University of Udine
March 25th, 2025

UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

Optimizing Privacy and Retrieval: Document and Query Sanitization Strategies

Marco Viviani

marco.viviani@unimib.it

University of Milano-Bicocca
Department of Informatics, Systems, and Communication, IKR3 Lab

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

Ikr3 Lab
information and knowledge representation, retrieval, and reasoning LAB

UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

Motivation and Context

Challenges

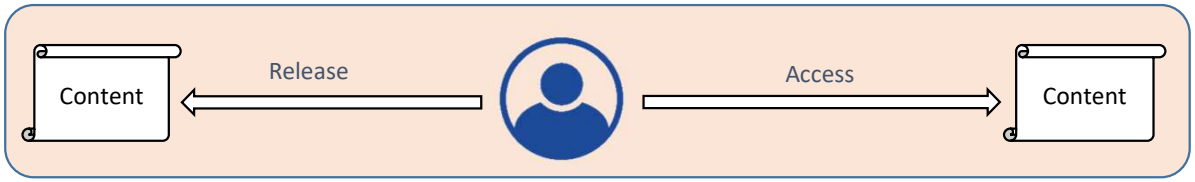
- Users may face **various risks** in releasing and accessing content (structured, semi-structured, unstructured) in online environments.

- **Content release**: Uncontrolled release of personal/sensitive data (privacy).
  - How to protect privacy?
  - How to avoid microtargeting?

- **Content access**: Access to "incomplete"/fake information.
  - How to identify the utility of information protected from a privacy perspective?
  - How to avoid misinformation access?

Putting the User at the Center

- In the **trade-off** between releasing personal/sensitive data and accessing useful/reliable information, **users must play a central role**.

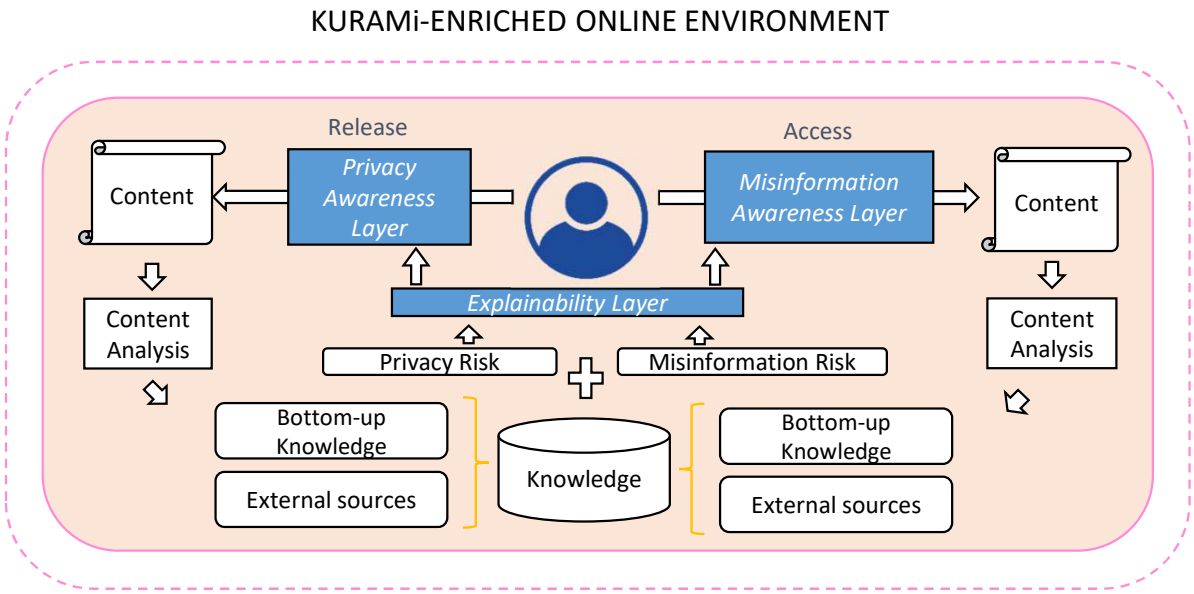


  - Provide users with automated and effective approaches promoting user autonomy.
  - Easily interpretable results without the decision-making process being left only to algorithms.

The KURAMi Project

- **KURAMi**: Knowledge-based, explainable User empowerment in Releasing private data and Assessing Misinformation in online environments.

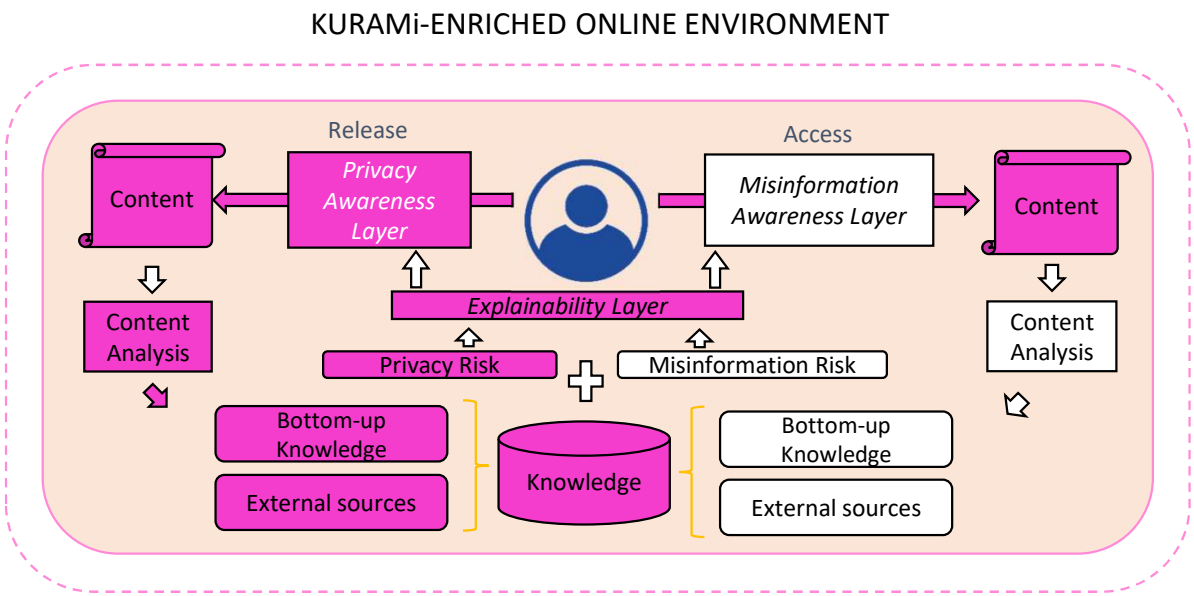
- PRIN 2022: Research project funded by the Italian EU - Next Generation EU, Mission 4, Component 2, CUP D53D23008480001 and Italian MUR.



KURAMi and Privacy: Some Tasks

- Various tasks are involved in KURAMi.

- In today's **seminar**:
  - Privacy risk assessment → Highlights.
  - Document sanitization → Data Marketplaces.
  - Query sanitization → Generative IR.



UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

Privacy Risk Assessment

Some Highlights

Privacy and Structured Data

- Assessing privacy risk in **structured data** (E.g., relational databases, spreadsheets, …) typically involves several approaches, each with distinct techniques to evaluate risks of re-identification or private/sensitive information disclosure.
  - *k*-anonymity, *l*-diversity, *t*-closeness: Approaches to reduce re-identification risk or sensitive information disclosure by ensuring indistinguishability within groups.
    - Linkage attacks: Assessing risks from linking with external, not de-identified data sources.
  - Differential Privacy (DP): Typically, adding "noise" to data to protect aggregated outputs.

A Concrete Example (1)

- Obtaining 3-anonymity and 3-diversity.

SSN	Name	DoB	Sex	ZIP	Disease
		64/09/27	M	94139	Chest pain
		63/09/30	F	94139	Broken arm
		64/04/18	M	94139	Gastritis
		63/04/15	F	94139	Ulcera
		63/03/13	F	94138	Short breath
		64/09/15	*M*	*94142*	*Stomach cancer*
		64/09/13	M	94141	Broken leg

(a) De-identified medical data

DoB	Sex	ZIP	Disease
1964	M	941**	Chest pain
1964	M	941**	Gastritis
1964	M	941**	Broken leg
1963	F	941**	Broken arm
1963	F	941**	Ulcera
1963	F	941**	Short breath

3-anonymous and 3-diverse table

Name	Address	City	ZIP	BirthDate	Sex	Education
...	...	...	...	...	...	...
John Doe	250 Market St.	San Francisco	94142	64/09/15	male	secondary
...	...	...	...	...	...	...

(b) Municipality register

A Concrete Example (2)

- An interactive privacy mechanism achieving **ε-differential privacy**.
  - The epsilon (ε) parameter quantifies the privacy-utility trade-off, with smaller values indicating stronger privacy protection.

- The mechanism works by **adding appropriately chosen random noise** to the answer $a = f(X)$, where f is the query function and X is the database.

- E.g., *query*: "*Compute the median of each column*".

  a: Noisy versions of the medians.

Privacy and Unstructured Data

- Assessing privacy risk in **unstructured data** (e.g., images, audio, videos, text documents such as emails, and social media posts) is more complex due to the lack of predefined structure, like rows and columns, and the diversity of potentially sensitive information.
  - Image and video analysis: Identifying faces or sensitive objects using computer vision.
  - Audio analysis: Identifying voice biometrics, contextual clues, or environmental sounds, as well as inadvertent leakage of personal, location, or behavioral information.
  - Text analysis: Identifying sensitive entities in text also based on semantic context and auxiliary data.

Text Analysis and Privacy

- **Step 1: Identifying sensitive entities** using Named Entity Recognition (NER) and other NLP methods.
  - Common entities include: personal data like names, addresses, emails, phone numbers, etc., including sensitive data like health conditions, political or religious affiliations, financial information, and other metadata like timestamps, geolocation, etc.

- **Step 2: Assigning risk scores** to entities based on:
  - Entity sensitivity: Certain entities (e.g., health conditions) are inherently more sensitive than others.
  - Uniqueness: Evaluates how rare and identifiable an entity is.
  - Exposure: The probability of exposure due to attacks or misuse.

- **Step 3: Aggregating risk scores** by aggregating the risk scores of individual entities, often using weighted sums, averages, or maximum-based aggregation.

UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

Document Sanitization

Cassani, L., Livraga, G., & Viviani, M. (2024, September). Assessing document sanitization for controlled information release and retrieval in data marketplaces. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (CLEF 2024) (pp. 88-99). Cham: Springer Nature Switzerland.

The Context: Data Marketplaces

- **Data Marketplaces** (DMs) are specialized virtual spaces that allow the exchange of various kinds of data that can range from highly specific and niche data to more general and broadly applicable information.
  - **Data owners** offer them for a fee on a DM.
  - **Registered users** can explore the platform to retrieve the data they need and, should they find data of interest, proceed with the purchase.
  - **DMs generate revenue** usually through commissions from processed transactions.

- In **marketplaces** for **physical items**, products can be presented with accurate descriptions and photographs and are subject to return and warranty policies.

Open Issues and Possible Solutions

- **Digital information** presents different characteristics by its nature.
  - **Data stored** within DM platforms must be protected so that they are only visible to users who have purchased them.
  - These platforms must also equip **potential buyers** with the tools needed to determine whether the data they find are indeed useful for them, without exposing the entire content before the sale is concluded.

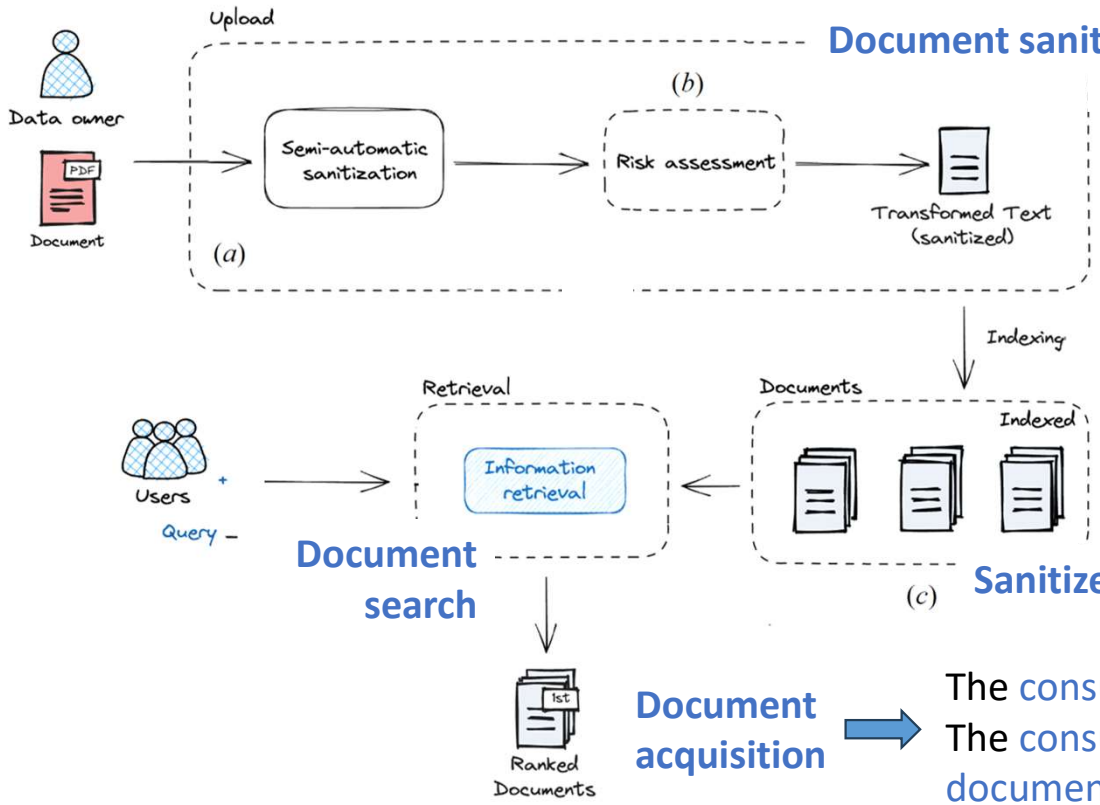
- Modern DMs also include **unstructured data**.
  - The objective of providing an accurate description remains the same.
  - Need for tailored strategies (*blurring* for images, *key frames* for videos).
  - What about textual documents? → **Text sanitization**.

Text Sanitization

- **ALERT**: The sanitized text should:
  - Protect the content not meant for disclosure.
  - Be sufficiently representative of the original text → Sufficiently match the buyer's information needs.

- A **twofold objective**:
  - **Various sanitization techniques** applied to textual documents within the DM context → Masking and/or summarization.
  - **Assessing retrieval effectiveness** of sanitized documents to verify that data sanitization, while concealing confidential content, compromises neither retrieval effectiveness nor data saleability.

The Proposed Architecture



Document sanitization *in-house*

Original documents can:

- Either be encrypted and sent to the platform (*honest-but-curious*).
- Or kept securely *in-house* by the owner.

Document search

Sanitized documents

Document acquisition

The consumer selects a subset of documents of interest.
The consumer purchases the original versions of these documents.

Document Masking and Summarization

- **Document masking**: selectively masking parts of the document (in terms of tokens) as deemed necessary by the owner.
  - E.g., if the owner does not wish to include the word ‘cat’ in a masked document, the original document can be sanitized by masking all occurrences of the token ‘cat’.
  - We can apply or not **Coreference Resolution (CR)** → Next slide.
- **Document summarization**: generating a summary of a document.
  - Keeping just the most important sentences in the summary, i.e., **extractive summarization**.
  - Rephrasing the original documents in a shorter version, i.e., **abstractive summarization**.
  - Extractive summarization preserves the original document’s representativeness by including original sentences in the summarized document.

Coreference Resolution

- **Coreference Resolution (CR)** is the task of finding all linguistic expressions (called mentions) in a given text that refer to the same real-world entity.

- ***The mouse*** *and* ***the elephant*** *are two animals, belonging to the class of mammals.* ***The former*** *has an average weight of 20 g, while* ***the latter*** *can weigh up to 6,000 kg. In addition,* ***the latter****, unlike* ***the former****, has a proboscis.*

CR

- ***The mouse*** *and* ***the elephant*** *are two animals, belonging to the class of mammals.* ***The mouse*** *has an average weight of 20 g, while* ***the elephant*** *can weigh up to 6,000 kg. In addition,* ***the elephant****, unlike* ***the mouse****, has a proboscis.*

Document Masking and CR

- **[MASK]** *and* **[MASK]** *are two animals, belonging to the class of mammals.* ***The former*** *has an average weight of 20 g, while* ***the latter*** *can weigh up to 6,000 kg. In addition,* ***the latter****, unlike* ***the former****, has a proboscis.*

The latter →

- **[MASK]** *and* **[MASK]** *are two animals, belonging to the class of mammals.* **[MASK]** *has an average weight of 20 g, while* **[MASK]** *can weigh up to 6,000 kg. In addition,* **[MASK]***, unlike* **[MASK]***, has a proboscis.*

(?)

Confidentiality Risk Assessment

- We intend the **confidentiality risk** as the possibility of demasking tokens that have been obfuscated by the data owner.

- Assessed by means of a **demasking resistance measure**:

$$dr(d) = 1 - \frac{n_{inf}}{n_{max}}$$

  - n_{inf}: the number of inferred tokens from the sanitized document.
  - n_{max}: the total number of obfuscated tokens in the sanitized document.

The «Online News» Scenario

- Showing only the title or the first portion of an article may not be the best choice for a customer interested in purchasing the article itself.
- **Data**: a subset of the articles from the Washington Post collected as part of TREC.
  - The collection includes 595,037 articles, stored in a JSON Lines format file, collected around 50 different topics.
  - A qrels.txt file is also provided for performance evaluation in IR.
  - Only documents with a length of less than 512 tokens (a limit imposed by BERT) were considered for evaluation → 3,776 articles.

Implementing the Solution

- **Summarization techniques**:
  - Luhn,
  - KLSummarizer,
  - Latent Semantic Analysis (LSA),
  - LexRank,
  - SBertSummarizer.
- **Masking assumption**: tokens to be obfuscated are entities in the original documents.
  - Those extracted by means of Named-Entity Recognition (NER).
- **Demasking**: performed using the DistilRoBERTa model.
  - LLMs can be employed to infer masked tokens.
- **Retrieval models**:
  - TF-IDF, BM25, DLH, DPH, InL2, MDL2.
- **Metrics**:
  - Average demasking resistance → $adr(D)$,
  - Mean Average Precision → MAP,
  - Normalized Discounted Cumulative Gain → nDCG.

Some Results: Masking Alone

Table 1. Evaluation metrics considering document masking for sanitization.

Model	MAP_{bl}	$nDCG_{bl}$	Masked	Demasked	$adr(D)$	MAP	nDCG
TF-IDF	0.234	0.411	41,816	8,867	0.788	0.211	0.386
BM25	0.234	0.411	41,816	8,867	0.788	0.212	0.386
DLH	0.226	0.403	41,816	8,867	0.788	0.204	0.38
DPH	**0.249**	**0.423**	41,816	8,867	0.788	**0.220**	**0.393**
InL2	0.238	0.413	41,816	8,867	0.788	0.216	0.389
MDL2	0.201	0.373	41,816	8,867	0.788	0.183	0.357

Some Results: Summarization + Masking

Table 2. Evaluation metrics considering summarization and masking for sanitization (average over 8 summary lengths; IR model: DPH).

Model	$adr(D)_{as}$	MAP_{as}	nDCG_{as}
Luhn	0.884	0.194	0.357
KLSummarizer	**0.918**	0.174	0.339
Latent Semantic Analysis (LSA)	0.885	**0.205**	**0.374**
LexRank	0.891	0.195	0.357
TextRank	0.877	0.204	0.371
SBertSummarizer	0.899	0.184	0.351

Average scores obtained over distinct document summary lengths (i.e., 10%, 20%, …, 80%) for the considered evaluation metrics, denoted in this case as $adr(D)_{as}$, MAP_{as}, and nDCG_{as}.

Some Results: Summarization + Masking + CR + Query Expansion

Table 3. Evaluation metrics considering summarization and masking for sanitization, CR, and QE (average over 8 summary lengths; IR model: DPH; summarizer: LSA).

CR	QE	$adr(D)_{as}$	MAP_{as}	nDCG_{as}
No	No	0.885	0.205	0.374
Yes	No	**0.892**	0.199	0.366
No	KLQE	0.885	**0.222**	**0.416**
Yes	KLQE	**0.892**	0.214	0.405

Some Takeaways

- Simple **token masking alone** is less effective at mitigating the risk of demasking compared to the **combination** of token masking with text summarization.
- While improving confidentiality, this approach **negatively impacts retrieval effectiveness**.
  - A balanced approach can be achieved by incorporating Coreference Resolution during the masking process and employing Query Expansion during retrieval.
- **Further research** (some ideas):
  - More sophisticated summarization algorithms that inherently incorporate data confidentiality principles could be developed.
  - Conducting comprehensive testing across various marketplace scenarios and datasets could validate the applicability and resilience of our approach.

UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

Query Sanitization

Herranz-Celotti, L., Guembe, B., Livraga, G., & Viviani, M. (2025, April). Can Generative AI Adequately Protect Queries? Analyzing the Trade-off Between Privacy Awareness and Retrieval Effectiveness. In *Proceedings of the 47th European Conference on Information Retrieval* (ECIR 2025), Lucca, Italy, 2025

The Context: Generative IR

- Users increasingly interact with generative AI tools and Information Retrieval Systems → **Generative Information Retrieval**.
- Queries have **longer** and more **elaborate prompts** → Likely containing more detailed and personal/sensitive information.
- Useful for users, yet **risky for privacy** → Several approaches for query protection, both security/privacy and IR research fields.

Background

- **Query protection** can entail protecting:
  - User identity (break link between identity and query).
  - User intent (blur query content).
    - Focus on query intent protection → Queries are modified before being sent to the IRS.

- **Protection** is typically **enforced on the user side** (no intervention from not fully trusted IRSs).
  - Query modification:
    - Generalization: lexical resources (e.g., WordNet) or heuristics to remove private/sensitive intent details.
    - Noise addition (e.g., DP-based text generation).
  - Query multiplication: confuse intent with dummy queries.

The Investigated Open Issues

- **Question 1**: Can generative LLMs mimic query protection approaches?
- **Question 2**: Can generative LLMs offer a viable solution for balancing user privacy with retrieval effectiveness?
- **Approach**: Study the feasibility of applying query protection through generative LLMs.
  - Six different prompts.
  - Inspired by SOTA confusion-based methods (query modification and multiplication).

Prompt Engineering

- Let us consider the following **Original Query (OQ)**.

- Let us consider the following **Preamble (P)**.

OQ. *gay clubs in New York*

PB. *You are a privacy expert tasked with protecting the user's privacy when performing the following task.*

Prompt 1 (P1): Simple Rewriting

OQ. *gay clubs in New York*

PB. *You are a privacy expert tasked with protecting the user's privacy when performing the following task.*

P1. [PB] *Rewrite this query*: [OQ]

E1. *Social venues in New York for LGBTQ+ community*

Prompt 2 (P2): Generalization

OQ. *gay clubs in New York*

PB. *You are a privacy expert tasked with protecting the user's privacy when performing the following task.*

P2. [PB] *Rewrite this query by applying generalization*: [OQ]

E2. *Entertainment venues for diverse communities in urban areas*

Prompt 3 (P3): Differential Privacy*

OQ. *gay clubs in New York*

PB. *You are a privacy expert tasked with protecting the user's privacy when performing the following task.*

P3. [PB] *Rewrite this query by applying Differential Privacy*: [OQ]

E3. *Gay clubs in New York, Los Angeles, and Chicago*

*With a caveat

Prompt 4 (P4): Dummy Queries

OQ. *gay clubs in New York*

PB. *You are a privacy expert tasked with protecting the user's privacy when performing the following task.*

P4. [PB] *Generate* [k] *dummy, random queries, given this query:* [OQ]

E4 ($k = 3$)**.** *Art galleries to visit in urban settings | Cultural festivals happening in the summer | Best coffee shops with outdoor seating*

Prompt 5 (P5): Dummy Queries + Semantics

OQ. *gay clubs in New York*

PB. *You are a privacy expert tasked with protecting the user's privacy when performing the following task.*

P5. [PB] *Generate* [k] *dummy queries, which are semantically related to this query*: [OQ]

E5 ($k = 3$). *LGBTQ+ events happening in New York City | Nightlife options for the LGBTQ+ community in urban areas | Social gatherings for LGBTQ+ individuals in major cities*

Prompt 6 (P6): Dummy Queries + Generalization*

OQ. *gay clubs in New York*

PB. *You are a privacy expert tasked with protecting the user's privacy when performing the following task.*

P6. [PB] *Generate* [k] *dummy queries, which generalize this query*: [OQ]

E6 ($k = 3$). *LGBTQ+ nightlife options in major cities | Social venues for diverse communities in urban areas | Inclusive entertainment spots in metropolitan regions*

*With a caveat

Implementing the Solution

- **Goal**: compare prompt-driven LLM methods with SOTA baselines.
  - Lexicon-based (WordNet)
  - Differential Privacy-based
- Different **retrieval models**:
  - Sparse (BM25).
  - Dense (MonoT5).
- **Datasets**:
  - NFCorpus (medical IR).
  - TREC-COVID (pandemic-related research).
  - Touché (controversial topics).
- **Metrics**:
  - Retrieval effectiveness (MAP, nDCG).
  - Query syntactic (Jaccard index) and semantic (cosine similarity among BERT embeddings) similarity.

Some Results for Sparse Retrieval

	NFCorpus					TREC-Covid					*Touché*				
QM	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**CS$_B$↓**	**JI↓**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**CS$_B$↓**	**JI↓**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**CS$_B$↓**	**JI↓**
NONE	0.149	0.322	0.273	-	-	**0.198**	**0.626**	**0.474**	-	-	0.225	**0.343**	**0.455**	-	-
WordNet	0.057	0.120	0.114	0.687	0.201	0.033	0.123	0.111	0.615	0.209	0.019	0.027	0.065	0.643	0.153
DP CMP$_1$	0.000	0.002	0.001	0.416	**0.000**	0.000	0.000	0.000	0.347	**0.000**	0.000	0.000	0.000	0.269	**0.000**
DP CMP$_5$	0.001	0.003	0.004	0.430	0.005	0.000	0.000	0.000	0.366	**0.000**	0.000	0.000	0.000	0.277	**0.000**
DP CMP$_{10}$	0.035	0.075	0.075	0.563	0.166	0.011	0.025	0.027	0.448	0.067	0.024	0.033	0.065	0.426	0.119
DP CMP$_{50}$	0.149	0.322	0.273	1.000	0.999	0.182	0.573	0.438	0.984	0.784	0.225	**0.343**	**0.455**	1.000	1.000
DP M$_1$	0.000	0.001	0.002	0.398	**0.000**	0.000	0.000	0.001	0.352	**0.000**	0.000	0.000	0.000	0.274	**0.000**
DP M$_5$	0.001	0.002	0.003	0.411	0.002	0.000	0.002	0.001	0.366	**0.000**	0.000	0.002	0.002	0.274	0.002
DP M$_{10}$	0.019	0.047	0.048	0.497	0.106	0.012	0.025	0.035	0.420	0.035	0.008	0.004	0.032	0.387	0.069
DP M$_{50}$	0.149	0.322	0.273	1.000	0.999	0.183	0.576	0.440	0.985	0.784	0.225	**0.343**	**0.455**	1.000	1.000
DP V$_1$	0.000	0.001	0.001	0.404	**0.000**	0.000	0.000	0.000	**0.346**	**0.000**	0.000	0.000	0.000	0.272	0.003
DP V$_5$	0.002	0.004	0.007	0.412	0.002	0.000	0.001	0.002	0.348	0.002	0.002	0.003	0.005	0.289	0.002
DP V$_{10}$	0.017	0.044	0.046	0.490	0.067	0.020	0.014	0.043	0.412	0.046	0.018	0.020	0.049	0.373	0.051
DP V$_{50}$	0.094	0.209	0.190	0.814	0.466	0.098	0.321	0.276	0.761	0.350	0.131	0.206	0.301	0.791	0.471
GPT$_{P1}$	0.092	0.202	0.188	0.696	0.291	0.098	0.337	0.265	0.893	0.424	0.145	0.240	0.344	0.879	0.427
GPT$_{P2}$	0.049	0.114	0.129	0.683	0.127	0.047	0.205	0.156	0.794	0.316	0.044	0.068	0.138	0.769	0.198
GPT$_{P3}$	0.093	0.194	0.188	0.597	0.196	0.078	0.320	0.223	0.747	0.289	0.110	0.174	0.282	0.800	0.296
GPT$_{P4}^{k=1}$	0.122	0.250	0.240	0.571	0.371	0.152	0.402	0.358	0.693	0.541	0.179	0.251	0.368	0.724	0.480
GPT$_{P4}^{k=3}$	0.146	0.299	0.271	0.548	0.205	0.121	0.333	0.291	0.677	0.304	0.191	0.283	0.394	0.731	0.253
GPT$_{P4}^{k=5}$	0.158	0.327	0.288	0.536	0.146	0.170	0.425	0.365	0.757	0.261	**0.226**	0.325	0.452	0.766	0.193
GPT$_{P5}^{k=1}$	0.144	0.307	0.274	0.759	0.424	0.167	0.519	0.406	0.871	0.566	0.200	0.300	0.419	0.890	0.516
GPT$_{P5}^{k=3}$	0.157	0.319	0.290	0.625	0.196	0.166	0.476	0.391	0.786	0.299	0.220	0.328	0.444	0.799	0.242
GPT$_{P5}^{k=5}$	**0.163**	**0.333**	**0.299**	0.610	0.155	0.173	0.473	0.397	0.781	0.230	0.225	0.339	0.452	0.774	0.183
GPT$_{P6}^{k=1}$	0.148	0.305	0.280	0.798	0.451	0.156	0.547	0.391	0.865	0.588	0.185	0.292	0.402	0.894	0.522
GPT$_{P6}^{k=3}$	0.143	0.293	0.272	0.682	0.214	0.131	0.411	0.338	0.796	0.286	0.205	0.316	0.430	0.817	0.247
GPT$_{P6}^{k=5}$	0.141	0.291	0.272	0.651	0.167	0.148	0.416	0.353	0.793	0.239	0.206	0.328	0.435	0.788	0.180

The best results are in **bold**. The second-best results are underlined.

Some Results for Dense Retrieval

	NFCorpus					TREC-Covid					*Touché*				
QM	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**CS$_B$↓**	**JI↓**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**CS$_B$↓**	**JI↓**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**CS$_B$↓**	**JI↓**
NONE	**0.156**	**0.346**	0.286	-	-	**0.088**	**0.709**	**0.492**	-	-	**0.250**	**0.392**	**0.489**	-	-
WordNet	0.060	0.137	0.122	0.687	0.201	0.016	0.222	0.132	0.615	0.209	0.018	0.042	0.076	0.643	0.153
DP CMP$_1$	0.000	0.001	0.001	0.416	**0.000**	0.000	0.000	0.000	0.347	**0.000**	0.000	0.000	0.000	0.269	**0.000**
DP CMP$_5$	0.002	0.005	0.005	0.430	0.005	0.000	0.000	0.000	0.366	**0.000**	0.000	0.000	0.000	0.277	**0.000**
DP CMP$_{10}$	0.033	0.077	0.076	0.563	0.166	0.003	0.056	0.033	0.448	0.067	0.016	0.016	0.060	0.426	0.119
DP CMP$_{50}$	**0.156**	**0.346**	0.286	1.000	0.999	0.079	0.680	0.458	0.984	0.784	**0.250**	**0.392**	**0.489**	1.000	1.000
DP M$_1$	0.000	0.001	0.002	0.398	**0.000**	0.000	0.001	0.000	0.352	**0.000**	0.000	0.000	0.000	0.274	**0.000**
DP M$_5$	0.000	0.002	0.003	0.411	0.002	0.000	0.000	0.001	0.366	**0.000**	0.000	0.000	0.001	0.274	0.002
DP M$_{10}$	0.018	0.049	0.049	0.497	0.106	0.005	0.052	0.040	0.420	0.035	0.011	0.026	0.044	0.387	0.069
DP M$_{50}$	**0.156**	**0.346**	0.286	1.000	0.999	0.080	0.692	0.462	0.985	0.784	**0.250**	**0.392**	**0.489**	1.000	1.000
DP V$_1$	0.000	0.001	0.001	0.404	**0.000**	0.000	0.000	0.000	**0.346**	**0.000**	0.000	0.000	0.000	0.272	0.003
DP V$_5$	0.001	0.003	0.006	0.412	0.002	0.000	0.004	0.003	0.348	0.002	0.000	0.000	0.003	0.289	0.002
DP V$_{10}$	0.016	0.047	0.047	0.490	0.067	0.008	0.061	0.052	0.412	0.046	0.013	0.018	0.047	0.373	0.051
DP V$_{50}$	0.098	0.229	0.198	0.814	0.466	0.047	0.471	0.304	0.761	0.350	0.147	0.241	0.328	0.791	0.471
GPT$_{P1}$	0.111	0.258	0.217	0.696	0.291	0.051	0.608	0.315	0.893	0.424	0.162	0.277	0.371	0.879	0.427
GPT$_{P2}$	0.059	0.152	0.146	0.683	0.127	0.025	0.354	0.182	0.794	0.316	0.056	0.116	0.164	0.769	0.198
GPT$_{P3}$	0.101	0.240	0.207	0.597	0.196	0.036	0.426	0.244	0.747	0.289	0.131	0.240	0.321	0.800	0.296
GPT$_{P4}^{k=1}$	0.113	0.256	0.240	0.571	0.371	0.063	0.612	0.397	0.693	0.541	0.184	0.293	0.394	0.724	0.480
GPT$_{P4}^{k=3}$	0.093	0.217	0.224	0.548	0.205	0.046	0.430	0.310	0.677	0.304	0.161	0.255	0.379	0.731	0.253
GPT$_{P4}^{k=5}$	0.084	0.190	0.218	0.536	0.146	0.067	0.521	0.381	0.757	0.261	0.168	0.244	0.403	0.766	0.193
GPT$_{P5}^{k=1}$	0.131	0.312	0.274	0.759	0.424	0.076	0.659	0.435	0.871	0.566	0.207	0.324	0.443	0.890	0.516
GPT$_{P5}^{k=3}$	0.120	0.276	0.266	0.625	0.196	0.071	0.599	0.414	0.786	0.299	0.189	0.284	0.426	0.799	0.242
GPT$_{P5}^{k=5}$	0.107	0.252	0.256	0.610	0.155	0.075	0.619	0.425	0.781	0.230	0.168	0.268	0.408	0.774	0.183
GPT$_{P6}^{k=1}$	0.153	0.334	**0.294**	0.798	0.451	0.068	0.653	0.411	0.865	0.588	0.199	0.332	0.426	0.894	0.522
GPT$_{P6}^{k=3}$	0.133	0.299	0.275	0.682	0.214	0.057	0.567	0.366	0.796	0.286	0.166	0.278	0.412	0.817	0.247
GPT$_{P6}^{k=5}$	0.112	0.253	0.255	0.651	0.167	0.066	0.561	0.382	0.793	0.239	0.152	0.257	0.391	0.788	0.180

The best results are in **bold**. The second-best results are underlined.

Some Takeaways

- Lexicon-based SOTA achieves privacy protection in spite of retrieval effectiveness.
- DP methods achieve reasonable effectiveness with epsilon values too high (~50).
- LLM-based query multiplication seems to balance protection and retrieval effectiveness → "Query expansion" effect?
- LLM-based methods tend to perform better in sparse retrieval → To be investigated.
- By observing the queries generated, it seems that the LLM perceives the protection mechanism as a blurring of the query with more general terms (unless explicitly instructed otherwise).

Overall Takeaways

Challenges and Open Issues

- **Text-based sanitization techniques and limitations**: Sanitization techniques that modify or remove sensitive text can degrade the semantic structure. This reduces the quality and relevance of the information retrieved by IR systems.
- **Privacy-precision trade-off**: Advanced mechanisms such as Differential Privacy (DP) mechanisms often sacrifice too much precision to protect sensitive data. This leads to lower retrieval accuracy and less useful results.
- **Inadequate privacy risk assessment**: Current privacy risk measures, based on semantic similarity between sanitized and original text, are often insufficient. They may fail to detect deeper vulnerabilities or hidden risks.
- **Generative AI and inversion attacks**: Generative AI models can help design privacy solutions but are also prone to inversion attacks. These attacks may reconstruct private data from anonymized outputs using learned patterns.

University of Udine
March 25th, 2025

UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura

Thank you for your attention!

Grazie per la vostra attenzione!

Questions? *Domande?*

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

Ikr3 Lab
information and knowledge representation, retrieval, and reasoning LAB