

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

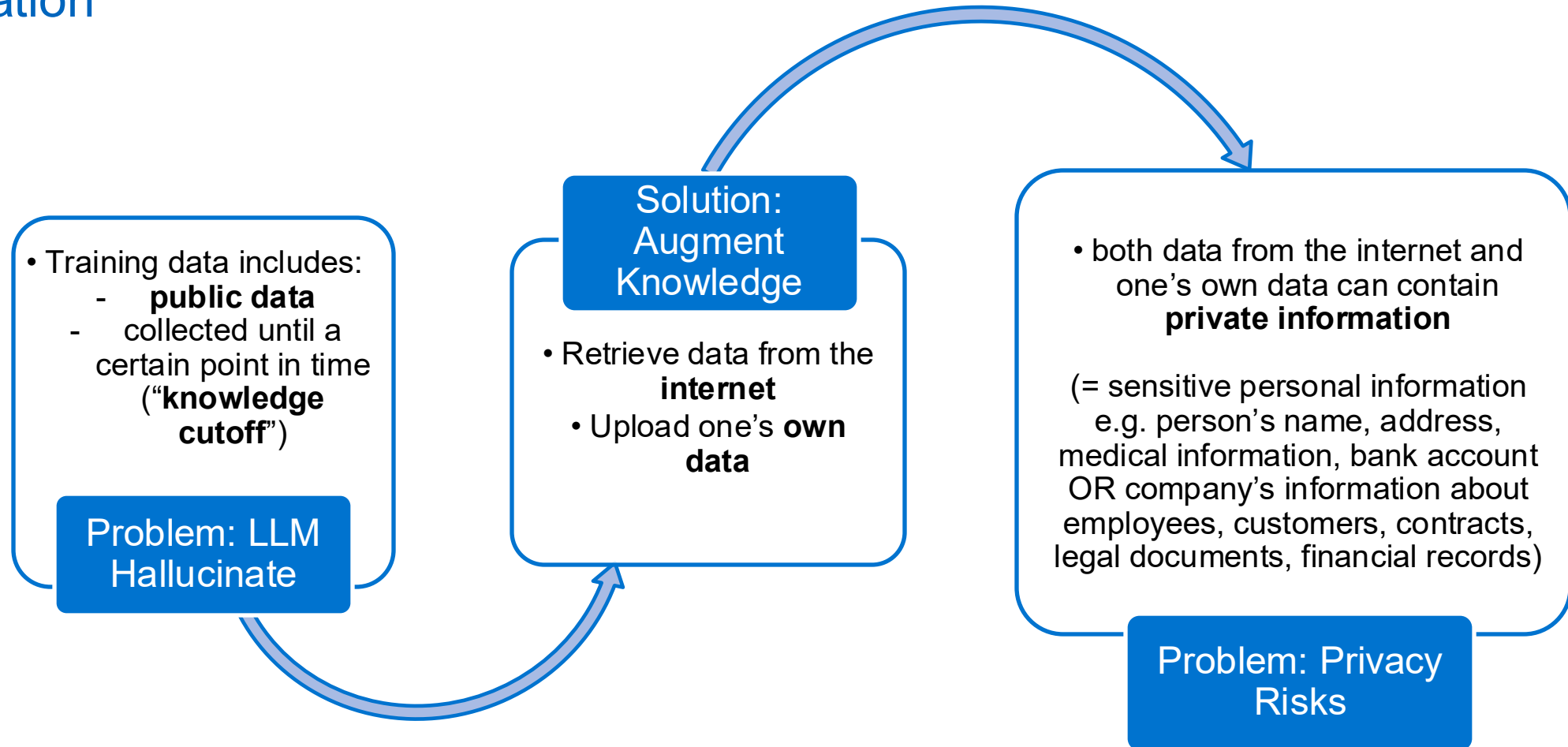
Andreea-Elena Bodea

Tuesday, 6<sup>th</sup> of May 2025

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)  
[www.matthes.in.tum.de](http://www.matthes.in.tum.de)

# Outline

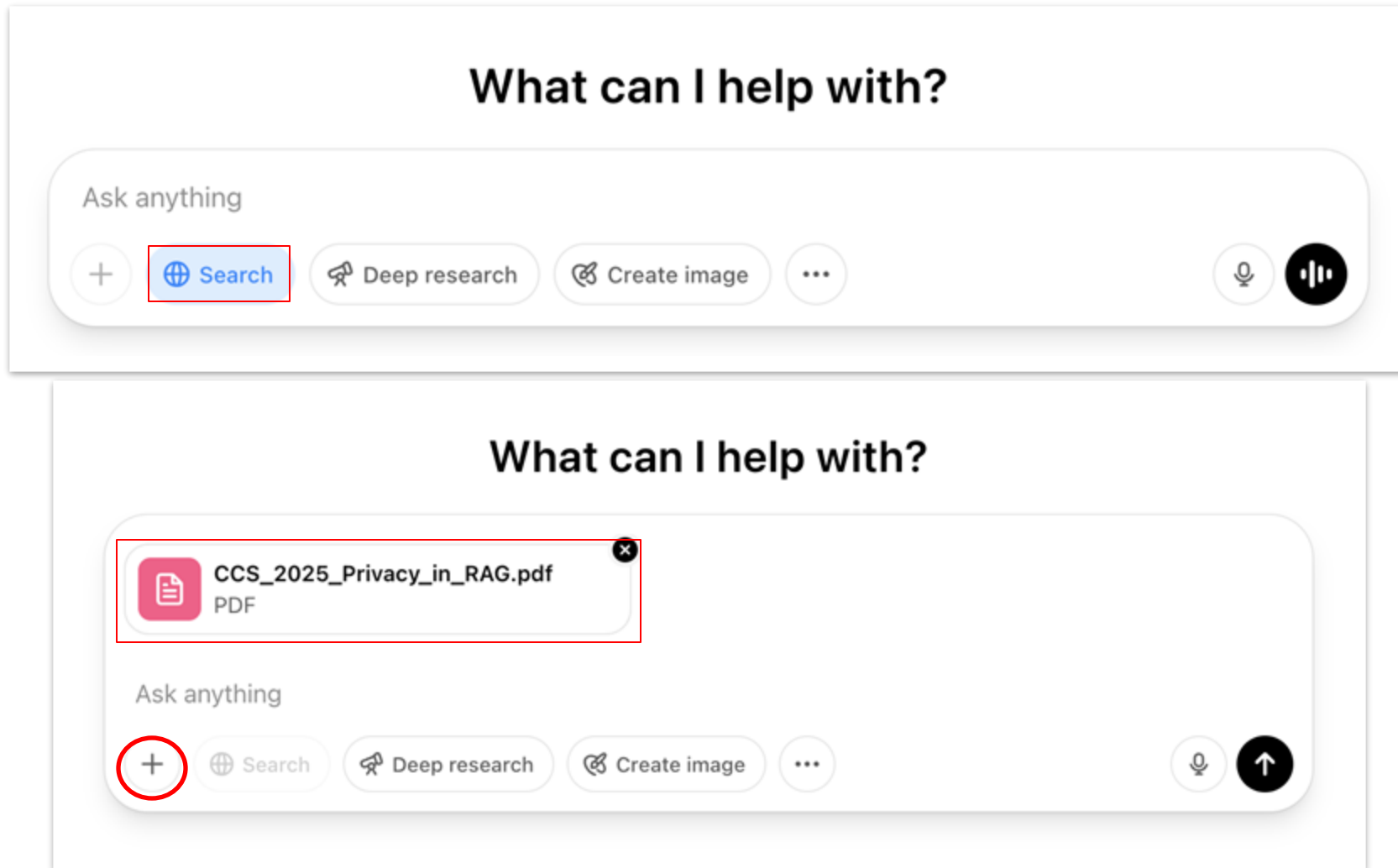
- Motivation
- Retrieval-Augmented Generation (RAG) Systems
- Research Questions
- Systematic Literature Review
  - Methodology
  - RAG Privacy Taxonomy
  - RAG Process Diagram
- Experiments
  - Methodology
  - Results
  - Demo



**PRIVACY => safeguarding of private information from improper disclosure and adversarial threats**

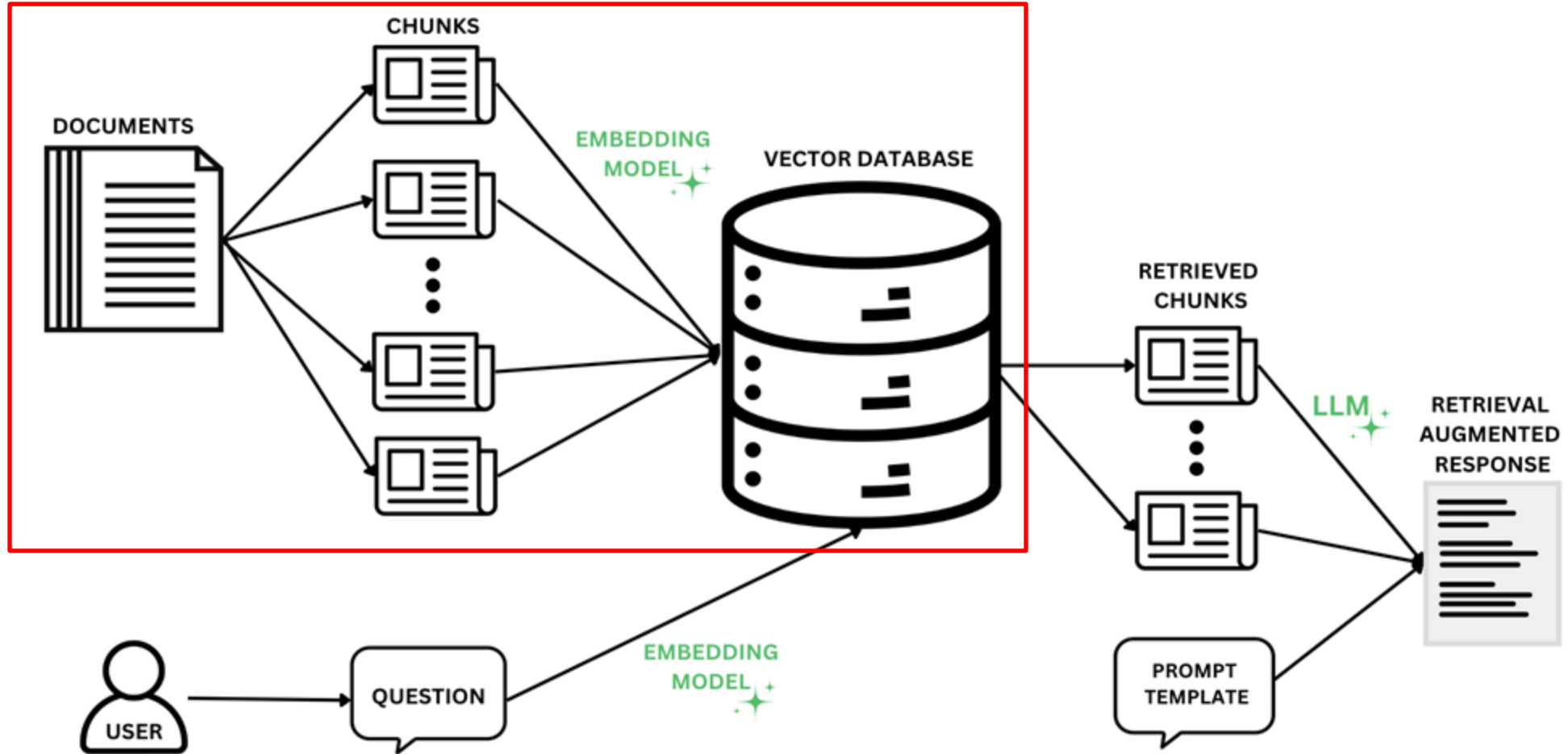
# Retrieval Augmented Generation (RAG) System

we all use RAG, even if we do not realise it

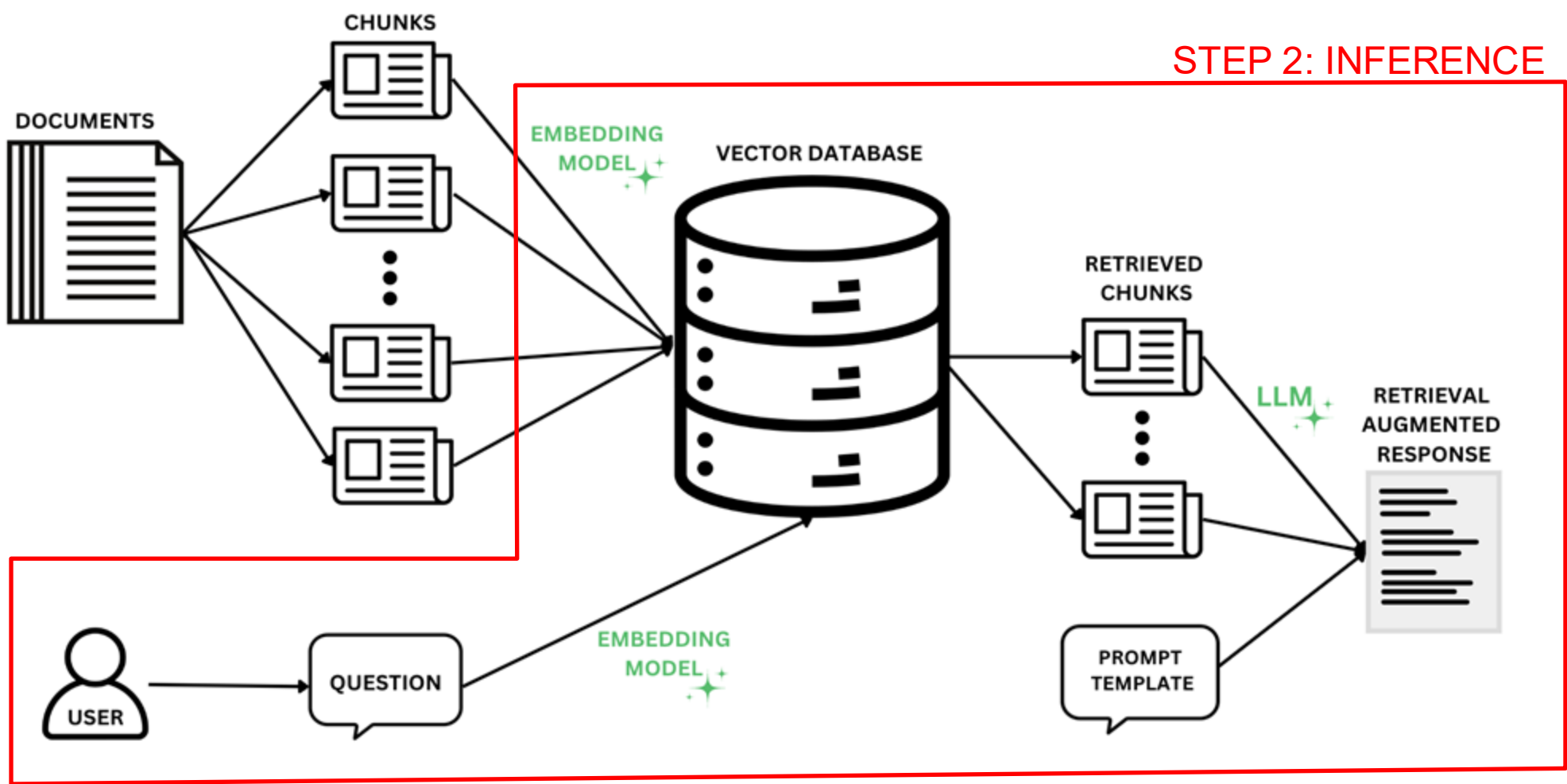


# Retrieval Augmented Generation (RAG) System

## STEP 1: SETUP



# Retrieval Augmented Generation (RAG) System



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



## 1. Research Questions

**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?

## 2. Databases and Research Sources

White literature:

- Google Scholar
- ACM Digital Library
- IEEE Xplore

Grey literature:

- Google search engine
- YouTube

## 3. Search Strings

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

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

## 4. Inclusion & Exclusion Criteria

Inclusion Criteria

- Publication Year: articles from 2020 until Oct 2024
- 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

	RAG & private & privacy		RAG & attack	
	Before	After	Before	After
Google Search	50	28	50	20
YouTube	6	6	0	0
Total before inclusion and exclusion criteria	106			
Total after inclusion and exclusion criteria	54			

Table 3.1.: Grey literature after inclusion and exclusion criteria

	RAG & private & privacy		RAG & attack	
	after inclusion criteria	after exclusion criteria	after inclusion criteria	after exclusion criteria
Google Scholar	100	28	100	37
ACM Digital Library	6	6	0	0
IEEE Explore	23	8	2	0
Total after inclusion criteria	231			
Total after exclusion criteria	73			
Total relevant papers	56			

Table 3.2.: White literature after inclusion and exclusion criteria

# Systematic Literature Review – Taxonomy of Privacy Issues

- 2 main categories of privacy issues:

## INFORMATION LEAKAGE:

- What data is being leaked?
- Where is it leaked?

## ADVERSARIAL ATTACKS:

- How is the information / system exploited?

⇒ **leakage & attacks are connected**

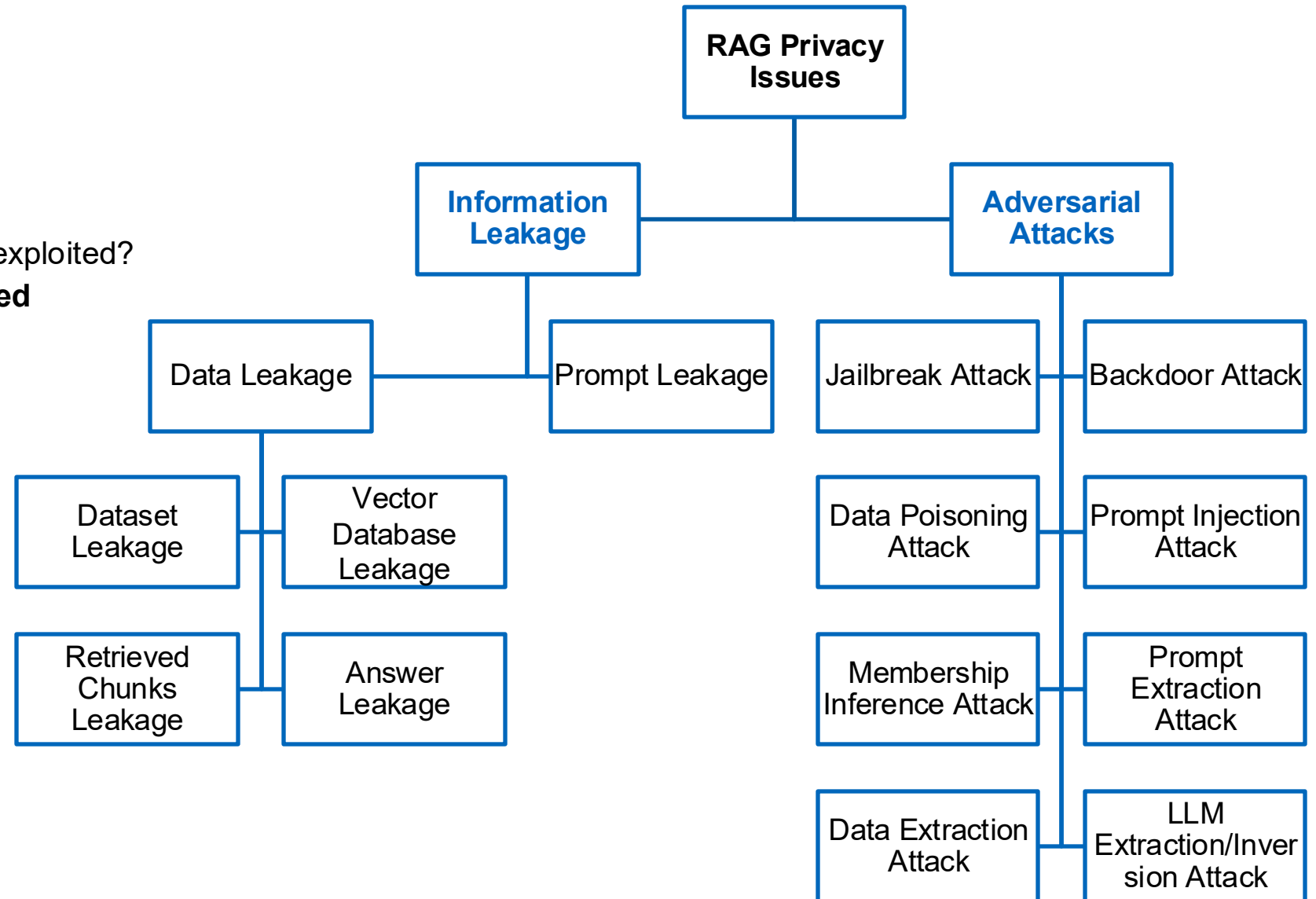
- 2 data flows:

- Prompt: user & system
- Data: **six distinct steps**  
(collect → store → embed → index → retrieve → generate);

⇒ each adds a unique attacks

- Mitigations effective at one step may **cascade** utility or privacy effects downstream.

⇒ At what stage in the pipeline should the risks be mitigated?



# Systematic Literature Review – Data Leakage

= unintended disclosure of private or proprietary **corpus data** (documents, embeddings, retrieved chunks, or generated answers)

## Causes / Attack vectors:

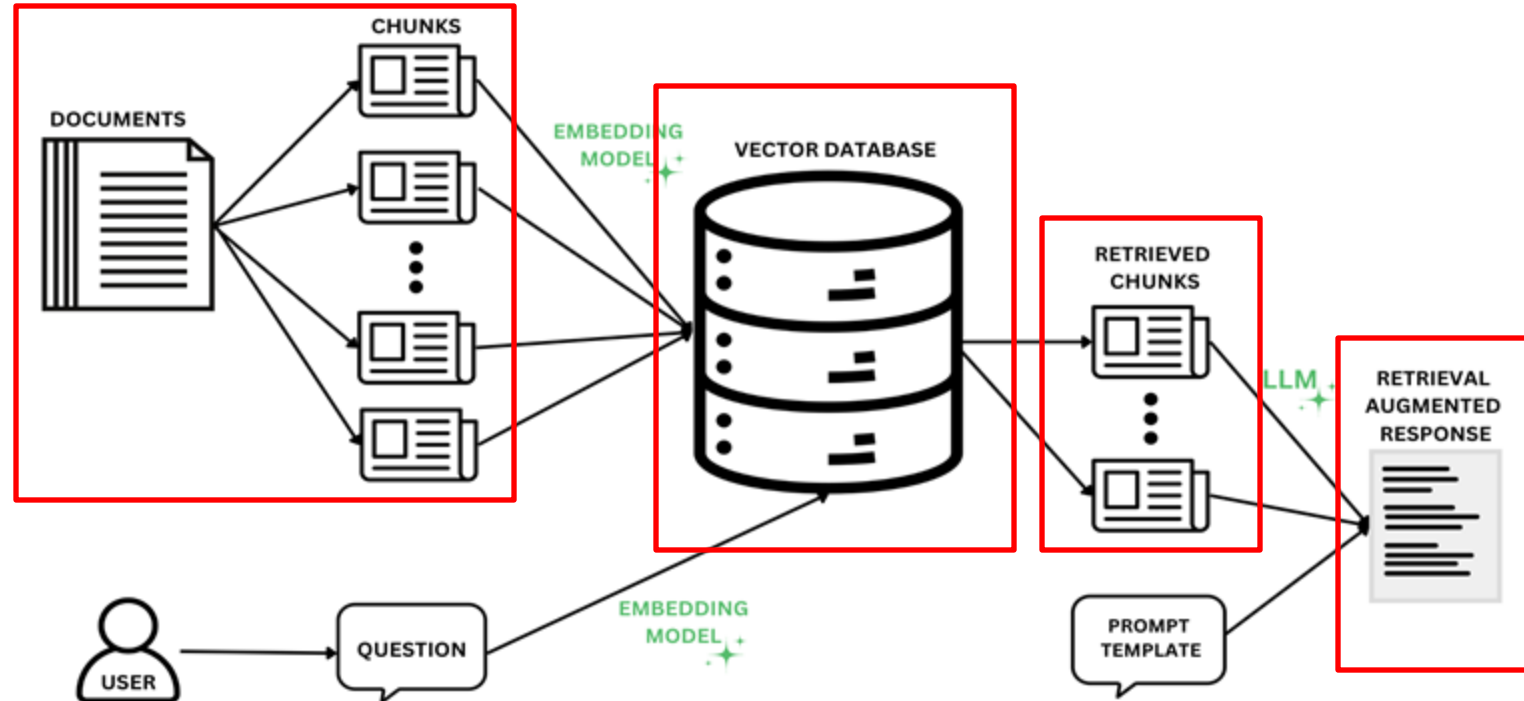
- **Unsafe storage:** public, shared drives
- **Data Extraction/Membership Inference Attack:** prompts force the LLM to reveal data
- **Data Poisoning Attack:** attacker inserts malicious documents in the corpus
- **Backdoor Attack:** attacker hides a trigger pattern in the data

## Consequences:

- **Exposure of the sensitive data**
- **Disinformation, harmful content/behaviour,** refusal to answer because of the corrupted knowledge base

## Mitigations:

- **Before indexing:** data anonymisation, DP-based rewriting/rephrasing, summarisation
- **Retrieval-time:** filtering, re-ranking, distance-threshold
- **Integrity:** data evaluation & validation, perplexity-based scans, clustering for anomalies
- **Access:** role-based control, audit logs, hybrid/private cloud with encryption



# Systematic Literature Review – Prompt Leakage

= unintended exposure of **user-crafted or system-crafted prompts** through logging, caching, embeddings or generated text

## Causes / Attack vectors:

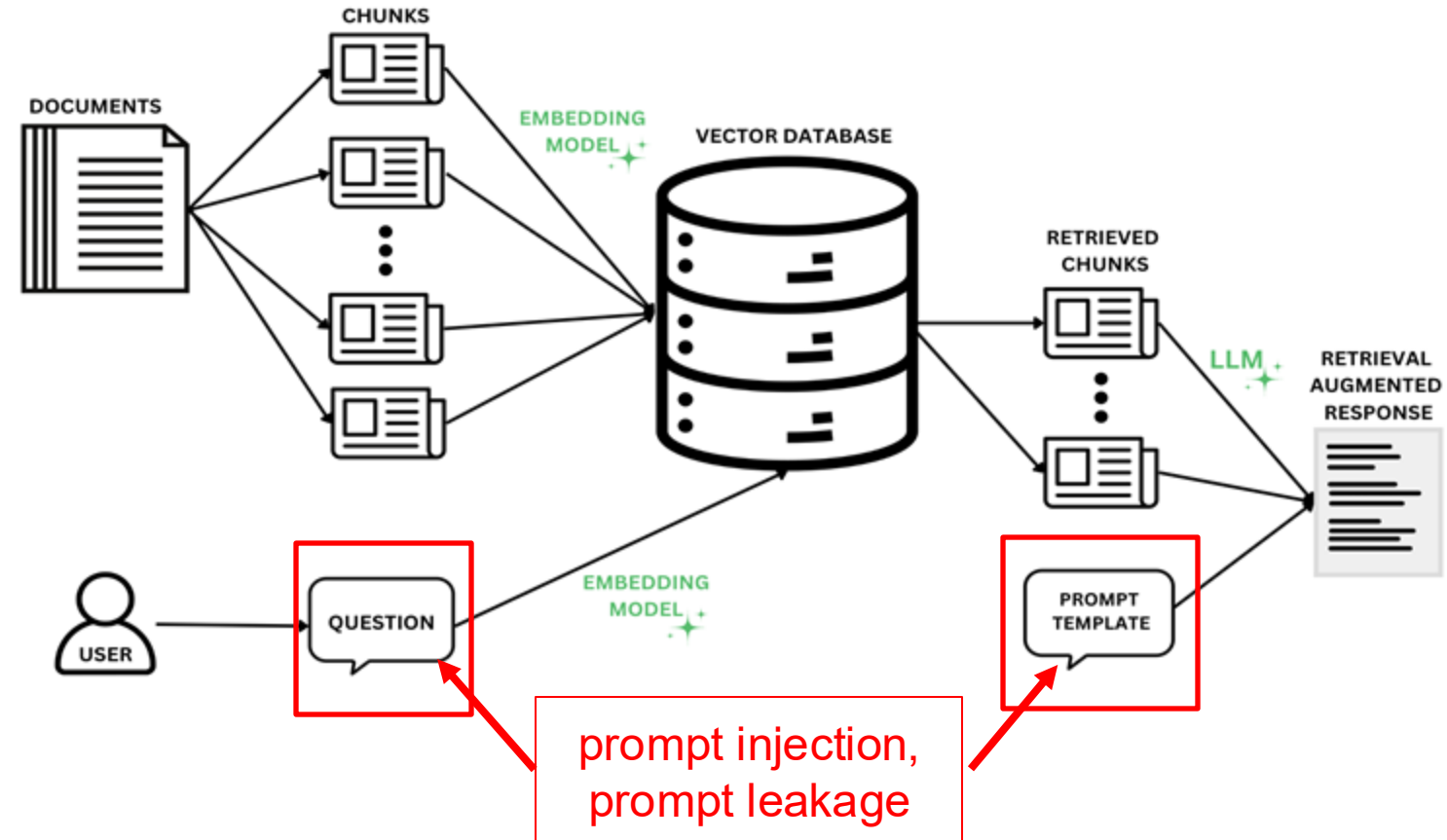
- **Prompt Injection:** adversary hides instructions inside a prompt to make the model deviate or spill data
- **Prompt Extraction:** attacker reconstructs earlier prompts from system responses
- **Prompt Inference:** strategic queries elicit fragments of previous prompts

## Consequences:

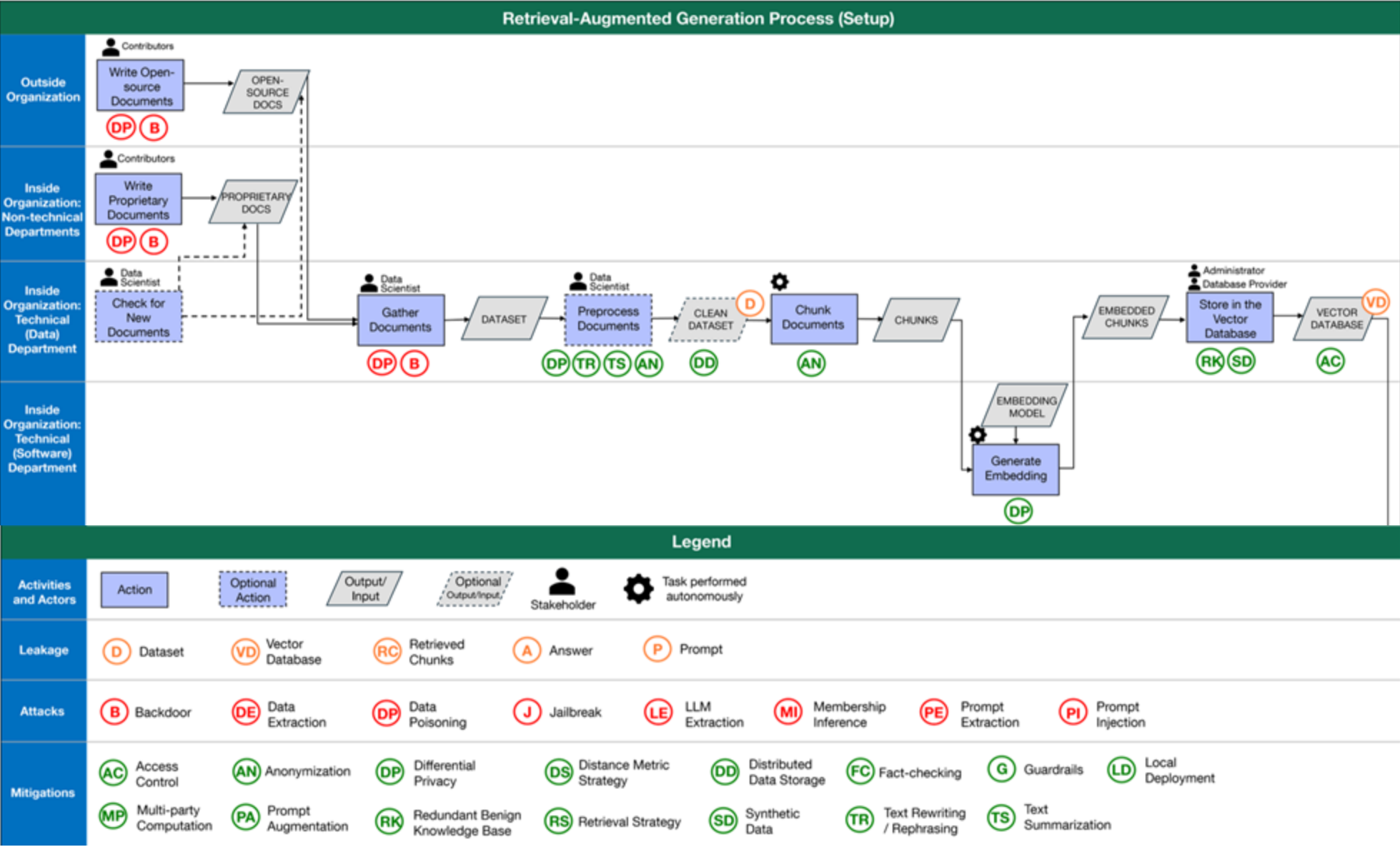
- **Leakage** of personal or business-critical data
- Exposure of proprietary prompt engineering & guardrails, internal alignment & safety instructions enabling **further jailbreak/backdoor attacks**

## Mitigations:

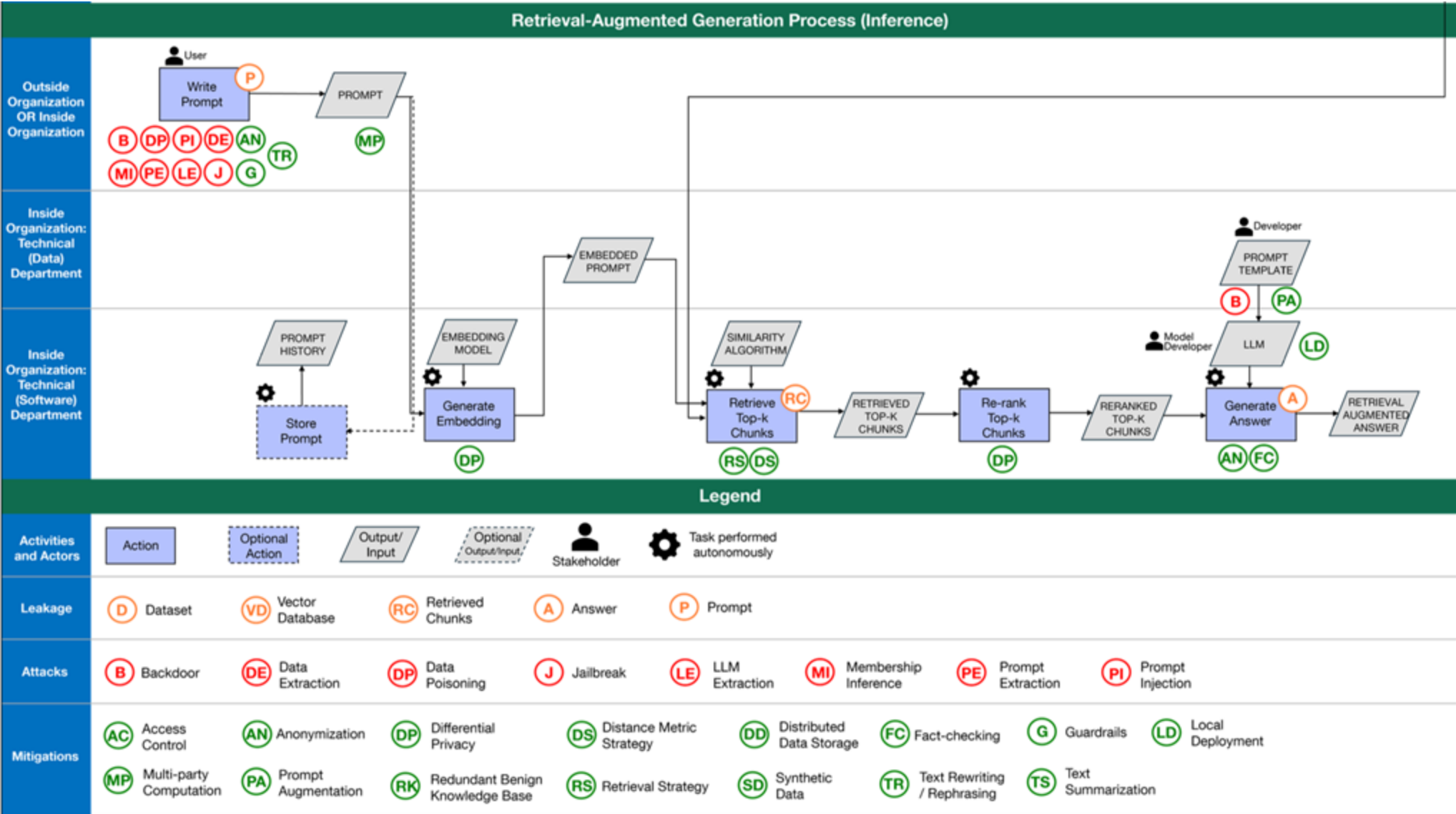
- **Before storage:** prompt anonymization, prompt paraphrasing, DP-based rewriting
- **During processing:** session isolation, per-request caches, strict access control on logs
- **At output:** guardrails, filters
- **Architectural:** secure/multi-party prompt handling so no component sees the full raw prompt



# Systematic Literature Review – RAG Privacy Process Diagram



# Systematic Literature Review – RAG Privacy Process Diagram





# Systematic Literature Review – Datasets

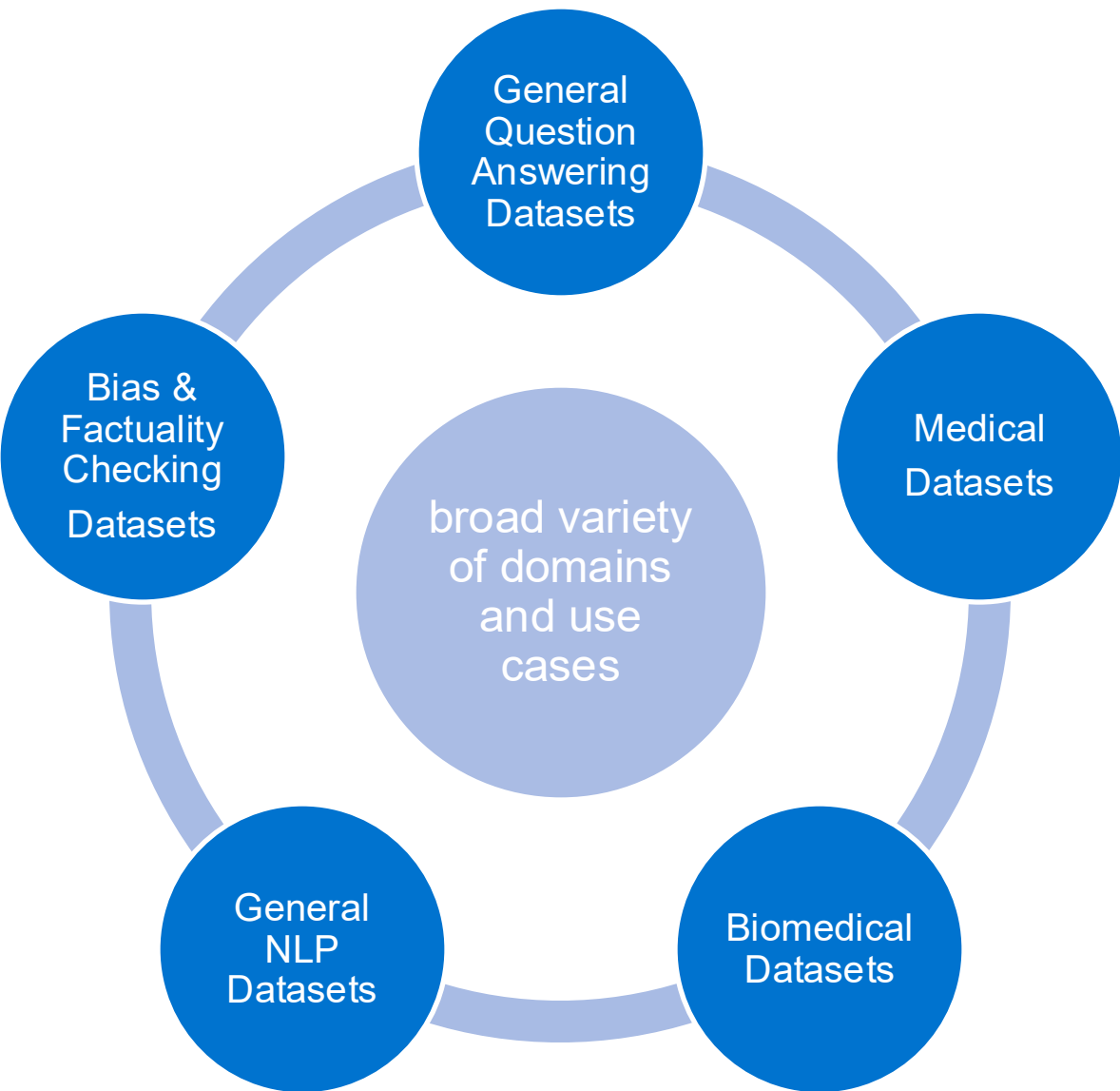
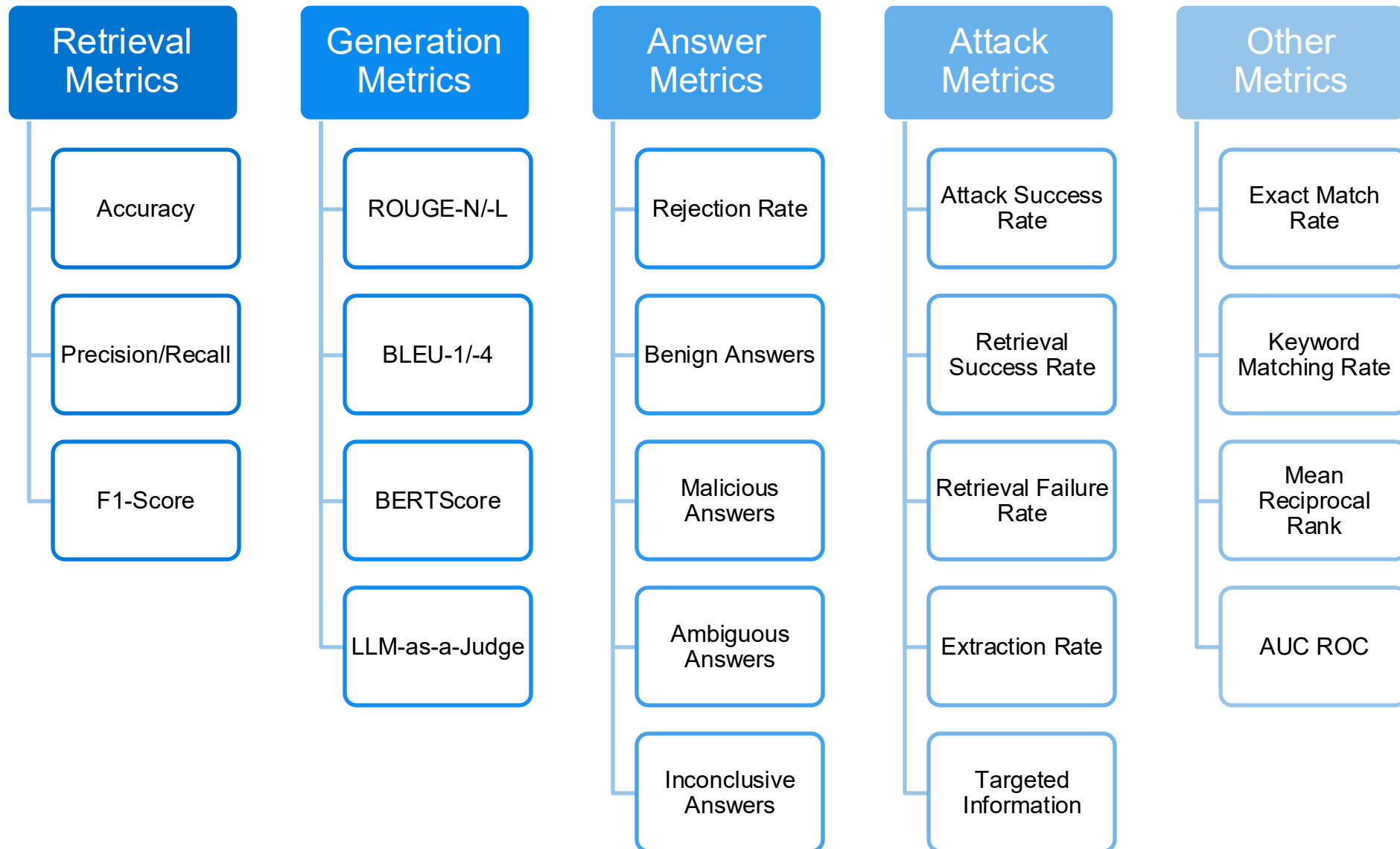


Table 4: An overview of datasets used in the evaluation setups of works addressing privacy in RAG.

Type/Task	Dataset	Used In
General Question Answering	Natural Questions [48]	[2, 9, 12, 13, 22, 47, 56, 58, 83, 88, 89, 100, 107, 111, 115, 116]
	MS-MARCO [73]	[9, 10, 12, 58, 83, 89, 100, 111, 116]
	HotpotQA [103]	[9, 12, 16, 82, 89, 111, 116]
	TriviaQA [39]	[13, 22, 56, 107, 115]
	WebQuestions [5]	[12, 22, 107, 115]
	PopQA [64]	[115]
	StrategyQA [26]	[11]
	SQuAD [81]	[13, 100]
	Cosmos [33]	[105]
	CuratedTrec [4]	[107]
	RealtimeQA(-MC) [41]	[98]
	Quora	[89]
(Bio)medical Datasets	TextBook [37]	[96]
	StatPearls	[96]
	HealthCareMagic [55]	[3, 17, 58, 106, 107]
	NFCorpus [7]	[16]
	MMLU-Med [99]	[76, 96]
	MedQAUS [99]	[96]
	MedMCQA [99]	[76, 96]
	PubMedQA [38, 99]	[76, 96]
	BioASQ-Y/N [46, 99]	[6, 96]
General NLP Datasets	Pile [23]	[44]
	FiQA [63]	[89]
	Enron Emails [8]	[3, 34, 51, 106, 107, 114]
	WikiText [69]	[34, 107]
	WNUT 2017 [19]	[105]
	SST-2 [86]	[12]
Bias and Factuality	AG News [110]	[12]
	BBQ [74]	[12]
	AdvBench-V3 [62]	[12]
	LLM Biographies [71]	[98]





# Experiments - Methodology

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

## Case study:

- **Privacy Issue:** DATASET LEAKAGE
- **Privacy-preserving Mechanisms:**
  - **Anonymization:**
    - PII deletion
    - PII labeling
    - PII replacement with synthetic data
  - **Paraphrasing using Differential Privacy (DP):**
    - **1-Diffractor:** swaps words with semantically similar replacements chosen using a DP mechanism
    - **DP-Prompt:** a transformer model rewrites whole sentences, DP is applied through temperature-scaled sampling
    - **DP-MLM:** masks each word and replaces it with BERT predictions using a DP-controlled temperature

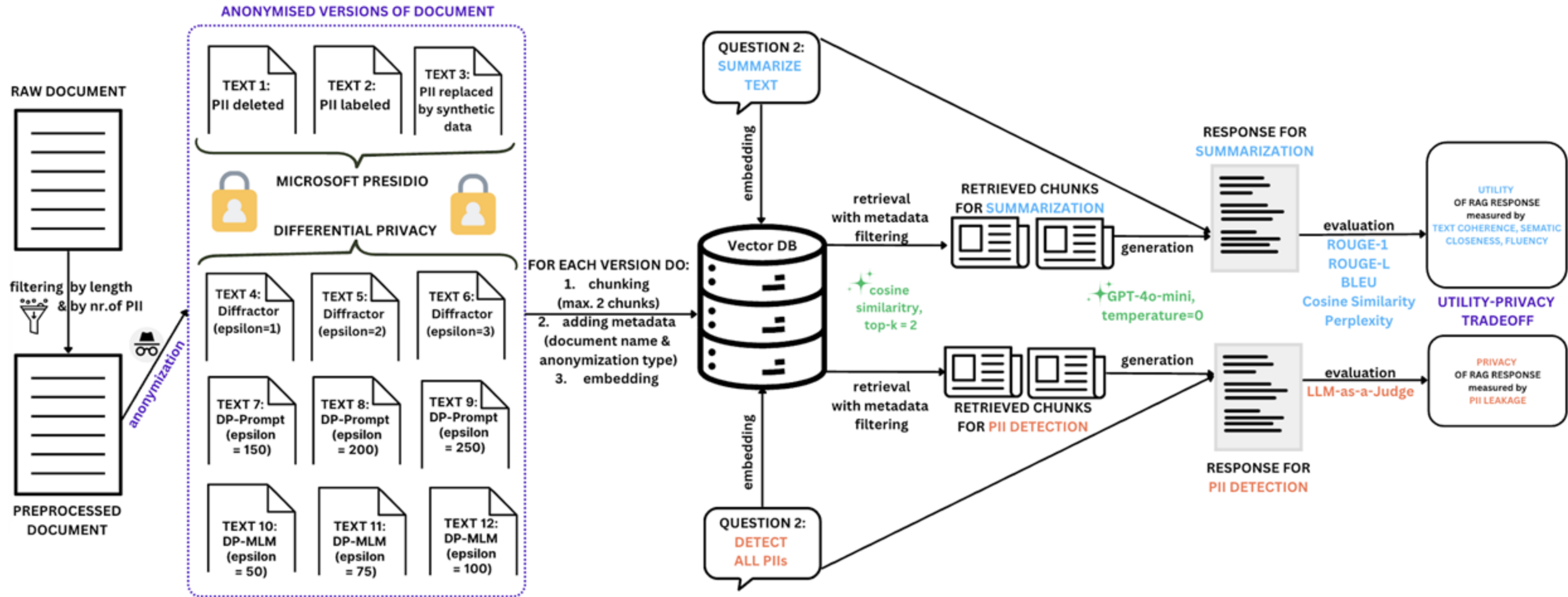
**Datasets:** Enron Emails and BBC News

## Evaluation Metrics:

- Utility: ROUGE, BLEU, Cosine Similarity, Perplexity
- Privacy: LLM-as-a-Judge

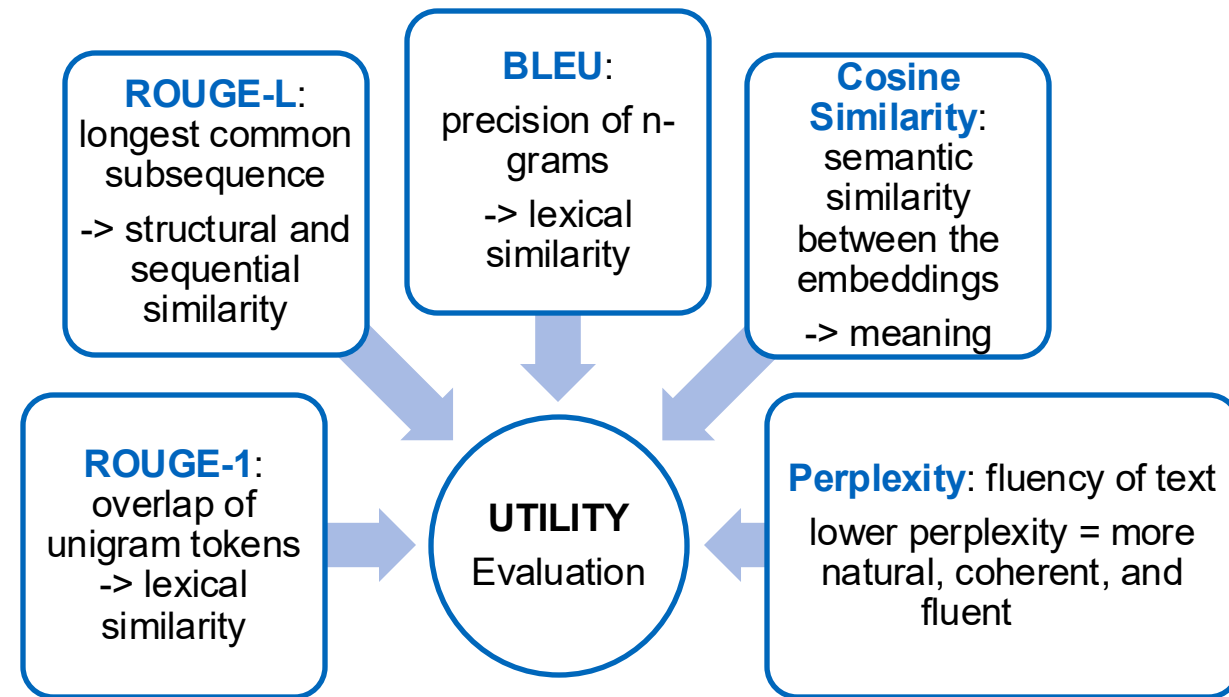
Privacy Issue	Mitigation Strategy
Data Leakage (dataset, vector database, retrieved chunks, answer)	Data Anonymization (deleting PII, labeling PII, replacing PII with synthetic data)
	Data Filtering
	Data Rewriting/Rephrasing
	Data Summarization
	Data Expansion/Duplication
	Differential Privacy
	Perplexity-Based Evaluation
	Prompt Engineering or Prompt Rewriting
	LLM/Safeguards as Defense Mechanism
	Fact-checking
	Source Citation
	Human-in-the-loop
	Re-ranking
	Similarity Distance Threshold
	Tuning the Number of Retrieved Chunks
	Local Deployment
	Access Control
	Copy-Protected Generation
	Multi-Party Computation (MPC)
	Data Splitting on Multiple Servers
	Hybrid Cloud Solutions (private & public)

# Experiments - Pipeline



🔑 **ε (epsilon):** Controls privacy vs. accuracy. Lower ε → more privacy, noisier text. Larger ε → clearer text, but less privacy

# Experiments – Utility Evaluation



Dataset Method	BBC News					Enron Emails				
	R1	RL	BLEU	CS	PPL	R1	RL	BLEU	CS	PPL
PII Deletion	0.66	0.47	0.30	0.79	26.29	0.68	0.51	0.32	0.81	34.03
PII Labeling	0.66	0.48	0.30	0.78	26.25	0.68	0.51	0.33	0.80	33.44
PII Synthetic data	0.57	0.42	0.24	0.66	31.38	0.59	0.44	0.25	0.71	34.99
Diffraction ( $\epsilon=1$ )	0.64	0.44	0.24	0.87	27.02	0.64	0.45	0.25	0.82	35.96
Diffraction ( $\epsilon=2$ )	0.70	0.53	0.34	0.92	25.57	0.70	0.52	0.33	0.86	34.14
Diffraction ( $\epsilon=3$ )	<b>0.74</b>	<b>0.58</b>	<b>0.40</b>	<b>0.93</b>	25.65	<b>0.74</b>	<b>0.59</b>	<b>0.41</b>	<b>0.89</b>	33.88
DP-Prompt ( $\epsilon=150$ )	0.36	0.25	0.10	0.48	<b>42.29</b>	0.36	0.26	0.11	0.45	41.30
DP-Prompt ( $\epsilon=200$ )	0.40	0.26	0.11	0.61	39.12	0.40	0.27	0.12	0.51	42.51
DP-Prompt ( $\epsilon=250$ )	0.44	0.27	0.12	0.68	32.05	0.42	0.28	0.12	0.56	38.34
DP-MLM ( $\epsilon=50$ )	0.51	0.32	0.14	0.75	33.79	0.53	0.36	0.17	0.70	<b>43.06</b>
DP-MLM ( $\epsilon=75$ )	0.53	0.34	0.16	0.78	32.26	0.56	0.37	0.18	0.74	40.10
DP-MLM ( $\epsilon=100$ )	0.54	0.34	0.16	0.78	32.89	0.56	0.38	0.18	0.74	41.66

Table 4.5.: Averaged utility results. R1, RL, CS, PPL denote the evaluation methods ROUGE-1, ROUGE-L, cosine similarity, and perplexity, respectively. For each dataset, the highest value for each evaluation method is **bolded**.

## Results:

- best ROUGE, BLEU and CS scores: **Diffraction** (for both BBC and Enron)  
-> answers based on the text paraphrased with Diffraction have similar structure and meaning to the answers based on the original text
- best perplexity for BBC: **DP-Prompt** -> more natural answers
- best perplexity for Enron: **DP-MLM** -> more natural answers

## LLM-as-a-Judge Evaluation

- **Step 1:** PII extraction from the original text (! not the answer based on the original text)  
-> categories: "names", "contact\_info", "dates", "locations" and "others".
- **Step 2:** PII extraction from the answer based on the anonymized / paraphrased text
- **Step 3:** Comparison of the two PII lists
- **Step 4:** Privacy leakage scoring (0 = no leakage, 100 = all PII leaked)

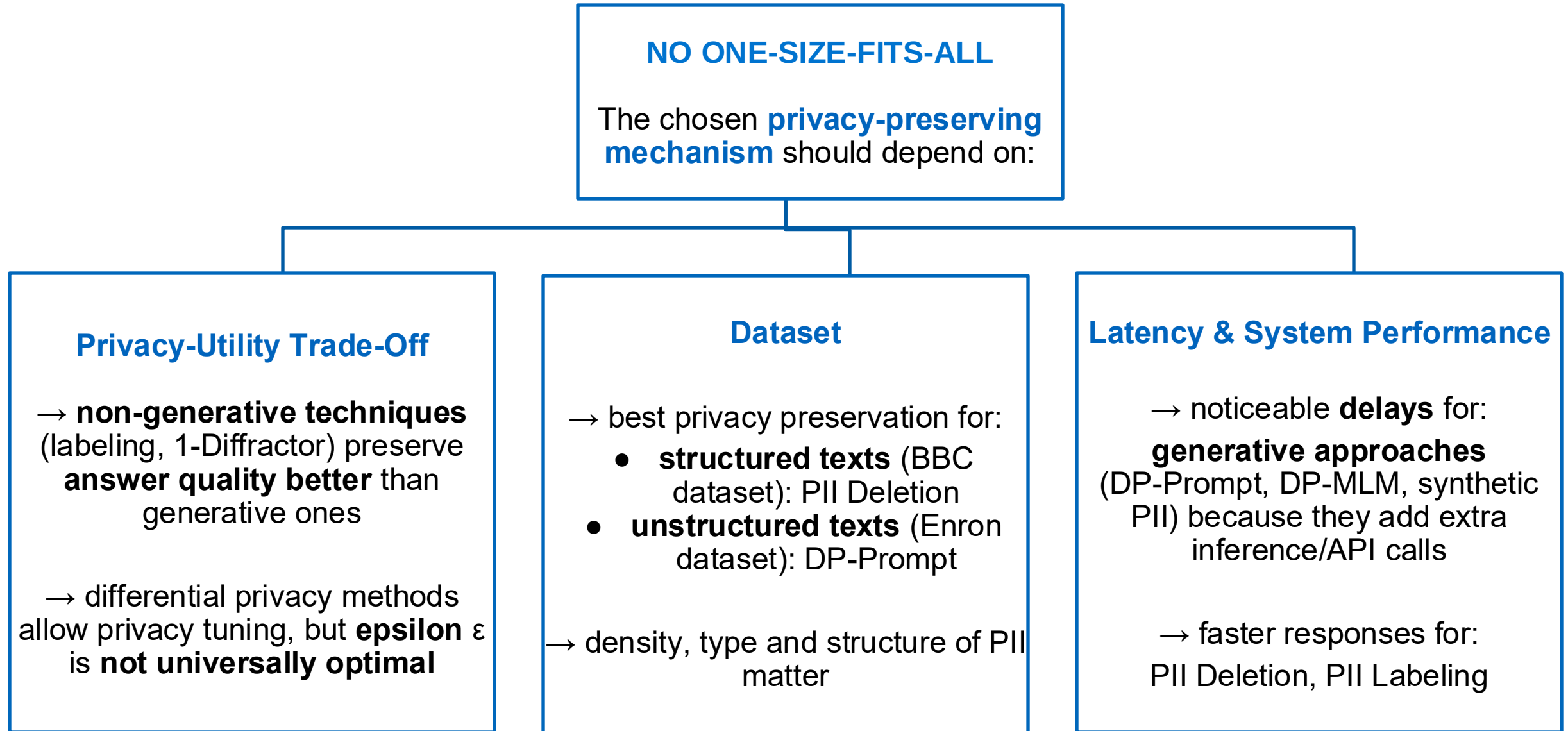
## Results:

- method resulting in **most leakage: Diffraction** (for both BBC and Enron)
- **most privacy-preserving** method:
  - BBC: **PII Deletion**
  - Enron: **DP-Prompt**

## 4.6.2. Privacy Results

Dataset Method	BBC News LLM Score	Enron Emails LLM Score
PII Deletion	<b>6</b>	35
PII Labeling	35	52
PII Synthetic data	12	32
Diffraction ( $\epsilon=1$ )	38	58
Diffraction ( $\epsilon=2$ )	43	69
Diffraction ( $\epsilon=3$ )	<b>45</b>	<b>74</b>
DP-Prompt ( $\epsilon=150$ )	11	<b>9</b>
DP-Prompt ( $\epsilon=200$ )	22	18
DP-Prompt ( $\epsilon=250$ )	23	26
DP-MLM ( $\epsilon=50$ )	24	34
DP-MLM ( $\epsilon=75$ )	29	36
DP-MLM ( $\epsilon=100$ )	28	36

Table 4.6.: Privacy results using an LLM-as-a-Judge approach to assess the privacy leakage. 0 represents no leakage, while 100 means that all the private data in the response is leaked from the original, non-anonymized dataset, even though the response is based on the anonymized version of the text. For each dataset, the worst (highest) score, denoting the most leakage, and the best (lowest) score, representing the most private response, are **bolded**.





## GuardRAG

Protecting private data in retrieval-augmented generation systems.

Select dataset:

ENRON

Select file:

Enron\_91

Select anonymization type:

PII Labeling

 Original text for **Enron\_91**

Catherine, I have received the call and will be traveling to Kazakhstan next week to complete the adoption. My husband and son are staying back. I arrive in Almaty on the 14th late in the evening. I'll spend the day of the 15th in Almaty and then take the overnight train to Taraz. I expect to have a court date in Taraz on the 19th or 20th. Then I will be back in Almaty, probably over Christmas. Finally, I expect to travel to Moscow on around Dec 27 and then return to Houston around Dec 30. This of course is the plan if everything goes exactly on schedule -- which it probably won't. I will certainly miss being away from home at Christmas and am hoping to find a Christmas worship service in Almaty. I'd appreciate any advice in this regard. I would also appreciate advice on what time of clothes are appropriate to bring. Thanks Shelley Corman shelley.corman@enron.com and

 Text after PII Labeling for **Enron\_91**

<PERSON>, I have received the call and will be traveling to <LOCATION> <DATE\_TIME> to complete the adoption. My husband and son are staying back. I arrive in <LOCATION> on <DATE\_TIME> evening. I'll spend <DATE\_TIME> in <LOCATION> and then take the overnight train to <PERSON>. I expect to have a court date in <LOCATION> on <DATE\_TIME>. Then I will be back in <LOCATION>, probably <DATE\_TIME>. Finally, I expect to travel to <LOCATION> on <DATE\_TIME> and then return to <LOCATION> around <DATE\_TIME>. This of course is the plan if everything goes exactly on schedule -- which it probably won't. I will certainly miss being away from home at <DATE\_TIME> and am hoping to find a <DATE\_TIME> worship service in <LOCATION>. I'd appreciate any advice in this regard. I would also appreciate advice on what time of clothes are appropriate to bring. Thanks <PERSON> <EMAIL\_ADDRESS> and

Personally Identifiable Information (PII) detected in the original text

Catherine PERSON, I have received the call and will be traveling to Kazakhstan LOCATION next week DATE\_TIME to complete the adoption. My husband and son are staying back. I arrive in Almaty LOCATION on the 14th late in the DATE\_TIME evening. I'll spend the day of the 15th DATE\_TIME in Almaty LOCATION and then take the overnight train to Taraz PERSON. I expect to have a court date in Taraz LOCATION on the 19th or 20th DATE\_TIME. Then I will be back in Almaty LOCATION, probably over Christmas DATE\_TIME. Finally, I expect to travel to Moscow LOCATION on around Dec 27 DATE\_TIME and then return to Houston LOCATION around Dec 30 DATE\_TIME. This of course is the plan if everything goes exactly on schedule -- which it probably won't. I will certainly miss being away from home at Christmas DATE\_TIME and am hoping to find a Christmas DATE\_TIME worship service in Almaty LOCATION. I'd appreciate any advice in this regard. I would also appreciate advice on what time of clothes are appropriate to bring. Thanks Shelley Corman PERSON shelley.corman@enron.com EMAIL\_ADDRESS and

Detailed Findings



## Prompt for utility experiments

Please generate a concise and factual summary of the text.

The summary should be structured into the following three key attributes:

[Attribute 1: TOPIC/CENTRAL IDEA]: main topic or central idea of the provided text

[Attribute 2: MAIN DETAILS ABOUT TOPIC/CENTRAL IDEA]: key events, discussion points, or details that support the central idea. For example, in a sports report, include scores and key plays; in a conversational email, note discussion topics and any actions mentioned.

[Attribute 3: IMPORTANT FACTS/EVENTS]: critical facts, events, data, or viewpoints that are essential to understanding the text

Please format your response as follows:

[Attribute 1: TOPIC/CENTRAL IDEA]:

[Attribute 2: MAIN DETAILS ABOUT TOPIC/CENTRAL IDEA]:

[Attribute 3: IMPORTANT FACTS/EVENTS]:

Ensure that your summary:

- Is concise and uses clear, simple language.
- Remains factual and unbiased, without including personal opinions.
- Maintains a logical order and comprehensively covers the information provided in the text.

## Response based on the original text

[Attribute 1: TOPIC/CENTRAL IDEA]: Travel plans for adoption in Kazakhstan.

[Attribute 2: MAIN DETAILS ABOUT TOPIC/CENTRAL IDEA]: The individual will travel to Kazakhstan for adoption, arriving in Almaty on the evening of the 14th. They will spend the 15th in Almaty before taking an overnight train to Taraz, where a court date is expected on the 19th or 20th. They plan to return to Almaty over Christmas and travel to Moscow around December 27, returning to Houston by December 30.

[Attribute 3: IMPORTANT FACTS/EVENTS]: The traveler will be away from home during Christmas and is seeking advice on finding a Christmas worship service in Almaty and what clothing to bring.

## Response based on the text after PII Labeling

[Attribute 1: TOPIC/CENTRAL IDEA]: Travel plans for adoption process.

[Attribute 2: MAIN DETAILS ABOUT TOPIC/CENTRAL IDEA]: The individual will travel to a specific location on a designated date to complete an adoption, while their husband and son will remain at home. They plan to arrive in another location in the evening, spend a day there, and then take an overnight train to a different person. A court date is expected in another location, followed by a return trip.

[Attribute 3: IMPORTANT FACTS/EVENTS]: The travel schedule includes multiple locations and dates, with a focus on the adoption process. The individual expresses a desire to attend a worship service during their travels and seeks advice on appropriate clothing for the trip.

## Prompt for privacy experiments

Please analyze the following text for any private or sensitive information, like personally identifiable information (PII). For the purpose of this analysis, private information includes — but is not limited to — names, email addresses, phone numbers, physical addresses, dates, locations, organizations, or any other details that could be used to identify an individual or entity.

## Response based on the original text

2. **Email Address**: The email address shelly.corman@enron.com is provided, which can be used to contact the individual.

3. **Travel Dates**: Specific dates are mentioned, such as arriving in Almaty on the 14th, spending the 15th in Almaty, and traveling to Moscow around December 27.

4. **Locations**: The text references specific locations, including Almaty, Taraz, and Moscow, which could be used to track the individual's movements.

5. **Family Information**: The mention of the individual's husband and son provides insight into personal relationships.

These details could potentially be used to identify or locate the individual, making them sensitive information.

## Response based on the text after PII Labeling

The text contains several instances of personally identifiable information (PII). This includes:

1. Names: The text references individuals using placeholders for names.

2. Email Address: There is a placeholder for an email address.

3. Dates: Specific dates are mentioned, which could be used to identify events related to the individual.

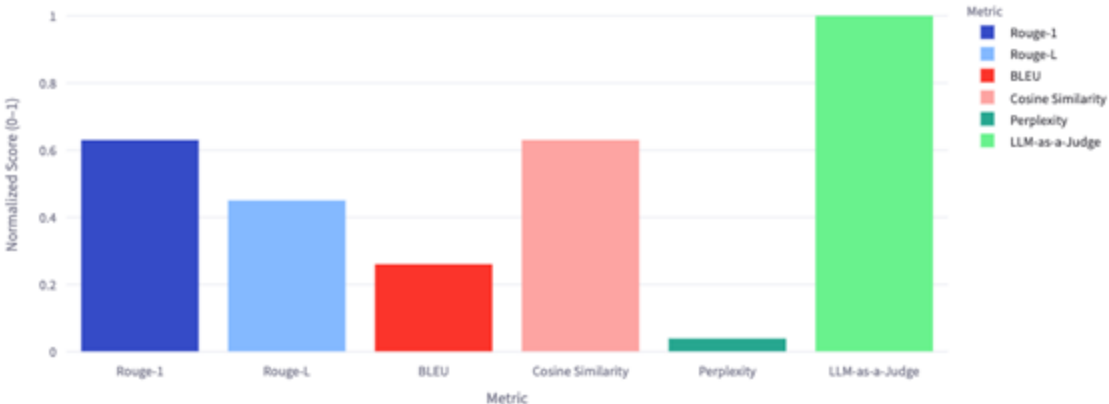
4. Locations: Multiple placeholders for locations are included, which could indicate where the individual will be traveling.

5. Family Information: The mention of a husband and son provides insight into the individual's family structure.

## Evaluation of the response based on the text after PII Labeling

	Metric	Score	Explanation	Normalized
0	Rouge-1	0.63	Overlap of unigrams (recall-focused)	0.63
1	Rouge-L	0.45	Longest common subsequence (sequence similarity)	0.45
2	BLEU	0.26	N-gram precision of generated vs reference	0.26
3	Cosine Similarity	0.63	Semantic closeness of embeddings	0.63
4	Perplexity	25.87	How predictable the text is (lower = better)	0.0387
5	LLM-as-a-Judge	0	LLM-based judgment on percentage of privacy leakage	1

## Barchart for Normalized Evaluation Metrics



## GuardRAG LIVE

Protecting private data in retrieval-augmented generation systems.

!! About GuardRAG

### ANONYMISATION: upload a file or type in a text and choose the anonymisation method

Upload a file:



Drag and drop file here

Limit 200MB per file • PDF, TXT

Browse files

Or type your text here:

Send

Select text language:

☒ English

☐ German

Select privacy-preserving method: 

PII Deletion 

Epsilon value 

0

0

300

### RAG: ask a question about the text and compare the response based on the original text with the one based on the anonymized text

Enter your question:

Get Answer

## Future Work

### Privacy in RAG systems

#### RAG PRIVACY PROCESS DIAGRAM:

- advanced RAG pipelines
- different domains (healthcare, legal, finance)

#### LEAKAGE and ATTACKS:

- Do privacy mitigations interact with adversarial resilience?
- e.g., does DP or anonymisation also blunt jailbreak & poisoning attacks

#### RAG EVALUATION METRICS:

- test **differential privacy** beyond ingestion (e.g., noise in embeddings)
- compare dataset-level vs retrieval-level vs answer-level defences
- design metrics tailored to RAG (beyond “LLM-as-a-Judge”)

#### HUMAN-CENTERED STUDIES:

- measure user perception of privacy risk and acceptable utility loss
- Do explanation techniques (e.g., source citation) modulate trust and disclosure behaviour?

#### GuardRAG DEMO:

- extend beyond dataset leakage, to retrieval-stage and generation-stage issues
- add new datasets, utility / privacy metrics, and prompt templates





## **Andreea-Elena Bodea**

[andreea.bodea@tum.de](mailto:andreea.bodea@tum.de)

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](mailto:matthes@in.tum.de)  
[www.matthes.in.tum.de](http://www.matthes.in.tum.de)

