Privacy Issues and **Privacy-preserving Mechanisms** in **Retrieval-Augmented Generation** (RAG) Systems

Andreea-Elena Bodea

Monday, 25 November 2024

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

- Motivation
- Research Questions
- Methodology and Expected Outcomes
- Systematic Literature Review
- Plan
- Timeline

© sebis

Motivation

Privacy in RAG systems

Benefits of RAGs:

- extend the knowledge base of LLMs
- improve accuracy & relevance of responses
- reduce hallucination

=> growing adoption

Importance of privacy:

- user trust
- adoption rate
- ethical responsibility
- legal compliance

Implications of privacy breaches and disclosure of sensitive information:

- identity theft
- financial loss
- lower adaption and trust in AI systems

Research in RAGs:

- Focused on improving the RAG architecture (retrieval generation methods)

=> research gap

© sebis

Research Questions

RQ1: What are the privacy-related issues in RAG systems and how can one systematized them?

RQ2: What privacy-preserving mechanisms can be implemented in RAG systems to mitigate the privacy-related issues and how can one systematize them?

RQ3: What are the trade-offs between privacy guarantees and the performance of the RAG systems when implementing privacy-preserving mechanisms?

© sebis

RQ1 - Methodology and Expected Outcomes

RQ1: What are the privacy-related issues in RAG systems and how can one systematized them?

Methodology	Expected Outcomes
• conduct a comprehensive **literature review** by surveying existing academic papers • investigate how RAG systems might leak sensitive information during the retrieval and generation processes • develop a framework or **taxonomy** to systematically categorize the privacy-issues	• first **systematic overview** of privacy issues in RAGs • **categorization** of areas where privacy is at risk • identification of **research gaps**

© sebis

RQ2 - Methodology and Expected Outcomes

RQ2: What privacy-preserving mechanisms can be implemented in RAG systems to mitigate the privacy-related issues and how can one systematize them?

Methodology	Expected Outcomes
• conduct a comprehensive **literature review** by surveying existing academic papers • analyze the applicability and effectiveness of privacy-preserving mechanisms for RAGs • develop a framework or **taxonomy** to systematically categorize the privacy-preserving mechanisms • identify open issues and challenges in applying those mechanisms to RAGs	• first **systematic overview** of privacy-preserving mechanisms in RAGs • **categorization** of privacy mechanisms for easier adoption and adaptation • assessment of the **strengths and weaknesses** of each privacy mechanism in the context of RAG • identification of **research gaps**

© sebis

RQ3 - Methodology and Expected Outcomes

RQ3: What are the trade-offs between privacy guarantees and the performance of the RAG systems when implementing privacy-preserving mechanisms?

Methodology	Expected Outcomes
• **experimental evaluation** through simulations of privacy-preserving RAG models to measure the performance impact in different settings • **comparative analysis** of privacy levels and their effect on performance in various RAG configurations • **case study** to apply findings in practical, real-world scenario	• impact of privacy-enhancing mechanisms on performance metrics such as latency, accuracy, and relevance • quantitative analysis showing the **relationship between privacy guarantees** (e.g., differential privacy levels) and **performance trade-offs** (e.g., slower response times, decreased retrieval quality) • practical guidelines and best practices for selecting privacy mechanisms based on specific performance needs and privacy requirements

© sebis

Systematic Literature Review

1. Research Questions
2. Databases and Research Sources
3. Search Strings
4. Inclusion & Exclusion Criteria

RQ1: What are the privacy-related issues in RAG systems?

RQ2: What privacy-preserving mechanisms can be implemented in RAG systems to mitigate the privacy-related issues?

White literature:

- Google Scholar
- ACM Digital Library
- IEEE Xplore

Grey literature:

- Google search engine
- YouTube

("rag" OR "retrieval augmented" OR "augmented generation")
AND
("private" OR "privacy")

("rag" OR "retrieval augmented" OR "augmented generation")
AND
("attack")

Inclusion Criteria

- Publication Year: articles from 2020 until present
- Search Result Number: articles in the first 100 results

Exclusion Criteria

- Irrelevance: do not address RAG or privacy explicitly (e.g. RAG abbreviation that stands for sth else)
- Duplicate Articles: removed

Systematic Literature Review - 5. Search Results: White & Grey Literature

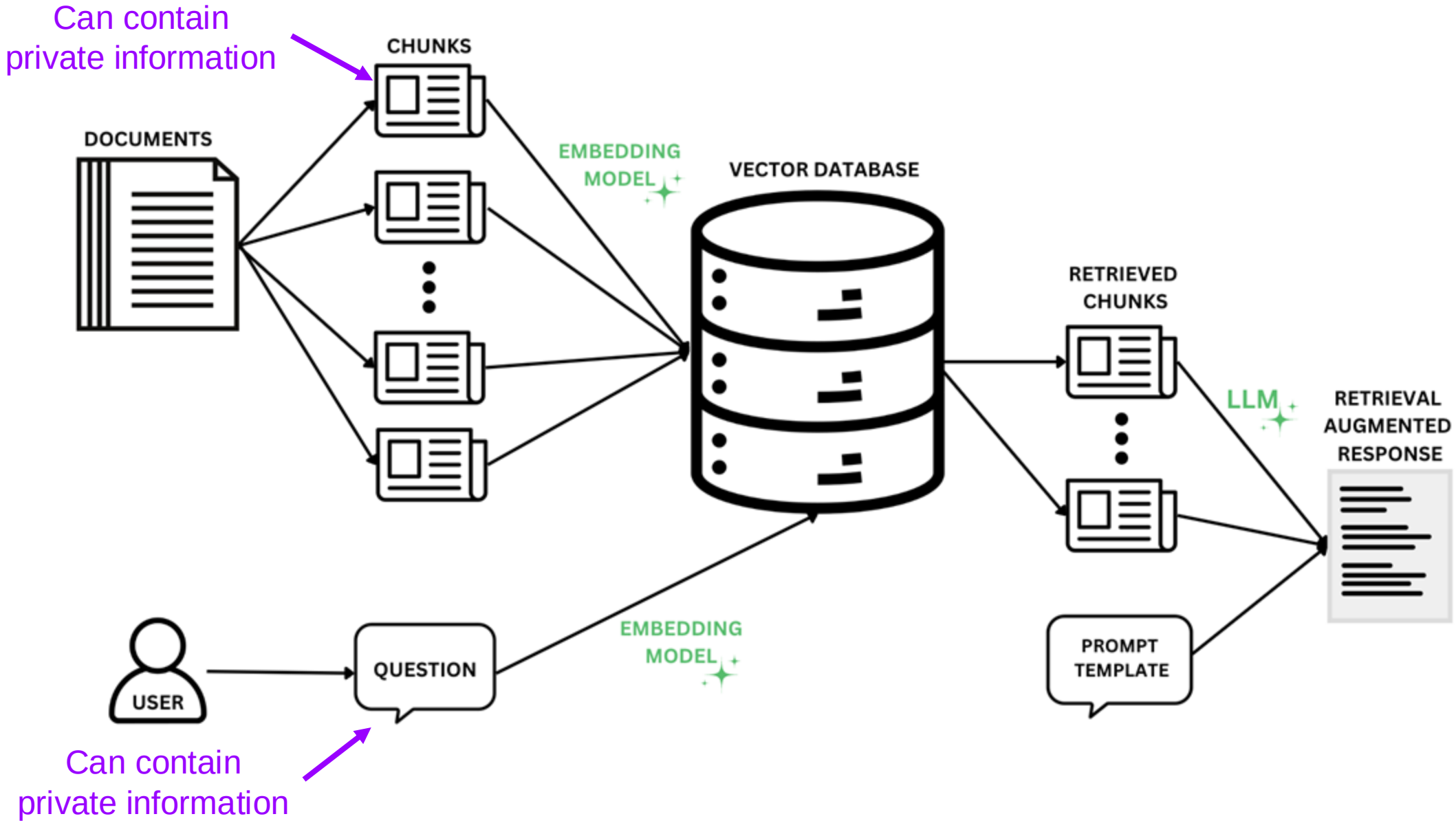
	Google Scholar	ACM	IEEE	Google Search	YouTube
RAG & private & privacy	("rag" OR "retrieval augmented" OR "augmented generation") AND ("private" OR "privacy") • Since 2020 => 21.100 results • Looked at the first 100 results (nothing relevant in the last results) • **Saved 28 papers**	[[Title: "rag"] OR [Title: "retrieval augmented"] OR [Title: "augmented generation"]] AND [[Title: "private"] OR [Title: "privacy"]] • 0 results [[Abstract: "rag"] OR [Abstract: "retrieval augmented"] OR [Abstract: "augmented generation"]] AND [[Abstract: "private"] OR [Abstract: "privacy"]] • 5 results -> 5 relevant	("All Metadata":"rag" OR "All Metadata":"augmented generation" OR "All Metadata":"retrieval augmented") AND ("All Metadata":"private" OR "All Metadata":"privacy") • 23 results -> 8 relevant	28 relevant pages in the first 50 results	6 relevant videos
RAG & attack	("rag" OR "retrieval augmented" OR "augmented generation") AND ("attack") • Since 2020 => 16.400 results • Looked at the first 70 results (nothing relevant in the last results) • **Saved 37 papers** • Irrelevant results: using RAG on documents about attacks	[[Title: "rag"] OR [Title: "retrieval augmented"] OR [Title: "augmented generation"]] AND [Title: "attack"] AND [E-Publication Date: (01/01/2020 TO *)] [[Abstract: "rag"] OR [Abstract: "retrieval augmented"] OR [Abstract: "augmented generation"]] AND [Abstract: "attack"] AND [E-Publication Date: (01/01/2020 TO *) • Both 0 results	("Abstract":"rag" OR "Abstract":"retrieval augmented" OR "Abstract":"augmented generation") AND ("Abstract":"attack") • 2 results -> 0 relevant ("Publication Title":"rag" OR "Publication Title":"retrieval augmented" OR "Publication Title":"augmented generation") AND ("Publication Title":"attack") • 0 results	20 relevant pages in the first 50 results	0 relevant videos

White literature: 78 papers - 5 duplicates + 3 survey papers = 76 papers

Grey literature: 48 links + 6 videos

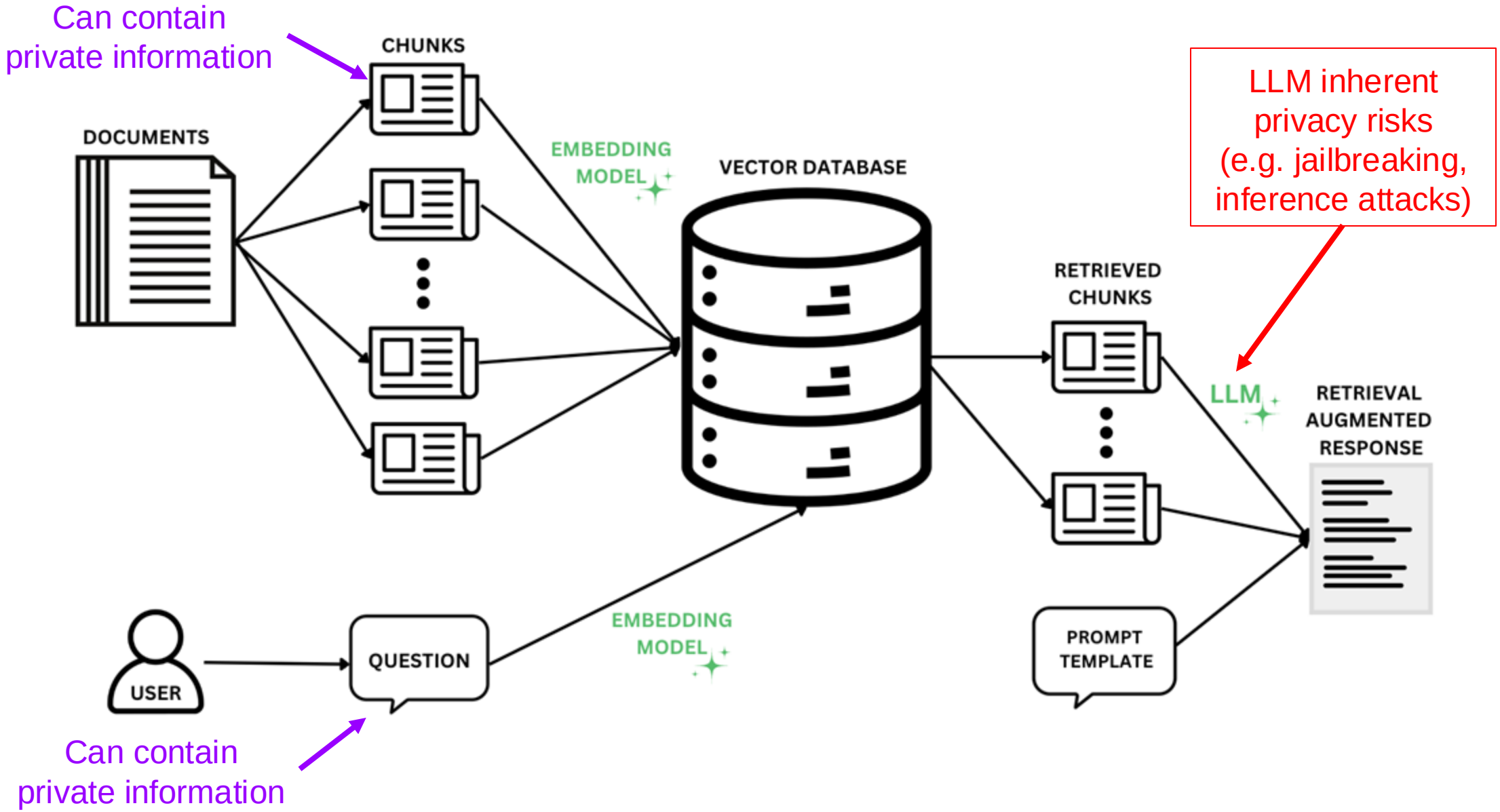
© sebis

Retrieval Augmented Generation (RAG) System



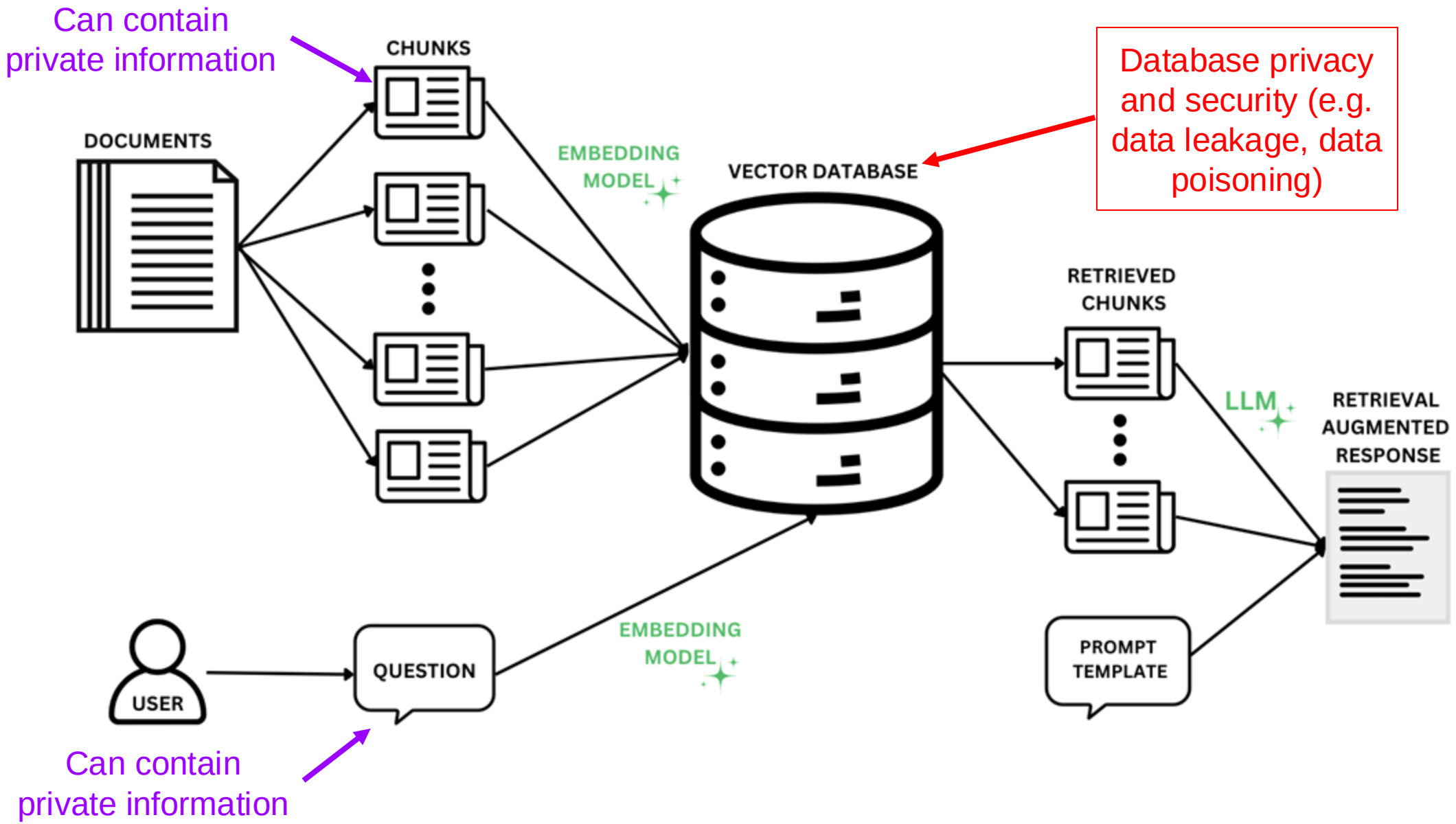
© sebis

Systematic Literature Review - 7. Synthesisation of Findings: RQ1



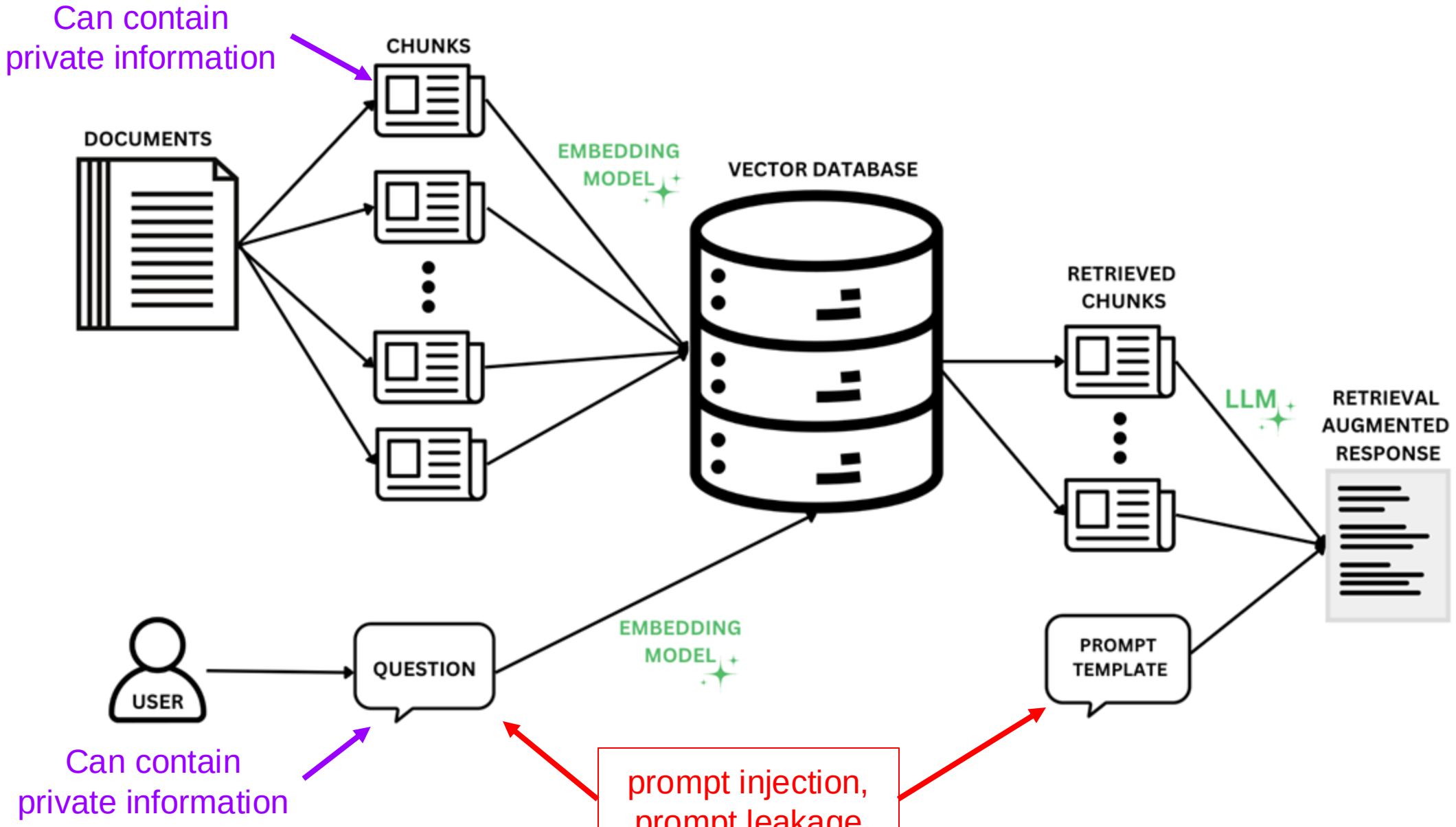
© sebis

Systematic Literature Review - 7. Synthesisation of Findings: RQ1



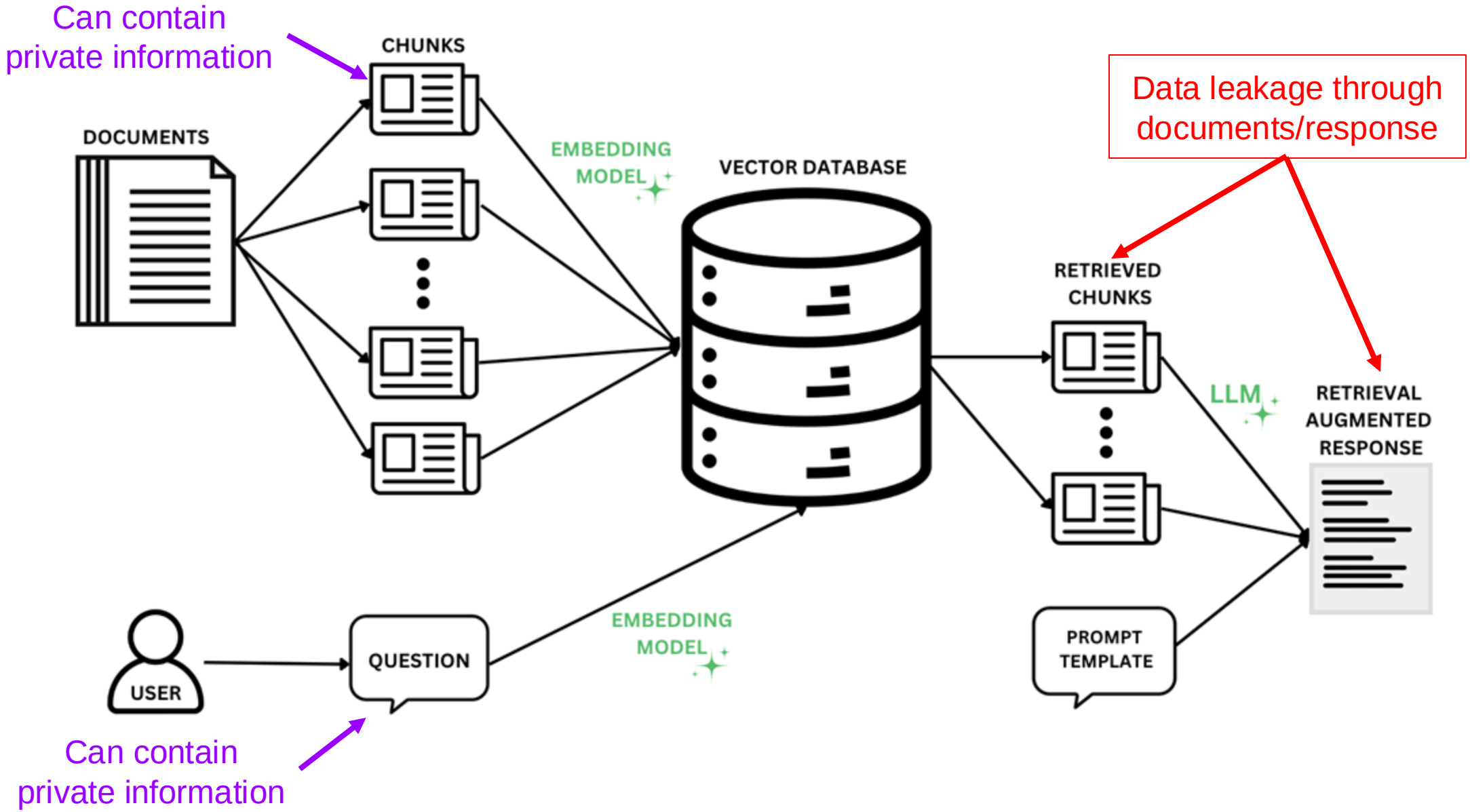
© sebis

Systematic Literature Review - 7. Synthesisation of Findings: RQ1



© sebis

Systematic Literature Review - 7. Synthesisation of Findings: RQ1



© sebis

Systematic Literature Review - 7. Synthesisation of Findings: RQ1

Research in RAG systems focuses on several aspects:

Purpose of RAG:
- For internal use
- For mixed use
- For external use

Type of user

Type of private information

Type of questions asked:
- YES/NO questions
- Open ended / close ended
- Controversial / non-controversial topics

Domain:
- General
- Medicine/Healthcare
- Image synthesis
- Autonomous driving
- Mission & life critical systems
- Education
- Research

Attack:
- Attackers' objective
- Attackers' capabilities
- Type (black-box, grey-box, white box)

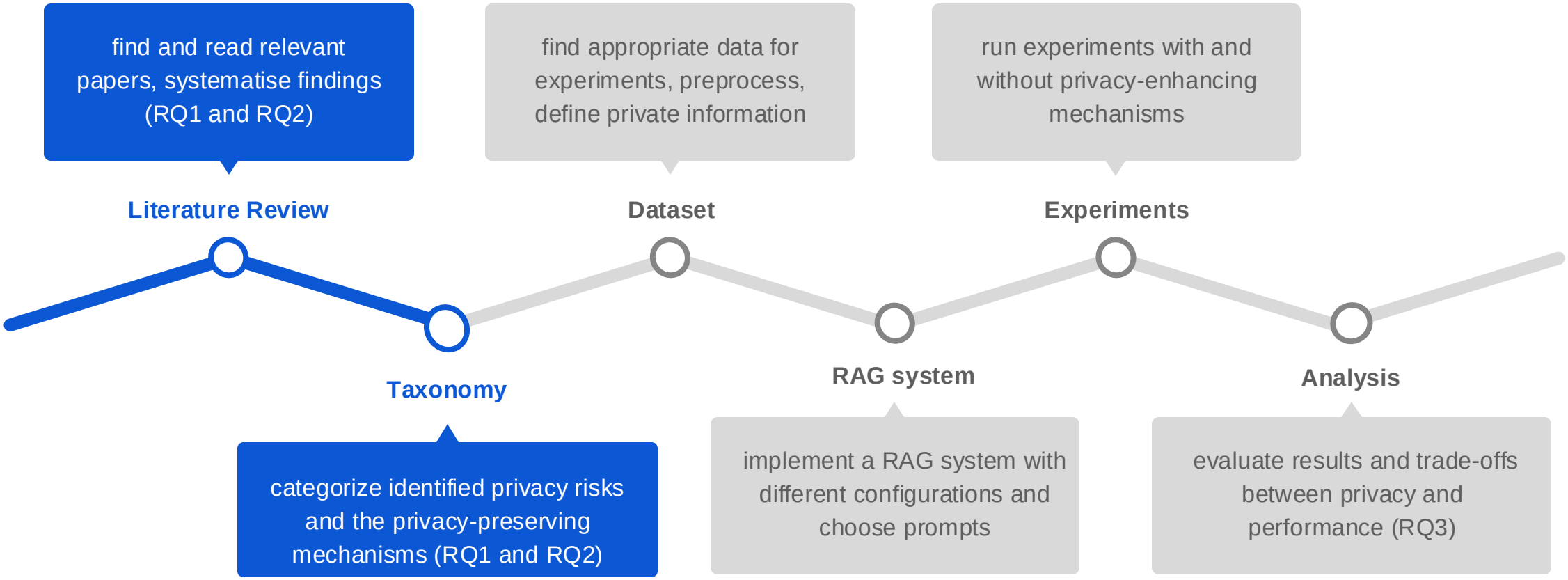
Task:
- Question answering
- Fact checking
- Summarization
- Database SQL queries

Data used in experiments:
- Real world data
- Synthetic data

Type of risk:
- Pre-retrival
- Post-retrieval
- Inherent to LLMs
- Inherent to databases
- Specific to RAGs

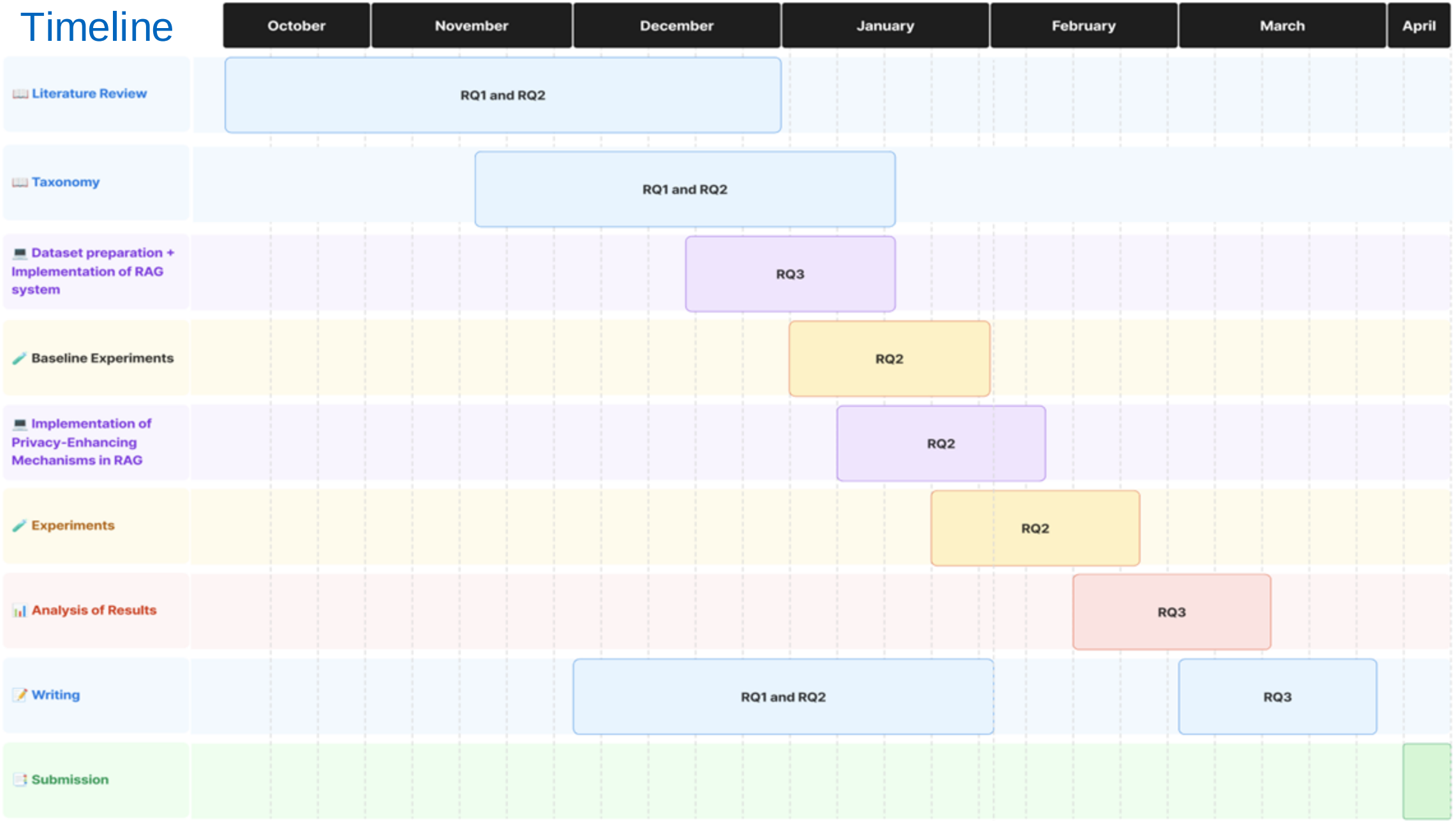
© sebis

Plan


find and read relevant papers, systematise findings (RQ1 and RQ2)
Literature Review
find appropriate data for experiments, preprocess, define private information
Dataset
run experiments with and without privacy-enhancing mechanisms
Experiments
Taxonomy
categorize identified privacy risks and the privacy-preserving mechanisms (RQ1 and RQ2)
RAG system
implement a RAG system with different configurations and choose prompts
Analysis
evaluate results and trade-offs between privacy and performance (RQ3)

© sebis

Timeline



© sebis

Prof. Dr.
Florian Matthes

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de

Notes

Private information

= sensitive information that helps in identifying a person
= sensitive personal information

BUT does NOT necessarily need to be confidential (secret, involving an agreement of nondisclosure)

PRIVATE information	CONFIDENTIAL information
= belonging to or for the use of one particular person or group of people only	= intended to be kept **secret** = sensitive data that is shared under an explicit or implicit agreement that it will not be disclosed without permission
Personal information is private. Personal information is that which is about an identified individual, or an individual who is reasonably identifiable: ● whether the information is true or not; and ● whether the information or opinion is recorded in a material form or not	Confidentiality refers to a relation between a two parties that guarantees any information shared by the first party is treated as private and as such cannot be divulged to third parties without the second party's consent. ● Hospitals and doctors ● Therapists ● Law firms ● Businesses ● Religious authorities ● Financial institutions
Examples: person's name, home address, email address, date of birth, medical information, and bank account details **Examples for companies**: Employee Records, Customer Information, usage data, sensitive corporate data (intellectual property - trade secrets, product development plans, contractual agreements with clients or partners, legal documents, internal communications, and financial records)	Examples: customer names and information, proprietary information such as branding guidelines and databases, supplier names and information, and contract terms
It is still classified as private information even if it enters the public domain.	It is no longer considered confidential if it goes out into the public sphere.
sth can be **private BUT NOT confidential** (ex: name is private but not secret; people you share private information - daily habits, vacation plans etc - with do not have a legal duty to keep it a secret)	sth can be **confidential BUT NOT private** (ex: information that is not personal but still has value that requires restricted access or secrecy -> research and development data of a company is not related to a person so it's not private but it may be secret)

Notes

PRIVACY issues	SECURITY issues
who has access to data and **how it's used**, often involving policies and regulations about handling personal information	**protecting data** from unauthorized access and attacks, focusing on the technical and procedural means to prevent data breaches or theft
Privacy Depends on Security: You can't have privacy without security because if data is not secure, unauthorized access can compromise privacy -> ex: data breach may expose sensitive personal information, directly violating privacy	**Security Does Not Guarantee Privacy**: Strong security measures can protect data, but privacy also involves the ethical use of data. -> ex: if a company securely collects personal information but uses it without proper consent, privacy is still compromised, even though the data may be technically secure
• **Data Collection**: What information is being collected, how much is collected, and whether individuals are informed about it. • **Data Usage**: How collected data is used, whether it aligns with what users have agreed to, and whether data is repurposed without consent. • **Data Sharing**: Whether personal information is shared with third parties without proper consent. • **Individual Control**: How much control a person has over their own data, including the ability to view, edit, or delete it. • **Transparency**: How much individuals know about what data is collected, how it's used, and by whom. • **Legal and Ethical Standards**: Adherence to privacy laws, such as the GDPR (General Data Protection Regulation) which sets guidelines for how personal data should be handled.	• **Data Breaches**: Ensuring that personal and sensitive data is not stolen, leaked, or accessed without permission. • **Unauthorized Access**: Protecting systems and information from unauthorized intrusions, such as hacking attempts or insider threats. • **Encryption**: Using methods to ensure that data is unreadable to unauthorized entities both during storage and transmission. • **Authentication and Authorization**: Mechanisms that ensure only the right people can access certain information (e.g., strong passwords, two-factor authentication). • **Vulnerability Management**: Identifying and fixing security loopholes to prevent exploitation by malicious actors.