

ChatGPT vs. Google Gemini: Assessing AI Frontiers for Patent Prior Art Search Using European Search Reports

Renukswamy Chikkamath^{1,*}, Ankit Sharma², Christoph Hewel² and Markus Endres¹

¹University of Applied Sciences Munich, Munich, Germany

²PAUSTIAN & PARTNERS, Munich, Germany

Abstract

Accurately identifying paragraphs in prior art documents that may compromise the novelty of claims in patent applications is crucial but challenging. While recent advancements in Large Language Models (LLMs) demonstrate impressive language understanding and analysis capabilities, their efficacy in legal contexts, such as patent examination, remains underexplored. This study addresses this gap by evaluating the effectiveness of ChatGPT and Google Gemini in patent prior art search, specifically in assessing novelty. We constructed a test dataset based on European search reports to assess the models' ability to retrieve the closest examiner-cited paragraphs from a set of candidate paragraphs. Our findings also highlight the potential of LLMs for patent classification across various hierarchical levels. Additionally, we explored the divergence between these LLMs and state-of-the-art embedding-based (patent-specific and general models) similarity functions in novelty identification. We show that optimized prompting enables ChatGPT and Google Gemini to excel in passage retrieval, surpassing state-of-the-art embeddings even without explicit fine-tuning. Despite their success, these models still face challenges in retrieving examiners' cited paragraphs that may diminish the novelty of a given prior art.

Keywords

ChatGPT, Google Gemini (BARD), Prior art search, Patent search reports, Patent retrieval

1. Introduction


Inventions meeting the criteria of novelty and inventive step can be granted a patent, providing legal protection for a limited period. During patent examination, two key tasks such as classification and prior art search are conducted by patent offices to ensure the invention's uniqueness and inventiveness. The substantial volume of patent applications (e.g., 3.4 million worldwide in 2022¹) poses challenges, prompting the need for Artificial Intelligence (AI) tools to assist in the manual processes. Global patent offices are actively exploring efficient AI tools² and models for various patent analysis tasks. For example, the European Patent Office (EPO) is considering an

Second International Workshop on Semantic Technologies and Deep Learning Models for Scientific, Technical and Legal Data May 26th or 27th, 2024 / Hersonissos, Greece held at ESWC 2024

*Corresponding author.

✉ renukswamy.chikkamath@hm.edu (R. Chikkamath); ankit.sharma33@outlook.com (A. Sharma); hewel@paustian.de (C. Hewel); markus.endres@hm.edu (M. Endres)

ORCID iD 0000-0002-6010-670X (R. Chikkamath)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹WIPO IP Facts and Figures 2023, WIPO, <https://www.wipo.int/publications/en/details.jsp?id=4686>

²Index of AI initiatives in IP offices, https://www.wipo.int/about-ip/en/artificial_intelligence/search.jsp

embedding methodology [1] to vectorize patent data for an AI-based prior art search, while concurrently utilizing their own AI-based classification³ tool. In the field of Natural Language Processing (NLP), a significant shift in language understanding has sparked growing interest and developments in text classification, summarization, and information retrieval. With the advent of LLMs like ChatGPT (released by OpenAI⁴ in December 2022), researchers, industry professionals, and technology enthusiasts have taken notice. Continued interest in LLMs is evident with Google's⁵ release of the public version AI tool BARD in March 2023 for text and image understanding. Additionally, the recent launch of the AI tool Gemini in February 2024 further contributes to the evolving landscape of LLMs.

The importance of these AI advancements, exemplified by ChatGPT and Google Gemini, has captured the attention of various industries (e.g. in the medical domain [2]) to assist human labor. Even patent professionals aimed to leverage these tools for tasks such as summarization, claims drafting, artificial patent drafting, feature mapping, and infringement⁶ search. However, some patent professionals express concerns about the limitations of ChatGPT, likening it to “stochastic parrots” that lack critical thinking abilities necessary for patent prior art search [3]. Similar reservations exist regarding tools like Google Gemini, with concerns about hallucinations⁷ and false answers with high confidence. Despite the widespread interest and potential conflicts in the usage of these language models, detailed investigations in the *patent domain*, particularly on standardized patent datasets based on examiners' search reports, are lacking. The absence of comprehensive studies, especially in challenging tasks like patent classification and prior art search (e.g., patent passage retrieval), highlights the need for empirical experiments. To address this gap, we propose to assess the effectiveness of ChatGPT and Google Gemini on standard patent tasks using a gold standard dataset developed for this purpose.

In particular, the objectives of this work are: i) Manually preparing a test dataset for patent classification and passage retrieval using European examination search reports. ii) Evaluating ChatGPT and Google Gemini for patent classification. iii) Evaluating ChatGPT and Google Gemini for novelty passage retrieval. iv) Exploring the divergence between these LLMs and state-of-the-art embedding-based (patent-specific and general models) similarity functions in novelty identification.

The remaining sections of the paper are organized as follows: Section 2 explains the related work. Section 3 outlines the dataset preparation, Section 4 presents the proposed methodology for investigating ChatGPT and Google Gemini, along with the evaluation strategy. Section 5 contains the recorded results for classification and passage retrieval, accompanied by a detailed discussion. Finally, we conclude our work and suggest future research directions in Section 6.

2. Related work

In the realm of prior art search, efforts to enhance the default BM25-based retrieval [4] have led to increased interest in leveraging semantic information [5, 6, 7]. In the patent domain, word

³<https://www.epo.org/en/news-events/news/new-cpc-text-categoriser-powered-ai>

⁴<https://openai.com/blog/chatgpt>

⁵<https://gemini.google.com/updates?hl=en>

⁶<https://havingip.com/patent-infringement-search-ai-chatgpt/>

⁷<https://havingip.com/bard-ai-patent-search-infringement-classification-drafting-prior-art/>

embeddings play a crucial role in representing patent text more efficiently for classification tasks [8]. Researchers, recognizing the importance of domain knowledge and semantic understanding, are also incorporating examiner knowledge to train models using search reports. A noteworthy dataset for understanding patent paragraphs, developed by Risch et al. [9], has sparked immediate interest in utilizing it for novelty prediction, providing a new perspective on training models for novelty [10]. While cross-encoder architectures are deemed unsuitable for large-scale retrieval settings, the use of bi-encoders in training BERT architectures has reshaped sentence embeddings [11]. The patent domain has adopted SBERT architectures, resulting in improved performance across various tasks, including classification [12]. Building on SBERT training methods, the EPO has recently leveraged search report citations to claim efficient embeddings that outperform BM25 and other general-purpose state-of-the-art embeddings in retrieval settings. These embeddings, however, remain proprietary and are not publicly accessible [1].

In addition to employing domain-specific language models [13], there is a growing interest in using AI for claim scoping [14], aiding prior art search for novelty [15], classification [16], and improving model explainability [17, 18]. Patent similarity approaches have undergone a recent shift [19, 20], with attention extending beyond domain-specific models. Large Language Models (LLMs) and other generative models, such as ChatGPT and Google Gemini, have gained traction across various industries.

The patent community is increasingly interested in assessing the effectiveness of LLMs in aiding prior art search [21, 22]. Despite their importance and demand, investigations based on standard datasets and empirical evidence regarding the use of LLMs (ChatGPT and Google Gemini), especially for classification and novelty passage retrieval, are lacking to the best of our knowledge. Hence, we propose the development of a test examination dataset in order to evaluate LLMs for classification and novelty passage retrieval, and in the subsequent section, we discuss the importance and strategies adopted for its creation.

3. ClaimCiteRetrieval Test Dataset for LLMs

To evaluate the effectiveness of ChatGPT and Google Gemini for novelty passage retrieval, a gold standard test dataset, which also supports free web-based chatbot versions, is currently unavailable. Therefore, we developed the ClaimCiteRetrieval dataset based on European examiners' search reports. These gold labels, originating from patent examiners, replicate artificial patent examinations for novelty using a passage retrieval task. The objective is to showcase whether and how it is feasible to identify the essential paragraphs for evaluating novelty or inventive step in a given test query (independent claim of an application).

The dataset preparation involves two stages: *selecting patents* and *matching query-paragraphs*. There are approximately 1000 text units, including patent numbers, abstracts, independent claims, IPC codes, and various paragraphs (both cited and non-cited). In total, we conducted 49 manual examinations to compile our dataset. The collection of data units for each examination requires approximately 35-40 minutes of human labor. The data units, employed for retrieval and classification tasks, exhibit both quality and an ample quantity sufficient for handling free versions to conduct artificial examinations.

- (i) **Selection of Patents:** (accounting semantic relatedness) Given the emphasis on semantic technologies, Deep Learning (DL) models, and LLMs for legal data handling in information retrieval, we opted for patents falling under international patent classification (IPC) class G06F (digital data processing). To intensify the retrieval challenge for LLMs and embedding models, we narrowed down to the last child node level within the hierarchical patent classification, for instance, “G06F 40/00 (handling natural language data).” To enhance semantic relatedness in test queries, making the retrieval task more demanding (as it is for human examiners) even at the passage or paragraph level, we selected codes beneath G06F 40/00, such as “G06F 40/10 (text processing),” “G06F 40/20 (natural language analysis),” and “G06F 40/30 (semantic analysis).” Moreover, the technological intricacies in inventions under G06F 40/00 (section G) are relatively less complex and shorter compared to more detailed patents in areas like chemical (C), mechanical (F), or electrical (H), avoiding further processing challenges for LLMs, especially in free-versioned models (focus of this work). We refined our search to $((ic = "g06f40/10" \text{ or } ic = "g06f40/20") \text{ or } ic = "g06f40/30") \text{ and } pd = 2023$, resulting in 173 patents. This study explores the complex procedures at the EPO, where examiners use codes (‘X’, ‘Y’, ‘A’, etc.) in the European search report to reference prior art paragraphs during searches. Out of 173 patents found, we selected 49 patent applications (AP), requiring a search report of AP that includes at least one ‘X’ citation (indicating cited prior art capable of negating novelty). We excluded ‘X’ citations⁸ to non-patent literature (NPL) due to their extensive text, often exceeding token limitations for free versions of ChatGPT and Google Gemini. Similarly, we excluded complete document citations to patent literature⁹ to avoid computational issues.
- (ii) **Selection of Citations and Candidate Paragraphs:** (accounting contextual relatedness) Specifically concentrating on ‘X’ citations, the research investigates one-to-many text matching for aligning an independent claim (C1) with a list of candidate prior art paragraphs [P1, P2, ... Pn]. Figure 1 (part a) visually illustrates an example search report, showcasing category codes on the top left. Figure 1 (part b) outlines the methodology for novelty paragraph retrieval, discussed in Section 4. The ‘X’ category paragraphs such as [0024]-[0025], [0037] are listed in the search report of Figure 1. These paragraphs impact novelty or inventive step independently, ‘Y’ category paragraphs do so when combined with other ‘Y’ paragraphs, and ‘A’ category paragraphs provide background without affecting novelty or inventive step. Identifying these paragraphs is time-consuming and demands domain expertise. LLMs must effectively identify these paragraphs, even when confronted with non-cited ones. Given the impracticality of using lengthy lists as inputs for free versions of ChatGPT and Google Gemini, we opt for a selection strategy involving 3 cited paragraphs and 6 contextually similar, non-cited paragraphs from the same prior art (PA), guided by specific rules as follows.
- (a) **Case 1:** For a given ‘X’ citation, if only one list of cited paragraphs within the same PA document, choose 3 cited paragraphs (each at beginning, mid, and end) and 3 non-cited paragraphs each immediately before and after those cited paragraphs (for overall and continuing contextual coverage). This results in 9 candidate paragraphs

⁸EP4276675, <https://data.epo.org/publication-server/pdf-document?pn=4276675&ki=A1&cc=EP&pd=20231115>

⁹EP22181654, <https://data.epo.org/publication-server/pdf-document?pn=4270238&ki=A1&cc=EP&pd=20231101>

- for each examination query in our dataset.
- (b) **Case 2:** For a given 'X' citation with multiple lists of paragraphs cited, select lists referencing claim 1 and apply Case 1. If there are multiple references to multiple lists, choose 3 random cited paragraphs from 3 distinct lists and 6 uncited paragraphs in total before and after the lists, as shown in Figure 1.
 - (c) **Case 3:** If neither Case 1 nor Case 2 is satisfied, select 3 cited and 6 non-cited paragraphs immediately adjacent to cited paragraphs randomly.

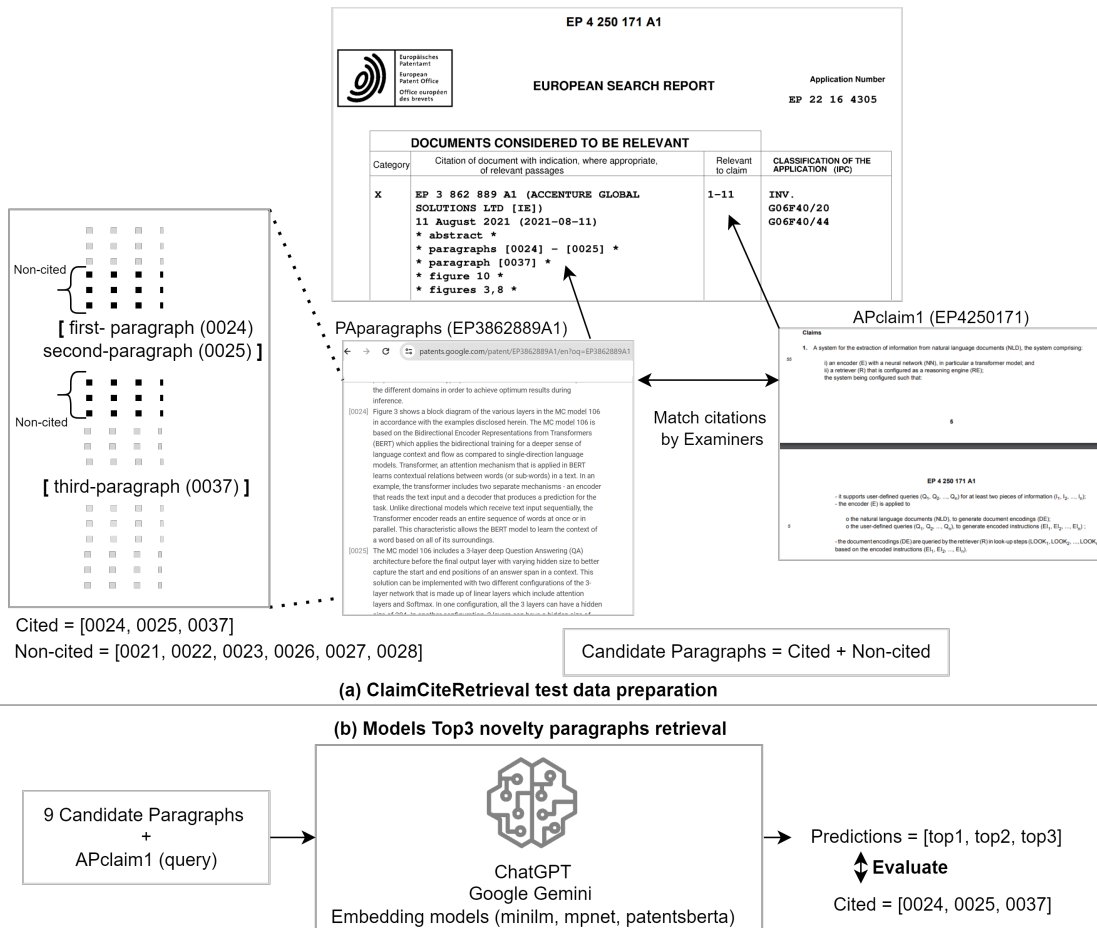


Figure 1: Test Data Preparation and Novelty Paragraph Retrieval Workflow

Analyzing Table 1 in our dataset reveals word counts for different patent text types. Notably, Table 1 indicates significant differences in word counts between abstracts and claims for both applications (AP) and prior art (PA).

Further examination of PA distribution across patent offices in our test data shows that U.S. (approximately 47%) and Chinese (approximately 41%) patent documents generally have longer abstracts and shorter claims compared to European patent documents. To establish an optimal text limit for using free versions of ChatGPT and Google Gemini in retrieval and classification,

Table 1
Text Type Word Counts in the ClaimCiteRetrieval Dataset

Text Type	Min	Max	Avg
APabstract	23	201	121.26
PAabstract	58	216	137.63
APclaim	51	337	166.16
PAclaim	58	421	151.79
PAparagraphs	386	1581	946.38

we tested ranges from 1000 to 4000 tokens (OpenAI¹⁰ claims 100 tokens is around 75 words). Beyond a 1500-word limit, both tools raised “text limit errors.” Therefore, in information retrieval processing, we set an optimal text limit of approximately 1600 words (query + paragraphs). Surprisingly, extractive question answering (e.g., query: when was the electric bulb invented?) can be conducted up to approximately 4000 words. However, given our focus on passage retrieval, we opt to work with approximately 10 paragraphs for each examination for a given patent. Our ClaimCiteRetrieval dataset consists of metadata such as publication numbers, IPC codes, along with different text types mentioned in Table 1.

The usage of this data type spans various settings and methodologies adopted for classification, passage retrieval, and similarity investigation, which we discuss in detail in the next section.

4. Methodology

This section discusses experiments investigating the effectiveness of tools such as ChatGPT and Google Gemini in patent prior art search, covering patent classification, text similarity, and novelty passage retrieval. Two types of tests can be conducted: Online (real-time connectivity of LLMs to web sources, e.g., Gemini, focused on the whole document level) and Offline (e.g., ChatGPT, at least in free versions). Despite Gemini’s real-time web connectivity, in many cases, sources and results cannot be verified, indicating the nature of hallucinations. For uniformity, we chose only the offline method (fine-grained evaluation, focused at the passage level) where, given contextual text and queries to LLMs, models can answer queries solely based on their training data knowledge and language understanding.

Classification and passage retrieval are explicitly experimented with ChatGPT and Google Gemini, while text similarity experiments, although implicit, are conducted to evaluate the complexity of classification and retrieval. State-of-the-art open-source embedding models, including patent-specific (PatentSBERTa) and general-purpose (all-MiniLM-L6-v2 and all-mpnet-base-v2), are employed for similarity and novelty passage retrieval experiments. Since these embedding models lack fine-tuned knowledge on patent classes, they are not used for classification in comparative analysis with ChatGPT and Google Gemini. Selection of these general-purpose embeddings is based on their superior performance and speed, as indicated on the leaderboard of sentence-BERT¹¹ models.

¹⁰Open AI tokenizer:<https://platform.openai.com/tokenizer>

¹¹https://www.sbert.net/docs/pretrained_models.html

For similarity, we compare independent claims and abstracts of given applications and their cited prior art documents using pairs (APclaim-PAclaim and APclaim-PAabstract). Abstracts and independent claims are chosen due to their concise representation of patent innovation. Understanding their semantic similarity is crucial for AI models, as these documents, cited by examiners, demonstrate novelty destruction. AI models should retrieve the cited prior art when the application is used for searching. Calculating similarities between cited and non-cited paragraphs by examiners to APclaim (APclaim-CitedParagraphs and APclaim-NonCitedParagraphs) reveals the complexity of novelty paragraph detection using AI models. Examiner search reports provide gold standard labels for training novelty detection models. Our similarity measures, averaged over cited and non-cited paragraphs, compared with the independent claim, offer empirical evidence on whether examiners' citations alone are sufficient for prior art search training.

For patent classification, we instruct ChatGPT and Google Gemini to classify patent text across various hierarchical levels, such as section, class, sub-class, and sub-groups. This test provides insights into the tools' understanding and differentiation capabilities when handling different subject matters. With nearly 70,000 IPC classification codes in the patent domain, covering diverse technical subjects, our dataset includes text types outlined in Table 1. For each application, we have two texts i.e. APabstract and APclaim. There is no training involved; we solely test these tools by writing effective prompts. The tools are already pre-trained on a large patent corpus, including classification codes and their descriptions. Thus, based on their training knowledge and understanding of patent language, the tools predict IPC classes for a given APabstract and APclaim.

For novelty passage retrieval, as shown in Figure 1 (part b), we gather 9 candidate paragraphs (both cited and non-cited) from prior art for each examination with APclaim (query). These are inputted to ChatGPT and Google Gemini, aiming for the tools to predict the Top 3 most similar paragraphs that could challenge the novelty of the APclaim. The predicted Top 3 paragraphs are compared against the cited (gold truth from examiners) to calculate the accuracies of the models at different levels. To conduct a comparative analysis, we consider three sentence embedding models from the state of the art. Additionally, embeddings are calculated for each candidate paragraph and APclaim. Cosine distance between APclaim and candidate paragraphs is used, and the top 3 closest with the highest similarity values are selected. We then assess these top 3 predictions of all three embeddings modes with ChatGPT and Google Gemini against the truth value cited by examiners for retrieval accuracies.

For evaluation, we assess accuracies in both classification and passage retrieval. In classification, we extract ground truth class codes from each patent application. Given that we predict the top 5 class codes for a given patent text, our goal is to determine how many of the Top 5 values match the ground truth (APipc). For instance, if ChatGPT predicts IPC codes as [p1, p2, p3, p4, p5], and APipc contains codes like [a1, a2, ..n], which can vary in number (assigned by patent offices), at the top 1 level, we check whether p1 matches any code in APipc. Similar accuracy variations are calculated at different top n levels (1 to 5). Metrics like precision, recall, and F1 may not be suitable here due to the limited patent domains covered, and the absence of complete and fixed class labels in the ground truth classification codes. In passage retrieval, we adopt a similar approach where models predict the top 3 values from 9 candidate paragraphs, and accuracies are calculated using PACited as the ground truth. For example, if ground truth

paragraphs of prior art PA_{cited} are [PA1, PA2, PA3], and ChatGPT predicts paragraphs as [Pr1, Pr2, Pr3], we check if Pr1 matches any item in the PA_{cited} list, corresponding to Top 1. We do not consider metrics like Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG), since the selection of 9 candidate paragraphs does not explicitly provide ranks; we randomly place the 3 relevant cited paragraphs in the list. However, since we have ranks in the top 3 most relevant paragraphs predicted by ChatGPT and Gemini, we calculate accuracies at different Top n levels (1 to 3). This demonstrates the accuracy of retrieval systems in making correct predictions at various rank levels. The findings of our work are presented in the next section along with a detailed discussion.

5. Results and Discussions

In this section, we show the results achieved for the experiments relating to similarity, classification, and passage retrieval.

Similarity: Semantic similarities between patent applications and their respective cited prior art patents, based on different text type combinations, are recorded in Table 2. It is clearly visible that the usage of independent claims from both the patent application (AP) and the prior art (PA) is more suitable compared to abstracts for various patent search activities. Even though claims contain legal jargon, they also encompass the most important subject matters for which patents are desired; both patent-specific and non-patent-specific embedding models indicate that claims can be more preferred. In cases where a robust match between AP claims and PA claims is crucial (APclaim-PAclaim), the PatentSBERTa model stands out, exhibiting a narrower range (45.02 to 98.92) and a relatively higher average similarity score (67.60). This model appears particularly well-suited for scenarios where precise alignment of claim language is paramount.

Table 2

Text Type Similarity between Patent Application (AP) and Cited Prior Art (PA)

Model	APclaim-PAclaim			APclaim-PAabstract		
	Min	Max	Avg	Min	Max	Avg
all-MiniLM-L6-v2	18.02	98.73	60.35	25.15	88.63	57.31
PatentSBERTa	45.02	98.92	67.60	<u>43.84</u>	89.18	61.28
all-mpnet-base-v2	27.95	98.51	62.48	31.29	89.93	59.56

On the other hand, if the focus is on comparing AP claims with PA abstracts (APclaim-PAabstract), the all-MiniLM-L6-v2 model demonstrates a competitive edge with a broader range (25.15 to 88.63) and a slightly lower average score (57.31) compared to other two models. The all-mpnet-base-v2 model, while providing respectable performance across both scenarios, falls in between the other two models in terms of similarity scores. The low semantic similarity score (43.84, underlined) achieved by the best model (PatentSBERTa) as shown in Table 2 suggests that the probability of prior art undermining the novelty of a patent application can be even lower than a random guess (50%). This underscores the intricate nature of novelty identification and the associated challenges faced by AI models.

Table 3 indicates that PatentSBERTa outperforms other embedding models in the average similarity when comparing APclaim to both cited and non-cited paragraphs. However, all-mpnet-base-v2 excels in terms of maximum similarities. When considering the best scores (indicated in bold), Table 3 clearly demonstrates a very minimal difference in average similarity between APclaim and cited/non-cited paragraphs. This suggests that distinguishing paragraphs in reference to an independent claim under a novelty test is indeed a challenging task.

Table 3
Text Similarity between APclaim and Cited/Non-Cited Paragraphs by Examiners

Model	APclaim-CitedParagraphs			APclaim-NonCitedParagraphs		
	Min	Max	Avg	Min	Max	Avg
all-MiniLM-L6-v2	19.51	73.42	44.03	17.03	70.49	39.87
PatentSBERTa	36.88	80.57	51.88	33.55	69.60	48.21
all-mpnet-base-v2	22.82	81.93	49.85	27.13	71.24	45.71

Classification: In IPC classification¹², there are nearly 70,000 codes available at the sub-group level and around 7314, 640, 129, and 8 at the main-group, sub-class, class, and section levels, respectively. Out of the total mentioned codes, ChatGPT and Gemini are tasked with predicting the Top 5 most suitable codes for a given claim or abstract. As we delve deeper into the levels, the challenge in prediction increases. For simplicity, after the sub-class level, we did not predict the main-group; instead, we prompted models to predict sub-groups.

Both ChatGPT and Gemini were evaluated for classifying APabstract and APclaim, and the results for these four combinations are presented in Table 4. For each combination of the model and text type, the best average scores are highlighted, revealing that Google Gemini outperforms (at Top1) all others with an average accuracy of 72.96% when an independent claim is used. ChatGPT outperforms other combinations at the Top levels (2-4) when used with abstract and claim. The best average scores at each section, class, sub-class, and sub-group levels are underlined. ChatGPT and Gemini performed equally at the section level, but for the rest of the levels, ChatGPT outperformed Gemini. Both ChatGPT and Gemini demonstrate their inadequacy and unsuitability for classification at the sub-group level. This can be particularly challenging as models need to predict out of nearly 70,000 sub-group codes without being fine-tuned for classification. Therefore, these models performed well only until the sub-class level. In summary, Table 4 shows, that ChatGPT performed well compared to Gemini with an overall performance (boxed scores) of 66.28% accuracy when claims are used. This work also indicates that claims contribute better to classifying the patent compared to abstracts.

Novelty passage retrieval: Table 5 shows the average accuracies of all models in novelty passage retrieval. It presents individual accuracies at each Top level (1-3), indicating that ChatGPT outperformed all other models. In general, both ChatGPT and Google Gemini surpass other embedding models. Specifically, among the embedding models, PatentSBERTa outperforms the other two and is almost as effective as Google Gemini.

For additional information, this work includes details on unusual examples in European examinations and user experiences with ChatGPT and Google Gemini in the GIT¹³ repository.

¹²<https://www.wipo.int/classifications/ipc/en/>

¹³<https://github.com/Renuk9390/ChatGPTvsGoogleGemini>

Table 4
Classification Accuracies (%) of ChatGPT and GoogleGEMINI

Model	Top Level	Section	Class	Sub-Class	Sub-Group	Average
APipc_claim_ChatGPTpred	Top1	100	100	48.98	10.20	64.79
	Top2	100	100	75.51	8.16	70.91
	Top3	93.88	93.88	83.67	0.00	67.85
	Top4	95.92	87.76	65.31	10.20	64.79
	Top5	100	89.80	59.18	2.04	62.75
	Average	<u>97.96</u>	94.28	<u>66.53</u>	<u>6.12</u>	
Overall performance						66.28
APipc_claim_GEMINIpred	Top1	100	100	89.80	2.04	72.96
	Top2	89.88	83.67	53.06	2.04	57.16
	Top3	65.31	65.31	48.98	0.00	44.90
	Top4	83.67	71.43	59.18	8.16	55.61
	Top5	100	73.47	55.10	0.00	57.14
	Average	87.77	78.77	61.22	2.44	
Overall performance						57.55
APipc_abstract_ChatGPTpred	Top1	95.92	95.92	59.18	4.08	63.77
	Top2	100	97.96	67.35	0.00	66.32
	Top3	97.96	97.96	91.84	0.00	71.94
	Top4	100	93.88	42.86	0.00	59.18
	Top5	95.92	87.76	44.90	0.00	57.14
	Average	<u>97.96</u>	<u>94.69</u>	61.22	0.81	
Overall performance						63.67
APipc_abstract_GEMINIpred	Top1	100	79.59	75.51	0.00	57.14
	Top2	100	95.92	44.90	4.08	46.94
	Top3	95.92	87.76	73.47	2.04	64.79
	Top4	69.39	67.35	51.02	0.00	61.22
	Top5	100	65.31	61.22	2.04	63.77
	Average	93.06	79.18	61.22	1.63	
Overall performance						58.57

Table 5
Novelty Passage Retrieval Accuracies (%) across Various Models

Model	Top1	Top2	Top3	Average
ChatGPT	55.10	48.98	44.90	49.66
GoogleGEMINI	48.98	40.82	51.02	46.94
all-MiniLM-L6-v2	51.02	34.69	30.61	38.77
PatentSBERTa	51.02	36.73	44.89	44.21
all-mpnet-base-v2	53.06	38.77	34.69	42.17

The repository also contains codes, dataset, and optimized prompts for use with the API of ChatGPT and Gemini. The conclusion of our work introduces potential areas for future research, which are discussed in the next section.

6. Conclusion

To assess AI frontiers' effectiveness (ChatGPT and Google Gemini), we created a test dataset for artificial examination called ClaimCiteRetrieval. We tested these tools for patent classification at various hierarchies and novelty passage retrieval. For comparison, we used state-of-the-art embedding methods. ChatGPT outperformed all other models in both classification and passage retrieval. Despite their success, these models still face challenges in retrieving examiners' cited paragraphs that may diminish the novelty of a given prior art. Based on empirical evidence, the average similarities between claims and paragraphs in European search reports are considerably lower compared to the similarities between claims and abstracts in both patent applications and prior art. This comparison leads to the conclusion that abstracts and claims from prior art can be more effective than individual paragraphs in training models for patent retrieval or general prior art search applications focused on novelty.

The core idea is to consider matching the abstracts and claims of cited prior art with those of the claimed invention. Subsequently, these matched pairs can be used in various semantic matching methods, such as using a triplet text format to generate embeddings [11]. Since independent claims constitute the core invention part, collectively representing the entire patent and its description, they are crucial. Independent claims also use legal terminology, which is an essential aspect for handling legal data. On the other hand, the abstracts of patent documents provide a concise technical summary of the entire patent and usually contain little or no legal jargon, making them suitable for simulating or mimicking user queries for prior art search engines. Therefore, we view this proposed research direction as potential future work for this study. We anticipate that exploring this research agenda will generate interest among researchers to develop intelligent tools for prior art search.

Acknowledgments

This research is part of the project "BigScience", which is funded by the Bavarian State Ministry for Economic Affairs, Regional Development, and Energy under the grant number DIK0259/01.

References

- [1] K. Vowinckel, V. D. Hähnke, SEARCHFORMER: Semantic Patent Embeddings by Siamese Transformers for Prior Art Search, *World Patent Information* 73 (2023) 102192.
- [2] M. Masalkhi, J. Ong, E. Waisberg, A. G. Lee, Google DeepMind's Gemini AI versus ChatGPT: A Comparative Analysis in Ophthalmology, *Eye* (2024) 1–6.
- [3] J. Liang, Can ChatGPT Be Used for Patent Search Work?, 2023. URL: <https://ipwatchdog.com/2023/03/20/can-chatgpt-used-patent-search-work/id=158018/>, iPWatchdog.
- [4] S. Robertson, H. Zaragoza, et al., The Probabilistic Relevance Framework: BM25 and Beyond, *Foundations and Trends® in Information Retrieval* 3 (2009) 333–389.
- [5] V. Stamatis, End to End Neural Retrieval for Patent Prior Art Search, in: *European Conference on Information Retrieval*, Springer, 2022, pp. 537–544.

- [6] R. Setchi, I. Spasić, J. Morgan, C. Harrison, R. Corken, Artificial Intelligence for Patent Prior Art Searching, *World Patent Information* 64 (2021) 102021.
- [7] L. Helmers, F. Horn, F. Biegler, T. Oppermann, K.-R. Müller, Automating the Search for a Patent's Prior Art With a Full Text Similarity Search, *PloS one* 14 (2019) e0212103.
- [8] J. Risch, R. Krestel, Domain-Specific Word Embeddings for Patent Classification, *Data Technologies and Applications* 53 (2019) 108–122.
- [9] J. Risch, N. Alder, C. Hewel, R. Krestel, Patentmatch: A Dataset for Matching Patent Claims & Prior Art, *arXiv preprint arXiv:2012.13919* (2020).
- [10] R. Chikkamath, M. Endres, L. Bayyapu, C. Hewel, An Empirical Study on Patent Novelty Detection: A Novel Approach Using Machine Learning and Natural Language Processing, in: *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, IEEE, 2020, pp. 1–7.
- [11] N. Reimers, I. Gurevych, Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks, *arXiv preprint arXiv:1908.10084* (2019).
- [12] H. Bekamiri, D. S. Hain, R. Jurowetzki, Patentsberta: A Deep NLP Based Hybrid Model for Patent Distance and Classification Using Augmented Sbert, *arXiv preprint arXiv:2103.11933* (2021).
- [13] R. Srebrovic, J. Yonamine, Leveraging the BERT Algorithm for Patents With TensorFlow and BigQuery, *White paper, Google* (2020).
- [14] S. Ragot, A Novel Approach to Measuring Patent Claim Scope Based on Probabilities Obtained From (Large) Language Models, *arXiv preprint arXiv:2309.10003* (2023).
- [15] K. Vaish, P. Rawat, S. Kathuria, R. Singh, K. Joshi, A. Verma, Artificial Intelligence Reducing the Intricacies of Patent Prior Art Search, in: *2023, CISES, IEEE, 2023*, pp. 978–982.
- [16] R. Chikkamath, V. R. Parmar, Y. Otiefy, M. Endres, Patent Classification Using BERT-for-Patents on USPTO, in: *Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing, 2022*, pp. 20–28.
- [17] H. Jang, S. Kim, B. Yoon, An EXplainable AI (XAI) Model for Text-Based Patent Novelty Analysis, *Expert Systems with Applications* (2023) 120839.
- [18] R. Chikkamath, R. F. Ali, C. Hewel, M. Endres, Explainable Artificial Intelligence for Highlighting and Searching in Patent Text, *PatentSemTech23,; 4th Workshop on Patent Text Mining and Semantic Technologies, colocated with the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 27th, 2023, Taipei, Taiwan.* (2023).
- [19] Z. Wang, Y. Liu, SEA-PS: Semantic Embedding with Attention to Measuring Patent Similarity by Leveraging Various Text Fields, *Journal of Information Science* (2022) 01655515221106651.
- [20] K. A. Younge, J. M. Kuhn, Patent-to-Patent Similarity: A Vector Space Model, Available at SSRN 2709238 (2016).
- [21] I. Ahmed, A. Roy, M. Kajol, U. Hasan, P. P. Datta, M. R. Reza, ChatGPT vs. Bard: A Comparative Study, *Authorea Preprints* (2023).
- [22] S. K. Singh, S. Kumar, P. S. Mehra, Chat GPT & Google Bard AI: A Review, in: *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*, IEEE, 2023, pp. 1–6.