Literature Review for MyPocketLawyer

Introduction

MyPocketLawyer is a chatbot that uses AI to help people understand legal topics and documents. It can answer legal questions using real laws and legal resources. Users can also upload their legal documents (like contracts, notices, agreements) and ask questions about them.The chatbot gives simplified answers that are easy to understand.

High-Level Problem and Sub-Problems

The central problem this project addresses is the lack of accessible and affordable legal aid in Nepal. Many people cannot afford legal help or have difficulty understanding complex legal documents and language. Legal aid centers are often overburdened, leading to long waiting times and unmet demand, especially in rural areas.

To tackle this, the project breaks down the problem into several sub-problems:

- **Accessible Q&A:** Answering legal questions in a simple, understandable way, using real laws and legal resources.
- **Document Understanding:** Allowing users to upload and ask questions about their own legal documents, such as contracts or notices.
- **Factual Accuracy:** Ensuring the AI's answers are factually correct and do not generate false information, which is a key risk with large language models (LLMs).
- **Multilingual Support:** Providing support for both English and Nepali translations, which is crucial for the target audience.

Key Concepts and Ideas

The project's core technology is a Retrieval-Augmented Generation (RAG) system. This method combines a retriever and a generator to produce answers. The retriever finds relevant information from a trusted knowledge base, and the generator (a large language model) synthesizes a coherent answer based on the retrieved information. This approach is used to ensure factual accuracy and reduce "hallucinations," a common issue with standalone LLMs. The project will also need to address multilingual support, as legal terminology and queries will be in both Nepali and English.

Datasets and Models

Recent research on retrieval-augmented generation (RAG) for legal applications emphasizes the importance of high-quality, curated datasets and effective retrievers and generators. Bhusal and Baral (2025) demonstrated that combining semantic and rule-based retrieval methods with GPT-4 improves QA performance for Nepali legal documents. Similarly, studies by Hindi et al. (2025) and Kabir et al. (2025) highlight the role of multilingual retrievers, such as sentence-transformers and dense vector models, in supporting domain-specific information retrieval. Sparse retrievers like BM25 remain a strong baseline for keyword matching, complementing semantic retrieval in hybrid approaches.

Data challenges are common across these studies. Legal documents often exist in scanned PDFs, requiring optical character recognition (OCR) for extraction, and public APIs for legal retrieval are limited. Moreover, inconsistent or outdated legal terminology, the need for bilingual support (e.g., Nepali-English), and the scarcity of structured legal resources complicate dataset preparation.

For the generation component, LLMs such as GPT-4, Llama-3-Instruct, and Mistral 7B have been successfully employed to synthesize coherent answers from retrieved information. These studies show that integrating curated datasets with powerful generative models and optimized retrieval strategies can mitigate hallucinations and improve factual accuracy, which is critical for legal applications (Bhusal & Baral, 2025; Hindi et al., 2025; Kabir et al., 2025).

Major Issues and Challenges

The literature review and project documents identify several key issues that need to be addressed:

- **Data Challenges**
  Many legal documents in Nepal are in scanned PDF formats, requiring optical character recognition (OCR) for data extraction. There is also a lack of open APIs for law retrieval, making data collection challenging.

- **Legal and Ethical Issues**
  Nepal currently lacks a robust legal framework for data protection, which is a major concern when handling sensitive user information. There is also the risk of algorithmic bias, which could unfairly affect different groups within Nepal’s diverse population.

- **Socio-technical Barriers**
  The adoption of AI in Nepal's legal sector faces significant challenges, including a lack of investment and advanced technological infrastructure. Furthermore, there is notable

resistance from some legal professionals who hold misconceptions that AI will replace them (Silwal, 2022).

- **Inference Latency and Performance Optimization**
  Advanced agentic frameworks involving multiple agents require several back-and-forth iterations, which can increase response time. Optimized platforms are needed for rapid data processing and low-latency responses to ensure LLMs deliver required performance in different operational environments (Singh et al., 2025).

- **Data Privacy and Regulatory Compliance**
  Massive language models handle data from multiple sources, which may include sensitive user information. Leakage of such data can violate privacy regulations like GDPR or HIPAA, harming trust. Ensuring regulatory compliance through LLMOps processes, audit logging, and data governance is critical (A3Logics, 2025; Zeng et al., 2024).

- **Integration with Legacy Ecosystems**
  Deploying LLMs often requires integration with legacy systems, which may have rigid, rule-based architectures incompatible with flexible, data-driven models. Bridging this gap requires significant engineering effort (A3Logics, 2025).

- **Model Monitoring, Governance, and Maintenance**
  LLM deployment demands constant monitoring to prevent hallucinations or drift, enforce governance policies, and maintain performance. Models must be regularly tuned, updated, and aligned with user needs.

- **User Trust and Explainability Issues**
  Users may hesitate to trust LLM outputs if the reasoning behind responses is unclear, especially in sensitive domains like law. Developing explainable AI that provides interpretable reasoning is technically challenging but essential.

Paper Review

Paper 1: Enhanced Retrieval for QA System Using GPT-4 and RAG Framework for Nepali Legal Documents - Nabin Bhusal

Link:
https://www.scribd.com/document/844123533/Revised2-Unmasked-ENHANCED-RETRIEVAL-FOR-QA-SYSTEM-TAILORED-FOR-NEPALI-LEGAL-DOCUMENTS-FOCUSING-ON-PSC-EXAMS-USING-GPT-4-AND-RAG-FRAMEWORK-pdf

Background & Motivation :
The legal documents in Nepali are very sophisticated and most times unavailable to the masses. The conventional retrieval systems lack the ability to retain contexts. The reason was to create a system that can support domain specific legal queries in a better manner.

The research question / purpose:
Does a better RAG model using GPT-4 as the generator generate a more accurate QA system on Nepali legal text?

Conceptual Framework:
A hybrid retrieval framework with rule-based keyword matching and semantic embeddings, that GPT-4 can be used to generate natural answers.

Data / Resources Used:
Legal texts from Public Service Commission (PSC) examinations, representing standardized Nepali legal content.

Methodology / Approach:

- Precise match retrieval (rule-based).
- Contextual matching Semantic retrieval Semantic retrieval.
- GPT-4 to produce final response.
- Evaluation through F1 scores

Techniques / Configurations:

- Two retrieval systems (rule-based and semantic) are integrated.
- The generation of contextual embedding to semantic search.
- Legal question-answering prompting strategies.

Analysis & Findings:

- Achieved F1 score of 0.69
- And beat simpler retrieval only systems.
- Revealed success in specific areas of the law.

Contribution:

Demonstrates that RAG systems under the condition of LLM can improve QA in Nepali law. Initial solution to AI-based legal support, Nepal.

Limitations:

- Small scope of data (PSC documents alone).
- None on case law or statutory texts.
- Limited reproducibility (code is not shared).

Future Directions:

- Expand to various jurisdictions.
- Establish open code and standards.
- Examine cross-domain QA outside of exams.

Paper 2: A.I. and Law: Scope and Possible Challenges in Nepal - Puja Silwal

Link: https://nepjol.info/index.php/kslr/article/view/64103

Background & Motivation:
The uses of AI in legal systems are growing across the globe. Nepal however has specific socio-economic, infrastructural, and legal challenges. The reason was to examine the preparedness of Nepal to adopt AI.

Research Question / Objective:

What opportunities and challenges are there to be considered with the application of AI into the legal system in Nepal?

Conceptual Framework:

An examination of the legal environment, data protection preparedness, and professional inclination toward AI by a policy and social analysis framework.

Data / Resources Used:No datasets conceptual review of: qualitative review of:

- Nepal’s legal frameworks
- Existing digital infrastructure.
- Socio-economic and cultural preparedness.

Methodology / Approach:

- Thematic analysis of policy records.

- International experiences of AI.
- Determination of Nepal-specific limitations.

Techniques / Configurations:

- Legal framework analysis
- Socio-economic assessment
- Resistance mapping amongst professionals.

Analysis & Findings:

- Absence of laws to protect and secure privacy.
- Underinvestment in AI infrastructure.
- Legal actor resistance to professionals.
- The barriers to AI are high, yet AI may increase access to justice.

Contribution:

Gives a macro-environmental insight into the blockers of AI legal adoption in Nepal. Identifies other policy gaps to technical innovation.

Limitations:

- None of the empirical testing or prototyping of systems.
- Wide in perspective, shallow in technical analysis.

Future Directions:

- Data protection framework development.
- AI pilot projects in access to law.
- Survey of social trust and approval.

Paper 3: From Hallucinations to Help: Can Retrieval-Augmented Generation (RAG) Deliver Trustworthy Clinical Artificial Intelligence - Prajita Niraula, Mallika Upreti

Link:
https://www.researchgate.net/publication/392192428_From_Hallucinations_to_Help_Can_Retrieval-Augmented_Generation_RAG_Deliver_Trustworthy_Clinical_Artificial_Intelligence

Background & Motivation:

LLMs can be very untrustworthy in the healthcare as they do give hallucinations. The paper examines the possibility of RAG structures minimizing errors through the basis of outputs on verifiable clinical data.

Research Question/Objective:

Is it possible to reduce hallucinations and bias in AI outputs to promote trustworthy clinical decision support with the help of RAG?

Conceptual Framework:

RAG is a hybrid of finding authoritative clinical knowledge and generate responses based on the use of LLM, and responses are anchored to trusted knowledge bases.

Data / Resources Used:

- Managed medical knowledge bases.
- Clinical guidelines
- Medical peer-reviewed literature.

Methodology / Approach:

- RAG architecture review in clinical situations.
- Clinical queries testing Prototyping.
- Empirical comparison to standalone LLMs.

Techniques / Configurations:

- Introducing knowledge base into RAG pipeline.
- Prompt engineering domain-specific to clinical.
- Assessment of the accuracy of response that is based on guidelines.

Analysis & Findings:

- RAG increased reliability of 58 percent clinically appropriate responses.
- Outperformed models with LLM only.
- Minimized prejudice and illusions.

Contribution:

Shows the significance of RAG in creating trustable AI in sensitive areas. Offers legal AI transferable lessons in Nepal.

Limitations:

- Still at a prototype stage
- Small scale of evaluation data.
- Needs more extensive clinical validation.

Future Directions:

- Extend RAG to larger medical areas.
- Low-resource healthcare setting test.
- Expand concepts to areas of law and governance.

Paper 4: Multilingual Grammatical Error Correction with Pre-trained Translation Models - Luhtaru, E.Korotkova, M.Fishel (2024)

Link: https://aclanthology.org/2024.eacl-long.73/

Background & Motivation:

The absence of annotated corpora is a hindering factor to grammatical error correction (GEC) of the low-resource languages. The present paper will explore the possibility of adapting multilingual translation models, which already encode cross-lingual grammatical patterns, to enhance the performance of GEC.

Research Question / Objective:

Are massively multilingual machine translation models discoverable to be fine-tuned to GEC and be more effective as compared to special-purpose GEC models, especially in low-resource languages?

Conceptual Framework:

- Multilingual multitasker models learn grammar and syntax of 100 or more languages.
- Those are fine-tuned to use them again on GEC work.
- Cross-lingual transfer can be used to enhance the performance of high-resource languages in low-resource situations.

Data / Resources Used:

- Estonian → ESEC corpus (2,000 pairs)
- German Falko-MERLIN (24,000 pairs) dataset.
- Czech → GECCC corpus (3,200 pairs)
- BEA-2019 and English English Connell 2014 datasets
- Other translation corpora (in addition to GEC datasets) (not entirely mentioned).

Methodology / Approach:

- Error correcting multilingual MT models.
- Both translation and error correction data were being used in the training.
- Assessment plan: F0.5 measure and human quality measure.
- A comparison with mT5-base, mT5-large, and GPT4-reference performance.

Techniques / Configurations:

- Optimizer: AdamW, learning rate = 3e-5, batch size = 32, and epochs = 10.
- Strategy on joint training (translation + GEC data).
- Language-specific fine-tuning procedures.
- Hugging Face Transformers application.

Analysis & Findings:

- Best mT5 baselines in all test languages.
- F0.5 scores:
  - Estonian = 58.7% (+4.2 vs mT5-base)
  - German = 61.4% (+2.8)
  - Czech = 52.3% (+3.1)
  - English = 69.8% (+1.4)
- Defined good cross-lingual transfer effects.

Contribution:

- Multilingual models that have been proven can be re-applied towards GEC.
- Produces high-end results at fewer parameters than mT5.
- Offered a viable route to low-resource languages (e.g., Nepali).

Limitations:

- No release or public code or model.
- Corpus size of translation is not mentioned.
- Minimal breakdown due to grammatical errors types.

Future Directions:

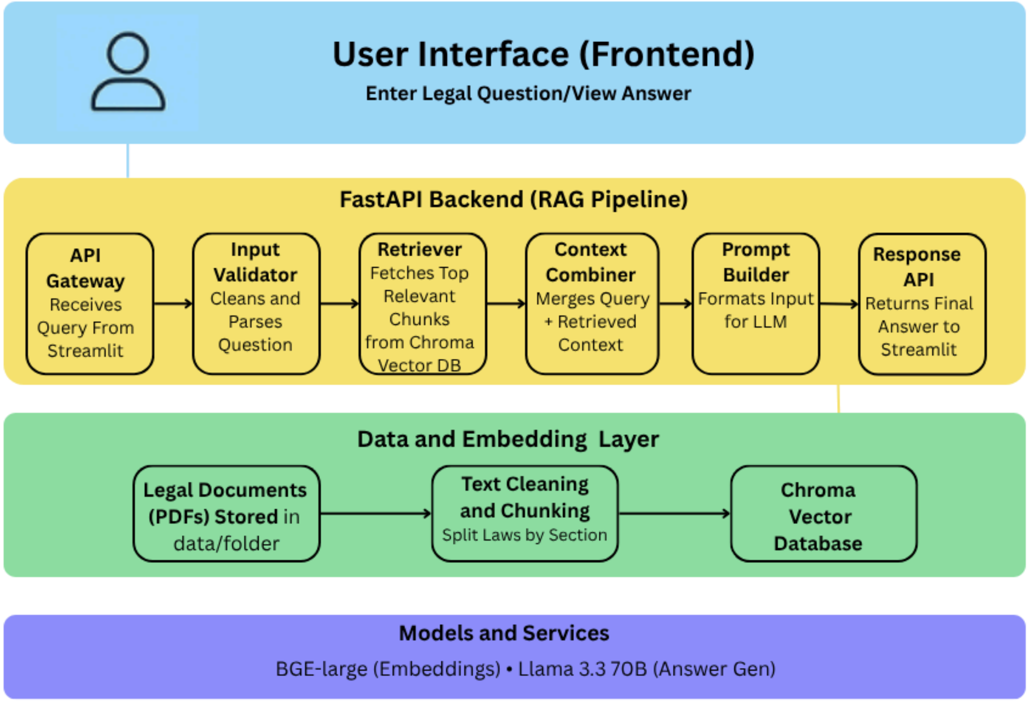
- Increase assessment of additional low-resource languages.
- Fine-tuning with domain-specific (e.g., education, healthcare).
- Make transparency more open-source and open-data.

Literature Review Matrix

Title/Author/ Date	Conceptual Framework	Research Question(s)/ Hypotheses	Datasets	Methodology	Analysis& Results	Conclusions	Implications for Future Research
1. Enhanced Retrieval for QA System... Using GPT-4 and RAG Framework/ Nabin Bhusal	A RAG framework with an enhanced retrieval component, using GPT-4 as the generative model, to create a QA system for Nepali legal documents.	Can an enhanced RAG framework using a powerful LLM like GPT-4 improve the accuracy of a legal QA system for domain-specific Nepali legal documents?	Legal documents from Nepal's Public Service Commission (PSC) examinations.	Integrated rule-based text retrieval with semantic retrieval to form an enhanced RAG system. Used GPT-4 for answer generation.	Achieved an F1 score of 0.69, demonstrating the system's effectiveness for a specific legal domain.	Advanced retrieval methods can significantly boost performance for specialized legal Q&A systems.	Future work should focus on incorporating a wider range of legal documents and testing with a more general legal context.
2. A.I. and Law: Scope and Possible Challenges in Nepal/ Puja Silwal/2022-04-01	A conceptual framework analyzing the current state and potential future of AI in Nepal's legal sector from a policy and social perspective.	What are the opportunities and challenges of applying AI to Nepal's legal system?	N/A (conceptual paper).	A qualitative review of existing legal frameworks, technological infrastructure, and socio-economic factors in Nepal.	Identified a lack of legal framework for data protection, low investment, and professional resistance as key challenges.	While AI has great potential to improve legal access, its implementation in Nepal faces significant legal, ethical, and practical barriers.	Future research is needed to develop clear data protection policies and strategies to build trust and acceptance of AI tools among legal professionals and the public.

3. From Hallucinations to Help: Can Retrieval-Augmented Generation (RAG) Deliver Trustworthy Clinical Artificial Intelligence/ Prajita Niraula, Mallika Upreti/ May 8, 2025	A conceptual framework for using RAG to ensure factual trustworthiness and reduce "hallucinations" in AI systems.	Can RAG mitigate the risks of inaccuracies and bias in AI, making it more trustworthy for critical applications like healthcare?	Discusses the use of curated knowledge bases such as clinical guidelines and peer-reviewed literature.	The paper reviews RAG's architecture and discusses its application in healthcare as a method to ground AI outputs in verifiable sources.	Preliminary studies showed a RAG-based model provided clinically appropriate responses in 58% of queries, a notable improvement over standalone LLMs.	RAG is a crucial method for developing trustworthy AI systems by ensuring their outputs are grounded in reliable data, a principle that can be applied to other domains like law.	The findings imply that RAG can be used to build reliable AI for low-resource settings and complex domains where accuracy is paramount.
4. Multilingual GEC / A. Luhtaru et al. / 2024	Adapting multilingual machine translation models to handle grammatical error correction in low-resource languages.	Is it possible to adapt multilingual machine translation models to perform grammatical error correction in multiple languages effectively?	Estonian, German, Czech, and English GEC corpora augmented with parallel translation data.	Training jointly on translation and grammatical error correction data, followed by language-specific fine-tuning.	Achieved higher F0.5 scores than mT5 models: Estonian +4.2%, German +2.8%, Czech +3.1%, English +1.4%.	Multilingual MT models can serve as a practical solution for GEC in languages with limited resources.	Paves the way for applying GEC in underrepresented languages using cross-lingual adaptation.

MyPocketLawyer RAG Chatbot – System Architecture


User Interface (Frontend)
Enter Legal Question/View Answer
FastAPI Backend (RAG Pipeline)
API Gateway Receives Query From Streamlit
Input Validator Cleans and Parses Question
Retriever Fetches Top Relevant Chunks from Chroma Vector DB
Context Combiner Merges Query + Retrieved Context
Prompt Builder Formats Input for LLM
Response API Returns Final Answer to Streamlit
Data and Embedding Layer
Legal Documents (PDFs) Stored in data/folder
Text Cleaning and Chunking Split Laws by Section
Chroma Vector Database
Models and Services
BGE-large (Embeddings) • Llama 3.3 70B (Answer Gen)

The **MyPocketLawyer** system follows a **Retrieval-Augmented Generation (RAG)** architecture, designed to answer legal questions using information from Nepali and English law documents.

- **User Interface (Frontend):**
  Built with Streamlit, this layer allows users to enter legal questions and view the final answers. It communicates with the backend through a REST API endpoint (`/ask`).

- **Backend (FastAPI Server):**
  The FastAPI application manages user requests and runs the RAG pipeline. It retrieves relevant text chunks from the Chroma vector database using BAAI/bge-large-en-v1.5 embeddings, combines them with the user's query, and sends them to the Groq Llama 3.3 70B model. The backend then formats the model's output and returns a structured, cited answer.

- **Data & Embedding Layer:**
  Contains stored legal PDF documents, cleaned and split into sections. These are

converted into vector embeddings and saved in the Chroma DB for fast retrieval during queries.

- **Models & Services:**
 The BAAI/bge-large-en-v1.5 model handles text vectorization, while the Groq-hosted Llama 3.3 70B model generates the final legal answer.

Project Solution Proposals

Based on the literature review, the design and implementation of MyPocketLawyer can be guided by the following strategies:

1. **Validating the RAG Approach:** Prior studies (Bhusal & Baral, 2025; Hindi et al., 2025) demonstrate that RAG frameworks are effective in domain-specific legal QA systems, reducing hallucinations and improving factual accuracy. Implementing a RAG architecture that combines reliable retrievers and LLM generators ensures that MyPocketLawyer provides trustworthy legal responses.

2. **Refining Data Strategy:** The literature emphasizes the importance of high-quality, curated datasets for legal AI (Kabir et al., 2025). Our plan to source data from official government sources (Nepal Law Commission, Supreme Court) and NGOs (Legal Aid Nepal) aligns with best practices. Attention to bilingual support, document parsing, and OCR for scanned PDFs will address common data preparation challenges.

3. **Optimizing Retrieval Methods:** Advanced retrieval methods, including hybrid approaches that combine semantic retrievers (e.g., sentence-transformers, multilingual-e5) and sparse retrievers (BM25), have been shown to improve answer relevance and accuracy (Bhusal & Baral, 2025; Hindi et al., 2025). Incorporating these strategies will enhance the system’s ability to locate pertinent legal information efficiently.

4. **Selecting Appropriate Generative Models:** Generators such as GPT-4o, Llama-3-Instruct, and Mistral 7B have been successfully used to produce coherent, contextually accurate responses from retrieved data. Using a mix of open-source and proprietary LLMs allows flexibility in balancing cost, performance, and domain adaptation.

5. **Addressing Deployment and Ethical Challenges:** The literature identifies key deployment issues including data privacy, algorithmic bias, and infrastructure limitations (A3Logics, 2025; Zeng et al., 2024). The system design should include measures for secure data handling, compliance with relevant privacy standards, explainable outputs, and mechanisms for continuous monitoring and updates to maintain trustworthiness and

reliability.

6. **Future-Proofing and Scalability:** Considering edge deployment, model distillation, and efficient retrieval strategies can enhance performance while minimizing latency and computational costs. Integrating these strategies ensures that MyPocketLawyer can scale to handle increased user demand without compromising accuracy or reliability.

References

A3Logics. (2025, January 15). *The challenges of deploying LLMs*. A3Logics. Retrieved from https://www.a3logics.com/blog/challenges-of-deploying-llms/

Barron, R. C., Eren, M. E., Serafimova, O. M., Matuszek, C., & Alexandrov, B. S. (2025). *Bridging legal knowledge and AI: Retrieval-augmented generation with vector stores, knowledge graphs, and hierarchical non-negative matrix factorization*. arXiv. https://doi.org/10.48550/arXiv.2502.20364

Bhusal, N., & Baral, D. S. (2025). *Enhanced retrieval for QA system tailored for Nepali legal documents focusing on PSC exams using GPT-4 and RAG framework* [PDF]. Department of Electronics and Computer Engineering, IOE, Tribhuvan University. Retrieved from https://www.scribd.com/document/844123533/Revised2-Unmasked-ENHANCED-RETRIEVAL-FOR-QA-SYSTEM-TAILORED-FOR-NEPALI-LEGAL-DOCUMENTS-FOCUSING-ON-PSC-EXAMS-USING-GPT-4-AND-RAG-FRAMEWORK-pdf

Chung, T., & Yang, C. J. (2024, September 7). *Legal document RAG: Multi-graph multi-agent recursive retrieval through legal clauses*. Medium. Retrieved from https://medium.com/enterprise-rag/open-sourcing-the-whyhow-knowledge-graph-studio-powered-by-nosql-edce283fb341

Hindi, M., Mohammed, L., Maaz, O., & Alwarafy, A. (2025). *Enhancing the precision and interpretability of retrieval-augmented generation (RAG) in legal technology: A survey*. IEEE Access, 13, 46171–46189. https://doi.org/10.1109/ACCESS.2025.3550145

Kabir, M. R., Sultan, R. M., Rahman, F., Amin, M. R., Momen, S., Mohammed, N., & Rahman, S. (2025). *LegalRAG: A hybrid RAG system for multilingual legal information retrieval*. arXiv. https://doi.org/10.48550/arXiv.2504.16121

nexos.ai. (2025, February 11). *6 biggest LLM challenges and possible solutions*. nexos.ai. Retrieved from https://nexos.ai/blog/llm-challenges/

Patil, B., Alam, R., Kalal, T. V., & Porwal, P. (2025, April). *BNS Mitra: RAG-optimized LLM based AI-powered legal virtual assistant*. In *2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI)*. IEEE. https://doi.org/10.1109/ICCSAI64074.2025.11063757

Rahman, S. M. W., Kim, S., Choi, H., Bhatti, D. S., & [Author]. (2025). *Legal query RAG*. IEEE Access, 99(PP), 1–1. https://doi.org/10.1109/ACCESS.2025.3542125

Singh, A., Ehtesham, A., Kumar, S., & [Author]. (2025). *Agentic retrieval-augmented generation: A survey on Agentic RAG*. arXiv. https://doi.org/10.48550/arXiv.2501.09136

Zeng, S., Zhang, J., He, P., et al. (2024). *The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG)*. arXiv. https://doi.org/10.48550/arXiv.2402.16893