Universiteit Leiden
The Netherlands

BSC. Data Science and AI

Hands-on Retrieval Insight and Trust in Legal AI: A Study of Law Students Interacting with RAG Systems

Nadia Xie Xiaoxin Koghee

Supervisors:
Zane Kripe & Wessel Kraaij

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

07/07/2026

Abstract

Retrieval-Augmented Generation (RAG) systems are increasingly adopted in the legal domain as AI tools are integrated into professional settings. Despite the increase in adoption, the legal domain has concerns about whether it can trust and implement AI tools that appear to be black boxes to them. This study investigates how acquiring practical insight into the retrieval process of a RAG system influences law students' trust in its outputs for legal research. Law students participated in a hands-on experiment in which they manipulated the retrieval process and conducted pre- and post-interviews to examine changes in trust. Results indicate that the insight enables more calibrated trust and verification of outputs, while acknowledging the efficiency benefits. These findings suggest that transparency through practical interaction may not directly increase overall trust, but can foster more informed and appropriately calibrated reliance on AI systems in legal research and support situational trust.

Contents

1 Introduction

In recent years, there has been widespread use of Artificial Intelligence (AI)- based applications across professional, educational and personal settings. In the legal domain, there was a 12% increase in the use of Generative AI (GenAI) in 2025 [Thomson Reuters, 2025]. One of the most used Large Language Models (LLMs) is ChatGPT [Orca Security, 2025]. LLMs provide personalised answers to user questions by generating tokens with statistical calculations based on the received input. They can be characterised as probabilistic predictors that generate text [IBM, 2024a]. LLMs enable everyday tasks to be automated and made easier. This allows users to be more efficient with their time and focus on more important tasks. To further optimise the performance of LLMs, researchers and developers continuously improve models and integrate them into newer systems, such as a Retrieval-Augmented Generation (RAG) system. Through combining an LLM with a retrieval component that provides relevant information to the LLM, RAG systems aim to reduce factual hallucinations and biases and improve factual accuracy by grounding responses in retrieved sources [Gupta et al., 2024]. RAG systems are currently employed across various fields, such as law, healthcare, finance, real estate, and others [ORQ.ai, 2024, Jaswal, 2026].

Specifically for the legal domain, AI tools can assist with legal research, case law summarisation, compliance monitoring, and risk assessment [ORQ.ai, 2024]. Outside of these more specialised tasks, people also use LLMs for more mundane tasks such as helping with written communication and translating texts into a preferred language [ORQ.ai, 2024].

Despite their ability to help, risks such as biases and hallucinations arise while using the LLMs, which may result in users' trust decreasing. For instance, if lawyers present incorrect AI-generated facts in court, consequences may have long-lasting effects on both clients and lawyers. Since users cannot rely on LLMs to be correct, their trust in the systems may deteriorate. To mitigate this issue, researchers have developed RAG systems to reduce the probability of generating inaccurate information. Given these risks associated with legal information, users must be able to develop an appropriate level of trust in these systems.

While no study has examined lawyers' trust in GenAI, several reports have examined its adoption. Thompons Reuters' Institute reports an 8% increase among law firms in their report on GenAI in professional services; law firms' main reasons for disagreeing with AI adoption are concerns about reliability and accuracy [Thomson Reuters, 2025]. These reasons are echoed in the 2025 Law360 Pulse AI Survey, with additional recurring arguments addressing legal ethics and standards, client confidentiality, security, data privacy, and the loss of institutional knowledge [Martinson, 2025]. This suggests that some legal professionals do not trust GenAI.

Researchers Lee et al. have identified transparency, along with competence, benevolence, and integrity, as factors that may influence users' trust in AI systems, based on the interpersonal trust relations [Lee and Cha, 2025, Söderström, 2009]. In another study, Zhou et al. define transparency as "Making RAG system processes and decisions clear and understandable to users, fostering trust and accountability" [Zhou et al., 2024]. This definition will also be used in the scope of this study. Ravi and Sindhgatta have examined how transparency in high-stakes domains substantially improved trust along with user understanding [Ravi and Sindhgatta, 2025]. In the field of Explainable AI (XAI), research has continually sought to address the lack of transparency and understanding of model decisions, defining XAI as the ability to solve trust issues in AI. For example, Guttikonda introduces a framework of a RAG system with XAI enabling non-technical professionals to interpret the outputs [Guttikonda et al., 2025].

As a result, understanding how the system retrieves information may influence trust in RAG systems, as this is a major component of the system.
Therefore, given the increasing adoption of AI tools within the legal domain, understanding how trust is formed is becoming increasingly important. Prior studies have examined trustworthiness in Retrieval-Augmented Generation systems, such as Zhou et al. and Ni et al. [Zhou et al., 2024, Ni et al., 2025]. However, Zhou et al. introduce a trustworthiness evaluation model for benchmarks, whereas this study examines how people perceive and evaluate RAG systems in the legal domain. As for the measurement of trust, explainability, fairness and other factors, which are identical to the described factors concluded by Law360's survey, are challenges to fostering trust in the legal domain, as researchers Ni et al. concluded [Ni et al., 2025]. Conversely, no research has examined how trust in AI systems among the future generation of legal professionals is fostered.
The younger generation of legal professionals is currently familiarising themselves with the legal domain. They represent future legal professionals and are continuously exposed to LLMs during their education. Understanding how they develop trust in such systems may provide insight into possible future adoption of legal AI technologies. For wide adoption within the legal domain, RAG systems would need to be consistently accurate, ensure confidentiality, and adhere to existing and proposed legal frameworks.
The objective of this study is to understand how non-expert users in a high-stakes professional domain develop appropriately calibrated trust in RAG systems they cannot fully verify or control. This thesis focuses on law students gaining hands-on insight into the inner workings and accuracy of a RAG system. This might provide a more in-depth view of how RAG systems work, or, conversely, confuse participants.
Guttikonda et al. argue that, within the field of XAI, explanations of models are limited to technical users due to the models' technical complexity [Guttikonda et al., 2025]. This may also occur with hands-on insight. If the explanations and RAG system are confusing, they might not make the system understandable to non-technical users.
This study aims to examine the following question: Does acquiring practical insight into an RAG system's retrieval process affect law students' trust in its outputs for legal research?

1.1 Thesis overview

This thesis follows the following structure. Chapter 1 introduces the research context and research question. Chapter 2 reviews the relevant literature about trust and transparency. Furthermore, it describes how perceptions of trust change as information about the inner workings of a RAG system becomes available. Additionally, it will briefly discuss the limitations of RAG systems. Chapter 3 describes the methodology of the study, including the experimental design and data collection process. Chapter 4 presents the results of the experiment, and 5 discusses the findings in relation to the literature. Finally, Chapter 6 concludes the thesis and illustrates the limitations and future directions for future research.

2 Background

2.1 RAG systems

RAG systems are architectures combining information retrieval with LLM generation. RAG systems are developed to enhance LLMs by incorporating external knowledge sources, thereby improving factual accuracy and reducing hallucinations. The architecture of a RAG system grounds outputs in retrieved information and can reduce the likelihood of hallucinations and biases occurring (Section 2.6.1 and Section 2.6.2).
As a result, RAG systems are useful in high-stakes domains, as inaccurate or unsupported information can have significant consequences. In the legal domain, inaccurate information can lead to significant legal, financial, or societal consequences for an individual or organisation. With an RAG implementation, the likelihood of inaccurate information is lower than if they were to use LLMs.
No empirical studies have examined the adoption or use of RAG systems specifically in the legal domain; however, researchers studied this topic within the boundary of GenAI. Thomson Reuters Institute researched GenAI in professional Services. They reported an increase in 2025 in the usage of GenAI in the legal domain compared to 2024. Thomson's highest reported implementation of GenAI in the legal domain is in document reviews, legal research and document summarisation [Thomson Reuters, 2025].

2.2 Limitations of RAG

The earliest methodology has limitations, resulting in inconsistencies in accuracy within outputs [Gao et al., 2024]. To address the limitations of previous systems, newer architectures introduce methods that surpass the limitations of previous RAG types. However, limitations might still occur, since the improvements do not eliminate them. Major differences occur in retrieval, indexing, document processing, memory and more.

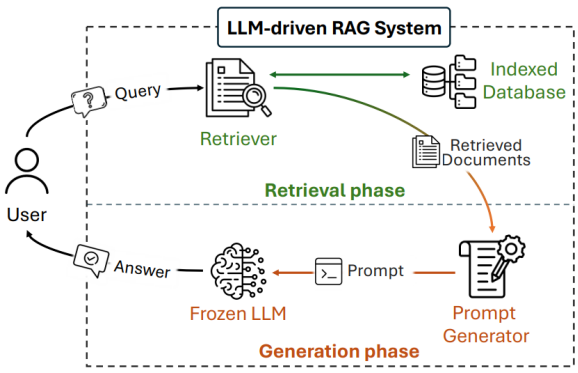

Figure 1: The typical workflow of an LLM-driven RAG system [Zhao, 2025]

Before implementing a RAG system, a user or developer adds external knowledge to the knowledge base. The system cleans diverse text formats containing data, such as PDF or HTML, and extracts raw data. It converts the data into a uniform plain-text format and segments it into smaller chunks to accommodate language models' limitations. An embedding model encodes the chunks into vector representations and stores them in a vector database [Gao et al., 2024].

A RAG system consists of two phases: the retrieval phase and the generation phase. In the first phase, the retrieval phase, the retriever component retrieves relevant information from the knowledge base (an indexed database, as illustrated in Figure 1). In the second phase, to correctly answer the user query, a prompt constructor combines the retrieved information with the user task. Subsequently, an LLM generates a response to the combined prompt [IBM, 2024c, Zhao, 2025].
A single iteration, as illustrated by Figure 1, starts with the user prompt. This query is transformed into a vector representation with the same encoding model used during the document processing phase. After computing similarity scores between the query vector and the chunk vectors in the indexed knowledge base, the system retrieves the top K chunks with the highest similarity score to the query [Gao et al., 2024].
Using the retrieved chunks as context, the prompt generator combines the initial user prompt and retrieved chunks into a prompt for the LLM. Subsequently, the user receives an answer to their prompt [Gao et al., 2024].
However, the retrieval phase often encounters problems with precision and recall. This leads to the selection of irrelevant chunks and the absence of crucial information [Gao et al., 2024]. Furthermore, during generation, the model may hallucinate (Section 2.6.1), producing text not supported by the retrieved information. In addition to hallucinations, RAG systems may suffer from irrelevance, toxicity, or bias (Section 2.6.2) in their outputs [Gao et al., 2024]. Moreover, systems may experience difficulties integrating retrieved information with the user task. Additionally, the systems may excessively rely on the retrieved information. In both cases, this can result in disjointed or incoherent outputs [Gao et al., 2024].
Therefore, this might result in complications when implementing RAG systems' outputs for legal professionals. Since accuracy is inconsistent, system reliability diminishes. This might lead users to trust the outputs improperly and, therefore, the consequences.

2.3 Transparency and Explainability in RAG Systems

The initial trust users might experience with RAG systems may develop into a more experienced type of trust, as explained in Section (2.8.2. However, to properly develop trust, various factors have a role, including transparency, which Section 2.8 will explain. When users use LLMs, they might not understand the model's inner workings. Developers of LLMs may refrain from maintaining transparency regarding the system's procedures, capabilities, and objectives to preserve competitive advantages [Hassija et al., 2024]. This results in a "black-box".
The field of XAI attempts to make these black-box models understandable for humans. Hassija explains that significant efforts are made to transform black-box models into more interpretable models. These models with increased transparency are referred to as white-box models. These terms are commonly used in contrast to one another [Hassija et al., 2024]. With the conversion of black-box models, XAI attempts to mitigate biases and hallucinations.
By attempting to explain the algorithm's behaviour with (global) interpretability, i.e. comprehensibility, XAI aims to explain the behaviour in understandable terms to an individual [Guidotti et al., 2018]. By doing this, XAI aims to increase algorithmic literacy. Furthermore, this can result in calibrated trust. This trust evolves through trust calibration. Okamura et al. state that, to calibrate trust, a user has to adjust their level of trust based on the actual reliability of an AI agent [Okamura and Yamada, 2020].
However, Renftle concludes that it may not be entirely possible to translate the behaviour into

comprehensible language for an individual. Stating that both the approximation and translation of models will be challenging in future research [Renftle et al., 2024]. Therefore, it might not be possible to properly calibrate a user's trust.
Furthermore, for each prompt tasked to an LLM, local interpretability is needed to explain the outcome of a specific output [Guidotti et al., 2018, Hassija et al., 2024]. This provides even more insight into the specific reasoning behind the model's behaviour for a single output.
While RAG systems have ensured more grounded outputs for users, ensuring transparency in how retrieved information is selected and used in generation is an important factor for fostering trust [Gupta et al., 2024]. This is also extended to LLMs. By providing more insight into AI models, XAI seeks to achieve a goal, such as trust [Renftle et al., 2024]. This is crucial for the integration of AI, as greater trust in AI is associated with higher acceptance of AI technologies, which, in turn, leads to more frequent use [KPMG, 2025].
Moreover, Shin argues that alongside trust, the concept of FAT (Fairness, Accountability, and Transparency) is important for AI adoption. Additionally, Shin explains that it should be expanded to include Explainability in FATE. FAT(E) is a method for judging literacy in algorithms exploiting AI while avoiding biased and discriminatory decisions. Shin describes how explainability, along with the other FAT components, is a core component of algorithmic literacy and a direct antecedent to credibility. This allows users to foster trust and to turn more towards AI without the burden of doubt [Han et al., 2026].

2.4 User Understanding of RAG Systems

Various cases illustrate negative interactions with algorithms, LLMs or other GenAI models' outputs, such as Eslami et al.'s case. In this study, the authors concluded that prolonged or passive use of algorithmic systems did not necessarily lead to understanding of the underlying algorithmic processes. Instead, users required explicit interventions to understand how the algorithm operated [Eslami et al., 2015]. Moreover, since users were aware of the algorithm, they drew incorrect conclusions about their relationships while using the News Feed. Therefore, to stop users from drawing incorrect conclusions, hidden algorithms, including their effects, should be explained to them.
In the media, more negative interactions with LLMs are reported, such as news articles about the Mata v. Avianca case. Two lawyers submitted six hallucinated law cases by ChatGPT [Neumeister, 2026, Brodkin, 2023]. Or in the following news article: Glenda Raap reports how some clients send in legal documents that are generated with AI, and Danny Vesters states that lawyers spend increasingly more time reading through these documents to fact-check what information is useful [nlt, 2026].
Besides LLMs being viewed negatively, these cases also show users may not understand the risks of inaccuracies with black-box outputs. As for RAG systems, they can be more transparent than LLMs, since their architectures are usually provided. However, how the LLM generates the responses is not explained. Additionally, users may still not understand their behaviour, such as why documents are retrieved or how relevance is determined.

2.5 RAG: Transparency through Practical Insight

To increase appropriately calibrated trust in RAG systems, XAI suggests making aspects of the system more understandable to users. In the context of RAG systems, a form of transparency could be to provide insight into the system's architecture, and more specifically, the retrieval process. Explaining how the system selects relevant information and incorporates it into the generated response may help users better understand how the response is generated, enabling them to evaluate its reliability in depth. These explanations may improve algorithmic literacy and help users distinguish between situations in which the system is likely to be reliable and those in which additional verification is required. Despite numerous studies on the trustworthiness of RAG in the legal domain, there has been no research on whether providing users with practical insight into an RAG system's retrieval process actually influences their trust. This thesis will therefore investigate whether exposing users to the retrieval process changes their trust in an RAG system.

2.6 Problems

RAG systems need to be reliable to be implemented within the legal domain. This section discusses problems occurring in RAG systems and LLMs.

LLMs have multiple problems, resulting in hallucinations and biases appearing in outputs. Moreover, some LLMs are presented as black-box models [Mittal, 2023], leading to uncertain reasoning about why these outputs occur.

RAG systems implement these LLMs into their architecture. This can result in RAG systems' outputs also containing hallucinations and biases. Furthermore, researchers have identified several emerging issues with RAG systems. Particularly, there is concern about the static knowledge base. Moreover, retrieval mechanisms may struggle to retrieve the most relevant information, and integration between retrieval and generation may fail in practice.

Further, apart from the technical challenges, the application of RAG systems raises important ethical and legal concerns. In particular, regarding confidentiality and accountability.

The law requires legal professionals to maintain the confidentiality of their clients. When using RAG systems or other AI tools, external providers may process sensitive documents, increasing the risk of data breaches or unauthorised access. To protect personal data, the General Data Protection Regulation (GDPR) governs the processing of personal data and imposes obligations on data controllers and processors [European Parliament and Council of the European Union, 2016].

Concerning AI, the current legislation is still evolving. The AI Act is the EU's AI legislation. It is a risk-based framework that distinguishes between providers and deployers of AI systems. It assigns providers additional responsibilities to help protect deployers [European Parliament and Council of the Europea

However, there is also other legislation that applies to AI systems, such as the GDPR and the Data Service Act.

Despite this, clear liability rules for damages caused by AI systems remain under development and are therefore unclear. A proposed AI Liability Directive aims to address these challenges by adapting existing liability rules and easing the burden of proof for individuals harmed by AI systems, [European Commission, 2022]. However, it has not yet been adopted.

Furthermore, uncertainty persists about who is ultimately responsible for incorrect or misleading AI-generated outputs, posing significant challenges for the adoption of AI in high-risk areas, such as the legal field. These aspects may discourage lawyers from implementing and trusting RAG

systems and other AI tools.

2.6.1 Hallucinations

Text generation in LLMs occurs token by token, using statistical calculations to determine what is most likely to follow, resulting in arbitrary outcomes. Thus, the model may produce outputs not grounded in facts. "Hallucinating" arises when a model generates text that misleads users by presenting inaccurate outputs as factual and reliable information [IBM, 2026, Terzidou, 2025, Bender et al., 2021].
A model's architecture, prompt design and domain specificity may influence these hallucinations [Alansari and Luqman, 2026]. They may take different forms, such as incorrect facts, fabricated sources, flawed reasoning, or misinterpretation of context [Anh-Hoang et al., 2025].
When the LLM starts hallucinating, the resulting output might appear to be applicable and relevant despite lacking legitimacy. Even though the generated content is unusable, considering it is not factually correct [nlt, 2026].
Epistemology provides an analytical framework for understanding the implications of hallucinations in LLMs. As a branch of philosophy, epistemology examines the nature and justification of knowledge [Steup and Neta, 2005]. This perspective is relevant to LLMs because their responses are generated through statistical prediction rather than mechanisms that verify factual correctness or justify their claims. Consequently, hallucinated outputs may appear coherent and authoritative despite lacking truth and evidential support. From an epistemological perspective, these outputs represent failures of knowledge production rather than merely incorrect information, as they fail to meet the epistemic standards of truth and justification despite appearing to be reliable knowledge.
This epistemic limitation is problematic in the legal domain, where justification, traceability and reliability are essential to validating legal reasoning [Li et al., 2026]. Legal arguments must be supported by verifiable sources and factual reasoning. When hallucinated outputs are presented without factual grounding or justification, they undermine epistemic standards required for legal reasoning and decision-making.

2.6.2 Bias

Kordzadeh et al. defined algorithmic bias as the phenomenon of producing discriminatory results violating norms of justice and equality, adversely impacting particular people or communities in workplaces or society [Kordzadeh and Ghasemaghaei, 2022]. They can emerge in AI models, including RAG systems, for various reasons [IBM, 2024b]. Training data may contain social biases. The selection of algorithmic design with respect to features, weights, and the objective function may also lead to algorithmic biases [Kordzadeh and Ghasemaghaei, 2022].
Throughout the entire AI model pipeline, the Sources of Harm in Machine Learning identified seven types of biases [Suresh and Guttag, 2021]. Suresh et al. constructed a pipeline of two distinct phases: data generation and model building. During data generation, historical, representation, and measurement biases can be embedded in the collected data. Throughout model building and implementation, learning, evaluation, aggregation and deployment biases may occur [Suresh and Guttag, 2021].
Researchers define data containing social biases as flawed. Flawed data as non-representative, lacking information, historically biased or otherwise "bad" data [IBM, 2024b]. The embedded biases lead the model to produce unfair outcomes and amplify pre-existing biases in the data [IBM, 2024b].

The amplification may occur through a feedback loop, where the biases can dynamically evolve. With the deployment of an AI system, the post-deployment data and decision-making processes may also reinforce and amplify the pre-existing biases, in addition to reflecting them [Hari, 2025]. Furthermore, AI models may identify patterns in unbiased training data based on correlations rather than causations. This causes the model to fail to filter on the intended factors [IBM, 2024b]. Biases in training data also manifest in proxy data, which AI systems utilise as substitute variables for protected traits, such as gender [IBM, 2024b]. However, these proxies may unintentionally or deceptively correlate with the sensitive attributes they are meant to replace. This results in the AI model "learning" incorrect correlations, allowing bias to arise in outputs [Kordzadeh and Ghasemaghaei, 2022].

Moreover, AI models that process Natural Language (NL) often contain more implicit rather than explicit biases. As a result, the embedded biases are in associations rather than direct statements [Zhao et al., 2025]. As LLMs learn statistical relationships between words and concepts from training data, their generated outputs can reflect societal stereotypes. As a result, biased outcomes may arise through subtle associations between groups and specific attributes or roles [Zhao et al., 2025]. In addition, evaluation bias occurs when an algorithm's results are interpreted based on subjective perceptions rather than the objective findings. Consequently, although an algorithm may be impartial and based on data, the implementation of its outputs might nevertheless result in discriminatory consequences [IBM, 2024b].

Apart from biases embedded in data, algorithmic design may also introduce bias. Programming errors may unknowingly transfer into the system. Developers may apply weighting to reduce bias by adjusting the data to better reflect the population. It can inadvertently introduce bias and inaccuracy through improper weights. Another possibility is for developers to intentionally or unintentionally implement algorithms with subjective rules based on their biases. Biased outputs may lead to direct or indirect discrimination, reinforcing social inequalities and resulting in unequal treatment of individuals or groups [IBM, 2024b].

These concerns of bias are relevant to RAG systems as well. Even though RAG enhances factual grounding by incorporating external knowledge during text generation, it does not necessarily eliminate algorithmic bias. Bias may already be present in the underlying LLM, while the documents in the indexed database may contain historical, societal, or other biases. Furthermore, during retrieval, depending on the algorithmic design, particular sources or perspectives may be favoured through ranking algorithms or indexing strategies. This may potentially reinforce existing biases in generated responses. Consequently, RAG systems may produce discriminatory or unfair outputs despite their improved retrieval capabilities. This is especially a concern in the legal domain, as biased outputs may influence legal decisions and risk legitimising unfair structures [Lendvai and Gosztonyi, 2025].

2.6.3 Black-box

A "black-box" model in XAI often has the internal workings and reasoning behind the predictions unavailable to the user.

Therefore, these systems are commonly referred to as "black box algorithms" [De Streel et al., 2020] or "black-box models" [Hassija et al., 2024]. Due to this lack of transparency, the sources of generated outputs are unverifiable. This results in increased caution towards AI adoption, particularly in high-stakes domains. In the legal domain, where legal outcomes can heavily influence an individual's life, such uncertainty may influence judicial decision-making and, consequently, case

outcomes.

2.6.4 Limitations Introduced by External Knowledge Bases in RAG Systems

Whilst RAG systems are developed to mitigate hallucinations and biases, they still have limitations. RAG systems are very dependent on their external database, as it justifies the reasoning behind the generated outputs. Yet, RAG systems may become obsolete since the database is static [RAGFlow, 2026b]. Subsequently, responses may become inaccurate over time [Petroni et al., 2024].

2.6.5 RAG: Retrieval Mechanisms and Integration

If the retrieval mechanisms fail to retrieve the most relevant information or the integration between retrieval and generation fails in a RAG system, the output will be ineffective for the user.
In addition, if the RAG system were to fail while the user presumes the outputs are correct, the user would accept inaccurate information and might apply it to their work.
In the legal domain, the same consequences may occur as with hallucinations (Section 2.6.1), as the output used in the legal domain would not pass the analytical framework.

2.7 Solutions

2.7.1 Hallucinations

Researchers attempt to mitigate hallucinations and biases by transforming black-box models into white-box models. White-box models allow insight into the behaviour behind the outputs. However, a limitation is that human capability to interpret transparent information from AI is a prerequisite for transparency to increase trust [Lee and Cha, 2025]. With this information, hallucinations and biases could be mitigated, as it may increase algorithmic literacy and provide insight into what needs to be adjusted [Hassija et al., 2024].
RAG systems mitigate hallucinations and biases by grounding their outputs in factual information. Additionally, by fine-tuning a RAG system for a specific domain, such as the legal domain, it can achieve exceptional levels of accuracy and reliability [Gangavarapu, 2025].

2.7.2 Bias

Explicit biases are generally easier to detect and mitigate, as they manifest as conscious attitudinal tendencies [Zhao et al., 2025]. They are typically more observable, and once identified, can be addressed through direct interventions, such as adjusting model weights or modifying decision thresholds.
In contrast, implicit biases pose a greater challenge as they reflect unconscious automated cognitive processes that operate through associations rather than explicit statements [Zhao et al., 2025]. However, there are various pre-, in- and post-processing methods to improve the data, such as human-in-the-loop approaches.
In RAG systems, it is also easier to locate the explicit biases in the data. This allows the outputs to remain neutral and non-discriminatory. However, in the LLM's training data, this is more complicated because the LLM is integrated into the RAG system. This could result in the biases being removed from the RAG's knowledge base, yet the possibility that the biases are present in the

LLM might still influence the RAG. Therefore, it has consequences similar to those of hallucinations in the legal domain (Section 2.6.1).

2.7.3 RAG: Knowledge Base

Users or developers of the RAG system can periodically update the external database during deployment to avoid using outdated information. Additionally, it can mitigate possible biases or hallucinations.

2.7.4 RAG: Retrieval Mechanisms and Integration

Various strategies have been implemented, including pre- and post-retrieval techniques such as prompt rewriting, to ensure that the retrieval component works.

2.8 Trust

Blöbaum expresses how Mayer et al. define trust as the willingness of a trustor, or trusting party, to be vulnerable to the actions of a trustee, or trusted party [Söderström, 2009], whilst the trustee is to perform a specific action [Blöbaum et al., 2016]. While Mayer's original model was for organisational trust, it has become a widely used model of trust [Kohn et al., 2021]. This model is widely adopted in the Trust in Automation (TiA) community and has also been extended to the AI community [Toreini et al., 2020, Kohn et al., 2021].
According to the model, trust can be interpreted as a person's assumption regarding another party's future behaviour. There is a subjective belief called benevolence, described in section 2.8.1, where a trustee will exhibit reliable behaviour to maximise the trustor's interest under uncertainty of a specific situation based on the subjective assessment of past experiences with the trustee [Cho et al., 2015]. As a consequence of this subjective belief, people become reliant on external entities —whether people, organisations or tools —and must accept the risks associated with this reliance [Justwan et al., 2018].
Justwan et al. illustrate this reliance as a willingness to be vulnerable [Justwan et al., 2018]. The trustor assumes that the delivered product is suitable for their intentions; if this condition is not met, the outcome may be pernicious to the trustor [Justwan et al., 2018]. In the legal domain, this is illustrated by a lawyer's willingness to use the output of a RAG system. They will assume that the given output is the most relevant to their prompt. However, if the output is incorrect, this will have consequences for their work and their client.

To distinguish among different forms of trust, Söderström identifies "organisation trust" (*organisational trust*), person trust (*interpersonal trust*) and "technology trust" (*technological trust*) in her research [Söderström, 2009]. More recently, researchers have argued that users of LLMs apply *interpersonal trust* to these systems, as they parallel aspects of the trust humans place in one another [Ng and Zhang, 2025]. This is because trust is based on antecedents such as perceived ability, benevolence, and integrity (section 2.8.1), while chatbots can create personalised interactions through tailored responses. As a result, users may anthropomorphise these systems and perceive them as social actors with human-like characteristics rather than merely technical tools.
In addition, an entire market has formed from these human-like chatbots. Multiple cases describe users developing emotional or romantic relationships with such companion-oriented chatbots, such

as Replika, entrusting them with personal information [Replika, 2026, Batty, 2025, Turnbull, 2023]. These cases demonstrate that users can anthropomorphise conversational AI. Although RAG systems are primarily designed as information retrieval tools, they also display conversational characteristics. This may change users' perceptions, leading them to view them more as an anthropomorphic system. Consequently, interpersonal trust frameworks may also apply to RAG systems.
Furthermore, the following section (Section 2.8.1) will provide insight into how we evaluate the trust types in Section 2.8.2.

2.8.1 Evaluation of Trust

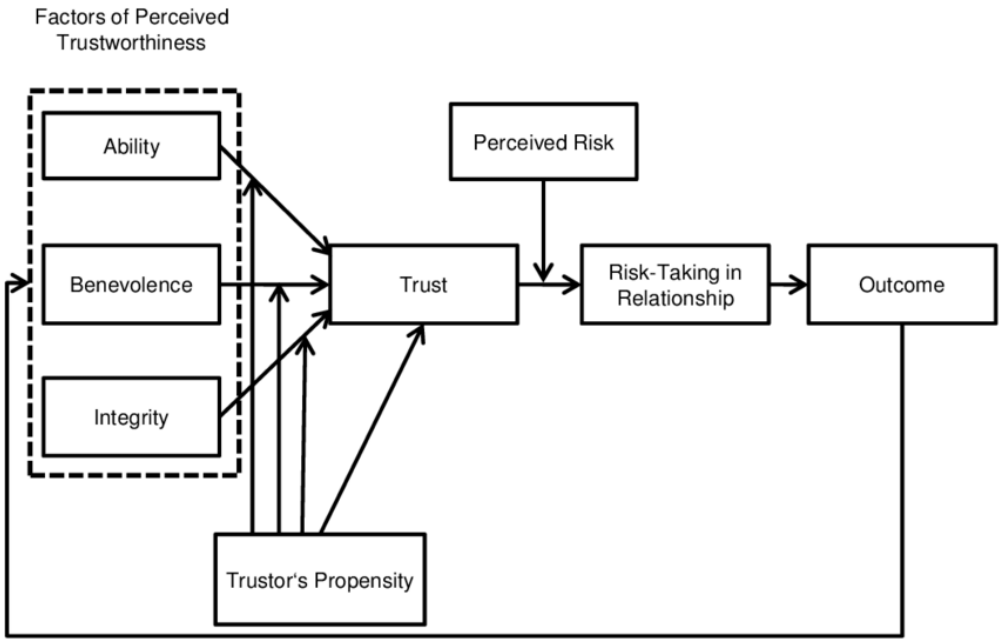

Figure 2: Trust model based on Mayer, Davis and Schoorman 1995 [Söllner et al., 2012]

Figure 2 is a formative trust model created based on Mayer, Davis and Schoorman [Söllner et al., 2012]. Within this model, *Ability*, *Benevolence*, and *Integrity* form the baseline of a model on how to foster trust and are supported by *Trustor's Propensity*. The three factors: Ability, Benevolence and Integrity, are considered factors that cause the trustor to trust the trustee and therefore are factors of perceived trustworthiness between humans [Söllner et al., 2012]. In addition, Söllner et al. applied the framework to research large language model-based coaching [Meywirth et al., 2026]. Söllner demonstrated the applicability of this trust model for AI. This research uses the same dimensions. However, in this study, ability will be the primary focus. Benevolence and Integrity are outside the scope of this study, as they reflect the extent to which the trustee is believed to assist the trustor and the trustee's adherence to the trustor's beliefs.
In the AI and RAG context, benevolence and integrity may be less applicable, since prior research suggests that some artificial agents may not be perceived as having intentions or self-directed goals [Kohn et al., 2021]. However, it may still be applicable as the trustor holds this perception. Moreover, Integrity captures adherence to moral and ethical principles. Whilst it can be argued

that these also apply to RAG systems working in the legal domain, this study examines whether increasing transparency in the retrieval component of an RAG system influences their beliefs about the system's accuracy and inner workings. Since this research only allows practical insight into the retrieval component, Benevolence and Integrity of the system fall beyond the scope of this thesis. Accuracy and transparency may solidify Ability, the third factor in trust. Shin attempts to increase algorithmic literacy and help towards credibility through the concept of FAT(E) [Shin, 2022]. High accuracy may indicate the system's ability to perform effectively and, consequently, its reliability. Therefore, transparency and accuracy influence perceived trust in a system and can be argued to support that trust. Lee et AL. show that transparency can positively influence perceived trustworthiness by strengthening perceptions of the above-mentioned dimensions [Lee and Cha, 2025].
By showing insight into the retrieval component of the RAG system. This study aims to increase participants' algorithmic literacy and enable them to calibrate trust appropriately. The participants will be able to visualise how accurate the retrieval process is and what influences it. Therefore, their experience will ground their calibrated trust rather than in a belief about whether the RAG system is working as intended.
However, in addition to these three dimensions, the Trustor's Propensity influences how the changes affect an individual's trust. It is an expectancy to be able to rely on the promises or statements of others [Söllner et al., 2012]. This expectancy is held by either an individual or a group [Söllner et al., 2012]. This variable shows how it influences an individual; some individuals allow their perceptions to adjust with less new information, whilst others require more information to adjust their trust and perceptions.

2.8.2 Interpersonal Relations of Trust

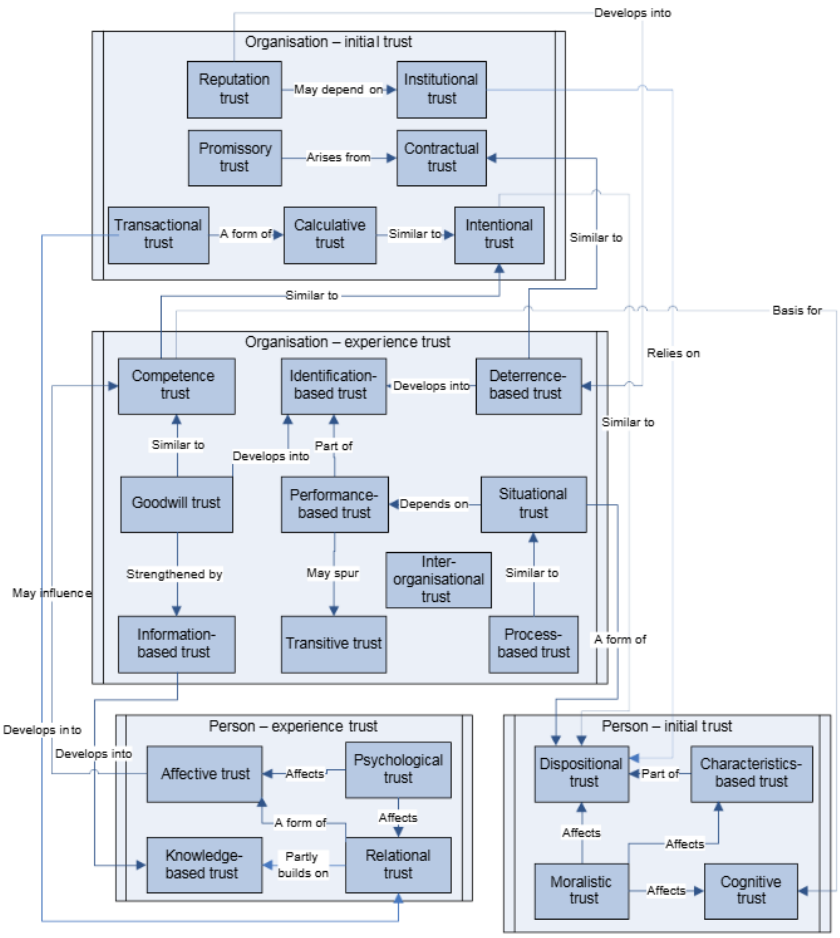

Figure 3: Trust relations, Söderström

Söderström divided person trust into two categories: initial trust and experience-based trust, as shown in Figure 3. Initial trust refers to the trust that formed before interaction with the other party, in this case, a RAG system. This is often influenced by reputation, prior knowledge about the technology, general attitude towards technology or the designer. Experience-based trust develops with repeated interaction with the system and is shaped by perceived behaviour, consistency, and reliability. Additionally, competence trust, information-based trust and situational trust are examined. Although Söderström classifies these under a different type of trust, they are directly influenced by, or influence, person trust. Therefore, they are more connected to people's perceptions than to the organisation [Söderström, 2009]. In the context of this study, these trust types provide additional information on how users form trust.

These different relations encompass the types of interpersonal trust this study will apply when discussing RAG systems. **Competence trust** refers to the expectation that the other party can

execute the stated requirements competently [Söderström, 2009]. **Cognitive trust** is similar to this; it refers to the expectations and first impressions a person holds of the other party.
In the context of the study, this implies that cognitive trust is a tentative expectation that a RAG system will retrieve the correct information, and competence trust develops when the RAG system consistently retrieves the correct information. This is strongly influenced by the system's ability and, therefore, its accuracy [Sassenberg et al., 2026].

Competence trust relates to information-based trust through another form of trust—goodwill trust, which reflects benevolence [Söderström, 2009]. However, this lies outside the scope of this study. **Information-based trust** is the prediction of behaviour after collecting information, when beginning direct interaction [Söderström, 2009].
The following example illustrates this: A user attempts to use a RAG system for the first time. Throughout their session, the user becomes familiar with how the system answers in the chat. They form an initial belief about how the system will behave. As interactions increase, the information-based trust develops into knowledge-based trust. When repeated interactions develop it into knowledge-based trust, the user knows what to expect from the system and can predict its behaviour after many interactions.

Cognitive trust is indirectly influenced by individuals' broader worldviews and moral values, which shape their perceptions of a system's integrity and trustworthiness [Söderström, 2009]. In the context of RAG systems, transparency and integrity play important roles in this process. By providing insight into how the system retrieves and generates information, users can better assess whether the behaviour aligns with their expectations and ethical and moral principles.

Information-based and knowledge-based trust are more strongly correlated with transparency, as information-based trust involves gathering information about a prospective partner, which develops into **knowledge-based trust** when partners become familiar through repeated interaction and therefore understand each other's strengths and weaknesses. In terms of AI, information-based trust involves finding information about the AI tool, understanding what the product does, and how to apply it in a professional setting. Gradually, the user can test the tool and apply it in their work, developing knowledge-based trust. It will then become clear where the AI tool can assist and where it becomes a hindrance.
Additionally, outside of these "external" relations, **situational trust** plays a significant role in utilising AI. Situational trust is shaped by several factors, such as the task, risk, and past outcomes [Söderström, 2009]. There are situations in which a user places their trust in the AI, and others in which that trust is withdrawn for specific reasons [Söderström, 2009]. Such situations can occur when an AI tool repeatedly returns incorrect information or is unable to assist. Therefore, trust is neither static nor linear; it can increase or decrease depending on the situation, reflecting calibrated trust.

2.8.3 Previous Studies on Trust

As prior research and this study show, a central assumption in XAI is that increasing the transparency of AI systems can foster trust. In their study of ChatGPT, Lee and Cha found that transparency positively influenced trust, even though trust is not solely reliant on transparency

[Lee and Cha, 2025]. In addition, accuracy reduces distrust when interacting with functionality and helpfulness [Lee and Cha, 2025]. Therefore, Lee and Cha argue that the relationship between trust and transparency is dynamic and that understanding the mechanisms underlying this relationship is essential for designing trustworthy AI systems [Lee and Cha, 2025]. This view is also reflected in the 2025 KPMG Global Trust in AI report, which identifies trust as a key prerequisite for AI acceptance and adoption. Although AI use continues to increase globally, only 46% of respondents reported being willing to trust AI systems, underscoring the need to improve transparency, governance, and AI literacy to support informed reliance on AI technologies [KPMG, 2025]. This view is also reflected in the 2025 KPMG Global Trust in AI report, which identifies trust as a key prerequisite for AI acceptance and adoption. Although AI use continues to increase globally, only 46% of respondents reported being willing to trust AI systems, underscoring the need to improve transparency, governance, and AI literacy to support informed reliance on AI technologies [KPMG, 2025].

In contrast, other studies argue that the effect of transparency is heterogeneous and depends on how users interpret the information they receive. Eslami et al. show that providing users with explanations of the Facebook News Feed algorithm did not uniformly increase trust [Eslami et al., 2015]. On the contrary, the same transparency intervention yielded varied outcomes depending on users' perceptions and interpretations of the algorithm's intentions. Some users reported greater understanding and trust; others became more sceptical after learning how the content was filtered and prioritised [Eslami et al., 2015].

This finding is relevant to RAG systems, where retrieved documents are similarly selected based on relevance criteria that are often unknown to users.

Furthermore, Kizilcec also found that greater transparency does not necessarily translate into greater trust. Even though explanations can increase users' understanding and sense of control, they may also expose them to the complexity and arbitrariness of algorithmic decision-making, leading some users to become less trusting despite the increased information available [Kizilcec, 2016]. These studies suggest that transparency does not increase trust but rather influences it through users' existing beliefs, expectations, and interpretations (the Trustor Propensity in Section 2.8.1). This raises the question of whether exposing users to the retrieval process and components of a RAG system will increase, decrease, or recalibrate trust.

2.9 Research Statement

Mayer's theoretical model assumes that by improving the perceptions of the system's ability through transparency, users' trust will also increase. From this perspective, understanding how the RAG system's retrieval component works should increase trust in the system.

However, other research, as shown in Section 2.8.3, has shown that this relationship is less straightforward than Mayer's model implies. The research indicates that transparency on its own does not consistently increase trust.

Instead, it relies on other factors such as accuracy and risk perception, as well as the Mayer's trustor's propensity. However, transparency may improve trust calibration rather than increasing trust overall. This helps users distinguish when reliance on RAG systems is appropriate and when it is not.

In the legal domain, the distinction on when to apply RAG systems is important during the legal decision-making process. Legal professionals may rely on RAG systems to aid them. However, many

users utilise these systems without understanding how they work. This incomplete understanding may lead to miscalibrated trust, in which users blindly rely on RAG systems. This can have significant consequences for the legal professionals and their clients.
Existing studies have examined trust in RAG among legal professionals. However, researchers have not examined professional novices such as law students. Whilst they are still learning to navigate through the legal domain, they have access to GenAI throughout their education; however, they do not have in-depth knowledge of AI. Furthermore, researchers have not examined hands-on engagement with the retrieval mechanism among this participant group. Therefore, it remains unclear whether acquiring practical, hands-on insight into the retrieval process of a RAG system influences how these users calibrate their trust in its output.
This introduces the following research question. **RQ:** How does acquiring practical insight into the retrieval process of an RAG system affect the trust of law students in its outputs for legal research? With the following hypotheses:

- H_0: Acquiring practical insight into the retrieval process of a RAG system does not significantly affect the trust of law students in its outputs for legal research.
- H_1: Acquiring practical insight into the retrieval process of a RAG system significantly affects the trust of law students in its output for legal research.

3 Research Design

3.1 Research Approach and Participants

This study will measure law students' trust level before and after the interactive experiment using both an initial and a follow-up interview. The experiment is a within-subjects design; this allows individuals to be examined without requiring a large sample size for an between-subjects comparison.
The experiment uses a pre-post design, which allows evaluation of treatment effectiveness by comparing results measured before and after an intervention. To ensure a proper evaluation of trust, both quantitative and qualitative data are collected. As trust is a complex construct, limiting the data collected to only quantitative measures would erase the different perspectives participants may hold on the topics discussed. By allowing participants to provide both data types, the data will not lose any context, and the quantitative data enables measurements to compare the gathered data.
Participants are assured that the RAG system is designed to help them and is free of malicious intent or human error. This is intended to prevent participants' beliefs about Benevolence and Integrity from influencing the experiment.
For this experiment, the participants will consist of nine third-year bachelor's Rechten students and a first-year master's student who graduated from the bachelor's in the previous academic year. Participants will be recruited using snowball sampling. Since the students are Dutch, the answers provided by the participants were either in their native tongue, English or a combination of them. The experiment was carried out in English unless additional explanations were requested in Dutch. After data collection, all responses in Dutch or a combination of Dutch and English were translated into English. They were manually transcribed and translated, with assistance from DeepL for translating certain Dutch legal terms and expressions [DeepL SE, 2026].

During their education, the students develop the foundation of their legal expertise. Yet with the rise of AI in the past year, more students have been utilising LLMs. This provides the participants with a position of familiarity with AI and the legal domain during their education. This gives them a different perspective from that of experienced lawyers. With this perspective, the experiment will examine how trust in RAG systems changes after gaining insight into the retrieval component and its outputs.

3.2 Measurement Framework

The interviews are split up into several thematic groups: accuracy, transparency and overall trust. For each of these groups, several Likert-Scale questions were constructed in the same format as in the study of Meywrith et al. [Meywirth et al., 2026]. This way, the statements had their own indicator, and together they provided insight into it, allowing for structured questioning that ensured all important factors were tested. The indicators of this study are the thematic groups. While accuracy and transparency align with a dimension that fosters trust, as discussed in the implemented framework, overall trust does not. Yet it was included to provide a clearer assessment of participants' trust levels and perceptions of the RAG system.

The initial interview begins with two Likert-scale questions to assess participants' propensity for trust. Then the interview continues with a series of open-ended questions to establish a baseline of participants' experiences with LLMs or chatbots. In this study, it is assumed that most law students have not heard of RAG systems, as they are presented as chatbots, and a user would require in-depth knowledge of RAG systems to classify them. This is not expected of them, since their education is not in a technical or technical-related field.

Moreover, both pre- and post-experiment interviews follow with the themes discussed. For each theme, participants are asked to rate several Likert-scale statements to provide quantitative data for analysis, followed by open-ended questions that elicit the reasoning behind their responses. The Likert-scale questions are rated on a scale of 1 to 5, with 1 indicating the most negative and 5 the most positive. Conducting interviews both before and after the experiment enables assessment of changes while understanding the reasoning behind them. This enables students to effectively convey their perceptions before and after gaining hands-on experience with a RAG system.

This interview structure is applied to accuracy, transparency and overall trust. Accuracy is used as an indicator of ability, transparency is applied to itself, and overall trust is used to see how accuracy and transparency influenced participants' trust. In all cases, Likert-scale responses provide quantitative measures, while the open-ended questions explore participants' reasoning. Furthermore, any changes depicted by the quantitative data might contradict the participants' thoughts. Therefore, the qualitative data might provide insight into their actual views regardless of the values in the quantitative data.

Whilst neither benevolence nor integrity is a specific theme in this experiment, some of the questions may indicate the participants' perception of them. However, RAG systems are built as a tool and are not intended to be something humans form an emotional bond with. Therefore, benevolence or integrity will not be tested in this study.

After the experiment, the qualitative data will be analysed and categorised through inductive coding. By reading through the interview transcripts, patterns in the answers emerge and can be grouped into a singular term, which becomes an applied code in this study.

3.3 System Configuration

The experiment will be conducted on a self-hosted server running RAGFlow [RAGFlow, 2026a]. RAGFlow is an open-source project on GitHub that allows people to deploy their own servers and utilise Retrieval-Augmented Generation with agents [RAGFlow Team, 2026].
The researcher uploaded and processed case law in the system prior to the experiments. The general template divides documents into consecutive chunks based on a predefined token limit, as described in the RAGFlow documentation [RAGFlow Team, 2026]. An alternative approach is to parse the document using the law template. However, there is limited documentation for this template, and several attempts to parse the document using the law template produced impractical results or could not be configured successfully. In addition, two file formats were considered: PDF and TXT. However, because the TXT file removes all structure from the text, it produces incoherent chunks. Consequently, the parsing configuration was reverted to the general template and PDF as the file format.
During document processing, several chunk sizes are considered. Initially, smaller chunk sizes ranging from 100 to 256 tokens are evaluated. This results in fragments that are often too short, cut in the middle of a paragraph or lack sufficient context to preserve the meaning of the legal text. During testing, retrieval with these chunk sizes often yielded incorrect answers. However, a chunk size of 512 produces more coherent segments while maintaining adequate contextual information. Consequently, this limit was set, along with the delimiter `\n\n` set to ensure that chunks are separated at the paragraph boundaries rather than splitting individual paragraphs. In an attempt to preserve as much context as possible. Finally, the system is configured to use the following embedding model, *nnomic-embed-text*, and the LLM, *llama3.2:latest*. These models are chosen since they are locally deployable via Ollama, enabling reproducibility, and they are freely available.

The retrieval mechanism is controlled by two primary system parameters: the similarity threshold and the vector similarity weight. These are used to calculate three retrieval metrics: term similarity, vector similarity and hybrid similarity. The system displays these scores above each retrieved chunk, as each chunk receives its own set of calculated values [RAGFlow, 2026b]. Based on these scores, the system decides whether or not to retrieve the chunk as an answer. These two primary parameters are the variables that the participants in the experiment will manipulate to gain hands-on insight. Term similarity and vector similarity are directly related through the vector similarity weight. If the vector similarity weight is represented by x, then the corresponding term similarity weight is equal to $1\text{-}x$ [RAGFlow, 2026b]. The term similarity measures the degree of keyword overlap between the user's query and a text chunk. More specifically, it evaluates the extent to which the keywords extracted from the query correspond with the associated retrieved chunks [RAGFlow, 2026b].
Conversely, vector similarity measures the semantic relevance of a chunk to the query. Rather than relying on exact keyword matches, it evaluates the similarity in meaning between the query and the chunk based on their vector representations [RAGFlow, 2026b]. The hybrid similarity score is then calculated as a weighted sum of term similarity and vector similarity. The vector similarity parameter determines the weights, with a default sum of: $0.7\times$term similarity$+0.3\times$vector similarity [RAGFlow, 2026b].
No additional functionality was applied during the experiment, such as a reranking model, cross-language search, or a knowledge graph. Consequently, RAGFlow operated using its default retrieval strategy, which combines weighted keyword similarity with weighted vector cosine similarity

[RAGFlow, 2026b]. The cross-language search was disabled as the experiment was conducted exclusively in English. The knowledge graph was disabled to avoid introducing additional retrieval mechanisms that may influence the results and complicate the evaluation of the selected parameters [RAGFlow, 2026b].

3.4 Experiment Procedure

The experiment will proceed as follows: The examinees will be presented with a physical copy of the case France v. Menneson (ECtHR) [of Human Rights,]. This case is chosen because it is relatively concise and has been studied in depth in consultation with a legal professor, enabling the researcher to provide consistent procedural guidance and accurately interpret participant interactions during the experiment.
The researcher prepared this case law in advance using the specific configurations described in Section 3.3. Subsequently, they will receive an explanation of how the system processed this document and the retrieval testing page. After familiarising themselves, they are shown how the information retrieval works with two example questions listed in the appendix (Section 8).
They will be informed that term similarity correlates with syntax and, therefore, the exact wording of the text. Vector similarity refers to the meaning, semantics, of the chunk. Lastly, the hybrid score comprises these two metrics, prioritising term similarity. The hybrid score is calculated with the following formula: $0.7 \times \text{term similarity} + 0.3 \times \text{vector similarity}$.
With these parameters, the system user can manipulate retrieval. The official explanation of the parameters given by RAGFlow is as follows:

The similarity threshold is the lower bound for retrieving text chunks. Therefore, it seeks a closer match between the chunks. If a chunk of text has a lower score than the chosen threshold, the system will filter it out and not return it. The parameter ranges from 0.0 to 1.0. With a default setting of 0.2[RAGFlow, 2026b].
The vector similarity weight influences how much the semantic meaning of a text is considered during testing. By default, it will be 0.3. This parameter is also tied to the term similarity weight [RAGFlow, 2026b].

The two parameters, as specified, manipulate the retrieval of this RAG system. These parameters are tied to the hybrid, term, and vector similarity scores, and, based on these scores, the system decides which chunks to retrieve.
As stated before, the similarity threshold is the lower bound for retrieving text chunks. This means it is responsible for the number of chunks the system will present to the user after retrieving information. The vector similarity weight controls the influence of semantics in retrieval, rather than the semantics themselves, and therefore affects the order of the retrieved chunks. The task of the examinees will be to try to get the correct chunk that holds the answer to the top. They will be presented with 2 questions, where scenarios 1 and 2 occur.
With this experiment, they will be met with three possible scenarios, as scenario 3 is shown to them with the first example question

- Scenario 1: the right chunk will appear at the top immediately with the default values on both metrics

- Scenario 2: the right chunk will appear at the top after some playing around with the values of both metrics
- Scenario 3: the right chunk will never appear at the top, no matter what is done with the values of both metrics.

Based on these results, the examinee can obtain a better understanding of the system, its accuracy, and that optimising the system is not possible by making both metrics super strict, and that having strict metrics does not guarantee accuracy. For instance, they will be able to understand whether a chunk is relevant or irrelevant according to the system [RAGFlow, 2026b].
The three scenarios are designed to expose participants to different levels of retrieval effectiveness that may occur during implementation. Scenario 1 and Scenario 2 represent successful retrieval. In Scenario 1, the relevant chunk is ranked first without requiring any system manipulation. In Scenario 2, participants must manipulate the parameters, similarity threshold and vector similarity weight to improve retrieval results and obtain the correct answer. Scenario 3 represents unsuccessful retrieval: the system cannot rank the most relevant chunk first, despite adjusting the retrieval parameters. This demonstrates the limitations of the retrieval component, allowing the participants to gain a clear overview of all possible situations.
All scenarios together illustrate to the participant that there is no single correct answer about which values must exist for the metrics to ensure the system gives the right answer at the top. Therefore, the system cannot guarantee accuracy.
By doing this experiment and allowing them to understand how the system works and why the system cannot ensure accuracy, they can see how this affects them as the system's transparency becomes clearer. Its accuracy is called into question throughout the examinee's interactive part.

4 Experiments Results

4.0.1 During the Experiment

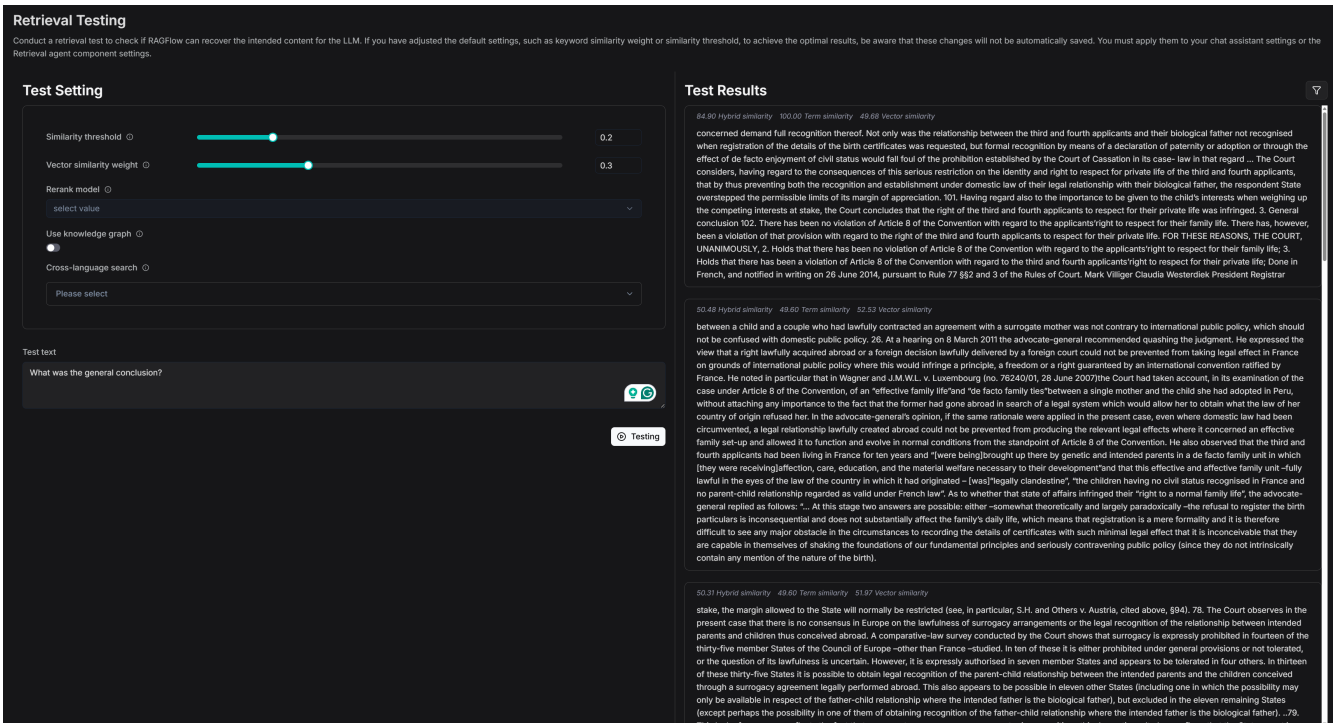

Figure 4: The retrieval testing page that the participants were seeing.

While the participants were performing the experimental tasks, their behaviour was observed. Most participants initially engaged in exploratory behaviour, testing different values at random to better understand the system. For some participants, this behaviour persisted throughout the experiment. However, certain participants appeared to make use of the information provided during the explanation. Rather than aimlessly searching for the correct parameters, these participants reflected on the system's underlying linguistic mechanisms, using the language in the question to reason about the values of the similarity threshold and vector similarity weight. By applying this knowledge, they were able to obtain the correct answer faster than other participants.

4.1 Descriptive Statistics

Both the before-and-after interviews collected quantitative and qualitative data. Quantitative data consisted of participants' responses to Likert-scale questions, while qualitative data were obtained through open-ended interview questions. In addition, during the interactive part of the experiment, the participants described their behaviour and interactions with the retrieval mechanisms.

4.1.1 Trustor's Propensity: Pre-experiment Interview

At the start of the interview, participants were asked to evaluate two questions about their beliefs concerning technology in general. The questions are in the appendix: question 1 statements a and b

(Section 8). Most participants indicated hesitation to trust new technology tools without personal experience, with responses ranging from 2 to 4 on the Likert scale and a mean of 2.7 (SD = 0.97). With 1 indicating "I strongly disagree" and 5 indicating "I strongly agree". The second question showed responses, scaling from 1 to 4, to whether they believed that the technology would work as intended, resulting in a mean of 2.3 (SD = 1.16). As Table 1 depicts.

Table 1: Descriptive statistics for general technology and LLMs

Label	Mean	SD
1-TG1	2.7	0.95
1-TG2	2.3	1.16
1-T1	2.7	0.95

The majority of participants were not familiar with chatbots in the legal sector or had only heard of them, except for one participant who had limited experience with them. Nevertheless, the participants were familiar with general-purpose LLMs. 8 participants reported using various LLMs for academic tasks. The most common LLMs were ChatGPT and Gemini, with occasional additions of Claude and some others. During the deployment of these LLMs, the participants generally described having limited to moderate familiarity with these tools. Some used AI regularly for their academic work, while others had little to no experience.
Participants reported using LLMs for summarisation and supporting tasks (e.g. writing feedback, drafting assistance and explaining legal concepts). They emphasised the need to verify the outputs because of concerns about inaccuracies and hallucinations. Additionally, the participants reported mixed experiences regarding the accuracy of the LLMs; many reported incorrect answers, hallucinations, and inconsistent performance.
In addition to the more general questions, the participants were asked about their understanding of how LLMs function. Again, the responses ranged from 2 to 4, on a scale of 1 indicating "no understanding" to 5 indicating "complete understanding". This resulted in a mean of 2.7 (SD = 0.95) on the five-point Likert scale (Table 1). Furthermore, to evaluate the baseline of participants' perceptions, the questions were divided into three thematic groups for both the before- and after-experiment interviews: Transparency (T), Accuracy (A), and Overall Trust (OT). The questions are found in the appendix (Section 8)

4.1.2 Intervention: Post-experiment Interview

In the follow-up interview, the theme Intervention (I) was introduced. For each theme, Likert-scale questions were asked, followed by corresponding open-ended questions. In addition to these questions, participants were asked several more open-ended questions on the same themes.

Question	Mean	SD
2-I1	4.4	0.97
2-I2	3.0	0.94

Table 2: Interventions: Mean and Standard Deviation

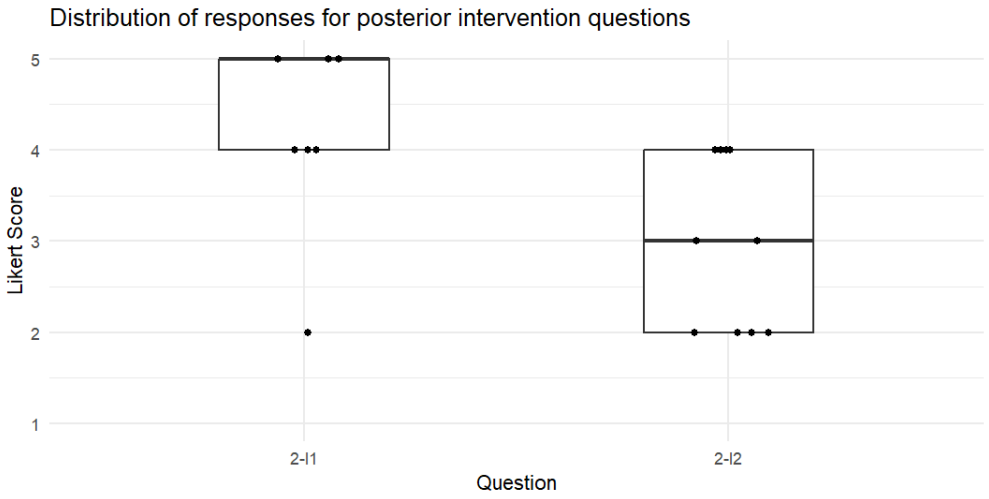

Figure 5: Distribution of responses for posterior intervention questions

Table 2 and Figure 5 present the intervention question results. Both questions were on a Likert scale of 1 to 5. For the first question, the scale ranged from "decreased significantly" to "increased significantly". The first question in this theme assesses whether participants' understanding of the retrieval process improved compared to before the experiment. The results showed a high mean score of 4.4 and an SD of 0.97. This indicates that interacting with the system significantly contributed to their understanding, as the scale's upper bound is 5. Participants' understanding of the retrieval component increased significantly. Except for one participant, who indicated that the adjustable parameters, the vector similarity weight and similarity threshold, increased confusion rather than improving their understanding.
The second intervention question was regarding the participants' confidence in the system's accuracy. This question received a lower mean score of 3.0 from the participants, with an SD of 0.94. The participants stated that their confidence in accuracy lessened, while others stated that their confidence increased. 6 participants understood why it produced the output, while others became more confused. One participant stated: "Like I said, confusing metrics. I already had low trust in these systems, and it only got worse". As Figure 5 shows, participants were divided into two groups, with two outliers in the middle. One group ranked lower at a score of 2, and another group ranked higher at a score of 4. This indicates that the intervention did influence their view and understanding of the RAG's retrieval component.

4.1.3 Transparency

Question	Mean	SD
1-T2	2.3	1.16
1-T3	2.2	1.03
1-T4	2.3	1.49

Table 3: Pre-experiment Transparency: the mean and SD

Question	Mean	SD
2-T2	3.7	0.82
2-T3	3.3	0.67
2-T4	3.5	1.27

Table 4: Post-experiment Transparency: the mean and SD

Table 5: Transparency questions: the mean and SD

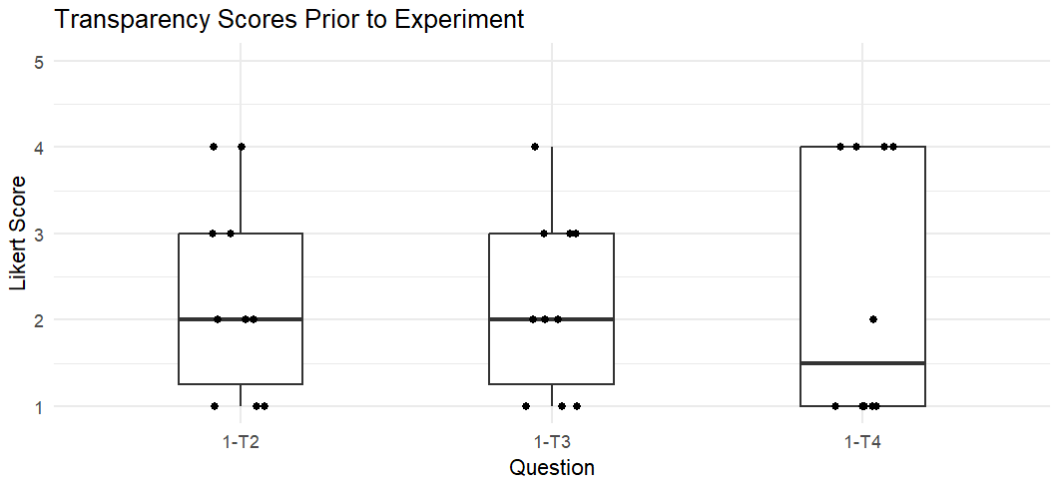

Figure 6: Participants' perceptions of transparency before the experiment

During the pre-experiment interview, participants received a brief explanation of the concept before being questioned about RAG systems. The explanation is included in the interview script in the Appendix (Section 8). All participants expressed having no familiarity with RAG systems. The responses to the Likert questions varied among participants, and Figure 6 shows these responses. The first and third transparency questions had a mean of 2.3, and for the second question, it was 2.2, as illustrated in Table 3. The first question illustrates the participants' estimation of their understanding of a RAG system. Some participants described it as similar to an LLM: "I think it won't be that different from, like, the LLM, so I think I understand it pretty well", with other participants offering similar explanations. Whilst others highlighted that they had never heard of it. Furthermore, one participant stated they understood the RAG system solely based on the explanation incorporated with their knowledge of LLMs. This question has a larger SD of 1.16, as Figure 6 reflects.

The second question (1-T3) describes how participants understood the retrieval mechanism of RAG systems. The mean for this question is 2.2, indicating that several participants did not understand how it worked. This is supported by Figure 6, where 6 participants rated this question 1 or 2, and the higher ratings are mainly 3, with one outlier on 4. During the explanations, participants repeatedly expressed the same answer in various ways: "I have no idea how it works". Whilst some participants attempted to imagine how it could work, they were unable to form a clear idea. Some of these attempts were based on the input, a combination of input, human behaviour, algorithms and exact wording.

The third question (1-T4) assesses how well participants understood RAG systems in order to evaluate whether the outputs were trustworthy. The SD on this question is 1.03, and the mean is the same as question one; Figure 6 illustrates how participants indicated that they either knew enough about the inner workings of an RAG system or thought they knew barely anything to evaluate how trustworthy the outputs were.

The median was similar for the first two questions and significantly lower for the last, suggesting

that participants believed they could not evaluate whether the outputs would be trustworthy. Furthermore, looking at participant-level averages revealed notable variation. Participants 2 and 7 exhibited the highest expectations for transparency with a mean of 3.67, while Participants 1 and 9 exhibited the lowest mean of 1.0, shown in Figure 8. Furthermore, 7 participants reported uncertainty regarding how RAG systems functioned before the experiment.

Table 6: Question 1-TB1: How well do you believe that you understand a RAG system, including the explanation provided?

Rating	Freq
2 - Limited understanding	1
3 - Moderate understanding	1
4 - Good understanding	6
5 - Complete understanding	2

During the pre-experiment interview and after following a more in-depth explanation of the RAG process, participants indicated that the majority were unfamiliar with Retrieval-Augmented Generation before the study. Only a small number had previously encountered the concept, and only one participant had experience using a RAG-based system. Notably, the participant was unaware that the AI used RAG technology. Their experience with the system was limited to observing that it tended to produce longer responses, yet its accuracy remained as inconsistent as that of other AI systems.
Lastly, the participants were asked to evaluate how well they believed they understood the RAG systems, as shown in Table 6. Eight Participants believed they understood the RAG systems well. The other two lower scores belong to participants 1 and 9.

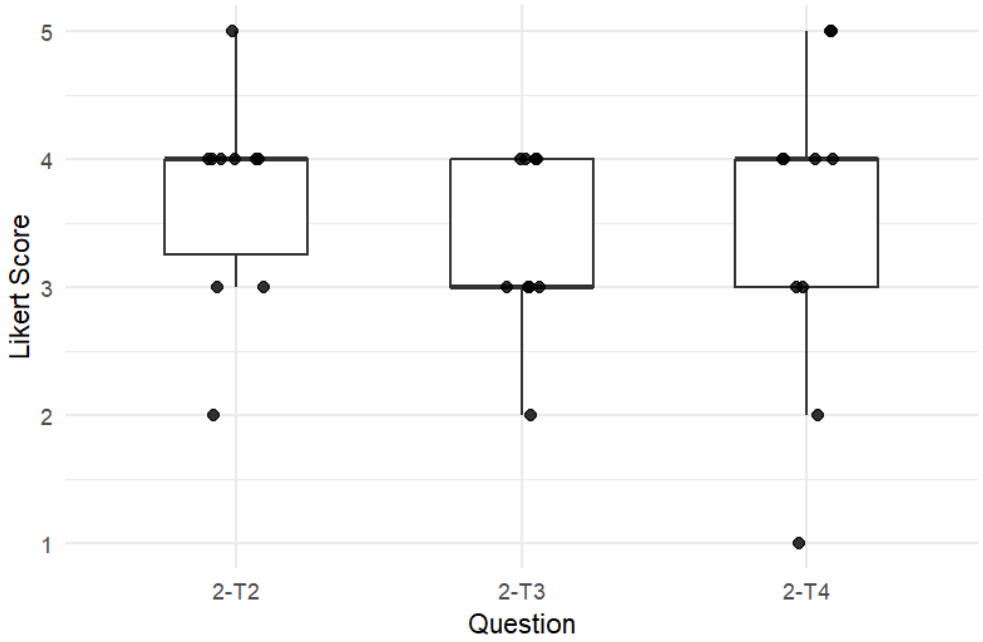

Figure 7: Distribution of responses for posterior transparency questions

The post-experiment interview repeated Questions 1-T2 and 1-T4. Question 1-T3 was replaced with a more specific question assessing whether participants could identify the cause of an unexpected RAG response.
Even though questions 1-T3 and 2-T3 do not fully overlap, 2-T3 further specifies 1-T3 by asking whether participants can identify why an unexpected output would be returned. This implies whether participants can identify why an incorrect answer is returned.
The Likert-scale transparency questions asked after the experiment are presented in Table 7. All transparency scores increased by at least one point compared to the pre-experiment interviews illustrated in Table 5. Furthermore, the SD also decreased: for the first two questions, it decreased by at least 0.3 (0.34 and 0.36), and for the last question, it decreased by 0.22. This indicates that the answers are more clustered.
For all three questions, responses clustered around scores of 3 and 4 on the Likert scale. However, the third question (2-T4) still shows greater variation among scores, as reflected in its SD of 1.27, compared to the first question's SD of 0.87 and the second question's SD of 0.67. When participants were asked whether there were any remaining uncertainties regarding the retrieval process. Six participants reported no uncertainties, while four stated that uncertainties remained. Remaining uncertainties included the similarity threshold and vector similarity weight in the coding.
During the open-ended questions, participants described the retrieval process in their own words. One of the more in-depth explanations was: "You ask a question and then the system will retrieve the information from the database with these parameters (Similarity threshold and Value similarity weight), and then it will give you a couple of text pieces back where it thinks your answer is written in.". Another example is: "I guess it looks for similarities and word-vector meaning in general to the database, and then it feeds me back depending on how high or low I let the vectors and how many chunks it feeds me back. And then the lesser the chunks, the better; the greater the chances I find the correct answer if I even put the correct prompts."
All participants provided descriptions that reflected the basic principles of the retrieval process. Furthermore, nine out of ten participants reported a change in their perception following the experiment.
One of the participants with an increase reported the following: "Yes, before this I had no idea how this works. I personally never had an explanation or looked for it myself, but now because of your explanation and because of the interactive experiment, I understand what it is looking for and thus how you can try things out with your own questions. So it is a lot more clear." And the participant without a change: "Not really, I already understood how it worked to a certain extent, and this was just more in depth"
When asked which aspect of the retrieval process they found most surprising, participants most often cited the use of semantic meaning in addition to the query's syntactic wording. Although participants offered suggestions about how they thought the RAG system retrieved information prior to the experiment, none of them included semantics. While some did report the exact wording, they had not thought of the meaning of a word and the relation to another word.

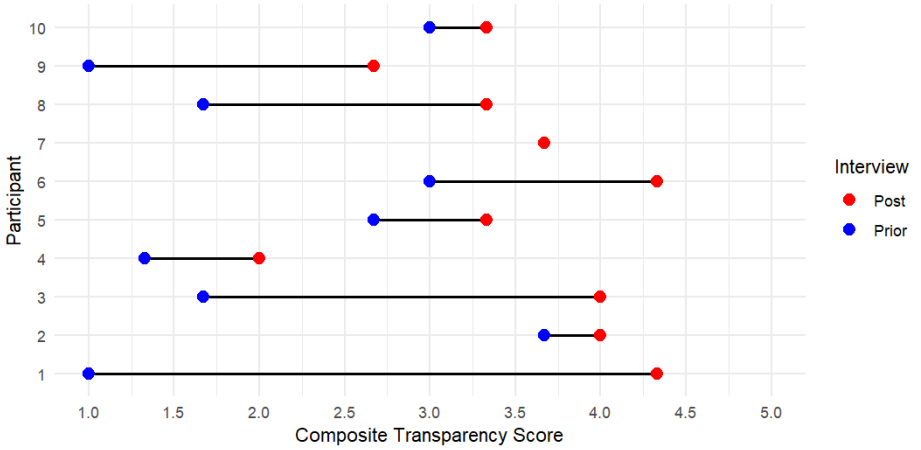

Figure 8: Composite transparency mean score before and after the experiment

Except for Participant 7, the participants show an increase in their transparency, in Figure 8. This figure shows the mean for participants across all three questions and how it changes before and after the experiment. Participant 1 had the greatest change from before to after the experiment.

4.1.4 Accuracy

Question	Mean	SD
1-A1	2.9	0.99
1-A2	2.9	1.10
1-A3	3.0	0.82

Table 7: Pre-experiment Accuracy: the mean and SD

Question	Mean	SD
2-A1	3.2	0.79
2-A2	2.4	1.07
2-A3	3.1	0.74

Table 8: Post-experiment accuracy: the Mean and SD

Table 9: Accuracy questions: the mean and SD

In the pre-experiment interview, three questions assessed participants' perceptions of the system regarding accuracy. The questions were answered on a scale from 1 to 5. Where 1 indicates "I strongly disagree" and 5 indicates "I strongly agree".

The first two questions assessed their belief in the system's ability to retrieve the most relevant passage (1-A1) and to do so consistently (1-A2); variation in the answers is shown in Figure 9. The mean for these questions was 2.9. The second question had the highest SD, indicating that participants have different expectations about the system's ability to consistently retrieve the most relevant passage.

For the third question (1-A3), responses clustered around a score of 3 on the Likert scale with an SD of 0.82. For this question, participants were asked whether they were confident that the retrieval process would prioritise the information most useful for them. This resulted in a mean of 3.0.

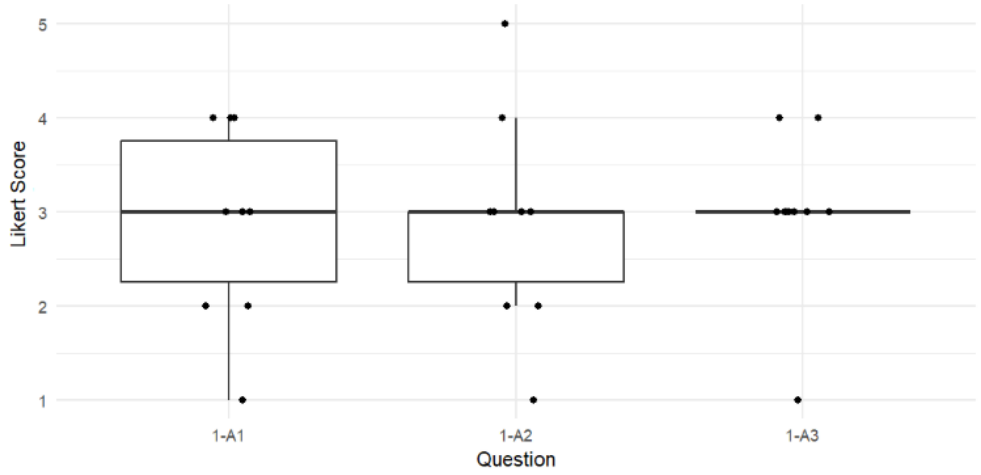

Figure 9: Participants' perceptions of accuracy before the experiment

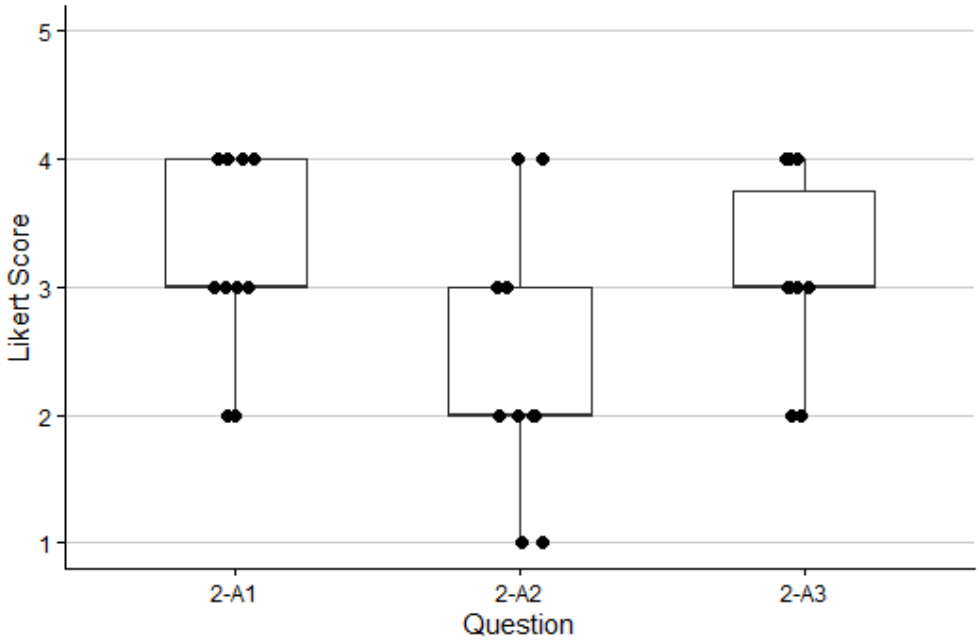

Figure 10: Distribution of responses for posterior accuracy questions

The post-experiment interview showed a similar pattern for Questions 2-A1 and 2-A3, whereas perceptions of retrieval consistency (2-A2) declined. All accuracy questions in the post-experiment interview remained the same as 1-A1, 1-A2 and 1-A3.

Question 2-A1 and 2-A3 received similar mean scores of 3.2 and 3.1, respectively, with SDs of 0.79 and 0.74. These questions asked participants whether the system would retrieve correct and relevant information for the user and prioritise the information most useful to them. Moreover, the second question focused on the consistency of retrieval. This resulted in a lower mean of 2.4 and had more diverse ratings, as shown in Table 8 and Figure 10.

Although participants generally considered the system somewhat reliable, many also recognised its limitations. Eight participants reported inconsistent retrieval accuracy during the experiment, contributing to uncertainty about the system's reliability. Consequently, nine participants stated that the experiment made them more cautious when using RAG systems. One participant described that the RAG system would be better than an LLM; however, it would still be AI, and therefore they should be careful. Another participant described how the system can retrieve the correct information; however, it is not possible to know whether it is correct without verification. The other participant stated that it does not influence them as they would not like to use the RAG systems or AI in general.

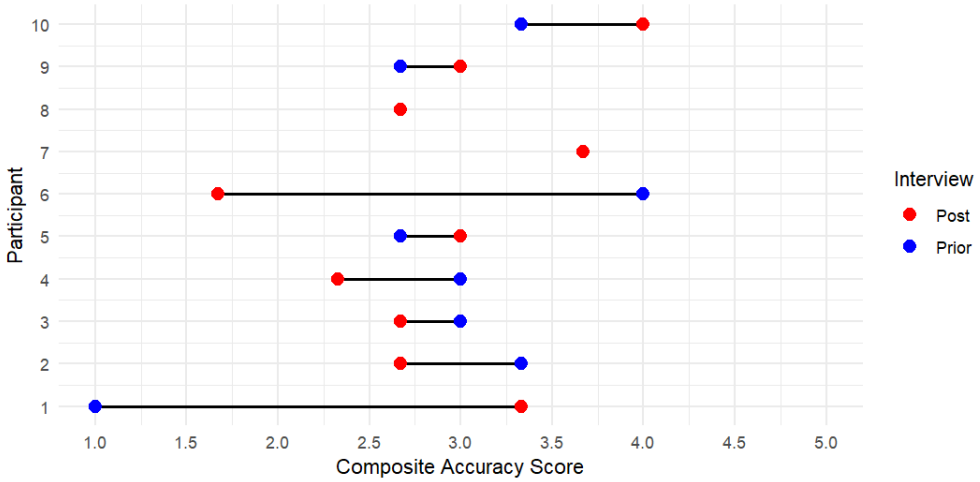

Figure 11: Composite accuracy mean score before and after the experiment

In Figure 11, the changes in the accuracy are reflected. The figure shows the same change as in Figure 8: the mean for participants across all three questions, along with its change before and after the experiment.
Figure 11 presents the composite mean accuracy scores before and after the experiment. The scores ranged from 1.00 to 4.00, with most participants reporting scores between 2.67 and 3.67. Additionally, after the experiment, four participants showed an increase in their composite mean, two showed no change, and four showed a decrease.

4.1.5 Overall Trust

Question	Mean	SD
1-OT1	3.6	1.17
1-OT2	1.4	0.70
1-OT3	1.6	0.70

Table 10: Pre-experiment Overall Trust: the mean and SD

Question	Mean	SD
2-OT1	3.6	1.43
2-OT2	1.6	0.97
2-OT3	1.9	1.20

Table 11: Post-experiment Overall Trust: the Mean and SD

Table 12: Overall Trust questions: the mean and SD

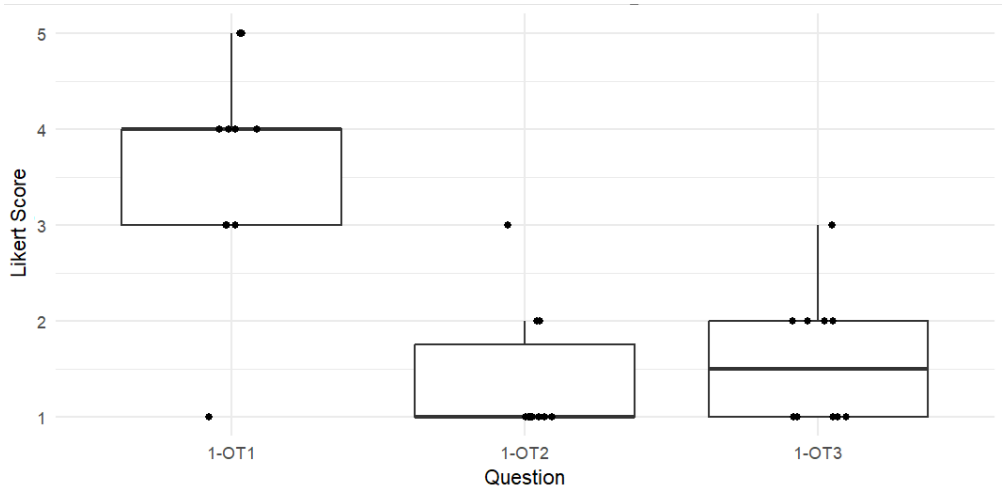

Figure 12: Participants' perceptions of their overall trust before the experiment

The final group of questions in the pre-experiment interview examined the willingness to use a RAG system during their research process. For these questions, the Likert scale ranged from 1 to 5, from I strongly disagree to I strongly agree.

The first question reflects participants' willingness to use a RAG system in their legal research. Responses ranged on the higher end of the Likert scale, from 3 to 5, for this question, with a single outlier in this range at 1 (Figure 12). The mean was 3.6, and the SD was 1.17.

The second question asked whether participants would trust the output of an RAG system enough to implement it without manually verifying every result. For this question, there were various qualitative negative answers which are also reflected in the quantitative data. The participant who chose 3 said that while they would like to double-check, if it works properly, you can actually trust it. However, other participants discussed ethical and professional concerns and the recurring argument that it is AI. The mean was 1.4, with an SD of 0.7.

The last question in this thematic group reflected participants' comfort level while relying on a RAG system when there would be real legal outcomes. This received nine participants responding with 1 or 2 and a mean score of 1.6 with an SD of 0.7.

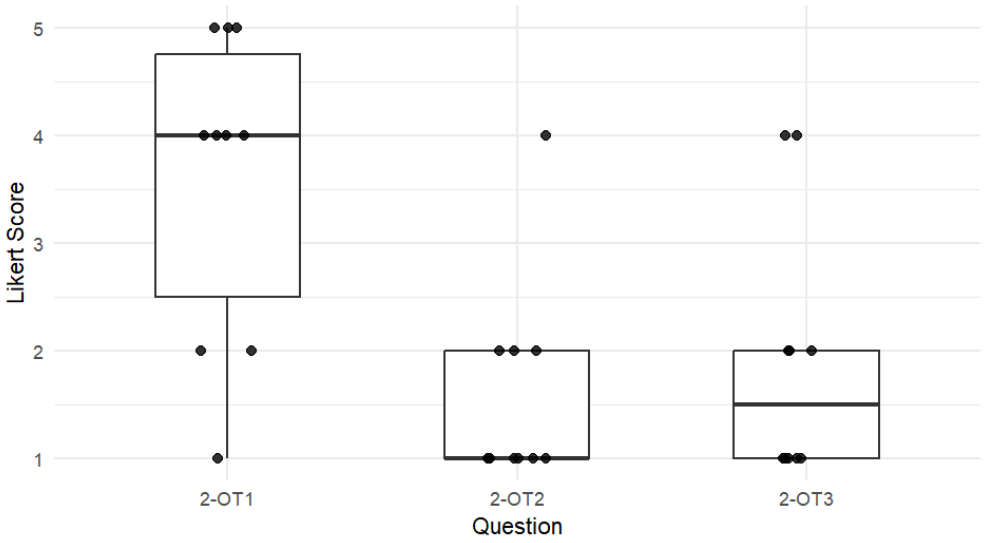

Figure 13: Distribution of responses for posterior overall trust questions

The overall trust after-experiment interview results are shown in Table 11 and Figure 13. After the experiment, participants reported a relatively high willingness to use a RAG system in their legal process, as reflected in the responses to the first question with a mean of 3.6 and an SD of 1.43. However, this result did not differ from the pre-experiment mean score. Nine participants frequently stated that they would verify the retrieved information before using it. Professional responsibility also influenced participants' willingness to trust the system. One participant gave accountability as a reason: "... I wouldn't trust it. Just because I'd rather that the blame falls on me instead of the system".

In the second question, which had a mean of 1.6 and an SD of 0.97, participants indicated that greater trust would only be possible if they verified the outputs while familiarising themselves with the system's behaviour.

In the open-ended responses, all participants stated that they would not use the outputs without verification, commonly repeating concerns about inconsistent accuracy. Nevertheless, if used as an assistive tool, participants reported that it could increase efficiency and, therefore, they would still use it. For example, a participant stated: "Yes, I think I would. You could implement it then as a time-saving tool. By asking questions like, is this a legal rule from which judgement is it, and then you'd get an answer from the RAG system, you could verify that really easily and continue working".

Furthermore, the last question in this thematic group received a mean of 1.9 with an SD of 1.20, which addresses the use of the outputs for real legal purposes. Participants indicated that AI increased the likelihood of errors and therefore they did not want to rely on an AI system; however, if they are allowed to manually verify it, two participants would consider it.

Additionally, five participants estimated that their understanding of the system changed their perception of it, four reported no change, and one was uncertain. Several participants provided reasons for this change, with the most frequently mentioned factors being the adjustable parameters: vector similarity weight and similarity threshold.

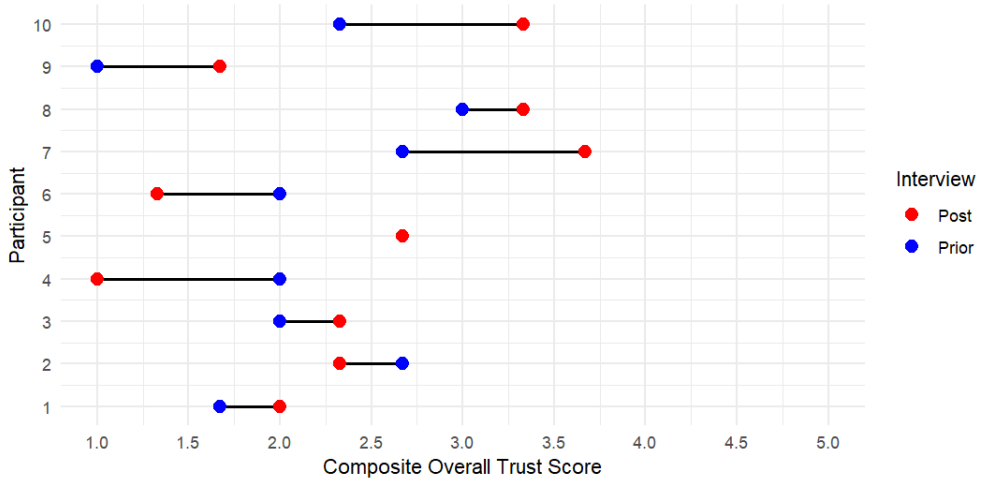

Figure 14: Composite overall trust mean score before and after the experiment

The composite mean scores ranged from 1 to 3, as shown in Figure 14. Participant 8 has the highest score, whereas participant 9 has the lowest score. Most participants ranged between 2 and 2.67. Finally, the most common reason reported was a lack of understanding of AI tools, either in general or in their inconsistent accuracy.

4.1.6 Participant Changes

Participant	Transparency	Accuracy	Overall Trust	Reported change
Participant 1	+3.33	+2.33	+0.33	Increase
Participant 2	+0.33	-0.67	-0.33	Slight change
Participant 3	+2.33	-0.33	+0.33	Increase
Participant 4	+0.67	-0.67	-1.00	Same
Participant 5	+0.67	+0.33	0.00	Slight change
Participant 6	+1.33	-2.33	-0.67	Decreased
Participant 7	0.00	0.00	+1.00	Increase
Participant 8	+1.67	0.00	+0.33	Same
Participant 9	+1.67	+0.33	+0.67	Same
Participant 10	+0.33	+0.67	+1.00	Increase

Table 13: The changes in the participants' scores before and after the experiment based on a scale of 1 to 5. **Added note:** the participants did not indicate positive or negative with the label "Slight change"

Table 13 shows the differences between the composite mean score before and after the experiment. These differences were also shown per category in each subsection in Figure 8, Figure 11 and Figure 14.

The last column, Reported change, represents the qualitative responses participants gave when asked about the experiment's effect on their trust. In the table, six participants described how they felt their trust evolved, and these specific changes are reflected in the data. In the other cases, the qualitative and quantitative data differed.
A notable case is Participant 1, in which all quantitative data increased, with larger differences than in the others. Furthermore, Participant 5 showed no change in Overall Trust. Although both transparency and accuracy increased in the quantitative data, the participant indicated a change in their trust.
Furthermore, if the Transparency and Accuracy score increased or remained unchanged, Overall Trust did not decrease, according to the quantitative results. However, in three cases the Overall Trust decreased if Accuracy also decreased.
Only Participant 3 was an outlier, in which Accuracy decreased, and Overall Trust increased. However, Transparency increased significantly for this participant.

4.2 Wilcoxon and Paired T-test

For the statistical testing, two separate tests have been conducted: a standard paired t-test and a Wilcoxon signed-rank exact test. The paired t-tests were initially conducted; however, due to the small sample size (n = 10) and the ordinal nature of the data, the Wilcoxon signed-rank test was considered the primary inferential test and is therefore highlighted while reporting the results. The effect sizes were calculated using the rank-biserial correlation.
To conduct these tests, the composite mean scores have been calculated and are reflected in Table 14. The results of the paired t-tests and Wilcoxon tests are documented in Table 15.

Table 14: Composite scores before and after the experiment

Transparency

Time	Mean	SD
Prior	2.47	1.03
Post	3.50	0.74

Accuracy

Time	Mean	SD
Prior	2.93	0.81
Post	2.90	0.67

Trust

Time	Mean	SD
Prior	2.20	0.59
Post	2.37	0.90

The transparency scores increased for nine out of ten participants after the experiment with the RAG systems as shown in Table 13 and Figures 8, 11 and 14). The paired t-test indicated a significant effect with a p-value of 0.014. Additionally, the Wilcoxon signed-rank test indicates a significant effect (V = 2.0, p = 0.008). With a rank-biserial correlation also indicating a large effect (r=1.00, 95% CI [1.00, 1.00]).
Neither the accuracy nor the overall trust scores showed a statistically significant difference between before and after the experiment. The accuracy scores did not differ significantly between pre- and post-test measures (V = 27.5, p = 0.906, r = -0.08), and the paired t-test of accuracy resulted in a p = 0.931.
The results for the overall trust indicate no significant effect either. The t-test p-value was 0.453, and the Wilcoxon test p-value was 0.356 (V = 17.5). The rank-biserial correlation (r = 0.36, 95% CI [-0.32, 0.79]), indicating that the posterior scores tended to be higher than the prior scores, yet there was no significant effect recorded, which is also reflected in the table and the figures.

Table 15: Results of paired-sample t-tests and Wilcoxon signed-rank tests

Construct	Mean Diff.	t	df	p (t-test)	V	p (Wilcoxon)	r	95% CI
Transparency	-1.03	-3.05	9	0.014	2.0	0.008	1	[1.00, 1.00]
Accuracy	0.03	0.09	9	0.931	27.5	0.906	-0.08	[-0.65, 0.55]
Overall Trust	-0.17	-0.79	9	0.453	17.5	0.356	0.36	[-0.32, 0.79]

5 Analysis

5.1 Transparency

Familiarity with LLMs did not necessarily correspond to an understanding of their underlying operations, nor did it explain their behaviour. Participants who regularly used GenAI or other AI tools showed no greater technical comprehension than those who rarely used them. This dissociation between usage and understanding suggests that familiarity alone is insufficient to develop the kind of AI literacy that enables calibrated trust evaluation.

This finding is consistent with prior work on algorithmic transparency, where users often required explicit interventions to understand how the algorithm operated [Eslami et al., 2015].

Increased transparency increased understanding in the system, allowing participants to distinguish between relevance and consistency. The experiment led participants to develop a more sophisticated understanding of the RAG retrieval process, rather than seeing it as a black-box model.

However, participants did not always fully understand the technical side of the RAG systems. Before the experiment, participants perceived the system as a black box. After gaining insight, they found the use of semantic meaning in addition to syntactic wording difficult to understand. Especially when combined with parameters that manipulate the retrieval process. This reflects a limitation in XAI and can result in miscalibrated trust, where something is assumed to work a certain way. However, it does not do so due to a technical misunderstanding.

5.2 Accuracy

Participants' confidence in the systems' accuracy did not increase trust in their reliability. On the contrary, exposure to the retrieval components emphasised the system's inconsistencies and potentially unexpected outcomes. This reduced participants' confidence in their consistency, leading them to gain a more realistic understanding of the systems' workings and limitations, thereby calibrating their trust.

Throughout the interviews, concerns about accuracy, hallucinations and consistency emerged as recurring themes. Additionally, ethical and professional concerns played a part, a recurring issue in AI in general for the legal domain, as shown by Law360 Pulse [Martinson, 2025].

Furthermore, participants distinguished between usefulness and correctness. They acknowledged that RAG systems could improve efficiency. However, participants frequently emphasised the importance of keeping a human in the loop and independently verifying outputs if they intended to use them. This indicates that the participants acknowledged the inconsistencies in accuracy, yet they were not severe enough to outweigh the perceived efficiency benefits of the systems.

However, this perception did not necessarily make them distrust the systems. Their propensity

for trust and initial mental model influence how they take observed accuracy into account. This supports the calibrated trust more; participants obtained a clearer view on when it is appropriate to rely on the outputs, and when it can be used as an assistive tool.
The experiment solidified the opinion that the participants shared in the initial interview. Participants observed that the LLMs were inconsistent in their accuracy. Through the experiment, participants became more informed about the retrieval process and how the system operates. While they still observed retrieval failures and inconsistencies, the participants concluded that RAG systems could still be beneficial to them, provided that appropriate verification measures were in place.

5.3 Overall Trust

The RAG system is not reliable as an authoritative decision-making resource; however, it is useful as an assistive research tool. The willingness to rely on these systems was very low. Additionally, participants did not want it to affect real-life cases. This implied that participants calibrated their trust in the RAG system as an assistive tool. Furthermore, they preferred human oversight while using the RAG system. This reflects recommendations within the explainable AI literature that transparency should support accountability and informed human oversight [Felzmann et al., 2019]
The statistical analysis showed a significant increase in transparency but no significant change in either accuracy or overall trust (Table 15). When considered together with the qualitative responses, these findings suggest that participants developed and solidified a more calibrated trust in the systems. In previous work, it was suggested that explainability should enable users to determine when reliance on AI is appropriate, as trustworthiness is not a stable quality [Papagni et al., 2023].
The increase in transparency and accuracy resulting in an increase of overall trust is the most common pattern. However, it is not always the case. This finding aligns with Shin, who argued that explainability should not be viewed solely as a mechanism for increasing trust [Shin, 2021].
Participant 6 provides a clear example of this relationship. Although their understanding of the RAG systems increased, their perceptions of accuracy and how they would implement it decreased compared to the beginning of the experiment. Consequently, both their statistical overall trust and their reported perception of trust declined. Demonstrating that increased understanding may not lead to higher overall trust.
Furthermore, the discrepancies between the statistical data and reported perceptions may suggest that quantitative measures alone do not fully capture all aspects of trust. The results suggest that increases in overall trust are often accompanied by increases in transparency and positive perceptions of accuracy; however, this pattern was not consistent across all participants.
Conversely, participant 7 represented an interesting exception. Their statistical scores showed no change in either transparency or accuracy, yet their overall trust did increase by an entire point. This increase correlates with their reported answers, which indicate increased trust. However, the participant did not provide an additional explanation, making it difficult to identify the motivation for the increase in trust. This illustrates how factors such as the trustor's propensity or the perceived usefulness might influence trust.
Overall, the findings in this study suggest that the experiment primarily improved participants' ability to calibrate their trust rather than increasing trust itself. Participants became more knowledgeable about how RAG systems operate and recognised their limitations. Consequently, they viewed the systems as valuable assistive research tools, yet still required human verification before

using their outputs, especially in the legal domain. These findings are consistent with prior research on explainable AI [Papagni et al., 2023, Felzmann et al., 2019].
This study extends existing research by showing that transparency in RAG systems does not increase trust; rather, it calibrates trust and improves AI literacy among law students. Furthermore, it describes how law students are willing to use it even when it is not always correct, as the implementation might still increase efficiency.

6 Discussion and Further Research

6.1 Limitations

This study relies on self-reported measures of trust, which are more susceptible to response biases such as social desirability, acquiescence, and naysaying biases. Consequently, the reported trust levels may not fully reflect the participants' actual attitude or behaviour. This also applies during the experiment, since it was an unconcealed experiment; participants' behaviour may have changed based on my reactions. Furthermore, due to the limited time period of the data collection, participant recruitment and data collection were restricted, resulting in a smaller sample than would be ideal for this study. Moreover, the study sample consists of only 10 participants. The statistical power is limited, and the risk of undetected meaningful effects increases. Additionally, the sample might not be a proper representative of the broader population, restricting its external validity.
The adopted theoretical framework in this study focuses on trust in RAG systems based on a limited background theory of interpersonal trust relationships applied to RAG systems. Consequently, broader perspectives from fields such as human-computer interaction are not examined in-depth. Therefore, the findings of this study should only be interpreted based on the background theory written in this study.

6.2 Future Research

To address the limited statistical power of the current study, future research could replicate the same study on a larger sample size. This would increase the statistical power and improve the generalisability of the findings. A larger sample size would be more likely to reflect the characteristics of the population. Additionally, similar studies could implement alternative interaction methods with RAG systems to investigate whether different interaction designs influence users' perceptions of trust.
Future research could investigate trust in RAG systems by using legal documents or cases that participants are already familiar with, thereby creating a more personalised experimental setting. Although such an approach would require additional facilities and a large participant pool to analyse the data properly, familiarity with the source material may reduce uncertainty in understanding the legal content. This would allow participants to focus more directly on evaluating the system's retrieval process, explanations, and outputs. In combination with their prior experience with the materials, participants may be able to assess the system in areas they previously found challenging, determine whether it retrieves the correct information, and assess whether its responses align with their expectations and understanding of the material.
A second option for future research is to conduct a longitudinal study. This could be done by allowing participants to become familiar with RAG systems over an extended period, for example,

while completing a case-based assignment. This may provide insight into how trust develops through repeated exposure and experience. Such a design could investigate whether increased familiarity with the system, together with continued observations of its accuracy, influences perceptions over time.
A third option would be to use a larger knowledge base by increasing either the number of documents, the size of a single document, or both. This method would allow the participant to see more clearly how information retrieval is applied in practice.
Collectively, these suggestions may provide a more comprehensive understanding of the various factors influencing trust in RAG systems. Larger sample sizes would strengthen statistical findings and better reflect the population; personalised experimental settings could improve trust assessments, and longitudinal studies may indicate how trust develops through repeated use of RAG systems.

6.3 Conclusion

This research explored how law students' practical understanding of a Retrieval-Augmented Generation (RAG) system's retrieval process impacts their trust in its legal research outputs. Based on quantitative and qualitative analyses of law students' responses, it can be concluded that accuracy and transparency do not increase overall trust. The quantitative results indicate that gaining practical insight into a RAG system's retrieval process did not significantly affect law students' trust in its outputs for legal research.
The experiment illustrated how transparency and accuracy calibrate trust and increase AI literacy. Depending on the model's competence and the circumstances, users decide whether to rely on the RAG system's outputs. Therefore, calibrated trust becomes a condition for situational trust observed in Söderström's trust model.
Furthermore, concerns about accuracy and, therefore, the system's ability remained. However, concerns about the system's usefulness did not influence AI adoption based solely on the system's ability. The calibrated trust was further adapted, with the potential for increased efficiency in legal research. These findings demonstrate how explainability contributes to trust calibration and adoption of a RAG system.
The participant established a context-dependent relationship in which trust was calibrated per situation, accompanied by perceived risk, and remained dependent on human oversight.
This study contributes to existing research by demonstrating that explainability within a RAG system's retrieval component supports AI literacy and calibrated trust. Although practical insight did not produce a statistically significant change in law students' overall trust, the qualitative finding indicated it influenced how students evaluated and would implement the system in their professional setting. Rather than increasing trust, transparency enables users to better understand both the capabilities and limitations of AI systems. Therefore, it allows them to make more informed decisions about their use.

7 Declaration of Use of Generative AI

ChatGPT was used in this thesis to assist with LaTeX/Overleaf formatting of the tables, the minipage template for the tables and Overleaf errors.

References

[nlt, 2026] (2026). Lawyers increasingly have to convince clients that ai chatbots give bad advice — nltimes.nl. https://nltimes.nl/2026/02/16/lawyers-increasingly-convince-clients-ai-chatbots-give-bad-advice.

[Alansari and Luqman, 2026] Alansari, A. and Luqman, H. (2026). Large language models hallucination: A comprehensive survey.

[Anh-Hoang et al., 2025] Anh-Hoang, D., Tran, V., and Nguyen, L.-M. (2025). Survey and analysis of hallucinations in large language models: Attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence*, 8:1622292.

[Batty, 2025] Batty, D. (2025). 'she helps cheer me up': the people forming relationships with ai chatbots. *The Guardian*.

[Bender et al., 2021] Bender, E., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? pages 610–623.

[Blöbaum et al., 2016] Blöbaum, B. et al. (2016). Trust and communication in a digitized world. *Models and Concepts of Trust Research. Heidelberg et al.: Springer*.

[Brodkin, 2023] Brodkin, J. (2023). Lawyers have real bad day in court after citing fake cases made up by chatgpt.

[Cho et al., 2015] Cho, J.-H., Chan, K., and Adali, S. (2015). A survey on trust modeling. *ACM Computing Surveys (CSUR)*, 48(2):1–40.

[De Streel et al., 2020] De Streel, A., Bibal, A., Frénay, B., and Lognoul, M. (2020). Explaining the black box: When law controls ai.

[DeepL SE, 2026] DeepL SE (2026). DeepL Translator. https://www.deepl.com/translator. Machine translation tool. Accessed: 2026-06-30.

[Eslami et al., 2015] Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., and Sandvig, C. (2015). " i always assumed that i wasn't really that close to [her]" reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 153–162.

[European Commission, 2022] European Commission (2022). Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence (ai liability directive). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496. COM(2022) 496 final, Accessed: 2026-04-20.

[European Parliament and Council of the European Union, 2016] European Parliament and Council of the European Union (2016). Regulation (eu) 2016/679 (general data protection regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj. Accessed: 2026-04-20.

[European Parliament and Council of the European Union, 2024] European Parliament and Council of the European Union (2024). Artificial intelligence act. https://eur-lex.europa.eu. Regulation laying down harmonised rules on artificial intelligence (AI Act), Accessed: 2026-04-20.

[Felzmann et al., 2019] Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1):2053951719860542.

[Gangavarapu, 2025] Gangavarapu, R. (2025). Taming ai hallucinations: Innovations in retrieval-augmented generation and evaluation. In *Mastering AI Governance: A Guide to Building Trustworthy and Transparent AI Systems*, pages 29–33. Springer.

[Gao et al., 2024] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.

[Guidotti et al., 2018] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., and Giannotti, F. (2018). A survey of methods for explaining black box models.

[Gupta et al., 2024] Gupta, S., Ranjan, R., and Singh, S. N. (2024). A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions.

[Guttikonda et al., 2025] Guttikonda, D., Indran, D., Narayanan, L., Pasarad, T., and Sandesh, B. (2025). Explainable ai: A retrieval-augmented generation based framework for model interpretability. In *ICAART (3)*, pages 948–955.

[Han et al., 2026] Han, S., Zhang, Y., Huang, Y., Li, X., Liu, C., and Guo, Y. (2026). Reimagining legal fact verification with genai: Toward effective human–ai collaboration. https://arxiv.org/abs/2602.06305. arXiv preprint arXiv:2602.06305 (not peer-reviewed / not finalised); accessed 2026-02-19.

[Hari, 2025] Hari, S. S. (2025). Understanding feedback loops in machine learning systems. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(2):2810–2823.

[Hassija et al., 2024] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., and Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74.

[IBM, 2024a] IBM (2024a). What are large language models (llms)? https://www.ibm.com/think/topics/large-language-models. Accessed: 2026-03-18.

[IBM, 2024b] IBM (2024b). What is algorithmic bias? https://www.ibm.com/think/topics/algorithmic-bias. Accessed: 2026-03-18.

[IBM, 2024c] IBM (2024c). What is retrieval-augmented generation (rag)? https://www.ibm.com/think/topics/retrieval-augmented-generation. Accessed: 2026-03-18.

[IBM, 2026] IBM (2026). What are ai hallucinations? https://www.ibm.com/think/topics/ai-hallucinations. Accessed: 2026-02-19.

[Jaswal, 2026] Jaswal, A. (2026). 10 real-world examples of retrieval augmented generation. Accessed: 2026-05-11.

[Justwan et al., 2018] Justwan, F., Bakker, R., and Berejikian, J. D. (2018). Measuring social trust and trusting the measure. *The Social Science Journal*, 55(2):149–159.

[Kizilcec, 2016] Kizilcec, R. F. (2016). How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2390–2395.

[Kohn et al., 2021] Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology*, 12:604977.

[Kordzadeh and Ghasemaghaei, 2022] Kordzadeh, N. and Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 31(3):388–409.

[KPMG, 2025] KPMG (2025). Trust, attitudes and use of ai: A global study 2025. https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2025/05/trust-attitudes-and-use-of-ai-global-report.pdf. Global report. Accessed: 2026-02-19.

[Lee and Cha, 2025] Lee, C. and Cha, K. (2025). Toward the dynamic relationship between ai transparency and trust in ai: A case study on chatgpt. *International Journal of Human–Computer Interaction*, 41(13):8086–8103.

[Lendvai and Gosztonyi, 2025] Lendvai, G. F. and Gosztonyi, G. (2025). Algorithmic bias as a core legal dilemma in the age of artificial intelligence: Conceptual basis and the current state of regulation. *Laws*, 14(3).

[Li et al., 2026] Li, Z., Yi, W., and Chen, J. (2026). Accuracy paradox: addressing epistemic, manipulative, and societal risks of hallucination in ai governance. *Computer Law & Security Review*, 61:106311.

[Martinson, 2025] Martinson, S. (2025). What attorneys really think of ai - law360 pulse.

[Meywirth et al., 2026] Meywirth, S., Janson, A., and Söllner, M. (2026). Do human trust models transfer to ai? a necessary condition analysis of interpersonal trust antecedents in large language model-based coaching. *Behaviour Information Technology*, pages 1–19.

[Mittal, 2023] Mittal, A. (2023). The Black Box Problem in LLMs: Challenges and Emerging Solutions — unite.ai. https://www.unite.ai/the-black-box-problem-in-llms-challenges-and-emerging-solutions/.

[Neumeister, 2026] Neumeister, L. (2026). Lawyers blame chatgpt for tricking them into citing bogus past cases in court.

[Ng and Zhang, 2025] Ng, S. W. T. and Zhang, R. (2025). Trust in ai chatbots: A systematic review. *Telematics and Informatics*, 97:102240.

[Ni et al., 2025] Ni, B., Liu, Z., Wang, L., Lei, Y., Zhao, Y., Cheng, X., Zeng, Q., Dong, L., Xia, Y., Kenthapadi, K., et al. (2025). Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*.

[of Human Rights,] of Human Rights, E. C. Case of mennesson v. france.

[Okamura and Yamada, 2020] Okamura, K. and Yamada, S. (2020). Adaptive trust calibration for human-ai collaboration. *Plos one*, 15(2):e0229132.

[Orca Security, 2025] Orca Security (2025). Most popular ai models in 2025. https://orca.security/resources/blog/most-popular-ai-models-2025/. Accessed: 2026-03-18.

[ORQ.ai, 2024] ORQ.ai (2024). Llm use cases: Real-world applications of large language models. https://orq.ai/blog/llm-use-cases. Accessed: 2026-03-18.

[Papagni et al., 2023] Papagni, G., de Pagter, J., Zafari, S., Filzmoser, M., and Koeszegi, S. T. (2023). Artificial agents' explainability to support trust: considerations on timing and context. *Ai & Society*, 38(2):947–960.

[Petroni et al., 2024] Petroni, F., Siciliano, F., Silvestri, F., and Trappolini, G. (2024). Ir-rag @ sigir24: Information retrieval's role in rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 3036–3039, New York, NY, USA. Association for Computing Machinery.

[RAGFlow, 2026a] RAGFlow (2026a). Ragflow. https://ragflow.io/. Accessed: 2026-03-18.

[RAGFlow, 2026b] RAGFlow (2026b). Ragflow documentation. https://ragflow.io/docs. Accessed: 2026-03-18.

[RAGFlow Team, 2026] RAGFlow Team (2026). Ragflow github repository. https://github.com/infiniflow/ragflow. Accessed: 2026-03-18.

[Ravi and Sindhgatta, 2025] Ravi, D. and Sindhgatta, R. (2025). Exploring trust and transparency in retrieval-augmented generation for domain experts. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.

[Renftle et al., 2024] Renftle, M., Trittenbach, H., Poznic, M., and Heil, R. (2024). What do algorithms explain? the issue of the goals and capabilities of explainable artificial intelligence (xai). *Humanities and Social Sciences Communications*, 11(1):1–10.

[Replika, 2026] Replika (2026). Replika: The ai friend to do life with. Accessed May 28, 2026.

[Sassenberg et al., 2026] Sassenberg, A.-M., Acharya, N., Kar, P., and Eshaghi, M. S. (2026). Beyond accuracy: The cognitive economy of trust and absorption in the adoption of ai-generated forecasts. *Forecasting*, 8(1).

[Shin, 2021] Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551.

[Shin, 2022] Shin, D. (2022). How do people judge the credibility of algorithmic sources? *AI & Society*, 37:81–96.

[Söderström, 2009] Söderström, E. (2009). Trust types: an overview. *Discourses Secur Assur Priv*, 15(16):1–12.

[Steup and Neta, 2005] Steup, M. and Neta, R. (2005). Epistemology.

[Suresh and Guttag, 2021] Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9.

[Söllner et al., 2012] Söllner, M., Hoffmann, A., Eike, M., Hirdes, Rudakova, L., Leimeister, S., and Leimeister, J. M. (2012). Towards a formative measurement model for trust.

[Terzidou, 2025] Terzidou, K. (2025). Generative ai systems in legal practice offering quality legal services while upholding legal ethics. *International Journal of Law in Context*, 21(3):431–452.

[Thomson Reuters, 2025] Thomson Reuters (2025). 2025 generative ai in professional services report: Ready for the next step of strategic applications. Special report, Thomson Reuters. Accessed: 2026-06-26.

[Toreini et al., 2020] Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C., and {van Moorsel}, A. (2020). The relationship between trust in ai and trustworthy machine learning technologies. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 272–283. Association for Computing Machinery. Funding Information: This work was funded in part by the UK Engineering and Physical Sciences Research Council for the projects titled "Fintrust: Trust Engineering for the Financial Industry" (EP/R033595/1) and "EPSRC Centre for Doctoral Training in Cloud Computing for Big Data" (EP/L015358/1). Publisher Copyright: © 2020 Copyright held by the owner/author(s). Publication rights licensed to the Association for Computing Machinery.; 3rd ACM Conference on Fairness, Accountability, and Transparency, FAT* 2020 ; Conference date: 27-01-2020 Through 30-01-2020.

[Turnbull, 2023] Turnbull, T. (2023). Replika users fell in love with their ai chatbot companions. then they lost them. *ABC News*.

[Zhao, 2025] Zhao, S. (2025). Understanding rag systems performance by profiling key factors.

[Zhao et al., 2025] Zhao, Y., Wang, B., Wang, Y., Zhao, D., He, R., and Hou, Y. (2025). Explicit vs. implicit: Investigating social bias in large language models through self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1–12.

[Zhou et al., 2024] Zhou, Y., Zhang, W., Shao, J., Liu, Y., Li, X., Jin, J., Qian, H., Liu, Z., Li, C., Zhang, J. C., et al. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102*.

8 Appendix

8.1 Interview Questions

8.1.1 Interview: Pre-experiment

1. Can you please rate the following statements on a scale from 1 to 5, where 1 indicates I strongly disagree, and 5 indicates I strongly agree
   (a) In general, I tend to trust new technology tools before I have much experience with them.
   (b) I generally assume that a new tool will work as intended until proven otherwise
2. Are you familiar with chatbots or LLMs specifically made for the legal field?
3. Have you previously used legal-specific chatbots or LLMs for working with case laws or other legal documents? (e.g. LegalGPT, LexisNexis, Legora)
4. Have you previously used general-purpose chatbots or LLMs for working with case laws or other legal documents?
5. How familiar are you with the use of chatbots or LLMs for legal assistance or retrieval of legal information?
6. To what extent do you trust chatbots or LLMs to assist with your legal academic work? Why, and in what way, do you believe that?
7. What are your primary reasons for not having used chatbots or LLMs for legal academic work?
   - Was a concern regarding accuracy a factor?
   - Did a lack of understanding of how the system operates make you less inclined to use it?
   - Was it a combination of factors, or were there other reasons?
8. What legal tasks have you used chatbots for in the context of your academic studies?
9. What has been your experience regarding the accuracy of LLM-generated responses (or chatbots-generated responses)?
10. On a scale from 1 to 5. Where 1 indicates no understanding, and 5 indicates complete understanding. How well would you estimate you understand how a chatbot or LLM functions? Please explain your rating.
11. Can you please rate the following statements on a scale from 1 to 5, where 1 indicates I strongly disagree and 5 indicates I strongly agree. And please explain your rating.
    (a) I understand how a RAG system decides which information to return in response to a question.
    (b) I understand how a RAG system decides what information to retrieve from the documents.

(c) I feel like I have enough information about how RAG systems work to evaluate whether their outputs are trustworthy.

12. Do you currently have any uncertainty about how RAG systems work?

13. Can you please rate the following statements on a scale from 1 to 5, where 1 indicates I strongly disagree, and 5 indicates I strongly agree. And please explain your rating.

    (a) I believe a RAG system would return correct and relevant information in response to a legal question.

    (b) I think a RAG system would consistently retrieve the most relevant passage from a legal document.

    (c) I am confident that a RAG system's retrieval process would prioritise the information most useful to me.

14. Is there anything specific that concerns or reassures you about the accuracy of RAG systems?

15. Can you please rate the following statements on a scale from 1 to 5, where 1 indicates I strongly disagree, and 5 indicates I strongly agree. And please explain your rating.

    (a) I would be willing to use a RAG system as part of my legal research process.

    (b) I would trust the output of a RAG system enough to use it without manually verifying every result.

    (c) I would feel comfortable relying on a RAG system when preparing work that could affect real legal outcomes.

16. What is the main reason for your decisions you just described?

17. Are you familiar with Retrieval-Augmented Generation (RAG) systems, either within the legal field or more generally?

18. To your knowledge, have you previously used a RAG system in the context of your academic work?

    - (If yes) How would you describe your experience using the system?

19. On a scale from 1 to 5. Where 1 indicates no understanding, and 5 indicates complete understanding. How well do you believe that you understand a RAG system, including the explanation provided? Please explain your rating.

8.1.2 Interview: Post-experiment

1. Can you please rate the following statements on a scale from 1 to 5, where 1 indicates I strongly disagree, and 5 indicates I strongly agree. And please explain your rating.

    (a) I understand how a RAG system decides which information to return in response to a question.

(b) If a RAG system returned an unexpected answer, I would be able to identify why.

(c) I feel like I have enough information about how RAG systems work to evaluate whether their outputs are trustworthy.

2. Do you currently have any uncertainty about how RAG systems work?

3. Can you please rate the following statements on a scale from 1 to 5, where 1 indicates I strongly disagree and 5 indicates I strongly agree. And please explain your rating.

   (a) I believe a RAG system would return correct and relevant information in response to a legal question.

   (b) I think a RAG system would consistently retrieve the most relevant passage from a legal document.

   (c) I am confident that a RAG system's retrieval process would prioritise the information most useful to me.

4. Is there anything specific that concerns or reassures you about the accuracy of RAG systems?

5. Can you please rate the following statements on a scale from 1 to 5, where 1 indicates I strongly disagree, and 5 indicates I strongly agree. And please explain your rating.

   (a) I would be willing to use a RAG system as part of my legal research process.

   (b) I would trust the output of a RAG system enough to act on it without manually verifying every result.

   (c) I would feel comfortable relying on a RAG system when preparing work that could affect real legal outcomes.

6. What is the main reason for your decisions you just described?

7. Compared to before the experiment, my understanding of how the retrieval process works has ...

   (a) Decreased significantly

   (b) Decreased slightly

   (c) Stayed the same

   (d) Increased slightly

   (e) Increased significantly

8. Seeing the retrieval process directly influenced my confidence in the system's accuracy,...

   (a) It made me much less confident

   (b) It made me somewhat less confident

   (c) No change

   (d) It made me somewhat more confident

(e) It made me much more confident

9. After experiencing how the retrieval process works, how would you describe in your own words how the system selects the information it returns?

10. Has your perception of how the system works changed compared to before the experiment? If so, what specifically changed, and if not, what did not change?

11. Was there anything about the retrieval process that surprised you, or that you did not expect?

12. How do you feel about the accuracy of the results you saw during the retrieval process?

13. We discussed that the retrieval will not behave in the same way with different questions. And that there are multiple ways to find the correct values and get the correct chunk as the highest-ranking result, if this is even possible. What did that tell you about the system's reliability?

14. Knowing that the accuracy cannot be guaranteed regardless of how the system parameters are set, does this influence your approach to using the system? And how would it influence you?

15. Has your overall level of trust in this system changed compared to before the experiment? What influenced the change? Or what kept it the same?

16. Would you be willing to use the output of this system without manually verifying the information yourself? And please explain your decision.

    - What role did accuracy play in that direction
    - What role did your understanding of how the system works play in that decision?

17. Would you be willing to use it as an assistive tool if you could manually verify the retrieved information yourself? Please explain your decision.

18. Does knowing how the retrieval process works change your answer compared to what you might have said before?

19. Was there anything specific about the experiment that particularly affected your current understanding and feeling about the system, either positively or negatively?

20. Is there anything else about your experience today that you feel is relevant to how much you would trust or use a system like this?

8.2 Interactive Experiment

8.2.1 Example Questions

1. What was the conclusion in the case of Mennesson v. France?

2. What was the general conclusion?

8.2.2 Task Questions

1. Who are the applicants in this case?
    - Correct chunk holds the word: PROCEDURE
2. What role did the "will of the French people" play in the Government's argument?
    - Correct chunk begins with: actual extent

8.2.3 Script

Prior to the start
Alright, first off, I want to thank you so much for doing this. I have an information sheet here with more information about what this study is about and what will happen with the data. I will not collect any personal data or identifiable information. And all data will be used only in this thesis and deleted from my personal devices as well. I do want to ask if you are comfortable with my voice recording this session? The voice recording will be deleted the moment I document your answer on paper for my thesis. Again, there is no identifiable information, just the answers to the questions, and they are mostly how you feel about chatbots, your experience with them, your possible experience with RAG systems and afterwards your experience and belief in this RAG system, which we will be doing in the interactive part.
If you feel uncomfortable at any point, you can always stop with the experiment. If you wish for me not to use the data, you can tell me up to 1 month later, and I will delete all the data of your participation. As said before, I will be voice recording this, so I can listen back to it and make sure I wrote it down properly.

Interview PT1.

I ask questions 1 - 10 of section 8.1.1
The respondent will answer So, a Retrieval Augmented Generation system, or otherwise called the

RAG system, retrieves information from a document about a specific situation. And uses that information to answer a user's question.
For example, if I wanted to know something about guitars, I could upload a document about guitars and then when I ask questions about them, the RAG system will use the information from the document to answer.

I ask questions 11 - 16 of section 8.1.1
The respondent will answer

So now, I'll explain the RAG system a bit more in-depth.

I will show the graph

This is an RAG system. So it works as we can see in the graph in 2 phases. The retrieval phase

and the generation phase. When the user asks a question, the retriever will try to find the correct information. Retrieving the information based on a question you asked. And using that information to answer your question. It retrieves the information from a database. The user provides the documents to this database, so that's how the system will have documents on that specific topic.

I ask questions 11 - 13 of section 8.1.1
The respondent will answer

Interactive part

Now we will continue with the interactive part of this experiment. I will explain a bit more in-depth how the system works, and then I will give you a task for you to try out.

I show a physical copy of the case law

This is the case law of Menneson v. France. Have you heard of it before? Well, this is the document that we will be using during this experiment. This case law is about a couple in France who used a surrogate mother in California and wanted to bring the baby back to France. But surrogacy is forbidden in France, so the baby was not allowed to go home with the parents. And because of that, the parents argue that Article 8 of the European Convention on Human Rights (the right to private and family life) was violated.

I show the chunks on the screen

This text has been processed by this RAG system. The system has split the text into pieces of around 500 characters. These pieces of text are called chunks. During the retrieval of information, the system is retrieving these chunks. So far, clear? Or do you have any questions?

The respondent will answer
I answer and switch the screen out for the retrieval testing page

This is the retrieval testing page. So that is this step.

I point on the graph

I will explain how it works with 2 example questions, and then you can try it on your own.

I input example question 1 into the text input and press on test

Alright, we have two sliding bars here that we are going to be adjusting during this experiment, and we can then see what effects this will have on the results. On the right, you can see which specific chunks have been retrieved, with the chunks that the system considers best at the top and downwards. First, we have the similarity threshold, which manipulates how many chunks you will retrieve based on the hybrid similarity score, which you can see attached to each chunk. If the value

of the chunk is higher than the threshold, it will be retrieved; it won't be retrieved. And then we have the vector similarity weight, which basically tells the system how much you would like the meaning to influence the rating.
What I mean by that is the following. As we can see, each chunk has 3 scores. The hybrid similarity, term similarity and vector similarity. The term similarity rates on how similar the wording is. So in this case, it can not find a good match within this text chunk. The vector similarity rates the meaning. And the hybrid similarity is a combination score between the meaning and the exact wording, but with a preference for the exact wording. So with the vector similarity, we can adjust how much influence the meaning will have. So if we set the vector similarity weight to a higher value, the meaning will count more, and the hybrid score will adjust to this.
Just a small recap: the similarity score manipulates how many chunks you retrieve, and the vector similarity weight adjusts the order of the chunks as the meaning of the text becomes more or less important. Alright so far, okay? Or do you have any questions?

The respondent will answer
I answer

Then this question, the right answer to this question is written in the following chunk *CTRL + F: Unanimously.* As you can see, it is not the top result. And I adjusted the metrics beforehand, trying to get it to be the top result, except no matter how you adjust the values, it will never appear as the top result. And because of this, the designers of the system made sure to never only use the top result. The default is the top 8 results. Otherwise, the system would never be able to answer correctly.

I input example question 2 into the text input and press on test

However, with this question, which is almost the exact same question, you do immediately get the right chunk as the top result. In this case, you would not like to use the other chunks, and a stricter similarity threshold would be better. Is this all understandable? Do you have any questions?

The respondent will answer
I answer
I give them a list with the questions

Now, I want you to try this for yourself. I'm going to give you 2 questions, where you have to try to get the correct chunk as the top result or at least as high as possible. And to get back little amount of chunks as possible. I will give you the identification to know which chunk is the right one. And while you try this out, I want to ask you if you can say aloud why you are adjusting that metric.
And just as a small reminder, after all this information. The similarity threshold manipulates how many chunks you are retrieving, and the vector similarity weight will adjust how much the meaning influences the chunks. And if you adjust both, it will be a combination. I have a sheet here with the task again, what the metrics mean and the questions.

I wait and discuss with them

Note: make it explicitly clear that there are multiple ways to get question 1 to the top with different values
Alright, that was the interactive part!

Interview PT2.

Then lastly, I want to ask a couple of last questions about your experience with the RAG system, and then we are all done!
I will ask all questions of section 8.1.2
The respondent will provide an answer to each of the questions