International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.131, August 2026

Enhancing Patent Readability: Leveraging Large Language Model-Generated Taxonomies for Prior Art Analysis

Rohan Kummaraguntla
IO-Innovation
Austin, TX, USA

Andrew J. Ouderkirk
IO-Innovation
Seattle, WA, USA

ABSTRACT

Patent documents are notoriously difficult to read because of their technical jargon, strict formatting, and lack of semantic structure. This paper studies the application of large language models (LLMs) to produce multi-level hierarchical taxonomies as a strategy to make patents more readable and applicable. By converting unstructured language into structured hierarchies, automated taxonomies offer an intuitive and scalable solution to navigating dense legal text for inventors, researchers, and intellectual property professionals. Patents from various fields including software, medical devices, and materials science were analyzed to evaluate the consistency, depth, and readability of the generated taxonomies. The results indicate that LLMs can effectively restructure complex legal documents into understandable, layered forms—providing an accurate and consistent tool for enhancing information retrieval, prior-art analysis, and a step towards human-artificial intelligence collaboration in intellectual property applications.

General Terms

Natural Language Processing, Information Retrieval, Human-Computer Interaction

Keywords

Large Language Models, Natural Language Processing, Patent Analytics, Intellectual Property, Information Retrieval, Document Readability, Taxonomy Generation, Prior Art Analysis

1. INTRODUCTION

The patent system is increasingly facing stark tension between technical disclosure and accessibility. Patents are valuable tools in storing technical knowledge and promoting innovation, but their complex structure and domain-specific language tend to prevent the transfer of knowledge that they are intended to facilitate. Issues of vague language to support broad claims and defensive writing to prevent replications hinder a central aspect of the patents: accessible disclosure [1]. The challenge of maintaining readability while preserving vital specifications impacts a broad range of industries across the innovation ecosystem, from patent examiners conducting prior-art searches to researchers trying to build upon existing developments to inventors writing patents in interdisciplinary fields.

As the number, sophistication, and value of patents continue to grow, particularly in interdisciplinary fields, the demand for more intuitive and semantically structured documentation is only growing in strength. This research utilizes large language models (LLMs), which have shown great progress in natural language comprehension and generation [2, 3], to create structured, multi-level taxonomies of patent documents. These taxonomies provide an alternative method for the arrangement of technical information using hierarchical, readable taxonomies of patent content. By comparing LLM-generated taxonomies against baseline readability metrics and comprehension tasks, this work offers insight into several ways LLMs can significantly enhance the usability and readability of patent documentation.

1.1 Problem Statement

Contemporary patents reflect several characteristics that limit their function as tools of knowledge dissemination. Firstly, the organizational structure of patents, while legally standardized, does not necessarily align with how technical professionals comprehend the information within them. Legal jargon, complex terminology, and lengthy paragraphs that lack semantic guidance make patents extremely difficult to process, even for legal experts. This issue becomes especially pressing in situations that demand speedy information extraction like legal due diligence, infringement analysis, and innovation scouting where long delays can result in severe financial or legal burdens.

Additionally, the traditional classification systems used to classify patents, such as the International Patent Classification (IPC) and the Cooperative Patent Classification (CPC), are manually regulated and often lack the semantic structure needed for effective document comprehension. These methods of classification are relatively outdated, and do not focus on user requirements. Therefore, they neither help inventors or analysts with document comprehension, nor do they attempt to reveal any conceptual relationships within the document, both of which are essential to prior-art analysis and inventing processes.

These structural limitations to current patent documentation are barriers to modern-day innovators, researchers, and patent examiners. Examiners are faced with strict deadlines that limit their ability

to perform extensive and comprehensive prior-art searches. Similarly, legal practitioners must invest significant time and effort to understand dense legal documents, a necessary commitment in order to conduct high-level analysis.
Issues of time and effort are disproportionately large for resource-poor inventors like early-stage companies, independent inventors, and small legal firms. Without the economic ability to hire experts or purchase expensive legal tools, these organizations tend to struggle in leveraging patents for a competitive strategy. Thus, a new method for structuring patent data is needed, one that enhances usability without sacrificing technical accuracy.
The scope of this problem has only expanded as technological complexity has increased. Specifically, cross-disciplinary innovations increasingly challenge traditional patent classification schemes, since they make it difficult for both the inventor and patent examiner to comprehend jargon. These trends suggest that existing approaches to patent readability are underdeveloped and badly-suited for modern-day applications.

1.2 Motivation

Structured taxonomies offer a promising solution to improving patent readability and accessibility. Taxonomies consist of hierarchical structures that outline relationships between concepts at varying levels of abstraction, thereby allowing for an easy searching mechanism [4]. Specifically, effective patent taxonomies could serve multiple functions: efficient identification of technical domains, systematic comparison of related inventions, and revealing conceptual frameworks for understanding applications. Taxonomic organization has empirically proven to be effective. For instance, in fields like biology and information science, taxonomies have been commonly utilized to organize and simplify complicated systems of knowledge [4, 5].
Recent developments in natural language processing offer new opportunities in automating taxonomic generation [2]. Large language models are capable of contextualizing the semantic relationship between a document and its usage. Unlike rule-based approaches, LLMs can process unstructured text and produce a more tailored organizational approach. This research seeks to suggest that automated taxonomies can enhance not just understanding but also the efficiency and quality of prior-art analysis and legal research.

1.3 Significance of This Research

This research seeks to address several important, upcoming needs in intellectual property. From a practical perspective, enhanced patent access could lessen barriers to innovation via a more efficient process for prior-art analysis. Small organizations and individual inventors would greatly benefit from a taxonomic approach to patents, especially if the process is automated.
This research can also benefit patent examiners who might utilize taxonomies to improve the efficiency of their patent examination process. Patent examiners could perform equally thorough prior-art searches within a shorter time-frame if the analyzing process is streamlined. This extends to corporate researchers and legal practitioners who could allocate more time to strategic analysis rather than basic document comprehension.
This research also extends beyond those practical considerations by expanding on the use of LLMs for automating organizational structures to improve document readability. The methods used in this research for generating patent taxonomies could possibly be leveraged in other fields that also have highly technical documents including scientific literature, regulatory documents, permit review, and technical standards.
The broader implications of this research might strengthen the relationship between machine learning and human comprehension of complex information. As technical knowledge continues to grow, tools that can improve cognitive capabilities will become increasingly important.

2. BACKGROUND AND RELATED WORK

This work builds on existing research by applying proven metrics of readability to patent documentation and automating the process through the use of large language models [1]. Even as advancements in LLMs have significantly improved natural language processing (NLP), their application to patent-text remains underexplored. Moreover, existing LLMs are not geared to analyze legal texts. This research seeks to address that by using Clinical Concept Extraction (CCE) for the structured extraction and organization of technical concepts within patents.

2.1 Patent Taxonomies and Classification

Patent classification has traditionally been based on manual, rules-based systems [6]. Standard classification systems like International Patent Classification (IPC), created by the World Intellectual Property Organization (WIPO), and the Cooperative Patent Classification (CPC), which is jointly maintained by the European and U.S. patent offices, classify patents into thousands of technical categories. These systems offer a standardized method of indexing patents and enabling international coordination, but their requirement for rigid, one-off structures makes them inflexible and unable to adapt to new, interdisciplinary technologies.
There are certain machine learning methods that have been applied to automate classification. PatentBERT, a transformer-based model, enhances the scalability of patent classification by translating text into pre-defined classification codes with relatively high accuracy [7]. However, their function is generally limited—they are metadata assignment tools and not knowledge abstraction tools. They do not generate human-readable taxonomies nor do they contribute to improved readability.
Developments like PatentLMM, a large language model used to generate patent figures, have shown the viability of using LLMs for automation [8]. Other works, like large-scale literature reviews of LLMs, indicate how LLMs can help complement human modes of thought. This research builds upon these works by assessing how LLMs can enhance patent readability while remaining accessible.
Additionally, current classification systems serve as backend support for patent databases. They are not created to help users comprehend the content or context of an individual patent. As such, the disconnect between the way patents are classified and the way they are interpreted by users remains an issue.

2.2 Patent Readability Metrics

This research seeks to address patent readability since patents are among some of the most difficult technical texts to comprehend [1]. Research regarding patent readability such as the Flesch–Kincaid Grade-Level (FKGL) model, the Gunning Fog Index, and the Simple Measure of Gobbledygook (SMOG) have revealed the usefulness of taxonomic organization.
Prior attempts at enhancing readability have involved creating summaries, annotating claims, and highlighting important sections. While these approaches are useful, they do not solve the underlying issue of structural disorganization. Specifically, they do not make

clear the conceptual relationships among various components of a patent, leaving users with a disjointed understanding.
The use of taxonomies offers an alternative approach that addresses issues regarding content organization and conceptual clarity. Rather than simply simplifying vocabulary, taxonomies reorganize the document by using an intuitive structure (hierarchical categorization) that is more aligned with human understanding.
By using these metrics to compare the taxonomy to the patent, this research will attempt to quantify the usefulness of a taxonomy for patent documentation.

2.3 Gap Identified

While there has been a surge of research on the application of LLMs to legal and technical documents [2], the majority of that research concentrates on extractive tasks like claim parsing, named entity recognition, or summarization [9, 10]. Far fewer researchers have looked into the generation of novel, structured representations of the documents themselves.
This paper attempts to address this gap by:

—Generating multi-level taxonomies from patent texts using prompt-engineered LLMs

—Conducting readability assessments to evaluate prose literacy, document literacy, and quantitative literacy

—Automating taxonomic generation for future patent classification and prior-art analysis

3. MODEL COMPARISON

This section compares transformer-based models such as GPT-3.5, GPT-4o, Hugging Face LLaMA 2, Gemini 2.5 Pro, and Gemini 2.5 Pro Deep Research with recurrent neural networks such as LSTMs and BiLSTMs in their ability to generate structured taxonomies from patent texts. This comparison explains the trade-offs between recurrent and transformer-based architectures, as well as offering some insight into which large language models are best suited for this task.

3.1 Recurrent Neural Networks

The first attempt at extracting structured information from text relied on recurrent neural networks (RNNs), specifically Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) architectures [10]. The models have the capability to be used for named entity recognition and simple information extraction, but they are worse at performing strongly on longer sequences.
This is clear when analyzing the internal operations of an LSTM cell which are shown below.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$

$$\tilde{C}_t = \tanh(W_C x_t + U_C h_{t-1} + b_C) \tag{2}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{4}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{5}$$

$$h_t = o_t \odot \tanh(C_t) \tag{6}$$

These equations control the gating mechanisms (input i_t, forget f_t, output o_t) and the hidden state h_t and cell state C_t, enabling the model to maintain memory over time. However, since this memory is local, it relies on recurrent connections to transfer information one step at a time. Thus, LSTMs are often challenged by long-range dependencies especially when the input sequence gets more complicated.
Taxonomy generation requires global structure awareness, since output elements are not just sequential but hierarchically related. LSTMs and BiLSTMs lack the ability to generate tree-structured multi-level dependencies without extensive architectural modifications (like tree-LSTM variants). Even with these modifications however, they remain outperformed by many transformer-based models that capture long-range dependencies far more effectively [11–13]
There was a notable difference between LSTM and BiLSTM. While BiLSTM was able to analyze the patent document more consistently through backpropagation, it still failed to generate a useful hierarchical taxonomy. On the other hand, LSTM had difficulties analyzing the patent text itself in the first place.

3.2 Comparative Analysis of Transformer-Based Models

This research also compared the performance of five large language models (LLMs): GPT-3.5, GPT-4o, HuggingFace LLaMA 2, Gemini 2.5 Pro, and Gemini 2.5 Pro Deep Research. These models were compared on several fronts including prompt compliance, contextual understanding, and the quality of the generated results. The findings indicated considerable performance variations, with Gemini 2.5 Pro being the strongest model for structured taxonomy generation, while others had mixed success.
GPT-3.5 (GPT-3 Pro) was capable of producing outputs that were well-structured and followed the instructions in the prompts fairly well. The output had a clear hierarchical relationship with a strong accuracy to the patent. However, GPT-3.5 was inconsistent with regard to depth and formatting, and its output lacked quantitative information (e.g., data, values) which are essential for patent examination. Additionally, it occasionally generated taxonomies that were not semantically deep and failed to map well onto the specific terms used in the domain. The information gathered using clinical concept extraction was going to waste since GPT-3.5 glossed over domain-specific terminology.
GPT-4o performed below standard expectations in this instance as well. While it is strong with conversational fluency, it tended to struggle with maintaining structural coherence in taxonomy generation. The model's output often consisted of incomplete hierarchies or mislabeled nodes. Similar to GPT-3.5, GPT-4o also had difficulties in maintaining the critical terminology necessary in the taxonomy.
The Hugging Face LLaMA 2 model also performed sub-optimally. Its generation lacked depth and structural coherence for hierarchical representation, and did not follow logical parent-child relationships at several levels. Fine-tuning prompts for better structural output produced only incremental benefits, suggesting possible limitations in either model size, training data, or alignment tuning. Although open-source models provide important benefits in transparency, they are not yet an adequate substitute for highly optimized commercial LLMs for structured generation tasks and domain-specific reasoning.
Gemini 2.5 Pro outperformed all of the other models consistently. It was able to generate multi-level taxonomies that consistently met predetermined formatting and structural requirements reliably. The outputs of Gemini 2.5 Pro were accurate and contextually appropriate hierarchies with the correct inclusion of comments and parent-child relationships. It was also incorporating the data gathered using clinical concept extraction to make informed decisions on the

nodes for taxonomy. A crucial aspect of Gemini 2.5 Pro was that it was including quantitative information consistently, which the other models had trouble with. Additionally, Gemini 2.5 Pro illustrated a smaller amount of hallucinations than other models, especially at the lower levels of the taxonomy, where hallucinations are common due to a lack of sufficient contextual information. This showed that the model is strong at understanding technical terminology as well as producing structured, interpretable taxonomies.
Gemini 2.5 Pro Deep Research, a web-augmented variant of the Gemini model, performed worse than its standard version. Gemini 2.5 Pro Deep Research tended to create inconsistent taxonomies and overlook structural prompts. The model's general focus on open-ended synthesis seemed to trade-off with producing structured content precisely. These results indicate that, while Gemini 2.5 Pro Deep Research is strong at completing exploratory tasks, it is less appropriate for narrowly scoped generation prompts.

3.3 Model Configuration

All language models were accessed via their respective APIs, and prompt engineering was standardized to ensure comparability. The temperature parameter was set to 0.2 to reduce output randomness. Since the taxonomy is intended to be a tool to simplify patent documentation, not standardize it, slight variation was acceptable, which is why the temperature was not set to 0.0.
The `max_tokens` setting ranged between 4,096 and 8,192 tokens depending on the input chunk size and the token limits imposed by each of the models.

4. METHODOLOGY

A multi-stage methodology was used to investigate the utility of large language models in generating structured patent taxonomies. This section outlines the approach, including how patents were selected and processed, how prompts and model parameters were designed for LLM generation, and how a robust evaluation framework was developed to assess both structural outputs and their impact on user comprehension.

4.1 Patent Selection

To ensure the model was accurately generating taxonomies for all patents regardless of the industry, a diverse selection of patent documents was used. By assembling a large sample of patents from various fields including medical devices, semiconductor manufacturing, polymer chemistry, and software systems, the model was able to effectively process patent data. Selection criteria prioritized patents published within the past decade, with a focus on capturing technical and structural variation. Twenty patents spanning five technological domains (AI/software, medical devices, semiconductors, energy/materials, and manufacturing/robotics) were selected. Appendix A (Table 2) provides the complete list of patents used in the study. Importantly, because only twenty patents were analyzed over 5 domains, the sample size is limited so additional research could be necessary.
Patents were retrieved from the United States Patent and Trademark Office (USPTO) database using Google Patents. Every chosen patent was downloaded in a PDF format to maintain structural integrity while allowing for general accessibility.

4.2 LLM Prompt Engineering

Prompt engineering was a crucial aspect of this research. The prompts requested hierarchical taxonomies, specifically 7- and 10-level classification schemes in order to maintain technical information but display it in a readable manner. The model was instructed to produce an output in strict JavaScript Object Notation (JSON). Each node in the taxonomy was required to have a label and a brief descriptive statement called a "Technical Summary." This provision served the function of organizing the content and making each item self-descriptive. Prompts included contextual cues drawn from patent metadata, as well as structural cues from preceding sections, thus ensuring consistency across different levels.
Both structured prompts with sample taxonomies and rigid formatting guidelines and unstructured prompts with open-ended natural-language requests were tested. This provided the necessary information to measure how prompt formality impacted generation consistency and token length. Structured prompts tended to produce more consistent and highly nested responses, whereas unstructured prompts had results that differed more in depth and category structure. Although structured prompts were more consistent in terms of formatting, the results indicated that over-constraining the model could lead to lower creativity or absence of domain specificity. Thus, producing effective outputs required balancing constraint with flexibility.

4.3 Clinical Concept Extraction

Clinical concept extraction (CCE) is commonly used in biomedical natural language processing (NLP), focusing on the identification of certain words and phrases such as diseases, drugs, and anatomical references from unstructured clinical narratives (e.g., progress notes, discharge summaries) [9, 14]. This process allows for downstream applications including decision support and knowledge graph construction, and this method can also be applied in other domains for different documents. A necessary element in generating the taxonomy is extracting technical content and normalizing it to industry-specific terminology, both of which clinical concept extraction excels at.
Given a sequence of input tokens $T = \{t_1, t_2, \ldots, t_n\}$, a concept extraction model will assign a label sequence $Y = \{y_1, y_2, \ldots, y_n\}$, where each y_i corresponds to a label from a predefined schema (e.g., `A-Term`, `B-Term`, `O`).
CCE typically utilizes a parameterized function f_θ, such as a BiLSTM-CRF or transformer-based model (in this study, the model was transformer-based) [10]:

$$\hat{Y} = f_\theta(T)$$

The output labels are then used to extract entity spans $s_i = (t_j, \ldots, t_k)$, which are normalized to a controlled vocabulary (e.g., IPC, CPC, PATENTSCOPE Metadata) using a linker ϕ:

$$\phi(s_i) \rightarrow c_i \in \mathcal{C}$$

where $\mathcal{C}$ is the set of standardized concepts.
After concept normalization, the set of concepts $C = \{c_1, c_2, \ldots, c_m\}$ can be structured into a hierarchical taxonomy. Let $\mathcal{T}$ denote a directed acyclic graph (DAG) representing the taxonomy, with nodes as concepts and edges as semantic relationships. Taxonomies can be modeled as a relation prediction task between pairs of concepts:

$$r_{ij} = g_\psi(c_i, c_j)$$

where g_ψ predicts the relation type $r_{ij} \in \mathcal{R}$ (e.g., `is-a`, `part-of`) based on embedding similarity or context dependence.
Figure 1 shows how clinical concept extraction is traditionally administered in the medical field. This same process can easily

be translated to patent documentation to identify key words and phrases.

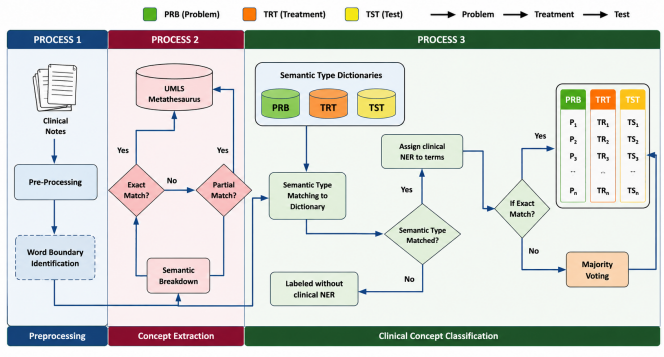

Fig. 1: Recreated clinical concept extraction workflow adapted from [14].

4.4 Taxonomy Generation Pipeline

The generation was done in several passes per patent. First, a clinical concept extraction algorithm was used to highlight key words and phrases that the taxonomy should include. This was a recursive algorithm that scanned related segments of the patent. Next, the entire document was divided into logically separated text blocks that matched conventional patent sections (claims, specifications, etc.). These were then fed into the model using engineered prompts, yielding partial taxonomies.
After processing individual chunks, a recursive merging process was used to combine a single consolidated taxonomy per patent. This was done through semantic clustering to remove duplication and find any overlaps, and then a term normalization algorithm was used to ensure the terminology being used in the taxonomy was similar to that of the patent.
Post-processing included:

—Validation of JSON structure through automated scripts.

—Normalization of terminology (e.g., converting "CVD" to "Chemical Vapor Deposition").

—Hierarchical organization where one element is related to the next

The process produced 7- to 10-level taxonomies per patent, annotated with technical details describing the function or applicability of the lowest-level nodes. The taxonomies were exported in spreadsheet and graph formats for assessment in later phases. An example generated taxonomy is provided in Appendix B (Table 3).

4.5 Evaluation Metrics

To quantify the distinction in readability between the patent and the taxonomy, the following models were used:

—Flesch–Kincaid Grade-Level (FKGL)

$$\text{FKGL} = 0.39\left(\frac{\text{total words}}{\text{total sentences}}\right) + 11.8\left(\frac{\text{total syllables}}{\text{total words}}\right) - 15.59 \quad (7)$$

The Flesch–Kincaid Grade Level estimates the U.S. school grade required to understand a text. It is calculated using the average number of words per sentence and syllables per word.

—Gunning Fog Index (GFI)

$$\text{GFI} = 0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right] \quad (8)$$

The Gunning Fog Index measures the years of formal education needed to comprehend a passage on the first reading. It focuses on sentence length and the percentage of complex words, which are words with three or more syllables.

—Simple Measure of Gobbledygook (SMOG)

$$\text{SMOG Grade} = 1.0430 \times \sqrt{X} + 3.1291 \quad (9)$$

$$X = \text{polysyllabic words} \times \left(\frac{30}{\text{sentences}}\right) \quad (10)$$

The SMOG formula is designed to predict the grade level required to understand a text, especially in professional and technical writing. It calculates readability based on the number of polysyllabic words in a set of 30 sentences.

Using these formulas allowed for a holistic understanding of the taxonomy by applying three well-known readability metrics.
Additionally, a survey was administered to assess how well the generated taxonomies reflected the content of the patents. There were two distinct forms that were created: one that provided the original patent document for reference, and another that provided the taxonomy for reference. Each form contained multiple-choice questions designed to test comprehension across three literacy domains: prose literacy, document literacy, and quantitative literacy. These domains were selected because of their relevance to technical documentation.
Prose literacy questions evaluated the participant's ability to interpret information from text. These questions were related to the implications and relevant summaries of core ideas in the patent. A question in this domain was "Can the graded refractive index increase or decrease with distance from the substrate?"
Document literacy questions focused on determining the participant's ability to navigate information presented in the documents. These questions were related to identifying relationships between elements, following logical paths, or locating specific terms. Questions like "How much of the solvent may be removed from the polymer matrix?" were used to test document literacy.
Quantitative literacy questions assessed the participant's ability to interpret numerical or data-related elements. This included understanding numeric references, frequencies, or patterns related to claims and features. For example, participants were asked, "How can the pore volume be reduced?" which could only be answered by referring to specific data.
To ensure objectivity, the person responsible for drafting the multiple-choice questions was never shown the taxonomies themselves. This separation helped prevent bias in question phrasing or emphasis that could favor or disadvantage any specific output. This allowed the questions to be impartial but still specific to the content of the document.

5. RESULTS

The results from the conducted study indicate that the taxonomy is more readable in prose, document, and quantitative literacy. Moreover, the taxonomy also produced lower scores than the patent when using metrics such as the Flesch-Kincaid Grade Level, the Gunning Fog Index, and the Simple Measure of Gobbledygook. This section details the results of the comparison between the patent and the taxonomy.

5.1 Quantitative Metrics

To evaluate the readability of patent documentation, three established metrics were applied: the Flesch-Kincaid Grade Level (FKGL), the Gunning Fog Index, and the Simple Measure of Gobbledygook (SMOG Index). These metrics were calculated for both the taxonomy and the original patent document. The results from these metrics show there is a significant difference in readability between the two formats.
As shown in Table 1, the taxonomy is considerably more accessible, with a Flesch-Kincaid score of 6.80, indicating middle school-level readability. In contrast, the patent document scored 11.06 on the same scale, suggesting the need for an 11th-grade reading level, approximately 63% higher than that of the taxonomy. Both the Gunning Fog and SMOG indices for the full patent document exceed college-level thresholds, underscoring its complexity. The summary, while still technical, is significantly easier to comprehend by a wider audience.

Table 1. : Readability Metrics Comparison Between Taxonomy and Patent

Metric	Taxonomy	Full Patent PDF
Flesch-Kincaid Grade Level	6.80	11.06
Gunning Fog Index	14.48	16.55
SMOG Index	10.05	14.43

5.2 Survey Analysis

To evaluate how effectively large language models support taxonomy generation, a survey was conducted to target human comprehension. A total of 1000 participants completed the surveys (n=1000), with 500 respondents assigned to read the original patent and 500 assigned to read the taxonomy. Participants were recruited through Amazon Mechanical Turk (MTurk) and Google Surveys. This survey method captured a variety of responses from both general readers and people with experience reading patent documentation. Participants were either given the original patent or the taxonomy, and were asked to answer a series of multiple choice questions testing document literacy, prose literacy, and quantitative literacy. In this study, subjects with the LLM-generated taxonomies outperformed those with the original patent documents on all literacy dimensions.

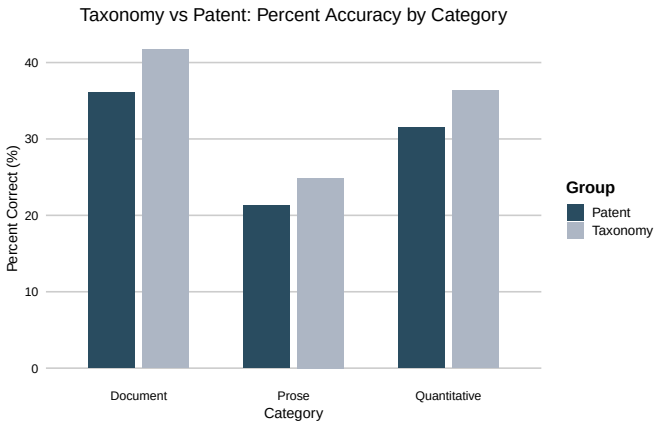

Fig. 2: Taxonomy vs Patent: Percent Accuracy by Category

Figure 2 shows that across all three literacy categories—document, prose, and quantitative—the LLM-generated taxonomy outperformed the original patent. Participants achieved higher average accuracy scores using the taxonomy, indicating the taxonomy effectively covers critical information disclosed in the patent.

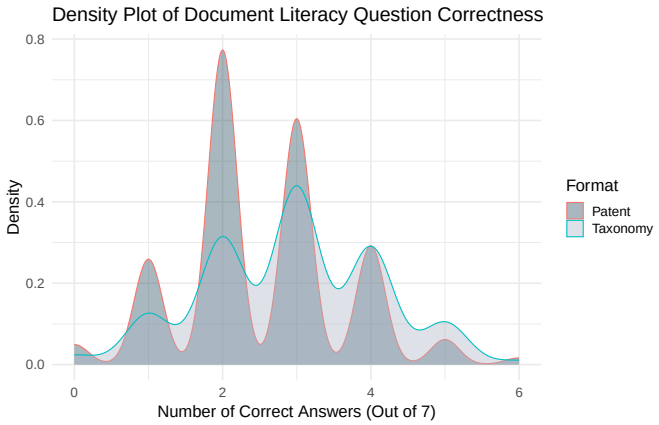

Fig. 3: Document Density Plot: Taxonomy vs Patent.

Figure 3 shows that in questions assessing document literacy, users who used the taxonomy achieved an average accuracy of 43.7%, compared to 36.1% for those who used the patent. Document literacy questions revealed the largest difference between the patent and the taxonomy. This is supported by data that indicates it is far easier to locate information on a taxonomy than a standard document.
To assess whether the distinction was statistically significant, an asymptotic two-sample Kolmogorov–Smirnov (K–S) test was performed. The test yielded a test statistic of $D = 0.172$ with a corresponding p-value of 7.533×10^{-7}, indicating a statistically significant difference between the two distributions at the 5% significance level.

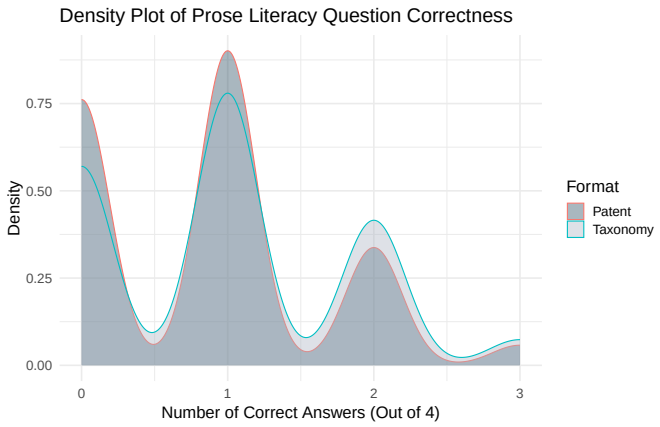

Fig. 4: Prose Density Plot: Taxonomy vs Patent.

Figure 4 shows that in questions assessing prose literacy, users who used the taxonomy achieved an average accuracy of 26.9%, compared to 21.2% for those who used the patent. Prose literacy questions revealed the second-largest difference between the patent and the taxonomy.
To assess whether the distinction was statistically significant, an asymptotic two-sample Kolmogorov–Smirnov (K–S) test was performed. The test yielded a test statistic of $D = 0.074$ with a corresponding p-value of 0.01998, indicating that the difference between the two distributions is statistically significant at the 5% level.

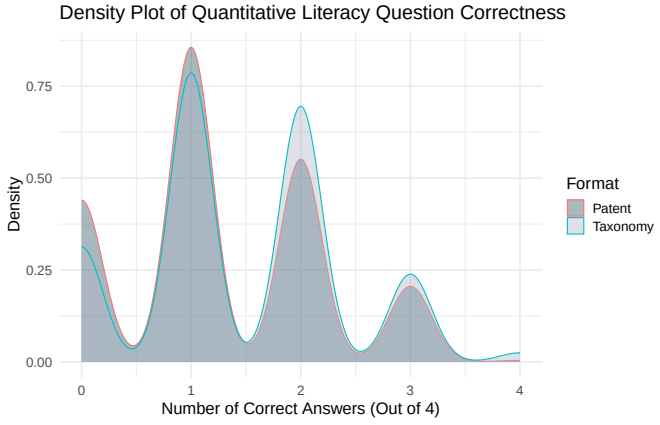

Fig. 5: Quantitative Density Plot: Taxonomy vs Patent.

Figure 5 shows that in questions assessing quantitative literacy, users who used the taxonomy achieved an average accuracy of 36.4%, compared to 31.5% for those who used the patent. Quantitative literacy questions revealed the smallest difference between the patent and the taxonomy. This is substantiated by tests from the various models; LLMs often tend to ignore quantitative data and statistics, and those values are necessary to answer quantitative literacy questions. Even then, it was easier for respondents to use the taxonomy to locate the quantitative information that was available.
To assess whether the distinction was statistically significant, an asymptotic two-sample Kolmogorov–Smirnov (K–S) test was performed. The test yielded a test statistic of $D = 0.096$ with a corresponding p-value of 0.01994, indicating a statistically significant difference between the two distributions at the 5% significance level.

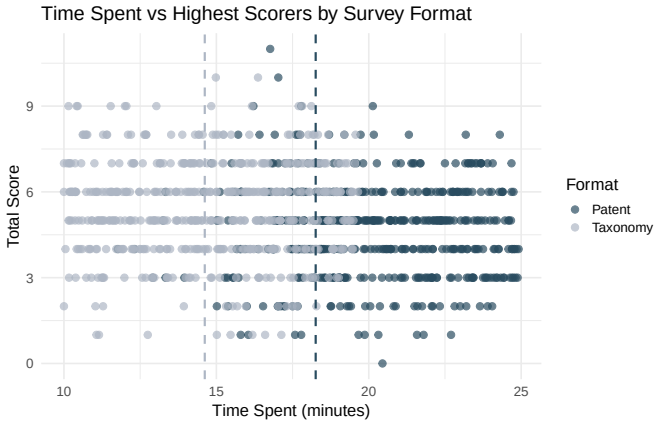

Fig. 6: Scatter Plot with High-Performer Density Markers.

Figure 6 reveals that task completion time was on average shorter for people who used the taxonomy. Even among the highest-scoring participants (those in the top 25%), those in the taxonomy group completed the survey 30% more efficiently, showing the taxonomy is useful for people who are experienced patent readers.

6. DISCUSSION

The results of this research demonstrate the possibilities of large language models in translating complex technical documents. Based on literacy-survey performance and quantitative metrics, taxonomies have the potential to offer significant improvements in intellectual property readability. This section will discuss the implications of these results and current limitations to the method.
Large language models are able to create hierarchical, semantically-nested taxonomies, allowing users to efficiently comprehend technical knowledge. The observed improvements in user comprehension metrics, particularly the large improvements in document literacy and reductions in time-to-task completion, suggest that LLM-generated taxonomies could be a valuable addition to patent documentation. For patent examiners, LLM-generated taxonomies could reduce the time necessary for prior-art searches and improve prior-art search accuracy. For inventors and researchers, LLM-generated taxonomies provide a more readable, efficient way of processing patent documentation especially in a world with greater interdisciplinary research.
From an information architecture perspective, LLM-generated taxonomies have a major advantage over conventional classification systems like the CPC or the IPC. These classification systems are inflexible, rigid, and often misaligned with fast-paced interdisciplinary fields. LLM-generated taxonomies, however, are adaptive and context-specific since they have a unique structure that breaks down the concepts standard documentation tends to miss.
However, there are certain drawbacks to using LLM-generated taxonomies. One such issue is hallucination, where the LLM will add additional, inaccurate information into the taxonomy [15]. Although mitigating hallucinations is possible through prompt engineering and post-processing routines, hallucinations could have major implications resulting in incorrect analysis.

Despite the challenges, the overall conclusion is optimistic. This study indicates that LLMs can process natural language to improve how technical knowledge is organized and shared.

7. CONCLUSION

This paper provides insight into the utility of using large language models to construct structured taxonomies from the unstructured text of patents. The taxonomies created were hierarchically sound and significantly more readable than traditional patent prose. Results from both the surveys and the models indicate that using the taxonomy can lead to higher comprehension rates, with the additional benefit of lowering the time it takes to comprehend the document. By transforming complex technical language into structured, accessible formats, large language models can help users navigate legal documentation more effectively.

Another application of this research is the comparison between various large language models. By comparing the performance of five LLMs (GPT-3.5, GPT-4o, HuggingFace LLaMA 2, Gemini 2.5 Pro, and Gemini 2.5 Pro Deep Research), this paper explains how each model fared in information extraction.

8. FUTURE WORK

While this study provides a strong starting point for efforts to use large language models to facilitate the generation of structured taxonomies from patent documents, it also provides numerous opportunities for future research. One priority should be applying a similar method to a much larger and more diverse set of patents over a wider variety of domains. This could explore whether the benefits of the taxonomy generation observed are preserved across different legal and technical norms.

Expanding the dataset could also help identify potential edge cases where the core patent documentation is necessary to pinpoint certain claims and descriptions. Including patents from countries other than the U.S., such as from European jurisdictions and the World Intellectual Property Organization, might also give a better sense of how well large language models generalize to legal traditions and documentation.

Finally, the methods used in this research could be applied beyond patents to other technical document types including scientific literature, engineering manuals, and clinical trial protocols. Broadening the scope of this research would further prove the utility of automating taxonomies to improve readability.

References

[1] Jin-Dong Li, Kazuaki Seki, and Hisashi Ueda. A novel approach for improving patent readability. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014. URL https://aclanthology.org/W14-1208.pdf.

[2] Jiacheng Li, Yi Zhang, Yichi Zhan, Bill Yuchen Lin, and Xiang Ren. Gptpedia: How gpt understands and uses wikipedia. *arXiv preprint arXiv:2505.19345*, 2024. URL https://arxiv.org/pdf/2505.19345.

[3] Hongyi Fu, Fuwen Luo, Bill Yuchen Lin, Yijia Liu, and Xiang Ren. Are long context language models reading between the lines? *arXiv preprint arXiv:2403.04105*, 2024. URL https://arxiv.org/abs/2403.04105.

[4] Richard Pak, Scott Pautz, and Robert Iden. Information organization and retrieval: An assessment of taxonomical and tagging systems. *Cognitive Technology*, 12(1):31–44, 2007.

[5] Edward O. Wilson. Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 359(1444):739, 2004. doi: 10.1098/rstb.2003.1440. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2003.1440.

[6] Nishant Garg, Bibek Adhikari, and et al. Semantic embeddings improve the classification and retrieval of patent documents. *Scientometrics*, 129(2):375–400, 2024. doi: 10.1007/s11192-024-05206-w. URL https://link.springer.com/article/10.1007/s11192-024-05206-w.

[7] Jieh-Sheng Lee and Jieh Hsiang. Patentbert: Patent classification with fine-tuning a pre-trained bert model. *arXiv preprint arXiv:1906.02124*, 2019. URL https://arxiv.org/abs/1906.02124.

[8] Shreya Shukla, Nakul Sharma, Manish Gupta, and Anand Mishra. Patentlmm: Large multimodal model for generating descriptions for patent figures. *arXiv preprint arXiv:2501.15074*, 2025. URL https://arxiv.org/abs/2501.15074.

[9] Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J. Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, Yiqing Zhao, Sunghwan Sohn, and Hongfang Liu. Clinical concept extraction: a methodology review, 2019. URL https://arxiv.org/abs/1910.11377.

[10] Xi Yang, Jiang Bian, William R. Hogan, et al. Clinical concept extraction using transformers, 2020. URL https://academic.oup.com/jamia/article-abstract/27/12/1935/5943218.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[12] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. URL https://arxiv.org/abs/2001.03708.

[13] Yifan Huang, Yicheng Jiang, Siyuan Wang, and et al. Hierarchical memory transformer for long document understanding. *arXiv preprint arXiv:2404.18255*, 2024. URL https://arxiv.org/abs/2404.18255.

[14] Asim Abbas, Muhammad Afzal, Jamil Hussain, T. Ali, H.S.M. Bilal, Seokhee Lee, and Seokhee Jeon. Proposed clinical concept extraction methodology, 2021. URL https://www.researchgate.net/figure/Proposed-clinical-concept-extraction-methodology-proces fig1_355166145.

[15] Ziwei Ji et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

APPENDIX

A. PATENT DATASET

Table 2 provides the complete list of patents used in this study.

B. EXAMPLE GENERATED TAXONOMY

Table 3 presents an excerpt of the generated taxonomy for Patent US11103892.

Table 2. : Patent Dataset Used for Taxonomy Generation Evaluation

Patent Number	Title	Assignee	Year	CPC	Domain
US11103892	Initiated Chemical Vapor Deposition Method for Forming Nanovoided Polymers	Meta Platforms Technologies LLC	2021	C23C	Energy / Materials
US10452785	Translation Assistance System, Translation Assistance Method and Translation Assistance Program	Rakuten, Inc.	2019	G06F	AI / Software
US10471699	Three-Dimensional Printing With Reciprocal Feeding of Polymerizable Liquid	Carbon, Inc.	2019	B33Y	Manufacturing / Robotics
US10789543	Functional Object-Oriented Networks for Manipulation Learning	University of South Florida St. Petersburg	2020	G06N	AI / Software
US10022406	Compositions for Fecal Microbiota Transplantation and Methods and Devices for Their Preparation and Delivery	Finch Therapeutics Holdings LLC	2018	A61B	Medical Devices
US10743779	Monitoring Mode of an Implantable Medical Device	Mozarc Medical US LLC	2020	A61N	Medical Devices
US11127831	Transistor Structure with an Overlying Gate on a Polysilicon Gate Structure and Related Method	GlobalFoundries US Inc.	2021	H01L	Semiconductors
US10847121	Display Apparatus and Method for Controlling Display Apparatus Displaying Image with Superimposed Mask	Seiko Epson Corp.	2020	H01L	Semiconductors
US10930562	Internal Via With Improved Contact for Upper Semiconductor Layer of a 3D Circuit	French Alternative Energies and Atomic Energy Commission (CEA)	2021	H01L	Semiconductors
US11329052	Method of Processing DRAM	Applied Materials Inc.	2022	H01L	Semiconductors
US9744012	Power Toothbrush With a Modulated Drive Signal	Koninklijke Philips NV	2017	A61C	Medical Devices
US10592783	Risky Transaction Identification Method and Apparatus	Advanced New Technologies Co. Ltd.	2020	G06N	AI / Software
US11288470	Scanning Device and Method for Automatically Triggering Operation Based on a Specific Motion Trajectory	Quanta Computer Inc.	2022	G06K	AI / Software
US11720797	Differential Recurrent Neural Network	Microsoft Technology Licensing LLC	2023	G06N	AI / Software
US10978764	Battery Air Dryer, Battery System and Electric Vehicle Battery	Mann and Hummel GmbH	2021	H01M	Energy / Materials
US11264638	Increasing Gravimetric Capacity in Batteries	University of California San Diego (UCSD)	2022	H01M	Energy / Materials
US10777894	Mechanically Reconfigurable Patch Antenna	Mitre Corp.	2020	H01Q	Energy / Materials
US10946354	Dynamic Mixer Dispense Valve for Two-Component High-Viscosity High-Ratio Compounds	Individual	2021	B29B	Manufacturing / Robotics
US11253893	Suction Unit for a Robotic Sorter	Bollegraff Patents and Brands BV	2022	B07C	Manufacturing / Robotics
US11806235	Prosthetic Heart Valves	Laplace Interventional Inc.	2023	A61F	Medical Devices

Table 3. : Example Excerpt of the Generated Taxonomy for US11103892

L1	L2	L3	L4	Technical Summary
Nanovoided Polymeric Thin Films				
	Formation Process			
		Precursor Mixture		
			Gaseous Monomer, Gaseous Polymerization Initiator, and Solvent	The patent describes a method of forming a nanovoided polymeric thin film by introducing a precursor mixture containing a gaseous monomer, a gaseous polymerization initiator, and a solvent.
			Gaseous Initiator	The gaseous initiator may include hydrogen peroxide, hydroperoxides, alkyl or aryl peroxides, halogens, or azo compounds.
			Flow Rate	The flow rates of the monomer, polymerization initiator, and solvent may each be independently selected from approximately 0.5 sccm to 100 sccm.
		Solvent Removal		
			Extraction Solvent	The patent describes using an extraction solvent, such as supercritical carbon dioxide, to remove the solvent without mechanically damaging the polymer matrix.
			Removal During or After Film Formation	The solvent may be removed during or after formation of the polymeric thin film.
		Control of Void Characteristics		
			Varying Process Conditions	Void size and void-size distribution may be controlled by varying process conditions, such as the ratio of gaseous monomer to solvent.