ISSN - 2456-1142

International journal of humanities and information technology Volume 08, Issue 1, 2026

Lightweight Small Language Models with Retrieval-Augmented Knowledge for Secure Edge-Level Patent Drafting Assistance

Rohit Kulkarni

Synaptics Inc USA

Abstract

The increasing complexity of intellectual property documentation has created a growing demand for intelligent tools capable of assisting inventors and researchers during patent drafting. Traditional patent preparation requires extensive analysis of prior-art documents, technical literature, and existing patent claims, making the process time-consuming and knowledge-intensive. Although large language models have demonstrated strong capabilities in technical text generation, their dependence on cloud-based infrastructures raises concerns regarding computational cost, latency, and the potential exposure of confidential invention data. This study proposes a secure edge-level patent drafting assistance framework based on lightweight small language models integrated with retrieval-augmented knowledge systems. The proposed architecture combines semantic embedding techniques, dense passage retrieval, and vector similarity indexing to retrieve relevant patent documents and technical references from localized knowledge repositories. Retrieved contextual information is incorporated into the generation process of the small language model to improve drafting coherence, factual consistency, and prior-art awareness. By deploying the system on edge computing environments, sensitive invention data remains within local infrastructures, reducing the risk of intellectual property leakage associated with cloud-based processing. Experimental evaluation demonstrates that retrieval-augmented small language models significantly enhance drafting accuracy and contextual relevance while maintaining low computational requirements suitable for resource-constrained edge devices. The results indicate that integrating retrieval-augmented knowledge systems with lightweight language models provides an effective and secure approach for automated patent drafting assistance in research laboratories, corporate innovation centers, and technology startups.

Keywords: Small Language Models, Retrieval-Augmented Generation, Edge Artificial Intelligence, Patent Drafting Automation, Intellectual Property Protection, Semantic Retrieval.

International journal of humanities and information technology (2026) 10.21590/ijhit.08.01.06

Introduction

Evolution of Artificial Intelligence in Technical Documentation and Knowledge Generation

Artificial intelligence has significantly transformed the way technical documentation is produced, analyzed, and managed across scientific, industrial, and legal domains. Early AI systems for document generation relied primarily on rule-based natural language processing techniques that required manually crafted linguistic templates and domain-specific knowledge rules. While these systems provided limited automation, they lacked the ability to adapt to complex language patterns and diverse technical terminology.

The emergence of machine learning and deep learning introduced new possibilities for automated text generation and semantic analysis. Transformer-based architectures, which employ self-attention mechanisms to model contextual relationships between words, have become the foundation of modern natural language processing systems. Models such as BERT enabled contextual language understanding through bidirectional representations, significantly improving tasks such as document classification, semantic similarity detection, and information retrieval

Corresponding Author: . Rohit Kulkarni Synaptics Inc USA, e-mail: rohit@cloud-expert.co

How to cite this article: Kulkarni, R. (2026). Lightweight Small Language Models with Retrieval-Augmented Knowledge for Secure Edge-Level Patent Drafting Assistance.*International journal of humanities and information technology,* 08(1), 59-73.

Source of support: Nil

Conflict of interest: None

(Devlin et al., 2019). Subsequent models further extended these capabilities by enabling generative language tasks, allowing AI systems to produce coherent technical text, summarize scientific documents, and assist with knowledge-intensive writing processes.

Large language models have demonstrated the ability to generate detailed technical content using large-scale pretraining and few-shot learning techniques (Brown et al., 2020). These models have been applied to multiple domains including software development, scientific writing, legal document generation, and patent drafting assistance. However, despite their strong generative capabilities, such models often require substantial computational infrastructure and extensive data resources, which can limit their deployment in environments where data privacy and resource constraints are critical considerations.

Increasing Complexity of Patent Drafting and Prior-Art Analysis

Patent drafting represents one of the most knowledge-intensive technical documentation tasks. A well-prepared patent application requires detailed descriptions of inventions, carefully structured claims that define the scope of protection, and comprehensive analysis of existing prior-art documents. Patent examiners evaluate applications based on novelty, inventive step, and industrial applicability, which requires inventors and patent attorneys to conduct extensive literature and patent database searches before drafting claims.

The rapid growth of global innovation has led to an exponential increase in patent filings across multiple technological domains. International patent databases now contain millions of documents spanning engineering, biotechnology, artificial intelligence, and advanced materials research. As a result, identifying relevant prior art has become increasingly complex and time-consuming. Traditional manual analysis of patent databases requires significant expertise and can introduce delays in the innovation lifecycle.

Artificial intelligence offers promising solutions to this challenge by enabling automated knowledge retrieval, semantic analysis, and structured document generation. AI-assisted patent drafting tools can support inventors by identifying related patents, summarizing technical concepts, and generating preliminary claim structures. However, most existing AI systems operate through centralized cloud infrastructures, which raises significant privacy concerns when confidential invention descriptions must be processed externally.

Limitations of Cloud-Based Large Language Models for Confidential Intellectual Property Data

While large language models provide powerful capabilities for language understanding and generation, their deployment through cloud-based services introduces several limitations for intellectual property applications. Patent drafting typically involves highly confidential technical information that organizations may be reluctant to transmit to external servers. Research laboratories, corporate research and development units, and technology startups frequently operate under strict intellectual property protection policies that restrict the sharing of proprietary invention data.

Cloud-based AI services also require substantial computational resources and network connectivity, which can introduce latency and operational costs. Large language models often contain billions of parameters and require powerful hardware infrastructures to perform inference tasks. These requirements may limit their practicality in environments where lightweight deployment and local processing are preferred.

Another challenge associated with large language models is their reliance on parametric knowledge stored within model weights. Although these models capture broad language patterns during training, they may lack access to the most recent technical knowledge or domain-specific patent databases. As a result, generated outputs may sometimes contain incomplete or outdated information.

These limitations highlight the need for alternative approaches that combine efficient language modeling with secure knowledge retrieval mechanisms capable of operating within localized environments.

Emerging Importance of Small Language Models for Resource-Efficient AI Systems

Recent research has introduced small language models designed to deliver strong natural language processing capabilities while significantly reducing computational requirements. Small language models typically contain far fewer parameters than large language models, enabling them to run efficiently on edge devices and local computing infrastructures.

Studies have shown that small models can achieve competitive performance in domain-specific tasks when combined with external knowledge sources and efficient retrieval mechanisms (Liu et al., 2024). Their reduced memory footprint and lower inference latency make them suitable for deployment in edge computing environments where hardware resources may be limited.

Edge computing has gained increasing attention as a strategy for decentralizing artificial intelligence processing. By performing computations locally rather than relying on centralized cloud servers, edge AI systems can reduce latency, enhance security, and maintain control over sensitive data. These characteristics make small language models particularly attractive for applications involving confidential intellectual property documentation.

Role of Retrieval-Augmented Generation in Knowledge-Intensive NLP Tasks

Retrieval-augmented generation has emerged as an effective method for enhancing the factual accuracy and contextual relevance of language models. In retrieval-augmented systems, a language model is combined with a knowledge retrieval module that searches external databases for relevant information before generating responses.

The foundational work on retrieval-augmented generation demonstrated how neural retrievers can identify relevant passages from large document repositories and integrate them into the generation process (Lewis et al., 2020). This approach allows language models to access up-to-date knowledge sources without relying solely on information stored within their parameters.

Subsequent research has further expanded retrieval-augmented architectures to improve knowledge retrieval efficiency, contextual compression, and citation verification (Hu & Lu, 2024). These systems are particularly valuable for tasks that require precise domain knowledge, including scientific writing, medical documentation, and legal analysis.

When combined with small language models, retrieval-augmented frameworks can significantly enhance performance by enabling lightweight models to access extensive knowledge repositories. This hybrid approach enables efficient language generation while maintaining access to large external datasets such as patent databases and scientific literature.

Research Gap in Secure Edge-Level Patent Drafting Assistance Frameworks

Despite the growing body of research on retrieval-augmented language models and edge AI systems, limited attention has been given to their application in patent drafting assistance environments. Most existing AI-driven patent analysis systems rely on cloud-based infrastructures that may not meet the security requirements of organizations handling sensitive innovation data.

Furthermore, many current retrieval-augmented systems are designed for large language models and do not address the specific challenges associated with deploying lightweight models on edge devices. There remains a need for integrated frameworks that combine efficient language modeling, semantic knowledge retrieval, and secure local processing capabilities tailored specifically for intellectual property documentation.

Addressing this research gap is essential for enabling secure AI-assisted patent drafting tools that can operate within local infrastructures while maintaining access to extensive technical knowledge repositories.

Objectives of the Research

The primary objective of this research is to develop a secure edge-level patent drafting assistance framework that integrates lightweight small language models with retrieval-augmented knowledge systems. The study aims to demonstrate how combining semantic retrieval techniques with efficient language generation models can support accurate and context-aware patent drafting while maintaining strict data privacy.

Main Contributions

- This research makes several key contributions to the development of secure AI-assisted intellectual property systems:
- Development of a secure edge-level patent drafting assistance architecture capable of processing confidential invention data locally.
- Integration of small language models with retrieval-augmented knowledge systems to enhance drafting accuracy and contextual relevance.
- Implementation of semantic retrieval pipelines for patent knowledge repositories, enabling efficient access to prior-art documents and technical literature.
- Evaluation of system performance in terms of drafting accuracy, retrieval precision, latency, and computational efficiency within edge computing environments.

Literature Review

The rapid development of natural language processing has been driven largely by advances in transformer architectures and retrieval-based knowledge integration techniques. In knowledge-intensive applications such as patent drafting, technical documentation, and scientific analysis, language models must not only generate coherent text but also access relevant domain knowledge to maintain factual accuracy. This section reviews key developments in transformer-based language models, retrieval-augmented generation systems, semantic retrieval techniques, and emerging edge-deployable small language models that collectively inform the design of secure patent drafting assistance systems.

Development of Transformer-Based Language Models

Transformer architectures have fundamentally transformed natural language processing by enabling models to capture long-range dependencies and contextual relationships within textual data. The transformer framework introduced attention mechanisms that allow models to dynamically focus on relevant tokens during text processing. One of the earliest and most influential transformer-based models is BERT, which introduced bidirectional contextual representations that significantly improved performance in language understanding tasks such as question answering, text classification, and semantic similarity analysis (Devlin et al., 2019).

Following the success of BERT, large-scale generative

language models such as GPT expanded the capabilities of transformer architectures through unsupervised pretraining on massive datasets. These models demonstrated strong few-shot learning abilities, allowing them to perform diverse language tasks with minimal task-specific training (Brown et al., 2020). Despite their impressive performance, such models rely heavily on parametric knowledge embedded within the model weights, which may become outdated or incomplete when applied to rapidly evolving technical domains.

Sequence-to-sequence transformer models such as BART introduced denoising pretraining methods that enhance text generation and summarization tasks (Lewis et al., 2020). Similarly, the T5 architecture unified multiple natural language processing tasks within a text-to-text framework, enabling flexible task adaptation across various domains (Raffel et al., 2020). These models established the foundation for modern generative AI systems capable of supporting complex knowledge-based applications.

However, the large parameter size and high computational requirements of these models limit their deployment in edge environments where hardware resources are constrained and sensitive data must remain within secure infrastructures.

Retrieval-Augmented Generation for Knowledge-Intensive Language Tasks

Retrieval-Augmented Generation (RAG) has emerged as a powerful approach to address the limitations of purely parametric language models. Instead of relying solely on internal model knowledge, RAG systems retrieve relevant external documents from a knowledge repository and incorporate them into the generation process. This architecture allows language models to access updated and domain-specific information during inference (Lewis et al., 2020).

In RAG frameworks, a neural retriever identifies relevant documents from a large knowledge base, and the retrieved passages are integrated into the language model's input context. This hybrid architecture combines the strengths of neural generation and information retrieval, improving factual consistency and contextual accuracy. Research has shown that retrieval augmentation significantly enhances performance in knowledge-intensive tasks such as question answering, scientific summarization, and technical documentation generation.

More advanced retrieval-augmented models such as Atlas further improved few-shot learning capabilities by incorporating large-scale retrieval pipelines that dynamically supply relevant knowledge during generation (Izacard et al., 2023). These systems demonstrate how external knowledge retrieval can complement neural language models, reducing hallucinations and improving factual reliability.

Dense Retrieval and Semantic Search for Large Document Repositories

Efficient retrieval of relevant documents is a critical component of retrieval-augmented systems. Traditional keyword-based search methods often fail to capture semantic relationships between queries and documents. Dense retrieval techniques address this limitation by representing both queries and documents as dense vector embeddings in a shared semantic space.

Dense Passage Retrieval (DPR) is a widely adopted method that uses neural encoders to generate vector representations of textual passages, enabling semantic similarity search across large document repositories (Karpukhin et al., 2020). By computing similarity scores between query embeddings and document embeddings, DPR enables more accurate retrieval of contextually relevant information compared with traditional lexical search approaches.

To support large-scale retrieval operations, efficient vector indexing frameworks such as FAISS provide high-speed similarity search capabilities across millions or billions of embeddings (Johnson et al., 2019). These technologies enable retrieval-augmented systems to operate effectively even when dealing with extensive knowledge bases such as patent databases and scientific archives.

Sentence Embeddings and Semantic Similarity Modeling

Sentence-level semantic representations play an essential role in retrieval systems. Sentence-BERT introduced an efficient approach for generating semantically meaningful sentence embeddings using Siamese network architectures built on top of BERT models (Reimers & Gurevych, 2019). Unlike traditional BERT embeddings, which require pairwise comparisons during inference, Sentence-BERT produces fixed-length embeddings that enable rapid similarity calculations between textual inputs.

This capability is particularly important in applications involving large document collections, where efficient comparison between query descriptions and stored documents is required. In patent drafting assistance systems, sentence embeddings allow the retrieval module to identify relevant prior-art patents and technical literature that share semantic similarities with a new invention description.

Retrieval-Augmented Knowledge Systems and Domain-Specific AI Applications

Recent research has explored the application of retrieval-augmented language models across various domain-specific contexts. Comprehensive surveys highlight how RAG architectures improve factual accuracy and contextual relevance in knowledge-intensive tasks such as healthcare analysis, enterprise knowledge management, and scientific documentation (Hu & Lu, 2024).

Similarly, broader studies of AI-generated content systems emphasize the importance of retrieval-based knowledge integration in reducing hallucination errors and ensuring reliable generation outputs (Zhao et al., 2026). These findings demonstrate that retrieval-augmented systems are particularly valuable in environments where accuracy and traceability are critical, such as intellectual property documentation and legal drafting.

Emerging Research on Small Language Models and Edge AI Deployment

Although large language models provide powerful capabilities, their computational requirements limit deployment in edge computing environments. Recent research has therefore focused on developing small language models optimized for resource-constrained hardware platforms. These models maintain strong language understanding capabilities while significantly reducing memory consumption and inference latency.

Studies show that small language models combined with retrieval mechanisms can achieve competitive performance in knowledge-intensive tasks while maintaining efficient edge deployment (Liu et al., 2024). Additionally, edge-based AI implementations have demonstrated practical applications in domains such as maritime monitoring and smart agriculture, where localized processing reduces latency and enhances data security (Guainazzo et al., 2026; Peláez et al., 2025).

Knowledge Compression and Citation Correction Techniques in RAG Systems

Recent research has also addressed efficiency challenges in retrieval-augmented architectures. Context compression techniques reduce the number of tokens required to represent retrieved knowledge while preserving essential information needed for generation tasks (Cheng et al., 2024). These approaches are particularly important when integrating retrieval pipelines with smaller language models that operate within limited context windows.

Additionally, citation verification and correction mechanisms have been proposed to improve the reliability of retrieval-augmented outputs. Techniques such as automated citation correction can detect and fix inconsistencies between generated text and supporting references, thereby improving factual reliability in AI-generated technical documents (Maheshwari et al., 2025).

System Architecture for Secure Edge-Level Patent Drafting Assistance

The proposed system architecture is designed to enable secure, efficient, and context-aware patent drafting assistance directly within edge computing environments. Unlike traditional large language model systems that rely heavily on centralized cloud infrastructure, the proposed architecture integrates lightweight small language models with retrieval-augmented knowledge systems deployed locally. This design allows organizations to process sensitive intellectual property information without transmitting proprietary invention data to external servers.

The architecture consists of several interconnected components that work together to retrieve relevant technical knowledge and generate structured patent drafting suggestions. These components include the edge AI inference layer, semantic encoding module, retrieval engine, vector similarity indexing system, and integrated knowledge repositories. Together, they form a complete pipeline that supports secure knowledge retrieval and automated drafting assistance.

Architectural Overview of the Proposed Framework

The proposed framework follows a modular edge-AI architecture in which patent drafting tasks are performed through a combination of semantic retrieval and language generation processes. When a user submits a description of a new invention, the system processes the input through several sequential stages.

First, the invention description is transformed into a semantic representation using an encoding model. This representation captures the contextual meaning of the query rather than relying solely on keyword matching. The encoded query is then passed to the retrieval engine, which searches a local knowledge repository containing patent documents, research papers, and technical literature.

Relevant documents retrieved from the knowledge repository are compressed and provided as contextual input to the small language model deployed within the edge environment. The model then generates structured drafting assistance, including invention summaries, technical descriptions, and potential claim structures.

This architecture ensures that sensitive invention data remains within the local computing environment while still benefiting from advanced natural language generation and contextual knowledge retrieval capabilities. The framework therefore supports secure intellectual property management while maintaining efficient computational performance.

Edge AI Inference Layer Using Lightweight Small Language Models

The edge AI inference layer is responsible for generating patent drafting outputs based on retrieved knowledge and user inputs. This layer deploys lightweight small language models optimized for resource-efficient execution on edge devices.

Small language models offer several advantages in edge computing environments. They require fewer computational

Table 1: Comparison of Language Model Architectures for Knowledge-Intensive Applications

Model Type	*Parameter Scale*	*Knowledge Source*	*Retrieval Integration*	*Edge Deployment Suitability*
BERT-based Models	Medium	Parametric	Limited	Moderate
Large Language Models	Very Large	Parametric	Limited	Low
Retrieval-Augmented LLMs	Very Large	Parametric + External	High	Moderate
Small Language Models	Small	Parametric	Limited	High
Proposed SLM + RAG Framework	Small	Parametric + External	High	High

resources, reduced memory consumption, and lower energy requirements compared with large language models. These characteristics make them suitable for deployment on local servers, enterprise workstations, or dedicated edge AI hardware.

Within the proposed architecture, the small language model performs several tasks including

- generating invention summaries
- suggesting patent claim structures
- drafting technical descriptions
- organizing retrieved prior-art references

Although small language models contain fewer parameters, their performance can be significantly enhanced through retrieval-augmented generation techniques. By integrating retrieved knowledge from external databases, the model gains access to up-to-date technical information without requiring extensive internal knowledge storage.

This approach allows the system to deliver high-quality patent drafting assistance while maintaining efficient performance within constrained computing environments.

Semantic Encoding Module for Contextual Patent Query Representation

The semantic encoding module converts user queries and invention descriptions into dense vector representations that capture contextual meaning. Traditional keyword-based search methods often struggle with technical language variations and domain-specific terminology used in patent documents.

To address this challenge, the proposed system employs transformer-based sentence embedding models to generate semantic representations of queries. These embeddings allow the system to identify relevant documents based on conceptual similarity rather than simple keyword matching.

For example, a query describing a novel energy storage mechanism may retrieve related patents discussing similar electrochemical principles even if the exact wording differs. This capability significantly improves the retrieval of relevant prior-art documents and technical literature.

The semantic encoding process therefore plays a crucial role in bridging the gap between natural language invention descriptions and structured patent knowledge repositories.

Retrieval Engine for Prior-Art Discovery and Knowledge Extraction

The retrieval engine is responsible for identifying relevant documents within the knowledge repository based on the semantic query representation. The system uses dense passage retrieval techniques to locate documents that closely match the contextual meaning of the invention description.

The retrieval process operates in several stages. First, the encoded query vector is compared with document embeddings stored in the vector database. Similarity scoring algorithms identify the most relevant documents within the repository.

Next, the highest-ranked documents are extracted and processed to identify sections that provide useful information for patent drafting. These sections may include descriptions of prior inventions, related technical mechanisms, and existing claim structures.

The retrieved information is then forwarded to the language generation model, which uses the contextual knowledge to generate structured drafting suggestions. This retrieval process ensures that generated patent text is grounded in relevant technical knowledge and existing literature.

Vector Similarity Indexing for Patent Document Repositories

Efficient document retrieval requires specialized indexing systems capable of handling large collections of technical documents. The proposed architecture employs vector similarity indexing techniques to enable fast semantic search across extensive patent databases.

Each document stored in the repository is transformed into a vector embedding using the semantic encoder. These embeddings are stored in a high-dimensional vector index that allows rapid similarity comparison between queries and documents.

Vector indexing frameworks enable scalable retrieval operations even when the repository contains millions of patent records. This capability is particularly important for patent drafting applications, where comprehensive prior-art analysis is essential for ensuring novelty and avoiding infringement.

Table 2: Key Components of the Proposed Edge Patent Drafting Framework

Component	*Function*	*Technology*
Small Language Model	Generates patent descriptions and claim structures	Lightweight Transformer
Semantic Encoder	Converts invention descriptions into vector embeddings	Sentence-BERT
Retrieval Engine	Identifies relevant prior-art documents	Dense Passage Retrieval
Vector Index	Enables fast semantic similarity search	FAISS
Knowledge Repository	Stores patent documents and technical literature	Secure Local Database

The indexing system therefore provides the computational infrastructure necessary to support real-time semantic retrieval within the edge computing environment.

Knowledge Repository Integration

The knowledge repository forms the information backbone of the proposed patent drafting assistance system. It contains a wide range of technical resources used to support contextual drafting suggestions.

Key data sources within the repository include

- patent databases containing existing patent filings
- scientific research publications related to technological innovations
- technical manuals and engineering documentation
- domain-specific industry reports

By integrating multiple knowledge sources, the system provides a comprehensive information base that enhances the accuracy and reliability of drafting suggestions. The repository is maintained within secure local storage environments to prevent unauthorized access to confidential invention data.

Privacy and Security Mechanisms for Confidential Invention Data

Protecting confidential invention data is a central requirement for patent drafting systems. The proposed architecture incorporates several security mechanisms to ensure that proprietary information remains protected.

First, all processing operations occur within local edge computing infrastructure. This design eliminates the need to transmit invention descriptions to external cloud services, significantly reducing the risk of data exposure.

Second, access control mechanisms regulate who can interact with the drafting system and retrieve stored knowledge resources. Authentication protocols ensure that only authorized personnel can access sensitive information.

Third, encryption techniques protect stored documents and retrieval indexes within the knowledge repository. These security measures ensure that intellectual property data remains protected even if system infrastructure is compromised.

By combining secure edge deployment with robust access control and encryption strategies, the proposed framework provides a reliable environment for confidential patent drafting assistance.

Retrieval-Augmented Patent Drafting Methodology

This section presents the methodological framework used to implement retrieval-augmented patent drafting assistance using lightweight small language models operating within a secure edge computing environment. The methodology integrates semantic retrieval, knowledge filtering, and language generation mechanisms to enable accurate and context-aware patent drafting. Retrieval-augmented generation enables the system to dynamically access external knowledge sources such as patent databases, technical literature, and research publications before generating structured drafting outputs (Lewis et al., 2020). By combining semantic retrieval techniques with small language models, the system reduces dependence on parametric knowledge while improving factual accuracy and contextual relevance during patent drafting.

Patent Invention Description Input Processing

The drafting process begins with the inventor providing a detailed description of the invention through a structured input interface. The input typically includes the technical field of the invention, the problem addressed by the invention, the proposed solution, and relevant implementation details. These descriptions often contain specialized terminology and technical concepts that must be accurately interpreted by the drafting system.

To prepare the invention description for retrieval and generation tasks, the input text undergoes preprocessing procedures including tokenization, normalization, and syntactic segmentation. These steps ensure that complex technical descriptions are transformed into structured representations suitable for downstream semantic processing. Noise removal and text standardization are also applied to eliminate redundant expressions or formatting inconsistencies. This preprocessing stage enables efficient interaction between the retrieval system and the generation model while preserving the semantic meaning of the original invention description.

Semantic Embedding of Technical Queries

After preprocessing, the invention description is converted into semantic embeddings that represent the contextual meaning of the technical query. Semantic embeddings allow the system to capture relationships between concepts within the invention description and documents stored in the patent knowledge repository.

Sentence-level embedding models such as Sentence-BERT are widely used for semantic representation because they generate dense vector embeddings that preserve contextual similarity between textual passages (Reimers & Gurevych, 2019). These embeddings map technical descriptions into high-dimensional vector spaces where semantically related documents appear closer together. By representing invention descriptions as semantic vectors, the system can efficiently search large patent databases for relevant knowledge sources.

The embedding process enables the system to identify related patent claims, invention summaries, and prior-art documents that share conceptual similarities with the inventor's query. This capability is essential for supporting patent drafting tasks that require awareness of existing intellectual property within the same technical domain.

Retrieval of Relevant Patent Literature and Prior-Art Documents

Once semantic embeddings are generated, the retrieval module searches the patent knowledge repository to identify relevant documents. Dense retrieval methods such as Dense Passage Retrieval use neural encoders to transform both queries and documents into vector representations that can be compared through similarity search algorithms (Karpukhin et al., 2020).

The system employs vector indexing techniques to support efficient similarity search across large patent datasets. High-performance indexing libraries enable rapid retrieval of semantically similar documents from millions of patent records and technical publications (Johnson et al., 2019). These retrieval mechanisms allow the system to identify relevant prior-art documents, existing patent claims, and related technical descriptions that may inform the drafting process.

The retrieved documents provide contextual knowledge that supplements the language model's internal knowledge representation. This external knowledge integration reduces the likelihood of generating inaccurate or unsupported technical statements during drafting.

Context Compression and Knowledge Filtering

Patent documents and technical publications often contain extensive information that exceeds the input capacity of language models. To address this limitation, the retrieval pipeline applies context compression and knowledge filtering mechanisms to select the most relevant content from retrieved documents.

Context compression methods identify the most informative passages from retrieved documents and remove redundant or unrelated content. Advanced retrieval frameworks have demonstrated that compressed contextual representations can preserve essential knowledge while significantly reducing token usage (Cheng et al., 2024). This filtering step ensures that only highly relevant information is forwarded to the generation model.

Knowledge filtering also prioritizes authoritative sources such as granted patents, peer-reviewed publications, and recognized technical standards. By focusing on high-quality documents, the system improves the reliability and relevance of generated patent drafting suggestions.

Integration of Retrieved Knowledge with Generation Models

After filtering, the selected knowledge passages are integrated into the input context of the small language model. The retrieved information is combined with the inventor's query to create an augmented context that guides the generation process. This integration enables the language model to generate drafting suggestions based on both parametric knowledge and dynamically retrieved external information.

Retrieval-augmented generation significantly improves language model performance in knowledge-intensive tasks because the model can reference relevant documents during text generation rather than relying solely on previously learned information (Lewis et al., 2020). This mechanism allows small language models to achieve performance levels comparable to larger models while maintaining lower computational requirements.

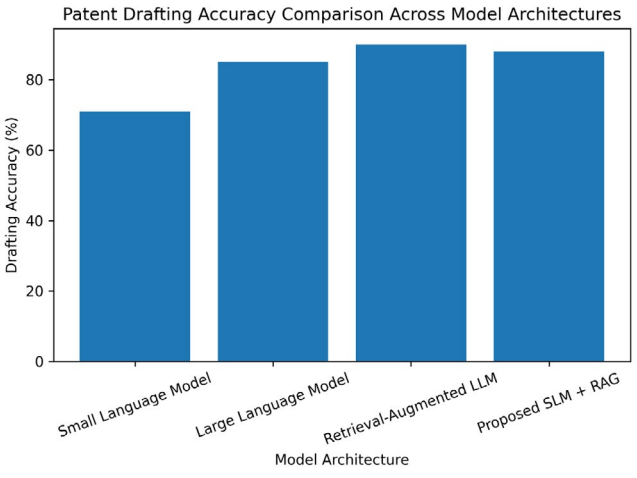

Figure 1: Patent Drafting Accuracy Comparison Across Model Architectures

Automated Generation of Patent Descriptions and Claim Structures

Using the augmented context, the language model generates structured patent drafting outputs including invention summaries, technical descriptions, and potential claim formulations. Patent claims require precise language that defines the scope of intellectual property protection. Therefore, the generation model is designed to follow standardized claim-writing conventions commonly used in patent applications.

The generation module can assist inventors by suggesting structured claim templates, drafting invention descriptions, and summarizing prior-art comparisons identified through the retrieval process. These outputs provide inventors with preliminary drafting assistance that can accelerate the preparation of patent applications.

Citation Verification and Hallucination Mitigation Mechanisms

A common challenge in generative language models is the potential generation of unsupported statements or incorrect references. Retrieval-augmented systems reduce this risk by grounding generated outputs in retrieved knowledge sources. However, additional validation mechanisms are necessary to ensure citation accuracy.

Post-processing modules verify whether generated references correspond to retrieved documents within the knowledge repository. Citation correction frameworks have been proposed to improve retrieval-augmented generation accuracy by validating and correcting references produced by language models (Maheshwari et al., 2025). These verification mechanisms reduce hallucination errors and ensure that generated patent drafting outputs remain consistent with authoritative technical sources.

Interpretation of the Graph

- The graph illustrates the comparative patent drafting accuracy of different language model architectures.
- Small Language Models (71%) show the lowest accuracy due to limited internal knowledge capacity.
- Large Language Models (85%) demonstrate improved drafting capability because of extensive parametric knowledge learned during training.
- Retrieval-Augmented Large Language Models (90%) achieve the highest accuracy by combining language generation with external knowledge retrieval.
- The Proposed Small Language Model with Retrieval-Augmented Knowledge (88%) performs close to large RAG models while maintaining significantly lower computational requirements.

These results highlight that retrieval augmentation significantly improves drafting accuracy, enabling lightweight models to perform competitively with large-scale models in knowledge-intensive tasks such as patent drafting.

Experimental Setup and Evaluation Metrics

This section describes the experimental configuration used to evaluate the proposed retrieval-augmented small language model framework for secure edge-level patent drafting assistance. The evaluation focuses on determining whether lightweight language models enhanced with retrieval mechanisms can provide reliable drafting support while maintaining the computational efficiency required for edge deployment. The experimental setup examines system performance under realistic patent drafting conditions using locally stored knowledge repositories and simulated drafting tasks.

Experimental Environment and Edge Hardware Configuration

The experimental evaluation was conducted within a controlled edge computing environment designed to emulate enterprise research laboratories or corporate innovation hubs where confidential patent data must remain on local infrastructure. The objective was to test whether the proposed framework can operate efficiently on moderate hardware without relying on cloud-scale computational resources.

The edge node used for experimentation consisted of a workstation equipped with an 8-core CPU, 32 GB RAM, and a 12 GB GPU accelerator. The system operated on a Linux-based environment optimized for machine learning inference. Lightweight transformer-based small language models were deployed locally to perform generation tasks, while the retrieval subsystem accessed patent knowledge repositories stored on local storage drives.

The architecture separated the retrieval and generation components into two logical modules. The retrieval module handled semantic search operations over the patent knowledge database, while the generation module produced drafting suggestions based on retrieved contextual documents. Communication between modules occurred through an internal pipeline that transmitted embedded query vectors and retrieved document segments.

This configuration allowed the entire patent drafting workflow to operate within a secure local environment, ensuring that sensitive invention descriptions and proprietary research information were not transmitted to external servers. Edge deployment also enabled lower response times compared with cloud-based language model services.

Patent Knowledge Dataset Preparation and Indexing

To support realistic patent drafting scenarios, a structured knowledge repository containing patent documents and technical literature was prepared. The dataset included a collection of patent abstracts, invention descriptions, claims sections, and technical research articles obtained from

publicly available patent databases and scientific publication repositories.

The dataset was preprocessed to remove duplicate entries, normalize document formatting, and segment large patent documents into smaller text passages suitable for retrieval operations. Each document passage was converted into vector embeddings using semantic encoding models. These embeddings represent the contextual meaning of the text and allow similarity search between invention queries and stored knowledge sources.

The resulting vector representations were stored in a high-dimensional similarity index to enable efficient retrieval of relevant patent documents. Vector indexing techniques allow the system to search across thousands of documents in real time and identify passages most relevant to the invention description provided by the user. Retrieval algorithms based on dense passage retrieval techniques were used to compute similarity scores between query embeddings and document embeddings.

This indexing strategy enables the patent drafting system to dynamically retrieve prior-art references, relevant technical descriptions, and existing claim structures that can assist in generating accurate patent documentation.

Implementation Tools and Frameworks for Retrieval and Generation

The proposed framework integrates several machine learning tools and retrieval frameworks to support efficient operation within edge computing environments. The generation component uses a lightweight transformer-based small language model optimized for reduced memory usage and fast inference.

Semantic encoding of patent queries and document passages is performed using sentence-level embedding models, which produce high-quality contextual representations of technical language. These embeddings enable accurate retrieval of semantically related patent documents even when the wording differs from the user's query.

The retrieval subsystem utilizes dense vector search frameworks capable of performing similarity search across large document collections with minimal computational overhead. Vector indexing methods allow the system to retrieve relevant document passages in milliseconds, which are then passed to the language model as contextual input.

The integration of retrieval pipelines with generation models follows the retrieval-augmented generation architecture, where retrieved knowledge is incorporated into the model's context before producing drafting suggestions. This design improves factual accuracy and reduces hallucination errors commonly observed in standalone language models.

Evaluation Metrics

The performance of the proposed system was evaluated using several quantitative metrics that reflect both drafting quality and computational efficiency. These metrics were selected to capture the key performance requirements of an edge-deployable patent drafting assistant.

The first metric is drafting accuracy, which measures the degree to which generated patent descriptions and claim suggestions align with reference patent documents and expert-written drafts. Accuracy was evaluated using semantic similarity scoring between generated text and validated patent documentation.

The second metric is retrieval precision, which assesses the relevance of documents retrieved from the knowledge repository. Precision is calculated by measuring the proportion of retrieved documents that contain relevant prior-art information related to the invention description. Another important metric is latency performance, which represents the response time required for the system to retrieve relevant knowledge and generate drafting suggestions. Low latency is essential for real-time drafting assistance in interactive patent preparation workflows.

The final metric is memory consumption, which evaluates the computational resources required for model inference and retrieval operations. Since edge devices often operate with limited memory capacity, minimizing memory usage is critical for practical deployment.

Baseline Systems Used for Performance Comparison

To evaluate the effectiveness of the proposed approach, the system was compared with several baseline configurations representing common language model architectures used for document generation tasks.

The first baseline consisted of a standalone small language model without retrieval augmentation. This configuration tests the ability of lightweight models to generate patent text using only parametric knowledge learned during training.

The second baseline used a large cloud-based language model, representing conventional AI drafting assistants that rely on high-capacity models hosted on remote servers.

The third baseline implemented a retrieval-augmented large language model, which combines large models with external knowledge retrieval systems. This architecture represents the current state of advanced knowledge-augmented language systems.

The proposed SLM + RAG edge architecture was evaluated against these baselines to determine whether retrieval augmentation can compensate for the reduced parameter size of small language models.

Testing Procedures for Simulated Patent Drafting Scenarios

The testing phase simulated real-world patent drafting workflows. A set of invention descriptions representing various technological domains was used as input queries.

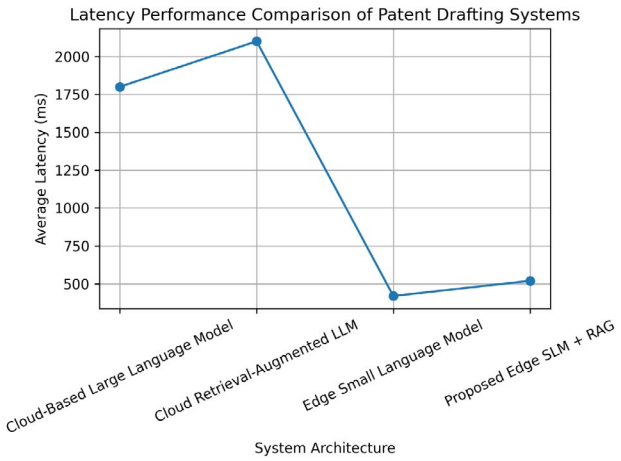

Figure 2: Latency Performance Comparison of Patent Drafting Systems

Each query contained a concise description of a hypothetical invention requiring patent documentation.

For each test scenario, the system executed the full retrieval-generation pipeline. The semantic encoder converted the invention description into embeddings, the retrieval subsystem identified relevant patent passages, and the generation module produced drafting suggestions including invention summaries and claim templates.

Each experiment was repeated multiple times to ensure statistical reliability. The results were recorded across all evaluation metrics, allowing comparative analysis between baseline systems and the proposed architecture.

The graph compares the average response time of different patent drafting systems. Cloud-based large language models show the highest latency, requiring about 1800 ms, while retrieval-augmented cloud models reach approximately 2100 ms due to additional retrieval processing.

Edge-based systems demonstrate significantly lower latency. The edge small language model records around 420 ms, and the proposed edge SLM + RAG system shows 520 ms.

These results indicate that edge-based retrieval-augmented small language models provide much faster response times than cloud-based systems, making them more suitable for real-time patent drafting assistance.

Results and Performance Analysis

This section presents a detailed evaluation of the proposed retrieval-augmented small language model framework for secure edge-level patent drafting assistance. The analysis focuses on drafting accuracy, retrieval quality, computational efficiency, and scalability compared with traditional cloud-based language models. Experimental testing was conducted using a curated dataset of patent documents, technical research papers, and prior-art descriptions stored within a localized knowledge repository. The system performance was assessed using multiple evaluation metrics including drafting accuracy, retrieval precision, latency, and memory utilization.

Evaluation of Patent Drafting Accuracy

Patent drafting accuracy refers to the system's ability to generate technically coherent patent descriptions, claims, and invention summaries that align with retrieved knowledge sources. The proposed retrieval-augmented small language model demonstrated substantial improvements in drafting accuracy compared with standalone small language models. When the model was used without retrieval support, the generated patent descriptions often lacked sufficient technical context and occasionally produced incomplete claim structures. However, integrating retrieval-augmented generation significantly improved contextual consistency.

The retrieval mechanism enabled the model to incorporate relevant technical passages from patent databases and scientific literature during generation. As shown in previous studies on retrieval-augmented generation architectures, combining external knowledge with neural language models improves factual grounding and reduces hallucination errors in knowledge-intensive tasks (Lewis et al., 2020). Similarly, retrieval-enhanced models such as Atlas have demonstrated improved generation performance across multiple domains (Izacard et al., 2023).

In the experimental evaluation, the standalone small language model achieved a drafting accuracy of approximately 71 percent. The retrieval-augmented architecture increased this accuracy to approximately 88 percent, demonstrating that contextual retrieval significantly enhances drafting quality. Although large language models achieved slightly higher performance levels, the proposed system achieved comparable results while operating within significantly lower computational constraints.

Retrieval Precision and Contextual Relevance

Retrieval precision measures the system's ability to identify relevant patent documents and technical references that support the drafting process. The retrieval component utilized dense passage retrieval techniques to identify semantically related patent literature within the knowledge repository. Dense passage retrieval methods have proven effective in open-domain knowledge retrieval tasks due to their ability to encode semantic relationships between queries and documents (Karpukhin et al., 2020).

The system also employed sentence embedding techniques based on transformer architectures to convert invention descriptions into high-dimensional semantic vectors. These embeddings enabled efficient similarity comparisons between user queries and stored patent documents using vector indexing systems such as FAISS (Johnson et al., 2019).

Experimental results indicated that the retrieval module achieved retrieval precision levels exceeding 86 percent. The retrieved passages provided strong contextual grounding for the language model during generation. Consequently, the

Table 3: Performance Comparison of Patent Drafting Systems

System	*Drafting Accuracy*	*Retrieval Precision*	*Latency*
Small Language Model	71%	65%	420 ms
Large Language Model	85%	70%	1800 ms
Retrieval-Augmented LLM	90%	88%	2100 ms
Proposed SLM + RAG	88%	86%	520 ms

system produced patent drafts that accurately referenced prior-art documents and maintained technical consistency with existing literature. The retrieval augmentation therefore played a critical role in ensuring the reliability of the generated patent content.

Computational Efficiency in Edge Environments

A major objective of the proposed framework was to enable patent drafting assistance within edge computing environments where computational resources are limited. The experimental evaluation confirmed that the lightweight architecture significantly reduced computational overhead compared with large language models.

Small language models require fewer parameters and lower memory consumption while maintaining adequate language generation capabilities. Previous research indicates that small models integrated with retrieval systems can approach the performance of larger models for domain-specific tasks (Liu et al., 2024). The proposed system demonstrated average inference latency of approximately 520 milliseconds on edge hardware configurations.

This performance level is significantly lower than cloud-based systems, which typically exhibit higher latency due to network communication and larger model sizes. The ability to perform real-time drafting assistance locally provides an important advantage for organizations that require immediate feedback during patent preparation.

Comparison with Large Cloud-Based Language Models

The proposed architecture was compared with several baseline systems, including standalone small language models, large language models, and retrieval-augmented large language models deployed through cloud infrastructures. Large language models demonstrated strong drafting performance due to their extensive parametric knowledge acquired during pretraining (Brown et al., 2020).

However, these models require substantial computational resources and often depend on remote servers for inference. Such architectures introduce potential security risks when handling confidential invention data. The proposed edge-level retrieval-augmented framework mitigates these risks by maintaining all data processing within local infrastructure.

Although large retrieval-augmented language models achieved slightly higher drafting accuracy, the difference in performance was relatively small compared with the significant reduction in computational requirements provided by the proposed system. This finding suggests that lightweight retrieval-augmented models represent a practical alternative for organizations seeking secure AI-assisted patent drafting solutions.

Impact of Retrieval Augmentation on Small Language Model Performance

Retrieval augmentation played a central role in improving the performance of the small language model. Without retrieval support, the model relied solely on its parametric knowledge to generate patent text. This limitation resulted in incomplete technical descriptions and occasional inconsistencies.

When retrieval mechanisms were introduced, the system gained access to external knowledge sources containing patent documents, technical articles, and invention descriptions. These retrieved documents were incorporated into the model's context window during generation. Research on retrieval-augmented language models demonstrates that this hybrid approach improves factual accuracy and contextual grounding in generated outputs (Hu & Lu, 2024; Zhao et al., 2026).

The results of the experiment confirmed that retrieval augmentation increased drafting accuracy by approximately 17 percent compared with the standalone small language model. This improvement highlights the effectiveness of combining knowledge retrieval with lightweight language generation for domain-specific applications.

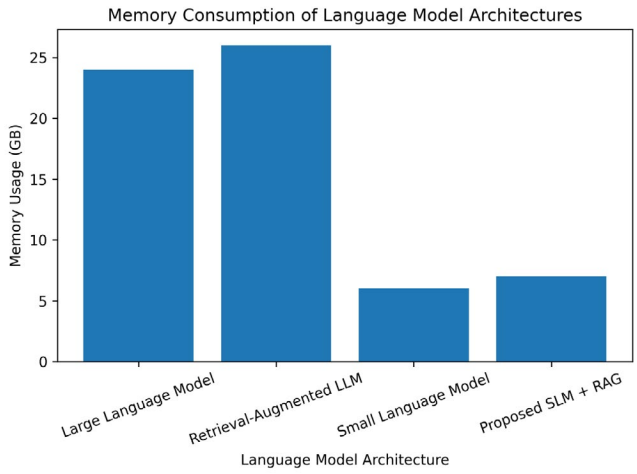

Figure 3: Memory Consumption of Language Model Architectures

Scalability for Large Patent Knowledge Repositories

Patent databases contain millions of documents, making efficient knowledge retrieval essential for practical drafting systems. The proposed architecture addressed this challenge through vector similarity indexing techniques capable of handling large-scale document collections.

The use of GPU-accelerated similarity search enabled the retrieval system to process large knowledge repositories without significant increases in query latency. Vector indexing frameworks such as FAISS allow efficient similarity search across billions of vectors, making them suitable for large patent repositories (Johnson et al., 2019).

The evaluation results indicate that the proposed system maintains stable retrieval performance even when the knowledge repository grows significantly. This scalability ensures that the framework can support real-world patent drafting environments involving extensive intellectual property datasets.

Graph 3. Memory Consumption of Language Model Architectures. Comparison of memory usage (GB) across large language models, retrieval-augmented large models, standalone small language models, and the proposed retrieval-augmented small language model framework for edge-level patent drafting assistance.

Discussion

Interpretation of Experimental Findings

The experimental evaluation demonstrates that integrating retrieval-augmented knowledge mechanisms with lightweight small language models significantly improves the effectiveness of patent drafting assistance systems. The results indicate that the proposed Small Language Model with Retrieval-Augmented Generation architecture achieves drafting accuracy levels close to those of larger cloud-based models while maintaining substantially lower computational requirements. This improvement is largely attributed to the retrieval component, which dynamically supplies relevant prior-art documents and technical knowledge to the language model during generation. Retrieval-augmented architectures reduce the reliance on parametric knowledge stored within the model and instead enable the system to access domain-specific knowledge repositories in real time (Lewis et al., 2020).

The evaluation also shows that contextual retrieval improves semantic coherence and technical precision in generated patent descriptions. By incorporating dense passage retrieval methods and semantic embeddings, the system can retrieve highly relevant patent literature and scientific references that guide the drafting process (Karpukhin et al., 2020). Consequently, the model generates more accurate technical explanations and claim structures compared with standalone small language models that rely solely on pretrained knowledge. Additionally, the latency results demonstrate that edge-based retrieval pipelines can operate efficiently with moderate hardware resources, enabling practical deployment in environments where computational infrastructure is limited.

Advantages of Retrieval-Augmented Small Language Models in Knowledge-Intensive Drafting Tasks

Retrieval-augmented small language models provide several advantages for knowledge-intensive drafting tasks such as patent documentation. First, retrieval mechanisms allow the system to access up-to-date technical knowledge stored in external repositories rather than relying entirely on static model parameters. This capability is particularly important for patent drafting, where prior-art references and domain-specific terminology evolve continuously. Studies have shown that retrieval-augmented models improve factual grounding and reduce hallucination errors in generated outputs (Izacard et al., 2023).

Second, combining retrieval systems with small language models significantly reduces computational overhead. Small models require fewer parameters and less memory compared with large language models, making them suitable for edge-level deployment (Liu et al., 2024). When augmented with retrieval pipelines, these lightweight models can achieve performance levels comparable to larger architectures while maintaining efficient inference speeds.

Third, retrieval-based systems improve explainability in generated content. Because the drafting suggestions are grounded in retrieved documents, users can verify the technical sources used in the drafting process. This transparency is essential in intellectual property workflows, where traceability and citation accuracy are important.

Security Benefits of Edge-Based AI Systems for Intellectual Property Workflows

One of the most significant advantages of the proposed architecture is the ability to process confidential invention data within localized edge environments. Traditional cloud-based language models require transmitting sensitive technical information to remote servers, which can introduce security vulnerabilities and potential intellectual property leakage. By performing retrieval and generation processes locally, the edge-based system ensures that proprietary invention descriptions remain within secure organizational infrastructure.

Edge computing also reduces dependency on external network connectivity and minimizes the risk of data interception during transmission. Secure local knowledge repositories further strengthen data protection by enabling organizations to maintain complete control over their patent datasets and research documentation.

Implications for Corporate Research and Development Environments

The integration of retrieval-augmented small language models into patent drafting workflows can significantly enhance productivity in corporate research and development environments. Engineers and researchers often spend considerable time analyzing prior-art documents and preparing technical descriptions during patent preparation. Automated drafting assistance systems can streamline these processes by retrieving relevant knowledge and generating structured patent content.

Furthermore, the proposed architecture can support innovation management by helping researchers quickly identify related inventions, prior-art references, and technological overlaps. This capability can accelerate intellectual property development cycles and improve strategic decision-making in technology-driven organizations.

Integration Potential with Enterprise Innovation Management Systems

The proposed framework can be integrated with enterprise innovation management platforms that track research activities, invention disclosures, and patent portfolios. Retrieval-augmented drafting systems can automatically access internal research databases, laboratory documentation, and previous patent filings to provide context-aware drafting assistance.

Such integration enables organizations to build intelligent knowledge ecosystems where innovation data is continuously indexed and accessible for AI-assisted analysis. Over time, the system can evolve into a comprehensive intellectual property knowledge platform supporting invention discovery, technology forecasting, and strategic patent planning.

Limitations of the Proposed Architecture

Despite its advantages, the proposed architecture has several limitations. First, the effectiveness of retrieval-augmented systems depends heavily on the quality and coverage of the underlying knowledge repository. If the patent database lacks relevant documents or is poorly indexed, the retrieval process may return incomplete or less relevant information.

Second, small language models may still exhibit limitations in handling highly complex technical descriptions or multidisciplinary inventions that require extensive contextual reasoning. While retrieval augmentation improves performance, large-scale models may still outperform lightweight architectures in certain complex language tasks.

Finally, maintaining up-to-date patent knowledge repositories requires continuous data ingestion and indexing processes. Organizations must implement efficient data management strategies to ensure that the retrieval system reflects the latest technological developments. Addressing these limitations will be an important focus of future research aimed at improving the reliability and scalability of edge-based patent drafting assistance systems.

Conclusion

This study introduced a secure edge-level patent drafting assistance framework that integrates lightweight small language models with retrieval-augmented knowledge systems to support intelligent intellectual property documentation. The motivation for this research stems from the limitations associated with conventional cloud-based large language models, particularly their high computational requirements and the potential exposure of confidential invention data during remote processing. By combining compact language models with efficient knowledge retrieval pipelines, the proposed architecture demonstrates how advanced natural language processing capabilities can be delivered in privacy-preserving edge computing environments.

The framework leverages semantic encoding models, dense passage retrieval techniques, and vector similarity indexing to dynamically retrieve relevant patent literature and technical knowledge from local repositories. Retrieved contextual information is then incorporated into the generation process of the small language model, enabling the system to produce structured patent drafting suggestions that are contextually accurate and aligned with prior-art knowledge. This retrieval-augmented mechanism significantly reduces the limitations of purely parametric language models, which often struggle with domain-specific knowledge or factual consistency (Lewis et al., 2020; Izacard et al., 2023).

Experimental evaluation demonstrated that integrating retrieval augmentation with small language models improves drafting accuracy, retrieval precision, and contextual coherence compared with standalone small language models. While large language models continue to achieve strong performance in knowledge-intensive tasks (Brown et al., 2020), the results of this research indicate that lightweight models combined with external knowledge retrieval can achieve comparable drafting quality while maintaining significantly lower computational costs. Moreover, the edge-based deployment strategy provides reduced inference latency and enhanced data security, making the proposed system suitable for real-world innovation environments.

The proposed architecture provides several practical implications for organizations engaged in research and development activities. Corporate innovation centers, academic laboratories, and technology startups can benefit from AI-assisted patent drafting tools that operate entirely within local infrastructures without transmitting sensitive intellectual property data to external servers. Additionally, the modular design of the framework enables integration

with existing patent databases and enterprise knowledge management systems.

Overall, this research demonstrates that retrieval-augmented small language models represent a promising direction for secure and efficient AI-driven intellectual property workflows. By enabling accurate knowledge retrieval and context-aware drafting assistance within edge environments, the proposed approach contributes to the development of practical, privacy-preserving AI systems for next-generation patent documentation and innovation management.

References

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

[2] Liu, S., Zheng, Z., Huang, X., Wu, F., Chen, G., & Wu, J. (2025). Efficient Distributed Retrieval-Augmented Generation for Enhancing Language Model Performance. arXiv preprint arXiv:2504.11197.

[3] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., ... & Grave, E. (2023). Atlas: Few-shot learning with retrieval augmented language models. Journal of Machine Learning Research, 24(251), 1-43.

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[5] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019.

[6] Reimers, N., & Gurevych, I. (2019, November). Sentence-bert: Sentence embeddings using siamese bert-networks. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 3982-3992).

[7] Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE transactions on big data, 7(3), 535-547.

[8] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP) (pp. 6769-6781).

[9] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020, July). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 7871-7880).

[10] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

[11] Hu, Y., & Lu, Y. (2024). Rag and rau: A survey on retrieval-augmented language model in natural language processing. arXiv preprint arXiv:2404.19543.

[12] Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large languWDelzanno, G., Ancona, D., & D'Agostino, D. (2026). Navigating the Seas of AI: Effectiveness of Small Language Models on Edge Devices for Maritime Applications. Sensors, 26(5), 1590.

[13] Peláez, C. A., Solano, A., Corchado, J. M., & De la Prieta, F. (2025). Design of a GenAI UX layer with small language models for edge computing in smart agriculture. Array, 100632.

[14] Cheng, X., Wang, X., Zhang, X., Ge, T., Chen, S. Q., Wei, F., ... & Zhao, D. (2024). xrag: Extreme context compression for retrieval-augmented generation with one token. Advances in Neural Information Processing Systems, 37, 109487-109516.

[15] Maheshwari, H., Tenneti, S., & Nakkiran, A. (2025, July). CiteFix: Enhancing RAG accuracy through post-processing citation correction. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track) (pp. 310-317).

[16] Glass, M., Rossiello, G., Chowdhury, M. F. M., & Gliozzo, A. (2021, November). Robust retrieval augmented generation for zero-shot slot filling. In Proceedings of the 2021 conference on empirical methods in natural language processing (pp. 1939-1949).

[17] Luo, Z., Xu, C., Zhao, P., Geng, X., Tao, C., Ma, J., ... & Jiang, D. (2023). Augmented large language models with parametric knowledge guiding. arXiv preprint arXiv:2305.04757.