PatentMind: A Multi-Aspect Reasoning Graph for Patent Similarity Evaluation

Yongmin Yoo, Qiongkai Xu, Longbing Cao
Frontier AI Research Centre, Macquarie University
School of Computing, FSE, Macquarie University
yooyongmin91@gmail.com {qiongkai.xu, longbing.cao}@mq.edu.au

Abstract

Patent similarity evaluation is essential for intellectual property analysis, yet existing methods struggle to capture the multifaceted structure of patent documents encompassing technical specifications, legal boundaries, and application contexts. We propose PatentMind, a framework that performs patent similarity assessment through a Multi-Aspect Reasoning Graph (MARG). PatentMind decomposes patent documents into three dimensions: technical features, application domains, and claim scopes, and computes dimension-specific similarity scores, which are then integrated via a context-aware dynamic weighting mechanism that emulates expert-level judgment. To facilitate evaluation, we introduce PatentSimBench, an expert-annotated benchmark comprising 500 patent pairs. Experiments show that PatentMind achieves a Pearson correlation of $r = 0.938$ with expert annotations, substantially outperforming embedding-based, patent-specific, and prompt engineering baselines. Our framework offers interpretable, multi-dimensional assessment applicable to downstream tasks such as prior art search and infringement risk analysis.

1 Introduction

Patent documents pose significant challenges for NLP-based similarity evaluation due to specialized domain knowledge, intricate legal language, and complex structural formats (Aslanyan and Wetherbee, 2022; Indukuri et al., 2007; Casola and Lavelli, 2021; Lupu and Hanbury, 2013). Accurate evaluation of patent similarity is crucial for identifying prior art, assessing infringement risks, and determining the novelty of innovations (Feng, 2020; Helmers et al., 2019; Arts et al., 2018; Wang et al., 2023). Given that inaccuracies can lead to overlooked prior art, invalid patent grants, or undetected infringements, robust and interpretable evaluation methods are essential. Moreover, the rapidly growing volume of patent applications across diverse technological fields further intensifies the need for reliable similarity evaluation techniques (Jiang and Goetz, 2024). With the increasing reliance on patent analytics for R&D strategy, legal risk assessment, and competitive intelligence, developing accurate and interpretable models for patent analysis has become critically important in both industry and academia (Hain et al., 2022; Yoo et al., 2023, 2025).

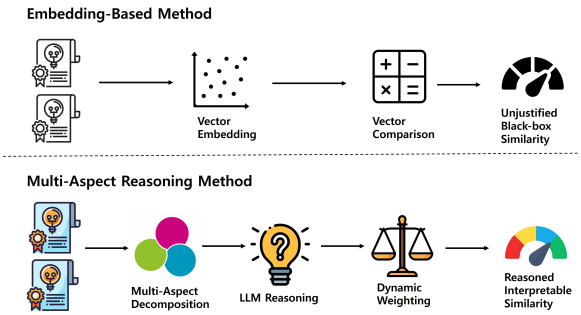

Figure 1: The comparison between multi-aspect reasoning and embedding-based similarity methods.

In the early stages of patent similarity evaluation, keyword-based methods such as bag-of-words were widely used. However, these approaches struggled to capture the semantics and technical specificity of patent documents, resulting in ineffective representations (D'hondt et al., 2013; Ascione and Sterzi, 2024). To overcome these limitations, embedding-based methods such as static word embeddings and transformer-based contextual embeddings were subsequently introduced (Lee and Hsiang, 2020; Reimers and Gurevych, 2019). Nevertheless, these methods compress patent texts into dense vectors, which can obscure critical distinctions and reduce interpretability.

Recent advances in large language models (LLMs) have demonstrated considerable promise in reasoning-based tasks, offering a compelling alternative to traditional embedding-based approaches. Unlike fixed embedding methods, LLMs support dynamic reasoning and semantic comprehension,

Findings of the Association for Computational Linguistics: ACL 2026, pages 14947–14964

facilitating more sophisticated similarity assessments. However, directly applying general-purpose LLMs to patent similarity tasks remains challenging due to specialized terminology and complex features inherent in patent texts (Ikoma and Mitamura, 2025; Ascione and Sterzi, 2024). Moreover, prompt-based approaches to LLMs have partially addressed the limitations of embeddings, but still struggle due to the domain-specific language and structural complexity of patent documents.

To address these challenges, we propose PatentMind, a novel patent similarity evaluation framework structured as a Multi-Aspect Reasoning Graph (MARG). This framework emulates the analytical processes of patent experts through multi-step reasoning based on LLMs. PatentMind decomposes patent documents into three core dimensions: technical features, application domains, and claim scopes. Each dimension is independently evaluated for similarity, and the results are integrated through a context-aware reasoning process consisting of four stages: Domain Relationship Analysis, Information Distribution, Dimension Relevance, and Cross-Validation. This multi-dimensional approach is designed to closely reflect real-world patent examination practices (USPTO, 2024).

Consequently, PatentMind effectively captures subtle distinctions across the diverse attributes of patents. Our empirical evaluations demonstrate that PatentMind, implemented with GPT-4o-mini as the underlying LLM, achieves a Pearson correlation of 0.938 with expert evaluations, significantly surpassing embedding and LLM-based prompting baselines. To facilitate rigorous evaluation, we introduce PatentSimBench as the first expert-annotated patent similarity benchmark that reflects real-world criteria employed by patent examiners and attorneys. By explicitly modeling technical, legal, and contextual aspects through interpretable reasoning, PatentMind collectively addresses key limitations of prior approaches.

The key contributions of PatentMind are summarized as follows:

- **Multi-Aspect Reasoning Graph (MARG):** We propose a structured reasoning framework over LLMs that performs decomposed similarity evaluations across three constitutive dimensions of patent documents and synthesizes them through multi-stage inference.
- **Context-Aware Dynamic Weighting:** We introduce a dynamic weighting mechanism grounded in multi-step contextual reasoning, jointly leveraging domain relationship analysis, score distribution characteristics, dimension relevance estimation, and cross-validation for robustness.
- **PatentSimBench Benchmark:** We construct and publicly release the first expert-annotated benchmark for patent similarity evaluation, comprising 500 rigorously curated patent pairs with rationale-supported annotations.

2 Related Work

2.1 Patent Similarity Evaluation

Early approaches to patent similarity evaluation, such as keyword matching and TF-IDF, treated patents as generic texts and often failed to capture their domain-specific structure and terminology (Tseng et al., 2007; Ascione and Sterzi, 2024). To address these limitations, vector-based models like Word2Vec and Doc2Vec were introduced to encode semantic relationships within patent corpora (Jeon et al., 2022). However, these models struggled with challenges such as polysemy and limited domain adaptation. To improve semantic representation, contextualized language models like SciBERT (Beltagy et al., 2019) and PatentBERT (Lee and Hsiang, 2020) were developed, offering embeddings tailored for scientific and patent texts. While these models produce more accurate representations, their black-box nature and lack of interpretability hinder their practical application in legal and technical decision-making contexts (Miric et al., 2023; Ascione and Sterzi, 2024).

To overcome these challenges, we propose PatentMind, a structured and interpretable framework that mimics expert reasoning via modular, dimension-based analysis. By decomposing patents into multiple perspectives and reasoning over each dimension, PatentMind provides transparent, multi-dimensional similarity scores aligned with human judgment criteria, addressing the limitations of opaque, single-vector embeddings.

2.2 Prompt Engineering

Prompt engineering has emerged as a key strategy for enabling complex reasoning in LLMs. Chain-of-Thought prompting (Wei et al., 2022) improves sequential reasoning by guiding models through intermediate steps. However, its linear structure limits effectiveness in tasks requiring divergent reasoning. To overcome this, Yao et al. (Yao et al.,

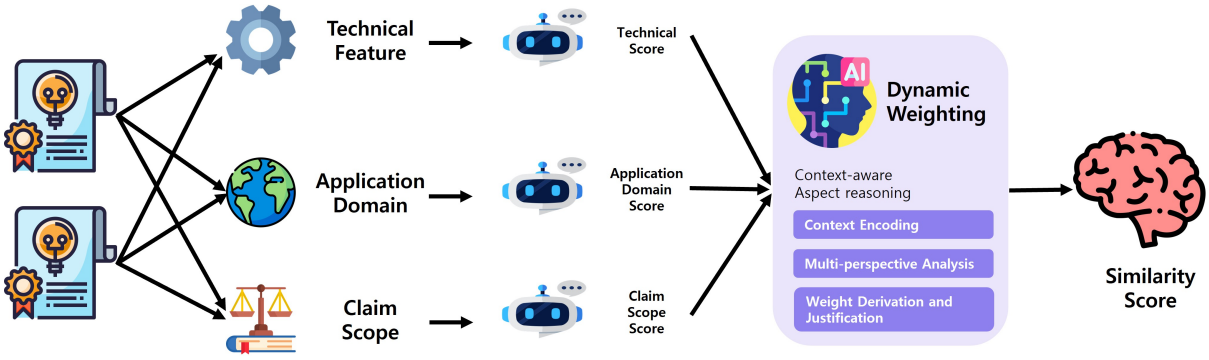

Figure 2: The workflow of PatentMind, structured as a Multi-Aspect Reasoning Graph (MARG).

2023) proposed the Tree-of-Thoughts framework, which enables exploration of multiple reasoning paths in parallel, enhancing performance on tasks requiring structured deliberation. Additional work demonstrates that carefully designed prompts enable strong few-shot (Brown et al., 2020) and zero-shot (Kojima et al., 2022) reasoning, using minimal task examples or simple cues. These prompting strategies significantly improve LLM performance on complex tasks, laying the foundation for PatentMind's structured reasoning approach, where multi-dimensional analysis is essential for capturing both technical and legal dimensions of patents.

3 Methodology

This section presents PatentMind, a novel framework for patent similarity evaluation structured as a Multi-Aspect Reasoning Graph (MARG). PatentMind systematically leverages the reasoning capabilities of LLMs to analyze patent documents across multiple critical dimensions. The framework decomposes patent texts into distinct dimensions, independently evaluates each dimension, and integrates the resulting similarity scores through a context-aware dynamic weighting mechanism. Figure 2 illustrates the MARG workflow, consisting of three core modules: Multi-Aspect Decomposition, Similarity Computation, and Context-Aware Dynamic Weighting.

3.1 Multi-Dimensional Feature Extraction

PatentMind captures the complex structure and domain-specific nuances of patent documents through LLM-based feature extraction. We decompose each patent into three dimensions: technical feature (T), application domain (D), and claim scope (C). These dimensions reflect how patent specialists evaluate patent similarity, focusing on what the invention is, where it applies, and how broadly it is claimed (USPTO, 2024).

To obtain these dimensions, we apply an LLM-based feature extraction function, T, D, or C, on a patent document P to obtain corresponding content. The function generates a structured representation composed of technical features $T(P)$, application domains $D(P)$, and claim scope $C(P)$. The title, abstract, and claims of the patent are used as input for the extractor. Note that these sections are legally mandated components for any valid patent application (e.g., 35 U.S.C. § 112), ensuring the availability of structured inputs. This structured prompting approach guides the LLM using aspect-specific instructions. Each prompt focuses on a distinct semantic aspect: methodologies for technical features, application contexts for domains, and legal boundaries for claim scope. These prompts are grounded in structured inputs from the title, abstract, and claims. Prompt designs for all dimensions are provided in Appendix B.

3.2 Dimension-wise Similarity Computation

Following feature extraction, we compute similarity between patent documents on these aspects. Given two patent documents P_1 and P_2 and their respective extracted features $T(P_i)$, $D(P_i)$, and $C(P_i)$, we calculate similarity scores across each dimension:

$$S = \{S_T, S_D, S_C\}, \tag{1}$$

where S_T, S_D, and S_C denote the similarity scores for the technical, domain, and claim dimensions, respectively, computed as $S_T = f_T(T(P_1), T(P_2))$, $S_D = f_D(D(P_1), D(P_2))$, and $S_C = f_C(C(P_1), C(P_2))$. These scores serve as the foundation for the final weighted similarity measure described in Section 3.3.

We implement similarity functions through targeted prompts that direct the LLM to assess simi-

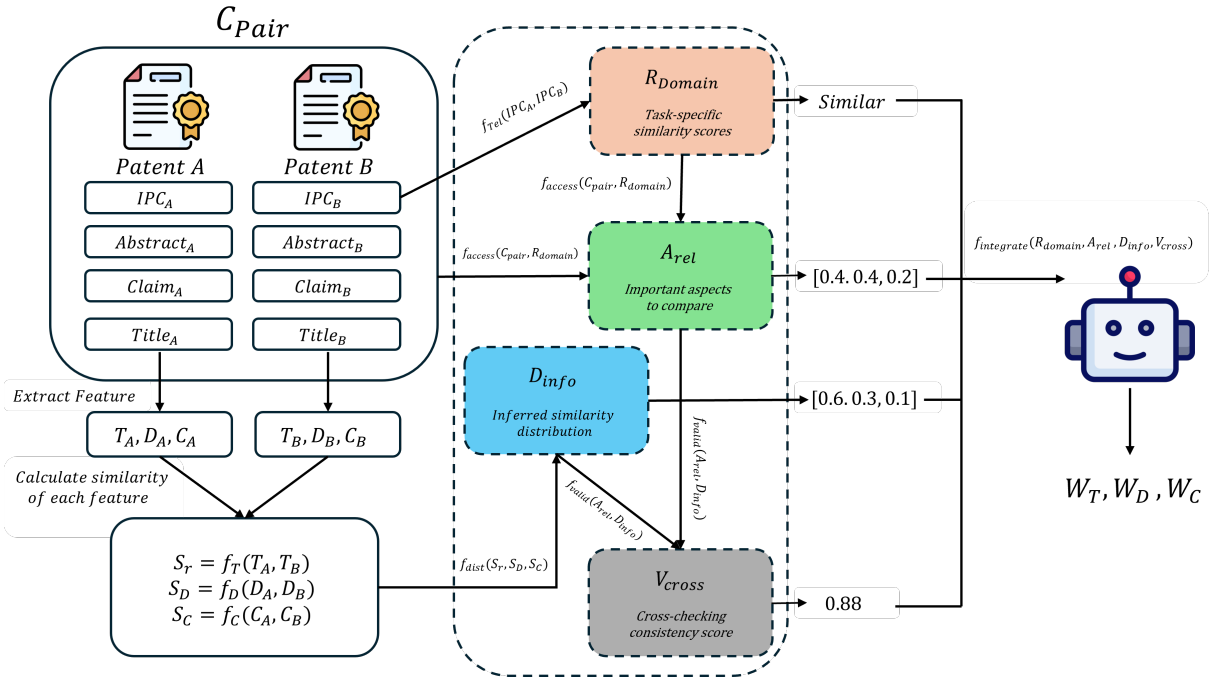

Figure 3: The computational graph of the Context-Aware Dynamic Weight Reasoning framework, with modules annotated by their notations: Domain Relationship Analysis (f_{rel}, R_{domain}), Information Distribution Analysis (f_{dist}, D_{info}), Dimension Relevance Assessment (f_{assess}, A_{rel}), and Cross-validation Reasoning (f_{valid}, V_{cross}).

larity using semantic relationships, contextual implications, and domain-specific equivalences. Each similarity score ranges from 0 to 1, with higher values indicating greater similarity. Appendix C details the prompt templates used.

3.3 Context-Aware Dynamic Weight Reasoning

PatentMind adapts to contextual nuances by computing weights through multi-step LLM reasoning. Instead of parameter-intensive fine-tuning, it functions as an open calculation process using reasoning prompts, highlighting both efficiency and adaptability. This process comprises four key stages.

3.3.1 Context Encoding

We construct a contextual representation of the patent pair by encoding key information from their titles, abstracts, claims, and IPC codes:

$$C_{pair} = \{t, a, c, IPC\}_{A,B}, \tag{2}$$

where t, a, c, and IPC represent the title, abstract, claims, and IPC codes of patents A and B.

3.3.2 Multi-perspective Analysis

The LLM performs multi-dimensional reasoning to guide the dynamic weighting process through four interdependent functions (detailed prompts in Appendix D). First, ***Domain Relationship Analysis*** (f_{rel}) leverages IPC codes to output a categorical label R_{domain}, determining the influence of domain similarity. Second, ***Information Distribution Analysis*** (f_{dist}) examines the patterns of aspect-wise scores (S_T, S_D, S_C) to generate D_{info}. Third, ***Dimension Relevance Assessment*** (f_{assess}) evaluates dimension importance (A_{rel}) based on the full context C_{pair} and R_{domain}. Finally, ***Cross-validation Reasoning*** (f_{valid}) checks the consistency between predicted relevance (A_{rel}) and actual score distribution (D_{info}), yielding a robustness score $V_{cross} \in [0, 1]$. We formalize this pipeline as a directed acyclic graph $G = (V, E)$, where nodes correspond to reasoning steps and edges represent dependencies facilitating the message-passing of contextual signals for the final weight derivation.

3.3.3 Weight Derivation and Justification

In this step, the integration function $f_{integrate}$ computes the final dimension weights w_T, w_D, and w_C by aggregating the outputs of the reasoning modules: the domain relationship label (R_{domain}), the information distribution signal (D_{info}), the dimension relevance scores (A_{rel}), and the robustness indicator (V_{cross}). It also generates a textual justification J that explains the rationale behind the derived weights, thereby enhancing interpretability. Detailed prompting instructions for each reasoning step, including $f_{integrate}$, are provided in

Attribute	Value
Patent Pairs	500 pairs
Technical Fields	IPC Sections
Similarity Range	1 to 5 (Likert scale)
Fleiss' Kappa	0.588
Cronbach's Alpha	0.967

Table 1: The statistics of PatentSimBench.

Appendix D. Appendix H provides concrete reasoning trace examples illustrating how these modules operate on patent pairs with divergent dimension scores.

3.4 Final Similarity Calculation

The overall similarity score combines the scores by

$$S_{\text{overall}} = w_T \times S_T + w_D \times S_D + w_C \times S_C. \quad (3)$$

Here, the weights are normalized such that $w_T + w_D + w_C = 1$. This weighted integration enables PatentMind to adaptively emphasize the most informative dimensions of each patent pair, reflecting the nuanced evaluation process of human patent experts. Appendix E details the prompts for this final step.

4 Dataset Construction

To facilitate rigorous evaluation of patent similarity methods, we introduce PatentSimBench, the first expert-annotated benchmark dataset tailored for patent similarity tasks. PatentSimBench provides rationale-supported similarity scores aligned with real-world legal and technical judgments, making it particularly suitable for evaluating multi-step reasoning models. The dataset consists of 500 patent pairs sampled from the USPTO database (USPTO, 2025), with a balanced representation across all IPC sections. The International Patent Classification (IPC) categorizes patents into eight top-level domains, such as human necessities, chemistry, physics, and digital technologies (WIPO, 1975), ensuring comprehensive technical coverage across diverse fields. Similarity annotations are based on a 5-point Likert scale, with scores ranging from 1 to 5. Our dataset is publicly available on GitHub.[1]

4.1 Dataset Composition

PatentSimBench is designed as an evaluation benchmark rather than a training corpus, as PatentMind requires no supervised training. Unlike standard NLI or STS annotation, patent similarity assessment demands simultaneous legal and technical expertise, resulting in substantially higher per-pair annotation costs. Given these constraints, we prioritized annotation rigor and broad IPC coverage over raw scale. PatentSimBench includes patent pairs evenly distributed across major technological domains, covering all IPC sections. Table 1 summarizes key dataset statistics. The moderate inter-annotator agreement scores, including Fleiss' Kappa (0.588) and Cronbach's Alpha (0.967), demonstrate robust consensus among expert annotators despite the inherent subjectivity in assessing patent similarity.

4.2 Annotation Process

Four domain experts with both legal and technical backgrounds annotated PatentSimBench using a 5-point Likert scale. The annotation guidelines were rigorously designed based on the USPTO Manual of Patent Examining Procedure (MPEP) §2141, which mandates a structured analysis of claim scope, technical content, and field of search. This ensures that our similarity criteria reflect actual legal standards rather than subjective judgments. The annotation process followed a two-phase protocol: (1) independent assessments by each expert, and (2) consensus refinement through structured discussions of divergent cases, during which annotators reviewed and adjusted their ratings based on shared rationales. For quality control, we excluded patent pairs with a post-discussion standard deviation greater than 2 ($\sigma_{\text{ratings}} > 2$). On a 5-point scale, such cases signal extreme disagreement indicative of unreliable annotation even after consensus discussion; a total of 32 pairs were excluded through this criterion. This process was guided by the predefined annotation guidelines described in Appendix F.

5 Experimental Results

We conducted comprehensive experiments on the PatentSimBench dataset to evaluate PatentMind against embedding-based models, prompt engineering strategies, and regression baselines. To assess the robustness and contribution of individual components, we performed ablation studies and cross-model evaluations using various LLMs. All experiments were carried out with standard evaluation metrics, using GPT-4o-mini as the backbone model. To ensure determinism and reproducibil-

[1] https://github.com/Yongmin-Yoo/PatentMind

ity, we used a fixed temperature of 0.2 and default sampling parameters (top_p=1.0). Furthermore, to verify result stability, we conducted 5 independent runs with different random seeds on the full PatentSimBench dataset (all 500 pairs). The observed standard deviation in Pearson correlation was negligible ($\sigma < 0.001$), confirming the robustness of our results without the need for reporting averaged metrics over multiple trials.

5.1 Comparison with Embedding-Based and Patent-Specific Methods

Model	Pearson (r) ↑	Spearman (ρ) ↑	MSE ↓	MAE ↓
Word2Vec	.761	.819	.145	.319
BERT	.835	.819	.228	.381
BERT-for-Patent	.762	.754	.115	.283
SciBERT	.766	.798	.194	.381
Patent-GPT-J	.892	.883	.124	.178
PatentMind	**.938**	**.923**	**.113**	**.092**

Table 2: The performance comparison of PatentMind with varying embedding-based and patent specific baseline methods.

Table 2 compares PatentMind with conventional embedding-based and patent-specific models. Baseline approaches exhibit only moderate correlation with expert judgments and show high error rates. In contrast, PatentMind achieves significantly better performance with a Pearson correlation of 0.938 and the lowest error rates, demonstrating superior alignment with human evaluations.

5.2 Comparison with Prompting-Based Methods

Prompting Strategy	Pearson (r) ↑	Spearman (ρ) ↑
I/O Prompting	.702	.672
Few-shot Prompting	.891	.897
Chain-of-Thought (CoT)	.866	.870
Self-Consistency w/ CoT	.887	.881
Tree-of-Thought (ToT)	.745	.654
Chain-of-Draft (CoD)	.834	.835
PatentMind	**.938**	**.923**

Table 3: The performance comparison of PatentMind with various prompting-based baseline methods.

We compare PatentMind with representative prompting strategies in Table 3. While advanced methods such as Few-shot, Chain-of-Thought (CoT), and Self-Consistency yield reasonable performance, they struggle to capture the multi-dimensional nature of patent documents. For instance, CoT typically follows a single reasoning path, limiting its ability to jointly model technical, legal, and domain-specific aspects. Self-Consistency introduces variation in reasoning but fails to incorporate context-sensitive judgment aligned with expert reasoning. In contrast, PatentMind integrates hierarchical reasoning with dynamic weighting, enabling more faithful modeling of patent semantics. As a result, it achieves the highest agreement with expert judgments (Pearson $r = .938$; Spearman $\rho = .923$). Notably, since all methods in Table 3 use the same backbone LLM (GPT-4o-mini), the consistent performance gap indicates that PatentMind's advantage stems from MARG's structured, dimension-wise decomposition rather than prompt phrasing or model capacity. These findings highlight the critical role of explicitly modeling multi-aspect patent contexts in enhancing both accuracy and interpretability.

5.3 Comparison of Various LLMs

LLM	Pearson (r) ↑	Spearman (ρ) ↑
Claude-3.5-Sonnet	.931	.928
Llama-3.3-70B-Versatile-128k	.922	.911
Deepseek-r1-distill-llama-70b	.914	.907
Qwen-QwQ-32B-Preview	.910	.893
PatentMind	**.938**	**.923**

Table 4: Performance of PatentMind with different backbone LLMs. GPT-4o-mini is the default backbone used throughout all other experiments.

To test PatentMind's robustness across LLM architectures, we evaluated its performance with multiple models. All tested LLMs showed high correlation with expert annotations, indicating that PatentMind's effectiveness is not model-dependent. GPT-4o-mini and Claude-3.5-Sonnet performed competitively, while Qwen-QwQ-32B showed slightly lower scores, likely due to architectural and reasoning differences. Overall, the consistent results highlight the model-agnostic strength of PatentMind's design.

5.4 Comparison of Regression Weighting

To validate the effectiveness of PatentMind's dynamic weighting strategy, we compared it against several regression-based baselines using the PatentSimBench dataset. Model performance was evaluated based on Pearson and Spearman correlations between predicted similarity scores and expert annotations. As shown in Table 5, all regression models, including Linear Regression, Lasso, SVM, Tree Boosting, Bayesian Ridge, and

Fold Type	Pearson	Spearman	RMSE	MAE
Linear Regression	.920	.897	.122	.095
Lasso Regression	.918	.901	.127	.104
SVM	.919	.895	.125	.102
Tree Boosting	.909	.886	.129	.097
MLP Regressor	.921	.896	.121	.095
Bayesian Ridge Regression	.920	.897	.122	.096
PatentMind	**.938**	**.923**	**.113**	**.092**

Table 5: The performance comparison of PatentMind using dynamic weighting and regression based baselines.

Method	Latency	Cost / Pair
Embedding (SciBERT)	< 0.1s	$\approx$ \$0.000
PatentMind (Ours)	20 30s	$\approx$ \$0.003
Human Expert	> 60 min	$>$ \$100.0

Table 6: Efficiency and cost comparison. Human cost is estimated based on average patent attorney hourly rates assuming 60 minutes per case.

a 2-layer MLP Regressor, underperformed relative to PatentMind. The best-performing baseline, MLP Regression (Pearson $r = .921$; Spearman $\rho = .896$), still fell short of PatentMind's performance (Pearson $r = .938$; Spearman $\rho = .923$). These results reveal that regression-based fixed-weighting methods struggle to capture the nuanced, context-sensitive variations essential for accurate patent similarity evaluation. Unlike these baselines, which require supervised training to learn fixed parameters, PatentMind functions as an open reasoning process driven by LLM prompts, avoiding extra training costs while dynamically adapting weights to each patent pair.

5.5 Efficiency and Cost-Benefit Analysis

To assess the practical viability of PatentMind, we evaluated its computational efficiency using GPT-4o-mini. On average, processing a single patent pair requires approximately 13,225 input tokens and 832 output tokens across 8 API calls. Through parallelized execution of independent extraction and scoring steps, the end-to-end latency is reduced to 20–30 seconds per pair. Financially, the operational cost is approximately \$0.003 USD per pair based on current API pricing. While computationally more expensive than simple embedding-based retrieval, this cost is negligible compared to human patent experts. As shown in Table 6, PatentMind offers expert-level reasoning at a fraction of the cost ($<$\$0.01), making it a cost-effective solution for high-stakes decision-making tasks such as infringement risk assessment.

We emphasize that this computational overhead is an intentional design choice. PatentMind is explicitly designed to function as a re-ranking and precision analysis module, typically applied after a broad initial retrieval phase. In a practical two-stage pipeline, a lightweight retriever (e.g., BM25 or embedding-based search) first identifies a set of candidate patents, and PatentMind subsequently re-ranks these candidates where accuracy and interpretability are critical. In high-stakes domains like patent infringement analysis or invalidity searches, where missing a critical prior art can result in significant legal consequences, accuracy and interpretability take precedence over millisecond latency. Therefore, the trade-off of higher computational cost for expert-level reasoning depth is justified for these targeted downstream applications.

6 Ablation Studies

6.1 Comparison of Aspect Impact

Model	Pearson	Spearman	Avg. Drop (%)
PatentMind Full	**.938**	**.923**	–
Equal Weighting	.904	.890	**3.35**
Two Dimensions Only			
w/o Claim Scope Dimension	.912	.906	2.15
w/o Technical Dimension	.901	.898	3.10
w/o Application Dimension	.903	.894	3.20
Average (Two Dimension)	**.905**	**.899**	**2.82**
Single Dimension Only			
Claim Scope Dimension Only	.903	.905	2.65
Technical Dimension Only	.908	.912	2.05
Application Dimension Only	.865	.848	7.40
Average (Single Dimension)	**.892**	**.888**	**4.03**

Table 7: The performance comparison of PatentMind using varying aspects.

To assess the contribution of each component in our framework, we conducted an ablation study across seven configurations, including the removal of dynamic weighting, exclusion of individual dimensions, and single-dimension-only variants. As shown in Table 7, replacing the context-aware dynamic weighting mechanism with equal weighting results in a noticeable performance drop (3.35%), demonstrating the importance of adapting weights to contextual relevance. Removing any one of the three core dimensions(Claim Scope, Technical, or Application)also led to performance degradation (avg. 2.82%), indicating their complementary roles in similarity evaluation. Notably, the removal of Technical (3.10%) or Application (3.20%) dimensions caused greater performance loss than removing Claim Scope (2.15%). Single-dimension vari-

ants showed even larger drops in performance (avg. 4.03%), with the Application-only model suffering the most (7.40%). Although the Technical-only model performed relatively better (2.05% drop), it still fell short of the full model's performance. These findings validate the effectiveness of our design choices: the integration of technical, legal, and domain-level reasoning, combined with context-aware dynamic weighting, leads to substantially improved accuracy and adaptability in patent similarity evaluation.

7 Error Analysis

In this section, we examine the performance boundaries of PatentMind. We analyze prediction accuracy and residual patterns, identify domain-specific error trends across IPC sections, and investigate high-error cases involving functional ambiguity and terminology mismatch.

7.1 Prediction and Residual Analysis

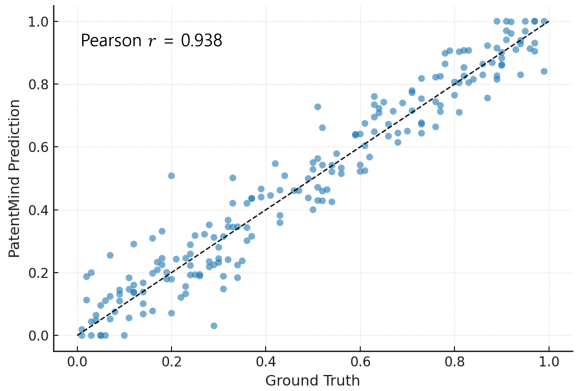

Figure 4: The correlation between expert judgments and scores predicted by PatentMind.

We assess PatentMind's reliability by comparing predicted similarity scores to expert annotations and analyzing residuals. Predictions show strong alignment, with most errors within the 0–0.1 range (MAE = 0.0918). Figure 4 shows predictions generally follow the ideal ($y = x$) line, though high-similarity pairs (ground truth > 0.8) tend to be slightly underestimated, and low-similarity pairs (< 0.2) slightly overestimated. The average residual (-0.0197) indicates minimal bias, with domain-specific variations discussed further below.

7.2 Error Analysis across Different IPC Codes

Our IPC-level analysis shows that PatentMind's prediction errors are more pronounced in certain technical domains, such as metallurgy (C22B), genetic engineering (C12N), and mechanical engineering (F16D). These errors stem from: (1) underrepresentation of domain-specific terms in LLM pretraining (e.g., "leaching" vs. "extraction" in C22B), (2) structural variations in patent drafting across domains, and (3) context-dependent similarity criteria. For instance, metallurgy patents (e.g., C22B300) often suffer from underestimated similarity due to synonym recognition issues (e.g., "leaching" vs. "extraction"), while audio processing patents (e.g., G06F316) tend to be overestimated due to language. Biotechnology patents (C12N) present additional challenges due to linguistic complexity and causal reasoning structures.

These findings underscore the need for domain-adaptive embeddings or knowledge graph integration to enhance synonym recognition and contextual understanding. PatentMind's modular design facilitates such enhancements without requiring extensive retraining, a key direction for future work.

7.3 High-Error Case Analysis

We further analyze the 50 patent pairs with the highest prediction errors to identify potential systematic limitations. We observe four recurring patterns: (1) underestimation of highly similar patents (e.g., predicted 0.53 vs. actual 0.833 in metallurgy); (2) overestimation of dissimilar patents (e.g., predicted 0.50 vs. actual 0.250 in marking tech); (3) difficulty distinguishing functional differences despite textual overlap (e.g., 0.48 vs. 0.750 in mechanical parts); (4) challenges in interpreting domain terms, particularly in biotech (e.g., 0.24 vs. 0.500). These findings suggest that, while PatentMind effectively captures general semantic similarity, performance may be further improved by incorporating more explicit representations of functional intent and domain-specific vocabulary. Additional examples and discussion are provided in Appendix I. Appendix H further illustrates the dynamic weighting reasoning process for representative cases.

8 Conclusion

We proposed PatentMind, a novel framework integrating linguistic, legal, and technical reasoning for patent similarity. By jointly modeling technical features, application domains, and claim scope, it effectively captures structural complexity. A key innovation is our context-aware weighting mechanism, which uses LLM reasoning to dynamically

adjust dimensional importance.

PatentMind achieves expert-level agreement ($r = 0.938$) and model-agnostic robustness, significantly outperforming embedding, prompting, and patent-tuned baselines. Notably, it delivers higher accuracy with greater efficiency than resource-intensive fine-tuning approaches.

Finally, we release PatentSimBench, the first expert-annotated similarity benchmark. These contributions extend beyond computational linguistics, offering a semantically grounded foundation for critical real-world tasks such as prior art search, infringement assessment, and novelty evaluation.

Limitations

Computational Efficiency. PatentMind's multi-stage reasoning structure requires 8 API calls and over 13,000 tokens per patent pair, resulting in 20–30 seconds latency per pair. While this is acceptable for re-ranking small candidate sets, it limits its applicability to large-scale retrieval scenarios. Future work will explore response caching, selective execution of reasoning steps, and lightweight model substitution through knowledge distillation or LoRA-based adaptation to reduce computational overhead while preserving reasoning quality.

Domain-Specific Opportunities. Our error analysis (Section 7.2) reveals that PatentMind's prediction errors are more pronounced in domains with complex terminology (e.g., metallurgy, biotechnology) or unique structural conventions. These errors primarily stem from vocabulary gaps in LLM pretraining, such as the failure to recognize "leaching" and "extraction" as synonyms in metallurgy patents. MARG's modular architecture enables targeted enhancements at individual reasoning nodes without reengineering the overall pipeline. For instance, domain-specific synonym lexicons or retrieval-augmented domain definitions can be incorporated into the feature extraction and similarity computation stages. Expanding dataset coverage to enable finer-grained domain-level evaluation is also a priority for future work.

Jurisdictional and Linguistic Scope. The current evaluation focuses exclusively on USPTO patents in English, as the largest English-language patent source commonly used in prior patent-NLP research. Patent similarity criteria are inherently tied to jurisdiction-specific legal standards; for example, the EPO employs a Problem-Solution approach, while the JPO and WIPO follow distinct drafting conventions. Although MARG's three-dimension decomposition reflects general factors in patent examination that are not inherently USPTO-specific, cross-jurisdictional and multilingual validation remains an important direction for future work.

Ethical Considerations

This research adheres to ethical principles in data utilization. The PatentSimBench dataset comprises exclusively publicly available patent documents, complying with intellectual property regulations and containing no personally identifiable information. Acknowledging potential biases in language models, we implemented expert annotations and structured reasoning methodologies to enhance the interpretability and fairness of the results. To facilitate transparency and reproducibility, we have made our research methodology and dataset publicly accessible. We have considered the potential misuse of AI-based patent similarity evaluation tools in legal disputes or competitive intelligence analysis. Consequently, we recommend using this system as a supplementary tool rather than as the sole basis for high-stakes decisions. Our future research will continue investigating bias mitigation strategies and enhancing ethical applications of patent evaluation systems. For transparency, we note that ChatGPT was used to improve grammar and clarity of expression.

Acknowledgments

This work is partially sponsored by the Australian Research Council Discovery projects DP260104429 and DP240102050, ARC LIEF project LE240100131, and ARC Linkage project LP230201022.

References

Sam Arts, Bruno Cassiman, and Juan Carlos Gomez. 2018. Text matching to measure patent similarity. *Strategic Management Journal*, 39(1):62–84.

Grazia Sveva Ascione and Valerio Sterzi. 2024. A comparative analysis of embedding models for patent similarity. *arXiv preprint arXiv:2403.16630*.

Grigor Aslanyan and Ian Wetherbee. 2022. Patents phrase to phrase semantic matching dataset. *arXiv preprint arXiv:2208.01171*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Silvia Casola and Alberto Lavelli. 2021. Summarization, simplification, and generation: The case of patents. *arXiv preprint arXiv:2104.14860*.

Eva D'hondt, Suzan Verberne, Cornelis Koster, and Lou Boves. 2013. Text representations for patent classification. *Computational Linguistics*, 39(3):755–775.

Sijie Feng. 2020. The proximity of ideas: An analysis of patent text using machine learning. *PLOS One*, 15(7):e0234880.

Daniel S. Hain, Roman Jurowetzki, Tobias Buchmann, and Patrick Wolf. 2022. A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177:121559.

Lea Helmers, Franziska Horn, Franziska Biegler, Tim Oppermann, and Klaus-Robert Müller. 2019. Automating the search for a patent's prior art with a full text similarity search. *PLOS One*, 14(3):e0212103.

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 preprint arXiv:2502.06316*.

K. V. Indukuri, A. A. Ambekar, and A. Sureka. 2007. Similarity analysis of patent claims using natural language processing techniques. In *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'07)*, pages 169–175. IEEE.

Daeseong Jeon, Joon Mo Ahn, Juram Kim, and Changyong Lee. 2022. A doc2vec and local outlier factor approach to measuring the novelty of patents. *Technological Forecasting and Social Change*, 174:121294.

Lekang Jiang and Stephan Goetz. 2024. Natural language processing in patents: A survey. *arXiv preprint arXiv:2403.04105*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Joon Lee and James Hsiang. 2020. Patentbert: Patent classification with fine-tuning a pre-trained bert model. *World Patent Information*, 62:101965.

Mihai Lupu and Allan Hanbury. 2013. Patent retrieval. *Foundations and Trends in Information Retrieval*, 7(1):1–97.

Marko Miric, Nan Jia, and Kevin G Huang. 2023. Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. *Strategic Management Journal*, 44(2):491–519.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yuen-Hsien Tseng, Chi-Jen Lin, and Yu-I Lin. 2007. Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247.

USPTO. 2024. *Manual of Patent Examining Procedure*, 9th edition, revision 01.2024 edition. United States Patent and Trademark Office. Accessed: 2025-03-22.

USPTO. 2025. Bulk data storage system (bdss) and open data portal (odp). `https://bulkdata.uspto.gov/`. BDSS retirement: April 11, 2025; ODP launched: February 12, 2025.

Meiyun Wang, Hiroki Sakaji, Hiroaki Higashitani, Mitsuhiro Iwadare, and Kiyoshi Izumi. 2023. Discovering new applications: Cross-domain exploration of patent documents using causal extraction and similarity analysis. *World Patent Information*, 74:102238.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

WIPO. 1975. International patent classification (ipc). `https://www.wipo.int/classifications/ipc/`. IPC established by the Strasbourg Agreement, entered into force October 7, 1975.

Shunyu Yao et al. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Yongmin Yoo, Cheonkam Jeong, Sanguk Gim, Junwon Lee, Zachary Schimke, and Deaho Seo. 2023. A novel patent similarity measurement methodology: Semantic distance and technological distance. *arXiv preprint arXiv:2303.16767*.

Yongmin Yoo, Qiongkai Xu, and Longbing Cao. 2025. PatentScore: Multi-dimensional evaluation of LLM-generated patent claims. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30727–30746, Suzhou, China. Association for Computational Linguistics.

A Appendix A: Patent Terminology and Concepts

This appendix provides detailed explanations of patent terminology and concepts referenced in the main paper, intended to help readers less familiar with intellectual property documentation.

A.1 Patent Document Structure

A patent document follows a standardized format consisting of several key components:

- **Title:** A brief description of the invention that typically indicates its function, mechanism, or purpose.
- **Abstract:** A concise summary (typically 150-250 words) outlining the invention's core technical contribution.
- **Background:** Explanation of the technical problem being addressed, limitations of existing solutions, and the context of the invention.
- **Detailed Description:** Comprehensive explanation of the invention's implementation, often including drawings, diagrams, specific embodiments, and working examples sufficient to enable reproduction by a person skilled in the relevant field.
- **Claims:** Precisely defined statements that establish the legal boundaries of protection. Claims are the most critical section for legal analysis and similarity assessment.

Claims are structured hierarchically, with independent claims defining the broadest protection and dependent claims adding specific limitations. For example, an independent claim might describe a general method for data processing, while dependent claims specify particular implementations, parameters, or use cases.

A.2 Patent Classification Systems

The International Patent Classification (IPC) system organizes patents using hierarchical alphanumeric codes that categorize technological domains. The IPC structure includes:

- **Section** (one letter, A-H): Broadest division of technology (e.g., A = Human Necessities, G = Physics)
- **Class** (two digits): Major technological divisions within a section (e.g., G06 = Computing)
- **Subclass** (one letter): Further division (e.g., G06F = Electric Digital Data Processing)
- **Group** (variable digits): Specific technological areas (e.g., G06F3/048 = Interaction techniques for graphical user interfaces)

Common IPC codes referenced in this paper include:

- **C12N:** Biochemistry - Microorganisms or enzymes
- **G06F:** Computing - Electric digital data processing
- **F16D:** Mechanical engineering - Couplings and brakes
- **C22B:** Metallurgy - Production or refining of metals

Each patent may be assigned multiple IPC codes to reflect its cross-disciplinary nature. These classifications are crucial for organizing patent literature, identifying relevant prior art, and contextualizing similarity assessments.

A.3 Legal Concepts in Patent Analysis

Several legal terms are fundamental to patent analysis:

- **Prior art:** Any evidence that an invention is already known before the filing date of a patent application. This includes existing patents, published applications, academic papers, public demonstrations, or commercial products. Prior art determines novelty and is a primary consideration in examining patent validity.
- **Novelty:** For an invention to be patentable, it must be new (novel) compared to prior art. A patent lacks novelty if all its essential elements are disclosed in a single prior art reference.
- **Non-obviousness:** Beyond novelty, an invention must involve an inventive step that would not be obvious to a person with ordinary skill in the relevant technical field.

- **Infringement:** Unauthorized making, using, selling, or importing of a patented invention. Infringement analysis examines whether a product or process incorporates all elements of at least one independent claim of a patent.

- **Claim construction:** The process of interpreting the meaning and scope of patent claims, which is essential for both infringement analysis and similarity assessment.

A.4 Multi-dimensional Nature of Patent Similarity

Patent similarity assessment is inherently multi-faceted, encompassing three critical dimensions that patent experts consider during evaluation:

- **Technical attributes:** The core invention mechanisms, algorithms, components, or methodologies. This dimension focuses on how the invention works and what technical problems it solves. Technical similarity might exist even when patents are applied in different domains.

- **Application contexts:** The fields, industries, problems, or use cases where the invention applies. Two patents may implement different technical approaches but address the same application problem, resulting in contextual similarity.

- **Legal boundaries:** The scope and limitations of protection defined by the claims. Legal similarity assessment considers the overlap in protection scope, which may differ from pure technical similarity. Broader claims typically encompass more potential similar patents than narrowly defined claims.

Patent examiners, attorneys, and analysts dynamically adjust the importance of these dimensions based on the specific context of the analysis task. For example, prior art searches emphasize technical similarity to assess novelty, while infringement analysis focuses on claim coverage and legal boundaries.

This multi-dimensional nature makes patent similarity assessment particularly challenging and distinguishes it from general document similarity tasks, justifying our development of the MARG framework that explicitly models these dimensions.

B Appendix B: Prompts for Feature Extraction

In this appendix, we provide the prompts used to guide the Large Language Model (LLM) in extracting the three essential dimensions from patent documents: Technical Features ($T(P)$), Application Domains ($D(P)$), and Claim Scope ($C(P)$).

B.1 Prompt for Technical Features ($T(P)$)

Prompt : *Summarize the technical features of the patent, focusing on methodologies, algorithms, and innovation points.*

B.2 Prompt for Application Domains ($D(P)$)

Prompt : *Identify the application domains of the patent, including industries, problem areas, and potential applications.*

B.3 Prompt for Claim Scope ($C(P)$)

Prompt : *Determine the claim scope of the patent, summarizing the legal protection boundaries and key rights asserted in the claims.*

C Appendix C: Prompts for Similarity Computation

In this appendix, we provide the prompts used to guide the Large Language Model (LLM) in computing the similarity scores for each of the three dimensions: Technical Features (S_T), Application Domains (S_D), and Claim Scope (S_C). These prompts enable the LLM to assess the degree of overlap between two patent documents based on their extracted features.

C.1 Prompt for Calculate Technical Features Similarity (S_T)

Prompt : *Given the technical feature summaries of Patent A and Patent B, assess the similarity of their technical contributions, focusing on methodologies, algorithms, and innovation points. Provide a similarity score between 0 and 1, where 0 indicates no overlap and 1 indicates identical technical features. Include a brief justification for your assessment. Output the result in the following format: Score: [numerical score], Reason: [justification].*

C.2 Prompt for Calculate Application Domains Similarity (S_D)

Prompt : *Given the application domain summaries of Patent A and Patent B, evaluate the similarity of*

their practical contexts, including industries, problem areas, and potential applications. Provide a similarity score between 0 and 1, where 0 indicates completely distinct domains and 1 indicates fully shared domains. Include a brief justification for your assessment.Output the result in the following format: Score: [numerical score], Reason: [justification].

C.3 Prompt for Calculate Claim Scope Similarity (S_C)

Prompt : *Given the claim scope summaries of Patent A and Patent B, analyze the similarity of their legal protection boundaries and key rights asserted in the claims. Provide a similarity score between 0 and 1, where 0 indicates no overlap in claim scope and 1 indicates identical claim scope. Include a brief justification for your assessment.Output the result in the following format: Score: [numerical score], Reason: [justification].*

D Appendix D: Prompt for Dynamic Weight

Below are the prompts used to guide the Large Language Model (LLM) through the five sequential reasoning stages of the context-aware dynamic weighting process in the MAGR framework. Each prompt specifies the input data and requires a structured output format.

D.1 Prompt for Domain Relationship Analysis (R_{domain})

Prompt: *Given the IPC codes of Patent A and Patent B from C_{pair} (titles, abstracts, claims, and IPC codes), assess the technical domain relationship between the two patents. Categorize the relationship as identical (same IPC subclass), hierarchical (one patent's domain subsumes the other), overlapping (shared IPC codes), or distinct (no common IPC codes). Output the result in the following format: Category: [relationship], Explanation: [justification].*

D.2 Prompt for Information Distribution Analysis (D_{info})

Prompt: *Given the similarity scores S_T, S_D, and S_C for Patent A and Patent B, as computed in Section 3.2, analyze the distribution pattern of these scores. Identify the pattern as uniform similarity (all scores are similar), dimension dominance (one score is significantly higher), or complementary dimensions (high similarity in one dimension offsets lower similarity in others). Output the result in the following format: Pattern: [pattern], Justification: [explanation].*

D.3 Prompt for Dimension Relevance Assessment (A_{rel})

Prompt: *Given the context of Patent A and Patent B from C_{pair} (titles, abstracts, claims, and IPC codes) and their domain relationship R_{domain}, assess the relative importance of technical features, application domains, and claim scope for evaluating their similarity. Assign relevance scores between 0 and 1 to each dimension, ensuring the sum equals 1. Output the result in the following format: Scores: [technical features: score, application domains: score, claim scope: score], Explanation: [justification].*

D.4 Prompt for Cross-validation Reasoning (V_{cross})

Prompt: *Given the dimension relevance scores (A_{rel}) and the actual similarity score distribution (D_{info}), assess how well these two align. If they strongly agree (i.e., the most important predicted dimension matches the dimension with the highest similarity), assign a robustness score close to 1. If they partially agree or conflict, assign a lower robustness score accordingly. Output the result in the following format: Metric: [score], Justification: [explanation].*

D.5 Prompt for Weight Derivation and Justification (w_T, w_D, w_C, J)

Prompt: *Given the domain relationship R_{domain}, similarity distribution pattern D_{info}, relevance scores A_{rel}, and robustness metric V_{cross} for Patent A and Patent B, integrate these inputs to determine the final weights for technical features w_T, application domains w_D, and claim scope w_C, ensuring $w_T + w_D + w_C = 1$. Provide a textual justification. Output the result in the following format: Weights: [w_T: score, w_D: score, w_C: score], Justification: [explanation].*

E Appendix E: Prompt for Similarity Score Calculation

Prompt : *Calculate the final similarity score S_{final} using the formula $S_{final} = w_T \times S_T + w_D \times S_D + w_C \times S_C$, where S_T, S_D, and S_C are the similarity scores for technical, application domains, and claim scope, respectively, and w_T, w_D, and*

w_C are their corresponding weights. Ensure that the result is a numerical value between 0 and 1, and return the value rounded to three decimal places. Output the result in the following format. Patent_Similarity_MAR :[score]

F Appendix F: PatentSimBench Annotation Guidelines

In this appendix, we provide the annotation guidelines used for constructing the PatentSimBench dataset. These guidelines were distributed to all expert annotators to ensure consistent evaluation of patent similarity.

F.1 Project Introduction

PatentSimBench is the first expert-annotated benchmark dataset for evaluating similarity between patent documents. The purpose of this project is to establish a reliable "gold standard" for patent similarity across various technical domains.

F.2 Rating Scale

Annotators evaluated patent pair similarity using a 5-point Likert scale:

Score	Level	Description
1	Very Low	Fundamentally different inventions
2	Low	Similar elements with substantial differences
3	Medium	Partial overlap in key dimensions
4	High	Substantial similarity in core technology
5	Very High	Nearly identical inventions

Table 8: Similarity rating scale for patent pairs.

F.3 Similarity Assessment Guidelines

Annotators were instructed to consider the following elements when evaluating similarity between patents:

F.3.1 Core Invention Concept

Definition: The essence of the problem being solved and its solution.

Assessment Guidelines:

- Fundamental problem-solving approach in both patents
- Core concepts and principles of the inventions
- Key innovation points and technical contributions
- Fundamental similarities and differences between inventions

F.3.2 Implementation Details

Definition: Specific methods and components that realize the core concept.

Assessment Guidelines:

- Physical/logical components and their arrangement
- Implementation details and mechanisms
- Performance parameters and operating conditions
- Comparison across various embodiments of the invention

F.3.3 Purpose and Effects

Definition: Intended effects and purposes of the invention.

Assessment Guidelines:

- Benefits and effects provided by the invention
- Intended usage environment and situations
- Degree of problem resolution achieved by the invention
- Impact on users or industries

F.3.4 Legal Protection Dimensions

Definition: The scope of legal protection sought by the patent document.

Assessment Guidelines:

- Scope and limitations of claims
- Characteristics of key rights assertions
- Clarity and specificity of legal protection
- Legal differentiation from existing patents

F.4 Annotation Procedure

Annotators followed a structured process:

1. Preliminary review of both patent documents
2. Structural analysis of each patent according to assessment guidelines
3. Comparative analysis to identify similarities and differences

4. Assessment of similarity across all evaluation elements

5. Determination of the overall similarity score (1-5)

6. Documentation of detailed rationale (minimum 100 words)

F.5 Assessment Principles

Annotators were guided by the following principles:

- **Objectivity**: Focus on document content rather than personal preferences
- **Consistency**: Assign consistent scores to patent pairs with similar characteristics
- **Detailed Review**: Analyze core technical/legal dimensions rather than superficial similarity
- **Patent Law Perspective**: Apply evaluation approaches based on USPTO MPEP guidelines

F.6 Examples and References

F.6.1 Example 1: High Similarity (Score 5)

Patent Pair: Patent A and Patent B (both describing a specific method for neural network acceleration)

Key Similarities:

- Both patents address the same problem of neural network computation acceleration
- Identical core technical approach using matrix decomposition
- Similar architectural components and data flow
- Equivalent performance claims and applications in mobile devices
- Nearly identical claim scope with minor variations in dependent claims

Annotator Rationale: "These patents are nearly identical in their core innovation, implementation approach, and intended applications. Both describe the same matrix decomposition technique for neural network acceleration, with the same architectural components and data flow patterns. While Patent B includes two additional dependent claims specifying memory management details, this represents a minor extension rather than a fundamental difference. The technical overlap is comprehensive, and the claims protect essentially the same invention."

F.6.2 Example 2: Medium Similarity (Score 3)

Patent Pair: Patent C and Patent D (both related to image processing systems)

Key Similarities/Differences:

- Both patents address image processing, but for different applications (medical imaging vs. autonomous vehicles)
- Similar preprocessing techniques but different core algorithms
- Partial overlap in component architecture but significant differences in implementation
- Different performance metrics and optimization goals
- Some overlap in claim scope but with substantially different limitations

Annotator Rationale: "These patents show moderate similarity in their approach to image processing, sharing common preprocessing techniques and some architectural elements. However, they diverge significantly in their core algorithms, with Patent C using a convolutional approach while Patent D employs a transformer-based method. Their application domains are distinct (medical diagnosis vs. autonomous navigation), leading to different optimization goals and performance metrics. The claim scope shows some overlap in general image processing methods but contains substantially different limitations reflecting their distinct applications."

F.6.3 Example 3: Low Similarity (Score 1)

Patent Pair: PAtent E and Patent F (blockchain system vs. wireless communication protocol)

Key Differences:

- Completely different technical domains and problem spaces
- No overlap in core technologies or methodologies
- Different implementation architectures and components

- Distinct application domains and user groups
- No similarity in claim scope or legal protection

Annotator Rationale: "These patents address fundamentally different technical domains with no meaningful overlap. Patent E describes a blockchain consensus mechanism for financial transactions, while Patent F details a wireless communication protocol for IoT devices. They employ different technologies, serve different purposes, target different users, and have no overlap in their implementation approaches. The claims seek protection for entirely unrelated inventions with no common elements that would create potential for infringement or prior art concerns."

F.6.4 Reference Materials Provided to Annotators

- USPTO Manual of Patent Examining Procedure (MPEP) Section 2141
- European Patent Office Guidelines G-VII, 5.1
- Prior art search methodology guides
- IPC classification reference materials

F.7 Quality Control Process

To ensure annotation quality:

- Cases with score differences >2 points were assigned to additional reviewers
- Weekly calibration meetings were held to discuss discrepancies
- Patent pairs with final standard deviation >2.0 were excluded from the dataset
- Senior patent experts provided final validation of all annotations

G Appendix G: Prompts for Baseline Methods

To ensure fair comparison and reproducibility, we provide the exact prompts used for the baseline methods, including Chain-of-Thought (CoT) and Few-shot prompting.

G.1 Prompt for Chain-of-Thought (CoT)

The CoT prompt encourages the model to generate intermediate reasoning steps before predicting the final score.

Prompt: *You are an expert patent analyst. Your task is to evaluate the similarity between two patent documents, Patent A and Patent B.*
Input:
- Patent A: [Title, Abstract, Claims]
- Patent B: [Title, Abstract, Claims]
Instructions:
1. Analyze the technical field, core innovation, and claim scope of both patents. 2. Compare the similarities and differences step-by-step. 3. Based on the reasoning, determine a similarity score between 0.0 (completely different) and 1.0 (identical).
Output the result in the following format:
Reasoning: [Your step-by-step analysis]
Score: [numerical score]

G.2 Prompt for Few-shot Prompting

For Few-shot prompting, we utilized a 3-shot setting. We selected three representative patent pairs with expert-annotated scores (Low, Medium, High) to guide the model.

Prompt: *Evaluate the similarity between two patents on a scale of 0.0 to 1.0. Here are three examples:*
Example 1:
Patent A: [Content of Patent A1]
Patent B: [Content of Patent B1]
Score: 0.2 (Low Similarity)

Example 2:
Patent A: [Content of Patent A2]
Patent B: [Content of Patent B2]
Score: 0.5 (Medium Similarity)

Example 3:
Patent A: [Content of Patent A3]
Patent B: [Content of Patent B3]
Score: 0.9 (High Similarity)

Task:
Patent A: [Target Patent A Content]
Patent B: [Target Patent B Content]
Score:

H Reasoning Trace Examples

This appendix illustrates how PatentMind's context-aware dynamic weighting operates on patent pairs

with divergent dimension scores.

Example 1: LED Electrode Patents (H01L25075). PatentMind produced the following dimension scores: $S_T = 0.72$ (shared nano-scale LED electrode technology), $S_D = 0.45$ (lamp vs. display applications), $S_C = 0.30$ (product claims vs. method claims). The reasoning modules operated as follows:

- f_{dist}: Detected *dimension dominance* pattern, as S_T significantly exceeds S_D and S_C.
- f_{assess}: Assigned $w_T = 0.45$, $w_D = 0.30$, $w_C = 0.25$.
- $f_{\text{integrate}}$ justification J: "Technical features receive the highest weight because the shared technological foundation is the primary basis for similarity. Application domain receives moderate weight because the lamp-versus-display distinction represents meaningful divergence. Claim scope receives the lowest weight as the product-versus-method distinction reflects drafting strategy rather than inventive divergence."

Final score: $0.45 \times 0.72 + 0.30 \times 0.45 + 0.25 \times 0.30 = 0.534$.

Example 2: Audio Processing Patents (G06F316 / H04L2906). Dimension scores: $S_T = 0.55$, $S_D = 0.62$, $S_C = 0.38$. Despite surface-level lexical overlap in audio-related terminology, the patents target distinct tasks (playback auditioning vs. content generation).

- f_{rel}: Classified domain relationship as *overlapping* based on partial IPC overlap.
- f_{assess}: Assigned $w_T = 0.35$, $w_D = 0.40$, $w_C = 0.25$.
- $f_{\text{integrate}}$ justification J: "Application domain receives the highest weight because the functional intent divergence is the most discriminative factor despite shared audio terminology."

Final score: $0.35 \times 0.55 + 0.40 \times 0.62 + 0.25 \times 0.38 = 0.536$.

I Detailed High-Error Case Analysis

We provide the error analysis of representative cases in Table 9.

Error Type	Patent Pair	Ground Truth	Model Prediction	Error
Underestimation	*METHOD FOR PLATINUM RECOVERY* (C22B300) *METHOD FOR PLATINUM RECOVERY* (C22B300)	0.833	0.53	0.303
Overestimation	*METHOD FOR APPLYING INK MARKINGS* (G06K1502) *METHOD FOR DETERMINING QUALITY OF MARKINGS* (G06T700)	0.250	0.50	0.250
Functional Differences	*DISC BRAKE* (F16D6500) *DISC BRAKE* (F16D65097)	0.750	0.48	0.270
Terminology	*DNA POLYMERASES WITH INCREASED 3'-MISMATCH DISCRIMINATION* (C12N912) *DNA POLYMERASES WITH INCREASED 3'-MISMATCH DISCRIMINATION* (C12N912)	1.000	0.70	0.300
Overestimation	*Audio Content Auditioning by Playback Device* (G06F316) *Systems and Methods for Automatically Generating Audio Content* (H04L2906)	0.500	0.24	0.260

Table 9: The representative cases by model mispredictions.