JURISMIND: Context-Driven Retrieval for Accurate and Relevant Legal Question-Answering in Patent Filings

Pandey Shourya Prasad[1,*], Vidhish Trivedi[1,*], Madhav Rao[1] and Srijoni Sen[2]
[1]*International Institute of Information Technology, Bangalore, India*
[2]*University of Tilburg, Netherlands*

Keywords: Large Language Model (LLM), Legal Domain, Retrieval Augmented Generation (RAG), Information Retrieval.

Abstract: Large Language Models (LLMs) have demonstrated strong performance in domain-specific conversational forums, but they often suffer from hallucinations—producing factually incorrect or contextually irrelevant responses. This issue is particularly critical in the legal domain, where accuracy is paramount. Existing solutions such as fine-tuning and static retrieval methods struggle to handle the complexities of legal language and often fail to provide sufficient contextual grounding. To address this, we propose JURISMIND, a context-driven retrieval-augmented generation (RAG) pipeline designed for the legal domain, with a focus on Patent Filing. Our approach retrieves relevant legal texts, case law, and statutes based on the input query. This retrieved context is combined with a base prompt and the user query, guiding the language model to respond using the provided legal context. This method significantly reduces hallucinations and improves the contextual accuracy of responses. Preliminary evaluation indicates that 56.32% of responses are in strong agreement and 27.59% in fair agreement with ground truth, totaling 83.91% alignment. Furthermore, JURISMIND achieves a BERTScore of 0.91, outperforming the 0.838 BERTScore of a pretrained LLaMA-based model. The code and dataset are publicly released to support adoption and further research in the developer community.

1 INTRODUCTION

In recent years, the application of AI in domain-specific tasks has seen significant progress in various domains like healthcare(Naz et al., 2024; Shokrollahi et al., 2023), agriculture (Cravero and Sepúlveda, 2021; Aashu et al., 2024) and others. There have also been many methods(Pandey et al., 2023b; Guo et al., 2016; Pandey et al., 2023a) proposed in various fields to enhance the efficiency of computations and results. Along similar lines, it can also be particularly useful in legal question-answering (QA) systems, where precision and context are critical. The legal domain is accumulated with the procedures of law, amendment to the same, and new ones inaddition to the cases where certain new decision was considered. This makes the retrieval process extremely tedious if one wants to know more about legal domain for a specific scenario. Hence technology usage in the AI tool will be preferred over any other automation choices. Retrieval-augmented generation (RAG) has emerged as a promising approach (Bayarri-Planas et al., 2024; Krishna et al., 2024), integrating retrieval mechanisms with generative large language models (LLMs) to improve response accuracy. However, they continue to face issues like hallucinations (Huang et al., 2023; Xu et al., 2024), where the formulated responses are the generation of content that is irrelevant, made up, or inconsistent with the input data. This challenge becomes especially substantial in the legal domain, where mis-interpretation or error has serious consequences. For example, a new applicant pursuing legal services in the form of filing patent is expected to find the most accurate approach to proceed with the services. However with mis-directed path, the applicant not only suffers from bearing extra cost but also squanders away the effort and time without achieving effective outcome. Hence in the patent filing process, it is essential to find appropriate responses towards submission of the application.

Existing mitigation strategies for hallucinations, such as using Epistemic Neural Networks (Verma et al., 2023), fine-tuning (Gekhman et al., 2024; Ballout et al., 2024) or static retrieval, often fail to address the intricate nuances of legal language and

[*] These authors contributed equally to this work.

Prasad, P. S., Trivedi, V., Rao, M. and Sen, S.
JURISMIND: Context-Driven Retrieval for Accurate and Relevant Legal Question-Answering in Patent Filings.
DOI: 10.5220/0013716600004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 323-329

the complex relationships within legal texts. Consequently, they lack the granularity necessary for assessing retrieval quality in RAG systems within the legal domain, where accurate retrieval and contextual relevance are paramount in minimizing hallucination. Techniques like cosine similarity with early exiting(Chen et al., 2024; Pandey et al., 2025) are also adopted to improve the results.

To address these issues, this paper introduces a context-driven RAG pipeline explicitly tailored for the Patent filing of the legal domain, focusing on minimizing hallucinations in legal QA. We have considered Patent filing as one such application within the legal domain to showcase the impact, so that it could be generalized in the future. Besides, Patent filing process is also not straight-forward considering the number of checks, an applicant has to perform based on the jurisdiction. Procedurally, the applicant needs to be aware of the holding and selling rights, apart from the duration of this rights, and its limitations based on the jurisdictions covered (Gaff and Dombrowski, 2013). The deadline for filing, public disclosure part, technology novelty, and cost for filing are few other potential factors that decides the filing process. A well articulated process along these lines will be valuable for the applicant (Raghupathi et al., 2018). Our proposed approach retrieves relevant legal documents, case law, and statutes based on the query and combines them with the language model prompt to guide the response. Preliminary evaluations of this approach demonstrate a high degree of alignment with ground truth answers, underscoring its potential to improve accuracy and reliability in legal applications. The proposed pipeline and dataset are made publicly available (Anonymous,), facilitating further research and development in legal QA.

2 RELATED WORKS

2.1 Retrieval Augmented Generation (RAG)

Zorik Gekhman et al. (Lewis et al., 2021) proposes their retrieval-augmented generation (RAG) models and highlight their effectiveness in knowledge-intensive NLP tasks by combining parametric and non-parametric memory. RAG frameworks integrating a pre-trained seq-2-seq model with dense vector retrieval mechanisms have shown improved factual accuracy and relevance in tasks like open-domain question answering. Their approach surpasses traditional models by generating diverse and factually grounded content, suggesting that RAG models offer a robust, general-purpose solution for applications requiring precise, knowledge-based language generation.

2.2 SMARThealth GPT

Al Ghadban et al. (Al Ghadban et al., 2023) presents a case study on using retrieval-augmented generation (RAG) models to improve healthcare education in low and middle-income countries (LMICs). Their SMARThealth GPT model aims to provide community health workers with accessible and relevant medical information to enhance maternal care. The authors use RAG techniques to demonstrate the effectiveness of the model to retrieve and generate context-specific information, addressing essential knowledge gaps. This work illustrates the potential of RAG models to support health workers in resource-limited settings, improve healthcare outcomes, and possibly other domain-specific tasks.

2.3 OwlMentor

D. Thüs et al. (Thüs et al., 2024) presents a study on OwlMentor, an AI-powered learning platform developed using principles from User-Centered Design (UCD) to focus on usability and usefulness. OwlMentor includes document-based chats, automated question generation, and quiz creation to facilitate interactive learning. The authors have used Technology Acceptance Model (Marangunić and Granić, 2015; Misirlis and Munawar, 2023), to assess system acceptance and investigated how general self-efficacy influences the use and perceived effectiveness of OwlMentor. This was a RAG based system to enhance the learning in education domain.

2.4 Indian Legal Assistant

The Indian Legal Assistant(007UT,), a LLaMA 8.03B based model finetuned on Indian legal texts. The model is designed to assist with various legal tasks and queries related to Indian law. It is a text generation model specifically trained to understand and generate text related to Indian law tasks such as Legal question answering, Case summarization, Legal document analysis and Statute interpretation.

3 METHODOLOGY

The proposed method comprises four stages as shown in Figure 1, namely Client Interface, Semantic Encoding, Context Retrieval, Prompt Generation and Natu-

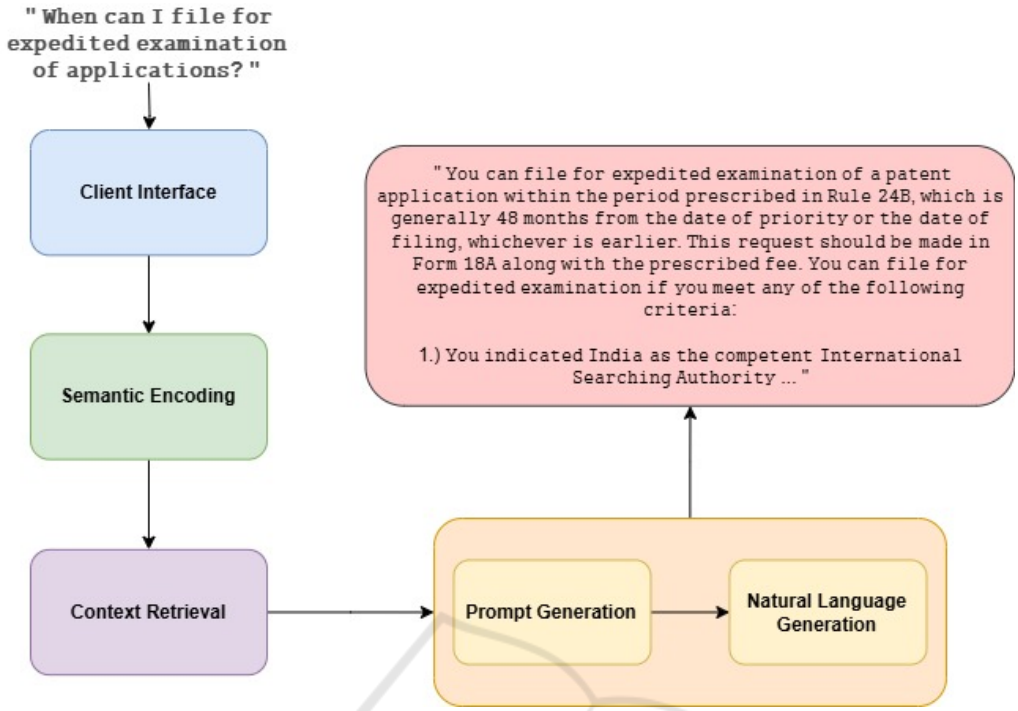

Figure 1: Flow Diagram representing an overview of the proposed method.

ral Language Generation (NLG). The initial stage offers a front-end interface that allows users to submit queries to the proposed system. Subsequent stages process these queries to formulate a response. The users query is converted into a vector embedding, which is then evaluated against the vector embeddings stored in the database using cosine similarity to assess similarity and retrieve relevant information for further usage.

The vector embeddings stored in our document database, specifically related to the domain of Indian Patent Law, are subjected to a similarity scoring process with the embedding generated in the preceding stage. Subsequently, the system identifies the most similar match through semantic search (Guha et al., 2003) to the query and retrieves the corresponding context, enabling the formulation of a more robust and relevant response. The context retrieved from the database is subsequently integrated with a base prompt to enhance the output and then provided to a large language model (LLM) to generate the response in a specific format. The base prompt guides the LLM in producing an answer that is succinct and consistent with the retrieved context. A sample of base prompt for LLM with retrieved-context and user query presented in Figure 3.

4 IMPLEMENTATION DETAILS

4.1 Dataset Details

Four publicly available documents sourced from public websites were utilized, each containing critical legal and procedural information related to intellectual property in India. These documents cover design rules, Indian patent rules, trademark rules, and a set of frequently asked questions addressing common queries in these domains. As the documents were available in PDF format, Optical Character Recognition (OCR) techniques were employed to extract the textual content efficiently. This extraction process ensured that the embedded text, including legal provisions and procedural guidelines, was accurately retrieved for further processing. The subsequent steps involved in structuring and refining this extracted text for effective retrieval-augmented generation (RAG) are described in detail below.

4.2 Database Construction

Milvus (Wang et al., 2021), an open-source vector database, facilitates efficient semantic search across domain-specific documents for data retrieval in a high-dimensional vector space.

A collection of documents about Indian Patent

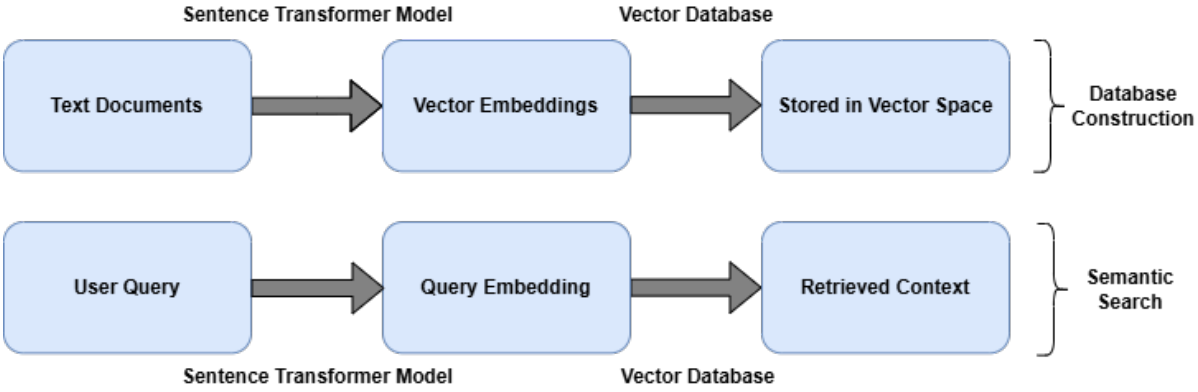

Figure 2: Flow Diagram representing an overview of the proposed method.

Law, denoted as the set $D = \{d_1, d_2, \ldots, d_n\}$, where each document d_i is chunked into smaller units, represented as pages $P_i = \{p_{i1}, p_{i2}, \ldots, p_{im_i}\}$, where m_i is the number of pages in document d_i. Each page p_{ij} is then tokenized into a sequence of words or phrases, represented as $T_{ij} = \{t_1, t_2, \ldots, t_k\}$, where k denotes the number of tokens on page p_{ij}. The tokenized sequences are further mapped to high-dimensional vector embeddings $\mathbf{v}_{ij} \in \mathbb{R}^{384}$, using a suitable embedding function $f : T_{ij} \rightarrow \mathbf{v}_{ij}$. A sentence transformer model, specifically `all-MiniLM-L6-v2` is employed. Formally, $v_{ij} = f(T_{ij})$, where f is an encoding function that projects the tokenized string T_{ij} into a 384-dimensional vector space $\mathbb{R}^{384}$. The resulting vector embeddings $\mathbf{v}_{ij}$ are inserted into the Milvus vector database, with each embedding $\mathbf{v}_{ij}$ represented as a point in a vector space of 384 dimensions. Milvus is designed to efficiently index and retrieve these embeddings using similarity search algorithms, such as cosine similarity (Rahutomo et al., 2012) or inner product, to compute similarity between query embeddings $\mathbf{q} \in \mathbb{R}^{384}$ and the stored embeddings. The similarity measure $\text{sim}(\mathbf{q}, \mathbf{v}_{ij})$ is given by the cosine similarity as stated in the Equation 1, where $\|\mathbf{v}_{ij}\|$ and $\|\mathbf{q}\|$ are the magnitudes of the respective vectors.

$$\text{sim}(\mathbf{q}, \mathbf{v}_{ij}) = \frac{\mathbf{q} \cdot \mathbf{v}_{ij}}{\|\mathbf{q}\| \|\mathbf{v}_{ij}\|} \quad (1)$$

Figure 2 shows the overall flow of database construction and information retrieval using semantic search over stored vector embeddings. These retrieved documents or document segments are semantically aligned with the users query vector $\mathbf{q}$ and provide a suitable context for generating an answer, thereby enabling a more robust and scalable semantic search capability over the corpus of domain-specific documents.

4.3 Information Retrieval and Answer Generation

The user provides a query transformed into a 384-dimensional vector embedding. This embedding is then compared against all stored embeddings in the Milvus vector database using a cosine similarity search. Based on cosine similarity, the top k matches are selected as contextual information for generating the final output. These context strings are integrated with a base prompt to produce a more structured and relevant response. The final structured output is generated using an LLM, `Gemini-1.5-flash`, and displayed to the user.

5 EVALUATION

To assess performance of our model for generating accurate responses, we compared the generated answers with the ground truth provided in the testing dataset, containing 87 questions. The comparison was performed by measuring the similarity between the generated answers and the ground truth using `GPT-4o` (OpenAI et al., 2024), with scores ranging from 0 to 100. These scores assess the model's output with the expected answer. For evaluation, the results were categorized into three distinct performance bands:

- **Low Accuracy (0-39)**: Responses that exhibit low similarity to the ground truth, indicating significant deviation or incorrectness.
- **Moderate Accuracy (40-69)**: Moderately similar Responses, indicating partial correctness or a reasonable attempt with some inconsistencies.
- **High Accuracy (70-100)**: Responses that are highly similar to the ground truth, demonstrating high accuracy and alignment with the expected answers.

As shown in Table 1, the model produced 14 answers that fell within the *Low Accuracy* band, 24 answers within the *Moderate Accuracy* band, and 49 answers within the *High Accuracy* band. This distribution indicates that while a considerable portion of the generated answers were moderately or highly accurate, there remain instances where the model's performance needs improvement. The overall distribution demonstrates that the majority of the generated answers (56.32%) fell within the *High Accuracy* band, reflecting a solid alignment with the ground truth.

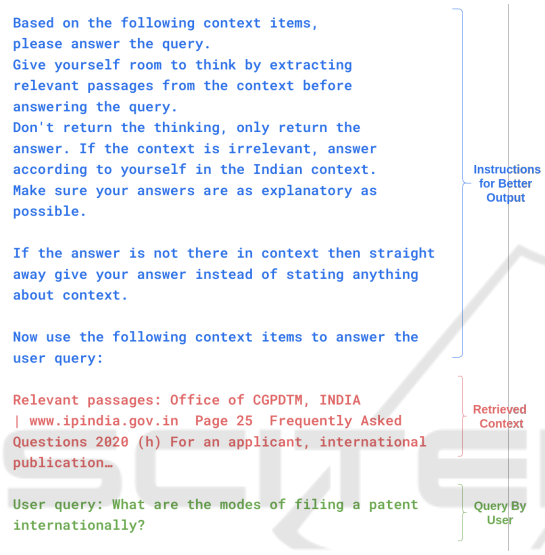

Figure 3: Base prompt for LLM, with retrieved-context and user query.

6 RESULTS

The evaluation results indicate that our proposed retrieval-augmented generation (RAG) pipeline enhances the accuracy of answers in the legal domain, reducing the frequency of hallucinations. Of the 87 tested questions, 49 answers (56.32%) achieved a high similarity score (70-100) with the ground truth, highlighting strong alignment with the expected legal responses. Additionally, 24 answers (27.59%) fell within the moderate accuracy range (40-69), indicating partial correctness and capturing relevant legal context, albeit with minor inconsistencies. However, 14 answers (16.09%) fell into the low accuracy band (0-39), reflecting a need for further refinement in complex or ambiguous cases. This distribution demonstrates that 83.91% of generated answers are either highly or moderately aligned with the ground truth.

7 COMPARISON WITH INDIAN LEGAL ASSISTANT

To assess the effectiveness **JurisMind**, RAG-based model, its performance against the Indian Legal Assistant model(007UT,), a pre-trained LLaMA 8.03B model specifically trained on Indian legal data. The evaluation is conducted using standard NLP metrics, including ROUGE, BLEU, and BERTScore.

Table 2 presents the **average performance scores** of both models across multiple legal question-answering tasks.

7.1 Evaluation Metrics

To quantitatively assess the performance of both models, the following NLP metrics:

7.1.1 (i) ROUGE Score:

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) measures the overlap of n-grams between the generated response and the reference answer. We report:

- **ROUGE-1:** Unigram (single word) overlap.
- **ROUGE-2:** Bigram (two-word sequence) overlap.
- **ROUGE-L:** Longest Common Subsequence (LCS) similarity.

The general formula for ROUGE-N is:

$$\text{ROUGE-N} = \frac{\sum_{s \in \text{Ref}} \sum_{w \in s} \text{Match}(w)}{\sum_{s \in \text{Ref}} \sum_{w \in s} \text{Total}(w)} \quad (2)$$

where $\text{Match}(w)$ represents the number of overlapping n-grams between the candidate and reference text.

7.1.2 (ii) BLEU Score:

Bilingual Evaluation Understudy (BLEU) measures how closely the generated response matches a set of reference responses by evaluating n-gram precision with a brevity penalty. The BLEU score is computed as:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \quad (3)$$

where:

- p_n is the precision for n-grams.
- w_n is the weight assigned to each n-gram order.
- BP (brevity penalty) is applied if the generated response is shorter than the reference.

Table 1: Distribution of Answer Similarity Scores.

Performance Band	Number of Answers	Percentage (%)
Low Accuracy (0-39)	14	16.09 %
Moderate Accuracy (40-69)	24	27.59 %
High Accuracy (70-100)	49	56.32 %

7.1.3 (iii) BERTScore:

BERTScore is a semantic similarity metric based on contextual word embeddings. Instead of token overlap, it computes cosine similarity between embeddings of words in the candidate and reference sentences:

$$\text{BERTScore} = \frac{1}{|S|} \sum_{w \in S} \cos(\text{Emb}(w)_{\text{ref}}, \text{Emb}(w)_{\text{cand}}) \tag{4}$$

where:

- $\text{Emb}(w)_{\text{ref}}$ and $\text{Emb}(w)_{\text{cand}}$ are the contextual embeddings of word w in the reference and candidate sentence, respectively.
- $\cos(\cdot,\cdot)$ denotes cosine similarity.
- S is the set of all words in the reference sentence.

7.2 Analysis of Results

The performance comparison in Table 2 demonstrates that **JurisMind** outperforms the Indian Legal Assistant model across all evaluation metrics. The key observations are:

- **Higher Precision and Recall:** JurisMind achieves a **ROUGE-1 score of 0.5378**, significantly surpassing ILA's 0.2259, indicating stronger word-level and phrase-level recall.
- **Better Coherence and Fluency:** With a **BLEU score of 19.8545**, JurisMind exhibits better fluency in legal language generation.
- **Stronger Contextual Understanding:** The **BERTScore of 0.9100** demonstrates superior semantic alignment between JurisMind's responses and ground truth answers.

Table 2: Average Performance Comparison of JurisMind and Indian Legal Assistant.

Metric	JurisMind	Indian Legal Assistant
ROUGE-1	0.5378	0.2259
ROUGE-2	0.2667	0.0501
ROUGE-L	0.4585	0.1819
BLEU	19.8545	10.8421
BERTScore	0.9100	0.8480

8 CONCLUSION

Our work proposes a novel RAG pipeline targeted at the legal domain, designed to address the persistent challenge of hallucinations in domain-specific question answering. By integrating dynamically retrieved, query-specific legal texts, our approach provides contextually accurate information that guides the language model, significantly reducing hallucinations. The high percentage of responses in agreement with the ground truth highlights the potential of this method for real-world applications, particularly in domains demanding high accuracy, such as legal systems. NLP metric results highlight the advantage of using a Retrieval-Augmented Generation (RAG) pipeline, which enables JurisMind to generate legally accurate and contextually rich responses. While the Indian Legal Assistant model benefits from its extensive pretraining on Indian legal documents, it struggles with precision and fails to structure responses effectively.

REFERENCES

007UT, V. Indian legal assistant. Hugging Face. model: https://tinyurl.com/KDIRR1.

Aashu, Rajwar, K., Pant, M., and Deep, K. (2024). Application of machine learning in agriculture: Recent trends and future research avenues.

Al Ghadban, Y., Lu, H. Y., Adavi, U., Sharma, A., Gara, S., Das, N., Kumar, B., John, R., Devarsetty, P., and Hirst, J. E. (2023). Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. *medRxiv*.

Anonymous. Dataset and model files of jurismind. GDrive. url: https://tinyurl.com/pipelineKDIR.

Ballout, M., Krumnack, U., Heidemann, G., and Kuehnberger, K.-U. (2024). Show me how it's done: The role of explanations in fine-tuning language models.

Bayarri-Planas, J., Gururajan, A. K., and Garcia-Gasulla, D. (2024). Boosting healthcare llms through retrieved context.

Chen, W.-C., Wong, H.-S. P., and Achour, S. (2024). Bitwise adaptive early termination in hyperdimensional computing inference. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, DAC '24,

New York, NY, USA. Association for Computing Machinery.

Cravero, A. and Sepúlveda, S. (2021). Use and adaptations of machine learning in big data—applications in real cases in agriculture. *Electronics*, 10:552.

Gaff, B. M. and Dombrowski, J. M. (2013). Ten things to know about applying for non-us patents. *Computer*, 46(8):9–11.

Gekhman, Z., Yona, G., Aharoni, R., Eyal, M., Feder, A., Reichart, R., and Herzig, J. (2024). Does fine-tuning llms on new knowledge encourage hallucinations?

Guha, R., McCool, R., and Miller, E. (2003). Semantic search. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, page 700–709, New York, NY, USA. Association for Computing Machinery.

Guo, R., Kumar, S., Choromanski, K., and Simcha, D. (2016). Quantization based fast inner product search. In *Artificial intelligence and statistics*, pages 482–490. PMLR.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Krishna, S., Krishna, K., Mohananey, A., Schwarcz, S., Stambler, A., Upadhyay, S., and Faruqui, M. (2024). Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

Marangunić, N. and Granić, A. (2015). Technology acceptance model: a literature review from 1986 to 2013. *Univers. Access Inf. Soc.*, 14(1):81–95.

Misirlis, N. and Munawar, H. B. (2023). An analysis of the technology acceptance model in understanding university students behavioral intention to use metaverse technologies.

Naz, A., Prasad, P., McCall, S., CHAN, C. L., Ochi, I., Gong, L., and Yu, M. (2024). Privacy-preserving abnormal gait detection using computer vision and machine learning.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., and et al., J. B. (2024). Gpt-4 technical report.

Pandey, N. P., Fournarakis, M., Patel, C., and Nagel, M. (2023a). Softmax bias correction for quantized generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1453–1458.

Pandey, N. P., Kulkarni, S., Wang, D., Gungor, O., Ponzina, F., and Rosing, T. (2025). Dpq-hd: Post-training compression for ultra-low power hyperdimensional computing.

Pandey, N. P., Nagel, M., van Baalen, M., Huang, Y., Patel, C., and Blankevoort, T. (2023b). A practical mixed precision algorithm for post-training quantization.

Raghupathi, V., Zhou, Y., and Raghupathi, W. (2018). Legal decision support: Exploring big data analytics approach to modeling pharma patent validity cases. *IEEE Access*, 6:41518–41528.

Rahutomo, F., Kitasuka, T., Aritsugi, M., et al. (2012). Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1. University of Seoul South Korea.

Shokrollahi, Y., Yarmohammadtoosky, S., Nikahd, M. M., Dong, P., Li, X., and Gu, L. (2023). A comprehensive review of generative ai in healthcare.

Thüs, D., Malone, S., and Brünken, R. (2024). Exploring generative ai in higher education: a rag system to enhance student engagement with scientific literature. *Frontiers in Psychology*, 15.

Verma, S., Tran, K., Ali, Y., and Min, G. (2023). Reducing llm hallucinations using epistemic neural networks.

Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X., Yu, K., Yuan, Y., Zou, Y., Long, J., Cai, Y., Li, Z., Zhang, Z., Mo, Y., Gu, J., Jiang, R., Wei, Y., and Xie, C. (2021). Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2614–2627, New York, NY, USA. Association for Computing Machinery.

Xu, Z., Jain, S., and Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models.