Citation-Driven Multi-View Training for Patent Embeddings: QaECTER and Sophia-Bench

Younes Djemmal, You Zuo, Kim Gerdes, Kirian Guiller

Citation-Driven Multi-View Training for Patent Embeddings: QaECTER and Sophia-Bench

Younes Djemmal[1], You Zuo[1,3], Kim Gerdes[2,4], Kirian Guiller[1]

Abstract

Patent retrieval underpins critical decisions in innovation, examination, and IP strategy, yet progress has been hampered by the absence of benchmarks that reflect the diversity of real-world search scenarios. We address this gap with two contributions. First, we introduce SOPHIA-BENCH, a large-scale patent retrieval benchmark comprising 10,000 queries and 75,000 corpus documents stratified across ten years, eight IPC technology sections, and twelve filing jurisdictions. Unlike prior benchmarks, SOPHIA-BENCH tests retrieval using 12 different query types-from structured patent fields to AI-generated summaries-and evaluates results against citation-based ground truth enhanced with a novel domain-relevance metric (InScope). Together, these enable systematic measurement of how well models perform across query types, technology domains, and jurisdictions. Second, we introduce QAECTER, a 344M-parameter embedding model trained on patent citation graphs and multi-view self-alignment. Despite its compact size, QaECTER establishes a new state of the art for patent retrieval. It outperforms the #1 model on the English retrieval text embedding benchmark (RTEB) [1], a model 23× larger, as well as all existing patent-specific models across every query type, IPC section, and jurisdiction on Sophia-bench, with gains of up to 7.2% average NDCG@10 over the next-best model. These results are confirmed on an independent external benchmark, where QAECTER surpasses all prior models without requiring task-specific instruction prompts. Both the benchmark and the model are designed for practical deployment in large-scale patent search systems.

[1]AI Lab, Questel, Paris, France
[2]Qatent, Paris, France
[3]Inria, Paris, France
[4]Université Paris-Saclay, Orsay, France
Email addresses: ydjemmal@questel.com, yzuo@questel.com, gerdes@lisn.fr, kguiller@questel.com

1 Introduction

Patent search is central to the work of inventors, examiners, and IP attorneys. Before filing an application, an inventor must determine whether the idea is novel. An examiner receiving that application must locate relevant prior art to make a proper determination. An attorney advising on freedom-to-operate needs to identify patents that might block a product launch. In all cases, the quality of the search results matters: missing a critical reference can lead to wasted R&D, invalidated patents, or costly litigation.

Patent retrieval presents fundamental challenges along four dimensions:

- **Query diversity.** Search scenarios vary substantially depending on the practitioner's role and task. An inventor submitting a full invention disclosure, a patent examiner working from the claims of a pending application, and a landscape analyst starting from a problem statement or keyword set each impose distinct retrieval demands. The increasing use of AI-generated technical summaries further broadens this range. A system that performs well on one query type may degrade significantly on others, yet operational settings require robust performance across all of them.
- **Document complexity.** Patent documents constitute a hybrid of technical disclosure and legal instrument, with heterogeneous sections that serve distinct communicative functions. Claims are drafted in deliberately broad language to maximize legal scope; descriptions interleave technical content with extensive background and enumerated embodiments. The pronounced linguistic divergence across these sections means that effective retrieval cannot rely on surface-level matching but must capture how each section relates to the underlying invention.
- **Jurisdictional variation.** Structural and linguistic properties of patent documents vary systematically across jurisdictions, reflecting differences in claim dependency conventions, fee structures, and examination procedures. Retrieval systems must account for these cross-jurisdictional patterns to maintain consistent performance across heterogeneous patent offices.
- **Temporal coverage.** The global patent corpus spans several decades and grows by millions of applications annually. Technical vocabulary shifts considerably over time, and entire domains—transformer architectures, CRISPR-based gene editing, and autonomous vehicles—were nascent or nonexistent two decades ago. A practical retrieval system must bridge this temporal gap, surfacing relevant prior art using contemporary terminology while adapting to fields whose technical language remains in flux.

Given these challenges, one would expect a well-developed ecosystem of retrieval benchmarks. In practice, existing resources either target patent-adjacent tasks rather than retrieval, or impose constraints that limit their applicability to real-world search scenarios:

- **CLEF-IP** [2]: provides a large European patent benchmark for prior-art candidate search, but it is based on an old corpus and limited to three languages, reducing its representativeness and suitability for broader cross-lingual or global evaluation.
- **TREC-Chem** [3]: restricted to chemical patents and journal articles, limiting both domain and document type coverage.
- **NTCIR** [4]: earlier editions (NTCIR-4–6) supported patent invalidity search over collections from the Japan Patent Office and, in one English subtask, the United States Patent and Trademark Office, but later editions shifted away from retrieval tasks. Consequently, invalidity

search is no longer maintained and remains limited to a small number of jurisdictions, reducing its suitability for general prior-art retrieval evaluation.
- **PatentMatch** [5]: frames prior art search as binary claim–passage classification rather than full-corpus retrieval.
- **DAPFAM** [6]: the closest work to ours, introducing cross-domain retrieval with section-level queries. However, the query set is small (1,247 families) and concentrated in 2000–2015, limiting coverage of recent technological developments.
- **MTEB** [1]: the dominant general-purpose embedding benchmark; includes legal retrieval tasks but no patent retrieval task.

No existing benchmark combines diverse query types, broad temporal and jurisdictional coverage, and domain-aware relevance labels in a single evaluation framework for patent retrieval.

We make two contributions. First, we introduce **Sophia-bench**, a large-scale patent retrieval benchmark covering the range of query types practitioners actually use: patent sections (titles, abstracts, claims, descriptions), structured extractions (object of the invention, independent claims set, advantages and disadvantages), and AI-generated summaries. The benchmark is built on citation-based ground truth spanning multiple years, technology domains, and jurisdictions, and introduces multiple testing axes (Figure 2) that enable measurement of both average performance and robustness across query types, IPC sections, and filing jurisdictions. Second, we introduce **QaECTER**, a 344M-parameter embedding model built specifically for patent retrieval. Despite its compact size, it outperforms state-of-the-art general-purpose embedding models up to 23× larger, as well as existing patent-specific models, across query types, domains, and time periods on both Sophia-bench and external benchmarks, establishing a new state of the art for patent search.

2 Introducing Qaecter

2.1 Overview

QaECTER is a 344M-parameter embedding model trained specifically for patent retrieval. The core training paradigm jointly optimizes two objectives: prior-art search and alignment of different *patent views*—distinct textual representations of the same patent, such as the abstract, claims, description, or structured extractions like the object of the invention.

The model is trained on a large-scale dataset drawn from Questel's Orbit database[1], one of the largest commercially available patent databases in the world. Training data spans patent applications published from 2002 to 2024, aggregated at the patent family level, and includes both original English text and translated English versions, making the model robust to linguistic and translation variation across jurisdictions.

Training pairs are derived from examiner citation relationships (X, Y and A categories), which provide complementary signals of semantic relevence at different levels of specificity. In addition, a proportion of training data consist of *self-alignment pairs*: two different views of the same patent, ensuring that different textual facets of the same invention map to nearby points in the embedding space. This multi-view alignment is a key design choice, continuously exposing the model to heterogeneous representations of the same underlying inventive concept.

[1]Questel, *Orbit Intelligence – Patent Search & Analytics Software*, 2026. https://www.questel.com/patent/ip-intelligence-software/orbit-intelligence/

The evaluation benchmark (SOPHIA-BENCH, Section 3) was drawn from a 10-year window (2016–2025) with no overlap with the training set.

2.2 Architecture

QaECTER builds on bert-for-patents [7], a BERT-Large model pre-trained via masked language modeling on a large-scale patent corpus. The tokenizer vocabulary is extended with special section tokens that delimit patent views during training, helping the model distinguish between section types; these tokens are not required at inference.

A projection head is used during training and discarded at inference: the final model outputs 1024-dimensional L_2-normalized embeddings via mean pooling over token representations.

We adopt a contrastive learning framework based on the InfoNCE objective with in-batch negatives and large effective batch sizes. Training proceeds for a single epoch with early stopping based on retrieval metrics.

3 SOPHIA-BENCH

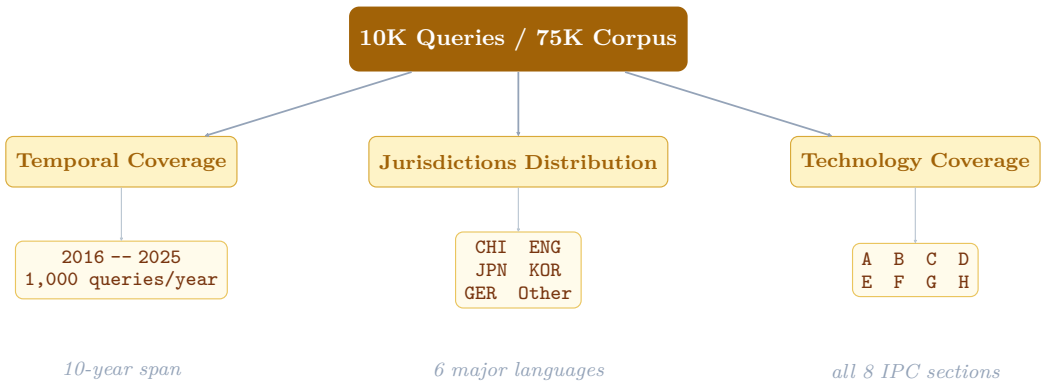

Fig. 1. Corpus design in SOPHIA-BENCH, comprising 10,000 patent queries and a 75,000-document corpus stratified across three dimensions: temporal coverage (2016–2025, with 1,000 queries per year), jurisdiction distribution (Chinese, English, Japanese, Korean, German, and other jurisdictions), and technology coverage (all eight IPC sections A–H).

SOPHIA-BENCH is a large-scale benchmark for evaluating patent embedding models on prior-art retrieval. Unlike existing benchmarks, SOPHIA-BENCH systematically varies the query text representation across 12 types while keeping the corpus fixed, enabling a controlled study of how different textual facets of a patent affect retrieval quality.

3.1 Dataset Construction

The benchmark is built from a corpus of 75,050 patent documents and a query set of 10,000 patents, both drawn from Questel's international patent database. All texts are in English: for patents originally filed in other languages, we use the English-language family representative (original or

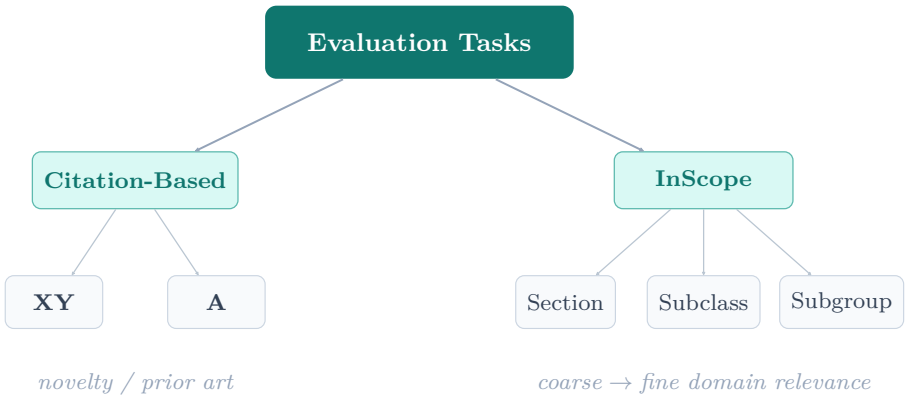

Fig. 2. Evaluation tasks in SOPHIA-BENCH: citation-based retrieval evaluates novelty and prior-art search using XY (cited by examiner) and A (cited by applicant) citations, while InScope tasks measure domain relevance at three IPC granularities from coarse Section-level to fine Subgroup-level classification.

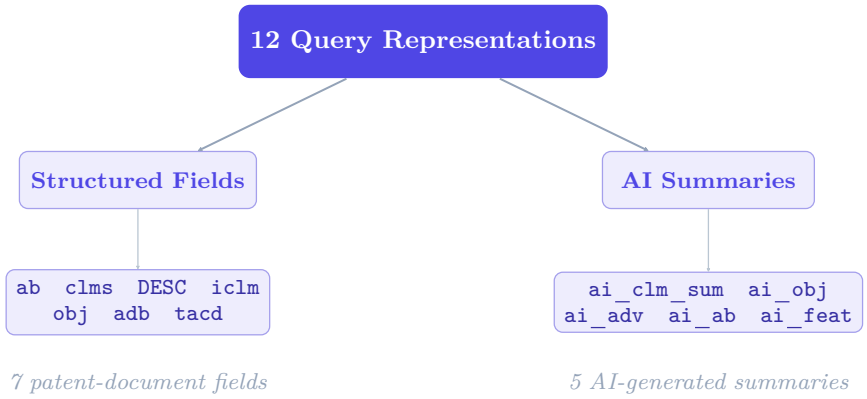

Fig. 3. The twelve query representations in SOPHIA-BENCH, comprising seven structured patent-document fields (abstract, claims, description, etc.) and five AI-generated textual summaries (claim summary, objective, advantages, abstract summary, and features).

translated). Each corpus document is represented by the concatenation of its title, abstract, claims, and description (`tacd`).

The query set contains 1,000 patents per year from 2016 to 2025. The set spans all eight IPC sections (A–H), covering technology domains from physics and electricity to chemistry, mechanical engineering, and human necessities. In terms of original filing jurisdiction, the queries reflect the global distribution of patent activity across 12 jurisdictions, including Chinese, English-speaking, Japanese, Korean and European patent offices. This diversity in both domain and jurisdiction enables evaluation of model robustness across patents originating from different technological fields and jurisdictional conventions.

3.2 Ground Truth

Relevance judgments are derived from patent citation graphs. For each query patent q, the relevant set $\mathcal{R}_q$ is constructed from four citation relationships:

- **Cited-XY**: patents cited by q as X or Y references (novelty or inventive step);
- **Citing-XY**: patents that cite q as X or Y references;
- **Cited-A**: patents cited by q as A references (general technological background);
- **Citing-A**: patents that cite q as A references.

The union of all four sets constitutes the full relevance set $\mathcal{R}_q$. On average, each query has 6.6 relevant documents (median 5, range 2–90). All 10,000 queries have at least one XY-relevant document; 9,196 have at least one A-relevant document. These two citation categories represent different levels of semantic similarity and are evaluated separately (Section 4.4).

Each document is also annotated with its International Patent Classification (IPC) codes, enabling domain-level analysis of retrieval performance. The IPC taxonomy is hierarchical: section (one of eight broad domains, e.g., G = Physics), subclass (e.g., G06K = Recognition and presentation of data), and subgroup (e.g., G06Q-010/026).

3.3 Query Representations

A central axis of Sophia-bench is the systematic variation of query text representation. While the corpus encoding is fixed (`tacd`), we evaluate 12 query text types grouped into two families.

3.30a Structured patent fields.: These correspond to standard patent sections or fields extracted via regular expressions:

- `ab`: abstract;
- `clms`: full claims;
- `DESC`: full description;
- `iclm`: independent claims only;
- `obj`: extracted object of the invention;
- `adb`: extracted advantages and disadvantages;
- `tacd`: title + abstract + claims + description (same representation as the corpus).

3.30b AI-generated summaries.: These are produced by a large language model applied to the full patent text, providing concise, semantically rich reformulations:

- `ai_claim_summary`: generated claims summary;
- `ai_obj`: generated object of the invention;
- `ai_adv`: generated advantages;
- `ai_ab`: generated abstract;
- `ai_features`: generated technical features of the invention.

This design isolates the effect of query representation: by fixing the corpus encoding and varying only the query side, we can directly measure how much information each text type carries for retrieval, and how robust each model is across the query types.

4 Evaluation Protocol

4.1 Models

We evaluate seven embedding models spanning patent-specific and general-purpose architectures. Table 1 summarizes their properties.

TABLE 1
Embedding models evaluated on Sophia-bench. All models produce L_2-normalized embeddings and are evaluated with a maximum sequence length of 512 tokens.

Model	Domain	Dim.	Params
Octen-Embed-8B	General	4096	8B
Octen-Embed-4B	General	2560	4B
PaECTER	Patent	1024	344M
bert-for-patents	Patent	1024	344M
patembed-large	Patent	1024	344M
modernbert-embed-base	General	768	100M
QaECTER (ours)	Patent	1024	344M

Octen-Embed-4B and **Octen-Embed-8B** [8] are large-scale general-purpose embedding models based on Qwen3 [9] decoder architectures, fine-tuned with LoRA for embedding tasks. **Octen-Embed-8B** ranks #1 on the Hugging Face RTEB Leaderboard [1] and the 4B variant ranks #2 (as of January 12, 2026).

PAECTER [10] is a patent-domain embedding model fine-tuned with a contrastive objective on patent citation pairs, built on BERT-Large for Patents.

BERT-for-Patents [7] is a BERT-Large model pre-trained from scratch on patent text via masked language modeling; we wrap it with mean pooling for embedding extraction.

PatEmbed-Large [11] is an instruction-tuned patent embedding model built on BERT-for-Patents; following the authors' recommendations, we prepend the specified query and document prefixes for the task (retrieval mixed) for fair comparison.

ModernBERT-Embed-Base [12] is a recent general-purpose embedding model based on a modernized BERT architecture.

QaECTER is our proposed model (Section 2).

4.2 Retrieval Setup

For each model, the entire corpus of 75,050 patent documents is embedded once using the `tacd` text representation, producing a single set of corpus vectors. Each query patent, however, can be represented using our 12 different text types, each yielding a distinct query embedding.

To measure how relevant each corpus document is to a given query, we use cosine similarity, a standard measure of how close two vectors are in the vector space. The score ranges from -1 (completely unrelated) to 1 (a perfect match). The higher the score, the more semantically similar the two patents are according to the model.

For each query, all 75,050 corpus documents are scored and ranked in descending order of similarity, producing one complete ranking per query. Since we evaluate 7 models, each with 12 query text types, this yields $7 \times 12 = 84$ distinct retrieval configurations, allowing us to systematically compare how the choice of model and query representation affects retrieval quality.

All embeddings are computed with a maximum sequence length of 512 tokens. Texts exceeding this limit are truncated by the tokenizer. No special tokens or instruction prefixes are used except for the model PatEmbed-Large, which requires task-specific prefixes as specified by its authors.

4.3 Metrics

We employ two families of metrics capturing complementary aspects of retrieval quality.

4.31 Citation-Based Retrieval Metrics: The objective here is to test how well the models can find the examiner's citations for each patent with the different types explained in (Section 3.2).

We use standard information retrieval metrics: Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), Recall@k, and NDCG@k computed against the citation-based ground truth at different cutoffs $k \in \{1, 5, 10, 20, 50, 100, 200, 500, 1\text{K}\}$.

4.32 Domain Relevance: InScope: Citation-based metrics only credit retrieval of documents that appear in the citation graph. A model may retrieve semantically relevant patents that are not among the known citations available in the ground truth search reports. Such results are penalized by citation metrics but may still be relevant and useful to a practitioner. To capture this, we introduce the **InScope** metric, which measures the proportion of top-k retrieved documents that share at least one IPC code with the query, truncated to a given granularity level g:

$$\text{InScope}_g@k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{|\{d \in D_k : \text{IPC}_g(d) \cap \text{IPC}_g(q) \neq \emptyset\}|}{k} \tag{1}$$

where D_k is the set of top-k retrieved documents and $\text{IPC}_g(d)$ denotes the IPC codes of document d truncated to granularity g.

In other words, InScope answers: *among the top-ranked results, what fraction belongs to the same technological domain as the query patent?* We evaluate this at three levels of IPC granularity, from broad to narrow:

- **Section**: one of eight broad domains (e.g., `G` = Physics);
- **Subclass**: a specific technology subclass (e.g., `G06Q`);
- **Subgroup**: the most specific IPC level (e.g., `G06Q-010/026`).

InScope is computed at $k \in \{1, 5, 10, 20, 50, 100\}$. Higher values indicate that the model's top results are more topically concentrated in the correct domain.

4.4 Evaluation Dimensions

All citation-based metrics are computed along four complementary axes:

1) **Overall**: all queries, all relevant documents.
2) **By citation category**: separate evaluation on XY citations (novelty/inventive step) and A citations (general background). These categories represent fundamentally different levels of semantic relatedness: XY references are closely related to the inventive concept, while A references provide broader technological context.
3) **By IPC section**: per-domain evaluation across the eight IPC sections (A–H), revealing whether models exhibit domain-specific strengths or weaknesses.
4) **By original jurisdiction**: per-jurisdiction evaluation across the 12 represented filing origins, assessing model robustness on patents originating from different patent office conventions and translation pipelines.

TABLE 2
Citation-based retrieval on Sophia-bench. Models are sorted by average NDCG@10. Best result per column is in **bold**.

		XY			A			All		
Query	**Model**	R@10	R@100	NDCG	R@10	R@100	NDCG	R@10	R@100	NDCG
`tacd`	QaECTER (ours)	**.582**	**.873**	**.473**	**.486**	**.831**	**.371**	**.529**	**.849**	**.538**
	Octen-Embed-8B	.536	.836	.436	.452	.798	.350	.491	.813	.502
	Octen-Embed-4B	.527	.833	.428	.445	.797	.344	.482	.811	.493
	PaECTER	.528	.834	.436	.428	.781	.334	.473	.802	.491
	patembed-large	.519	.824	.423	.432	.779	.334	.471	.797	.483
	modernbert-embed-base	.383	.684	.315	.324	.639	.258	.352	.659	.368
	bert-for-patents	.279	.521	.234	.208	.452	.172	.240	.481	.261
`ai_feat.`	QaECTER (ours)	**.577**	**.875**	**.472**	**.482**	**.833**	**.366**	**.527**	**.851**	**.535**
	Octen-Embed-8B	.546	.844	.445	.451	.804	.346	.494	.821	.505
	Octen-Embed-4B	.532	.838	.434	.441	.798	.338	.482	.814	.493
	PaECTER	.511	.825	.424	.417	.772	.322	.461	.795	.477
	patembed-large	.507	.822	.411	.424	.780	.323	.461	.797	.469
	modernbert-embed-base	.367	.682	.298	.309	.634	.241	.337	.657	.347
	bert-for-patents	.218	.475	.179	.171	.421	.136	.194	.448	.203

5 Results and Discussion

QaECTER establishes a new state of the art for patent retrieval, surpassing all existing patent-specific and general-purpose embedding models known to us. We demonstrate this through extensive evaluation on Sophia-bench, covering 12 query types, 7 models, and multiple evaluation axes, as well as on an independent external benchmark. The following subsections present and discuss these results in detail.

5.1 Citation-Based Retrieval

We begin with citation-based retrieval, selecting the best-performing query type from each group of queries based on average NDCG@10: `tacd` for structured patent fields and `ai_features` for AI-generated queries. Table 2 reports Recall@10, Recall@100, and NDCG@10, broken down by citation category: XY (novelty/inventive step), A (general background), and all citations combined.

QaECTER achieves the highest scores across all metrics, citation categories, and both query types. On the `tacd` query, it reaches an NDCG@10 of 0.538, outperforming the next-best model (Octen-Embed-8B, 0.502) by 3.6%. This gap is notable given that Octen-Embed-8B has 8 billion parameters-23× more than QaECTER's 344 million. QaECTER outperforms existing patent embeddings models PaECTER and patembed-large by an even larger margin of 4.7% and 5.5% respectively.

Across all models, XY citations are consistently easier to retrieve than A citations, with higher recall and ranking scores. This is expected: XY references are closely tied to the query patent's core inventive concept, making them more semantically similar, whereas A references provide broader technological context, making them harder to distinguish from other patents in the same domain.

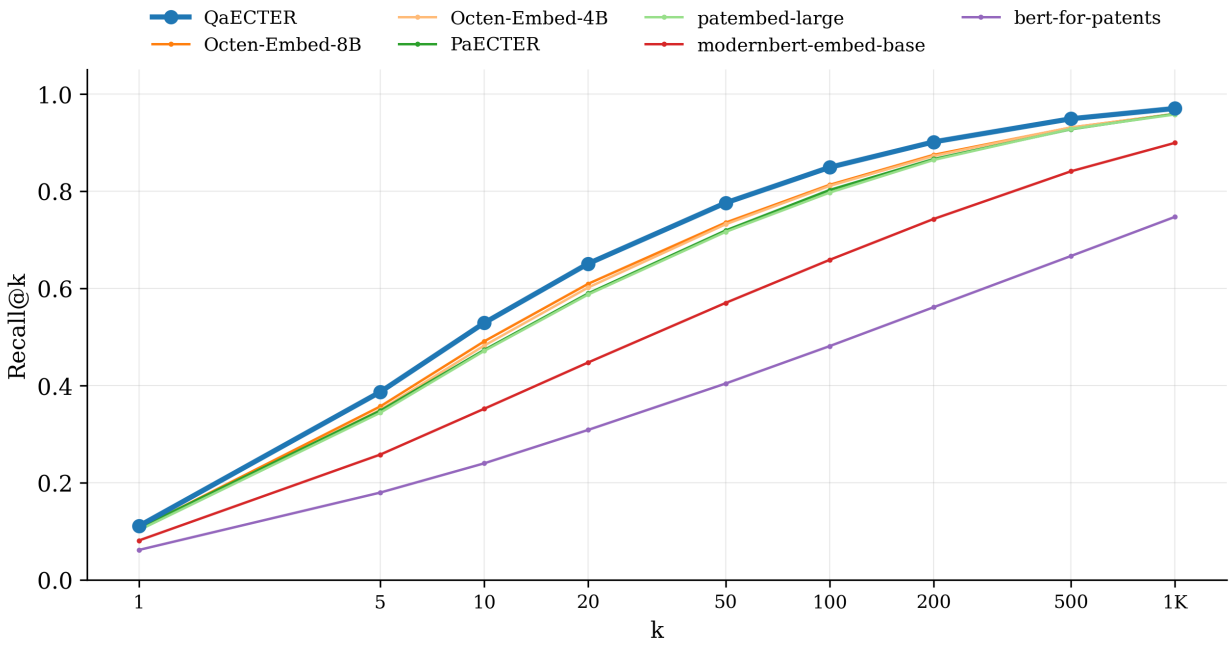

Fig. 4. Recall@k curves on SOPHIA-BENCH for the `tacd` query type (overall). QaECTER (blue) leads at every cutoff from $k = 1$ to $k = 1{,}000$.

QaECTER maintains the lead on both citation types across all general-purpose and patent embedding models.

The results are consistent across all query types. `tacd`, which uses the full patent text as query, achieves marginally higher scores than `ai_features`, but the model ranking is identical. This confirms that the generalization of QaECTER to different query types, and more importantly, it showcases its superior search capabilities to other models when searching using the technical features of an invention.

Figure 4 shows the full Recall@k curves for the `tacd` query type up to $k = 1{,}000$. QaECTER maintains a consistent lead at every cutoff, with the gap widening in the low-k regime where precision matters most for practitioners. At $k = 10$, QaECTER retrieves 53% of all relevant documents, compared to 49% for Octen-Embed-8B and 47% for PaECTER and patembed-large, a gap that is expected to grow larger when searching in larger production grade databases of patents.

5.2 Robustness Across Query Types

Table 3 reports NDCG@10 for all 12 query types, providing a comprehensive view of how each model handles different textual representations of the same underlying inventions.

QaECTER ranks first on all 12 query types, with an average NDCG@10 of 0.502. The margin over the second-best model (Octen-Embed-8B, 0.468) is consistent across query families, indicating a systematic advantage rather than one driven by specific text types. We can particularly see the superiority of QaECTER over the patent specific embeddings models here by looking at the low

TABLE 3
NDCG@10 across all 12 query types on Sophia-bench (overall). Best result per row in **bold**. Query types are grouped into structured patent fields (top) and AI-generated summaries (bottom).

Query type	QaECTER	Octen-Embed-8B	Octen-Embed-4B	PaECTER	patembed-large	modernbert-embed-base	bert-for-patents
`tacd`	**.538**	.502	.493	.491	.483	.368	.261
`clms`	**.524**	.490	.476	.469	.465	.335	.194
`ab`	**.517**	.480	.471	.474	.459	.358	.234
`iclm`	**.517**	.482	.471	.455	.461	.344	.207
`obj`	**.474**	.441	.431	.419	.424	.334	.185
`DESC`	**.481**	.446	.421	.403	.412	.343	.092
`adb`	**.383**	.348	.330	.322	.333	.247	.103
`ai_ab`	**.539**	.502	.493	.480	.472	.393	.139
`ai_feat`	**.535**	.505	.493	.477	.469	.347	.203
`ai_clm_sum`	**.537**	.501	.492	.480	.476	.325	.197
`ai_adv`	**.496**	.465	.453	.425	.433	.317	.078
`ai_obj`	**.482**	.456	.440	.418	.420	.351	.093
Average	**.502**	.468	.455	.443	.442	.339	.166
Max – Min	.156	.157	.163	.169	.150	.146	.183

generalization capabilities of PaECTER and patembed-large that suffer from a high variability in performance across query types.

The performance spread across query types (max minus min NDCG@10) is similar for the top models (0.15–0.17), with `adb` (advantages and disadvantages) being the most challenging and `ai_ab` or `tacd` the easiest. The difficulty of `adb` likely stems from its shorter length and narrower semantic scope compared to full-text representations.

AI-generated summaries perform comparably to their structured counterparts. For instance, `ai_ab` (0.539) slightly outperforms `ab` (0.517) for QaECTER, suggesting that the LLM reformulation adds useful semantic signal. Similarly, `ai_claim_summary` (0.537) outperforms `clms` (0.524). This finding validates the use of AI-generated queries in patent search workflows.

bert-for-patents exhibits the largest performance spread (0.183) and near-zero scores on `DESC` and `ai_adv`, confirming that without task-specific fine-tuning, general patent language modeling does not transfer effectively to retrieval.

5.3 Jurisdiction and Temporal generalisation

Figure 5 shows NDCG@10 averaged across all 12 query types, broken down by filing jurisdiction. We report individual results for the five jurisdictions with at least 100 queries in the benchmark. The remaining eight jurisdictions: French, Italian, Spanish, Dutch, Portuguese, Norwegian, and Czech, have fewer than 100 queries each and are grouped under ROW (Rest of World).

QaECTER achieves the highest NDCG@10 across all six jurisdiction groups. Its advantage holds consistently whether evaluated on Chinese, Japanese, and Korean patents, English-origin filings, or

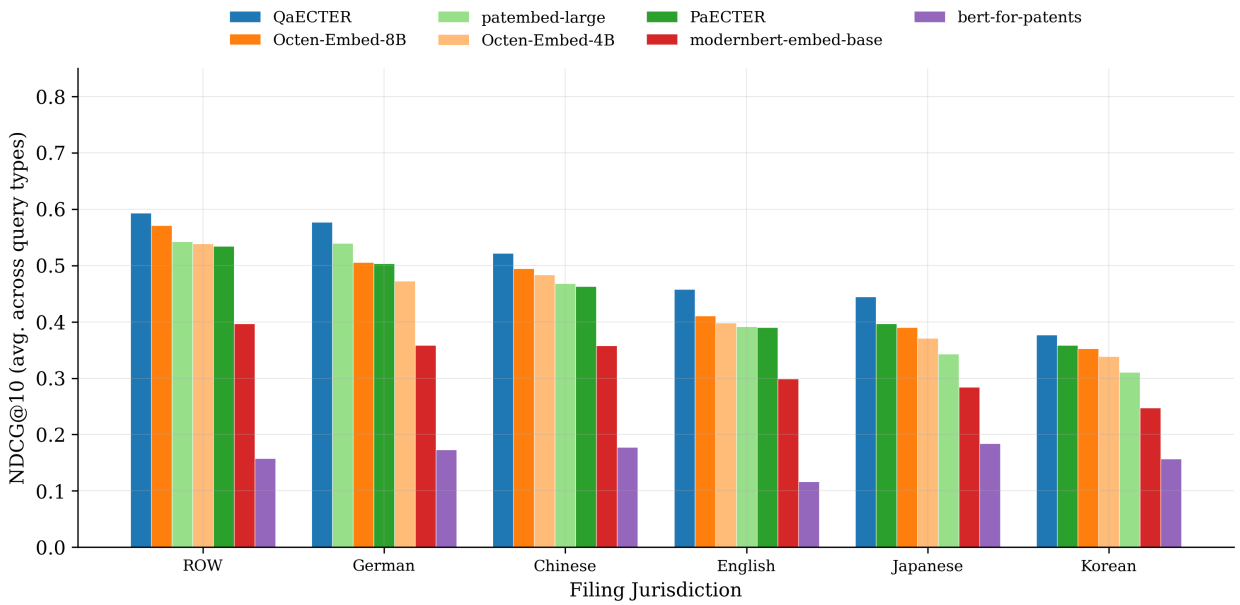

Fig. 5. NDCG@10 by filing jurisdiction on SOPHIA-BENCH, averaged across all 12 query types. Jurisdictions are sorted by QaECTER performance (descending). Four jurisdictions with fewer than five queries are excluded.

European jurisdictions such as Germany and ROW. Moreover, QaECTER exhibits lower performance variation across jurisdictions than competing models: while others show steeper declines as retrieval difficulty increases, QaECTER maintains a stable margin throughout. This suggests that its cross-jurisdictional robustness is a direct consequence of its training design rather than an artifact of strong results on any single patent corpus.

Performance remains stable across publication years for all models. Notably, QaECTER maintains a consistent lead throughout, and its advantage over competing models widens on the most recent patents (2024–2025). While the full benchmark was excluded from training, the training corpus only covers publications from 2002 to 2024, meaning 2025 patents are entirely unseen, yet our model achieves its strongest relative performance on this subset (Figure 6).

5.4 Robustness Across Technology Domains

Figure 7 reports NDCG@10 by IPC section, averaged across query types.

QaECTER ranks first in all eight IPC sections, delivering the best performance across the full range of patent technology domains. Its results remain consistently strong regardless of domain size or technical focus, demonstrating stable and reliable retrieval quality throughout the benchmark. This uniform leadership confirms that QaECTER's advantage is broad and systematic, rather than concentrated in specific or easier areas, establishing it as a robust general-purpose patent retrieval model.

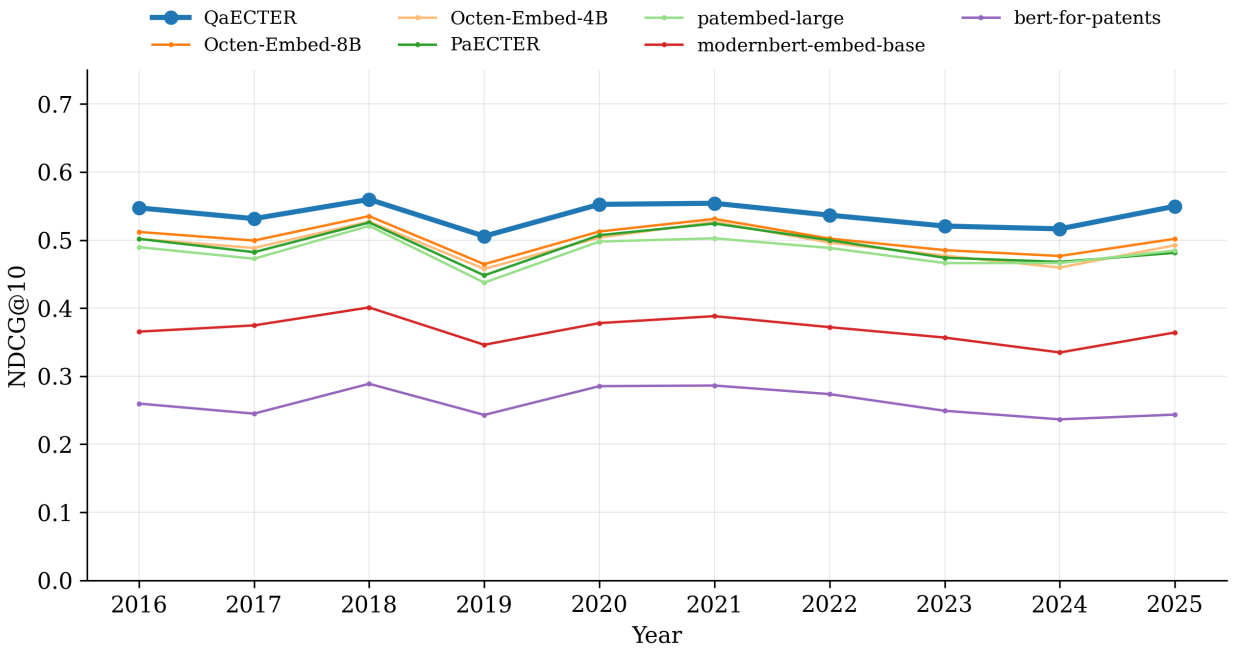

Fig. 6. NDCG@10 by publication year on SOPHIA-BENCH for the TACD query type. All models show stable performance across publication years, with no significant temporal degradation.

5.5 Domain Relevance (InScope)

Table 4 reports InScope at $k \in \{10, 100\}$ across three IPC granularity levels: section (8 broad domains), subclass (4-character codes), and subgroup (full IPC codes).

QaECTER achieves the highest InScope@10 at all three granularity levels, indicating that its top-ranked results are the most topically relevant. At the section level, over 91% of the top-10 retrieved documents share the query's broad technology domain.

At @100, patembed-large slightly edges ahead (by less than 1 percentage point), but QaECTER leads at @10, indicating stronger precision in the top ranks. This is notable because QaECTER was never trained on classification tasks, yet it matches and surpasses patembed-large, which explicitly included IPC classification during training, highlighting QaECTER's strong generalization beyond its training objectives.

Performance drops substantially from section to subgroup granularity: InScope@10 falls from approximately 0.91 to 0.49. This steep decline reflects the inherent difficulty of fine-grained IPC matching as retrieving documents that share the same specific subgroup code requires capturing highly detailed technical distinctions that are challenging even for specialized models. In practice, combining our model with an IPC-based pre-filter can fully address this limitation.

5.6 External Validation: DAPFAM

To assess generalization beyond SOPHIA-BENCH, we evaluate QaECTER on DAPFAM [11], an independent patent retrieval benchmark introduced alongside the PatEmbed model family. DAP-

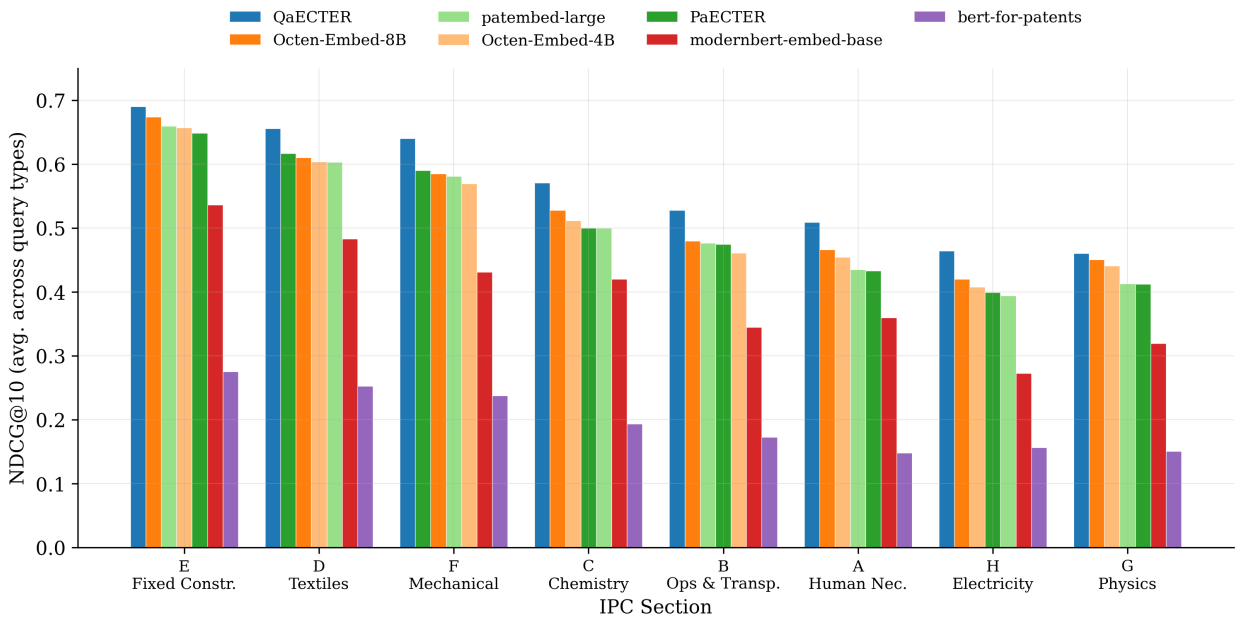

Fig. 7. NDCG@10 by IPC section on SOPHIA-BENCH, averaged across all 12 query types. Sections are sorted by QaECTER performance (descending).

FAM evaluates cross-view retrieval across six query–corpus configurations, combining title+abstract and title+abstract+claims as queries against three corpus representations. We report the average NDCG@100 across all six DAPFAM.ALL subtasks.

Results are outlined in Table 5. QaECTER achieves 0.379 NDCG@100 without any instruction prefix, surpassing every model in the DAPFAM comparison, including patembed-large's best score of 0.377, which requires task-specific prompt prefixes. The result is particularly striking for patembed-large, whose performance collapses to 0.044 without its prompt, revealing a strong dependency on instruction tuning. This strong dependency makes it impossible to use a single corpus embedding across tasks, which is a real burden in large scale production settings.

In contrast, QaECTER's prompt-free design and strong generalization capabalities makes it directly usable in production systems without per-task prompt engineering. Another important observation here is that on our independent benchmark sophia-bench, PaECTER has nearly identical retrieval results to patembed-large, surpassing it by a thin margin in most tasks, while on DAPFAM patembed-large scores higher than PaECTER, although the tasks in DAPFAM are not structurally different from the tasks of sophia-bench (both tasks are citation-based patent retrieval). This suggests the possibility of overfitting from patembed-large on their DAPFAM test data.

These results confirm that QaECTER's advantage is not specific to SOPHIA-BENCH but extends to independently constructed benchmarks with different ground truth and evaluation protocols establishing itself as the new state of the art patent embeddings model.

TABLE 4
InScope analysis on Sophia-bench. Best result per column is in **bold**.

		Section		Subclass		Subgroup	
Query	**Model**	@10	@100	@10	@100	@10	@100
`tacd`	QaECTER (ours)	**.915**	.844	**.784**	.628	**.490**	.297
	Octen-Embed-8B	.901	.818	.762	.592	.472	.279
	Octen-Embed-4B	.904	.828	.765	.606	.474	.287
	PaECTER	.902	.826	.760	.597	.465	.277
	patembed-large	.914	**.851**	.781	**.636**	.484	**.299**
	modernbert-embed-base	.889	.813	.730	.567	.430	.250
	bert-for-patents	.834	.744	.614	.454	.330	.187
`ai_feat.`	QaECTER (ours)	**.913**	.842	**.781**	.627	**.487**	**.297**
	Octen-Embed-8B	.901	.815	.758	.590	.468	.278
	Octen-Embed-4B	.901	.822	.759	.599	.467	.282
	PaECTER	.899	.823	.751	.592	.455	.274
	patembed-large	.911	**.848**	.776	**.633**	.474	.296
	modernbert-embed-base	.885	.811	.720	.566	.414	.246
	bert-for-patents	.824	.750	.599	.462	.303	.187

TABLE 5
DAPFAM.ALL benchmark (NDCG@100). Results for models other than QaECTER are from [11].

Model	**No Prompt**	**Best**
QaECTER (ours)	**.379**	**.379**
patembed-large	.044	.377
PatEmbed-Base	.352	.370
PaECTER	.343	.343
BGE-PatentMatch	.314	.314
bert-for-patents	.228	.228

6 Conclusion

We presented two complementary contributions to patent retrieval. Sophia-bench provides the first large-scale benchmark that systematically evaluates embedding models across 12 query representations, eight technology domains, multiple jurisdictions, and a decade of patent filings. By combining citation-based ground truth with the InScope domain-relevance metric, it enables fine-grained diagnosis of model strengths and failure modes that prior benchmarks could not capture.

QaECTER, our 344M-parameter embedding model, establishes a new state of the art on both Sophia-bench and the independent DAPFAM benchmark. Its consistent superiority over models up to 23× larger demonstrates that domain-specific training on patent citation graphs with multi-view self-alignment can be far more effective than scaling general-purpose architectures. Critically, QaECTER achieves this without instruction prompts or task-specific prefixes, making it directly

deployable in production search pipelines.

Our evaluation also yields several broader findings. AI-generated summaries match or exceed traditional patent fields as queries, validating their use in modern search workflows. The InScope analysis reveals that even the best models struggle with fine-grained IPC subgroup matching, pointing to an important direction for future work.

References

[1] Hugging Face, "MTEB Leaderboard," 2026. [Online]. Available: https://huggingface.co/spaces/mteb/leaderboard

[2] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz, "CLEF-IP 2011: Retrieval in the Intellectual Property Domain," in *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[3] M. Lupu, J. Huang, J. Zhu, and J. Tait, "TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at TREC," *ACM SIGIR Forum*, vol. 43, no. 2, 2009.

[4] A. Fujii, M. Iwayama, and N. Kando, "Overview of the Patent Retrieval Task at the NTCIR-6 Workshop," in *Proceedings of the NTCIR-6 Workshop*, 2007.

[5] J. Risch, N. Alder, C. Hewel, and R. Krestel, "PatentMatch: A Dataset for Matching Patent Claims & Prior Art," *arXiv preprint arXiv:2012.13919*, 2020.

[6] I. Ayaou, D. Cavallucci, and H. Chibane, "DAPFAM: A Domain-Aware Family-level Dataset to Benchmark Cross-Domain Patent Retrieval," *arXiv preprint arXiv:2506.22141*, 2025.

[7] Hugging Face, "anferico/bert-for-patents," 2022. [Online]. Available: https://huggingface.co/anferico/bert-for-patents

[8] Octen Team, "Octen Embedding Models," 2026. [Online]. Available: https://huggingface.co/Octen

[9] A. Yang *et al.*, "Qwen3 Technical Report," *arXiv preprint arXiv:2505.09388*, 2025.

[10] M. Ghosh, M. E. Rose, S. Erhardt, E. Buunk, and D. Harhoff, "PaECTER: Patent-level Representation Learning using Citation-informed Transformers," *arXiv preprint arXiv:2402.19411*, 2024.

[11] I. Ayaou and D. Cavallucci, "PatenTEB: A Comprehensive Benchmark and Model Family for Patent Text Embedding," *arXiv preprint arXiv:2510.22264*, 2025.

[12] Nomic AI, "nomic-ai/modernbert-embed-base," 2026. [Online]. Available: https://huggingface.co/nomic-ai/modernbert-embed-base