



Large language models in the patent lifecycle: A systematic review

Pingfei Jiang^{a,*}, Andrea Pietro Arena^b, Salvatore Sorce^b, Mark Atherton^c

^a Department of Innovation, Technology and Entrepreneurship, University of Exeter, Exeter, United Kingdom

^b Dipartimento di Ingegneria e Architettura, Università degli Studi di Enna "Kore", Enna, Italy

^c Department of Mechanical and Aerospace Engineering, Brunel University of London, London, United Kingdom

ARTICLE INFO

Keywords:

Large language models

Patent analysis

Generative AI

Systematic literature review

Retrieval-augmented generation

ABSTRACT

Patents constitute the world's largest repository of codified technical knowledge, yet their length, legal terminology, and linguistic ambiguity have limited the effectiveness of traditional text mining and keyword-based retrieval. Large Language Models (LLMs) offer new capabilities for processing this complexity, but the resulting literature is fragmented across disciplines, with no comprehensive synthesis of how these tools are being applied across the patent lifecycle. This study systematically reviews the literature on LLMs and Generative AI in patent analytics, following the PRISMA 2020 protocol, to map current applications, characterise dominant architectural trends, and identify priority research gaps. Using Scopus and Web of Science, we identified and screened a corpus of 129 peer-reviewed studies published from 2018 to early 2026, organised according to three patent lifecycle stages: creation and pre-filing ($N = 36$), examination and prosecution ($N = 34$), and post-grant management ($N = 60$). The review shows that research activity has grown sharply since 2024, accompanied by a shift from exploratory, single-task studies toward applied, multi-component systems. Across this literature, Retrieval-Augmented Generation (RAG) has become the dominant architectural response to hallucination risk, grounding generated content in verified external sources. The review further identifies persistent gaps in domain-specific evaluation metrics, multimodal reasoning, and governance frameworks for AI-generated patent content. For patent professionals and institutions, the findings indicate that LLMs are best positioned as assistive tools requiring human oversight rather than autonomous replacements, and that the central challenge going forward is not whether but how the patent system governs this transition.

Glossary

Agentic AI	AI systems that autonomously plan and execute multi-step tasks (retrieving, drafting, checking) in a coordinated loop, rather than responding to a single prompt.
Context degradation	The tendency of large language model performance to decline as input length grows, with models under-using or "losing track of" information placed in the middle of long documents (sometimes called "lost in the middle").
Hallucination	The tendency of generative AI to produce plausible but factually incorrect output, such as fabricated citations, invented technical details, or non-existent prior art.
Long-tail problem	The difficulty models have in reliably handling rare or infrequently occurring categories in a dataset (e.g.,

(continued on next column)

(continued)

Longitudinal validation	uncommon patent classes or figure types), because too little training data exists to learn them well. Evaluation of a system's performance over an extended period of real-world use, rather than against a single static test set.
Model collapse	Progressive degradation in a model's output quality and diversity that occurs when models are trained repeatedly on AI-generated (synthetic) rather than human-authored data.
N-gram metrics	Automated text-evaluation scores (e.g., ROUGE, BLEU) that measure overlap in word sequences between generated and reference text, they capture lexical similarity but not factual accuracy or legal validity.
Retrieval-Augmented Generation (RAG)	An architecture in which a model retrieves relevant documents from an external knowledge source before generating its response, grounding output in verified

(continued on next page)

This article is part of a special issue entitled: LLM/GenAI/Agentic AI for IP analytics and strategy published in World Patent Information.

* Corresponding author.

E-mail address: P.Jiang@exeter.ac.uk (P. Jiang).

<https://doi.org/10.1016/j.wpi.2026.102506>

Received 5 April 2026; Received in revised form 6 July 2026; Accepted 22 September 2026

Available online 26 September 2026

0172-2190/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

(continued)

evidence rather than relying solely on internal
(parametric) knowledge.

1. Introduction

Patents represent the world's largest repository of codified technical knowledge, providing essential insights for innovation management, competitive intelligence, and R&D strategy [1–3]. However, extracting value from patent data has historically presented significant computational and analytical challenges. The extreme document length, specialised legal terminology, and deliberate linguistic ambiguity characteristic of patent texts frequently undermine the effectiveness of traditional text mining and bibliometric approaches [4]. For decades, researchers have relied on keyword-based retrieval and static topic modelling to analyse patent landscapes [5]. While these methods have demonstrated effectiveness, they struggle to capture the semantic nuance essential for accurate patent analysis, often failing to identify emerging technological convergence or unoccupied technology spaces [6].

The emergence of Large Language Models (LLMs) represents a transformative development in how intellectual property is created, analysed, and managed. Unlike previous generations of Natural Language Processing (NLP) systems, Transformer-based architectures have enabled models to move beyond simple classification to complex reasoning and content generation tasks [7]. We are witnessing a fundamental transition from patent information retrieval toward knowledge synthesis. Recent studies demonstrate that LLMs can now automate tasks previously requiring domain expertise, including claim drafting [8,9], standard-essential patent identification [10], and structured data extraction from unstructured text [11]. The release of patent-specific benchmarks [12] and the publication of special issues on generative AI applications in patent analytics [13] signal that this technology has transitioned from experimental exploration to operational deployment.

Despite the exponential growth in research output, particularly evident in the surge of publications during 2024 and 2025, the literature remains fragmented [14]. Existing reviews, such as Jiang et al. [4], provide excellent coverage of generic NLP techniques like LSTM and Word2Vec. However, they do not sufficiently address the paradigm shift introduced by Generative AI, which moves beyond information retrieval to content creation. The majority of current research emphasises technical optimisation rather than the practical integration of LLMs into patent analytics workflows. While numerous studies propose architectures for improving patent classification [15,16] or evaluating summarisation quality [17,18], they often lack a unified framework for understanding how these tools integrate into organisational processes. Furthermore, significant challenges persist, including hallucination risks in prior art retrieval [19] and the opacity of AI decision-making processes [15,20,21], which present substantial barriers to widespread adoption. Innovation managers, therefore, lack a comprehensive framework for the systematic integration of LLMs, specifically regarding where these tools offer genuine value versus where they introduce unacceptable risks [22,23].

To address this gap, this article presents a systematic literature review of 129 studies following the PRISMA 2020 protocol [24], with the objective to systematically review and synthesise the empirical literature on the application of LLMs and Generative AI to patent analytics, in order to (i) map current applications onto the stages of the patent lifecycle, (ii) characterise the dominant architectural and methodological trends, and (iii) identify the principal research gaps and a forward research agenda for the field. We organise the literature using a patent lifecycle framework. Unlike machine learning taxonomies (e.g., classification vs. generation), this lifecycle approach aligns with the actual

workflow of patent attorneys and examiners [25,26], offering a more practical roadmap for organisational integration. The patent lifecycle framework encompasses three distinct phases: (1) creation and pre-filing, including invention disclosure generation and automated drafting; (2) examination and prosecution, including semantic prior art search and office action responses; and (3) post-grant management, including portfolio valuation, litigation analysis, and technology forecasting. Fig. 1 (Section 2) summarises the structural and legal characteristics of patent documents that motivate the architectural responses discussed throughout this review.

The remainder of this paper is structured as follows. Section 2 provides the theoretical background and technological foundations related to patent analysis and Generative AI. Section 3 details the review methodology, outlining the data sources, selection process, and the inclusion and exclusion criteria applied within the PRISMA 2020 protocol. Section 4 presents the core findings, organised around the three patent lifecycle stages. Section 5 offers a critical discussion and synthesis of the identified patterns and trends. Section 6 proposes future research directions, and Section 7 concludes with remarks on the strategic integration of generative AI into patent information management.

2. Theoretical background and technological foundations

To understand the transformative potential of Large Language Models (LLMs) in patent analytics, it is necessary to first characterise the unique constraints inherent to patent data and the technological developments that address them.

2.1. The unique challenge of patent data

Patents are structurally distinct from other forms of technical literature, such as scientific publications or technical reports [23]. They combine legal terminology with detailed technical specifications, deliberately crafted to maximise the scope of protection while satisfying statutory disclosure requirements [15]. Critically, relevant information is not concentrated in a single section but is dispersed across heterogeneous components: the claims, which delineate the legal boundaries of protection; the detailed description, which provides technical enablement; and the drawings, which offer visual representations of the invention [4,27,28].

This structural dispersion gives rise to a significant “long-context” challenge. Standard patent descriptions frequently exceed the token limits of conventional NLP models, often surpassing 11,000 tokens and reaching as high as 180,000 tokens in complex cases [4,29]. Effective processing, therefore, requires advanced chunking or segmentation strategies [30,31]. The domain is also inherently multimodal and multilingual. A single patent family may comprise documents across multiple languages and contain chemical structures or mechanical diagrams that text-only models are unable to interpret [32,33]. Consequently, generic models trained on general-purpose corpora frequently underperform when applied to the precision demands of intellectual property analysis without domain-specific adaptation [34].

Patent documents pose domain-specific modelling challenges that are less prominent in general-purpose NLP. Their length, legal–technical language, and multimodal structure create recurrent bottlenecks in context handling, factual grounding, and visual-semantic reasoning. To improve reliability in legally sensitive tasks, recent patent-oriented AI systems increasingly combine long-context processing, grounded generation (e.g. RAG), and multimodal modelling. Fig. 1 shows that patent documents are not simply another application domain for general-purpose NLP, since their structural and legal characteristics introduce specific modelling constraints. In this sense, it helps to explain why the literature has moved toward certain architectural responses and, more broadly, strengthens the transition of this paper from the background section to the later discussion of methodological developments.

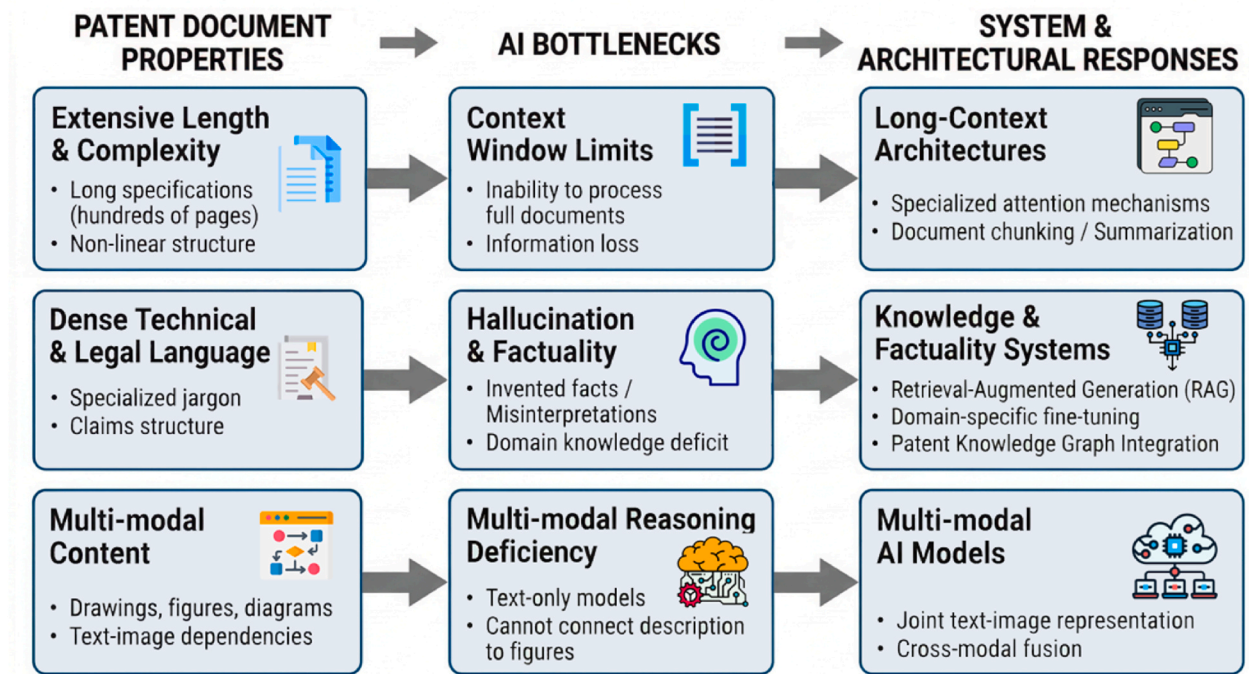


Fig. 1. Conceptual mapping between core patent document properties, recurrent AI bottlenecks, and corresponding architectural responses in patent analytics.

2.2. The shift from statistical to semantic analysis

The application of text mining to patent data has historically relied on bag-of-words approaches, including keyword counting and Boolean search operators [35]. While these methods have demonstrated utility for straightforward retrieval tasks, they are fundamentally limited by the broader vocabulary mismatch problem. This challenge encompasses not only synonymy but also polysemy, where identical terms carry distinct meanings across technical domains [36]. Patent applications routinely employ such techniques to circumvent prior art or to broaden the scope of claims. This linguistic variability renders keyword-dependent methods ineffective for capturing conceptual similarity across documents.

The introduction of the Transformer architecture addressed this limitation by enabling the generation of contextual semantic embeddings. Unlike statistical models that operate on word frequency [37–40], Transformer-based architectures such as BERT and RoBERTa map textual content into high-dimensional vector spaces, in which proximity reflects conceptual relatedness rather than lexical overlap [41,42]. This capability enables the identification of white space opportunities or prior art that is conceptually identical but linguistically distinct [6]. Empirical evaluations indicate that embedding-based retrieval approaches substantially outperform keyword-based baselines in patent classification and technology convergence tasks [16,43].

2.3. Generative AI and the RAG framework

While discriminative models have demonstrated strong performance in classification tasks, the emergence of generative models such as GPT-4 and Llama has introduced the capacity for content creation within the patent domain. Applications now include automated claim drafting, invention summarisation, and the identification of novel technological opportunities [29,44]. Nevertheless, generative AI presents a well-documented limitation in high-stakes professional contexts, e.g. the tendency to produce plausible but factually inaccurate outputs, commonly referred to as hallucination [19]. In patent analytics, this risk is particularly consequential, as erroneous patent citations or technical

details could have significant legal implications.

To address this challenge, the literature has increasingly converged on Retrieval-Augmented Generation (RAG) as a dominant architectural pattern [45,46]. In a RAG framework, the model does not rely solely on its parametric knowledge. Instead, it retrieves relevant documents from verified external sources, such as WIPO PATENTSCOPE or the USPTO database, to ground its output in the retrieved evidence. This approach substantially reduces hallucination rates while preserving the generative capability of LLMs. Although RAG does not eliminate the risk of inaccurate outputs entirely, empirical studies across legal and technical domains indicate that it offers a meaningful improvement in factual reliability, particularly for tasks such as prior art analysis and freedom-to-operate assessments [9].

2.4. Agentic AI and multimodal systems

The most recent developments in LLM research, predominantly from 2024 onwards, have introduced the concept of Agentic AI, systems capable of autonomous decision-making and tool use within a loop, to the patent domain. Unlike single-prompt interactions, agentic systems orchestrate multiple specialised AI components that perform distinct roles within a unified workflow [47,48]. For instance, a planner decomposes complex tasks into sub-tasks, a retriever sources relevant data, and a critic evaluates the output for accuracy and coherence. This modular architecture enables multi-step reasoning that mirrors the analytical workflow of a human patent professional, facilitating tasks such as comparative infringement analysis and technology road mapping [49]. In parallel, multimodal systems are evolving to process text, images, and structured data simultaneously. This integration is particularly relevant to patent analytics, where invention disclosures frequently rely on drawings and diagrams to convey technical content that cannot be fully captured through text alone [50,51].

3. Methodology

To capture the rapidly evolving landscape of LLMs in patent analytics, this study employs a systematic literature review methodology

following the PRISMA 2020 guidelines [24]. Scopus and Web of Science (WoS) were selected as the primary databases to ensure comprehensive coverage across computer science, engineering, legal, and management disciplines. The temporal scope was restricted to post-2017 publications to align with the emergence of the Transformer architecture and subsequent advancements in generative AI.

3.1. Data sources and selection process

Scopus and WoS represent the two most widely used databases for systematic reviews in interdisciplinary research. This dual-database strategy was designed to maximise coverage: Scopus provides broader indexing of technical conference proceedings, which is particularly important for capturing emerging AI developments, while WoS ensures inclusion of high-impact journals with rigorous selection criteria.

The search strategy was designed to capture the intersection of GenAI and Patent Analytics. Table 1 presents the exact search strings applied in each database. In Web of Science, the Topic (TS) field was used, covering Title, Abstract, Author Keywords, and Keywords Plus. In Scopus, the TITLE-ABS-KEY field was used, covering Title, Abstract, and Author Keywords. Both searches were restricted to Articles, Conference Papers, and Reviews published.

This search strategy was executed in January 2026, capturing publications through early 2026. The technology terms were selected to encompass both general LLM architectures and specific, widely adopted models. Domain terms were chosen to capture the full spectrum of patent-related applications, from creation through enforcement.

3.2. Inclusion and exclusion criteria

The initial search yielded 1289 documents: 892 from Scopus and 397 from Web of Science. Duplicate records across Scopus and Web of Science were identified manually, by matching on author names, titles, and publication years, and removed before screening. Title and abstract screening and full-text eligibility assessment were conducted by a single reviewer against the criteria in Table 1. An 8.7% overlap was identified between the databases ($n = 112$ duplicates), indicating that the dual-search strategy successfully captured distinct scholarly communities.

The main stages of the documented literature selection process are summarised in Fig. 2, which shows the progression from database identification and duplicate removal to screening and final study inclusion, thereby providing a concise overview of the selection pathway underpinning the review corpus.

The selection criteria applied during the screening process are summarised in Table 2. Studies were included if they applied Transformer-based or generative AI models to patent-related tasks and were published in English from 2018 onward. Studies were excluded if they employed only traditional machine learning approaches without Transformer architectures, focused exclusively on legal or policy discussions without technical implementation, or addressed AI-generated content outside the patent domain.

The final corpus of 129 studies represents research spanning computer science, information science, engineering, and management

disciplines. These studies were then systematically categorised according to the patent lifecycle framework for structured synthesis in Section 4.

3.3. Data extraction and synthesis

For each included study, the title and abstract were read to determine the primary patent-related task addressed (e.g., drafting, prior art search, classification, litigation analysis) and the study was categorised accordingly into one of the three patent lifecycle phases and associated sub-categories described in Section 1. Where a study addressed tasks spanning more than one sub-category, it was listed under each applicable category. A formal quality or risk-of-bias assessment was omitted because the included literature was highly heterogeneous, spanning NLP benchmarking, applied systems, and bibliometric analyses with differing reporting standards.

3.4. Methodology limitations

The review is restricted to peer-reviewed, indexed literature in English. Grey literature such as patent office technical reports and unindexed conference proceedings was excluded, which is a consequential scoping decision for a practitioner audience given that a meaningful share of operationally relevant work circulates in these channels. The English-only filter is a further relevant constraint, as CNIPA, JPO, and KIPO are among the most active deployers of LLM-based examination tools among major patent offices, meaning non-English literature from these jurisdictions is likely underrepresented in the corpus. Second, screening and eligibility assessment were conducted by a single reviewer following a structured protocol against pre-defined criteria (Table 1), independent dual screening was not performed, which is a limitation relative to PRISMA best practice and may introduce undetected selection bias. Finally, no formal quality or risk-of-bias assessment was applied to included studies. Given the high methodological heterogeneity of the literature, which spans machine learning benchmarking, applied systems development, and bibliometric analysis, a single evaluation instrument was not readily transferable across study types. We acknowledge this as a limitation relative to standard systematic review guidelines.

4. Findings

The temporal distribution of the selected studies (see Fig. 3) reveals that the application of LLMs to patent analytics represents an emerging but rapidly expanding research domain. The public release of ChatGPT in November 2022 and the introduction of Meta's Llama models in February 2023 mark a critical inflexion point in the field. After an absence of publications in 2018 and 2019, research activity between 2020 and 2023 remained modest, averaging fewer than five studies per year. In contrast, the field expanded rapidly in 2024 (41 studies) and 2025 (71 studies). This sharp increase, accounting for more than 88% of the total corpus, coincides temporally with the widespread availability of generative models via both proprietary platforms and open-source alternatives.

This volume shift is accompanied by a shift in research questions. Pre-2023 studies are predominantly foundational and single-task, exemplified by Lee and Hsiang's (2020–2022) work fine-tuning GPT-2 for claim generation and early embedding-based classification studies (Song et al., 2023; Zhang et al., 2023). In contrast, 2024–25 studies are predominantly applied, multi-component systems addressing verification, retrieval grounding, and end-to-end workflows, such as AWACo-pilot, LE-PARIS, JURISMIND, and MemGraph. This reflects a move from asking whether an LLM can perform a given task toward how to make an LLM-based system for that task reliable enough for operational deployment.

To understand where LLMs are generating value within patent

Table 1

Search strings for the academic databases with a temporal filter of publication year > 2017.

Database	Search String
Web of Science	TS=(“Large Language Model” OR “LLM” OR “Generative AI” OR “Transformer Model” OR “BERT” OR “GPT” OR “Llama”) AND TS=(“Patent” OR “Intellectual Property” OR “Prior Art” OR “Invention Disclosure” OR “Patent Claims”)
Scopus	TITLE-ABS-KEY(“Large Language Model” OR “LLM” OR “Generative AI” OR “Transformer Model” OR “BERT” OR “GPT” OR “Llama”) AND TITLE-ABS-KEY(“Patent” OR “Intellectual Property” OR “Prior Art” OR “Invention Disclosure” OR “Patent Claims”)

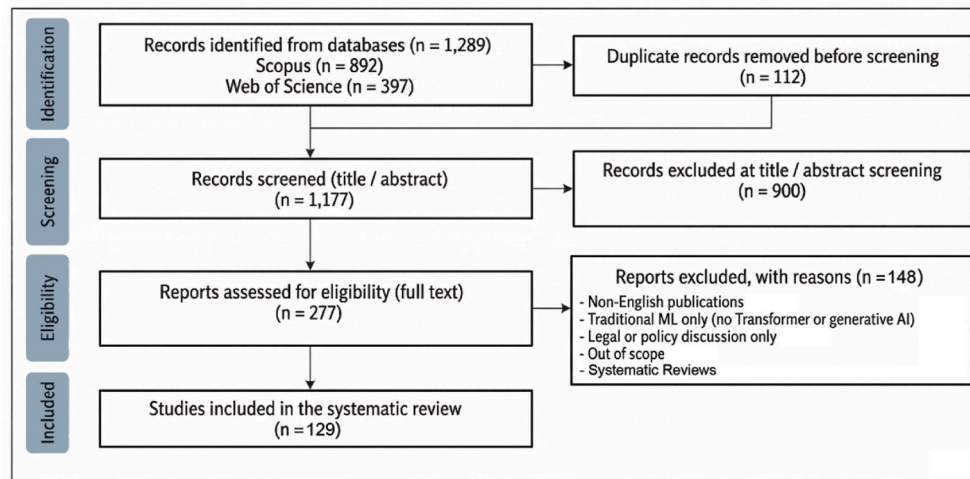


Fig. 2. Simplified PRISMA flow diagram showing the literature selection process.

Table 2
Literature selection criteria.

Criteria	Inclusion (Keep)	Exclusion (Discard)
Timeframe	2018 – Present	Papers published before 2018
Language	English	Non-English
Topic	Studies applying Transformer-based or generative models to patent tasks	Papers using only traditional ML (e.g., SVM, LDA, Word2Vec) or keyword-based methods
Domain	Patent applications (e.g., drafting, search, classification, examination)	Legal or policy discussions without technical implementation; AI-generated art or copyright issues unrelated to patents

information management, the literature was categorised according to the patent lifecycle framework. As illustrated in Fig. 4 and detailed in Table 3, research activities are distributed across all three lifecycle stages, with a notable concentration in post-grant applications. Worth noting that studies that address more than one lifecycle phase or sub-category are listed under each applicable sub-category in Tables 3–5.

Creation & Pre-filing (N = 36): A substantial portion of the literature examines the generative capabilities of LLMs, with automated patent drafting (20 studies) and idea generation (13 studies) representing the primary focus areas. This distribution reflects a shift from analytical information retrieval toward content creation applications.

Examination & Prosecution (N = 34): Research in this phase concentrates predominantly on prior art search and retrieval (21 studies), addressing the critical need to reduce search latency and improve semantic recall during patent examination. The remaining studies in this category focus on novelty assessment and office action response generation.

Post-Grant Management (N = 60): This phase constitutes the largest research domain, driven primarily by extraction, forecasting, and landscaping applications (31 studies). This emphasis suggests that researchers are prioritising the application of LLMs to mine existing patent repositories for competitive intelligence, technology trend analysis, and portfolio management. Additional post-grant studies address classification, litigation analysis, and infringement detection.

The distribution pattern indicates that while LLMs are being explored across the entire patent lifecycle, the majority of empirical research has focused on extracting value from existing patent databases rather than supporting the creation of new intellectual property. This may reflect both the lower risk associated with analytical applications and the greater availability of historical patent data for model training and validation. The following sections provide a detailed synthesis of the specific applications, performance metrics, and managerial implications within each lifecycle stage.

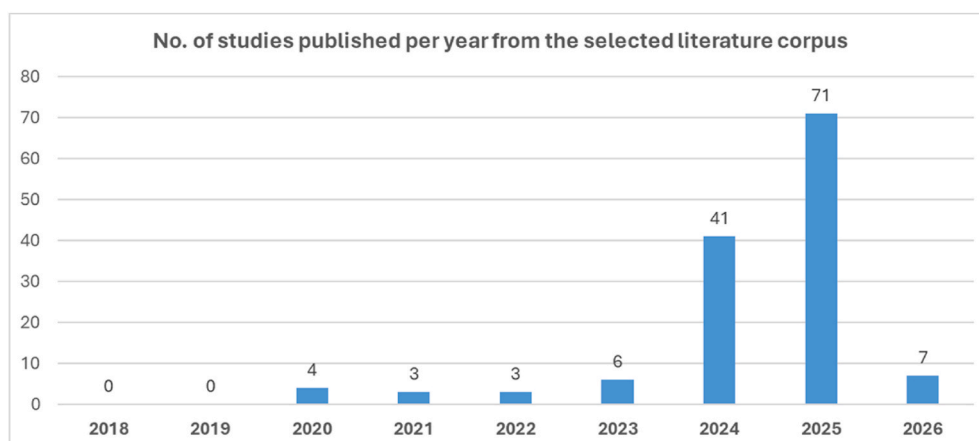


Fig. 3. Number of studies published per year from the literature reviewed.

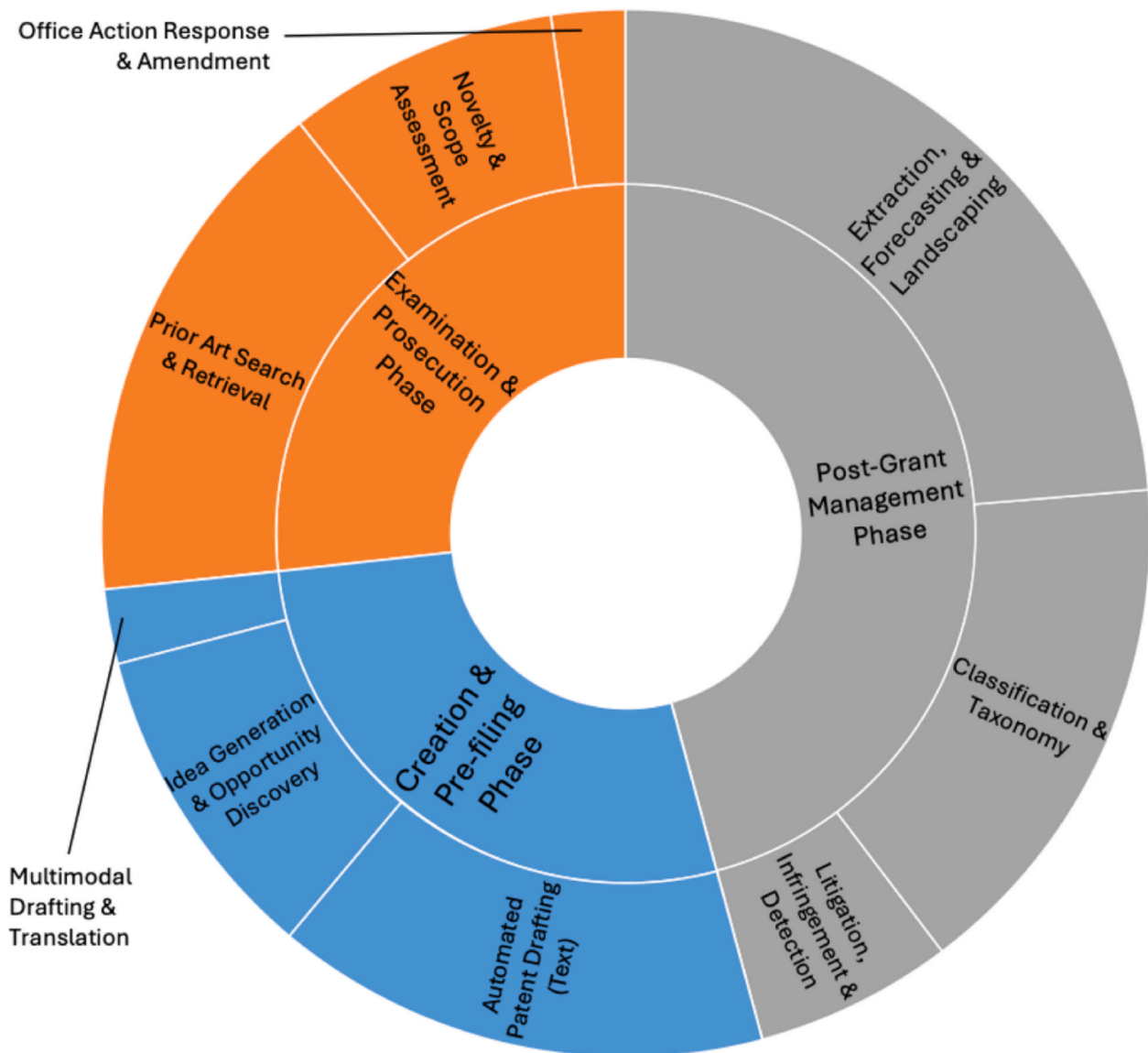


Fig. 4. Sunburst chart of studies categorised according to patent lifecycle.

4.1. Creation & Pre-filing phase

A detailed breakdown of the Creation & Pre-filing Phase is presented in Table 4, which categorises studies addressing idea generation, automated drafting, and multimodal applications.

The review demonstrates a transformation in the early stages of innovation, moving from passive opportunity identification toward active, generative ideation enabled by LLMs. While traditional approaches relied predominantly on expert intuition or keyword-based mapping, contemporary research increasingly employs generative AI to synthesise novel technical concepts directly from patent corpora.

4.1.1. Idea generation & opportunity discovery

A substantial portion of the literature focuses on structuring unstructured patent text to facilitate ideation. Siddharth et al. [45] propose a method to extract explicit engineering design knowledge from patent descriptions to support RAG. By fine-tuning language models for relation identification, they demonstrate how structured knowledge bases can guide LLMs to generate technically coherent and verifiable responses for design-intensive tasks. Park et al. [44] introduce a data-driven morphological analysis framework that reduces expert

subjectivity. By utilising Latent Dirichlet Allocation (LDA) for dimension extraction and named entity recognition (NER) for value extraction, they demonstrate how LLMs can systematically construct morphological matrices. Kim et al. [55] advance this approach by introducing a semi-supervised topic modelling pipeline combined with network analysis. Their framework specifically addresses cross-domain insights, enabling engineers to solve specific technological challenges by retrieving function-oriented solutions from unrelated technical fields.

Beyond structural analysis, researchers are leveraging the reasoning capabilities of LLMs to generate actionable invention concepts. Liu et al. [56] argue that while general-purpose LLMs often produce technically infeasible concepts, a customisation framework using RAG can align model outputs with specific design constraints, thereby enhancing concept feasibility. Ren et al. [58] employ fine-tuning and reinforcement learning from human feedback to create models capable of generating technically viable patent concepts that outperform baseline approaches. Bagdady et al. [52] provide empirical evidence of LLMs' creative utility by analysing ChatGPT-generated ideas for prosthetic devices, noting that while LLM accelerates the ideation process, the novelty of generated concepts requires careful verification against existing prior art.

Several studies propose comprehensive multi-agent or catalogue-

Table 3

Summary of studies reviewed categorised into patent lifecycles. The sum is > 129 because some papers belong to more than one sub-category.

Creation & Pre-filing Phase	36
Idea Generation & Opportunity Discovery	13
Automated Patent Drafting (Text)	20
Multimodal Drafting & Translation	3
Examination & Prosecution Phase	34
Prior Art Search & Retrieval	21
Novelty & Scope Assessment	10
Office Action Response & Amendment	3
Post-Grant Management Phase	60
Classification & Taxonomy	21
Litigation, Infringement & Detection	8
Extraction, Forecasting & Landscaping	31
Generic Systematic Reviews	6

Note: The sub-category counts sum to more than the unique corpus size of N = 129 because studies addressing more than one lifecycle phase or sub-category are listed under each applicable category.

Table 4

Studies that address the Creation & Pre-filing Phase.

Sub-category	Literature
Idea Generation & Opportunity Discovery	Bagdady et al. [52]; Chen et al. [53]; Jiang et al. [54]; Kanumolul et al. [47]; Kim et al. [55]; Liu et al. [56]; Nomaguchi et al. [57]; Park et al. [44]; Ren et al. [58]; Rydholm et al. [59]; Siddharth et al. [45]; Terao et al. [48]; Wang et al. [29];
Automated Patent Drafting (Text)	Ahmed et al. [32]; Berdnyk et al. [60]; Chen et al. [61]; Gao et al. [62]; Higashiyama [63]; Htun et al. [64]; Jiang et al. [65]; Lee [66]; Lee [8]; Lee et al. [67]; Lee et al. [68]; Lee et al. [69]; Lee et al. [70]; Mawaldi et al. [9]; Ren et al. [58]; Shomee et al. [71]; Wang et al. [29]; Xiao et al. [34]; Yoo et al. [28]; Zuo et al. [18];
Multimodal Drafting & Translation	Shukla et al. [33]; Wei et al. [72]; Yoo et al. [28]

Table 5

Examination and prosecution phase.

Sub-category	Literature
Prior Art Search & Retrieval	Ali et al. [35]; Awale et al. [27]; Chikkamath et al. [19]; David Raj et al. [73]; Ding et al. [74]; Jin et al. [75]; Kosonocky et al. [76]; Kosonocky et al. [77]; Lee et al. [78]; Liu et al. [79]; Lo et al. [80]; Prasad et al. [81]; Pustu-Iren et al. [82]; Shaghrouzi et al. [83]; Siow et al. [36]; Vangala et al. [84]; Waterstraat et al. [85]; Xiong et al. [86]; Xiong et al. [87]; Zhang et al. [88]; Zheng et al. [89]
Novelty & Scope Assessment	Gao et al. [90]; Herzberg [10]; Ikoma et al. [91]; Knappich et al. [92]; Lee et al. [93]; Parikh et al. [94]; Ragot [95]; Schmitt [96]; Zheng et al. [97]; Zhu [98]
Office Action Response & Amendment	Chu et al. [46]; Chu et al. [99]; Kawano et al. [100]

based systems for idea generation and opportunity discovery. Kanumolul et al. [47] introduce Agent Ideate, a multi-agent framework designed to automate product idea generation from patent texts, representing a move toward autonomous innovation systems. Nomaguchi et al. [57] utilise LLMs to extract problem-solution pairs from patents to create design catalogues, enabling designers to systematically explore solution spaces for specific challenges. Wang et al. [29] and Terao et al. [48] further explore collaborative invention processes, where LLMs guide the selection and refinement of ideas to optimise for novelty and

technical validity.

4.1.2. Automated patent drafting (text)

The application of LLMs to patent drafting has evolved from simple text completion toward sophisticated, multi-stage generation of complex legal documents. The literature in this category highlights the tension between automation efficiency and the strict formal requirements of patent offices, focusing predominantly on claim generation, specification drafting, and translation.

Claims constitute the legally defining component of a patent, and their automation represents a primary research focus. Lee et al. [68] established early baselines by fine-tuning GPT-2 for independent claim generation, a method subsequently refined by Lee [8], who introduced reinforcement learning to generate claims with higher semantic novelty rather than just linguistic fluency. Chen et al. [61] specifically evaluated LLMs in the biomedical domain, noting that while smaller models can learn claim style, they often struggle with the precise dependency structures required in complex chemical or biological claims. Jiang et al. [65] address the challenge of claim revision and generation from descriptions, creating the Patent-CR dataset to benchmark how effectively models can refine claims based on examiner feedback or newly discovered prior art.

Beyond claim generation, recent systems attempt to produce complete patent documents. Wang et al. [29] introduce Patentformer, an approach that generates specifications by synthesising information from both claims and drawing descriptions, addressing the challenge of maintaining coherence across lengthy documents. Shomee et al. [71] contribute PatentWriter, a benchmarking framework for abstract generation, demonstrating that modern LLMs can produce abstracts that often surpass human-written baselines in readability, though technical accuracy remains a concern requiring human verification. Gao et al. [62] present a pipeline for converting research papers directly into patent documents, utilising prompt engineering to map academic discourse into patent legal structure.

Patent translation is treated as a specialised form of drafting within this domain. Ahmed et al. [32] and Berdnyk et al. [60] highlight that patent translation requires distinct linguistic models compared to general translation, with models like GPT-3 demonstrating superior accuracy in preserving legal semantics compared to older recurrent neural network architectures. The shared translation tasks at WMT 2024, discussed by Higashiyama [63] and Htun et al. [64], confirm that fine-tuned LLMs now outperform traditional neural machine translation systems in Japanese-English and Chinese-Japanese patent translation tasks.

The deployment of LLMs in patent drafting raises concerns regarding data security and quality control. Mawaldi et al. [9] propose AWACo-pilot, an on-premise solution acknowledging that law firms cannot risk data leakage via cloud-based LLMs. On the evaluation front, Xiao et al. [34] investigate the distinction between AI-generated and human-written texts, constructing the Patent Abstract Detection dataset to help maintain the integrity of the patent system. Zuo et al. [18] and Yoo et al. [28] provide multi-dimensional metrics to assess not only fluency but also legal and technical faithfulness of generated content.

4.1.3. Multimodal drafting & translation

Patents are inherently multimodal, relying on figures to substantiate text claims. This emerging sub-category addresses the complex challenge of integrating vision and language models to automate drafting tasks that require visual understanding. For example, Wei et al. [72] focus on extracting visual descriptors from figure captions, facilitating better retrieval and indexing of patent images. Shukla et al. [33] introduce PatentLMM, a large multimodal model specifically designed to generate detailed descriptions for patent figures. Their work demonstrates that multimodal models can effectively “read” patent drawings to produce coherent textual embodiments. Complementing this, Yoo et al. [28] also contribute to the understanding of how multimodal

inputs (text + metadata/images) can refine classification and drafting accuracy.

While the generative capabilities in this phase are promising, methodological limitations persist. Most studies rely on n-gram metrics (ROUGE, BLEU), which correlate poorly with legal validity. Furthermore, few studies (e.g., Mawaldi et al. [9]) incorporate “human-in-the-loop” evaluations by qualified attorneys, leaving a gap in understanding the legal risks of automated drafting.

4.2. Examination & Prosecution Phase

A detailed breakdown of the Examination and Prosecution Phase is illustrated in Table 5, which categorises studies addressing prior art search, novelty assessment, and office action response.

4.2.1. Prior art search & retrieval

The literature on prior art search demonstrates a transition from keyword-centric approaches toward semantic frameworks that address the vocabulary mismatch problem inherent in patent texts. Jin et al. [75] specifically address this challenge for Chinese patent retrieval by optimising keyword-based element extraction, highlighting the necessity for language-specific processing strategies. David Raj et al. [73] and Ali et al. [35] propose query expansion techniques to bridge the gap between user queries and complex patent documents. The former employs feature concatenation of BERT and GPT-3 embeddings with optimal Bi-Clustering, while the latter utilises abstract and key term extraction to train domain-specific BERT models.

A significant architectural trend in prior art search and retrieval involves the integration of Retrieval-Augmented Generation (RAG) to ground LLMs in verified patent databases. Ding et al. [74] introduce an automated system combining vector retrieval with RAG to enhance semantic relevance, achieving substantial recall improvements over baseline LLM approaches. Lee et al. [78] propose PAI-NET, which incorporates prior art relationship information directly into the retrieval network, eliminating the need for manual ontology construction. Xiong et al. [87] advance this approach by introducing MemGraph, which activates an LLM’s parametric memory to traverse a memory graph, identifying related entities to expand queries for more accurate semantic matching. In the legal domain, Prasad et al. [81] present JURISMIND, a context-driven retrieval system designed to reduce hallucination in legal question-answering by ensuring that retrieved prior art satisfies legal criteria.

Recent work extends into multimodal and domain-specific search applications. For example, Pustu-Iren et al. [82] established early multimodal approaches combining visual and textual embeddings for patent retrieval. Lo et al. [80] address the long-tail problem in patent image retrieval by using LLMs to generate fine-grained descriptions of patent figures, enabling more effective visual-semantic alignment. Domain-specific extraction is explored by Kosonocky et al. [76], [77] and Vangala et al. [84], who utilise LLMs to extract chemical reaction datasets and functional landscapes from patent literature, demonstrating that LLMs can parse complex chemical syntax that previously challenged standard NLP models. Siow et al. [36] and Waterstraat et al. [85] demonstrate the necessity of tailored search approaches for specialised domains such as nursing care technology and sintered metals, where generic search strategies prove insufficient.

4.2.2. Novelty & scope assessment

This category focuses on the evaluative tasks of determining whether an invention is novel and defining the boundaries of its legal protection. The literature demonstrates a progression toward computational approaches that emulate the cognitive processes of patent examiners through logical decomposition and contrastive learning mechanisms.

Ikoma et al. [91] present a dataset and task formulation that simulates the examination process by comparing claims against cited prior art, finding that generative models outperform classification models in

generating explanations for novelty rejections. Zheng et al. [97] introduce TCLMA, a multi-agent framework utilising two-dimensional contrastive learning to enhance scientific novelty evaluation, demonstrating substantial reductions in model hallucination rates. In the telecommunications domain, Herzberg [10] employs generative AI to assess the standard-essentiality of 5G patents, a critical task for determining fair, reasonable, and non-discriminatory licensing terms.

Quantifying the scope of patent protection is addressed through several computational approaches. Ragot [95] proposes a mathematical framework measuring claim scope as the reciprocal of probability derived from language models, offering a computational proxy for legal breadth. Gao et al. [90] focus on prediction, using fine-grained claim dependency graphs to forecast patent approval, arguing that structural relationships between claims predict both scope and grantability. Lee et al. [93] frame patent novelty as a textual entailment problem with their PatentEdits dataset, enabling models to detect revisions made during prosecution. Schmitt [96] provides a systematic analysis using GPT-4 to disentangle the overlapping concepts of patent quality, value, and novelty, establishing a multidimensional assessment framework.

4.2.3. Office action response & amendment

Research on automated office action response represents an emerging domain in human-AI collaboration for patent prosecution, evolving from passive document analysis toward active legal reasoning and argument generation. Kawano et al. [100] developed ClaimBrush, a framework for automated patent claim refinement that constructs paired datasets of pre- and post-examination claims and employs preference optimisation techniques to align model outputs with examiner revision patterns. Chu et al. [99] presented LE-PARIS (LLM Enhanced Patent Office Action Response Intelligence System), extending their earlier template-driven approach through a hybrid recommendation mechanism that suggests response strategies and leverages large language models to formulate legal arguments. Longitudinal validation of this system demonstrated measurable improvements in attorney performance during prosecution.

Despite these advances, several methodological challenges persist in prosecution support systems. The reliance on proprietary benchmarks limits reproducibility, with approximately 20% of studies in this domain providing open-access code or datasets for validation (e.g., Prasad et al. [81], Knappich et al. [92]). Major systems, including AWACopilot [9] and LE-PARIS [99], remain closed-source. Furthermore, evidence regarding query expansion strategies remains inconclusive: while such techniques enhance retrieval recall, studies suggest they may introduce semantic drift, particularly in densely technical domains where irrelevant prior art is retrieved [74].

4.3. Post-Grant Management Phase

A detailed breakdown of the Post-Grant Management Phase is illustrated in Table 6, which covers topics such as patent classification, infringement detection, and technology forecasting and landscaping.

4.3.1. Classification & taxonomy

The literature on patent classification has transitioned from static keyword matching toward sophisticated, hierarchy-aware deep learning frameworks that address the structural complexities of patent classification schemes such as the Cooperative Patent Classification (CPC) and International Patent Classification (IPC). Rafieian et al. [110] introduce a transformer-based pipeline that preserves the graph-based taxonomy of CPC labels, achieving substantial improvements in precision and recall on the USPTO-70 k dataset compared to flat classification models. This hierarchical approach is complemented by Li et al. [43], who address large-grained classification for specific technological domains such as Smart Cities and Internet of Things. They propose a multilabel regression model designed for small-sample learning, arguing that for strategic policy analysis, broad categorisation often provides greater

Table 6
Post-grant management phase.

Sub-category	Literature
Classification & Taxonomy	Gavrilov et al. [101]; Kamateri et al. [15]; Li et al. [43]; Li et al. [102]; Lin et al. [103]; Manjunath et al. [104]; Mesa et al. [105]; Motamedi et al. [106]; Nanba et al. [107]; Nkolongo et al. [108]; Phyu et al. [109]; Rafeian et al. [110]; Shaghoezi et al. [83]; Song et al. [41]; Tang et al. [111]; Xiong et al. [87]; Yoshikawa et al. [112]; Yoshikawa et al. [113]; Yu et al. [114]; Zhang et al. [115]; Zhang et al. [42]
Litigation, Infringement & Detection	Koutsopoulos et al. [116]; Park et al. [117]; Trappey et al. [50]; Trappey et al. [118]; Trapp et al. [119]; Xiao et al. [34]; Xiong et al. [86]; Yu et al. [120]
Extraction, Forecasting & Landscaping	Ai et al. [11]; Afifuddin et al. [6]; Cano-Marín [22]; Campanile et al. [20]; Choi et al. [21]; Giordano et al. [121]; Gong et al. [122]; Hu [123]; Hu et al. [124]; Maghsoudi et al. [49]; Jiang et al. [125]; Kim et al. [55]; Kim et al. [126]; Kim et al. [127]; Kriesch et al. [128]; Lai [129]; Mejia et al. [2]; Mohammadi et al. [130]; Ni et al. [12]; Park et al. [23]; Pazhouhan et al. [131]; Pelaez et al. [132]; Petralia [133]; Rosa et al. [134]; Shi et al. [135]; Sohn et al. [136]; Trappey et al. [118]; Trapp et al. [119]; Urata et al. [137]; Vangala et al. [84]; Zheng [138]

utility than fine-grained subgrouping.

A notable trend in patent classification involves utilising LLMs to refine both input data and classification structures. Yoshikawa et al. [112], [113] demonstrate that feeding raw patent text into classifiers is inefficient; instead, they employ GPT-3.5 to generate concise summaries of claims and descriptions. Fine-tuning models such as RoBERTa on these AI-generated summaries yielded higher F1 scores than using original texts, indicating that LLMs can effectively distil classification-relevant features. Nanba et al. [107] support semantic interoperability by automating the construction of multilingual patent thesauri, using fine-tuned ChatGPT models to align hypernyms between Japanese and English patent documents. Furthermore, Motamedi et al. [106] employ LLMs to semi-automatically refine taxonomies themselves, iteratively consolidating overly detailed CPC groups into more abstract, practical taxonomies for innovation tracking while reducing human workload and maintaining semantic coherence.

To address annotation costs in patent classification, Xiong et al. [87] introduce an iterative LLM-assisted active learning framework. By using LLMs to estimate uncertainty and select informative samples, they reduced manual annotation effort by 60% in a drone technology dataset while maintaining high Macro-F1 scores. For institutional applications, Mesa et al. [105] propose a system utilising LLM function calling to streamline the retrieval and cleaning of university patent data, reducing hallucination errors common in generative outputs.

4.3.2. Litigation, infringement & detection

The application of LLMs to legal aspects of the patent lifecycle has transitioned from passive document retrieval toward active risk assessment and infringement detection. Trappey et al. [50], [118] provide a comprehensive framework for litigation mining using the unmanned aerial vehicle industry as a case study. They employ LLMs to construct litigation ontologies and extract inventive processes from judgment texts, enabling visualisation of legal disputes and identification of infringement risks through Technology Function Matrices. This work demonstrates the potential of LLMs to bridge the gap between technical patent claims and legal interpretations.

As generative AI increasingly serves as a tool for invention drafting, detecting AI-generated content has become a critical validity concern. Xiao et al. [34] address patent system integrity by developing a benchmark to detect AI-generated patent abstracts. They analyse linguistic patterns to distinguish between human-written and machine-generated texts, a necessary capability to prevent the

proliferation of low-quality, AI-generated filings. Park et al. [117] explore latent patterns in computer technology patents to trace LLM usage in drafting, identifying specific tokens that have increased in frequency following the release of ChatGPT. Addressing code infringement, Yu et al. [120] introduce CodeIPPrompt, a testing framework to evaluate whether code generation models inadvertently reproduce licensed code, thereby violating intellectual property rights.

4.3.3. Extraction, forecasting & landscaping

This category demonstrates the evolution of patent analytics from simple bibliometrics toward deep semantic forecasting and large-scale structured dataset creation. Gong et al. [122] contribute Deep-InnovationAI, a dataset mapping global AI innovation by linking 2.3 million patents with 3.5 million academic papers. By utilising dual-layer BERT classifiers and hypergraph analysis, they provide infrastructure for tracking knowledge transfer from academic research to industrial applications.

In technological forecasting, researchers are increasingly employing dynamic modelling to predict trajectories. Trappey et al. [118] utilise Dynamic Topic Modelling on 6G patents to reveal the evolution of technology clusters such as autonomous navigation and network slicing, identifying these as high-risk litigation zones and emerging innovation hotspots. Mohammadi et al. [130] perform network analysis on Generative AI patents, identifying a pivotal transition post-2016 from modular innovations toward integrated generative frameworks. Afifuddin et al. [6] focus on predictive modelling for technology convergence, while Kim et al. [55] propose a framework for Technology Opportunity Analysis that uses LLMs to synthesise cross-domain insights, enabling engineers to solve specific problems by retrieving functional solutions from unrelated technical fields.

The extraction of structured knowledge from unstructured text represents another significant research direction to offer insight in patent landscaping. Ai et al. [11] demonstrate the capability of LLMs to extract chemical reaction data and organic synthesis procedures from patents, unlocking information previously embedded in image-heavy or complex chemical texts. Choi et al. [21] introduce PatentHAN, a hierarchical attention network that enhances interpretability by identifying pivotal claims that drive future citations, facilitating early identification of breakthrough technologies. Campanile et al. [20] propose an AI-driven methodology for Internet of Things patents that combines relevance classification with impact estimation, moving beyond keyword search toward semantic value assessment.

5. Discussion and synthesis

This review indicates that the integration of Large Language Models (LLMs) into patent analytics has matured beyond experimental applications. The literature demonstrates a fundamental shift in the research trajectory: from static performance metrics toward architectural robustness, human-AI collaboration frameworks, and navigation of complex legal environments.

5.1. Architectural evolution

A critical synthesis of the reviewed literature consolidates and characterises the architectural shift toward Retrieval-Augmented Generation (RAG) as the primary mechanism for mitigating the inherent limitations of generative AI in high-stakes legal contexts [56,74]. While earlier studies prioritised classifier performance metrics such as F1 scores and accuracy on static datasets like USPTO-3M [110], current research emphasises faithfulness and verifiability [99]. This emphasis on faithfulness is difficult to verify independently, as approximately 80% of studies in the reviewed corpus rely on proprietary or closed benchmarks and code, a structural feature of the current literature with direct consequences for reproducibility and cross-study comparison (see Section 6.1.1). The issue of hallucination, where models generate

plausible but factually incorrect technical details or legal citations, has been identified as a critical barrier to operational deployment [19,81].

In response, researchers have increasingly adopted RAG frameworks that ground LLM outputs in verified external knowledge bases. This is evident in the evolution from purely generative models to systems such as PAI-NET, which explicitly incorporates prior art relationship information into the retrieval network, eliminating the need for manual ontology construction [78]. Similarly, JURISMIND demonstrates that combining base prompts with dynamically retrieved legal texts substantially reduces hallucination rates compared to static fine-tuning approaches [81]. The evolution of RAG frameworks is becoming increasingly sophisticated, such as MemGraph, which activates the parametric memory of LLMs to traverse memory graphs, identifying related entities to expand queries rather than relying solely on vector similarity [86]. This architectural trend suggests that advances in patent AI depend not primarily on model scale, but on architectures that effectively integrate semantic generation with evidence-based retrieval.

5.2. The managerial shift

The literature also signals a distinct transition from the objective of fully autonomous AI systems toward the practical deployment of AI as assistive tools [23,139]. While LLMs have demonstrated the capability to draft claims and specifications that approximate the style of patent attorneys, they fundamentally lack the domain expertise and strategic reasoning required for robust intellectual property protection [61,65]. Consequently, the value proposition has shifted from replacing human experts toward augmenting their productivity and consistency [9].

This human-in-the-loop paradigm is supported by empirical findings indicating that while AI can reduce drafting time substantially, it requires expert oversight to ensure technical specificity and legal compliance [139,140]. Systems such as AWACopilot are explicitly designed as assistive platforms to support patent attorneys, acknowledging that the nuanced interpretation of patentability criteria remains a human-centric task [9]. In the ideation domain, frameworks such as PatentGPT and data-driven morphological analysis are positioned as tools to stimulate human creativity by revealing unexpected technological combinations rather than functioning as autonomous inventors [44,58]. This synthesis suggests that the principal managerial implication of generative AI in intellectual property lies in the transformation of the patent professional's role, from routine drafter to strategic curator and evaluator of AI-generated outputs.

5.3. Systemic integrity

While the technological capabilities of AI in patent analytics have advanced rapidly, a substantial governance gap has emerged, wherein legal frameworks have not kept pace with technical developments. The literature highlights three critical areas of concern: inventorship, data privacy, and intellectual property infringement.

First, cases such as DABUS and subsequent litigation across multiple jurisdictions underscore the absence of international consensus on whether AI systems can be named as inventors, creating legal uncertainty for AI-generated innovations [58,139,140]. Second, operational deployment faces significant privacy challenges. For instance, the risk of data leakage when uploading confidential patent descriptions to cloud-based LLMs has necessitated the development of secure, on-premise solutions [61]. Third, the generative nature of these models raises intellectual property concerns. Tools such as CodeIPrompt have been developed to assess whether code generation models inadvertently reproduce licensed code, highlighting the risk that users may unknowingly violate intellectual property rights [120]. This governance gap is further complicated by the need for detection mechanisms. As AI-generated filings increase, patent offices face the challenge of distinguishing between human-authored and machine-generated texts to maintain examination process integrity [34,117]. The synthesis of these

issues indicates that widespread adoption depends as much on legal and regulatory clarity as on algorithmic advancement.

6. Future research directions

Analysis of selected studies reveals that while the technical feasibility of applying LLMs to patent tasks has been demonstrated, the systemic implications of this integration remain insufficiently understood. Drawing from the synthesis presented in Section 5, we identify eight critical research directions under three categories. To provide actionable guidance, these directions are organised into a temporal framework based on logical dependencies and technological maturity (see Fig. 5).

Immediate priorities include domain-specific validation and contextual architectures. Establishing legally valid evaluation metrics constitutes a prerequisite for rigorous comparative analysis. Reliance on surface-level metrics such as ROUGE and BLEU remains problematic, as these measures capture only lexical overlap while failing to assess semantic coherence, factual accuracy, or legal precision. Without advancement beyond n-gram matching approaches, claims of performance improvement lack empirical verifiability. Concurrently, determining optimal trade-offs between RAG and extended context windows represents an urgent architectural decision. Recent empirical studies demonstrate that while extended context windows reduce retrieval complexity, they introduce computational overhead and attention degradation, whereas retrieval-augmented approaches maintain precision at lower token costs but require careful retrieval system design.

Medium-term priorities encompass multimodal reasoning, innovation curation, and legal explainability. These areas depend on the maturation of vision-language models and the development of inherently interpretable architectures. Legal contexts demand transparency mechanisms that satisfy judicial requirements for accountability, necessitating glass-box models or robust post-hoc explanation frameworks rather than opaque systems.

Long-term research addresses systemic consequences, including professional role transformation, economic validation, and data stewardship. Effects on patent professional competencies and potential corpus degradation through model collapse represent lagging indicators requiring longitudinal observation. Model collapse, wherein systems trained recursively on synthetic data exhibit progressive performance degradation, poses particular risk to patent information quality and demands proactive governance frameworks.

6.1. Multimodal and architectural advancement

Current research exhibits a substantial bias toward text processing. However, patents are inherently multimodal documents in which the inventive concept is frequently conveyed through drawings, chemical structures, or flowcharts that cannot be fully captured through textual description alone.

6.1.1. Domain-specific validation

A consistent finding across the literature is the inadequacy of standard NLP evaluation metrics for patent applications. Metrics such as ROUGE and BLEU, which measure n-gram overlap, demonstrate weak or negligible correlations with expert assessments of patent claim quality. Recent efforts have introduced domain-specific evaluation frameworks such as PatentScore [28] and PatClaimEval [141] to assess legal validity, structural compliance, and feature completeness. The field requires standardisation of comprehensive benchmark datasets that evaluate legal reasoning, enablement, and definiteness assessment rather than solely measuring linguistic fluency or surface-level textual similarity. Such benchmarks should encompass the full range of patent quality dimensions, including novelty assessment, claim scope adequacy, and technical disclosure sufficiency.

A critical parallel challenge concerns the absence of shared, open-

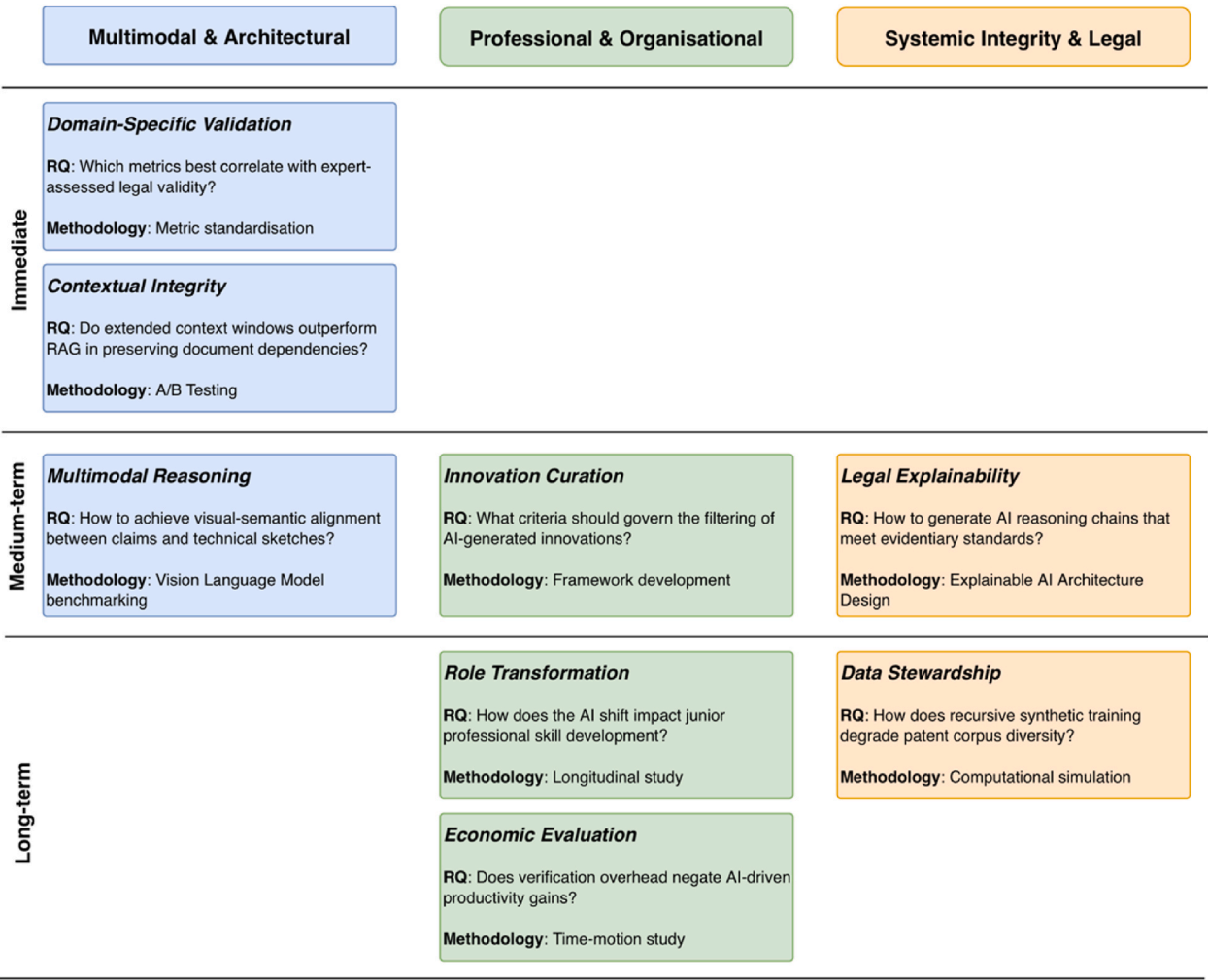


Fig. 5. Future research agenda for LLM integration in patent analytics.

access evaluation datasets. Approximately 80% of studies reviewed rely on proprietary or closed benchmarks, severely limiting reproducibility and cross-study comparison. The release of open datasets such as PEDANTIC, which targets definiteness assessment in patent claims [92], and MoZIP, which provides multilingual IP benchmarking [12], represent early steps toward shared infrastructure. However, these address isolated task types, a unified evaluation suite covering the full patent lifecycle remains absent. Future research should also address the disconnect between automated metrics and legal expert judgement identified across drafting [18] and retrieval tasks [19] alike. Establishing annotation protocols that incorporate qualified patent attorneys as evaluators, alongside automated metrics, would provide the empirical foundation necessary for meaningful claims of performance advancement in this domain. Recent work in legal NLP more broadly suggests that human-in-the-loop evaluation design is achievable at scale when structured annotation guidelines are shared openly.

6.1.2. Contextual integrity

Patent analytics exposes a coupled bottleneck that cannot be treated as two independent limitations. Patents are inherently long and structurally heterogeneous documents, and high-stakes legal–technical outputs amplify the impact of hallucinations. RAG has emerged as the dominant architectural response [142,143], yet chunk-based retrieval may fragment cross-sectional dependencies across claims, embodiments, and figures. Extended-context LLMs offer an alternative, but empirical evidence shows attention degrades over long sequences, with

models becoming “lost in the middle” [144]. Neither approach eliminates hallucination [145,146], and length extrapolation techniques do not guarantee attribution fidelity [147].

Future architectures should adopt structure-aware segmentation aligned with patent semantics (claim hierarchies, embodiment paragraphs, drawing references) rather than uniform token splitting, and combine this with hierarchical, multi-stage retrieval. Graph-augmented retrieval, where evidence is structured as a network of claim elements, citations, and classification ontologies, offers a promising path to maintaining legal coherence that unstructured vector retrieval cannot guarantee. Hallucination mitigation should be treated as a verification problem, incorporating entailment-style checks and critic models [148] rather than relying solely on prompting strategies. Evaluation frameworks must go beyond lexical overlap metrics, explicitly measuring long-context degradation and attribution fidelity [149].

6.1.3. Multimodal reasoning

Patents are inherently multimodal documents, yet the overwhelming majority of LLM applications reviewed rely exclusively on text. Future research should progress beyond text-only Transformers toward Multimodal Large Language Models capable of interpreting the specialised visual syntax of patent schematics. Current general-purpose multimodal models such as LLaVA and MiniGPT-4 encounter difficulties with patent figures, which typically consist of sparse, binary sketches involving node-link diagrams rather than natural photographic images. Recent work has begun addressing this challenge through specialised encoders

such as PatentMME [33] to capture the structural elements of patent figures and language-informed approaches for image retrieval that incorporate figure captions and classifications. Nevertheless, progress remains fragmented. Vision models are rarely evaluated against the full visual vocabulary of patent drawings, which encompasses mechanical assemblies, circuit diagrams, chemical structures, and flowcharts within a single document family [27,82].

A critical and largely unaddressed research gap concerns cross-modal alignment, specifically, grounding visual embeddings in claim language to identify infringement risks and novelty boundaries that textual analysis alone cannot detect. Patent examiners routinely use figures to resolve ambiguities in claim scope that the written description leaves open. Future architectures must therefore process text, images, and structured graph data simultaneously, and be evaluated on tasks requiring integrated reasoning across modalities rather than on unimodal proxies. Promising directions include contrastive pre-training adapted for the sparse-sketch domain, and the development of patent-specific multimodal benchmarks analogous to PatentScore [28] but incorporating figure-grounded assessment criteria.

6.2. Professional and organisational impact

The integration of agentic AI systems transforms the technology from a passive analytical tool into an active agent, raising critical questions for research and development management regarding workforce transformation and economic value creation.

6.2.1. Innovation curation

The capacity of generative AI to produce large volumes of plausible product concepts and patent ideas introduces a qualitatively new challenge of effective selection and prioritisation. Research is needed on AI-assisted innovation funnel management systems that can autonomously evaluate technical feasibility, market potential, and strategic fit to filter AI-generated concepts. Critical questions include how organisations should structure processes to manage the increased volume of candidate innovations, what criteria should govern selection decisions when ideas are machine-generated rather than expert-proposed, and whether traditional stage-based processes remain effective when dealing with AI-augmented ideation at scale.

6.2.2. Role transformation

As routine tasks such as data collection, preprocessing, and initial search query formulation become automated, the role of the patent professional is evolving from drafter toward strategic curator. LLMs are increasingly conceptualised not as replacements but as assistive systems that require human oversight to address hallucination risks and ensure technical accuracy. A critical research gap concerns the long-term impact on the profession: whether this evolution will result in deskilling of junior analysts who no longer perform foundational tasks, or upskilling whereby analysts must develop capabilities in prompt engineering and strategic IP portfolio management. Empirical longitudinal studies are needed to assess how these technological changes reshape professional development pathways and required competencies within patent analytics roles.

6.2.3. Economic validation

While technical capabilities continue to advance, the economic case for LLM adoption requires rigorous empirical validation. Preliminary estimates suggest that LLM-based analysis could reduce costs substantially compared to traditional expert-driven qualitative analysis. However, these potential savings must be evaluated against verification costs and the expenses associated with secure, on-premise deployments necessitated by data confidentiality requirements in cloud-based systems. A critical research gap concerns quantifying the “verification overhead”: whether the time and resources required to validate AI-generated outputs and prevent hallucination errors negate the

efficiency gains realised in initial content generation. Researchers should conduct controlled field experiments (A/B testing) comparing the billable hours and error rates of LLM-assisted attorneys versus traditional workflows.

6.3. Systemic integrity and legal governance

The widespread adoption of generative AI in patent systems introduces systemic risks concerning data integrity, legal enforceability, and the long-term quality of the patent corpus.

6.3.1. Data stewardship

As patent offices potentially encounter increasing volumes of AI-generated applications, a risk emerges that future models will be trained predominantly on synthetic, machine-generated text rather than human-authored disclosures. This circularity could lead to the homogenisation of patent language and the degradation of the global patent corpus as a knowledge resource. A critical research gap concerns the development of robust detection mechanisms to identify AI-generated patent texts while maintaining examination process integrity. Fundamental questions include: What is the long-term impact of the prevalence of synthetic data on prior-art quality and on the effectiveness of future patent analytics models? How does recursive training on AI-generated content affect model performance and the semantic diversity of patent language? Research is needed on sustainable data curation strategies that preserve the human-originated knowledge foundation of the patent system while accommodating AI-assisted drafting practices.

6.3.2. Legal explainability

When AI systems are employed to assess patent validity or determine standard-essential patent status during litigation, the opaque nature of model reasoning becomes a legal liability. Courts require transparent, auditable justifications for legal conclusions. Current research emphasises the necessity of explainability mechanisms that transcend binary classification outputs to generate structured reasoning that articulates why a claim satisfies or fails to satisfy patentability criteria. A critical research gap concerns developing Explainable AI architectures specifically designed for legal scrutiny. Can models provide interpretable attention mechanisms at the claim level that identify specific textual elements supporting legal conclusions? How can AI systems generate reasoning chains that satisfy evidentiary standards in judicial proceedings? Research should explore architectures such as hierarchical attention networks that reveal which patent sections and prior art references contribute to validity assessments, ensuring that AI-driven legal analyses can withstand cross-examination and judicial review.

6.4. Research direction summary

The research directions identified in this review form an interconnected ecosystem rather than isolated challenges, wherein technical advancements, professional workflows, and institutional integrity converge. Fig. 5 synthesises this research agenda by mapping primary questions and methodologies across three foundational pillars.

Effective integration of large language models into patent workflows depends on cascading interdependencies among these pillars. Advances in multimodal reasoning and contextual integrity constitute prerequisites for organisational value realisation. The verification overhead that currently constrains economic adoption, for instance, can be addressed only through domain-specific validation metrics and legal explainability frameworks developed within subsequent pillars.

Concurrently, as patent professionals transition toward strategic curation roles, a systemic vulnerability emerges: model collapse arising from recursive training on synthetic data. Research demonstrates that models trained exclusively on their own outputs exhibit progressive performance degradation, particularly affecting rare events in the

original data distribution. Mitigating this risk requires a socio-technical approach that aligns vision-language architectures with practitioner requirements and institutional transparency standards. By pursuing this integrated framework, future research can advance beyond fragmented performance metrics toward a resilient and trustworthy patent information ecosystem.

7. Conclusion

This systematic review of 129 studies demonstrates that the integration of Large Language Models into patent analytics represents a significant development in innovation management research and practice. The analysis reveals a transition from keyword-based retrieval systems toward AI-assisted knowledge synthesis, where advanced language models support tasks ranging from automated drafting to cross-domain classification and comparative technical analysis.

The analysis across the patent lifecycle consolidates and characterises a notable architectural trend: researchers are increasingly adopting Retrieval-Augmented Generation frameworks as a mechanism to balance the generative capabilities of LLMs with the legal requirement for verifiable, factually grounded outputs. The evidence indicates that while hallucination remains a challenge, it is being addressed through document chunking strategies, reranking algorithms, and grounded citation mechanisms that anchor model outputs in verified patent databases.

This review identifies several critical research directions. First, the field must progress beyond text-only analysis toward multimodal architectures capable of processing the visual and diagrammatic elements that constitute essential components of patent disclosures. Second, new evaluation frameworks are needed that assess legal reasoning quality, enablement adequacy, and claim definiteness rather than relying solely on surface-level linguistic similarity metrics. Third, longitudinal empirical studies are required to understand how AI integration affects professional skill development, organisational innovation processes, and patent corpus quality over time. Finally, research is needed on governance frameworks that address systemic risks, including recursive training on synthetic data and the explainability requirements of legal decision-making contexts.

For patent analytics practitioners and innovation managers, this review highlights two key implications. First, as AI systems reduce the time and cost associated with generating invention disclosures and conducting prior art searches, competitive advantage increasingly depends on effective curation and strategic evaluation of AI-generated outputs rather than on disclosure volume alone. Organisations must develop capabilities in prompt engineering, output verification, and strategic intellectual property portfolio management. Second, the deployment of AI-assisted tools necessitates workforce adaptation. The role of patent professionals is evolving from routine drafting toward strategic oversight, requiring the development of new competencies in AI system management, quality assurance, and cross-functional coordination between technical, legal, and business functions.

Large Language Models are reshaping patent information management by enabling capabilities that extend beyond automation of existing tasks. As generative AI systems become increasingly integrated into patent creation, examination, and enforcement processes, critical questions emerge regarding governance, quality assurance, and the preservation of innovation system integrity. The central challenge is not whether LLMs will transform intellectual property analytics, but how organisations and institutions will govern these technologies to ensure they enhance rather than compromise the quality and reliability of patent information as a knowledge resource for innovation. Patent offices are not merely observers of this governance challenge: several, including the USPTO, EPO, and UKIPO, have already issued guidance on AI-assisted examination and on applicants' use of generative AI in drafting, positioning offices as active governance actors whose evolving practice will shape, rather than simply respond to, this transition.

CRediT authorship contribution statement

Pingfei Jiang: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Andrea Pietro Arena:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Salvatore Sorce:** Writing – review & editing, Writing – original draft, Methodology. **Mark Atherton:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors wish to thank the anonymous reviewers for their time and constructive feedback. No funding was received for this research. No individuals were interviewed or otherwise involved as participants in this study.

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Data availability

No data was used for the research described in the article.

References

- [1] H. Ernst, Patent information for strategic technology management, *World Pat. Inf.* 25 (2003) 233–242, [https://doi.org/10.1016/S0172-2190\(03\)00077-2](https://doi.org/10.1016/S0172-2190(03)00077-2).
- [2] C. Mejia, Y. Kajikawa, Patent research in academic literature. Landscape and trends with a focus on patent analytics, *Front. Res. Metr. Anal.* 9 (2024) 1484685, <https://doi.org/10.3389/frma.2024.1484685>.
- [3] J. Lee, S. Park, J. Lee, A study on the monitoring of technology innovation through patent analysis, *Hum. Soc. Sci. Commun.* 13 (1) (2025), <https://doi.org/10.1057/s41599-025-06363-w>.
- [4] L.K. Jiang, S.M. Goetz, Natural language processing in the patent domain: a survey, *Artif. Intell. Rev.* 58 (7) (2025), <https://doi.org/10.1007/s10462-025-11168-z>.
- [5] L. Aristodemou, F. Tietze, The state-of-the-art on intellectual property analytics (IPA): a literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data, *World Pat. Inf.* 55 (2018) 37–51, <https://doi.org/10.1016/j.wpi.2018.07.002>.
- [6] M. Afifuddin, W. Seo, Developing a predictive model for anticipating technology convergence: a transformer-based model and supervised learning approach, *PLoS One* 20 (6 June) (2025), <https://doi.org/10.1371/journal.pone.0326417>.
- [7] A.J.C. Trappey, Y.Y.C. Lin, C.Y. Wu, Generative AI-based intelligent patent summarization for intellectual property knowledge communication and cooperation, *World Pat. Inf.* 83 (2025), <https://doi.org/10.1016/j.wpi.2025.102410>.
- [8] J.-S. Lee, Generating patent claims with semantic novelty, *World Pat. Inf.* 83 (2025), <https://doi.org/10.1016/j.wpi.2025.102404>.
- [9] M.H. Mawaldi, Z. Kastrati, A. Gustafsson (Eds.), *AWACopilot: a Secure On-Premise Large Language Model-based Solution for Enhanced Patent Drafting*, CEUR Workshop Proceedings, 2025.
- [10] A. Herzberg, Assessing the standard-essentiality of 5G technology patents by means of generative artificial intelligence, *World Pat. Inf.* 81 (2025), <https://doi.org/10.1016/j.wpi.2025.102363>.
- [11] Q. Ai, F. Meng, J. Shi, B. Pelkie, C.W. Coley, Extracting structured data from organic synthesis procedures using a fine-tuned large language model, *Digital Discov.* 3 (9) (2024) 1822–1831, <https://doi.org/10.1039/d4dd00091a>.
- [12] S. Ni, M. Tan, Y. Bai, F. Niu, M. Yang, B. Zhang, et al. (Eds.), *Mozip: a Multilingual Benchmark to Evaluate Large Language Models in Intellectual Property*, 2024.
- [13] T. Trippe, J.S.J. Lee, Special issue on applications of generative AI and large language models in the patent domain, *World Pat. Inf.* 82 (2025), <https://doi.org/10.1016/j.wpi.2025.102365>.
- [14] N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, et al., *The AI index 2025 annual report*, in: *Institute for Human-Centered AI, Stanford University, Stanford, CA*, 2025.
- [15] E. Kamateri, M. Salampasis, E. Pérez-Molina, Will AI solve the patent classification problem? *World Pat. Inf.* 78 (2024) <https://doi.org/10.1016/j.wpi.2024.102294>.

- [16] B. Rafieian, P.P. Vázquez, Improved multi-label hierarchical patent classification using LLMs, *World Pat. Inf.* 81 (2025), <https://doi.org/10.1016/j.wpi.2025.102356>.
- [17] J. Ding, H. Chen, S. Kolapudi, L. Pobbathi, H. Nguyen (Eds.), Quality Evaluation of Summarization Models for Patent Documents. IEEE International Conference on Software Quality, Reliability and Security, QRS, 2023, <https://doi.org/10.1109/QRS60937.2023.00033>.
- [18] Y. Zuo, K. Gerdes, E. de la Clergerie, Sagot B. PatentEval, Understanding errors in patent generation, in: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1*, 2024, pp. 2687–2710. LONG PAPERS.
- [19] R. Chikkamath, A. Sharma, C. Hewel, M. Endres (Eds.), ChatGPT Vs. Google Gemini: Assessing AI Frontiers for Patent Prior Art Search Using European Search Reports, *CEUR Workshop Proceedings*, 2024.
- [20] An AI-Driven methodology for patent evaluation in the IoT sector: assessing relevance and future impact, in: L. Campanile, R. Zona, A. Perfetti, F. Rosatelli (Eds.), *International Conference on Internet of Things, Big Data and Security, IoTBDs - Proceedings*, 2025, <https://doi.org/10.5220/0013519700003944>.
- [21] J. Choi, J. Yoon, C. Lee, Early screening of potential breakthrough technologies with enhanced interpretability: a patent-specific hierarchical attention network model, *Comput. Ind. Eng.* 203 (2025), <https://doi.org/10.1016/j.cie.2025.111034>.
- [22] E. Cano-Marin, The transformative potential of generative artificial intelligence (GenAI) in business: a text mining analysis on innovation data sources, *ESIC Market* 55 (2) (2024), <https://doi.org/10.7200/esicm.55.333>.
- [23] S. Park, G. Kim, S. Lee, Evaluating the value of LLMs in patent-based technology intelligence: toward increasing efficiency and reducing expert dependency, *Technol. Forecast. Soc. Change* 222 (2026), <https://doi.org/10.1016/j.techfore.2025.124375>.
- [24] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *PLoS Med.* 18 (3) (2021) e1003583, <https://doi.org/10.1371/journal.pmed.1003583>.
- [25] A. Ali, A. Tufail, L.C. De Silva, P.E. Abas, Innovating patent retrieval: a comprehensive review of techniques, trends, and challenges in prior art searches, *Appl. Syst. Innovat.* 7 (5) (2024), <https://doi.org/10.3390/asi7050091>.
- [26] A. Shahzeb, The role of artificial intelligence in enhancing patent lifecycle management, *Int. J. Scient. Res. Comput. Scie. Eng. Inform. Technol.* 11 (2) (2025) 2194–2205, <https://doi.org/10.32628/cseit23112582>.
- [27] S. Awale, E. Muller-Budack, R. Delaviz, R. Ewerth (Eds.), Exploring Patents Visually: an Interactive Search System for Multimodal Patent Image Search and Interpretation, *CEUR Workshop Proceedings*, 2025.
- [28] Y. Yoo, Q. Xu, L. Cao, PatentScore: multi-dimensional evaluation of LLM-generated patent claims, *arXiv:2505.19345* (2025), <https://doi.org/10.48550/arXiv.2505.19345>.
- [29] J. Wang, S.K.R. Mudhiganti, M. Sharma (Eds.), Patentformer: a Novel Method to Automate the Generation of Patent Applications, 2024, <https://doi.org/10.18653/v1/2024.emnlp-industry.101>.
- [30] F. Gobet, P.C.R. Lane, S. Croker, P.C.-H. Cheng, G. Jone, I. Oliver, et al., Chunking mechanisms in human learning, *Trends Cognit. Sci.* 5 (6) (2001) 236–243.
- [31] S. Musslick, L.K. Bartlett, S.H. Chandramouli, M. Dubova, F. Gobet, T.L. Griffiths, et al., Automating the practice of science: opportunities, challenges, and implications, *Proc. Natl. Acad. Sci. U. S. A.* 122 (5) (2025) e2401238121, <https://doi.org/10.1073/pnas.2401238121>.
- [32] M. Ahmed, A. Ouda, M. Abusharkh, S. Kohli, K. Rai, An optimized approach to translate technical patents from English to Japanese using machine translation models, *Appl. Sci. (Switzerland)* 13 (12) (2023), <https://doi.org/10.3390/app13127126>.
- [33] PATENTLMM: large multimodal model for generating descriptions for patent figures, in: S. Shukla, N. Sharma, M. Gupta, A. Mishra (Eds.), *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, <https://doi.org/10.1609/aaai.v39i19.34257>.
- [34] Z. Xiao, W. Hu, Y. Wu, W. Chen, H. Li, X. Liu (Eds.), On the Gap Between AI-Generated and Human-Written Patent Texts. Lecture Notes in Computer Science, 2025, https://doi.org/10.1007/978-981-96-2882-7_14.
- [35] A. Ali, M.A. Humayun, L.C. De Silva, P.E. Abas, Optimizing Patent Prior Art Search: an Approach Using Patent Abstract and Key Terms. *INFORMATION*, 2025, <https://doi.org/10.3390/info16020145>, 16(2).
- [36] Exploring the patent landscape of sintered metal technologies: an analysis using LLM-based AI patent search, in: K.S. Siow, W. Wang, R. Bahru, X. Long, H.W. Lee (Eds.), *Proceedings of the IEEE/CPMT International Electronics Manufacturing Technology (IEMT) Symposium*, 2024, <https://doi.org/10.1109/IEMT61324.2024.10875250>.
- [37] Y.-H. Tseng, C.-J. Lin, Y.-I. Lin, Text mining techniques for patent analysis, *Inf. Process. Manag.* 43 (5) (2007) 1216–1247, <https://doi.org/10.1016/j.ipm.2006.11.011>.
- [38] M. Atherton, P. Jiang, D. Harrison, A. Malizia, Design for invention: annotation of functional geometry interaction for representing novel working principles, *Res. Eng. Des.* 29 (2) (2017) 245–262, <https://doi.org/10.1007/s00163-017-0267-2>.
- [39] G. Kim, J. Bae, A novel approach to forecast promising technology through patent analysis, *Technol. Forecast. Soc. Change* 117 (2017) 228–237, <https://doi.org/10.1016/j.techfore.2016.11.023>.
- [40] P. Jiang, M. Atherton, S. Sorce, D. Harrison, A. Malizia, Design for invention: a framework for identifying emerging design-prior art conflict, *J. Eng. Des.* 29 (10) (2018) 596–615, <https://doi.org/10.1080/09544828.2018.1520204>.
- [41] B. Song, C. Luan, D. Liang, Identification of emerging technology topics (ETTs) using BERT-based model and semantic analysis: a perspective of multiple-field characteristics of patented inventions (MFCOPIs), *Scientometrics* 128 (11) (2023) 5883–5904, <https://doi.org/10.1007/s11192-023-04819-x>.
- [42] C. Zhang, S. Jian, L. Chao, C. Fan, L. Bo, L. Chen (Eds.), A Semantic Understanding Method for Patent Text Based on Large Language Model, 2023, <https://doi.org/10.1109/EI259745.2023.10513309>.
- [43] M. Li, L. Wang, Leveraging patent classification based on deep learning: the case study on smart cities and industrial internet of things, *J. Informetr.* 19 (1) (2025), <https://doi.org/10.1016/j.joi.2024.101616>.
- [44] J. Park, Y. Geum, LLM-enhanced idea generation: data-driven morphological analysis with LDA and NuNER, *Comput. Ind. Eng.* 208 (2025), <https://doi.org/10.1016/j.cie.2025.111426>.
- [45] L. Siddharth, J.X. Luo, Retrieval augmented generation using engineering design knowledge, *Knowl. Base Syst.* 303 (2024), <https://doi.org/10.1016/j.knsys.2024.112410>.
- [46] J.M. Chu, H.C. Lo, J. Hsiang, C.C. Cho (Eds.), Patent Response System Optimised for Faithfulness: Procedural Knowledge Embodiment with Knowledge Graph and Retrieval Augmented Generation, 2024.
- [47] G. Kanumolul, A. Urlana, C.V. Kumar, B.M. Garlapati, Agent ideate: a framework for product idea generation from patents using agentic AI. *PROCEEDINGS OF THE 2ND WORKSHOP ON AGENT AI FOR SCENARIO PLANNING*, 2025, pp. 50–57.
- [48] Y. Terao, Y. Tachioka, Collaborative invention: team ditlab at the AgentScen shared task - refining patent-based product ideation via LLM-guided selection and rewriting. *PROCEEDINGS OF THE 2ND WORKSHOP ON AGENT AI FOR SCENARIO PLANNING*, 2025, pp. 67–81.
- [49] M. Maghsoudi, N. Mohammadi, S.M.A. Mousavi Roudsari, A multi-layered patent analytics framework for technology roadmapping in the circular economy, *Clean. Eng. Technol.* 29 (2025), <https://doi.org/10.1016/j.clet.2025.101107>.
- [50] A.J.C. Trappey, S.C. Chou, G.K.J. Li, Patent litigation mining using a large language model—Taking unmanned aerial vehicle development as the case domain, *World Pat. Inf.* 80 (2025), <https://doi.org/10.1016/j.wpi.2024.102332>.
- [51] A survey on patent analysis: from NLP to multimodal AI, in: H.H. Shomee, Z. Wang, S.N. Ravi, S. Medya (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025.
- [52] K.R. Bagdady, A. Hoffman, S.S. Patel, P. Jagasia, I. Herzog, G.A. Dumanian, et al., Artificial intelligence-driven innovation: patent and publication analysis of ChatGPT-generated ideas for upper extremity prosthetics, *Plastic Reconstr. Surg., Glob. Open* 13 (11) (2025), <https://doi.org/10.1097/GOX.00000000000007309>.
- [53] L.Q. Chen, Y.X. Song, S.X. Ding, L.Y. Sun, P. Childs, H.Y. Zuo, et al., TRIZ-GPT: an LLM-augmented method for problem-solving, in: *PROCEEDINGS OF ASME 2024 INTERNATIONAL DESIGN ENGINEERING TECHNICAL CONFERENCES AND COMPUTERS AND INFORMATION IN ENGINEERING CONFERENCE 62024, IDETC-CIE*, 2024.
- [54] Z.M.J. Jiang, A. Liu, D.W. Zhang, X.W. Xu, Y. Dai, Customization and personalization of large language models for engineering design, *CIRP Ann. Manuf. Technol.* 74 (1) (2025) 191–195, <https://doi.org/10.1016/j.cirp.2025.03.001>.
- [55] J. Kim, S. Park, S. Lee, Technology opportunity analysis for creating innovative solutions: a framework for cross-field insights using patent data, *Adv. Eng. Inform.* 70 (2026), <https://doi.org/10.1016/j.aei.2025.104116>.
- [56] C. Liu, N.I. Fomin, S. Xiao, G. Sun, Y. Lyu, How ChatGPT is shaping next-generation patent solutions, *Buildings* 15 (13) (2025), <https://doi.org/10.3390/buildings15132273>.
- [57] Creating design catalog from patent documents with large language model for design concept generation, in: Y. Nomaguchi, K. Kuroishi, A. Syamimi, D. Tanaka, K. Okamoto, K. Fujita (Eds.), *Proceedings of the Design Society*, 2025, <https://doi.org/10.1017/pds.2025.10119>.
- [58] R. Ren, J. Ma, J. Luo, Large language model for patent concept generation, *Adv. Eng. Inform.* 65 (2025), <https://doi.org/10.1016/j.aei.2025.103301>.
- [59] E. Rydholm, T. Bastys, E. Svensson, C. Kannas, O. Engkvist, T. Kogej, Expanding the chemical space using a chemical reaction knowledge graph, *Digital Discov.* 3 (7) (2024) 1378–1388, <https://doi.org/10.1039/d3dd00230f>.
- [60] L. Berdnyk, T. Vysotska, S. Korotkova, N. Moskalenko, O. Cherkashchenko, Cognitive models of patent translation, *J. Theor. Appl. Inf. Technol.* 102 (24) (2024) 9031–9040.
- [61] F.-C. Chen, C.-L. Pan, Evaluating application of large language models to biomedical patent claim generation, *World Pat. Inf.* 80 (2025), <https://doi.org/10.1016/j.wpi.2025.102339>.
- [62] W. Gao, J. Yang, Q. Yu, X. Shangguan, N. Xiao, X. Ling (Eds.), An AI-Powered Patent Document Generation System Based on Large Language Model, 2025, <https://doi.org/10.1109/ICCAI66501.2025.00059>.
- [63] S. Higashiyama (Ed.), Results of the WAT/WMT 2024 Shared Task on Patent Translation, *Conference on Machine Translation - Proceedings*, 2024.
- [64] O. Htun, A. Poncelas (Eds.), Rakuten's Participation in WMT 2024 Patent Translation Task. *Conference on Machine Translation - Proceedings*, 2024.
- [65] L. Jiang, C. Zhang, P.A. Scherz, S. Goetz, Can Large Language Models Generate high-quality Patent Claims?, 2025, pp. 1272–1287, <https://doi.org/10.18653/v1/2025.findings-naacl.70>.
- [66] J.-S. Lee (Ed.), Patent Transformer: a Framework for Personalized Patent Claim Generation, *CEUR Workshop Proceedings*, 2020.
- [67] J.S. Lee, Acem. Measuring and controlling text generation by semantic search. *WWW'20: COMPANION PROCEEDINGS OF THE WEB CONFERENCE 2020*, 2020, pp. 269–273.

- [68] J.-S. Lee, J. Hsiang (Eds.), PatentTransformer-1.5: measuring patent claim generation by span relevancy, *Lect. Notes Comput. Sci.* (2020), https://doi.org/10.1007/978-3-030-58790-1_2.
- [69] J.-S. Lee, J. Hsiang (Eds.), Prior Art Search and Reranking for Generated Patent Text. *CEUR Workshop Proceedings*, 2021.
- [70] J.-S. Lee, J. Hsiang, Patent claim generation by fine-tuning OpenAI GPT-2, *World Pat. Inf.* 71 (2022).
- [71] H.H. Shomee, S.K. Maity, S. Medya, PATENTWRITER: a benchmarking study for patent drafting with LLMs, *arXiv:250722387v1* (2025), <https://doi.org/10.48550/arXiv.2507.22387>.
- [72] Visual descriptor extraction from patent figure captions: a case study of data efficiency between bilstm and transformer, in: X. Wei, J. Wu, K. Ajayi, D. Oyen (Eds.), *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 2022, <https://doi.org/10.1145/3529372.3533299>.
- [73] G.D. David Raj, S. Mukherjee, C.R.R. Robin, R.L. Jasmine, An intelligent feature concatenation process-based effective query expansion for patent retrieval approach using optimal Bi-clustering and enhanced social engineering optimizer, *Int. J. Comput. Intell. Syst.* 18 (1) (2025), <https://doi.org/10.1007/s44196-025-00963-9>.
- [74] Y. Ding, Y. Wu, Z. Ding, An automatic patent literature retrieval system based on LLM-RAG, *arXiv:250814064* (2025), <https://doi.org/10.48550/arXiv.2508.14064>.
- [75] Z.Y. Jin, Z.Y. Yang, G.B. Tang, T.C. Liu, E.D. Xun, Research on keyword-based element extraction for Chinese patent retrieval. 2024 INTERNATIONAL CONFERENCE ON ASIAN LANGUAGE PROCESSING, IALP 2024, 2024, pp. 67–73.
- [76] C.W. Kosonocky, A.L. Feller, C.O. Wilke, A.D. Ellington, Using alternative SMILES representations to identify novel functional analogues in chemical similarity vector searches, *Patterns* 4 (12) (2023), <https://doi.org/10.1016/j.patter.2023.100865>.
- [77] C.W. Kosonocky, C.O. Wilke, E.M. Marcotte, A.D. Ellington, Mining patents with large language models elucidates the chemical function landscape, *Digital Discov.* 3 (6) (2024) 1150–1159, <https://doi.org/10.1039/d4dd00011k>.
- [78] K.-Y. Lee, J. Bai, PAI-NET: retrieval-augmented generation patent network using prior art information, *Systems* 13 (4) (2025), <https://doi.org/10.3390/systems13040259>.
- [79] R. Liu, Z. Zou, S. Chen, Y. Liu, J. Wan, Harnessing AI for understanding scientific literature: innovations and applications of chat-agent system in battery recycling research, *Mater. Today Energy* 49 (2025), <https://doi.org/10.1016/j.mtener.2025.101818>.
- [80] H.C. Lo, J.M. Chu, J. Hsiang, C.C. Cho (Eds.), Large Language Model Informed Patent Image Retrieval, *CEUR Workshop Proceedings*, 2024.
- [81] P.S. Prasad, V. Trivedi, M. Rao, S. Sen (Eds.), JURISMIND: Context-Driven Retrieval for Accurate and Relevant Legal Question-Answering in Patent Filings. International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, ICK3 - Proceedings, 2025, <https://doi.org/10.5220/0013716600004000>.
- [82] K. Pustu-Iren, G. Bruns, R. Ewerth (Eds.), A Multimodal Approach for Semantic Patent Image Retrieval, *CEUR Workshop Proceedings*, 2021.
- [83] M. Shaghrouzi, B. Teimourpour, M.A. Soltanshahi (Eds.), Fine-Tuning BERT for Persian Patent Classification: a Dataset and Model Exploration, 2024, <https://doi.org/10.1109/AISP61396.2024.10475235>.
- [84] S.R. Vangala, S.R. Krishnan, N. Bung, D. Nandagopal, G. Ramasamy, S. Kumar, et al., Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature, *J. Cheminf.* 16 (1) (2024), <https://doi.org/10.1186/s13321-024-00928-8>.
- [85] J. Waterstraet, L. Walter, Designing tailored patent search approaches – a case study on nursing care technology, *World Pat. Inf.* 84 (2026), <https://doi.org/10.1016/j.wpi.2025.102420>.
- [86] Q. Xiong, Z. Xu, Z. Liu, M. Wang, Z. Chen, Y. Sun, et al. (Eds.), Enhancing the Patent Matching Capability of Large Language Models via the Memory Graph, 2025, <https://doi.org/10.1145/3726302.3729970>.
- [87] S. Xiong, S. Chen, J. He, Y. Liu, J. Mao, C. Liu, Scalable multi-label patent classification via iterative large language model-assisted active learning, *World Pat. Inf.* 82 (2025), <https://doi.org/10.1016/j.wpi.2025.102380>.
- [88] Z. Zhang, S. Liu, T. Ma, J. An, X. He, Personalized patent transaction recommendation based on enterprise profiles driven by large language model, *Inf. Stud.: Theory and Application* 48 (5) (2025) 177–186, <https://doi.org/10.16353/j.cnki.1000-7490.2025.05.021>.
- [89] W. Zheng, Q. Ding, N. Li, Y. Pan, X. Dong (Eds.), A Patent Fast Recommendation Algorithm Based on User Requirements Combined with an Inverted Table, 2024, <https://doi.org/10.1109/AIMERA59657.2024.10735767>.
- [90] Beyond scaling: predicting patent approval with domain-specific fine-grained claim dependency graph, in: X.K. Gao, F. Yao, K. Zhao, B. He, A. Kumar, V. Krishnan, et al. (Eds.), *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024, <https://doi.org/10.18653/v1/2024.acl-long.285>.
- [91] H. Ikoma, T. Mitamura, Can AI Examine Novelty of Patents?: Novelty Evaluation Based on the Correspondence Between Patent Claim and Prior Art, 2025, <https://doi.org/10.48550/arXiv.2502.06316>.
- [92] V. Knapppich, A. Friedrich, A. Häty, S. Razniewski (Eds.), PEDANTIC: a Dataset for the Automatic Examination of Definiteness in Patent Claims. *CEUR Workshop Proceedings*, 2025.
- [93] R. Lee, A. Spangher, X. Ma, PatentEdits: framing patent novelty as textual entailment, *arXiv:241113477v1* (2024), <https://doi.org/10.48550/arXiv.2411.13477>.
- [94] A. Parikh, S. Dori-Hacohen (Eds.), ClaimCompare: a Data Pipeline for Evaluation of Novelty Destroying Patent Pairs. *CEUR Workshop Proceedings*, 2024.
- [95] S. Ragot, A novel approach to measuring the scope of patent claims based on probabilities obtained from (large) language models, *World Pat. Inf.* 79 (2024), <https://doi.org/10.1016/j.wpi.2024.102321>.
- [96] V.J. Schmitt, Disentangling patent quality: using a large language model for a systematic literature review, *Scientometrics* 130 (1) (2025) 267–311, <https://doi.org/10.1007/s11192-024-05206-w>.
- [97] L. Zheng, Q. Xu, J. Ma, Y. Wang (Eds.), TCLMA: a two-dimension Contrastive Learning Based Multiagent Framework for Scientific Novelty Evaluation, 2025, <https://doi.org/10.1145/3766918.3766939>.
- [98] Z. Zhu (Ed.), The Application of Artificial Intelligence Technology to the Intellectual Property System, 2025, <https://doi.org/10.1109/DIPCA65051.2025.11042723>.
- [99] J.M. Chu, H.C. Lo, J. Hsiang, C.C. Cho, From PARIS to LE-PARIS: toward patent response automation with recommender systems and collaborative large language models, *Artif. Intell. Law* 33 (4) (2025) 955–981, <https://doi.org/10.1007/s10506-024-09409-7>.
- [100] S. Kawano, H. Nonaka, K. Yoshino (Eds.), Claimbrush: a Novel Framework for Automated Patent Claim Refinement Based on Large Language Models, 2024, <https://doi.org/10.1109/BigData62323.2024.10825172>.
- [101] N. Gavrilov, V. Korkhov, E. Pen, A. Tokarev (Eds.), Language model-based algorithm for constructing knowledge graphs from patent data, *Lect. Notes Comput. Sci.* (2026), https://doi.org/10.1007/978-3-031-97648-3_15.
- [102] J. Li, J. Zhou, M. Cao, An objective effect evaluation framework for vectorization models on patent semantic similarity measurement, *Electronics (Switzerland)* 14 (20) (2025), <https://doi.org/10.3390/electronics14204056>.
- [103] Y.X. Lin, J.H. Tang, W.J. Huang, Y.Y. Ding, J.G. Hu, Assoc computing M. Chinese named entity recognition for IC patent domain based on RoBERTa-wwm-ext, GCN and efficient global pointer. 2024 5TH INTERNATIONAL CONFERENCE ON COMPUTING, NETWORKS AND INTERNET OF THINGS, 2024, pp. 234–240. CNIOT 2024.
- [104] T. Manjunath, E. Appelmans, S. Balta, D. DiMercurio, C. Avalos, K. Stark, Topic analysis on publications and patents toward fully automated translational science benefits model impact extraction, *Front. Res. Metr. Analyt.* 10 (2025), <https://doi.org/10.3389/frma.2025.1596687>.
- [105] J.J.G. Mesa, Acm. Scalable patent analysis using function calling with local large language models. PRACTICE AND EXPERIENCE IN ADVANCED RESEARCH COMPUTING, 2025. PEARC 20252025.
- [106] E. Motamedi, I. Novalija, L. Rei, Semi-automatic hierarchical taxonomy creation from existing taxonomies with large language models, *Bus. Informat. Syst. Eng.* (2026), <https://doi.org/10.1007/s12599-025-00982-y>.
- [107] H. Nanba, K. Iwakuma, S. Fukuda (Eds.), Automatic Construction of a Patent Term Thesaurus with Fine-Tuned Chatgpt. *CEUR Workshop Proceedings*, 2025.
- [108] F.T. Nkolongo, S. Echchakoui, A. Mehdi (Eds.), Harnessing large language models for precision topic extraction and technology patent nomination: a GPT-Centric methodology, *Procedia Comput. Sci.* (2025), <https://doi.org/10.1016/j.procs.2025.03.116>.
- [109] S.L. Phyu, M. Uchkempirov, M.M. Proma, P. Kulkarni (Eds.), Augmenting Patent Summarization Using Large Language Model with Knowledge Graph. *CEUR Workshop Proceedings*, 2024.
- [110] B. Rafieian, P.-P. Vázquez, Improved multi-label hierarchical patent classification using LLMs, *World Pat. Inf.* 81 (2025), <https://doi.org/10.1016/j.wpi.2025.102356>.
- [111] J. Tang, S. Kong, Q. Xu, W. Shen, X. Yu (Eds.), Research on Patent Navigation in the Special Equipment Field Based on LLM Tools, 2025, <https://doi.org/10.1145/3772726.3772760>.
- [112] N. Yoshikawa, R. Krestel (Eds.), Improving Patent Classification Using AI-Generated Summaries. *CEUR Workshop Proceedings*, 2024.
- [113] N. Yoshikawa, R. Krestel, Do large language models understand patents? Enhancing patent classification through AI-generated summaries, *World Pat. Inf.* 81 (2025), <https://doi.org/10.1016/j.wpi.2025.102353>.
- [114] C. Yu, L. Chen, H. Xu (Eds.), Research on Named Entity Recognition from Patent Texts with Local Large Language Model. *CEUR Workshop Proceedings*, 2024.
- [115] J. Zhang, H. Ren, S. Guo, J. Sun (Eds.), Multi-Model Collaboration and prompt-driven Patent Classification Methods, 2024, <https://doi.org/10.1145/3690407.3690464>.
- [116] G. Koutsopoulos, N. Kniby, P. Wiman, I. Karina-Berzina, J. Stirna (Eds.), CLAIM: a Tool for Analyzing Legal Documents, *CEUR Workshop Proceedings*, 2025.
- [117] K. Park, H. Zo (Eds.), Exploration of Large Language Model Usage in Computer Technology Patent Texts. International Conference on ICT Convergence, 2024, <https://doi.org/10.1109/ICTC62082.2024.10827279>.
- [118] A.J.C. Trappey, C.Y. Wu, Y.C. Lee, L.P. Hung, Generative AI approach for inventive process visualisation-enhancing human-AI hybrid understanding and comparing of patents, *J. Eng. Des.* (2025), <https://doi.org/10.1080/09544828.2025.2518657>.
- [119] S. Trapp, J. Warschat (Eds.), LLM-based Extraction of Contradictions from Patents, IFIP Advances in Information and Communication Technology, 2025, https://doi.org/10.1007/978-3-031-75919-2_1.
- [120] Z. Yu, Y. Wu, N. Zhang, C. Wang, Y. Vorobeychik, C. Xiao (Eds.), CODEIPROMPT: intellectual property infringement assessment of code language models, *Proc. Mach. Learn. Res.* 202 (2023) 40373–40389, 1690.
- [121] Towards the extraction of semantic relations in design with natural language processing, in: V. Giordano, M. Consoloni, F. Chiarello, G. Fantoni (Eds.), *Proceedings of the Design Society*, 2024, <https://doi.org/10.1017/pds.2024.208>.

- [122] H. Gong, H. Zou, X. Liang, S. Meng, P. Cai, X. Xu, et al., A global dataset mapping the AI innovation from academic research to industrial patents, *Sci. Data* 12 (1) (2025), <https://doi.org/10.1038/s41597-025-05518-3>.
- [123] S. Hu (Ed.), Analysis of International Competitive Landscape for Large Language Model Technologies Based on Patent Metrics, 2023, <https://doi.org/10.1109/SCOUT58937.2023.00062>.
- [124] S. Hu, Y. Li (Eds.), International Competitive Landscape for Generative Artificial Intelligence Technology Based on Patent Metrics, *Smart Innovation, Systems and Technologies*, 2024, https://doi.org/10.1007/978-981-97-3210-4_3.
- [125] J.R. Jiang, F.L. Ying, R. Dhuny, Unveiling technological evolution with a patent-based dynamic topic modeling framework: a case study of advanced 6G technologies, *Appl Sci-Basel* 15 (7) (2025), <https://doi.org/10.3390/app15073783>.
- [126] C. Kim, J. Lee, Discovering patterns and trends in customer service technologies patents using large language model, *Heliyon* 10 (14) (2024) e34701, <https://doi.org/10.1016/j.heliyon.2024.e34701>.
- [127] M. Kim, J. Lee, What are the future trends in natural gas technology to address climate change? Patent analysis through large language model, *Energy* 312 (2024), <https://doi.org/10.1016/j.energy.2024.133644>.
- [128] L. Kriesch, Losacker S. A global patent dataset of bioeconomy-related inventions, *Sci. Data* 11 (1) (2024), <https://doi.org/10.1038/s41597-024-04163-6>.
- [129] M.Y. Lai (Ed.), How Do Patent Claims Reflect Scientific Knowledge: Insights from Human Genomics, 2022, <https://doi.org/10.23919/PICMET53225.2022.9882736>.
- [130] N. Mohammadi, J.H. Dahoie, A.A. Bengari, A. Rahimi, Mapping the technological evolution of generative AI: a patent network analysis, *Sci. Rep.* 15 (1) (2025), <https://doi.org/10.1038/s41598-025-26810-7>.
- [131] M. Pazhouhan, A. Karimi Mazraeshahi, M. Jahanbakht, K. Rezaeejad, M. H. Rohban, Wave and tidal energy: a patent landscape study, *J. Mar. Sci. Eng.* 12 (11) (2024), <https://doi.org/10.3390/jmse12111967>.
- [132] S. Pelaez, G. Verma, B. Ribeiro, P. Shapira, Large-scale text analysis using generative language models: a case study in discovering public value expressions in AI patents, *Quantit. Sci. Stud.* 5 (1) (2024) 153–169, <https://doi.org/10.1162/qss.a.00285>.
- [133] S. Petralia, Mapping general purpose technologies with patent data, *Res. Pol.* 49 (7) (2020), <https://doi.org/10.1016/j.respol.2020.104013>.
- [134] G.F. Rosa, J.O. Avelino, M.C. Cavalcanti, J.C. Duarte, TForMIX: a method that combines LLM and multidimensional modeling for technological foresight, *IEEE Access* 13 (2025) 153320–153339, <https://doi.org/10.1109/ACCESS.2025.3605116>.
- [135] Y. Shi, G. Chen, S. Li, Q. Zhou, X. Liang, H. Gao, et al. (Eds.), Research and Application of the Energy Science and Technology Intelligence Brain Integrating Large Model Technology and Patent Big Data Analysis, 2024, <https://doi.org/10.1109/ICEPG63230.2024.10775587>.
- [136] E. Sohn, J. Seo, C.H. Lee (Eds.), ANALYSIS OF GLOBAL R&D TRENDS IN NATURAL COSMETICS BASED ON PATENT BIBLIOMETRICS, 2024.
- [137] S. Urata, N. Miya, R. Hashimoto, T. Konishi, K. Fujii, Y. Murakami, et al., Leveraging LLMs to identify patent-free glass materials, *Comput. Mater. Sci.* 258 (2025), <https://doi.org/10.1016/j.commatsci.2025.114022>.
- [138] Y. Zheng, Can investors learn from patent documents? Evidence from textual analysis, *Contemp. Account. Res.* 42 (2) (2025) 1331–1358, <https://doi.org/10.1111/1911-3846.13036>.
- [139] L.V. Bui, Advancing patent law with generative AI: human-in-the-loop systems for AI-assisted drafting, prior art search, and multimodal IP protection, *World Pat. Inf.* 80 (2025), <https://doi.org/10.1016/j.wpi.2025.102341>.
- [140] N. Yadav, M.K. Kumawat, G. Ajmal, Mir K.B. AI-assisted patent drafting tools: a patent landscape & future prospectives, *World Pat. Inf.* 83 (2025), <https://doi.org/10.1016/j.wpi.2025.102402>.
- [141] L. Jiang, P.A. Scherz, S. Goetz (Eds.), Towards Better Evaluation for Generated Patent Claims. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2025.
- [142] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks. 34th Conference on Neural Information Processing Systems, *NeurIPS* 2020, 2020.
- [143] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, et al., Retrieval-augmented generation for large language models: a survey, *arXiv:2312.10997v5* (2023), <https://doi.org/10.48550/arXiv.2312.10997>.
- [144] N.F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, et al., Lost in the middle: how language models use long contexts, *Transact. Assoc. Computat. Linguist.* 12 (2024) 157–173, <https://doi.org/10.1162/tacl.a.00638>.
- [145] K. Shuster, S. Poff, M. Chen, D. Kiela, J. Weston, Retrieval augmentation reduces hallucination in conversation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 3784–3803, <https://doi.org/10.18653/v1/2021.findings-emnlp.320>.
- [146] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, H. Hajishirzi, When not to trust language models: investigating effectiveness of parametric and non-parametric memories, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics vol. 1, Long Papers, July; Toronto, Canada, 2023.
- [147] O. Press, N.A. Smith, M. Lewis, Train short, test long: attention with linear biases enables input length extrapolation, *arXiv:2108.12409* (2022), <https://doi.org/10.48550/arXiv.2108.12409>.
- [148] Y. Bai, S. Kadavath, S. Kundu, A. Askill, J. Kernion, A. Jones, et al., Constitutional AI: harmlessness from AI feedback, *arXiv:2212.08073* (2022), <https://doi.org/10.48550/arXiv.2212.08073>.
- [149] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, R. Anderson, The curse of recursion: training on generated data makes models forget, *arXiv:2305.17493* (2024), <https://doi.org/10.48550/arXiv.2305.17493>.

Pingfei Jiang is a Senior Lecturer in Design and Innovation at the Department of Innovation, Technology and Entrepreneurship, University of Exeter, where he carries out research on data-driven design, product functional modelling, sustainable design, and intellectual property analytics. He is the corresponding author of this review. Before joining Exeter, he held a lectureship in Engineering Design at Kingston University London, and earlier at Brunel University London, where he developed a long-standing research programme on design for invention and patent conflict detection. He holds a PhD in Design Engineering and an MSc in Advanced Mechanical Engineering, both from Imperial College London, and a BEng in Mechanical Engineering from the University of Sheffield. He is a Chartered Engineer (CEng), a Fellow of the Higher Education Academy (FHEA), and a member of the Design Society.

Andrea Pietro Areda holds an MSc in Artificial Intelligence and Cybersecurity. He is currently a PhD student in Intelligent Systems for Engineering at the Università degli Studi di Enna “Kore”, Italy. His research interests lie in human-centred artificial intelligence, human-computer interaction, and intelligent systems, with particular attention to the design and evaluation of AI-based systems that support effective, reliable, and meaningful interaction between humans and computing systems.

Salvatore Sorce holds an MSc and a PhD in computer engineering. He is currently an Associate Professor in Information Processing Systems in the Department of Engineering and Architecture at the Università degli Studi di Enna “Kore”, Italy.

His main research interests are in novel interfaces and novel human-computer interaction paradigms (both human-side and machine-side); pervasive displays (technologies, techniques, deployment, evaluation); ubiquitous computing and pervasive/distributed systems. Salvatore has published over 80 scientific papers, is a member and chair of several conference and journal committees, and is also a senior member of the ACM.

Mark Atherton is an Emeritus Professor of Mechanical Engineering at Brunel University of London and an accomplished expert in engineering robustness of physical prototype designs supported by simulation models. Working at the intersection of engineering vision and practical implementation has led to the development of tools that assist a designer to identify potential patent infringement early in the design process. Now researching the use of AI to extract design knowledge from patents for supporting early stage design, with a focus on extracting geometry-function interaction depicted in patent diagrams and images in order to improve searchable design image information. Authored over 120 scientific publications.