# A Survey on Patent Analysis: From NLP to Multimodal AI

Anonymous ACL submission

## Abstract

Recent advances in Pretrained Language Models (PLMs) and Large Language Models (LLMs) have demonstrated transformative capabilities across diverse domains. The field of patent analysis and innovation is not an exception, where natural language processing (NLP) techniques presents opportunities to streamline and enhance important tasks—such as patent classification and patent retrieval—in the patent cycle. This not only accelerates the efficiency of patent researchers and applicants, but also opens new avenues for technological innovation and discovery. Our survey provides a comprehensive summary of recent NLP-based methods—including multimodal ones—in patent analysis. We also introduce a novel taxonomy for categorization based on tasks in the patent life cycle, as well as the specifics of the methods. This interdisciplinary survey aims to serve as a comprehensive resource for researchers and practitioners who work at the intersection of NLP, Multimodal AI, and patent analysis, as well as patent offices to build efficient patent systems.

## 1 Introduction

The growing complexity and volume of textual data across various domains have driven significant advancements in NLP, particularly through PLMs (Devlin et al., 2019) and LLMs (Radford et al., 2019). The field of patents and technological innovation is not an exception. This advancement can streamline complex patent-related tasks such as classification, retrieval, and valuation prediction. For instance, for patent examination, patent offices often rely only on the examiner to judge whether a technology is innovative enough and, thus, patentable. However, it is challenging for the human examiner to stay updated on various domains due to the exponential growth in technology and apply the knowledge during evaluation. This intersection of NLP, Multimodal AI, and patent processes can accelerate the efficiency of the patent systems—patent reviewers as well as applicants—and help in a faster technological innovation to benefit our society.

The patent application and granting process involves complex textual analysis tasks that require significant human effort for both applicants and reviewers. To streamline this, NLP techniques can be helpful, particularly in patent classification, retrieval, and quality analysis (Krestel et al., 2021). Patent classification can benefit from multi-label classification tools for the hierarchical schemes: International Patent Classification (IPC) and the Cooperative Patent Classification (Roudsari et al., 2022; Althammer et al., 2021). To evaluate novelty and avoid infringement, the patent retrieval task becomes important while filing or reviewing a new patent application. On the other hand, quality analysis also requires a substantial amount of effort. NLP-based representation learning methods can be useful in both tasks (Chung and Sohn, 2020; Lin et al., 2018). Lastly, recent advanced LLMs can generate accurate and technical language descriptions for patents and, thus, are useful to optimize human resources and precision in patent writing (Lee and Hsiang, 2020a).

The existing patent surveys in the literature do not cover the recent studies in this area and fail to show the trends and methods in task specific manner. We introduce a novel taxonomy to categorize the methods based on the relevant tasks and the nature of the methods. Our taxonomy provides an in-depth view of the methods being used in specific tasks. Moreover, it captures the recent trends of using advanced methods (e.g., LLMs) that are missing from the existing surveys. This will be beneficial for researchers who aim to build task-specific methods.

**Overview.** Fig. 1 provides the hierarchical organization of patent tasks and methods. We organize the survey as follows: Sec. 2 provides background, Sec. 3 summarizes the methods for individual tasks, and Sec. 4 provides future research directions.

## 2 Background

A patent grants the owner or holder exclusive rights to an invention and can be a novel product or a process that usually offers a unique method or technical solution. In exchange for this right, inventors must publicly disclose detailed information about their invention in a patent application. The United States Patent and Trademark Office (USPTO[1]) issues three types of patents: utility, design, and plant. In this work, we focus on utility and design patents, considering their importance in innovation across industries. Utility patents protect the rights related to how the invention works or is used. It provides the entitlement to the functionality of a product. On the other hand, design patents protect the right of the look of an invention and are intended to safeguard the form of a product. Here, we outline the relevant tasks.

**Formulation.** We provide the problem formulations of these patent tasks in Appendix A.3.

**Datasets.** We describe the common benchmark patent datasets in Appendix A.6.

### 2.1 Patent Classification

Patent classification is an important but time-intensive task in the patent life cycle (Grawe et al., 2017; Shalaby et al., 2018; Risch and Krestel, 2018). This involves a multi-label classification for patents where the classification scheme is hierarchical, and a patent can get multiple labels. There are two widely used patent classification systems: International Patent Classification (IPC) and the Cooperative Patent Classification (CPC). The IPC comprises 8 sections, 132 classes, 651 subclasses, 7590 groups, and 70788 subgroups in a hierarchical order (i.e., sections have classes and classes have subclasses, and so on). CPC is an expansion of IPC and is collaboratively administered by the European Patent Office (EPO) and the USPTO. It consists of around 250,000 classification entries and is divided into nine sections (A-H and Y), which are further broken down into classes, subclasses, groups, and subgroups[2]. Table 7 (see Appendix) shows an example of CPC classification.

**Challenges.** Patent classification is challenging due to its multi-class and multi-label nature. A single patent can be assigned multiple CPC/IPC codes, which makes the classification process complex. Additionally, the hierarchical structure of patent taxonomies introduces dependencies that require models to capture relationships between broad and fine-grained categories. Moreover, patent documents have various sections such as titles, abstracts, and claims—each contains different information. Given the extensive length of these full-text patent documents, identifying the most relevant sections for classification also poses a significant challenge.

[1]https://www.uspto.gov/

[2]https://www.cooperativepatentclassification.org/

### 2.2 Patent Retrieval

Patent Retrieval (PR) (Kravets et al., 2017; Kang et al., 2020; Chen et al., 2020; Setchi et al., 2021) focuses on developing methods to efficiently retrieve relevant patent documents and images based on specific search queries. PR plays a crucial role in identifying new patents related to new inventions. It is essential for evaluating novelty of a patent as well as ensuring that it does not infringe on existing patents. Moreover, patent image retrieval can serve as a source of inspiration for design.

**Challenges.** Patent retrieval tasks involve both text and image retrieval with unique challenges. Text retrieval is complex due to the use of similar words to describe new inventions; an invention can be described using various synonyms and phrasings which make it difficult to retrieve crucial information for patent infringement analysis. On the other hand, image retrieval is particularly challenging due to the nature of the images involved, which are typically black and white sketches, including numbers to describe the inventions.

### 2.3 Patent Quality analysis

Businesses have shown great interest in evaluating patent value due to its significant impact in generating revenue and investment (Aristodemou, 2021). Investors usually aim to predict the future value of technological innovation from the target firm while making investment decisions. As a result, many companies hire professional patent analysts for quality analysis. This complex task demands substantial human effort as well as expertise in various domains (Lin et al., 2018). The quality of a patent can be assessed using various measures, including the number of forward or backward citations, the number of claims, the grant lag, patent family size, the remaining lifetime of the patent (Aristodemou, 2021; Erdogan et al., 2022).

**Challenges.** The challenge in analyzing patent quality is the ambiguity of the metrics to quantify

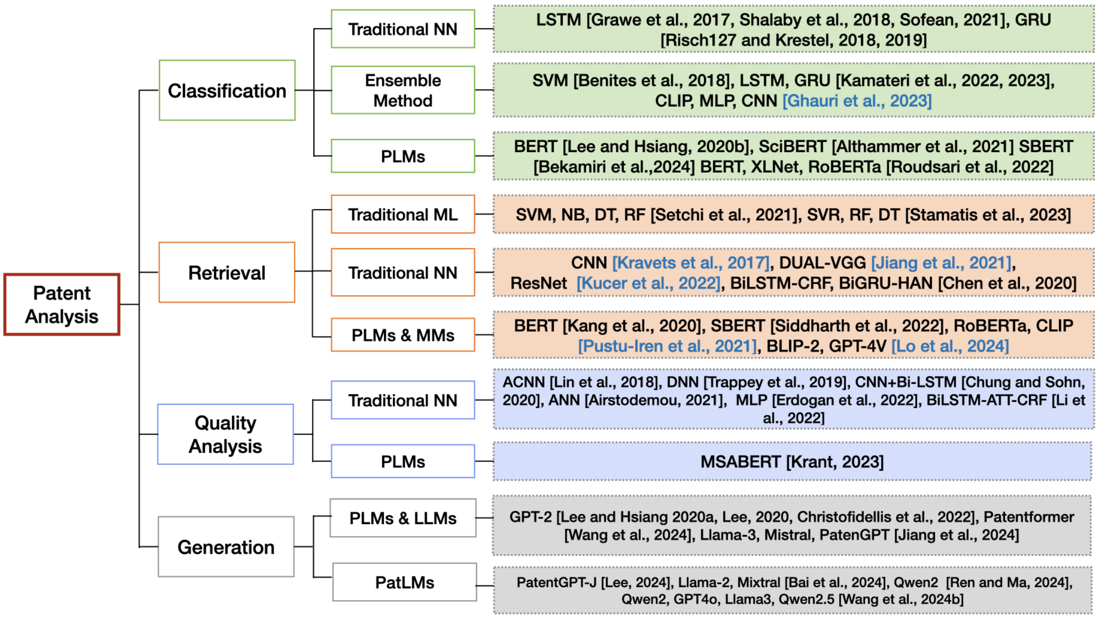


Figure 1: The schema of the main organization with the methods in each patent-related task. We summarize the methods for four individual tasks: patent classification, retrieval, quality analysis, and generation. “NN”, “MMs”, “PLMs”, and “PatLMs” denote neural networks, multimodal models, pre-trained language models, and patent language models, respectively. The works that use images are in blue.

the quality of a patent. Commonly used measures for the quality analysis are the number of citations (both forward and backward), the number of claims, and the grant lag. However, the weight of each of these measures remains unclear. Moreover, analyzing these information to perform a comprehensive study is non-trivial.

### 2.4 Patent Generation

Patents usually require a considerable amount of written text, which requires significant human resources. The patent generation task involves generating specific sections of a patent, such as abstract, independent claims, and dependent claims, based on instructions for an AI tool. Patent documents require precise and technical language to accurately describe the invention and its claims (Risch et al., 2021). AI-assisted patent generation will help automate the drafting process, which involves time, effort, and legal requirements. This will also reduce the amount of patent attorney time which will be a substantial cost saver.

**Challenges.** Though the patent document has certain structures, one major challenge is to evaluate the dependency—which can help in patent generation—among the parts of the patent. For instance, one part (e.g., abstract, claims) can be used as an input in a generative model (e.g., a LLM) to generate a different part of the patent. Additionally, it becomes non-trivial to construct effective instructions or prompts that guide the generation process. The generation also brings the question of evaluation of the generated content or text, i.e., how to judge whether the generated content is desired or not appropriate.

## 3 Methods

We organize the important patent tasks that can benefit from recent advancements in NLP and Multimodal AI. An overview of important patent tasks is shown in Figure 2 (Appendix A.1). The frequently used AI methods in the papers covered by this survey are in Table 6.

### 3.1 Patent Classification

In the literature, several models have been used to automate this process. We organize them based on the nature of the method into three major categories. Table 1 represents a summary of the methods for patent classification. We present the evaluation metrics and the results in Table 8 in Appendix A.4.

#### 3.1.1 Traditional Neural Networks

The commonality among these methods is that they follow a two-step approach: generate initial features and then use a classifier for the final classification. One of the initial studies (Grawe et al., 2017) implements a single-layer LSTM to classify

Table 1: Studies on patent classification. Hierarchy levels for classification include Section, Class (white), Subclass (blue), Group, and Subgroup (grey). The color green represents the category of visualizations. Table 8 provides more details on the performance in the Appendix.

| Papers | Embeddings | Methods | Components |
|---|---|---|---|
| (Grawe et al., 2017) | Word2Vec | Single layer LSTM | Description |
| (Shalaby et al., 2018) | Fixed Hierarchy Vectors | LSTM | ADC |
| (Risch and Krestel, 2018) | FastText | GRU | Full text |
| (Benites et al., 2018) | TF-IDF | SVM | Single Text Block |
| (Risch and Krestel, 2019) | FastText | GRU | Full text |
| (Lee and Hsiang, 2020b) | – | BERT-base | Claim |
| (Althammer et al., 2021) | – | BERT, SciBERT | Claim |
| (Sofean, 2021) | Word2Vec | Multiple LSTMs | Description |
| (Roudsari et al., 2022) | Word2Vec, FastText | BERT, XLNet, RoBERTa | Title, abstract |
| (Kamateri et al., 2022) | FastText, Glove, Word2Vec | CNN, LSTM, GRU | TADC |
| (Ghauri et al., 2023) | Vision Transformer | MLP | Image |
| (Kamateri et al., 2023) | FastText | Bi-LSTM, Bi-GRU, LSTM | Metadata |
| (Bekamiri et al., 2024) | SBERT | KNN | Claim, title, abstract |

patents at the IPC subgroup level where the initial features are obtained by the Word2Vec method. Similarly, (Shalaby et al., 2018) use LSTM for IPC subclass level classification. For the initial document representation, the method uses fixed hierarchy vectors that utilize distinct models for various segments of the document. (Risch and Krestel, 2018) and (Risch and Krestel, 2019) focus on training fastText word embeddings on a corpus of 5 million patent documents, then use Bi-GRU for classification. Similarly, (Sofean, 2021) applies text mining techniques to extract key sections from patents, train Word2Vec, and then use multiple parallel LSTMs for the classification task. These collectively show the usefulness of neural networks in patent classification.

#### 3.1.2 Ensemble Models

The models in this category are used to ensemble different word embeddings and deep learning models. (Benites et al., 2018) use SVM as a baseline method and experiment with various datasets, the number of features, and semi-supervised learning approaches. Meanwhile, (Kamateri et al., 2023) and (Kamateri et al., 2022) both investigate ensemble models incorporating Bi-LSTM, Bi-GRU, LSTM, and GRU. More specifically, (Kamateri et al., 2022) conduct experiments with different word embedding techniques, whereas (Kamateri et al., 2023) focus on applying various partitioning techniques to enhance the performance of the proposed framework. While the above methods heavily focus on texts, (Ghauri et al., 2023) classify patent images into distinct types of visualizations, such as graphs, block circuits, flowcharts, and technical drawings, along with various perspectives, including side, top, left, and perspective views. The approach utilizes the CLIP model with Multi-layer Perceptron (MLP) and various CNN models.

#### 3.1.3 Pre-trained Language Models (PLMs)

The first study (Lee and Hsiang, 2020b) which involves PLMs, fine-tune the BERT model on the USPTO-2M dataset and introducing a new dataset, USPTO-3M at the subclass level to aid in future research. Concurrently, (Roudsari et al., 2022) also fine-tune BERT, along with XLNet (Yang et al., 2019), and RoBERTa on the USPTO-2M dataset. They establish XLNet as the new state-of-the-art in classification performance, achieving the highest precision, recall, and f1 measure. (Althammer et al., 2021) implement domain adaptive pre-trained Linguistically Informed Masking and shows that SciBERT-based representations perform better than BERT-based representations in patent classification. SciBERT is pre-trained on scientific literature which helps the method to understand the technical language of patents. (Bekamiri et al., 2024) use Sentence BERT that takes into account entire sentences instead of word by word. On USPTO data, their method gives the highest recall and f1 score.

#### 3.1.4 Discussion and Suggestion

The evaluation measures for patent classification are accuracy, precision, recall, and the f1 score on the CPC or IPC. The earlier works on patent classification are mostly focused on simpler neural networks (Risch and Krestel, 2018, 2019). Applying models such as LSTM can capture the sequence and context in the text, which is suitable for the patent domain since the context is critical. However, these are comparatively simple models that might be limited to capturing complex technical

Table 2: Works on patent retrieval. The papers are white, blue, and gray based on the data type of text, image, and both, respectively. The dataset details are provided in Appendix A.6.

| Work | Method | Training | Datasets |
|---|---|---|---|
| (Kravets et al., 2017) | CNN | supervised | Freepatent, Findpatent |
| (Kang et al., 2020) | BERT | pre-trained | WIPS |
| (Chen et al., 2020) | BiLSTM-CRF, BiGRU-HAN | supervised | USPTO |
| (Jiang et al., 2021) | DUAL-VGG | supervised | - |
| (Setchi et al., 2021) | SVM, Naive Bayes, Random Forest, MLP | supervised | - |
| (Pustu-Iren et al., 2021) | RoBERTa, CLIP | pre-trained | EPO |
| (Siddharth et al., 2022) | Sentence-BERT, TransE | pre-trained, unsupervised | USPTO |
| (Kucer et al., 2022) | (ImageNet, Sketchy) ResNet50 | supervised, finetuned | DeepPatent |
| (Higuchi and Yanai, 2023) | Deep Metric Learning | self-supervised | DeepPatent |
| (Higuchi et al., 2023) | InfoNCE and ArcFace | self-supervised | DeepPatent |
| (Lo et al., 2024) | BLIP-2, GPT-4V | pre-trained, supervised | DeepPatent2 |

structures in patent documentation. This limitation is evident in the evaluation metrics; for instance, the highest accuracy at the subclass level is only 0.74 (Table 8 in Appendix). More advanced techniques, including PLMs, have become popular over time. PLMs could be powerful because of their pre-training step on a massive amount of data. Patent text is different from the usual text in scientific articles (e.g., research papers). Thus, fine-tuning PLMs on patent datasets might be able to address some of these concerns by providing context-aware representations for the patent domain. From Table 8, the early works have a low precision of 0.53 on USPTO data (Risch and Krestel, 2018). PLMs—such as BERT and RoBERTa—have significantly improved the performance to 0.82 (Roudsari et al., 2022). The language models used for classification tasks in the patent domain are generally simpler compared to advanced LLMs such as GPT and LLaMA. There is a significant gap between recent practices in the patent domain and the existing advanced AI models.

## 3.2 Patent Retrieval

We organize the relevant studies below based on the types of methods. Table 2 provides an overview of studies for patent retrieval. We present the results by these methods in Table 9 (Appendix A.4).

### 3.2.1 Traditional Machine Learning

Initial studies have used traditional machine learning methods for patent retrieval. (Setchi et al., 2021) describe five technical requirements to investigate the feasibility of AI for the task. These requirements include query expansion and identification of semantically similar documents. The study uses SVMs, Naive Bayesian learning, decision tree induction, and RF, along with word embeddings, to solve the prior art retrieval problem. Prior art usually implies the references which may be used to determine the novelty of a patent application. Patent data is searched through multiple resources and returns results based on the query and the database and these results need to be merged to create the final result. (Stamatis et al., 2023) employ techniques such as random forest, Support Vector Regression, and Decision Trees to merge the search findings effectively.

### 3.2.2 Traditional Neural Networks

The methods based on neural networks have been popular in recent years for patent retrieval. (Kravets et al., 2017), (Jiang et al., 2021), and (Kucer et al., 2022) implement CNN, DUAL-VGG, and ResNet, respectively, to retrieve patent images based on a query image. (Chen et al., 2020) aim to solve entity identification and semantic relation extraction by BiLSTM-CRF (Huang et al., 2015) and BiGRU-HAN (Han et al., 2019), respectively.

### 3.2.3 PLMs & Multimodal Models (MMs)

PLMs are useful in many text-related tasks and patent retrieval is not an exception. (Kang et al., 2020) use the BERT language model which includes the combinations of title, abstract, and claim. (Siddharth et al., 2022) incorporate Sentence-BERT (Reimers and Gurevych, 2019) for text embeddings as well as use the TransE method for the citation and inventor knowledge graph embeddings. They identify that the mean cosine similarity among the vector representations of the patents is effective in linking multiple existing patents to a target patent. Multimodal techniques have also been used in information retrieval (Pustu-Iren et al., 2021). Here, the visual features are extracted using vision transformers, while textual features are from sentence transformers. (Pustu-Iren et al., 2021) utilize CLIP for image embedding alongside RoBERTa for captur-

Table 3: Summary of the methods on patent quality: “Many" includes Linear regression, Ridge regression, Random Forest, XGBoost, CNN, and LSTM. “APR" stands for the measures of accuracy, precision, and recall. IncoPat is a global patent database. We denote Attribute Network Embedding, Attention-based Convolutional Neural Network, European Telecommunications Standards Institute, Derwent Innovation by ANE, ACNN, ETSI, and DI, respectively.

| Papers | Indicators | Methods | Evaluation Metrics | Datasets |
|---|---|---|---|---|
| (Lin et al., 2018) | Citations, meta features | ANE, ACNN | RMSE | USPTO, OECD |
| (Trappey et al., 2019) | Principal component analysis (PCA) | DNN | Accuracy | ETSI and DI |
| (Hsu et al., 2020) | Investor reaction, citations | Many | MAE | Patentsview |
| (Chung and Sohn, 2020) | Abstract, claims, predefined | CNN, Bi-LSTM | Precision, recall | USPTO |
| (Aristodemou, 2021) | 12 patent indices | ANN | APR, F1, FNR, MAE | USPTO, OECD |
| (Erdogan et al., 2022) | 9 patent indices | MLP | Accuracy, Kappa, MAE | USPTO |
| (Li et al., 2022) | Maintenance period | BiLSTM-ATT-CRF | APR, F1 | IncoPat |
| (Krant, 2023) | Patent text | MSABERT | MSE | USPTO, OECD |

ing textual features, and thus, enhances the search process by incorporating both visual and textual data. (Lo et al., 2024) use distribution-aware contrastive loss to improve understanding of class and category information which achieves robust representations even for tail classes. For captioning, they employ open-source BLIP-2 and GPT-4V, a frozen text encoder from CLIP for text feature, and various visual encoder backbones, including ViT variants, ResNet50, EfficientNetB-0, and SwinV2-B. Among other techniques, (Higuchi and Yanai, 2023), (Higuchi et al., 2023) employ a deep metric learning framework with cross-entropy methods such as InfoNCE (Oord et al., 2018) and ArcFace (Deng et al., 2019).

#### 3.2.4 Discussion and Suggestion

Patent retrieval process involves several subtasks, such as defining technical requirements and merging search outcomes from various databases. The early methods often use traditional techniques like SVM, Naive Bayes, Decision trees, etc. While the image retrieval methods apply a variety of CNNs to effectively handle and analyze the visual data, the text retrieval methods have shifted towards PLMs for advanced linguistic analysis. Traditional machine learning techniques are limited to capturing the complexity of both patent image and text. Although CNNs are popular for image retrieval tasks, the question remains in their effectiveness for patent image retrieval, as patent images are non-traditional and technical. On the other hand, combining Vision Transformer alongside RoBERTa, Sentence-BERT, TransE shows another approach that might be more suitable for handling the multimodal (e.g., text, images) aspect of patents. (Pustu-Iren et al., 2021) demonstrate that the image and text-based transformer models achieve the highest mean average precision in patent retrieval tasks.

### 3.3 Patent Quality Analysis

We organize the methods for patent quality analysis below and provide a summary in Table 3.

#### 3.3.1 Traditional Neural Networks

(Erdogan et al., 2022) apply an MLP-based approach for quality analysis, utilizing nine indices such as claim counts, forward citations, backward citations, the patent family size to measure the value of a patent, etc. (Li et al., 2022) classify patents based on their maintenance period in four categories. This study implements a Bi-LSTM along with the attention mechanism and Conditional Random Field (CRF) to predict the quality of a patent. (Trappey et al., 2019) use Deep Neural Networks with 11 quality indicators. (Hsu et al., 2020) predict forward citation and investor reaction to patent announcements implementing CNN-LSTM neural networks and various ML models. (Chung and Sohn, 2020), (Lin et al., 2018) and (Aristodemou, 2021) apply a variety of neural networks such as CNN, Bi-LSTM, Attention-based CNN (ACNN), deep and wide Artificial Neural Networks (ANN), respectively.

#### 3.3.2 Pre-trained Language Models (PLMs)

(Krant, 2023) proposes to use MSABERT to assess patent value based entirely on the textual data and use the OECD (Eurostat, O., 2005) quality indicators for evaluation. Building upon BERT, MSABERT handles the multi-section structure and longer texts of patent documents. The OECD index includes composite indicators and generality with other predominant indices.

#### 3.3.3 Discussion and Suggestion

While numerous measures are used in assessing the quality of a patent, the absence of universally accepted “gold standard” poses a challenge. Among

Table 4: Example of the works that used PLMs and LLMs to solve patent tasks. This shows the growing trend of incorporating large-scale language models to improve patent processing and analysis.

| Work | Model | Task | Year |
|---|---|---|---|
| (Lee and Hsiang, 2020b) | BERT | Classification | 2020 |
| (Kang et al., 2020) | BERT | Retrieval | 2020 |
| (Lee and Hsiang, 2020a) | GPT-2 | Generation | 2020 |
| (Lee, 2020) | GPT-2 | Generation | 2020 |
| (Althammer et al., 2021) | SciBERT | Classification | 2021 |
| (Pustu-Iren et al., 2021) | RoBERTa | Retrieval | 2021 |
| (Roudsari et al., 2022) | BERT, RoBERTa | Classification | 2022 |
| (Siddharth et al., 2022) | SBERT | Retrieval | 2022 |
| (Christofidellis et al., 2022) | GPT-2 | Generation | 2022 |
| (Krant, 2023) | MSABERT | Quality Analysis | 2023 |
| (Bekamiri et al., 2024) | Sentence-BERT | Classification | 2024 |
| (Lo et al., 2024) | BLIP, GPT-4 | Retrieval | 2024 |
| (Wang et al., 2024a) | GPT-J, T5 | Generation | 2024 |
| (Lee, 2024) | GPT-J | Generation | 2024 |
| (Jiang et al., 2024) | Llama-3, Mistral, and PatentGPT-J | Generation | 2024 |
| (Bai et al., 2024) | Llama-2 and Mixtral | Generation | 2024 |
| (Ren and Ma, 2024) | Qwen2 | Generation | 2024 |
| (Wang et al., 2024b) | Qwen2, LLAMA3, GPT-4o, Mistral | Generation | 2024 |

several used indices, only forward citations are directly associated with the value—both monetary and quality—of a patent. Even though applying different deep learning models has some success, the question of building a method to handle technical information, metadata, and images together remains open. While MSABERT on the entire dataset will be computationally costly, building upon it might be useful for quality evaluation.

### 3.4 Patent Generation

The generative models are becoming increasingly popular in many domains. The recent developments in LLMs have also led to novel methods for generating patents, thus reducing significant human effort. Sec. 3.4.1 presents the studies with LLMs and PLMs for generating patent texts, and Sec. 3.4.2 focuses on the pretrained and advanced methods used for patent-specific data. Table 4 shows the trend of using PLMs and LLMs to solve different patent tasks, and most patent-related tasks are shifting towards leveraging LLMs. Table 5 (see Appendix) shows the summary of patent generation. We also discuss the broader impact in App. A.8.

#### 3.4.1 Patent Text Generation with LLMs

(Lee and Hsiang, 2020a) implement GPT-2 (Radford et al., 2019) models to generate the independent claims in patents. The researchers fine-tune the model on 555,890 patent claims of the granted utility patents in 2013 from USPTO. Providing a few words, the method generates the first independent claim of the patent. However, the study is limited to providing quantitative metrics to evaluate the quality of the generated patent claims. In a separate study, (Lee, 2020) focuses on personalized claim generation by fine-tuning a pre-trained GPT-2 model with inventor-centric data to demonstrate greater relevance. The measure of personalization in the generated claims has been assessed using a BERT model. (Christofidellis et al., 2022) introduce the Patent Generative Transformer (PGT) that supports three tasks: part-of-patent generation, text infilling, and coherence evaluation. They train GPT-2 on a dataset of 11.6 million patents. PGT shows strong zero-shot capabilities for generating abstracts with high semantic similarities from keywords. Patentformer (Wang et al., 2024a) generates detailed patent specifications by fine-tuning T5 and GPT-J language models on a dataset that includes claims, drawings, and descriptions. It focuses on two tasks: Claim-to-Specification, which creates specification text from a single claim, and Claim+Drawing-to-Specification, which integrates claims, drawings, and descriptions to produce richer specifications. (Jiang et al., 2024) generate claims by incorporating descriptions instead of abstracts. It also demonstrates an interesting observation that the general-purpose models—such as Llama-3, GPT-4, and Mistral—outperform models specifically trained on patent data (e.g., PatentGPT-J). The authors also conclude that fine-tuning enhances clarity, but revisions are still necessary for legal robustness.

#### 3.4.2 Patent-Specific LLMs

(Lee, 2024) finetunes a pretrained model PatentGPT-J-6B using reinforcement learning from human feedback (RLHF) to align patent claim generation with drafting goals. The authors design a custom reward function where claim length up to a defined length and inclusion of limiting terms are rewarded. These limiting terms improve the chance of patent approval. However, further improvements in text quality and broader datasets are needed to meet legal and practical patent standards. (Bai et al., 2024) build a cost-effective LLMs for the intellectual property (IP) domain to handle domain-specific expertise and long-text processing. They finetune open-source models like LLaMA2 and Mixtral with over 240 billion multilingual IP-focused tokens, nearly half from patent data. The approach incorporates pretraining, fine-tuning, and reinforcement learning to align model outputs with human preferences. Similarly, (Ren and Ma, 2024) introduce a specialized LLM based on Qwen2-1.5b for automated patent drafting. The approach integrates domain-specific knowledge using knowledge graphs, supervised fine-tuning, and RLHF. A multi-agent framework for drafting patents using LLMs is introduced by (Wang et al., 2024b). They employ agents for planning, writing, and reviewing to generate comprehensive patents from inventor drafts.

#### 3.4.3 Discussion and Suggestion

The use of PLMs and LLMs for automating patent generation has grown rapidly. However, a critical challenge remains in evaluating the quality of generated patents. The existing studies focus only on pretraining LLMs on patent-specific data to better capture the domain's technical language and structure without rigorous evaluation techniques. As a result, human intervention becomes essential to ensure accuracy, legal validity, and compliance with patent standards. Additionally, most approaches for patent generation focus exclusively on the text and overlook the multimodal nature of patents. This is particularly important for design patents, which consist of images predominantly.

## 4 Future Directions

Many researchers have leveraged NLP and Multimodal AI for patent analysis, yet significant research opportunities remain going forward. We believe a foundation model (e.g., LLMs, MLMs) tailored for patent data will enhance understanding and performance across diverse tasks.

**Multimodal Learning on Patents.** The availability of multiple modalities (e.g., text, images) in patent documents offers a comprehensive understanding of the related patent tasks. One of the challenges is that the patent images are often more complex and use advanced domain related concepts compared to the natural (or RGB) images. Recent advances in multimodal learning would allow for more reliable and accurate patent analysis. Intuitively, drawings or sketches provide geometrical information about individual patents. In general, multimodal learning can be used to *align representations* derived from text descriptions with those derived from technical images.

**Generative AI for Patents.** In patent generation, LLMs can suffer from hallucination, where they generate incorrect information. They might produce repetitive and monotonous texts that will lack creativity. Further, to mitigate the risk of patent infringement, LLMs need up-to-date patent data. Thus, the generation process requires human oversight and feedback to ensure accuracy and relevance and cannot be fully automated yet. On the other hand, the assessment of the text generated by the generative models is also challenging. As patents include jargons and many domain specific words, evaluating generated patent text in terms of only natural language will not be sufficient. Thus, the important question remains—*how to construct domain-specific evaluation measures for the synthetic or the generated text from LLMs?*

Additional extended future directions are mentioned in App. A.7.

## 5 Conclusions

In this survey, we have provided a comprehensive overview of various patent analysis tasks. We have presented a novel schema with a detailed organization of the research papers, analyzing the corresponding methodologies, their advantages, limitations, and how they are applied to different patent-related tasks. Our survey also focuses on the recent advancements of PLMs and LLMs as well as their usefulness in the patent domain. We have offered several insights into some potential future directions. This survey aims to be a useful guide for researchers, practitioners, and patent offices all over the world in the multidisciplinary field of NLP, Multimodal AI, and patent systems.

## 6 Limitations

The life cycle of a patent—the time from its submission to acceptance—is lengthy as it undergoes significant scrutiny and multiple iterations of revisions. The advancements in Machine Learning (e.g., LLMs) can make this process faster and thus, can essentially accelerate technological innovation. For instance, while reviewing, recent tools can help retrieve relevant documents more efficiently and accurately than a human reviewer who often requires enough experience. Our work is a survey of the existing methods for such tasks in patents. Though the survey itself does not have limitations as such, we discuss the limitations of modern AI techniques in general for patent tasks.

There are a few limitations of using AI in patent analysis. First, the LLMs methods may lack the nuanced understanding that human experts possess. Second, evaluation scores in classification and retrieval indicate lower accuracy (see Tables 8 & 9) and thus, they still need human intervention to obtain relevant literature—which is important while reviewing—to prevent the patent infringement issues. Therefore, the entire process cannot be fully automated, and it is important to have human experts in the loop. This requirement also applies to generative models for patent drafting (Sec. 3.4) which needs human guidance for accuracy. Additionally, there are ethical concerns regarding the potential displacement of human workers by AI tools.

## 7 Ethics Statement

In this work, we have surveyed AI methods for patent tasks. We do not foresee any ethical issues from our study.

## References


Kehinde Ajayi, Xin Wei, Martin Gryder, Winston Shields, Jian Wu, Shawn M. Jones, Michal Kucer, and Diane Oyen. 2023. Deeppatent2: A large-scale benchmarking corpus for technical drawing understanding. *Scientific Data*.

Sophia Althammer, Mark Buckley, Sebastian Hofstätter, and Allan Hanbury. 2021. Linguistically informed masking for representation learning in the patent domain. *arXiv preprint arXiv:2106.05768*.

Leonidas Aristodemou. 2021. *Identifying valuable patents: A deep learning approach*. Ph.D. thesis.

Zilong Bai, Ruiji Zhang, Linqing Chen, Qijun Cai, Yuan Zhong, Cong Wang, Yan Fang, Jie Fang, Jing Sun, Weikuan Wang, et al. 2024. Patentgpt: A large language model for intellectual property. *arXiv preprint arXiv:2404.18255*.

Hamid Bekamiri, Daniel S Hain, and Roman Jurowetzki. 2024. Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert. *Technological Forecasting and Social Change*, 206:123536.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP*.

Fernando Benites, Shervin Malmasi, and Marcos Zampieri. 2018. Classifying patent applications with ensemble methods. *ALTA Workshop*.

Liang Chen, Shuo Xu, Lijun Zhu, Jing Zhang, Xiaoping Lei, and Guancan Yang. 2020. A deep learning based method for extracting semantic information from patent documents. *Scientometrics*, 125:289–312.

Yung-Chang Chi and Hei-Chia Wang. 2022. Establish a patent risk prediction model for emerging technologies using deep learning and data augmentation. *Advanced Engineering Informatics*, 52:101509.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *ACL*.

Seokkyu Choi, Hyeonju Lee, Eunjeong Park, and Sungchul Choi. 2022. Deep learning for patent landscaping using transformer and graph embedding. *Technological Forecasting and Social Change*, 175.

Dimitrios Christofidellis, Antonio Berrios Torres, Ashish Dave, Manuel Roveri, Kristin Schmidt, Sarath Swaminathan, Hans Vandierendonck, Dmitry Zubarev, and Matteo Manica. 2022. PGT: a prompt based generative transformer for the patent domain. In *ICML 2022 Workshop on Knowledge Retrieval and Language Models*.

Park Chung and So Young Sohn. 2020. Early detection of valuable patents using a deep learning model: Case of semiconductor industry. *Technological Forecasting and Social Change*, 158:120146.

DaVinci. 2024. Davinci ai. https://www.getdavinci.ai/. Accessed: 2024-04-27.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

Zulfiye Erdogan, Serkan Altuntas, and Turkay Dereli. 2022. Predicting patent quality based on machine learning approach. *IEEE Trans Eng Manag*.

Eurostat, O. 2005. *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data*. OECD, Paris. A joint publication of OECD and Eurostat.

Junaid Ahmed Ghauri, Eric Müller-Budack, and Ralph Ewerth. 2023. Classification of visualization types and perspectives in patents. In *TPDL*.

Vito Giordano, Giovanni Puccetti, Filippo Chiarello, Tommaso Pavanello, and Gualtiero Fantoni. 2023. Unveiling the inventive process from patents by extracting problems, solutions and advantages with natural language processing. *Expert Systems with Applications*, 229:120499.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Mattyws F Grawe, Claudia A Martins, and Andreia G Bonfante. 2017. Automated patent classification using word embedding. In *ICMLA*. IEEE.

Xu Han, Tianyu Gao, Yuan Yao, Demin Ye, Zhiyuan Liu, and Maosong Sun. 2019. Opennre: An open and extensible toolkit for neural relation extraction. In *EMNLP*.

Kotaro Higuchi, Yuma Honbu, and Keiji Yanai. 2023. Patent image retrieval using cross-entropy-based metric learning. In *IW-FCV*.

Kotaro Higuchi and Keiji Yanai. 2023. Patent image retrieval using transformer-based deep metric learning. *WPI*, 74:102217.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Po-Hsuan Hsu, Dokyun Lee, Prasanna Tambe, and David H Hsu. 2020. Deep learning, text, and patent valuation. *Text, and Patent Valuation*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Hyejin Jang, Sunhye Kim, and Byungun Yoon. 2023. An explainable ai (xai) model for text-based patent novelty analysis. *Expert Systems with Applications*.

Lekang Jiang, Caiqi Zhang, Pascal A Scherz, and Stephan Goetz. 2024. Can large language models generate high-quality patent claims? *arXiv preprint arXiv:2406.19465*.

Shuo Jiang, Jianxi Luo, Guillermo Ruiz-Pava, Jie Hu, and Christopher L Magee. 2021. Deriving design feature vectors for patent images using convolutional neural networks. *Journal of Mechanical Design*, 143(6):061405.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *EACL*.

Eleni Kamateri, Michail Salampasis, and Konstantinos Diamantaras. 2023. An ensemble framework for patent classification. *WPI*, 75:102233.

Eleni Kamateri, Vasileios Stamatis, Konstantinos Diamantaras, and Michail Salampasis. 2022. Automated single-label patent classification using ensemble classifiers. In *ICMLC*.

Dylan Myungchul Kang, Charles Cheolgi Lee, Suan Lee, and Wookey Lee. 2020. Patent prior art search using deep learning language model. In *IDEAS*.

Xabi Krant. 2023. *Text-based Patent-Quality Prediction Using Multi-Section Attention*. Ph.D. thesis.

Alla Kravets, Nikita Lebedev, and Maxim Legenchenko. 2017. Patents images retrieval and convolutional neural network training dataset quality improvement. In *ITSMSSM*.

Ralf Krestel, Renukswamy Chikkamath, Christoph Hewel, and Julian Risch. 2021. A survey on deep learning for patent analysis. *WPI*, 65:102035.

Michal Kucer, Diane Oyen, Juan Castorena, and Jian Wu. 2022. Deeppatent: Large scale patent drawing recognition and retrieval. In *WACV*.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Jieh-Sheng Lee. 2020. Patent transformer: A framework for personalized patent claim generation. In *CEUR Workshop Proceedings*, volume 2598. CEUR-WS.

Jieh-Sheng Lee. 2024. Instructpatentgpt: training patent language models to follow instructions with human feedback. *Artificial Intelligence and Law*, pages 1–44.

Jieh-Sheng Lee and Jieh Hsiang. 2020a. Patent claim generation by fine-tuning openai gpt-2. *WPI*, 62:101983.

Jieh-Sheng Lee and Jieh Hsiang. 2020b. Patent classification by fine-tuning bert language model. *WPI*, 61:101965.

Rongzhang Li, Hongfei Zhan, Yingjun Lin, Junhe Yu, and Rui Wang. 2022. A deep learning-based early patent quality recognition model. In *ICNC-FSKD*.

Shaobo Li, Jie Hu, Yuxin Cui, and Jianjun Hu. 2018. Deeppatent: patent classification with convolutional neural networks and word embedding. *Scientometrics*, 117:721–744.

Hongjie Lin, Hao Wang, Dongfang Du, Han Wu, Biao Chang, and Enhong Chen. 2018. Patent quality valuation with deep learning models. In *DASFAA*.

Xipeng Liu and Xinmiao Li. 2022. Early identification of significant patents using heterogeneous applicant-citation networks based on the chinese green patent data. *Sustainability*, 14(21):13870.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hao-Cheng Lo, Jung-Mei Chu, Jieh Hsiang, and Chun-Chieh Cho. 2024. Large language model informed patent image retrieval. *arXiv preprint arXiv:2404.19360*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Kader Pustu-Iren, Gerrit Bruns, and Ralph Ewerth. 2021. A multimodal approach for semantic patent image retrieval. In *PatentSemTech*.

Qatent. 2024. Qatent. https://qatent.com/. Accessed: 2024-04-27.

Questel. 2024. Ai classifier. https://www.questel.com/patent/ip-intelligence-software/ai-classifier/. Accessed: 2024-04-27.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1:9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP*.

Runtao Ren and Jian Ma. 2024. Patentgpt: A large language model for patent drafting using knowledge-based fine-tuning method. *arXiv preprint arXiv:2409.00092*.

Julian Risch, Nicolas Alder, Christoph Hewel, and Ralf Krestel. 2021. Patentmatch: A dataset for matching patent claims & prior art. In *PatentSemTech@SIGIR*.

Julian Risch and Ralf Krestel. 2018. Learning patent speak: Investigating domain-specific word embeddings. In *ICDIM*.

Julian Risch and Ralf Krestel. 2019. Domain-specific word embeddings for patent classification. *Data Technologies and Applications*, 53(1).

Arousha Roudsari, Jafar Afshar, Wookey Lee, and Suan Lee. 2022. Patentnet: multi-label classification of patent documents using deep learning based language understanding. *Scientometrics*, pages 1–25.

Rossitza Setchi, Irena Spasić, Jeffrey Morgan, Christopher Harrison, and Richard Corken. 2021. Artificial intelligence for patent prior art searching. *WPI*.

Marawan Shalaby, Jan Stutzki, Matthias Schubert, and Stephan Günnemann. 2018. An lstm approach to patent classification based on fixed hierarchy vectors. In *SDM*.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *ACL*.

Homaira Huda Shomee, Zhu Wang, Sathya N. Ravi, and Sourav Medya. 2024. IMPACT: A large-scale integrated multimodal patent analysis and creation dataset for design patents. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

L Siddharth, Guangtong Li, and Jianxi Luo. 2022. Enhancing patent retrieval using text and knowledge graph embeddings: a technical note. *Journal of Engineering Design*, 33(8-9):670–683.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Mustafa Sofean. 2021. Deep learning based pipeline with multichannel inputs for patent classification. *WPI*, 66:102060.

Vasileios Stamatis, Michail Salampasis, and Konstantinos Diamantaras. 2023. Machine learning methods for results merging in patent retrieval. *Data Technologies and Applications*.

Mirac Suzgun, Luke Melas-Kyriazi, Suproteem Sarkar, Scott D Kominers, and Stuart Shieber. 2024. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. *Advances in Neural Information Processing Systems*, 36.

Amy JC Trappey, Charles V Trappey, Usharani Hareesh Govindarajan, and John JH Sun. 2019. Patent value analysis using deep learning models—the case of iot technology mining for the manufacturing industry. *IEEE-TEM*, 68(5):1334–1346.

Juanyan Wang, Sai Krishna Reddy Mudhiganti, and Manali Sharma. 2024a. Patentformer: A novel method to automate the generation of patent applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1361–1380.

Qiyao Wang, Shiwen Ni, Huaren Liu, Shule Lu, Guhong Chen, Xi Feng, Chi Wei, Qiang Qu, Hamid Alinejad-Rokny, Yuan Lin, et al. 2024b. Autopatent: A multi-agent framework for automatic patent generation. *arXiv preprint arXiv:2412.09796*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*.

Wan Mohammad Faris Zaini, Daphne Teck Ching Lai, and Ren Chong Lim. 2022. Identifying patent classification codes associated with specific search keywords using machine learning. *WPI*, 71:102153.

Tao Zou, Le Yu, Leilei Sun, Bowen Du, Deqing Wang, and Fuzhen Zhuang. 2023. Event-based dynamic graph representation learning for patent application trend prediction. *IEEE TKDE*.

# A Appendix

## A.1 Overview of the tasks

Overview of the major patent tasks: patent classification, patent retrieval, patent generation, and patent quality analysis is shown in Figure 2. Popular AI methods in the literature covered by this survey are listed in Table 6.

## A.2 Search and inclusion criteria.

We have conducted our literature search using Google Scholar and Semantic Scholar, focusing on various categories of patent-related tasks. To align with the recent trends, we have limited our search to publications from 2017 to 2024. Our search criteria included various keywords such as 'patent', 'AI in patent', 'patent classification', 'patent tasks', 'patent retrieval', 'patent generation', 'patent quality analysis', and 'patent dataset'. This combination of search terms has yielded hundreds of patent-related research papers. We have excluded more than half of these papers after reviewing their titles and abstracts, as they have not met our criteria (e.g., they did not fall under any of the relevant categories). After thorough scrutiny and reorganization, we have included 50 papers for the survey.

## A.3 Background: Formulation of Patent Tasks

We provide the problem formulations of the popular patent tasks as follows.

### A.3.1 Patent Classification

Given patents as $(x_i, y_i)_{i=1}^{N}$, where $x_i$ denotes the features of the $i$-th patent, $C$ denotes the set of classes, $C = \{1, 2, \ldots, k\}$, and $y_i = \{y_{i1}, y_{i2}, \ldots, y_{iK}\}$ is a binary multi-label vector, where $y_{ik} \in \{0, 1\}$ is an indicator of whether class $k$ is the correct classification for the example patent $i$. Since a single patent can belong to more than one class in $C$, the goal is to predict $y_i$.
Table 7 shows an example of CPC classification.

### A.3.2 Patent Retrieval

Given a query patent as $q$ and a set of patents $X = \{x_i, \ldots, x_n\}$, where $x_q$ and $x_i$ are the features of the query and the patent $i$ in the set $X$. The goal is to compute a similarity score (e.g. cosine) $s(x_q, x_i)$ and return a set of patents $R(q) = \{x_j, \ldots, x_k\}$ based on top-k high similarities.

### A.3.3 Patent Generation

Given the patent $x_i$, where $x_i$ are the features constructed from the instruction, title, abstract, or any other part of the patent of the example patent $i$, the output $y_i$ can be another part of the patent (e.g., abstract, the first claim). The generation function $G$ can be denoted as $y_i$=$G(x_i; \theta)$, where $\theta$ is the parameter of the generation model $G$. The goal is to generate $y_i$ by learning $\theta$, or inferring from a pre-trained model with learned $\theta$.

Table 5 shows the summary of the models and datasets used to generate parts of the patent text.

## A.4 Evaluation results

We discuss all the studies and related methods in Section 3. We present the evaluation metrics and the results in Table 8 and 9.

## A.5 NLP and AI-based Methods for Other Relevant Patent Tasks

There are other interesting studies in the patent domain. Recent work focuses on patent infringement, such as (Chi and Wang, 2022) develop a model with different deep learning methods, such as CNN and LSTM, to predict the possibility of a patent application being granted and classify the reason for a failed application. Another work (Choi et al., 2022) applied a transformer and a Graph Neural Network (GNN) on patent classification for patent landscaping. (Zaini et al., 2022) present an unsupervised method to identify the correlations between patent classification codes and search keywords using PCA and k-means. These studies provide advanced deep learning methods to avoid the risks in patent application. Moreover, there are various studies on generating new ideas and evaluating novelty, such as identifying the inventive process of novel patent using BERT (Giordano et al., 2023), and an explainable AI (XAI) model for novelty analysis via (Jang et al., 2023). (Zou et al., 2023) propose a new task to predict the trends of patents for the companies, and also provide a solution for the task by training an event-based GNN. These studies bring new insights and directions for patent ideas and developments.

**Applications in Businesses.** The use of LLMs among businesses for patent related processes has significantly risen over time. The usage of the machine learning methods for these patents is growing at an impressive average annual rate of

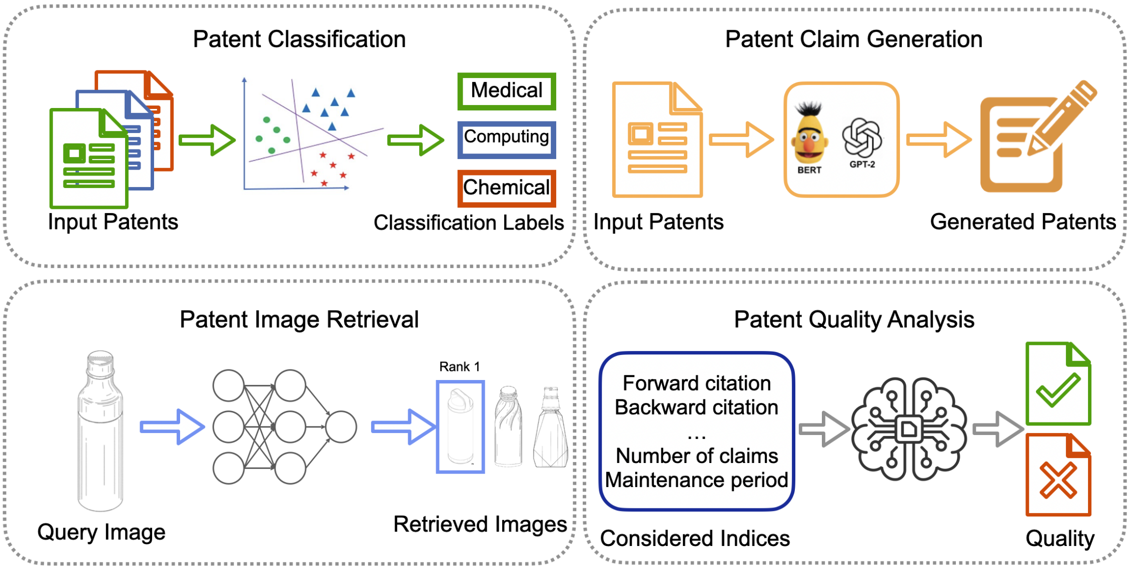


Figure 2: The overview of four major tasks of patent analysis. The patent retrieval task includes obtaining relevant patents (text and images). Please refer to the detailed descriptions of these tasks in Section 2.

Table 5: Summary of the works on patent generation. Here, "comprehensive" denotes patent claims, specification drafting, classification, translation, etc. IP data includes research papers, litigation records, web, news, etc.

| Papers | Model | Parts | Data |
|---|---|---|---|
| (Lee and Hsiang, 2020a) | GPT-2 | Independent Claims | USPTO |
| (Lee, 2020) | GPT-2 | Personalized Claims | USPTO |
| (Christofidellis et al., 2022) | GPT-2 | title, abstract, claim | – |
| (Wang et al., 2024a) | T5, GPT-J (Patentformer) | Claim-to-Specification, Claim+Drawing-to-Specification | USPTO |
| (Jiang et al., 2024) | Llama-3, Mistral, and PatentGPT-J | Claims | HUPD |
| (Lee, 2024) | PatentGPT-J | Claim | USPTO, PatentsView |
| (Bai et al., 2024) | LLaMA2 and Mixtral | Comprehensive | Both patent and IP data |
| (Ren and Ma, 2024) | Qwen2 | Comprehensive | USPTO |
| (Wang et al., 2024b) | Qwen2, LLAMA3, GPT-4o, Mistral | Comprehensive | HUPD |

28%[3]. Businesses are increasingly applying AI to enhance various aspects of the patent process, from drafting and classification to search and analysis. Some of the prominent examples include (Qatent, 2024), (DaVinci, 2024), and (Questel, 2024). (Qatent, 2024) leverages the latest NLP techniques to facilitate patent drafting for patent practitioners. It focuses on automating routine tasks—typing, automating renumbering of claims, and antecedence checking. It recommends various word and sentence alternatives during the claim drafting process, such as synonyms, broader or more specific terms, and other linguistic variations. Despite recent discussions around AI-generated inventions, Qatent maintains a human-centric approach which ensures all outputs are driven and controlled by human drafters. (DaVinci, 2024) is an advanced tool for drafting patents that uses generative AI to streamline the process. It supports a variety of document formats and lets users alter the AI's writing style to suit their needs. (Questel, 2024) offers AI powered patent classification, comprehensive patent search capabilities, efficient exploration of new markets, and opportunities such as management of patent fees and renewals.

### A.6 Patent Dataset and Repositories

Patent data are publicly available for bulk download from several sources in various formats such as XML, TSV, TIFF, and PDF. Examples include the USPTO, PatentsView[4], EPO[5], and WIPO[6]. Freepatent and Findpatent are patent data websites, where Findpatent includes patents registered in Russia. Beyond these resources, several patent datasets are available for benchmarking purposes. The datasets are detailed in Table 10.

[3] https://ip.com/blog/can-ai-invent-independently-how-ai-is-changing-the-patent-industry/

[4] https://patentsview.org/

[5] https://www.epo.org/

[6] https://www.wipo.int/classifications/ipc/en/ITsupport/

Table 6: Popular AI methods in the literature. We use the acronyms frequently in our survey.

| Acronym | Full Name | Paper |
|---|---|---|
| LSTM | Long short-term memory | (Hochreiter and Schmidhuber, 1997) |
| CNN | Convolutional Neural Networks | (LeCun et al., 1998) |
| Bi-LSTM | Bidirectional Long Short-Term Memory | (Graves and Schmidhuber, 2005) |
| Word2Vec | – | (Mikolov et al., 2013) |
| GRU | Gated Recurrent Units | (Cho et al., 2014) |
| Bi-GRU | Bidirectional Gated Recurrent Units | (Cho et al., 2014) |
| DUAL-VGG | Dual Visual Geometry Group | (Simonyan and Zisserman, 2015) |
| FastText | – | (Joulin et al., 2017) |
| BERT | Bidirectional Encoder Representations from Transformers | (Devlin et al., 2019) |
| RoBERTa | Robustly Optimized BERT Pre-training Approach | (Liu et al., 2019) |
| SciBERT | Scientific BERT | (Beltagy et al., 2019) |

Table 7: An example of Cooperative Patent Classification (CPC) Scheme for the section A and its hierarchical categorization.

| Level | Code | Category |
|---|---|---|
| Section | A | Human Necessities |
| Class | A61 | Medical or Veterinary Science: Hygiene |
| Sub-class | A61B | Diagnosis: Surgery: Identification |
| Group | A61B5 | Measuring for diagnostic purposes; Identification of persons |
| Sub-group | A61B5/0006 | ECG or EEG signals |

### A.7 Extended Future Directions

**Patent Assessment.** To asses patent's novelty, one of the major tasks is to retrieve similar patents to determine whether the patent is significantly different from existing patents. One of the important task in this case is to generate search queries. This often needs alternate search terms, related words, and synonyms which require domain knowledge. The quality and structure of queries directly impact the relevance of the search results. The current methods are yet to automate this entire process. Thus, it brings challenges to obtain adequate similar patents and correctly assess patent's innovativeness and novelty. On the other hand, the generic quality analysis are based on well-known measures (Aristodemou, 2021; Erdogan et al., 2022). Nonetheless, it remains unclear which of these indices are associated with the actual value of the patent (e.g., generated revenue).

**Building a Knowledge Graph.** Patents are represented as nodes connected by edges such as citations in a citation network (Liu and Li, 2022). This structured representation allows for detailed citation analysis which is considered a crucial metric in understanding a patent's value. One interesting future direction would be to build a knowledge graph using other important information such as metadata, semantic similarity of patents, etc. This may lead to a more organized landscape of patents. This knowledge graph can help with prior art searches, the identification of related patents, and identify valuable patents (e.g., patents with high citations) (Siddharth et al., 2022).

### A.8 Broader Impacts

The life-cycle of a patent—the time from its submission to acceptance—is lengthy as it undergoes significant scrutiny and multiple iterations of revisions. The advancements in LLMs can make this process faster and thus, can essentially accelerate technological innovation. For instance, while reviewing, recent tools can help retrieve relevant documents more efficiently and accurately than a human reviewer who often requires enough experience.

Some of the major benefits are as follows: (1) Speed: The inclusion of LLMs and Multimodal AI in patent analysis tasks will speed up the review process. For example, (Ghauri et al., 2023) use a vision transformer that classifies images much more efficiently than previous works, and (Bekamiri et al., 2024) achieve higher recall in classification tasks. Since patent classification is a time-consuming task for a human expert, incorporating these advancements into the review process will make the process faster. (2) Novelty: Another important task is retrieving similar patents which is essential to assess the novelty of a patent. (Higuchi and Yanai, 2023) show a satisfactory mAP in retrieving similar images, which can play a key role in patent infringement. (3)Innovation: (Lee and Hsiang, 2020a; Lee,

Table 8: Existing results on the patent classification task. Hierarchy levels for classification include Section, Class, Subclass , Group, and Subgroup. The tuple (Result 1, Reuslt 2) denotes the results using (Data 1, Data 2) for the papers that report the measures using multiple datasets separately. The WIPO-alpha is a dataset for automated patent classification systems, and ALTA2018 is a dataset from Language Technology Programming Competition.

| Papers | Hierarchy Level | Accuracy | Precision | Recall | F1 | Top-3 | Data |
|---|---|---|---|---|---|---|---|
| (Grawe et al., 2017) | Subgroup | 0.63 | 0.63 | 0.66 | 0.62 | – | USPTO |
| (Shalaby et al., 2018) | Subclass | – | – | – | 0.61 | 0.79: F1 | - |
| (Shalaby et al., 2018) | Class | – | – | – | 0.72 | 0.89: F1 | - |
| (Risch and Krestel, 2018) | Subclass | – | (0.49, 0.53) | – | – | (0.72,0.75): Precision | WIPO-alpha, USPTO |
| (Benites et al., 2018) | Class | – | – | – | 0.78 | – | ALTA2018, WIPO |
| (Risch and Krestel, 2019) | Subclass | – | (0.49, 0.53) | – | – | (0.72,0.75): Precision | WIPO-alpha, USPTO |
| (Lee and Hsiang, 2020b) | Subclass | – | 0.81 | 0.55 | 0.65 | 0.44: F1 | USPTO |
| (Althammer et al., 2021) | Subclass | 0.59 | 0.58 | 0.59 | 0.581 | – | USPTO |
| (Sofean, 2021) | Subclass | 0.74 | 0.92 | 0.63 | 0.75 | – | EPO, WIPO |
| (Roudsari et al., 2022) | Subclass | – | (0.82, 0.82) | (0.55, 0.67) | (0.63, 0.72) | – | USPTO, CLEF-IP 2011 |
| (Kamateri et al., 2022) | Subclass | 0.64 | – | – | – | – | CLEF-IP 2011 |
| (Ghauri et al., 2023) | Image type | 0.85 | – | – | – | – | CLEF-IP 2011, USPTO |
| (Kamateri et al., 2023) | Subclass | 0.68 | – | – | – | 0.89: accuracy | CLEFIP-0.54M |
| (Bekamiri et al., 2024) | Subclass | – | 0.67 | 0.71 | 0.66 | – | USPTO |

Table 9: Results of the papers for the Patent Retrieval task. Here, mAP denotes mean average precision. Freepatent and Findpatent are patent data websites, where Findpatent includes patents registered in Russia. WIPS is a patent information search system.

| Work | Data type | Data | Accuracy (%) | Precision | Recall | F1 | mAP |
|---|---|---|---|---|---|---|---|
| (Kravets et al., 2017) | image | Freepatent, Findpatent | 30 | – | – | – | – |
| (Kang et al., 2020) | text | WIPS | – | 71.74 | 94.29 | 81.48 | – |
| (Chen et al., 2020) | text | USPTO | – | 92.4 | 91.9 | 92.2 | – |
| (Pustu-Iren et al., 2021) | image+text | EPO | – | – | – | – | 0.715 |
| (Siddharth et al., 2022) | text | USPTO | 70.2 | 65.9 | 81.2 | 72.6 | – |
| (Kucer et al., 2022) | image | DeepPatent | 70.1 | – | – | – | 37.9 |
| (Higuchi and Yanai, 2023) | image | DeepPatent | – | – | – | – | 0.85 |
| (Higuchi et al., 2023) | image | DeepPatent | – | – | – | – | 0.622 |
| (Lo et al., 2024) | image | DeepPatent2 | – | – | – | – | 0.69 |

2020) explore generating new patents, which is an important component to foster new innovation. This research provides inspiration for further development in the field including creation of new and innovative patents.

Table 10: Overview of Patent Datasets: size, format, data type and intended tasks

| **Dataset** | **Size** | **Format** | **Data type** | Task |
|---|---|---|---|---|
| USPTO-2M (Li et al., 2018) | 2M | JSON | text | Classification |
| BIGPATENT (Sharma et al., 2019) | 1.3M | JSON | text | Summarization |
| USPTO-3M (Lee and Hsiang, 2020b) | 3M | SQL statement | text | Classification |
| PatentMatch (Risch et al., 2021) | 6.3M | JSON | text | Retrieval |
| DeepPatent (Kucer et al., 2022) | 350K | XML & PNG | text & image | Retrieval |
| DeepPatent2 (Ajayi et al., 2023) | 2M | JSON & PNG | text & image | Retrieval |
| HUPD (Suzgun et al., 2024) | 4.5M | JSON | text | Multi-purpose |
| IMPACT (Shomee et al., 2024) | 3.61M | CSV & TIFF | text & image | Multi-purpose |