Synergy of Artificial Intelligence and Boolean Search in Patent Prior Art Assessment

Asiful M. Islam, Robert Vacarneau, Jordan E. Rodriguez, Mihai Surdeanu, Marvin J. Slepian*

Asiful M. Islam
Department of Computer Sciences
University of Arizona
Tucson, Arizona, USA
asifulislam@arizona.edu

Robert Vacarneau
Department of Computer Sciences
University of Arizona
Tucson, Arizona, USA
rvacareanu@arizona.edu

Jordan E. Rodriguez
Department of Computer Sciences
University of Arizona
Tucson, Arizona, USA
jerodriguez@arizona.edu

Mihai Surdeanu
Department of Computer Sciences
University of Arizona
Tucson, Arizona, USA
msurdeanu@arizona.edu

Marvin J. Slepian
Department of Medicine and Biomedical Engineering
James E. Rogers College of Law
Arizona Center for Accelerated Biomedical Innovation
University of Arizona
Tucson, Arizona, USA
slepian@arizona.edu

* Corresponding author

Abstract

Intellectual property plays a crucial role in driving technological and economic advancement in the United States and worldwide. A key factor in translating scientific and technical breakthroughs into valid, useful, and patentable inventions is ensuring that a patent application's written description demonstrates novelty and non-obviousness. A critical step in this process involves comparing the new submission to existing technical work—known as "prior art"—found in both issued and published patents globally, as well as in academic literature. This necessitates the use of *effective* and *explainable* search tools. With the rapid increase in patent applications, artificial intelligence (AI) has emerged as a powerful tool to assist, streamline, and accelerate this process. However, AI search—especially that based on neural networks—is unexplainable due to the "black box" nature of its underlying models. This opacity conflicts with legal requirements to justify why prior art was identified. In contrast, older search algorithms based on Boolean methods are inherently explainable, as they rely on direct query-to-document matching ("what you see is what you get"). Here, we compare and contrast the efficacy and explainability of AI search versus traditional Boolean search. Specifically, we explore two research questions: (a) Can Boolean methods explain the results of AI search? and (b) Can Boolean search fully replace AI search? To address the first question, we develop a method that generates Boolean queries that approximate the outputs of AI search using a set of linguistic heuristics. For the second, we adapt these heuristics to construct Boolean queries based solely on the patent of interest and compare their outputs against AI search. Our experiments on a large patent database indicate that while Boolean methods can explain a majority of AI search results, they cannot fully replace AI search. This suggests that a hybrid approach—where AI retrieves prior art and Boolean methods provide explainability—may represent the future of patent search.

Introduction

Intellectual property in the form of patents, copyrights and trademarks are important property rights established in the founding of the United States under the Constitution and in subsequent legislation, e.g. the Lanham Act (Article I Sec 8, Clause 8, Lanham Act)[1,2]. Intellectual property is a vital and dynamic element for a society and its continued advance, fostering innovation and creativity, scientific and cultural advancement and economic growth[3–5]. Intellectual property, in the form patents, affords conversion of ideas and scientific discoveries with reduction to practice into new practical translational advances covering methods, devices, means of manufacture or compositions of matter[6,7]; or in the arts in the form of copyrights, for creative written, dramatic and related works[8]. In modern society the patent system in particular has been developed to assign this property right for development and commercialization of a technology for a defined period of time. This limited right is often catalytic in science and engineering to secure the greater funding needed to convert idea into reality, beyond that of initial grant funding[9]. In testimony to the technical advance of the United States and related countries around the world the rate of patent generation continues to accelerate at an exponential rate[10,11]. To secure a patent an inventor must search the massive scientific and patent literature to objectively determine the novelty of the proposed invention.[12,13] Relatedly, to grant a patent the United States Patent and Trademark Office (USPTO) and other National patent agencies must determine whether a submitted patent application addresses patentable subject matter, is novel and non-obvious[7,14]. This determination requires detailed technical understanding on the part of professional patent examiners schooled in science, technology as well as the law to determine the novelty and nonobviousness of the submission. Today, with over 850,000 patents being applied for annually, a clear need for the use and dependence upon adjunctive computational technologies in the terms of search engines and artificial intelligence to facilitate this process - for both inventor and examiner - has become critical[10].

A central step in determining the novelty of a patent application, and for that matter the overall inventive process, is examining the closeness, similarity and originality of the present work in comparison to the body of prior inventions broadly referred to as the *prior art* [7,13]. The strict definition of novelty is governed under 35 U.S. Code §102 which states that, “A person shall be entitled to a patent unless: (1) the claimed invention was patented, described in a printed publication, or in public use, on sale, or otherwise available to the public before the effective filing date of the claimed invention.”[15] So as a practical approach patent examiners first determine the overall technical area of the patent, then use various patent databases and related artificial intelligence (AI) tools to broadly identify similar patents as prior art, which brings them a large set of art to examine and compare to make a determination[13,14]. Despite these computational advances, current tools have limitations as they rely on search algorithms that are undisclosed or proprietary, or artificial intelligence tools that are similarly opaque. The issue and problem raised with such search opacity relates to the need in law to be able to explain *why* the prior art was found. In other words, *transparency* and *traceability* are vital so that a legally binding

pronouncement may be made as the reasoning for determining a piece of prior art either nullifies or allows a patent to meet the novelty standard.

However, this transparency is generally lost with the AI tools that have grown to dominate patent search lately[16–19]. In contrast, Boolean search tools such as Westlaw[20] have been a standard tool in patent examination for many years due to their explainability. Examiners are well-trained and experienced in using Boolean logic to perform searches, making it a trusted and reliable method for retrieving relevant prior art. Boolean search allows examiners to precisely control the search criteria by using specific keywords combined with operators such as AND, OR, and NOT. Examiners can easily understand why certain documents were retrieved based on the query they constructed, allowing them to trust the results and make informed decisions. For these reasons, the use of Boolean search aligns well with legal and regulatory standards in patent examination, where a systematic and reproducible search methodology is required to ensure fairness in the examination process.

With a long-term goal of developing tools for the scientist, inventor and examiner to enhance accuracy and ease of prior art search and information retrieval, in this paper, we investigate whether Boolean search can be used to address the lack of transparency of AI searches, while preserving the performance obtained by AI. We investigate two scenarios: (a) whether Boolean search can be used to *explain* the results of an AI prior art search, and (b) whether Boolean search can completely *replace* AI searches. Formally, the two hypotheses investigated here are:

Hypothesis 1: Given a set of reference documents (patents) selected as prior art for a given patent by an AI search tool, a Boolean query can be built to identify and retrieve the same reference documents.
Hypothesis 2: Given solely a patent of interest, a priori selection of prior-art documents from a large database via a Boolean search tool will identify the same documents that will be retrieved via an AI search.

Informally, the first hypothesis focuses on explaining a prior-art search carried out by AI, whereas the second hypothesis focuses on replacing AI search with Boolean tools. We translated these two hypotheses into the following research questions (RQ):

RQ 1: Can Boolean search concur/explain the results of an AI search through a single Boolean query?
RQ 2: Can Boolean search produce equivalent results and therefore replace an AI Search?

To answer these questions, we introduce a method that constructs a Boolean query that explains a collection of *k* patents using linguistic heuristics. In particular, our method: (a) ranks all terms in each patent using the tf-idf informativeness score; (b) selects the *n* most informative terms from each patent; and (c) aggregates them into a single disjunctive query, which is then optimized to remove redundant terms (see Methods for details on this process).

This method is used in two distinct experimental settings:

1. In the first scenario, the input to the proposed query building method consists of *k* similar patents retrieved using an AI search (in particular, we used the search tool on the Google Patent website[21]) for one patent of interest. Then this query is executed over a large collection of patents, and its results are compared against the *k* patents retrieved by AI. If the results are similar, we conclude that the Boolean query explains the results of the opaque AI search.

2. In the second setting, the input to the proposed method is *solely* one patent of interest. If this query produces *k* patents that are similar to the results of the AI search, we conclude that Boolean search can completely replace the AI search.

Results

Experimental Settings

For all experiments discussed in this paper, we used patents provided by IFI CLAIMS Patent Services and Google[21]. Specifically, we utilized a patent dataset made publicly available through the "USPTO - Explainable AI for Patent Professionals" Kaggle coding competition[22]. This dataset contains 13.3 million patents from 1790 to 2023. For each patent, the dataset includes data from multiple patent sections, i.e. title, abstract, claims, description, and Cooperative Patent Classification (CPC) codes. For 4,000 of these patents, the dataset includes a list of the top 50 most similar patents, as retrieved by the AI search tool on the Google Patents website.[1] To support our experiments, we constructed a searchable index containing all 4,000 × 50 = 200,000 patents.[2] Finally, we randomly selected 2,500 of the 4,000 patents as our test dataset.

Scenario 1: Explaining the Results of AI Search

In this scenario, we generate Boolean queries that aimed to explain the top *k* similar patents retrieved by AI. This experiment raised several sub-questions: (a) which patent sections are most informative?, and (b) how does the informativeness of Boolean queries vary with their length? To control for these parameters, we carried out two experiments.

In the first experiment, we generated Boolean queries using a single patent section at a time. In particular, we extracted the most informative 2 keywords from one section of each similar patent (see Methods for a walkthrough example of this process). Figure 1 summarizes the results of this experiment. In the figure, the X axis shows that length of the query (see Methods for the pruning strategy used to create short queries); the Y axis shows Mean Average Precision (MAP) for the top 50 patents retrieved by the Boolean queries.[3] Intuitively, MAP illustrates the overlap between

[1] https://patents.google.com
[2] We used the Whoosh indexing and searching library: https://whoosh.readthedocs.io/en/latest/indexing.html
[3] The results of the Boolean search, which are not ranked by default, were ranked using the tf.idf similarity function in Whoosh.

the top *k* patents retrieved by the Boolean query and the original *k* patents retrieved by AI. Figure 1 indicates that this simpler setting yields results that overlap up to approximately 42% with the AI search and the overlap between Boolean and AI searches increases as the Boolean queries increase in size. The top three most informative patent sections are: title, CPC codes, and description.

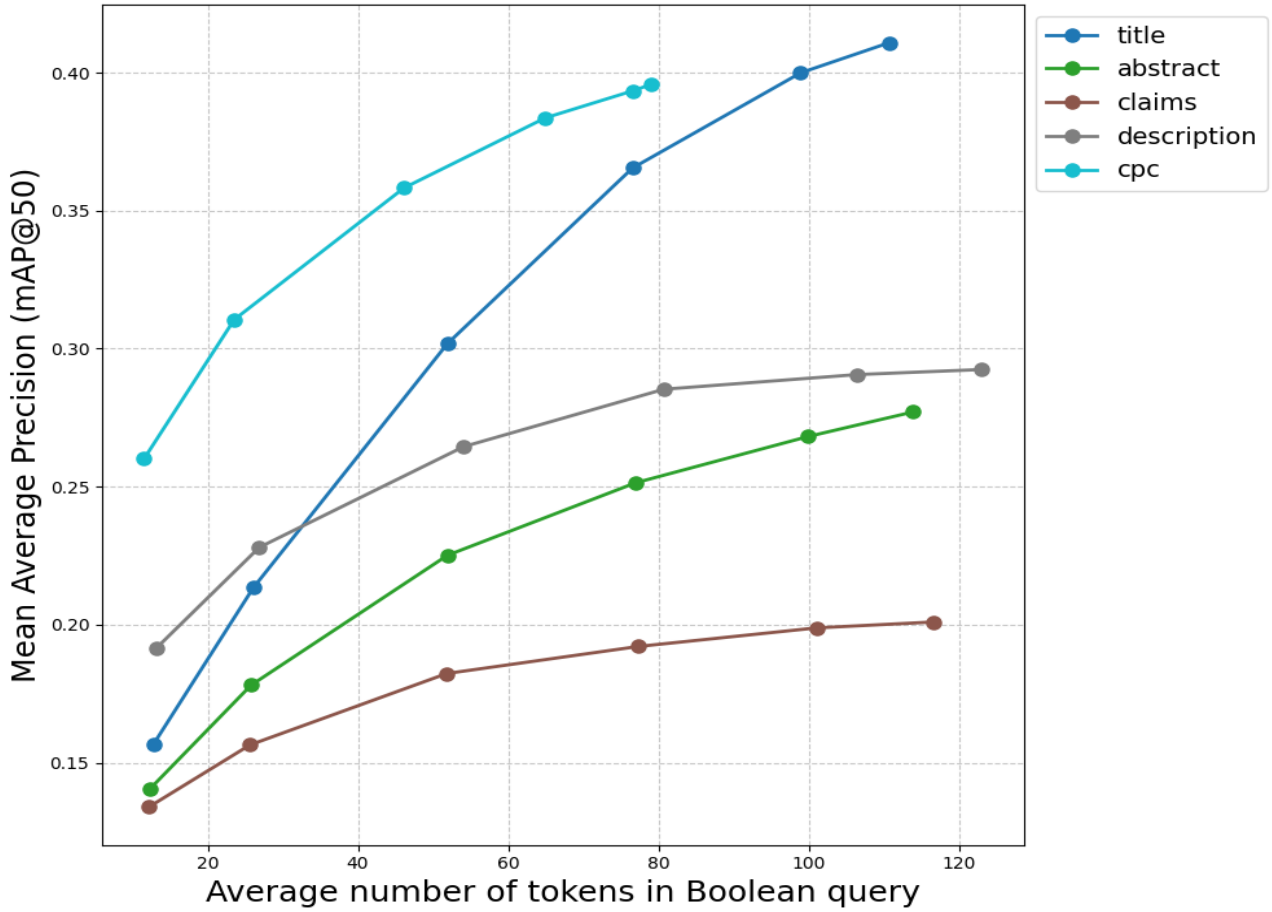

Figure 1: Mean Average Precision (MAP) of the top 50 results for Boolean queries generated from the *k* similar patents retrieved by AI search. Here, keywords are extracted from one section of each patent (e.g., title).

In the second experiment, we used the most informative 2 keywords from *combinations* of two sections from each patent. Figure 2 shows the results of this experiment under the same conditions. These results indicate that combining patent sections is beneficial, with the best performance obtained by combining keywords from the title and description. This configuration obtains 54% overlap with the results of AI search, a 12% increase over the best results obtained using keywords from individual sections (Figure 1).

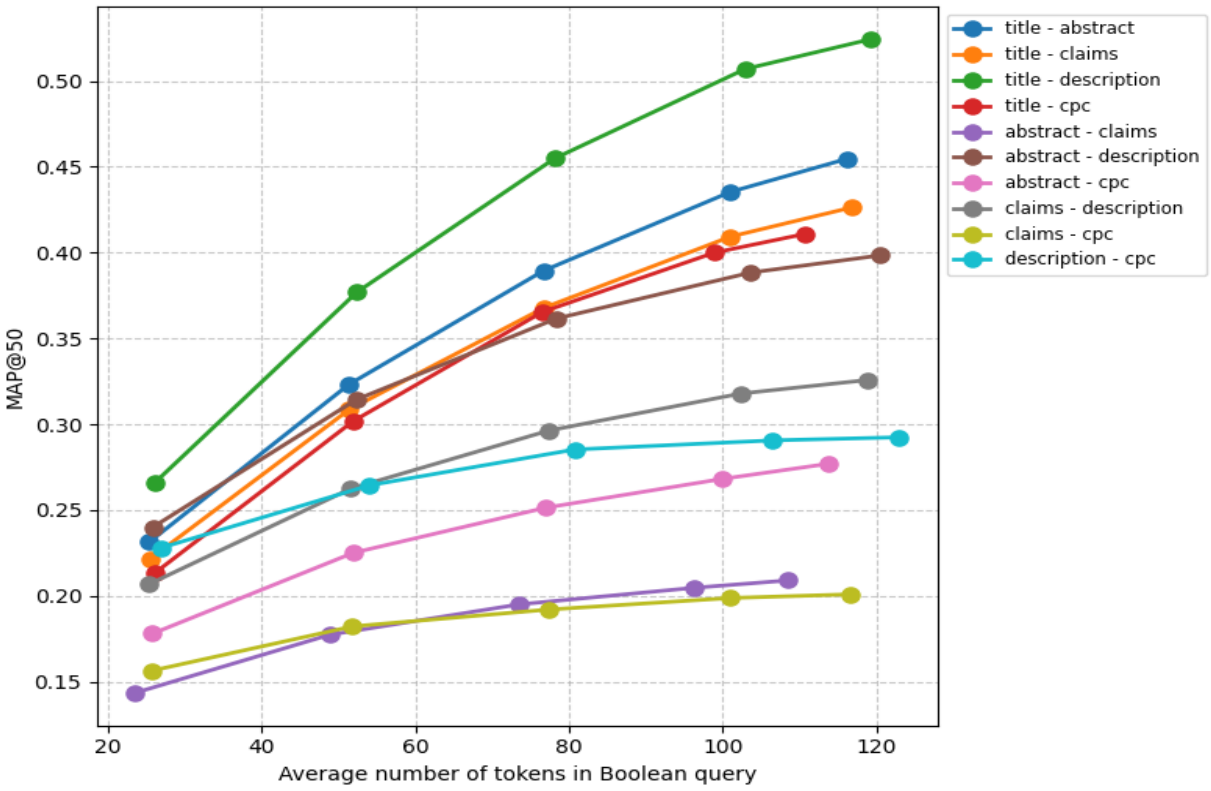

Figure 2: Mean Average Precision (MAP) of the top 50 results for Boolean queries generated from the *k* similar patents retrieved by AI search. Here, keywords are extracted from combinations of two sections of each patent (e.g., “title - description” contains keywords extracted from both the title and description of each patent).

Scenario 2: Replacing AI Search

In this scenario, we generated a Boolean query *only* from the patent of interest in order to verify whether Boolean search can replace AI-driven prior art search. Similar to the previous scenario, in the first experiment we generated Boolean queries using keywords from a single patent section. Figure 3 shows the results of this experiment. These results indicate that the best-performing keywords come from the CPC and description sections. However, the MAP scores indicate that all these results have minimal overlap with the AI results.

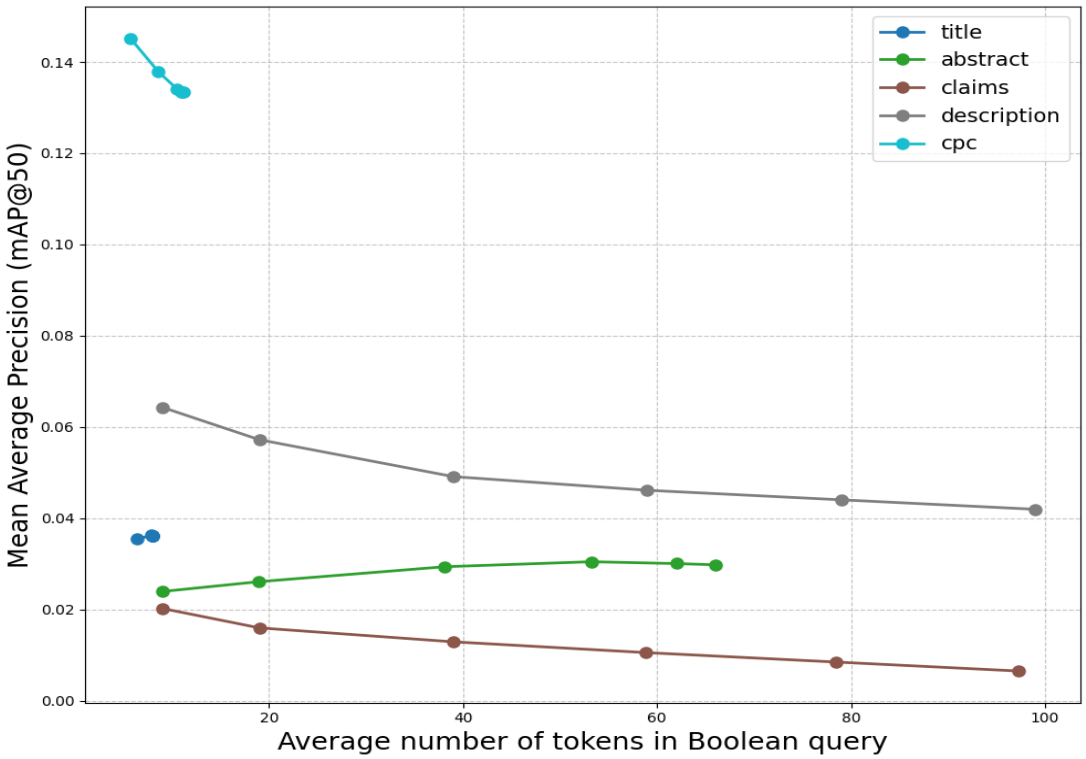

Figure 3: Mean Average Precision (MAP) of the top 50 results for Boolean queries generated solely from the patent of interest. Here, query keywords are extracted from one section of the target patent (e.g., title) and combined in a disjunctive query.

In the second experiment, we generated queries from two patent sections. For each query, 50% of the keywords come from one of the five patent sections. These results are summarized in Figure 4. While this figure indicates a higher overlap with AI results, the overall overlap percentage remains small.

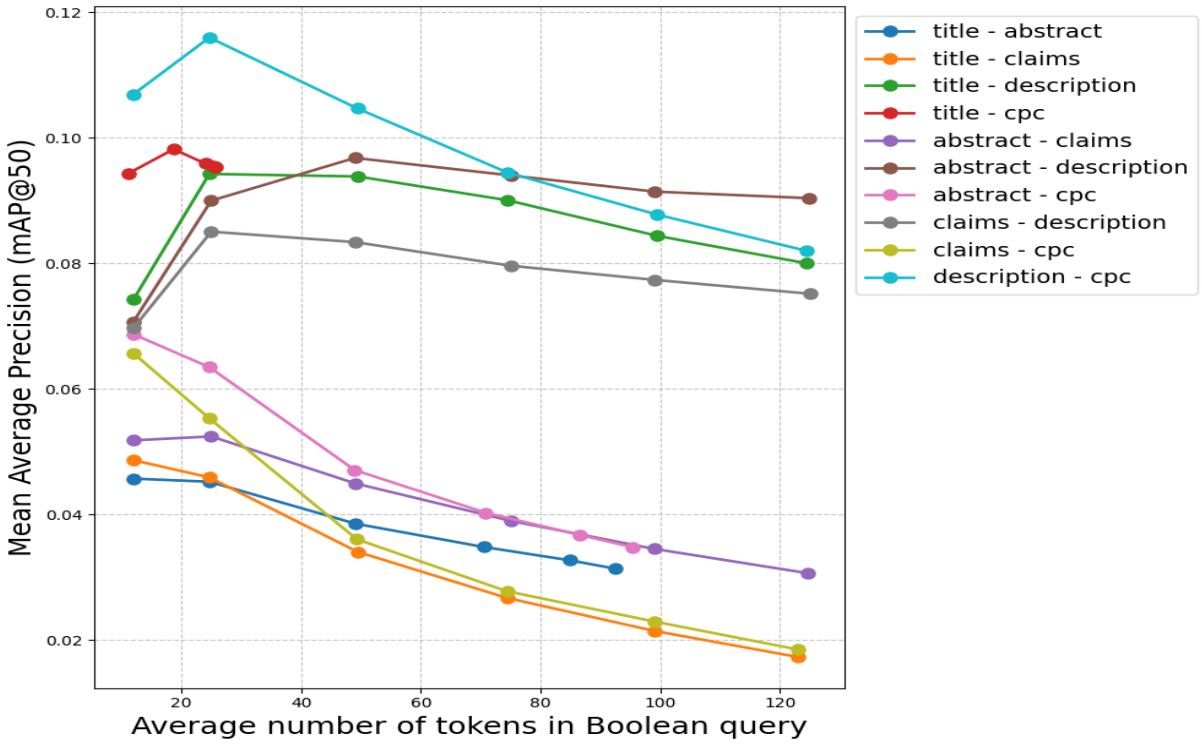

Figure 4: Mean Average Precision (MAP) of the top 50 results for Boolean queries generated solely from the patent of interest. Here, query keywords are extracted from two sections of the target patent (e.g., "title-abstract" indicates that the corresponding query uses keywords from the title and abstract sections). Within each section, keywords are paired using conjunction (AND), and the pairs are then combined using disjunction (OR).

Discussion

Based on the above experiments, we answer our two research questions as follows:

RQ 1: Can Boolean search concur/explain the results of an AI search through a single Boolean query? Answer: mostly yes.

The results in Figures 1 and 2 indicate that a (small) majority of the AI search results can indeed be explained through a Boolean query. This is an encouraging finding, as this direction offers transparency and interpretability to the opaque AI search systems. By leveraging Boolean logic to replicate the outputs of AI-based patent searches, our system bridges the gap between traditional methods familiar to patent professionals and the advanced capabilities of modern more approachable AI search tools. The best performance is obtained when the system generates structured Boolean queries based on keywords extracted from different sections of patents. These queries are designed to emulate the retrieval behavior of AI models while providing clear, human-readable search criteria. By doing so, the system empowers patent professionals to validate and

refine search results, ensuring that important contextual or technical nuances are not overlooked. This interpretability is particularly valuable in legal and professional settings, where the reasoning behind decisions must be clear and justifiable. While Boolean queries provide transparency, the study also highlights their limitations in scenarios where contextual relationships are essential but not readily available. AI models excel in capturing deeper semantic and abstract connections that go beyond the explicit keywords used in Boolean logic. Despite these limitations, the complementary strengths of Boolean queries and AI models present an opportunity for integration. Boolean queries ensure traceability and human oversight, while AI systems provide sophisticated retrieval capabilities, especially when contextual understanding is required. Our system offers a balance between interpretability and performance in the design of patent search tools.

Diving deeper into these results, we observe that the best queries combine keywords from the patent titles and descriptions. We hypothesize this is because the title offers a concise summary of each patent, while the description provides richer context, complementing the title with additional subject-specific elements. This combination enhances the model's ability to capture both broad and specific patent characteristics, resulting in more accurate retrieval. In contrast, claims performed the worst among all sections. This result, while at first glance may seem surprising, is likely because claims use specialized legal language that, while essential for defining the patent's scope, is less effective for retrieval purposes due to its non-standard vocabulary.

Further, both Figures 1 and 2 indicate that increasing the number of tokens in a Boolean query generally improves overall performance. However, this improvement comes with a trade-off. A longer query can reduce interpretability, as the added complexity makes it harder for users to understand and evaluate the query's structure and logic. In our method, query length is an adjustable parameter. That is users can set a maximum length based on their specific needs or comfort levels. In our experiments, we limited each subquery to a maximum of two keywords per patent to control for overall query length. This choice achieved a manageable balance between performance and interpretability, enabling effective retrieval while keeping the queries straightforward. While this approach was suitable for our study, users with different priorities may prefer adjusting the query length to achieve their desired trade-off between precision and interpretability.

RQ 2: Can Boolean search produce equivalent results and therefore replace an AI Search? Answer: no.

The results in Figures 3 and 4 indicate that, when only the target patent is provided, Boolean queries produce outputs that have little overlap with the results of AI search. Without the contextual guidance provided by similar patents, this approach lacks the deeper semantic understanding required to identify subtle similarities between patents. As a result, Boolean queries that rely solely on keywords extracted from the target patent result in poor retrieval performance. In contrast, AI search successfully leverages its advanced contextual understanding to extract rich semantic information from the target patent, enabling it to identify meaningful similar patents. AI models overcome the limitations of keyword-based approaches by

capturing implicit relationships and abstract concepts. As a result, an AI-based search system produces better retrieval performance without the need for explicit input from similar patents.

All in all, these findings highlight the complementary strengths of Boolean queries and AI search models. Boolean queries provide interpretability and transparency, making them valuable for explaining AI-driven search results. On the other hand, AI models excel in retrieval tasks, especially when contextual information must be derived from the target patent alone. Our method bridges the gap between these two approaches by combining the interpretability of Boolean queries with the contextual capabilities of AI models.

While this study demonstrates the potential of Boolean queries to explain AI-driven patent search results, there are several avenues for future exploration to enhance both the effectiveness and applicability of the proposed approach.

- *Incorporating semantic context:* The current method relies on keyword extraction from specific sections of patents, which can overlook deeper semantic relationships. Future work could integrate semantic embedding models or knowledge graphs to capture latent connections between patents, complementing the Boolean logic with richer contextual understanding.
- *Dynamic query optimization:* While the study focuses on static query formulation, introducing dynamic query optimization and adapting queries based on real-time feedback from retrieval performance could improve average precision. Techniques such as reinforcement learning could be employed to refine queries iteratively.
- *Exploration of multimodal inputs:* Patent documents often include images, diagrams, and non-textual metadata that are crucial for understanding the scope and novelty of an invention. Future systems could incorporate multimodal data to enhance query generation and retrieval accuracy.

By advancing this interplay between AI and explainability, we can pave the way for more robust, reliable, and transparent tools that support innovation in intellectual property growth and management in the rapidly evolving technological landscape.

Methods

This effort aims to answer the two research questions listed in *Introduction*, i.e., explaining the prior art results of an AI search with a single Boolean query, and replacing AI search with Boolean search. We discuss our approaches for answering these two questions below.

Scenario 1: Explaining the Results of AI Search with Boolean Queries

The overall architecture for this scenario is depicted in Figure 5. In this setting, our system takes as input: (a) a *target patent*—i.e., the patent for which we seek to retrieve prior art—and (b) a set of *k* similar patents, identified as *prior art* by an AI-powered search of a large patent repository.

Using this information, our Query Generator constructs a Boolean query designed to replicate the AI search's functionality—specifically, to retrieve results that closely match the *k* patents identified by the AI. For example, given a set of patents related to smartphones, our approach might generate a query such as:

```
('smartphone' OR 'mobile phone' OR 'cell phone') AND ('touchscreen' OR
'wireless communication') AND NOT ('landline' OR 'pager').
```

Finally, the patents retrieved via this Boolean search are compared against the AI search results by our Query Evaluator component. We describe these two components in detail below.

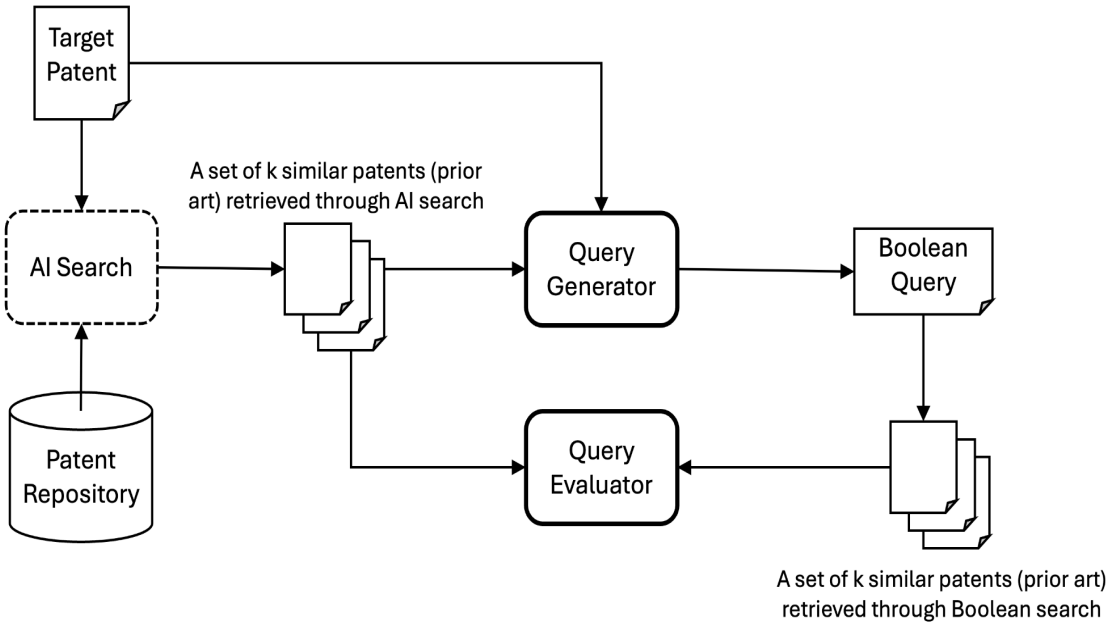

Figure 5: High-level view of the proposed architecture. The system takes as input a target patent and a set of *k* similar patents (prior art) retrieved using AI search, and produces a Boolean query designed to retrieve the same *k* patents using Boolean search.

Query Generator

The architecture of our query generation component is illustrated in Figure 6. As shown, our method begins by extracting keywords from each of the similar patents, followed by constructing individual subqueries. These subqueries are then combined into a single Boolean query that represents the entire set of patents. Finally, we optimize the query by reducing keyword redundancy. We detail these steps below.

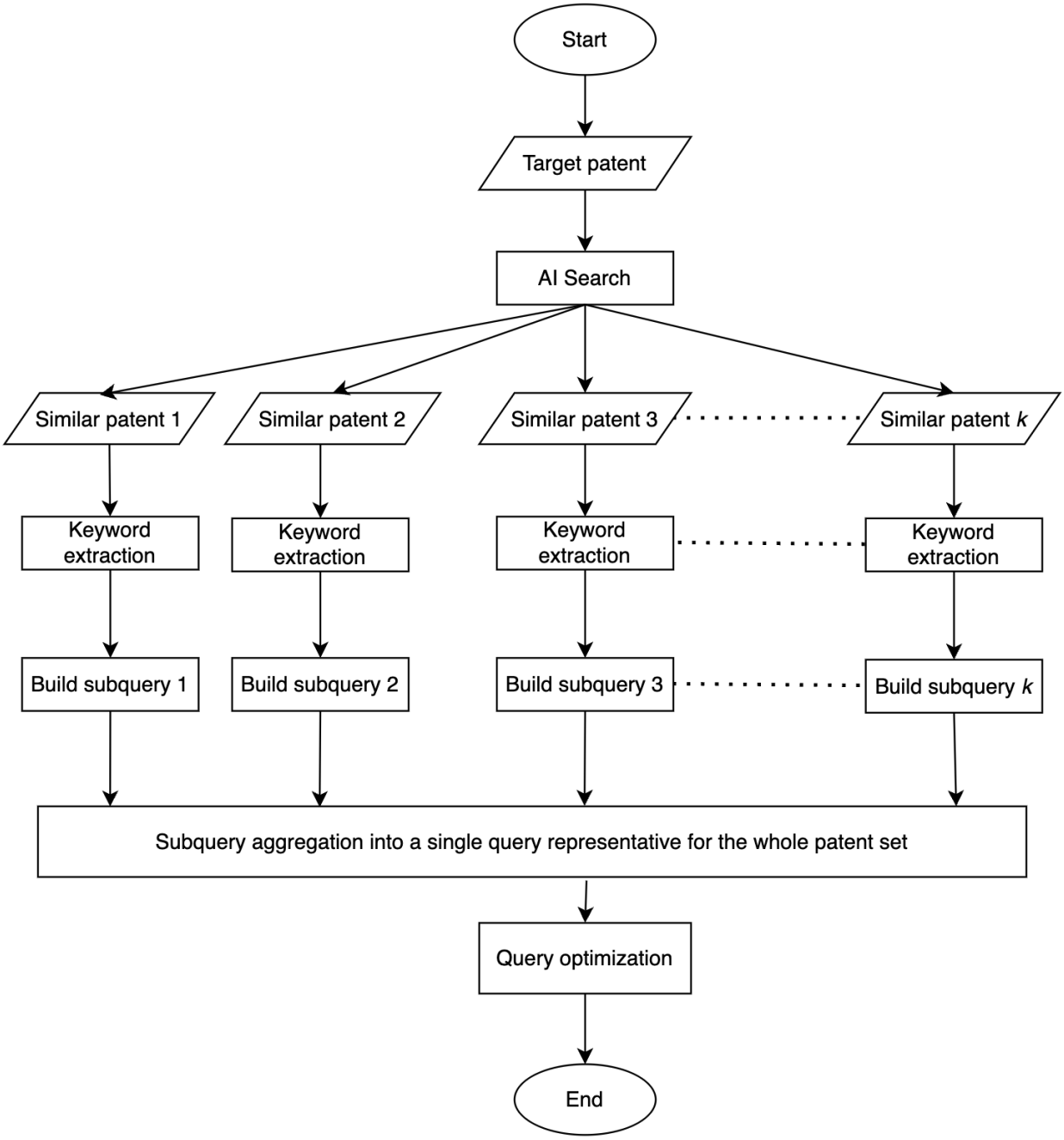

Figure 6: A detailed view of the query generator component. This component extracts representative keywords from each of the prior art patents, constructs individual subqueries from them, and then aggregates the corresponding subqueries into a single query representative of the whole set.

Step 1: Keyword extraction

The first step in our method is to extract keywords from each of the *k* patents previously retrieved by the AI search. To ensure that the extracted keywords are meaningful and representative of the content, we perform the following tasks:

- *Stop word removal:* We begin by removing common stop words, i.e., words such as “the,” “is,” “and” that provide little semantic signal, from each patent document.
- *Keyword ranking by tf.idf score:* After removing stop words, we compute the term frequency-inverse document frequency (tf.idf) score for each word in the document. The tf.idf score helps identify the most important words within the document by considering both the frequency of the word in the specific document and its rarity across the whole patent dataset[23]. Words with higher tf.idf scores are considered more relevant for representing the content of the patent.
- *Top keyword selection:* Based on the computed tf.idf scores, we extract the *N* keywords with the highest scores from each patent. These top-scoring keywords serve as the foundation for constructing the individual subqueries in the next step.

Step 2: Subquery construction

Once we have identified the key terms from similar patents, the next step is to construct a subquery that represents each patent. We do this by simply connecting the extracted keywords from the respective patent using the **AND** operator. This approach ensures that each subquery is specific enough to retrieve the given patent while also allowing for the retrieval of additional patents that share the same set of keywords.

Example: If the top keywords for a particular patent are *“smartphone,” “wireless,”* and *“camera,”* the corresponding subquery would be: “smartphone” AND “wireless” AND “camera”.

Step 3: Subquery combination

After constructing individual subqueries for each similar patent, the next step is to combine them into a single Boolean query that represents the entire set of patents retrieved by the AI search. To maximize the likelihood of retrieving all similar patents for the given target patent, we use the **OR** operator to connect the subqueries. The final query takes the form of multiple subqueries joined by OR, for example: (subquery$_1$) OR (subquery$_2$) OR (subquery$_3$) OR …

Step 4: Query optimization

The final step in our method is to optimize the combined Boolean query by minimizing redundant keywords caused by overlapping content in the input patents. Specifically, if two subqueries share common keywords, we restructure the query to eliminate unnecessary repetition. To optimize the Boolean query, we identify common keywords across subqueries and restructure the query using the distributive law of Boolean algebra. The distributive law states that performing the AND (multiplication) operation on a group of elements connected by the OR (addition) operation is equivalent to applying the AND operation to each element individually and then combining the results using OR. Formally, for any Boolean variables *A, B,* and *C*, the expression *A(B+C)* is logically equivalent to *AB+AC*. In our case, whenever we encounter a Boolean expression of the form *AB+AC*, we transform it into *A(B+C)* by identifying and factoring out the common terms. For instance, the query: (‘wireless’ AND ‘smartphone’ AND ‘antenna’) OR (‘wireless’ AND

‘smartphone’ AND ‘camera’) is simplified to: ‘wireless’ AND ‘smartphone’ AND (‘antenna’ OR ‘camera’).

This optimization reduces query complexity by eliminating redundant terms, shortening the query expression, and decreasing the number of logical operations required for evaluation. Search queries with fewer Boolean operations generally execute faster in the Boolean search system. A shorter and well-structured query is easier to interpret, thereby enhancing explainability.

Step-by-step example of the query generation process:

The process begins with a target patent and its *k* most similar patents, retrieved using AI search techniques. From each similar patent, the top *n* keywords are extracted to construct a subquery using the AND operator. Each subquery is designed to retrieve at least its corresponding similar patent. In the example below, the prefix **‘ti’** indicates a keyword from the title, while **‘detd’** indicates a keyword from the description. We illustrate the example using $k = 50$ and $n = 2$, though the method is flexible and works for any values of *k* and *n*.

Example:
Target patent: *US-2022196839-A1*
Title: *Procedurally generated three-dimensional environment for use in autonomous vehicle simulations*

We extract the top 2 keywords from each of the 50 similar patents and construct a subquery using the AND operator.

- **Similar patent 1:** *US-2023192136-A1*
  Title: *Filtering real-world objects for inclusion in a simulated environment*
  Top Keywords: ('ti:generate', 'detd:ridesharing')
  Subquery 1: ('ti:generate' AND 'detd:ridesharing')

- **Similar patent 2:** *US-2022358317-A1*
  Title: *Automatic detection of roadway signage*
  Top Keywords: ('ti:signage', 'detd:ridesharing')
  Subquery 2: ('ti:signage' AND 'detd:ridesharing')

- **Similar patent 3:** *US-2023186560-A1*
  Title: *Mapping data to generate simulation road paint geometry*
  Top Keywords: ('ti:map', 'detd:ridesharing')
  Subquery 3: ('ti:map' AND 'detd:ridesharing')
  ……

- **Similar patent 50:** *US-2023176200-A1*
  Title: *Deriving surface material properties based upon lidar data*

Top Keywords: ('ti:deriving', 'detd:ridesharing')
Subquery 50: ('ti:deriving' AND 'detd:ridesharing')

These 50 subqueries are then combined using the **OR** operator to form a single query.

Combined Query: *('ti:generate' AND 'detd:ridesharing') OR ('ti:signage' AND 'detd:ridesharing') OR ('ti:map' AND 'detd:ridesharing') OR .....OR (('ti:deriving', 'detd:ridesharing')*

Finally, an optimization step is applied to reduce query length by eliminating redundant keywords such as `ridesharing'.

Optimized query: *(detd:ridesharing (ti:generate OR ti:sinage OR ti:map OR ..... OR ti:deriving)) OR .....)*

Query Evaluator

The query evaluator compares the results obtained by the Boolean search to determine the overlap with the *k* most similar patents retrieved by the AI search. The query evaluator uses mean average precision (MAP)[23] as the evaluation metric. MAP measures both relevance and ranking to assess how well the Boolean queries replicate AI search results. First, it calculates the average precision (AP) of each query by considering the positions of relevant patents in the ranked list.

For example, consider a simple ranked list of search results as follows:

- AI search results (i.e., ground truth in our case): [*P1, P2, P3, P4, P5*]
- Boolean search result (retrieved by query): [*P3, P6, P1, P5, P7*]

We calculate the AP for the Boolean search results using the following equations:

$$Precision\ at\ rank\ i\ = \frac{number\ of\ relevant\ items\ retrieved\ up\ to\ rank\ i}{total\ number\ of\ items\ retrieved\ up\ to\ rank\ i}$$

$$AP\ = average\ of\ the\ precision\ values\ at\ rank\ i\ for\ all\ relevant\ items$$

In our example, the Boolean search retrieved three relevant patents (i.e., patents that also appear in the ground truth) at positions 1, 3, and 4 (*P3, P1,* and *P5*). First, we compute the precision at these ranks, and then, using the precision values, we calculate the AP.

Precision at rank 1: 1/1 = 1.0

Precesion at rank 3: ⅔ = 0.667

Precision at rank 4: ¾ = 0.75

AP = (1.0+0.667+0.75)/3 = 0.806

The final MAP score is then obtained by averaging the AP values across all queries in the test set. Intuitively, a higher MAP score indicates a stronger alignment between Boolean and AI search results. All calculations are performed for the top *k* ranked items, a metric commonly referred to in information retrieval as mean average precision at *k* (MAP@k).

Scenario 2: Replacing AI Search with Boolean Search

As discussed, this scenario explores whether Boolean search can fully *replace* AI search. Here, only the target patent is provided, and we construct a query using keywords extracted exclusively from it. Unlike the previous approach, this method does not rely on the *k* similar patents retrieved by the AI search system. Instead, it attempts to retrieve them directly using Boolean search with the generated query.

Aside from using the target patent instead of the *k* similar patents, the system components—i.e., the Query Generator and Query Evaluator—remain identical to those in the previous scenario.

Competing Interests:

All authors declare no competing financial and/or non-financial interests in relation to the work described.

References

1. U.S. Constitution - Article I | Resources | Constitution Annotated | Congress.gov | Library of Congress. https://constitution.congress.gov/constitution/article-1/.
2. Lanham (Trademark) Act (15 U.S.C.) Index, January 2023 (BitLaw). https://www.bitlaw.com/source/15usc/index.html.
3. Greenhalgh, C. & Rogers, M. *Innovation, Intellectual Property, and Economic Growth*. (Princeton University Press, 2010). doi:10.2307/j.ctt1zgwjjb.
4. Gambardella, A. Private and social functions of patents: Innovation, markets, and new firms. *Res. Policy* **52**, 104806 (2023).
5. Yin, J. *et al.* Intellectual property protection as catalyst for radical technological innovation in national research program teams through innovation milieu and group potentials. *Sci. Rep.* **14**, 25038 (2024).
6. Menell, P. S., Lemley, M. A., Merges, R. P. & Balganesh, S. *Intellectual Property in the New Technological Age:2023*. vol. I (Clause 8 Publishing, 2023).
7. Merges, R. P. & Duffy, J. F. *Patent Law and Policy: Cases and Materials*. (Carolina Academic Press, 2021).
8. Gorman, R. A., Ginsburg, J. C. & Reese, R. A. *Copyright: Cases and Materials*. (West Academic, 2024).
9. Hourihan, M. & Budget, D. *Public Research Investments and Patenting: An Evidence Review*. 14 https://www.aaas.org/sites/default/files/2020-05/AAAS%20Public%20Research%20and%20Patenting%20FINAL.pdf.
10. *2024 PPAC Annual Report*. 76 https://www.uspto.gov/sites/default/files/documents/ppac-2024-annual-report.pdf (2024).
11. World Intellectual Property Indicators 2024: Highlights - Patents Highlights. https://www.wipo.int/web-publications/world-intellectual-property-indicators-2024-

highlights/en/patents-highlights.html.

12. Search for patents. https://www.uspto.gov/patents/search.

13. Basics of Prior Art Searching. https://www.uspto.gov/sites/default/files/documents/Basics-of-Prior-Art-Searching.pdf

14. Patent process overview. https://www.uspto.gov/patents/basics/patent-process-overview.

15. Resources, M. MPEP. *2152 Detailed Discussion of AIA 35 U.S.C. 102(a) and (b) [R-11.2013]* https://www.uspto.gov/web/offices/pac/mpep/s2152.html.

16. Google Patents. Patent search engine. https://patents.google.com/.

17. PQAI. Patent Search Tool | PQAI. https://search.projectpq.ai/.

18. IPRally | AI Patent Search, Review & Classification. https://www.iprally.com/.

19. Patentfield. Patentfield | AI Patent Search, Analytics and investigating database for Japan and US. https://en.patentfield.com/.

20. Thomson Reuters. Search with Terms and Connectors. https://www.thomsonreuters.com/content/helpandsupp/en-us/help/westlaw-edge/searching/search-with-terms-and-connectors.html.

21. Google Patents. Google Patents Research Data.

22. Scott Beliveau *et al.* USPTO - Explainable AI for Patent Professionals. https://kaggle.com/uspto-explainable-ai.

23. Christopher D. Manning, Hinrich Schütze, & Prabhakar Raghavan. *Introduction to Information Retrieval*. (Cambridge University Press, 2008).