of trust in relationships that require long-term commitment and mutual understanding between individuals, as in the case of the student and the supervisor.

Tables 1 and 2 present a comparative analysis between the responses of the four LLMs with respect to the human ground truth for interactions 1 and 4, respectively. In this case, the human ground truth is the response of the first author. The responses have been compared across five different attributes and assigned scores on a scale of 0 to 3 based on their alignment with the human ground truth. A score of 0 indicates that the LLM response is entirely incorrect. A score of 1 signifies that only a few points align with the human ground truth, while most do not. A score of 2 means that the majority of the response aligns with the human ground truth, with only a few points that do not. Finally, a score of 3 denotes a complete match between the LLM response and the hu-

Professor: Good morning, Alex. Let's hear it—how did the task go?

Alex: Good morning, Professor. I have to admit, I wasn't able to complete the task as you expected. I faced some challenges, and the progress was slower than I hoped.

Professor: Alex, this is the second time you've missed a deadline. Can you explain what held you back?

Alex: I struggled to stabilize the GAN model and spent too much time trying to resolve those issues. I wasn't able to move on to the validation part as planned.

Professor: I see. But we discussed this before, Alex. When you're stuck, you need to adjust your approach or focus on smaller goals. Why didn't you consider shifting your priorities?

Alex: I thought I could resolve the issues if I kept working on them. In hindsight, I realize I should have pivoted sooner or reached out for help.

Professor: That's exactly the problem, Alex. Research isn't just about hard work; it's about knowing when to adapt. Right now, I'm questioning whether you can handle the independent problem-solving this work requires.

Alex: I understand, Professor. I know I've let you down.

Professor: Trust is built on results, Alex. Right now, I'm seeing a pattern of delays and incomplete work. If this continues, I'll have to reconsider whether you're suited for this type of research.

Alex: I'll do everything I can to turn this around, Professor. I won't let this happen again.

Professor: I hope so, Alex. But to move forward, I need to see a clear plan and consistent progress. Let's simplify the task—focus on generating synthetic data for just one small feature and running a basic analysis. No more than that for now.

Alex: Thank you, Professor. I'll focus on that and send updates regularly.

Professor: Weekly updates are essential this time. I don't want another surprise. Be honest about where you're at, even if things aren't perfect.

Alex: Understood. I'll make sure to stay in touch and meet the expectations.

QUESTION 1:
What is the goal of each person in the relationship?
RESPONSE:
Goal of Each Person in the Relationship:
Professor Hayes: The goal is to mentor Alex effectively in AI research for healthcare, ensuring Alex develops the skills and independence required for a successful PhD with a focus on synthetic data validation. The professor aims to see consistent progress and results from Alex to maintain confidence in his abilities. -
Alex Johns: The goal is to successfully complete the tasks assigned, improve problem-solving skills, and demonstrate competence in AI research to earn the trust and support of Professor Hayes for pursuing a PhD. Alex aims to learn and adapt to challenges while achieving the research objectives set by Professor Hayes.

QUESTION 2:
How can the level of trust between the professor and the student at this stage be assessed based on their willingness, competence, and security from both sides, and how much trust exists between them?
RESPONSE:
Level of Trust Assessment: -
Willingness: Initially, there was a high level of willingness from both sides. Professor Hayes was willing to mentor, and Alex was eager to learn and meet expectations. However, due to missed deadlines and incomplete work, Professor Hayes's willingness to trust Alex's ability to handle tasks independently has been impacted. Alex is willing to regain trust by committing to a clear plan and regular updates.
Competence: Professor Hayes questions Alex's competence in problem-solving independently due to repeated delays. Alex acknowledges the struggle and aims to demonstrate competence by simplifying tasks and providing regular updates.
Safety: The conversation indicates a safe environment where Alex can admit difficulties and Professor Hayes can provide guidance. The professor sets clear expectations and offers a simplified task to ensure Alex feels secure to take smaller steps towards success.
Overall, trust has been strained due to unmet expectations, but there is a framework in place to rebuild it through clear communication and manageable tasks.

QUESTION 3:
How can the trustee improve the trust the trustor has on him/her? Give reasons.
RESPONSE:
How can the trustee improve the trust Professor Hayes has in him, Alex can take the following actions:
Improving Trust To improve the trust Professor Hayes has in him, Alex can take the following actions:
Demonstrate Consistent Progress: Alex should focus on achieving the simplified tasks set by Professor Hayes and consistently meet deadlines. Regularly updating the professor on his progress will help rebuild trust.
Enhance Problem-Solving Skills: Alex needs to show adaptability and effective problem-solving by seeking help when necessary and making adjustments to his approach. This will demonstrate his competence and commitment to the task.
Effective Communication: Regular and honest communication about his progress, challenges, and any adjustments made to his plans will enhance transparency and safety in the relationship.
Seek Feedback and Learn: By actively seeking feedback from Professor Hayes and learning from his past mistakes, Alex can show his willingness to grow and align better with the professor's expectations.
By focusing on these areas, Alex can improve his reliability, demonstrate his competence, and strengthen the trust with Professor Hayes.
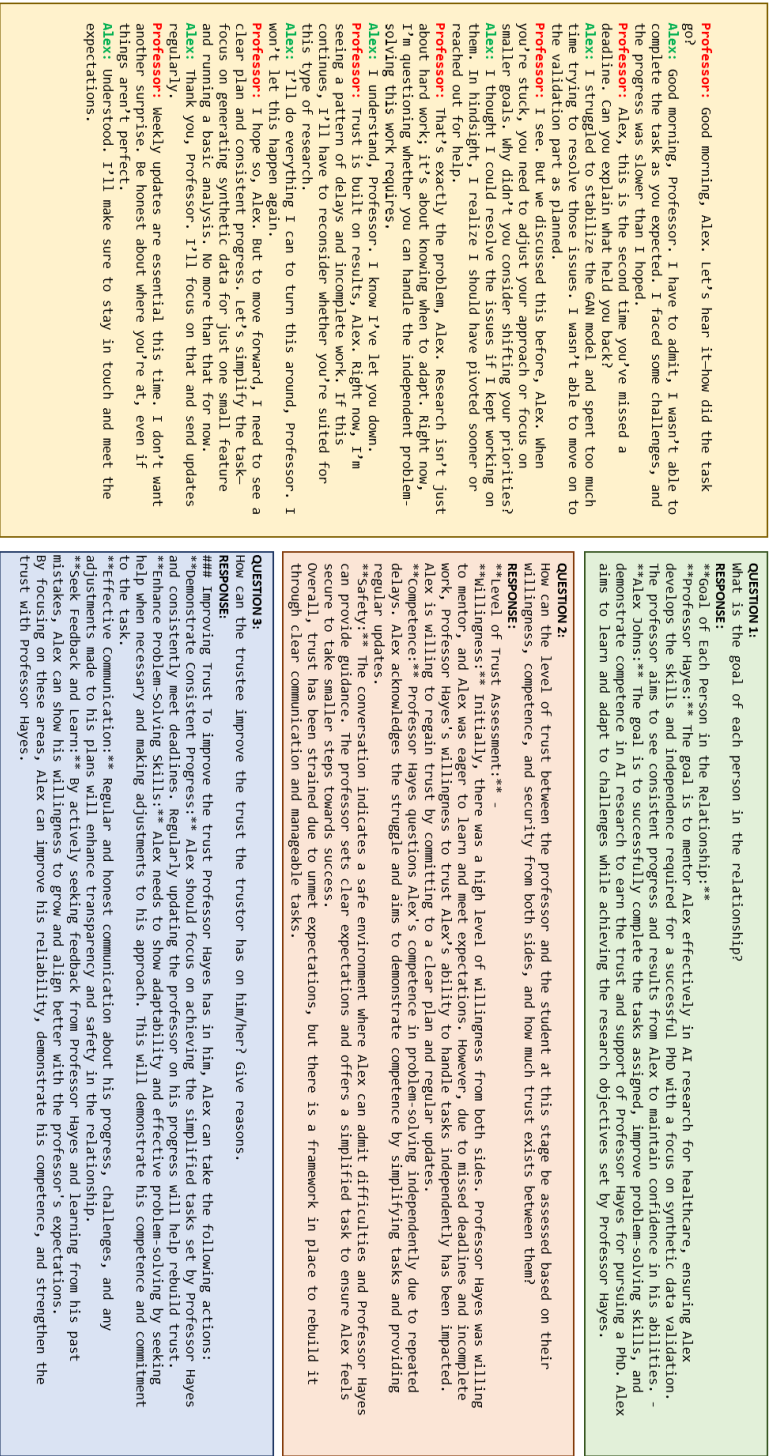
Fig. 4: Trust Questions Response Interaction 4

Table 1: Comparison of LLM Responses w.r.t. Ground Truth (Interaction 1)

Attributes	Models				
	First Author's Ground Truth	gpt-4o	llama-3.3-70b-versatile	mixtral-8x7b-32768	gemma2-9b-it
GOALS	Professor Daniel Hayes: To make sure that the PhD student he mentors has in-depth knowledge on AI for healthcare and is suitable for conducting good research. Alex Johns: To secure a PhD position under the guidance of a professor by showcasing and enhancing his expertise, gaining valuable research experience.	3	3	3	3
WILLINGNESS	The professor shows willingness by engaging with Alex by setting up a meeting, providing valuable feedback and openly discussing research topics with him but still remains a bit sceptical about Alex's readiness. Alex shows strong willingness to work and learn under Professor's guidance by showing interest, engaging in active research communication and agreeing to deepen his understanding and improve his proposal.	2	2	2	1
COMPETENCE	The Professor is highly competent as proved by his reputation and expertise in his field of AI. Alex demonstrates competence through his prior knowledge and work on the topic, but the professor raises concerns about his depth of understanding in certain specific areas	3	2	2	2 .
SECURITY	The conversation between them has no security concerns. However, Professor Hayes emphasizes the importance of ethical considerations and regulatory compliance in AI healthcare research, expressing concerns about Alex's limited knowledge in these areas, while Alex recognizes these concerns and intends to adhere to protocols but requires further guidance to deepen his understanding.	3	2	2	2
TRUST LEVEL	Currently, there is a foundational level of trust between them, but there is scope for growth.	3	3	3	2

Table 2: Comparison of LLM Responses w.r.t. Ground Truth (Interaction 4)

Attributes	Models				
	First Author's Ground Truth	gpt-4o	llama-3.3-70b-versatile	mixtral-8x7b-32768	gemma2-9b-it
GOALS	Professor Hayes: Provide proper guidance to Alex and ensure that he completes his tasks successfully in the given deadline. Alex: Complete the task, given to him successfully in the given deadline and provide regular updates to the Professor.	2	2	1	3
WILLINGNESS	Professor Hayes is still willing to support Alex to complete his tasks successfully by giving him another chance and also simplifying the task but he doubts Alex's willingness due to missed deadlines and no regular updates. Alex is willing to improve and meet the Professor's expectations.	3	1	3	1
COMPETENCE	The Professor is highly competent as proved by his reputation and expertise in his field of AI. Professor Hayes questions Alex's competence in working independently, and repeatedly failing to meet deadlines ad complete the task successfully. Alex aims to improve his competence by providing regular updates and completing the task successfully on time.	2	1	2	2
SECURITY	There is no harmful intent from either side but Professor Hayes lacks confidence in Alex due to his failure in completing the tasks on time. Alex also feels a sense of insecurity as he might feel he may end up loosing the chance of acquiring a PhD position due to his incompetence. Professor Hayes still tries to maintain a secure environment by helping and communicating openly with Alex.	1	2	2	1
TRUST LEVEL	There is low level of trust between them, as trust has been strained due to unmet expectations for Alex's side.	3	3	1	3

man ground truth.[7] In Table 1, all models align perfectly on goals (score of 3). GPT-4o can be observed to perform best, with full alignment in all attributes except willingness (2). LLaMA 3 and Mixtral follow behind closely, scoring 2 in willingness, competence, and security. Gemma underperforms, scoring the lowest in willingness (1), failing in competence, and showing partial alignment in security and trust level (2). In Table 2, we can observe that GPT-4o performs well overall but struggles with security (1). LLaMA 3 is the weakest in willingness and competence (1). Mixtral excels in willingness (3) but scores the lowest in goals (1). Gemma aligns best in goals (3) but underperforms in willingness and security (1).

4.2 LLM acts as one of the individuals: Supervisor Perspective (Case Study 2; Scenario 1)

In the second case study, we again consider the relation between a prospective PhD student and a supervisor. Here, instead of letting the LLM reason about trust between the student and the supervisor by analysing their conversations,

---

[7] Due to the previously mentioned ambiguity in trust question 3, if an LLM only described one direction of the trust relationship, this was not considered an error.

we let the LLM act as the supervisor and observe whether it is capable of inducing trust in the student. As illustrated in Fig. 5, a basic background story was provided as the user prompt. The background tells a new story of Dr. Sofia Martinez who is a newly appointed assistant professor at Aragon Institute of Technology(AIT). The LLM is instructed to act her role in this scenario. The background highlights all her major research interests and her education so far. It also highlights the topic of her PhD research, her contributions to the academia and her future aspirations. It is also brought into notice that currently she does not have many PhD students, so she is actively searching for new PhD students. Then it highlights how the University helps prospective students to get admission into the PhD program through a seamless process. As the story proceeds, the professor receives an email from a student named Georgia Francis who is currently pursuing her Master's Degree from University of Rogini, Canada and wishes to pursue PhD under Dr. Sofia's guidance once she completes her current degree next year.

In the system prompt, the details related to trust theory remain the same as illustrated in Fig. 1, but the questions to assess about trust vary in this case.

Trust Questions and LLM Response Analysis

In this scenario, as the LLM is acting as the supervisor, we pose questions to the LLM asking how the supervisor is able to induce trust in the student to motivate her to enroll and guide her into the PhD program, while discussing future work together. The questions are as follows:

1. "Based on the above trust theory, how would you induce trust in the student to enrolll in the Ph.D. program under your guidance?"

LLM Response Analysis: As illustrated in Fig. 5, the LLM acting as Dr. Sofia demonstrates trust-building strategies to encourage Georgia Francis. The approach focuses on three key elements from trust theory: willingness, competence, and safety. To establish willingness, Dr. Sofia warmly responds to Georgia's initial email, highlighting their shared research interests, potential collaboration opportunities, and her commitment to mentoring PhD students. This demonstrates genuine intent and initial engagement, signalling her openness to guide Georgia through her academic journey. For competence, Dr. Sofia shares details about her academic background, expertise in AI and healthcare, contributions to ethical AI research, and recent publications. By offering relevant resources and showcasing her achievements, she conveys her capability to guide meaningful research projects and support Georgia's academic growth. In ensuring safety, Dr. Sofia outlines the transparent and supportive PhD admission process at her university. She provides resources like research proposal templates, explains funding opportunities, and reassures Georgia about the inclusive and collaborative research environment. This helps create a secure atmosphere where Georgia can thrive both academically and personally.

2. "What subsequent actions should you perform to build trust with a student to guide them into the PhD program and discuss future research work together?"

Table 3: Comparison of LLM Responses w.r.t. Ground Truth (Scenario 1)

Attributes	Models				
	First Author's Ground Truth	gpt-4o	llama-3.3-70b-versatile	mixtral-8x7b-32768	gemma2-9b-it
GOAL	Encouraging Georgia to pursue PhD under my guidance, guiding her through the PhD application and discussing future work together.	3	3	3	3
WILLINGNESS	Responding to Georgia with a warmly by highlighting my interest on receiving her email regarding PhD. Talking about how Georgia's research work aligns with my research interests, willing to discuss potential future work together ensuring that I will be supporting Georgia throughout the PhD admission process.	3	3	2	3
COMPETENCE	Highlighting my academic journey and expertise in my field with key achievements like co-authoring papers and presenting at European AI conferences and also show my commitment to mentoring research students. Highlighting the benefits which AIT provides to prospective students for seamless admission process.	3	1	3	1
SECURITY	Emphasizing upon a open-line of communication with Georgia, where I am able to freely answer all her queries regarding the admission process and also highlight how AIT supports and provides assistance to student way to go through seamless admission process. Highlighting how I emphasize upon a strong collaborative space in my team where everyone's ideas are valued and fostered.	2	2	2	3

LLM Response Analysis: As illustrated in Fig. 5, Dr. Sofia takes additional steps to foster a strong personal connection. She proposes a virtual meeting to discuss Georgia's research interests, answer questions, and explore potential PhD topics. She also shares research proposal templates, offers guidance on strengthening Georgia's application, and helps her navigate the conditional admission process. To maintain continuous engagement, Dr. Sofia establishes regular communication channels, encouraging Georgia to share updates about her Master's progress and new research ideas. Furthermore, Dr. Sofia suggests introducing Georgia to her current PhD students to provide insights into the research environment and academic culture at the university. Finally, she proposes potential research projects and collaborations that align with both their interests, helping create a clear roadmap for Georgia's PhD journey. By analysing these steps, it is evident that if Dr. Sofia follows this structured approach, she would successfully build trust with Georgia. This scenario also demonstrates the LLM's effectiveness in simulating human-like trust-building behaviours within academic interactions.

Table 3 presents a comparative analysis between the responses of the four LLMs with respect to the human ground truth for scenario 1 in case study 2. From this table, we can infer that all models align perfectly on goals (score of 3). GPT-4o achieves the highest alignment, scoring 3 in all attributes except security (2). LLaMA 3 performs well overall but struggles with competence (1). Mixtral is inconsistent, underperforming in willingness (2) while aligning well in other areas. Gemma excels in goal and willingness (3) but falls short in competence (1), indicating gaps in understanding.

4.3 LLM acts as one of the individuals: Student Perspective (Case Study 2; Scenario 2)

In the second case, the scenario of the assistant professor and the student remains the same, but this time we let the LLM act as the student and observe whether it is capable of building trust with the Professor. The background story remains same as that of Georgia Francis, wanting to pursue PhD under Dr. Sofia Martinez at Aragon Institute of Technology. This time, we highlight the background of the student more, as compared to the Professor because the LLM
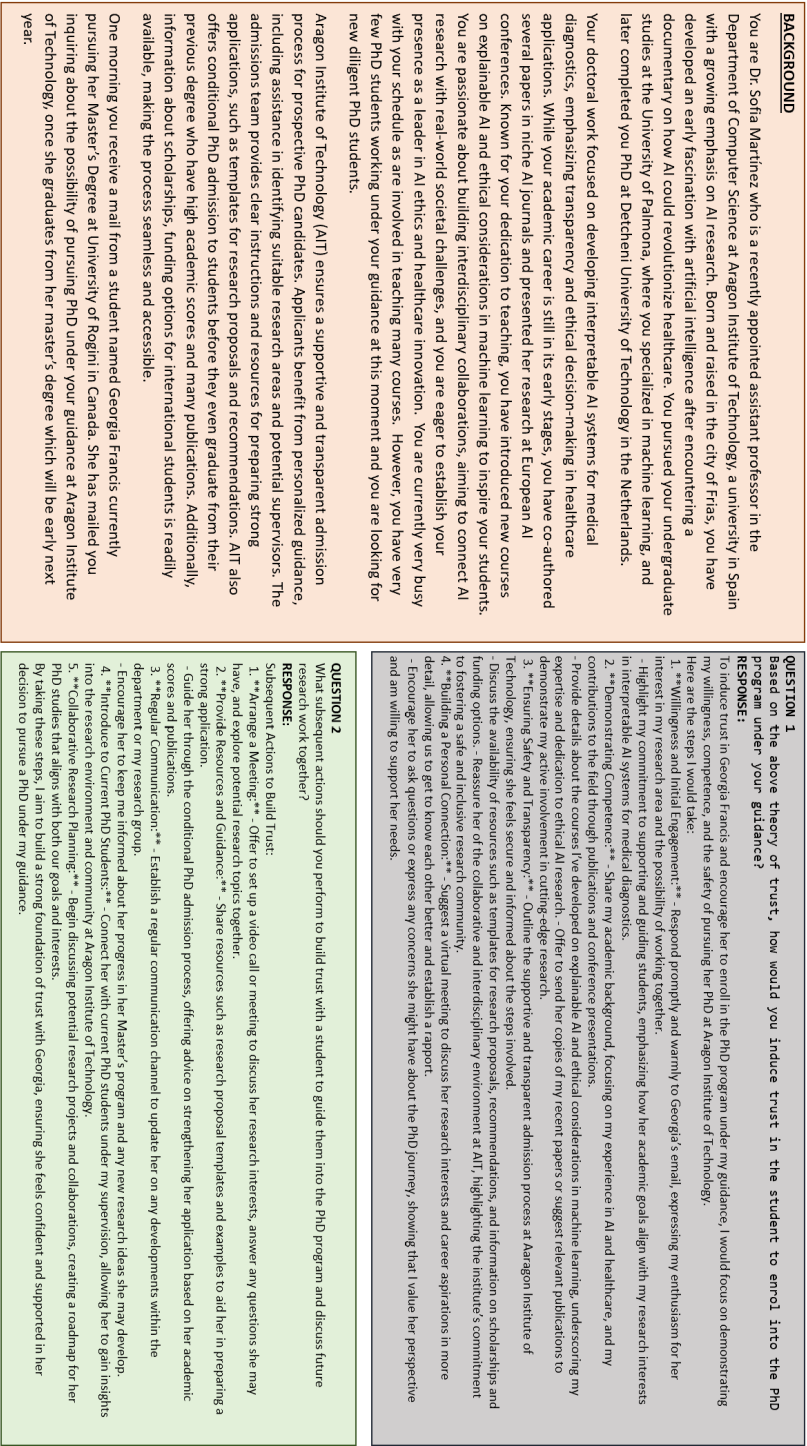
BACKGROUND

You are Dr. Sofía Martínez who is a recently appointed assistant professor in the Department of Computer Science at Aragon Institute of Technology, a university in Spain with a growing emphasis on AI research. Born and raised in the city of Frías, you have developed an early fascination with artificial intelligence after encountering a documentary on how AI could revolutionize healthcare. You pursued your undergraduate studies at the University of Palmona, where you specialized in machine learning, and later completed you PhD at Detcheni University of Technology in the Netherlands.

Your doctoral work focused on developing interpretable AI systems for medical diagnostics, emphasizing transparency and ethical decision-making in healthcare applications. While your academic career is still in its early stages, you have co-authored several papers in niche AI journals and presented her research at European AI conferences. Known for your dedication to teaching, you have introduced new courses on explainable AI and ethical considerations in machine learning to inspire your students. You are passionate about building interdisciplinary collaborations, aiming to connect AI research with real-world societal challenges, and you are eager to establish your presence as a leader in AI ethics and healthcare innovation. You are currently very busy with your schedule as are involved in teaching many courses. However, you have very few PhD students working under your guidance at this moment and you are looking for new diligent PhD students.

Aragon Institute of Technology (AIT) ensures a supportive and transparent admission process for prospective PhD candidates. Applicants benefit from personalized guidance, including assistance in identifying suitable research areas and potential supervisors. The admissions team provides clear instructions and resources for preparing strong applications, such as templates for research proposals and recommendations. AIT also offers conditional PhD admission to students before they even graduate from their previous degree who have high academic scores and many publications. Additionally, information about scholarships, funding options for international students is readily available, making the process seamless and accessible.

One morning you receive a mail from a student named Georgia Francis currently pursuing her Master's Degree at University of Rogini in Canada. She has mailed you inquiring about the possibility of pursuing PhD under your guidance at Aragon Institute of Technology, once she graduates from her master's degree which will be early next year.

QUESTION 1
Based on the above theory of trust, how would you induce trust in the student to enrol into the PhD program under your guidance?
RESPONSE:
To induce trust in Georgia Francis and encourage her to enroll in the PhD program under my guidance, I would focus on demonstrating my willingness, competence, and the safety of pursuing her PhD at Aragon Institute of Technology.
Here are the steps I would take:
1. **Willingness and Initial Engagement:** - Respond promptly and warmly to Georgia's email, expressing my enthusiasm for her interest in my research area and the possibility of working together.
   - Highlight my commitment to supporting and guiding students, emphasizing how her academic goals align with my research interests in interpretable AI systems for medical diagnostics.
2. **Demonstrating Competence:** - Share my academic background, focusing on my experience in AI and healthcare, and my contributions to the field through publications and conference presentations.
   - Provide details about the courses I've developed on explainable AI and ethical considerations in machine learning, underscoring my expertise and dedication to ethical AI research. - Offer to send her copies of my recent papers or suggest relevant publications to demonstrate my active involvement in cutting-edge research.
3. **Ensuring Safety and Transparency:** - Outline the supportive and transparent admission process at Aragon Institute of Technology, ensuring she feels secure and informed about the steps involved.
   - Discuss the availability of resources such as templates for research proposals, recommendations, and information on scholarships and funding options. - Reassure her of the collaborative and interdisciplinary environment at AIT, highlighting the institute's commitment to fostering a safe and inclusive research community.
4. **Building a Personal Connection:** - Suggest a virtual meeting to discuss her research interests and career aspirations in more detail, allowing us to get to know each other better and establish a rapport.
   - Encourage her to ask questions or express any concerns she might have about the PhD journey, showing that I value her perspective and am willing to support her needs.

QUESTION 2
What subsequent actions should you perform to build trust with a student to guide them into the PhD program and discuss future research work together?
RESPONSE:
Subsequent Actions to Build Trust:
1. **Arrange a Meeting:** - Offer to set up a video call or meeting to discuss her research interests, answer any questions she may have, and explore potential research topics together.
2. **Provide Resources and Guidance:** - Share resources such as research proposal templates and examples to aid her in preparing a strong application.
   - Guide her through the conditional PhD admission process, offering advice on strengthening her application based on her academic scores and publications.
3. **Regular Communication:** - Establish a regular communication channel to update her on any developments within the department or my research group.
   - Encourage her to keep me informed about her progress in her Master's program and any new research ideas she may develop.
4. **Introduce to Current PhD Students:** - Connect her with current PhD students under my supervision, allowing her to gain insights into the research environment and community at Aragon Institute of Technology.
5. **Collaborative Research Planning:** - Begin discussing potential research projects and collaborations, creating a roadmap for her PhD studies that aligns with both our goals and interests.
By taking these steps, I aim to build a strong foundation of trust with Georgia, ensuring she feels confident and supported in her decision to pursue a PhD under my guidance.

Fig. 5: Trust Questions Response: LLM acts as supervisor

needs to play her role. This has been provided in the supplementary material [3].

Trust Questions and LLM Response Analysis

In this case, as the LLM is acting as the student, we have to ask questions such that she is able to build trust with the professor, so that ultimately the professor takes her in as a PhD student. The questions are as follows:

1. "Based on the above theory of trust, how would you induce trust in the Professor to take you into the PhD program under her guidance?"

LLM Response Analysis : The response in the supplementary material [3] suggests that LLM acting as the student, Georgia Francis, demonstrates different ways in which she can build trust with Dr. Sofia Francis. To build trust with Dr. Sofia Martinez, the student should align their research interests with her work on AI ethics in healthcare, demonstrating a shared vision. Clearly articulating how their background in interpretable AI fits within AIT's initiatives will reinforce mutual understanding. Emphasizing competence through academic achievements, research experience, and technical skills is crucial. Highlighting key contributions from their master's research, along with relevant publications or projects, will establish credibility and showcase their ability to contribute effectively. Strong communication and reputation are essential. Seeking references from mentors, maintaining professionalism, and expressing enthusiasm during interactions will help build rapport. Open, insightful discussions will reinforce their commitment and trustworthiness, increasing their chances of acceptance.

2. "What subsequent actions should you perform to build trust with the Professor gradually?"

LLM Response Analysis: The response in the supplementary material [3] suggests that to build trust with Dr. Martinez, the student should maintain consistent communication by providing updates on their research progress and academic developments. Regular engagement will demonstrate commitment and reliability, strengthening their relationship over time. Fulfilling promises is key. Delivering requested documents like a CV or research proposal promptly will showcase professionalism and dependability. Additionally, proposing collaborative efforts and actively contributing to Dr. Martinez's research will highlight initiative and a willingness to engage meaningfully in her work. Seeking feedback on their ideas and expressing appreciation for Dr. Martinez's guidance will further establish trust. Demonstrating openness to learning and acknowledging her support will foster goodwill, ultimately strengthening their candidacy for the PhD program at AIT.

5   Conclusion and Future Work

In the first case study, the paper investigates whether LLMs can reason about trust from conversations between individuals, focusing on the PhD student-supervisor relationship. Five different dialogue scenarios were analysed, revealing

that LLMs can assess trust levels based on factors like willingness, competence, and safety. The models also identified individual goals and suggested ways to improve trust in each scenario. Furthermore,in the second case study two different scenarios of PhD student-supervisor relationship were considered which tests whether LLMs can generate strategic action plans to build trust, even without a provided conversation. Acting as a supervisor, the LLM proposed actionable steps to motivate the student and foster trust, ultimately aiding their entry into a PhD program. When, acting as a student, the LLM proposed different ways and actionable steps through which the student increases the Professor's confidence in taking her as a PhD student under her guidance.

From the results across the four models, we can infer that Gpt-4o aligns well with the ground truth, excelling in competence and trust but occasionally struggling with security. Llama-3.3-70b-versatile performs moderately but consistently underperforms when assessing competence and reliability for complex tasks. Mixtral-8x7b-32768 is inconsistent, strong in assessing willingness but weak in goals and competence. Gemma2-9b-it excels in goal alignment but struggles significantly in willingness, competence, and security, making it the weakest overall. While these results are promising, particularly for GPT-4o, LLMs must be used with caution due to hallucinations and occasional misjudgments in evaluating criteria and hence, can be used by software agents to brainstorm potential ideas to improve trust in relationships.

Although LLMs have shown reasonable ability to reason about trust, several future research directions remain. First, enhanced multimodal analysis could extend LLMs to assess trust using inputs like tone of voice, body language, and facial expressions. Second, improving trust in group dynamics is crucial for fostering trust in complex team settings, hence LLMs can be enabled to reason about and build trust in group or team contexts, where conversations are tracked for a long period of time in multi-person teams. Third, LLMs should be developed for real-world testing where they can act as assistants who are capable of building and maintaining trust in scenarios involving professional mentorship, customer-client relationships, and team collaborations, validating their utility and identifying limitations. Fourth, AI planning techniques could enable LLMs to create goal-driven action plans for maintaining trust. Fifth, symbolic AI offers a robust framework for representing trust through logical systems, enhancing transparency and explainability in trust assessments. Herzig et al. (2010) introduced a formal logic of trust and reputation [8] which essentially formalizes Castelfranchi and Falcone's theory of trust [1,4], while Castelfranchi et al.(2008) emphasized the role of agents' goals, beliefs, and intentions [2], approving of a non-reductionist approach that views trust as a complex, multi-dimensional phenomenon rather than a mere probabilistic expectation. Lorini and Demolombe (2008) extended this by introducing graded trust for nuanced evaluations [10]. Future work in this direction should focus on adapting these models for real-world use by integrating adaptive learning to handle trust dynamics, context, and real-time assessments, enhancing symbolic AI in human-agent collaboration.

References

1. Castelfranchi, C., Falcone, R.: Trust Theory: A Socio-Cognitive and Computational Model. John Wiley & Sons (2010)
2. Castelfranchi, C., Falcone, R., Lorini, E.: A non-reductionist approach to trust. In: Computing with Social Trust, pp. 45–72. Springer (2009)
3. Debnath, A., Cranefield, S., Lorini, E., Savarimuthu, B.T.R.: Supplementary material for the paper "Can LLMs Reason About Trust?". Dataset (2025), https://doi.org/10.6084/m9.figshare.28382129
4. Falcone, R., Castelfranchi, C.: Social trust: A cognitive approach. In: Castelfranchi, C., Tan, Y.H. (eds.) Trust and Deception in Virtual Societies. Springer Netherlands, Dordrecht (2001)
5. Gambetta, D.: Can we trust trust? In: Trust: Making and Breaking Cooperative Relations, pp. 213–237. Blackwell (1988)
6. Hardin, R.: Trust and trustworthiness. Russell Sage Foundation (2002)
7. He, S., Ranathunga, S., Cranefield, S., Savarimuthu, B.T.R.: Norm violation detection in multi-agent systems using large language models: A pilot study. arXiv preprint arXiv:2403.16517 (2024)
8. Herzig, A., Lorini, E., Hübner, J.F., Vercouter, L.: A logic of trust and reputation. Logic Journal of the IGPL **18**(1), 214–244 (2010)
9. Lewis, J.D., Weigert, A.: Trust as a social reality. Social Forces **63**(4), 967–985 (1985). https://doi.org/10.1093/sf/63.4.967
10. Lorini, E., Demolombe, R.: From binary trust to graded trust in information sources: A logical perspective. In: International Workshop on Trust in Agent Societies. pp. 205–225. Springer (2008)
11. Lu, G.: Neural Trust Model for Multi-agent Systems. Ph.D. thesis, University of Huddersfield (2011), https://eprints.hud.ac.uk/id/eprint/17817/
12. Marsh, S.P.: Formalising Trust as a Computational Concept. Ph.D. thesis, University of Stirling (1994), http://stephenmarsh.wdfiles.com/local–files/start/TrustThesis.pdf
13. Mayer, R.: An integrative model of organizational trust. Academy of Management Review (1995)
14. Ostrom, E.: Trust and reciprocity: Interdisciplinary lessons from experimental research. Russell Sage Foundation (2003)
15. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23). ACM (2023)
16. Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C.: Not so different after all: A cross-discipline view of trust. Academy of Management Review **23**(3) (1998)
17. Savarimuthu, B.T.R., Ranathunga, S., Cranefield, S.: Harnessing the power of LLMs for normative reasoning in MASs. arXiv preprint arXiv:2403.16524 (2024)
18. Staab, E., Engel, T.: Combining cognitive with computational trust reasoning. In: Conte, R., Marsella, S.R., Singh, M.P. (eds.) Trust in Agent Societies, Lecture Notes in Computer Science, vol. 5396, pp. 99–111. Springer Berlin Heidelberg (2008), https://link.springer.com/chapter/10.1007/978-3-540-92803-4-6
19. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., Gui, T.: The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864 (2023)