Is it novel and why? Fine-grained patent novelty prediction based on passage retrieval

Valentin Knappich, Anna Hätty, Simon Razniewski, Annemarie Friedrich

Angaben zur Veröffentlichung / Publication details:

Knappich, Valentin, Anna Hätty, Simon Razniewski, and Annemarie Friedrich. 2026. "Is it novel and why? Fine-grained patent novelty prediction based on passage retrieval." In *SIGIR '26: proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, VIC, Australia, July 20-24, 2026*, edited by Alistair Moffat, Falk Scholer, Hannah Bast, Marc Najork, and Min Zhang, 845–56. New York, NY: ACM. https://doi.org/10.1145/3805712.3809576.

Is It Novel and Why? Fine-Grained Patent Novelty Prediction Based on Passage Retrieval

Valentin Knappich
Bosch Center for AI
Stuttgart, Germany
University of Augsburg
Augsburg, Germany
valentin.knappich@de.bosch.com

Anna Hätty
Bosch Center for AI
Stuttgart, Germany
anna.haetty@de.bosch.com

Simon Razniewski
ScaDS.AI & TU Dresden
Dresden, Germany
simon.razniewski@tu-dresden.de

Annemarie Friedrich
University of Augsburg
Augsburg, Germany
annemarie.friedrich@uni-a.de

Abstract

Novelty assessment is a critical yet complex task in the examination process for patent acceptance, requiring examiners to determine whether an invention is disclosed in a prior art document. The process involves intricate matching between specific features of a patent claim and passages in the prior art. While prior work has approached novelty prediction primarily as a binary classification task at the claim level, we argue that this formulation is susceptible to spurious correlations and lacks the granularity required for practical application. In this work, we introduce FiNE-Patents (Fine-grained Novelty Examination of Patents), a novel dataset comprising 3,658 first patent claims annotated with fine-grained, feature-level prior art references extracted from European Search Opinion (ESOP) documents. We propose shifting the evaluation paradigm from simple binary classification to a joint retrieval and abstract reasoning task at the feature level, requiring models to identify specific passages from a prior art document that disclose individual claim features, and to identify which features of a claim make it novel. We implement and evaluate LLM-based workflows that decompose claims into features, analyze each feature against prior art, and finally derive a claim-level novelty prediction. Our experiments demonstrate that these workflows outperform embedding-based baselines on passage retrieval and novel feature identification. Furthermore, we show that unlike trained classifiers, LLMs are robust against spurious correlations present in the claim-level novelty classification task. We release the dataset and code to foster further research into transparent and granular patent analysis.

CCS Concepts

• **Computing methodologies** → **Information extraction**; **Natural language generation**; **Search methodologies**.

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809576

Keywords

Large Language Models, LLM, Passage Retrieval, Patent Analysis, Novelty Prediction, Explainable AI

ACM Reference Format:
Valentin Knappich, Anna Hätty, Simon Razniewski, and Annemarie Friedrich. 2026. Is It Novel and Why? Fine-Grained Patent Novelty Prediction Based on Passage Retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3805712.3809576

1 Introduction

Patents play a crucial role in protecting inventions and fostering innovation across various industries. Computational support for the patent domain has been actively studied for decades, with particularly strong activity in prior-art search and patent classification [2, 11, 30, 41, 45, 51]. Patent novelty prediction has recently received increased attention [23, 25, 36, 57] and aims to predict whether the invention defined by a patent has already been disclosed in prior work. Within patents, the set of claims defines the scope of the legal protection sought by the applicant. Each claim is a single sentence defining the invention as a combination of multiple *features*, where each feature is a component, step, or characteristic of the invention. For example, in a claim describing a new text classification method, individual features could include f_1 = *converting text into tokens*, f_2 = *applying a neural network to the tokens*, and f_3 = *mapping embeddings into the label space.* An invention is considered novel over a *prior art document* if at least one feature of the claim is not disclosed in the prior art document. Predicting patent novelty thus requires comparing each feature of a claim against the prior art, demanding both deep technical understanding and highly abstract reasoning.

Prior work has primarily evaluated novelty prediction at the claim level as a binary classification task [10, 23, 36, 44, 47]. However, in most setups, this is highly susceptible to spurious correlations leading to overly optimistic results of trained models [10, 23], even after stratifying for claim length and removing the strongest biases. To analyze the spurious correlations, we create an adversarial test set using adversarial filtering [32] to detect whether models

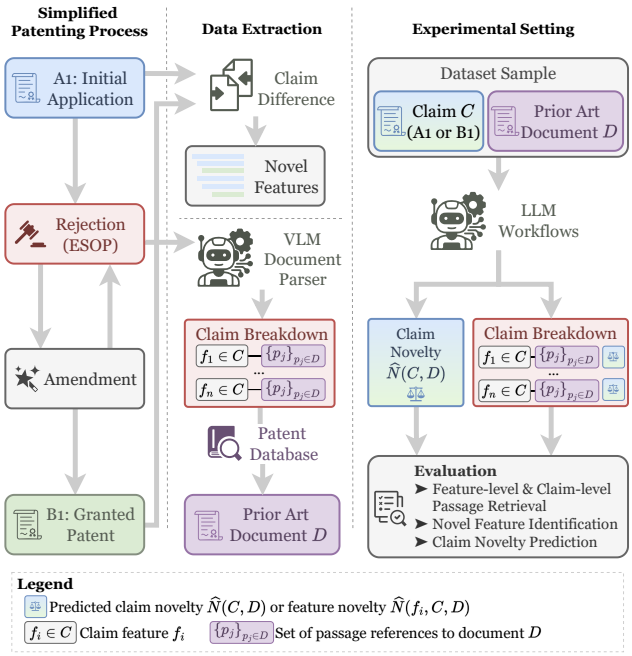

Figure 1: Methodology Overview. [Left] Simplified Patenting Process: The initial application (A1) is rejected due to lack of novelty in a European Search Opinion (ESOP) document and amended (typically at least once) until the granted patent (B1) is issued. [Middle] Data Extraction: We extract feature-level prior art references from the ESOP and retrieve the prior art document cited as the closest prior art D. [Right] Experimental Setting: Given a claim C (either from the initial or the granted version) and the closest prior art document, the model retrieves relevant paragraphs $\{p_j\}_{p_j \in D}$ per claim feature f_i, determines whether each feature is disclosed in D, and predicts whether the claim C is novel over D.

make use of them. Interestingly, we find that zero-shot Large Language Models (LLMs) do not rely on these shortcuts, achieving similar performance on the adversarial test set as on the regular test set.

Crucially, we argue that patent novelty examination should also be evaluated at the feature level. A claim novelty label provides little to no insight into the specific aspects of the invention that might be disclosed by a prior art document, making it of limited practical use. Instead, we propose a more nuanced evaluation protocol that requires models to (i) identify which passages are relevant for a given feature (*passage retrieval*), (ii) determine which features are novel (*novel feature identification*), and (iii) predict whether the claim is novel (*claim novelty prediction*). This approach provides patent professionals with more actionable and trustworthy support, as they can transparently trace and verify the model's reasoning by directly inspecting the cited source passages underlying each identified match. Figure 1 provides an overview of our methodology.

To support this evaluation framework, we introduce the FiNE-Patents dataset based on examination records from the European Patent Office (EPO). In the European patent system, the initial application document (referred to as *A1*) undergoes examination and is typically amended one or more times in response to examiner objections before a patent is finally granted (referred to as *B1*).[1] When a claim lacks novelty, the examiner writes a European Search Opinion (ESOP) document explaining the rejection. The ESOP cites the most relevant prior art document (referred to as D) and provides a detailed *breakdown* of the claim into features, describing which passages of D disclose which features of the claim. We extract these feature-level prior art references from the ESOP documents to evaluate models' novelty analyses. An example of the extraction process is shown in Figure 2.

To generate detailed novelty analyses, we implement workflows using LLMs to output both claim-level novelty predictions and feature-level passage references. We use either just a single LLM call, or analyze each feature separately and then aggregate the results. We find that our methods outperform baselines based on embedding similarity on retrieval metrics and that they do not rely on shortcuts based on spurious correlations for novelty prediction.

We summarize our contributions as follows:

(1) We propose to evaluate patent novelty examination across three sub-tasks: *passage retrieval* at the feature and claim level, *novel feature identification*, and claim-level *novelty prediction*.
(2) We present FiNE-Patents, a dataset of 3,658 first patent claims with feature-level prior art passage references and novelty labels.
(3) We implement LLM-based workflows to predict patent novelty that split the claims into features, analyze each feature separately, and then aggregate the results.
(4) We publicly release the dataset, code to create the dataset, and code to reproduce our experiments.[2]

2 Related Work

Novelty assessment is a fundamental challenge across multiple domains, from evaluating creative ideas [9, 43, 48] and scientific research contributions [1, 21, 38, 55, 56, 61] to assessing patent applications. In the patent domain, recent advances in large language models have opened new possibilities for supporting various tasks, including prior art search [15, 57], patent drafting [28, 53], analysis [27, 30, 51, 54], and novelty prediction [10, 22, 23, 33, 36, 40, 44, 47, 49, 50, 57]. Our work builds upon this emerging body of research by introducing fine-grained, feature-level supervision for patent novelty examination.

Patent Grant Prediction. Predicting the likelihood of a patent application being granted has been one of the most prevalent tasks in patent NLP [25]. Prior works have either predicted the final grant decision as a whole [22, 40, 49], or separately predicted novelty and/or inventiveness [10, 23, 33, 36, 44, 47, 50, 57], lack of definiteness [27], or insufficient enablement [29, 59]. Lim et al. [36] introduce the PANORAMA dataset for patent novelty and inventiveness prediction, which contains claim-level passage references from US rejection documents.

[1] https://register.epo.org/help?lng=en&topic=kindcodes
[2] https://github.com/boschresearch/fine-patents

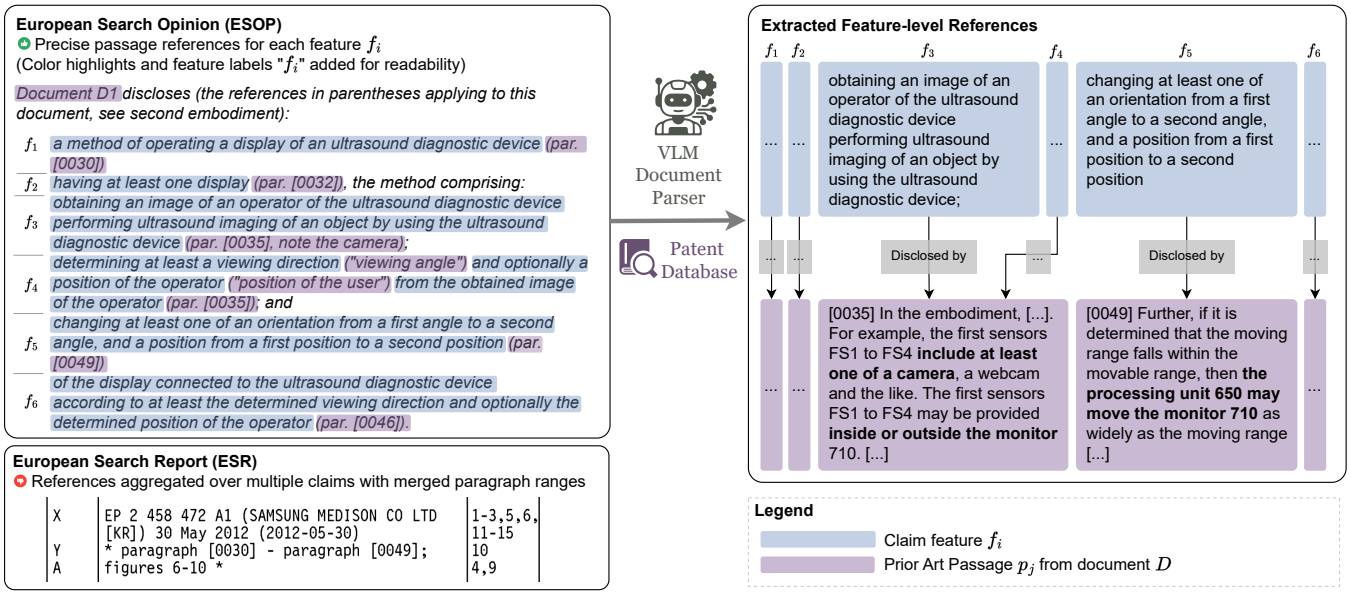

Figure 2: Example of the feature-level prior art references. [Top left] In the ESOP document, the examiner rejects claim 1 for lack of novelty over prior art document D (here: D1) and provides feature-level references to D (claim features highlighted in blue, references in purple). [Bottom left] The European Search Report (ESR) commonly used by prior work lists references aggregated over multiple claims with merged paragraph ranges and without mapping passages to features. [Right] The extracted feature-level references for claim 1, showing which passages in D disclose which features of the claim.

Patent Passage Retrieval. The retrieval of relevant patent prior art passages has been studied in various works [3, 5, 6, 17], most notably based on the CLEF-IP 2012 [4], CLEF-IP 2013 [46], and NTCIR 2005 [18, 19] datasets. They study the retrieval of relevant passages from a large corpus of patent documents given a query claim. In contrast, our work focuses on finding relevant passages from a single prior art document and uses feature-level labels extracted from office action full texts.

3 FiNE-Patents Dataset

In this section, we describe the FiNE-Patents dataset, whose statistics are summarized in Table 1. The dataset contains examination data from 3,163 patent applications filed at the EPO between 2012 and November 2025, comprising 3,658 first claims with an even split between novel and not novel claims. For each claim, the dataset includes the first ESOP document, written by the examiner and parsed into a structured format, as well as the full text of the closest prior art document cited in the ESOP. In the following, we provide details on the task definition and evaluation protocol (Section 3.1), the data source (Section 3.2), the dataset creation process (Section 3.3), an analysis of spurious correlations (Section 3.4), and an analysis of data contamination (Section 3.5).

3.1 Task Definitions and Evaluation Protocol

We propose to evaluate models on three sub-tasks: retrieval of relevant prior art passages from the closest prior art document, the identification of novel features in a claim, and claim novelty prediction. The input is a claim under examination C and its closest prior art document D consisting of passages $p_j \in D$. Each claim C consists of features f_i as defined by a segmentation $S(C) = \{f_1, f_2, \ldots, f_n\}$. A gold feature corresponds to the span of the claim for which the examiner has provided references in the ESOP document. For generality, we do not assume access to the gold segmentation at inference time, i.e., models must first produce their own segmentation $\hat{S}(C)$.

3.1.1 Passage Retrieval. The first sub-task is the retrieval of relevant passages from the prior art document D that disclose the features f_i of the claim C. That is, for each feature $f_i \in \hat{S}(C)$, the full claim C, and the prior art document D, the model must retrieve passages p_j that disclose f_i. We refer to the set of passages cited in the ESOP as $P(f_i, C, D) = \{p_j \in D \mid p_j \text{ discloses } f_i\}$ and to the set of passages retrieved by the model as $\hat{P}(f_i, C, D)$. For evaluation at the feature level, we compute the precision (P), recall (R), F1, and nDCG [24] scores between $P(f_i, C, D)$ and $\hat{P}(f_i, C, D)$ by mapping $f_i \in \hat{S}(C)$ to the feature in $S(C)$ with minimal edit distance. At the claim level, we aggregate the passages over all features $P(C, D) = \bigcup_{f_i \in S(C)} P(f_i, C, D)$, i.e., $P(C, D)$ contains all passages that disclose at least one feature of the claim.

While cited passages are reliably indicative of disclosure, the absence of a citation does not necessarily imply non-disclosure, i.e., negatives may not be true negatives. To mitigate this issue, we additionally report soft variants of precision, recall, and F1 ($\tilde{\text{P}}$, $\tilde{\text{R}}$, $\tilde{\text{F1}}$) [16]. To obtain them, we compute the ROUGE-L [37] score between each predicted passage and each cited passage, and aggregate:

Statistic		Count
# Claims	All	3658; 50.0% novel
	Train	1454 (39.7%); 50.1% novel
	Val	360 (9.8%); 52.2% novel
	Test	1844 (50.4%); 49.5% novel
	Adversarial Test	278 (7.6%); 50.0% novel
# Applications	All	3163
	Rejected claim only	1334 (42.2%)
	Granted claim only	1334 (42.2%)
	Both claims	495 (15.6%)
# Features per Claim	All	6.2 ± 2.4
	w/ References	4.5 ± 2.2
Cited Patents	All	3658
	US	2658 (72.7%)
	EP	996 (27.2%)
	# paragraphs	115.9 ± 90.7
	# claims	21.9 ± 23.4
	# words	11662.8 ± 8963.4
# Cited Passages	Total	21464
	Per Claim	11.7 ± 25.0
	Per Feature	1.9 ± 4.4
	Paragraph	20166 (94.0%)
	Claim	1020 (4.8%)
	Abstract	278 (1.3%)

Table 1: Dataset Statistics. Averages are reported as mean ± standard deviation. *Adversarial Test* is a subset of *Test*.

$$\tilde{\mathrm{P}} = \frac{1}{|\hat{P}(f_i, C, D)|} \sum_{p_j \in \hat{P}(f_i,C,D)} \max_{p_k \in P(f_i,C,D)} \text{ROUGE-L}(p_j, p_k)$$

$$\tilde{\mathrm{R}} = \frac{1}{|P(f_i, C, D)|} \sum_{p_k \in P(f_i,C,D)} \max_{p_j \in \hat{P}(f_i,C,D)} \text{ROUGE-L}(p_j, p_k)$$

This can be seen as a soft generalization of precision and recall, where each predicted passage receives partial credit based on its maximum overlap with any ground truth passage (for precision), and symmetrically, each ground truth passage is scored by its maximum overlap with any predicted passage (for recall). We argue that lexical overlap is a reasonable proxy for soft labels: while matching claim features to prior art passages across different applications requires bridging the vocabulary gap, the terminology used within a single prior art document is largely consistent, making ROUGE-L a suitable similarity measure. Since ESOP documents always refer to the initially filed version of the claim, we compute retrieval scores only for the claims labelled as not novel.

3.1.2 Novel Feature Identification. At the core of novelty assessment is the identification of novel features. Given the model's binary novelty prediction $\hat{N}(f_i, C, D) \in \{0, 1\}$ for each feature $f_i \in \hat{S}(C)$, we extract the character ranges of features predicted as novel and compare them to the character ranges actually added during prosecution. We report precision (P), recall (R), and F1 scores. Since added features are only available for the granted version of the claims, we compute these scores only for the claims labelled as novel.

3.1.3 Claim Novelty Prediction. The final sub-task is the binary classification of whether the claim C is novel over the prior art document D, referred to as $\hat{N}(C, D)$. We report the fraction of claims predicted to be novel, accuracy, and macro F1.

3.2 Data Source

Prior works have utilized the paragraph reference information from *European Search Reports (ESRs)* to construct patent novelty datasets [10, 47]. An example is shown in the bottom left of Figure 2. ESRs list the relevant prior art documents, categorize them by their relevance to novelty and inventive step,[3] and reference the relevant passages in the prior art. They are, unlike ESOPs, available at scale in a structured format. However, the references are often very coarse-grained. By comparing our parsed ESOP citations to the ESRs, we find that only 19% of the passages cited in the ESR are also cited in the ESOP, indicating a large discrepancy between the two sources of supervision. Upon inspection of examples, we find that this discrepancy has two main causes. First, the ESR citations are often aggregated over multiple claims, losing information about which passage relates to which claim. Second, examiners often merge ranges for brevity, e.g., citing paragraphs `10-20` instead of `10-13, 16-18, 20`. In contrast, ESOPs provide precise references at the feature level, and are thus the superior source of supervision at the cost of expensive PDF parsing.

3.3 Dataset Creation

In the following, we describe the dataset creation process, which consists of four main steps: (i) retrieval of seed patent applications, (ii) parsing of ESOP documents to extract feature-level prior art references, (iii) retrieval of the full text of the cited prior art documents, and (iv) extraction of newly added features from the initial to the granted claim.

3.3.1 Seed Patent Applications. We start the dataset creation process by retrieving all EPO patent applications from the CPC classes `G06`[4], `G10L`[5], `H04L`[6], and `H04N`[7] from 2012 onwards. These classes were chosen such that the information retrieval community is familiar with the domain of the patents, making it easier to understand the examination. To exclude low-quality applications, we filter for applications that were finally granted. Patent application and publication numbers are retrieved from the EPO's Linked Data service[8] and publication full texts are retrieved from the EPO's Publication Server API[9].

3.3.2 ESOP Parsing. For each seed patent application, we download, if available, the first *European Search Opinion* document from

[3] https://www.epo.org/en/legal/guidelines-pct/2025/b_x_9_2.html
[4] G06: Computing; Calculating; Counting
[5] G10L: Speech Analysis Techniques Or Speech Synthesis; Speech Recognition; Speech Or Voice Processing Techniques; Speech Or Audio Coding Or Decoding
[6] H04L: Transmission Of Digital Information
[7] H04N: Pictorial Communication
[8] https://data.epo.org/linked-data/about
[9] https://www.epo.org/en/searching-for-patents/technical/publication-server

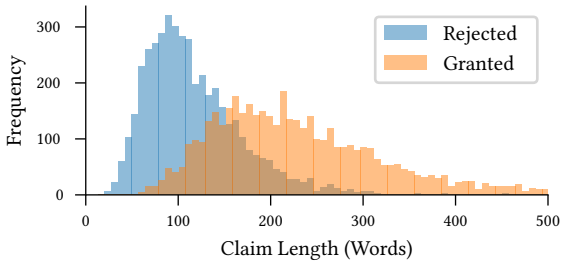

Figure 3: Histogram of claim lengths before stratification. The length of the claim under examination provides a strong shortcut for predicting novelty. After stratified sampling, each class has the distribution of the overlap.

the EPO register.[10] These documents contain the examiner's reasoning for rejecting a subset of the claims. For rejections based on lack of novelty, the examiner recites the rejected claim and provides precise references to the closest prior art document D for every feature of the claim. Our analysis centers on the first claim of each application, as it is generally the broadest and most significant independent claim, and is most frequently examined in depth within the ESOP. We extract these feature-level prior art references using the vision language model (VLM) `Qwen3-VL-235B-A22B-Thinking` [7] with entirely image-based input and structured output. This end-to-end approach avoids the accumulation of errors in multi-step pipelines including OCR and yields highly accurate results, but it is costly to scale. To minimize costs, we first filter for documents that contain a breakdown of the first claim using the highly efficient OCR model MinerU 2.5[11] [42] and regular expressions that match the claim text with additional references in parentheses. Figure 2 shows an example of an ESOP document on the left and a visualization of the structured output on the right.

3.3.3 Prior Art Documents. For each prior art document cited in the ESOP, we retrieve the full text from the EPO Publication Server API for all available jurisdictions. For U.S. patents cited as prior art, we obtain the full text from the USPTO bulk data dump.[12] To limit the complexity of the pipeline, we only consider prior art patents filed at the EPO or USPTO, leaving the inclusion of patents from other jurisdictions and non-patent literature to future work.

3.3.4 Newly Added Features. To evaluate novel feature identification, we leverage the edit operations performed by applicants during prosecution. Since applicants typically add new limitations to overcome novelty objections, character spans added between the initial and granted claim serve as a strong proxy for novel features. We determine the newly added character spans in the finally granted claim in B1 compared to the initially filed claim in A1 using a simple diff algorithm based on character-level Levenshtein distance [34]. That is, we determine the character ranges in the claim that were added during prosecution. To evaluate, we compute the overlap of these ranges with the features identified as novel by a model as described above.

3.3.5 Filtering Statistics. The dataset creation described above represents a complex pipeline with multiple steps, at each of which a percentage of the samples is discarded. The initial set of seed applications from the selected CPC classes contained 364k patent applications, out of which 132k were ultimately granted. Of these granted applications, 67k have complete publication full texts available in the EPO Publication Server. 65k additionally have at least one ESOP document available. Of these 65k ESOP documents, 19k recite the full first claim with feature-level prior art references, of which 17k are successfully parsed using our VLM document parser. After filtering out rejections due to lack of inventive step, samples with unavailable prior art documents and samples with unlocatable references, 4.4k applications remain. Since each application contributes both an initial and a granted claim, this amounts to 8.8k claims. We additionally perform a claim length stratification as described below, after which 3,658 claims remain. Lastly, we split the dataset into training, validation, and test sets with a 40/10/50 ratio.

3.4 Spurious Correlation Mitigation

A common concern is that models may exploit spurious correlations in the dataset to achieve high classification performance without performing the intended task [20]. In our setup, the claim-level novelty prediction task is particularly susceptible to this issue, because there are superficial differences between the initial and granted claims. In the following, we describe our efforts to mitigate spurious correlations in the dataset and analyze the remaining ones.

3.4.1 Length Stratification. As prior work has noted, the claim length is a strong signal for novelty ([23], see Figure 3) because applicants typically add limitations to overcome novelty objections. To mitigate this effect, we stratify the dataset based on the length of the claim into 100 bins and sub-sample the majority class, such that all bins contain an equal number of novel and not novel samples. Figure 3 shows the claim length distribution before stratification. As a result of our sampling procedure, both classes share the same claim length distribution, shown as the overlapping region in the figure.

3.4.2 Reference Numerals and Patent Numbers. Patent documents often use reference numerals to refer to specific parts of the invention across the claims, description, and drawings (e.g., "a device (10) comprising..."). Upon inspection of the dataset, we find that they are often omitted in the initial application and added during prosecution. On average, granted claims contain 12.1 numerals while initial claims contain only 0.3. Since they are easy to detect with regular expressions, we remove them from all claims in the dataset.

We additionally note that the descriptions of granted patents contain more patent number references than those of initial applications (2.1 vs. 0.2 on average). In particular, the closest prior art document cited in the ESOP is referenced 0.75 times on average in granted patents but almost never in initial applications. Since we do not provide the description of the application under examination to the models, this does not affect our experimental setup but should be considered in settings that include the description as input.

[10] https://register.epo.org/
[11] https://huggingface.co/opendatalab/MinerU2.5-2509-1.2B
[12] https://bulkdata.uspto.gov

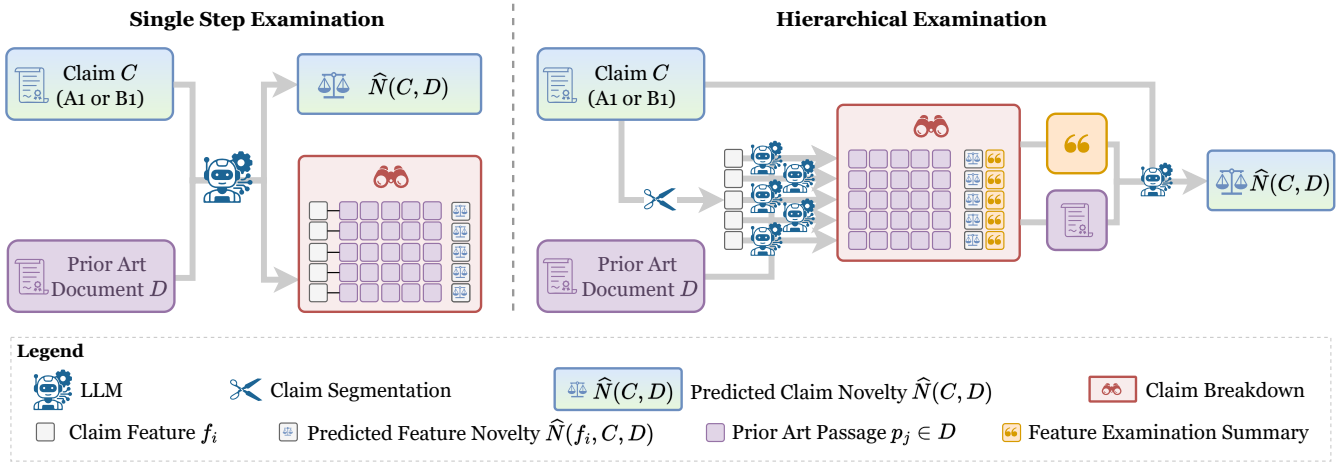

Figure 4: LLM Workflows. [Left] In *Single Step Examination*, the model receives the claim and full prior art document, and outputs the breakdown and claim novelty prediction. [Right] In *Hierarchical Examination*, the claim is first segmented into features, each feature is examined separately against the prior art document, and finally, the results are aggregated into a claim novelty prediction.

3.4.3 Adversarial Test Set. While the above steps mitigate the strongest shortcuts, more subtle spurious correlations still remain. We train a BERT-base classifier [14] on the training set using only the claim text as input and find that it achieves a test accuracy of 75%. Since meaningful predictions without considering the prior art are impossible, this classifier must be relying on shortcuts. To identify these shortcuts, we use an interpretable logistic regression model in Section 5.2.

To be able to determine whether a model relies on these shortcuts, we create an adversarial test set using adversarial filtering [32] where models overfitted to these spurious correlations fail. Specifically, we filter the test set, keeping only samples that are misclassified by the trained BERT model. To ensure comparability with results from the full test set, we balance the dataset by randomly removing samples from the majority class. Since it contains only 278 claims, it is not well-suited to evaluate and compare models in detail, but a substantial decrease in accuracy compared to the full test set is a strong indicator of overfitting to spurious correlations.

3.5 Data Contamination Analysis

Data contamination is a common concern when evaluating LLMs on domain-specific tasks, especially when the dataset is created from publicly available documents that may have been included in the pre-training data [8, 13, 35]. We use INFINI-GRAM [39] to search for contents of our dataset in the largest indexed corpus,[13] which was used to train OLMo 2 [52] and contains 4.6T tokens across 3M documents. Our analysis shows that the relevant public data of our benchmark is not leaked into the largest openly indexed pre-training corpus. First, we search for the phrase "*The subject-matter of claim*" which appears verbatim in the majority of ESOP documents. Only 58 documents in the index contain this phrase, the majority of which are EPO case law proceedings,[14] and none of them are ESOP documents. Second, we search for the exact text of the claims in our dataset. We sample 100 claims, search for the first 5 words of each claim to avoid false negatives due to whitespace and punctuation differences, and manually verify whether matches contain the claim text. Of the 100 sampled claims, none appear in the index; in many cases, even the first 5 words yield zero matches, despite consisting of generic phrases such as *"An apparatus for communicating in."* Future work should conduct further analysis like time-based data splits [28] to definitively rule out data contamination for models with closed pre-training corpora.

[13] `v4_olmo-2-0325-32b-instruct_llama`
[14] e.g. https://www.epo.org/en/boards-of-appeal/decisions/t130520eu1

4 LLM Workflows

In this section, we describe the LLM workflows that we implement to perform automatic patent novelty prediction with passage retrieval at the claim and feature level. An overview of the workflows is shown in Figure 4. Each workflow is implemented using DSPy [26] and generates a structured output containing (i) the relevant prior art passages for each feature of the claim, (ii) a novelty prediction for each feature (*"Fully Disclosed"*, *"Partially Disclosed"*, *"Not Disclosed"*), and (iii) an overall novelty prediction for the claim (*"Novel"*, *"Not Novel"*).

Single Step Examination. In the *Single Step Examination* workflow, the model receives the claim and prior art document as input and is prompted to output the structured response in a single step. On average, the prompt contains 17k tokens.

Hierarchical Examination. In the *Hierarchical Examination* workflow, each feature is analyzed separately, allowing the model to perform deeper reasoning on each feature. It follows three main steps:

(1) Claim Segmentation: First, the claim is split into individual features either using a dedicated LLM call or by splitting at semicolons and newlines.

Model	Passage Retrieval (N=932)														Novel Feature Identification (N=912)		
	Claim-level							Feature-level									
	P	$\tilde{P}$	R	$\tilde{R}$	F1	$\tilde{F1}$	nDCG	P	$\tilde{P}$	R	$\tilde{R}$	F1	$\tilde{F1}$	nDCG	P	R	F1
Baselines																	
Random	6.1	28.9	16.5	45.7	7.9	34.4	10.8	2.3	20.7	3.3	25.9	2.2	22.4	2.8	4.2	5.5	3.2
Rouge-L	11.3	43.1	25.5	55.2	13.6	46.9	17.7	4.6	26.6	11.9	34.3	5.7	28.9	7.9	32.2	36.8	30.9
Qwen3-8B-Embeddings	13.7	40.2	35.7	58.4	17.4	45.8	23.5	8.3	28.9	22.7	42.3	10.6	32.9	15.4	27.0	12.6	13.5
Single-step Examination																	
Qwen3-30B-A3B-It	21.8	44.7	39.8	60.9	24.9	49.8	31.6	13.2	31.3	27.7	44.9	16.0	35.6	22.1	42.3	44.0	39.1
Qwen3-30B-A3B-Th	**22.7**	45.3	41.6	60.6	25.4	49.8	31.8	12.2	26.6	26.5	38.8	14.7	30.3	19.9	42.5	44.9	38.6
Qwen3-VL-235B-A22B-Th	20.4	44.0	52.4	70.3	26.1	52.3	35.1	13.8	32.4	36.0	52.1	17.7	38.4	25.1	43.4	47.2	41.3
Qwen3.5-397B-A17B	20.2	44.9	**57.6**	**74.8**	27.1	54.4	**36.6**	14.8	34.7	38.9	55.6	19.2	41.1	27.3	45.2	52.7	44.7
Hierarchical Examination																	
Qwen3-30B-A3B-Th	21.5	**46.3**	44.9	65.1	25.5	52.2	33.6	12.8	29.6	26.2	41.6	14.9	33.2	19.7	48.5	73.8	53.7
w/ LLM segmentation	21.4	**46.3**	44.4	64.7	25.2	52.1	32.7	12.7	28.8	27.1	41.2	15.0	32.4	19.9	50.9	56.1	48.2
Qwen3.5-397B-A17B	20.7	46.2	56.8	74.2	**27.4**	**55.1**	35.5	**17.0**	**38.4**	**39.1**	**57.3**	**20.9**	**44.1**	**28.6**	**53.5**	**81.8**	**59.7**

Table 2: Experimental Results for *Passage Retrieval* and *Novel Feature Identification*. *Passage Retrieval* metrics are only computed for rejected samples and *Novel Feature Identification* metrics are only computed for granted claims. $\tilde{P}$, $\tilde{R}$, and $\tilde{F1}$ denote the soft variants of precision, recall, and F1. Best results are in bold.

(2) Feature-level Retrieval & Examination: Next, for each feature, the LLM is called with the feature, the full claim for context, and the prior art document. The model analyzes the prior art document with respect to the given feature and outputs the identifiers of relevant prior art passages, a novelty prediction, and an examination summary for the given feature.
(3) Claim-level Novelty Prediction: Finally, the model is called with the claim under examination, the prior art document filtered according to the retrieval results of the previous step, and the feature-level examination summaries. The model outputs the overall novelty prediction for the claim.

This approach requires a total of 129k prompt tokens per claim on average (over 7x of *Single Step Examination*), with 95% of the tokens used for feature-level examination. However, each feature's examination can be performed in parallel. We order the elements in the prompt for feature-level examination in a way that maximizes prefix cache hit rate (i.e., prior art, claim, feature). Overall, despite requiring over 7x more tokens, it runs only 2.5x longer than *Single Step Examination* on a dedicated vLLM deployment with prefix caching.

5 Experiments

In this section, we describe our experimental setup, including several baselines (Section 5.1) and results (Section 5.2). We focus our analysis on open-weight models because patent examination requires high standards of data privacy and security that are difficult to guarantee when using closed-weight models via API. We deploy `Qwen/Qwen3-30B-A3B-Thinking-2507` [58], `Qwen/Qwen3-VL-235B-A22B-Thinking-FP8` [7], and `Qwen/Qwen3.5-397B-A17B-FP8`[15] via vLLM [31].

[15]https://huggingface.co/Qwen/Qwen3.5-397B-A17B-FP8

5.1 Baselines

We implement several baselines for a better understanding of the task and metrics: *Random* across all sub-tasks, *Logistic Regression* and *BERT Classifier* for claim-level novelty classification, and *Embedding Similarity* and *ROUGE-L* for passage retrieval and novel feature identification.

5.1.1 Random. To establish a lower bound, we implement a random baseline that assigns claim-level and feature-level novelty labels and retrieves passages uniformly at random.

5.1.2 ROUGE-L. To establish a simple baseline for passage retrieval and novel feature identification using only lexical overlap, we compute the ROUGE-L score [37] between each feature and each passage in the prior art document. We consider a feature to be disclosed if any passage has a similarity score above 0.4.

5.1.3 Embedding Similarity. In addition, we implement an embedding-based baseline for passage retrieval and novel feature identification using `Qwen/Qwen3-Embedding-8B` [60]. Similar to the ROUGE-L baseline, we compute the cosine similarity between each feature and each passage in the prior art document and consider a feature to be disclosed if any passage has a similarity score above 0.5.

5.1.4 Logistic Regression. To analyze the spurious correlations in the dataset, we implement a logistic regression model using various features based solely on the claim under examination without considering the prior art document. These features include the number of words, the number of features, the assigned IPC[16] classes (indicating the technical domain), and the top 500 TF-IDF features with n-grams up to length 4. We use a standard feature scaler to ensure

[16]https://www.wipo.int/en/web/classification-ipc

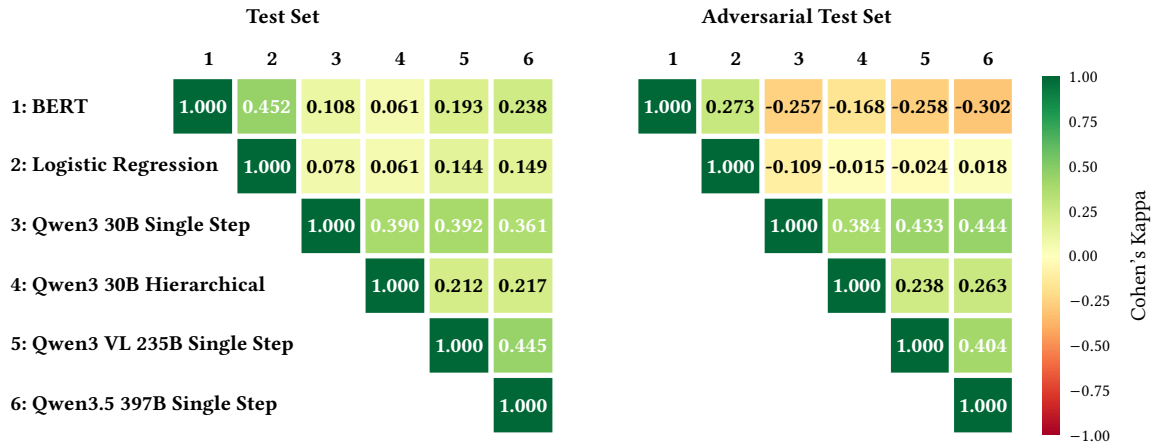

Figure 5: Cohen's kappa agreement between different models' predictions on the test and adversarial test sets.

that the model coefficients are interpretable. This baseline is only applied to the claim-level novelty classification task.

5.1.5 BERT Classifier. Like the logistic regression model, this baseline only considers the claim under examination without the prior art document, and should thus not be able to make any meaningful predictions. We fine-tune a BERT-base model [14] on the training set for claim-level novelty classification for 3 epochs.

5.2 Results

Table 2 shows the results for *Passage Retrieval* and *Novel Feature Identification*, and Table 3 shows the results for *Claim Novelty Prediction*.

5.2.1 Passage Retrieval. The LLM workflows clearly outperform lexical and semantic similarity baselines across all passage retrieval metrics. At both claim-level and feature-level, the best-performing model is Qwen3.5-397B-A17B using hierarchical examination, achieving a claim-level F1 score of 27.4 and a feature-level F1 score of 20.9 compared to 17.4 and 10.6 respectively for Qwen3-8B-Embeddings. This represents a substantial relative improvement of over 50% over the strongest non-LLM baseline. Hierarchical examination yields moderate but consistent retrieval improvements for Qwen3.5-397B-A17B, whereas Qwen3-30B-A3B-Thinking shows no clear benefit, suggesting that the advantages of task decomposition for retrieval may grow with a model's capabilities. Overall, performance differences across configurations are more pronounced at the feature level than at the claim level, confirming that fine-grained evaluation provides a more discriminative signal for comparing models. Finally, retrieval performance scales consistently with model size across all three models.

5.2.2 Novel Feature Identification. All LLM workflows substantially outperform the baselines on novel feature identification, even in the single-step setting. Notably, hierarchical examination yields large gains over single step examination, with Qwen3-30B-A3B-Thinking improving from 38.6 to 53.7 F1 and Qwen3.5-397B-A17B from 44.7 to 59.7 F1. Furthermore, hierarchical examination with the 30B model outperforms single-step examination with the larger 397B model (53.7 vs. 44.7 F1), revealing that structured decomposition is more critical than model scale for this sub-task.

5.2.3 Claim Novelty Prediction.

Fraction of Predicted Novel Claims. A notable observation across models is the significant variation in the fraction of claims predicted as novel. The Random baseline predicts approximately 49.5% of claims as novel, closely matching the balanced dataset distribution. In contrast, Qwen3-30B-A3B-Instruct (without thinking mode) exhibits extreme bias and predicts merely 3.2% as novel, indicating a strong tendency to classify claims as lacking novelty when not engaging in extended reasoning. The reasoning-enabled models demonstrate more balanced prediction distributions. Using single-step examination, Qwen3-30B-A3B-Thinking, Qwen3-VL-235B-A22B-Thinking, and Qwen3.5-397B-A17B predict 69.1%, 48.8%, and 50.8% as novel. Using hierarchical examination, the predicted novel fractions vary widely across models (30%–88%), suggesting that the effect of task decomposition on prediction bias is model-dependent. Overall, these variations show that even minor changes to the setup can substantially impact model behavior, underscoring the importance of comprehensive evaluation and calibration in real-world applications.

Spurious Correlations. On the test set, the fine-tuned BERT classifier achieves the highest accuracy of 75.2%, despite not having access to the prior art document. This indicates that, even after stratifying the dataset based on claim length and removing reference numerals, subtle but effective spurious correlations remain that models can exploit to achieve high claim novelty prediction performance without performing the intended task. The logistic regression model captures a substantial portion of BERT's performance gain over random guessing, suggesting that BERT relies heavily on simple, interpretable features. Although this does not definitively prove that BERT uses identical features to the logistic regression model, their predictions show moderate agreement (Cohen's Kappa of 0.45). We inspect the learned weights of the logistic

regression model, and find that among the most important features are (i) the presence of specific n-grams that are used to narrow a claim's scope (e.g., "*wherein*", "*characterized*", "*in that the*"), (ii) the indicator variables for certain IPC classes (technical domains) with imbalanced label distributions, (iii) the number of features in the claim, and (iv) the counts of punctuation characters.

LLMs' Robustness to Spurious Correlations. Interestingly, we find that the implemented LLM workflows do not appear to exploit these spurious correlations. As described in Section 3.4, we create an adversarial test set using adversarial filtering to identify samples that are misclassified by the BERT classifier. On this adversarial subset, the LLM workflows show only minor accuracy loss or even gains compared to the full test set. We note that due to the small size of the adversarial test set (278 samples), these results have very high variance. Therefore, we propose not to use this subset to evaluate the performance of models directly, but rather to verify that models do not exploit the identified spurious correlations. Additionally, we compute the agreement between different models' predictions using Cohen's Kappa [12], as shown in Figure 5. LLM workflows show only low to moderate agreement with the BERT classifier and the logistic regression model on the test set, and negative agreement on the adversarial test set, further indicating that they do not rely on the same spurious correlations.

Cross-Dataset Analysis. To validate that the issue of spurious correlations is not specific to our dataset, we conduct a similar analysis on the NOC4PC task of the PANORAMA dataset [36]. We train an XGBoost model on their claim novelty prediction task (NOC4PC) using TF-IDF and structural features from the claim text alone. The model achieves a test macro F1 of 41%, substantially outperforming most prompted LLMs in their evaluation. This confirms that spurious correlations are a prevalent issue in patent novelty datasets and that future work should carefully validate that models do not exploit them.

Prediction Accuracy. While Qwen3-30B-A3B-Instruct performs close to random chance, the thinking-enabled models achieve substantially better results. Using single-step examination, Qwen3-30B-A3B-Thinking, Qwen3-VL-235B-A22B-Thinking, and Qwen3.5-397B-A17B achieve 61.4%, 65.1%, and 69.1% accuracy respectively, representing meaningful improvements over the random baseline of 50.1%. Sampling multiple reasoning trajectories and averaging the results (self-consistency) yields further improvements in accuracy, with the 397B model achieving 70.7% accuracy. Hierarchical examination does not aid claim-level novelty prediction, i.e., the added reasoning effort does not translate into better claim-level predictions. Furthermore, adding feature examination summaries actually reduces accuracy to 57.9% and makes predictions even more unbalanced (88.5% predicted as novel), suggesting that the intermediate summaries may introduce noise or bias into the final aggregation step. Even using the label references from the ESOP document directly (i.e., using an oracle for the retrieval step; see Table 3 "with label references") does not improve claim novelty prediction accuracy, indicating that passage retrieval is not the main bottleneck for the binary prediction. Overall, while absolute accuracies leave substantial room for improvement, the robustness of LLMs against spurious correlations is particularly encouraging for practical deployment.

Model	Predicted Novel	Accuracy	Macro F1
Baselines			
Random	49.5%	50.1 / 49.6	50.1 / 49.6
Logistic Regression	47.7%	66.3 / 36.3	66.2 / 36.3
BERT (claim only)	39.7%	**75.2** / 0.0	**74.9** / 0.0
Single-step Examination			
Qwen3-30B-A3B-It	3.2%	52.0 / 51.1	38.2 / 36.9
Qwen3-30B-A3B-Th	69.1%	61.4 / 62.2	60.0 / 60.9
w/o prior art	75.2%	53.7 / 55.0	50.7 / 52.3
Qwen3-VL-235B-A22B-Th	48.8%	65.1 / 62.8	65.1 / 62.8
Qwen3.5-397B-A17B	50.8%	69.1 / 65.1	69.1 / 65.1
w/ self-consistency	53.2%	70.7 / **67.5**	70.7 / **67.5**
Hierarchical Examination			
Qwen3-30B-A3B-Th	88.5%	57.9 / 57.2	50.8 / 49.8
w/o summaries	75.3%	61.5 / 61.7	59.1 / 59.0
w/ label references	69.0%	61.6 / 58.3	60.3 / 56.9
Qwen3.5-397B-A17B	30.2%	63.3 / 65.5	61.8 / 64.0

Table 3: Experimental Results for *Claim Novelty Prediction*. For accuracy and macro F1, we report the performance on the full test set and on the adversarial subset (after the slash). Best results are in bold.

6 Discussion and Limitations

Our experimental results demonstrate both the promise and challenges of automated patent novelty examination at the feature level. While LLM-based workflows substantially outperform embedding-based baselines on passage retrieval and novel feature identification, absolute performance levels indicate room for improvement. Our multi-faceted evaluation reveals that different design choices yield varied improvements across evaluation facets: hierarchical examination enhances passage retrieval and novel feature identification but hurts claim-level novelty prediction, whereas model scale consistently improves performance across all facets.

Reliability of ESOP Documents. ESOP documents reflect the immense effort of highly trained patent examiners. In particular, EPO examination records are widely considered to be among the highest quality, given that each application is, unlike in other jurisdictions, examined by at least three examiners in order to reach a joint decision.[17] Nevertheless, they are not infallible and may contain errors or omissions, as exemplified by opposition procedures. Given the complexity and technical depth of patent examination, future work should further investigate how reliable ESOP documents are as a source of supervision, how a team of expert humans would perform on the proposed tasks, and how large their inter-annotator agreement would be.

[17] https://www.epo.org/en/legal/epc/2020/a18.html

Extension to Multi-Document Settings. The examination of novelty is formally always performed with respect to a single prior art document, thus our work has focused on this setting. However, it is likely that examiners perform such analysis across multiple documents, and record only the best match. Future work should explore extending the proposed approaches to multi-document settings, where workflows must first identify relevant documents, perform novelty examination with respect to each document, and aggregate results into a final examination report.

Generalizability Across Domains. While our work focuses on patents from specific technical domains (computing, speech processing, telecommunications, and pictorial communication) filed at the EPO, the proposed methodology is applicable more broadly. The feature-level examination approach is domain-agnostic and could be extended to other technical fields, though domain-specific language understanding may present additional challenges. Furthermore, while we focus exclusively on patent literature as prior art, novelty examination in practice also considers non-patent literature such as scientific publications, technical manuals, and public disclosures. Incorporating such diverse sources would require addressing challenges in document retrieval, format heterogeneity, and citation extraction, but would enable more comprehensive novelty assessment systems.

7 Conclusion

In this work, we have challenged the prevailing paradigm of treating patent novelty examination as a simple binary classification task. We introduced FiNE-Patents, a novel dataset comprising 3,658 first patent claims with fine-grained, feature-level prior art references extracted from European Search Opinion documents. This dataset enables evaluation at fine granularity to assess models' ability to identify specific passages disclosing individual claim features and to recognize which features make a claim novel. Our analysis revealed significant spurious correlations in existing approaches to novelty prediction, including claim length, reference numerals, and domain-specific linguistic patterns. We demonstrated that trained classifiers readily exploit these shortcuts, achieving high accuracy without genuinely reasoning about claim-prior art relationships. To address this, we created an adversarial test set using adversarial filtering, providing a diagnostic tool to verify model robustness. We proposed and evaluated LLM workflows that decompose claims into features, analyze each feature against prior art, and derive claim-level predictions. Our experiments demonstrated that these workflows substantially outperform embedding-based baselines on passage retrieval and novel feature identification tasks. Crucially, unlike fine-tuned classifiers, LLMs proved robust against spurious correlations, maintaining consistent performance on the adversarial test set. Our findings have important implications for both the research community and patent examination practice. For researchers, we provide a challenging benchmark that requires genuine semantic understanding and abstract reasoning, along with fine-grained supervision enabling more nuanced model development. For practitioners, our work points toward AI systems that can provide transparent, verifiable assistance to patent examiners, patent attorneys, and domain experts by explicitly linking predictions to supporting evidence in prior art documents. We release the FiNE-Patents dataset, the dataset creation pipeline, and all experimental code to foster further research into transparent and granular patent analysis. Future work should explore extending these approaches to multi-document retrieval settings, designing robust training pipelines, and establishing human expert performance.

References

[1] Lin Ai, Ziwei Gong, Harshsaiprasad Deshpande, Alexander Johnson, Emmy Phung, Ahmad Emami, and Julia Hirschberg. 2025. NovAScore: A New Automated Metric for Evaluating Document Level Novelty. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 3479–3494. https://aclanthology.org/2025.coling-main.234/

[2] Amna Ali, Ali Tufail, Liyanage Chandratilak De Silva, and Pg Emeroylariffion Abas. 2024. Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches. *Applied System Innovation* 7, 5 (2024). doi:10.3390/asi7050091

[3] Linda Andersson, Mihai Lupu, João Palotti, Allan Hanbury, and Andreas Rauber. 2016. When is the time ripe for natural language processing for patent passage retrieval?. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 1453–1462.

[4] Linda Andersson, Parvaz Mahdabi, Allan Hanbury, and Andreas Rauber. 2012. Report on the CLEF-IP 2012 Experiments: Exploring Passage Retrieval with the PIPExtractor. In *Conference and Labs of the Evaluation Forum*. https://api.semanticscholar.org/CorpusID:5914491

[5] Linda Andersson, Parvaz Mahdabi, Allan Hanbury, and Andreas Rauber. 2013. Exploring patent passage retrieval using nouns phrases. In *European Conference on Information Retrieval*. Springer, 676–679.

[6] Iliass Ayaou, Denis Cavallucci, and Hicham Chibane. 2025. DAPFAM: A Domain-Aware Patent Retrieval Dataset Aggregated at the Family Level. *arXiv preprint arXiv:2506.22141* (2025).

[7] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631* (2025).

[8] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 67–93. doi:10.18653/v1/2024.eacl-long.5

[9] Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior research methods* 53, 2 (2021), 757–780.

[10] Matthias Blume, Ghobad Heidari, and Christoph Hewel. 2024. Comparing Complex Concepts with Transformers: Matching Patent Claims Against Natural Language Text. In *Proceedings of the 5th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech@SIGIR)*.

[11] Zhenhai Chi, Wuquan Lin, Zhanhao Xiao, Huihui Li, Weiqi Chen, and Xiaoyong Liu. 2026. A review of patent analysis based on machine learning. *Applied Soft Computing* 186 (2026), 114063. doi:10.1016/j.asoc.2025.114063

[12] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[13] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating Data Contamination in Modern Benchmarks for Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 8706–8719. doi:10.18653/v1/2024.naacl-long.482

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[15] Yao Ding, Yuqing Wu, and Ziyang Ding. 2025. An automatic patent literature retrieval system based on llm-rag. *arXiv preprint arXiv:2508.14064* (2025).

[16] Pasi Fränti and Radu Mariescu-Istodor. 2023. Soft precision and recall. *Pattern Recognition Letters* 167 (2023), 115–121.

[17] Atsushi Fujii and Tetsuya Ishikawa. 2005. Document Structure Analysis for the NTCIR-5 Patent Retrieval Task.. In *NTCIR*.

[18] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2005. Overview of Patent Retrieval Task at NTCIR-5. In *NTCIR Conference on Evaluation of Information Access Technologies*. https://api.semanticscholar.org/CorpusID:267804819

[19] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2006. Test Collections for Patent Retrieval and Patent Classification in the Fifth NTCIR Workshop.. In *LREC*. 671–674.

[20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[21] Tirthankar Ghosal. 2022. Studies in aspects of peer review: novelty, scope, research lineage, review significance, and peer review outcome. *SIGIR Forum* 55, 2, Article 26 (March 2022), 2 pages. doi:10.1145/3527546.3527577

[22] Shohei Hido, Shoko Suzuki, Risa Nishiyama, Takashi Imamichi, Rikiya Takahashi, Tetsuya Nasukawa, Tsuyoshi Idé, Yusuke Kanehira, Rinju Yohda, Takeshi Ueno, Akira Tajima, and Toshiya Watanabe. 2012. Modeling Patent Quality: A System for Large-scale Patentability Analysis using Text Mining. *Inf. Media Technol.* 7, 3 (2012), 1180–1191. doi:10.11185/IMT.7.1180

[23] Hayato Ikoma and Teruko Mitamura. 2025. Can AI Examine Novelty of Patents?: Novelty Evaluation Based on the Correspondence between Patent Claim and Prior Art. arXiv:2502.06316 [cs.CL] https://arxiv.org/abs/2502.06316

[24] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[25] Lekang Jiang and Stephan M Goetz. 2025. Natural language processing in the patent domain: a survey. *Artificial Intelligence Review* 58, 7 (2025), 214.

[26] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=sY5N0zY5Od

[27] Valentin Knappich, Annemarie Friedrich, Anna Hätty, and Simon Razniewski. 2025. PEDANTIC: A Dataset for the Automatic Examination of Definiteness in Patent Claims. In *Proceedings of the 6th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech@SIGIR)*.

[28] Valentin Knappich, Anna Hätty, Simon Razniewski, and Annemarie Friedrich. 2025. PAP2PAT: Benchmarking Outline-Guided Long-Text Patent Generation with Patent-Paper Pairs. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 9524–9554. doi:10.18653/v1/2025.findings-acl.496

[29] Nancy Kong, Uwe Dulleck, Adam B Jaffe, Shupeng Sun, and Sowmya Vajjala. 2023. Linguistic metrics for patent disclosure: Evidence from university versus corporate patents. *Research policy* 52, 2 (2023), 104670.

[30] Ralf Krestel, Renukswamy Chikkamath, Christoph Hewel, and Julian Risch. 2021. A survey on deep learning for patent analysis. *World Patent Information* 65 (2021), 102035.

[31] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[32] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *International Conference on Machine Learning*. Pmlr, 1078–1088.

[33] Ryan Lee, Alexander Spangher, and Xuezhe Ma. 2024. PatentEdits: Framing Patent Novelty as Textual Entailment. arXiv:2411.13477 [cs.CL] https://arxiv.org/abs/2411.13477

[34] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union, 707–710.

[35] Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. An Open-Source Data Contamination Report for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 528–541. doi:10.18653/v1/2024.findings-emnlp.30

[36] Hyunseung Lim, Sooyohn Nam, Sungmin Na, Ji Yong Cho, June Yong Yang, Hyungyu Shin, Yoonjoo Lee, Juho Kim, Moontae Lee, and Hwajung Hong. 2025. PANORAMA: A Dataset and Benchmarks Capturing Decision Trails and Rationales in Patent Examination. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=JWewCpdjjq

[37] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[38] Ethan Lin, Zhiyuan Peng, and Yi Fang. 2025. Evaluating and Enhancing Large Language Models for Novelty Assessment in Scholarly Publications. In *Proceedings of the 1st Workshop on AI and Scientific Discovery: Directions and Opportunities*, Peter Jansen, Bhavana Dalvi Mishra, Harsh Trivedi, Bodhisattwa Prasad Majumder, Tom Hope, Tushar Khot, Doug Downey, and Eric Horvitz (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, USA, 46–57. doi:10.18653/v1/2025.aisd-main.5

[39] Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. In *First Conference on Language Modeling*. https://openreview.net/forum?id=u2vAyMeLMm

[40] Hao-Cheng Lo and Jung-Mei Chu. 2021. Pre-trained Transformer-based Classification for Automated Patentability Examination. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 1–5.

[41] Mihai Lupu, Atsushi Fujii, Douglas W Oard, Makoto Iwayama, and Noriko Kando. 2017. Patent-related tasks at ntcir. In *Current challenges in patent information retrieval*. Springer, 77–111.

[42] Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, et al. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186* (2025).

[43] Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity* 49 (2023), 101356.

[44] Arav Parikh and Shiri Dori-Hacohen. 2024. ClaimCompare: A Data Pipeline for Evaluation of Novelty Destroying Patent Pairs. In *Proceedings of the 5th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech@SIGIR)*.

[45] Florina Piroi and Allan Hanbury. 2017. Evaluating information retrieval systems on European patent data: The clef-ip campaign. In *Current Challenges in Patent Information Retrieval*. Springer, 113–142.

[46] Florina Piroi, Mihai Lupu, and Allan Hanbury. 2013. Overview of CLEF-IP 2013 Lab: Information Retrieval in the Patent Domain. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 232–249.

[47] Julian Risch, Nicolas Alder, Christoph Hewel, and Ralf Krestel. 2021. PatentMatch: A Dataset for Matching Patent Claims & Prior Art. In *Proceedings of the 2nd Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech@SIGIR)*.

[48] Janika Saretzki and Mathias Benedek. 2026. Investigating the Validity Evidence of Automated Scoring Methods for Divergent Thinking Assessments. *Creativity Research Journal* (2026), 1–17.

[49] Jinzhi Shan, Qi Zhang, Chongyang Shi, Mengting Gui, Shoujin Wang, and Usman Naseem. 2024. Structural Representation Learning and Disentanglement for Evidential Chinese Patent Approval Prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2014–2023.

[50] Yongluo Shen and Zexi Lin. 2024. PatentGrapher: A PLM-GNNs Hybrid Model for Comprehensive Patent Plagiarism Detection Across Full Claim Texts. *IEEE Access* 12 (2024), 182717–182725. https://api.semanticscholar.org/CorpusID:274416597

[51] Homaira Huda Shomee, Zhu Wang, Sathya N. Ravi, and Sourav Medya. 2025. A Survey on Patent Analysis: From NLP to Multimodal AI. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 8545–8561. doi:10.18653/v1/2025.acl-long.419

[52] Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 OLMo 2 Furious (COLM's Version). In *Second Conference on Language Modeling*. https://openreview.net/forum?id=2ezugTT9kU

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

[54] Suyuan Wang, Xueqian Yin, Menghao Wang, Ruofeng Guo, and Kai Nan. 2024. Evopat: A multi-llm-based patents summarization and analysis agent. *arXiv preprint arXiv:2412.18100* (2024).

[55] Wenqing Wu, Chengzhi Zhang, Tong Bao, and Yi Zhao. 2025. SC4ANM: Identifying optimal section combinations for automated novelty prediction in academic papers. *Expert Syst. Appl.* 273 (2025), 126778. https://api.semanticscholar.org/CorpusID:276296422

[56] Long Xia, Jun Xu, Yanyan Lan, Jiafeng Guo, and Xueqi Cheng. 2016. Modeling Document Novelty with Neural Tensor Network for Search Result Diversification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and*

Development in Information Retrieval (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 395–404. doi:10.1145/2911451.2911498

[57] Qiushi Xiong, Zhipeng Xu, Zhenghao Liu, Mengjia Wang, Zulong Chen, Yue Sun, Yu Gu, Xiaohua Li, and Ge Yu. 2025. Enhancing the Patent Matching Capability of Large Language Models via the Memory Graph. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 337–347. doi:10.1145/3726302.3729970

[58] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).

[59] Yongmin Yoo and Kris W Pan. 2026. Pat-DEVAL: Chain-of-Legal-Thought Evaluation for Patent Description. *arXiv preprint arXiv:2601.00166* (2026).

[60] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

[61] Yi Zhao and Chengzhi Zhang. 2025. A review on the novelty measurements of academic papers. *Scientometrics* 130, 2 (2025), 727–753.