Patent Representation Learning via Self-supervision

You Zuo, Kim Gerdes, Eric Villemonte de La Clergerie, Benoît Sagot

Patent Representation Learning via Self-supervision

You Zuo[1,2] **Kim Gerdes**[2,3] **Éric de la Clergerie**[1] **Benoît Sagot**[1]
[1]Inria, Paris, France
[2]Qatent, Paris, France
[3]Université Paris-Saclay (LISN, CNRS), Orsay, France
{you.zuo, eric.de_la_clergerie,benoit.sagot}@inria.fr
gerdes@lisn.fr

Abstract

This paper presents a simple yet effective contrastive learning framework for learning patent embeddings by leveraging multiple views from within the same document. We first identify a patent-specific failure mode of SimCSE-style dropout augmentation: it produces overly uniform embeddings that lose semantic cohesion. To remedy this, we propose section-based augmentation, where different sections of a patent (e.g., abstract, claims, background) serve as complementary views. This design introduces natural semantic and structural diversity, mitigating over-dispersion and yielding embeddings that better preserve both global structure and local continuity. On large-scale benchmarks, our fully self-supervised method matches or surpasses citation- and IPC-supervised baselines in prior-art retrieval and classification, while avoiding reliance on brittle or incomplete annotations. Our analysis further shows that different sections specialize for different tasks—claims and summaries benefit retrieval, while background sections aid classification—highlighting the value of patents' inherent discourse structure for representation learning. These results highlight the value of exploiting intra-document views for scalable and generalizable patent understanding.

1 Introduction

Patents are a vital source of technical knowledge, with hundreds of thousands of new filings each year. The USPTO[1] alone processed over 430k applications in 2024, with a backlog exceeding 800k documents (Ellis, 2025). This makes tools for partial automation essential, for examination itself as well as for leveraging this vast database of human inventiveness.

To support patent-related tasks such as prior-art search, classification, or trend analysis, recent efforts have focused on *patent representation learning*—encoding a document's content into a dense vector using transformer models. Supervised methods, fine-tuned on IPC[2] codes (Lee and Hsiang, 2020; Bekamiri et al., 2024) or citation graphs (Vowinckel and Hähnke, 2023; Ghosh et al., 2024), achieve strong performance in classification and retrieval. However, they rely on curated labels that are often sparse, biased, or unstable. IPC annotations suffer from class imbalance and taxonomic shifts (van Hoewijk and Holmström, 2022; Haščič et al., 2015), while citation links may be incomplete or strategically manipulated (Blume et al., 2024).

Self-supervised contrastive learning offers a way to sidestep these issues by deriving training signals directly from the data itself (Chen et al., 2020; Gao et al., 2021). Yet in the patent domain, naïve augmentations such as dropout, as used in SimCSE, are insufficient: dropout performs only *feature-level* perturbations that leave lexical, syntactic, and discourse content unchanged, yielding views that are nearly identical—especially problematic for long, complex documents. As we will demonstrate later with embedding diagnostics, this insufficiency manifests as *over-dispersion of embeddings*—an instance of the curse-of-dimensionality effect—where representations spread across higher dimensions but lose semantic cohesion. We hypothesize that patents' long, heterogeneous discourse, unlike short sentences, calls for augmentations that introduce genuine *discourse- or structural-level* diversity. Our experiments confirm that such augmentations mitigate over-dispersion.

In our work, we propose a fully self-supervised framework that leverages the inherent structure of patents to construct diverse, semantically meaningful data augmentations. Different sections of a patent (e.g., abstract, claims, background) describe the same invention with distinct styles and levels of

[1]United States Patent and Trademark Office (USPTO): https://www.uspto.gov

[2]International Patent Classification (IPC): https://www.wipo.int/en/web/classification-ipc

detail, naturally serving as complementary views. This section-based augmentation mitigates over-dispersion and yields embeddings that maintain semantic cohesion while making fuller and more balanced use of the representation space.

Our contributions are threefold: (1) Diagnose a patent-specific failure mode—dropout-only augmentation causes *over-dispersion of embeddings*, quantified via alignment, uniformity, and singular spectrum divergence; (2) Propose section-based augmentation that mitigates this issue and improves retrieval/classification; (3) Release code, checkpoints, and evaluation queries for reproducibility[3].

2 Related Work

Patent representation learning. Recent work on patent representation has focused on adapting transformer models to the patent domain. PatentBERT (Lee and Hsiang, 2020) fine-tuned BERT on claims for IPC/CPC[4] classification, while PatentSBERTa (Bekamiri et al., 2024) leveraged silver labels from a cross-encoder to improve similarity search. Other encoders have been trained from scratch on large patent corpora: BERT for Patents (Srebrovic and Yonamine, 2020) (100M+ documents) captured technical vocabulary, while SciBERT (Beltagy et al., 2019), though trained on scientific texts, transferred effectively to patents with improvements from linguistically-informed masking (Althammer et al., 2021). More recently, ModernBERT (Yousefiramandi and Cooney, 2025) extended pretraining to patents with long-context modeling, improving efficiency over earlier PatentBERT.

Beyond pretraining, contrastive learning has advanced retrieval by constructing positives from metadata. (Li et al., 2022) used IPC codes, while (Xiao et al., 2023) combined topic prompting with contrastive learning for more isotropic embeddings. More recent work exploits examiner citations: (Björkqvist and Kallio, 2023) modeled patents as graphs, pairing first claims with cited descriptions; PaECTER (Ghosh et al., 2024) fine-tuned BERT-for-Patents on EPO citation pairs, setting a strong benchmark for prior-art retrieval; and SearchFormer (Vowinckel and Hähnke, 2023) adopted a similar citation-driven strategy for title–abstract embeddings, though it remains proprietary.

Self-supervised sentence representation learning. Self-supervised learning avoids external labels by constructing signals directly from the data. Early NLP work drew on vision-based contrastive framework (Chen et al., 2020), with methods such as CLEAR (Wu et al., 2020) and ConSERT (Yan et al., 2021) using discrete augmentations. SimCSE (Gao et al., 2021) showed that dropout noise alone can outperform such heuristics, establishing a strong baseline.

Momentum encoders, inspired by MoCo and BYOL, enlarged negative pools and stabilized training (Fang et al., 2020; Zhang et al., 2021). Loss variants introduced angular margins (Zhang et al., 2022b), discriminator-guided alignment (Chuang et al., 2022), or differentiable augmentation via prefix tuning (Wang and Lu, 2022). Hard-negative strategies ranged from Gaussian perturbations (Wu et al., 2021; Zhou et al., 2022) and clustering-aware sampling (Deng et al., 2023) to intermediate-layer negatives (Chen et al., 2023). MixCSE (Zhang et al., 2022a) interpolated features between positives and negatives, while SimCSE++ (Xu et al., 2023) improved gradient flow by disabling dropout on negatives.

Recent non-contrastive objectives such as Barlow Twins (Zbontar et al., 2021) and VICReg (Bardes et al., 2021) align positive views while minimizing redundancy. NLP adaptations (Pappadopulo and Farina, 2024; Çağatan, 2024) achieve performance competitive with SimCSE, offering alternatives without explicit negatives.

Despite this progress, patent-specific self-supervised methods remain underexplored. We address this gap by exploiting the inherent multi-view structure of patents to design stronger data augmentation strategies for more effective contrastive learning.

3 Preliminaries and Background

Patent documents serve not only as legal instruments, but also as structured repositories of technical knowledge. Each patent document organizes complex technical content into standardized sections—such as claims, abstract, and detailed description—that collectively define the novelty, scope, and applicability of the invention. This rich and multi-faceted structure makes patents an ideal yet underutilized resource for learning meaning-

[3] https://github.com/ZoeYou/patentmapv1

[4] Cooperative Patent Classification (CPC): https://www.cooperativepatentclassification.org/home

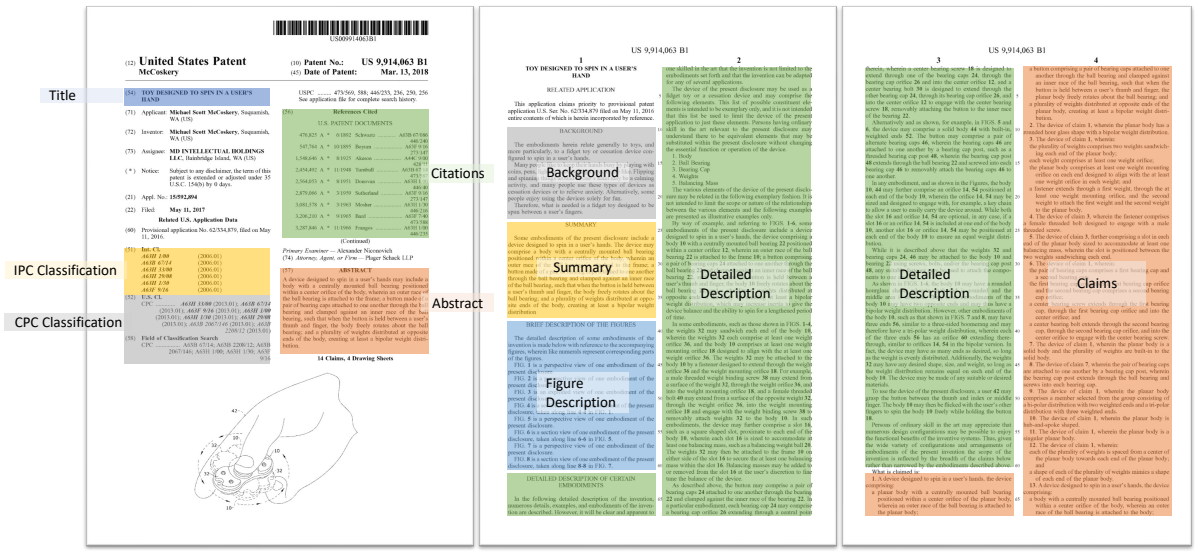

Figure 1: An example of granted patent document (US Patent 9,914,063 B1). (Section structure and formatting may vary across jurisdictions.)

ful textual representations. Fig. 1 shows a typical granted US patent with its key sections annotated.

In this work, we utilize the following sections from each patent document: **Title** (a concise and specific identifier of the invention), **Abstract** (a brief summary of novelty and technical contributions, typically 50–150 words, often used in search and retrieval), **Claims** (the legally binding component of the patent, written in formal, highly structured language to define the scope of exclusive rights), **Background** (provides technical context and motivation for the invention, sometimes referencing existing methods or limitations in prior art)[5], **Summary** (an overview of key features, objectives, and advantages), **Figure Description** (brief explanations of each figure to aid understanding of the invention), and **Detailed Description** (comprehensive disclosure with examples and embodiments, enabling skilled practitioners to reproduce the invention).

4 Methodology

Our framework builds on contrastive learning but is motivated by a domain-specific failure mode we observe in patents, which we term *over-dispersion of embeddings*. We first outline the baseline contrastive setup (Sec. 4.1), then analyze this effect (Sec. 4.2), and finally present our section-based augmentation remedy (Sec. 4.3).

[5]The Background section is not a formal disclosure of prior art but may describe known techniques or problems to motivate the invention.

4.1 Contrastive Learning Framework

Following standard contrastive learning, we construct a set of paired examples $\mathcal{D} = \{(x_i, x_i^+)\}_{i=1}^N$, where each x_i^+ is a *positive view* of x_i, semantically equivalent but differing through data augmentation. In the baseline setting, we adopt SimCSE's dropout-based augmentation: the positive pair consists of identical input sentences, where x_i and x_i^+ are encoded independently using different random dropout masks applied to model activations (dropout probability $p = 0.1$).

Both views are encoded by the same Transformer encoder f_θ. We use **[CLS] pooling**, i.e., we take the final-layer [CLS] vector as the sentence representation. A lightweight projector g_ϕ (linear + `tanh`) maps the pooled feature to the contrastive space, yielding $h = g_\phi(f_\theta(x))$. The projector is discarded at inference.

The model is trained using in-batch negatives with the InfoNCE objective. For a given pair (x_i, x_i^+) within a minibatch of N examples, the loss is defined as:

$$\mathcal{L}_i = -\log \frac{\exp\big(\mathrm{sim}(h_i, h_i^+)/\tau\big)}{\sum_{j=1}^{N} \exp\big(\mathrm{sim}(h_i, h_j^+)/\tau\big)}, \quad (1)$$

where τ is a temperature hyper-parameter and $\mathrm{sim}(h_i, h_j)$ denotes the cosine similarity: $\mathrm{sim}(h_1, h_2) = \frac{h_1^\top h_2}{\|h_1\|\|h_2\|}$.

4.2 Over-dispersion of Embeddings

Sentence embeddings from pre-trained BERT encoders are known to suffer from *anisotropy*: representations concentrate along a few dominant directions, forming a narrow cone with a rapidly decaying singular-value spectrum (Ethayarajh, 2019; Li et al., 2020). Recent studies debate whether this phenomenon is inherent to attention (Godey et al., 2024a) or can be mitigated by training dynamics such as normalization (Machina and Mercer, 2024).

SimCSE (Gao et al., 2021) showed that contrastive learning with dropout augmentation alleviates anisotropy in general-domain text. However, when applying the same setup to patents, we observe a different degeneration. As shown in Figure 2, dropout-only augmentation drives the singular spectrum divergence (SSD) toward a uniform spectrum: embeddings spread more evenly across dimensions, yet downstream performance on IPC classification declines.

We term this the *over-dispersion of embeddings*: representations become excessively isotropic—occupying the space broadly but drifting away from the semantic manifold. In this regime, dropout provides too little semantic variation, so the model learns invariance to random noise rather than meaningful linguistic diversity, resulting in dispersion without semantic anchors.

4.3 Patent Sections as Data Augmentation

To counter over-dispersion, data augmentation must introduce genuine semantic or structural variation rather than near-identical views. Patent documents naturally provide such diversity through their section structure. Abstracts summarize, claims legally delimit, and backgrounds situate the invention in broader context—yielding *semantically aligned but stylistically distinct* views of the same invention.

We therefore propose **section-based augmentation**, which pairs the standard Title+Abstract (TA) with other sections (e.g., claims, background, summary) from the same document. These section-based views act as natural augmentations, enforcing invariance at the document level while anchoring embeddings to invention-level semantics. Unlike surface-level perturbations, our method directly exploits the discourse structure of patents.

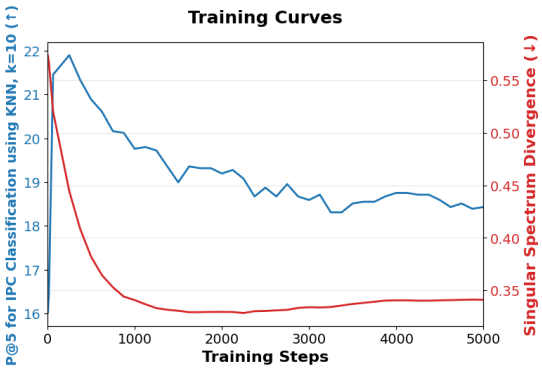

Figure 2: Training curves with dropout-only augmentation (input = **TA**). We report Precision@1 on IPC classification (KNN, k=10) in blue, and **singular spectrum divergence (SSD)** in red, evaluated every 250 steps. (The SSD is defined as the KL divergence between the normalized singular values of the embedding matrix and a uniform distribution.)

In practice, our training incorporates two types of augmentations: (i) **dropout views**, where TA is encoded twice with independent dropout masks, ensuring stability by aligning near-identical views; and (ii) **section views**, where TA is paired with another available section, introducing stronger discourse diversity. The combination is designed to balance consistency (from dropout views) with diversity (from section views).

We illustrate positive-pair construction as follows:

```
Input: patent document with predefined
sections 𝒮
Output: positive pair (x_i, x_i^+)

viewA = Title + Abstract (denoted
TA)
viewB = randomly sample one section
s ∈ 𝒮
if s == Abstract:
  viewB = TA with independent dropout
mask   // dropout-positive
else:
  viewB = full text of section s   //
section-positive
return (viewA, viewB)

// Negatives are constructed analogously
from other patents
```

To avoid anchoring bias, viewA and viewB are randomly swapped with 50% probability during training. For negatives, we consistently sample

from the viewA space (i.e., other patents' TA or their augmented variants), so that negative construction remains comparable across augmentation strategies.

5 Experiments

5.1 Training Data

We use the HUPD dataset (Suzgun et al., 2022), containing English-language USPTO patent applications (2004–2018). For training we select filings from 2010–2018, yielding 2.78M documents after filtering. From each patent we extract six sections: *abstract*, *claims*, *summary*, *background*, *figure description*, and *detailed description*, with *title* concatenated to the *abstract*. To facilitate section-aware learning, each section is prepended with a special token (e.g., `[abstract]`, `[claim]`), following (Srebrovic and Yonamine, 2020). Table 1 reports average lengths of each section. Detailed preprocessing rules (e.g., heading segmentation, boilerplate removal) are provided in Appendix A.

5.2 Implementation Details

Our encoder is initialized from BERT-for-Patents (Srebrovic and Yonamine, 2020) and fine-tuned with a maximum input length of 512 tokens. We use AdamW with learning rate 1×10^{-5}, cosine decay, and 10% warm-up. The temperature τ in Eq. 1 is fixed at 0.05.

Augmentation policies. To assess how different kinds of variation affect patent contrastive representation learning, we compare a spectrum of augmentation policies—ranging from shallow perturbations to discourse- and metadata-informed views. Each policy is applied consistently when forming training pairs, so that differences in performance can be attributed to the nature of the augmentation itself.

Concretely, we evaluate: (i) **Dropout** (SimCSE-style; p=0.1): stochastic feature noise; (ii) **Sentence shuffling** (Yan et al., 2021): surface reordering; (iii) **Random cropping**: structural variation via truncating a 10% continuous span; (iv) **Paraphrasing**: semantic rewriting of TA with an open-source LLM (Qwen3-0.6B (Team, 2025)); (v) **Section-based views** (ours): discourse-level variation by pairing Title+Abstract (TA) with another section from the same patent (e.g., claims, background, summary); (vi) **IPC-matched pairs**: cross-document positives using another patent that shares the *exact* IPC subgroup list[6].

We include a diverse set of augmentations not as competing methods per se, but as controlled comparisons to test our hypothesis that only structural or semantic diversity—rather than surface perturbations—is effective in mitigating over-dispersion and improving downstream performance.

Training setup. Models are trained on a single NVIDIA H100 GPU with batch size 512 for one epoch. We adopt DeepSpeed Stage 1 for memory efficiency and fp16 mixed-precision to accelerate training.

Baseline Models. We compare against four categories of baselines:

- **Domain-specific patent encoders**: BERT-for-Patents (Srebrovic and Yonamine, 2020) (unsupervised pretraining) and PatentBERT (Lee and Hsiang, 2020) (CPC-supervised).
- **Citation-trained models**: PaECTER and PatSPECTER (Ghosh et al., 2024) (patent citations), and SPECTER 2.0 (Singh et al., 2022) (scientific citations).
- **General-purpose embedding LLM**: gte-Qwen2-7B-instruct (Li et al., 2023).
- **Lexical retrieval**: BM25.

Together, these baselines span supervised, citation-driven, and general-purpose approaches, situating our method within both patent-specific and general embedding learning. Details are in Appendix B.

6 Evaluation and Results

6.1 Evaluation Tasks

We assess the learned embeddings on two complementary downstream tasks: (1) prior-art retrieval and (2) IPC classification. Retrieval evaluates local semantic continuity between related patents, while classification inquiries the global organization of embeddings by technology domain.

[6]The International Patent Classification (IPC) hierarchy contains roughly 650 **subclasses** (e.g., `A01B`) and more than 70 000 finer-grained **subgroups** (e.g., `A01B 1/02`). Patents are typically annotated with multiple labels at each level (on average 1–3 subclasses and 10–15 subgroups). When two patents possess an identical *subgroup list*, they usually disclose inventions from very similar technical domains.

Section	title	abstract	claims	background	summary	drawing	detailed description
# words	7.9	108.6	975.0	475.7	685.1	371.7	1582.7

Table 1: Average word counts of patent sections in our dataset.

Prior Art Retrieval. We construct a benchmark using 200 query patents filed in 2021–2022 from the EPO full-text corpus. For each query, we extract cited documents from the official examiner search reports as positive examples. Negative examples are sampled using three strategies: (i) documents retrieved via "more-like-this" (MLT) Elasticsearch queries[7], (ii) patents from the same IPC subclass but not cited, and (iii) cited-of-cited patents not directly linked to the query. The resulting retrieval pool includes 48,110 documents. Detailed sampling protocol is described in Appendix C.2.

To reflect diverse patent retrieval scenarios, we consider two query-to-document configurations:

- **Abstracts Only**: Both queries and documents are represented by their *title* + *abstract* embeddings. This reflects early-stage novelty search where only summaries are available.
- **Claims→All**: Queries use *claims* embeddings, while documents contribute three vectors (*TA*, *claims*, *description*). The final ranked list is de-duplicated at the patent level, mirroring the examiner workflows where claims are compared against the full disclosure of candidate patents.

We report Recall@K (K = 20, 50, 100), averaged over 200 queries.

IPC Classification via KNN. To evaluate global embedding geometry, we perform IPC subclass classification with a k-nearest neighbor ($k = 10$) classifier. We use 30k USPTO patents from 2005–2009 (6k per year), with an 85/15 train/test split from the HUPD dataset (Suzgun et al., 2022). Each document is represented by its *TA* embedding, and labels are assigned by majority vote among neighbors.

6.2 Main Results

Table 2 reports results[8] on prior-art retrieval and IPC classification. All our models (annotated as "Ours") use the same dropout-based contrastive framework, differing only in the augmentation policy used to construct training pairs. We report results from the final checkpoint without early stopping.

Prior Art Retrieval. As shown in the retrieval block of Table 2 (left two blocks), dropout-only training yields limited gains and collapses in the challenging Claims→All setting, confirming that nearly identical positives fail to capture cross-sectional semantics. Classical augmentations provide moderate improvements, with cropping combined with shuffling and paraphrasing outperforming shuffling alone, likely because they introduce length or semantic variation rather than mere surface reordering. However, **section-based augmentation delivers the largest gains**, raising R@100 from 56.21 (dropout) to 71.22 with *claims* as the second view—matching or surpassing citation-trained baselines such as PaECTER at several cutoffs. Summaries and background sections also improve performance, though less strongly, and combining sections yields competitive results. These findings confirm that discourse-level variation supplies the semantic diversity needed to mitigate over-dispersion and improve retrieval. Further analysis (§C.2.3) shows that section-augmented models retrieve a more balanced mix of document sections, indicating improved cross-structure generalization.

IPC Subclass Classification via KNN. The classification block of Table 2 (rightmost three columns) evaluates whether embeddings preserve global technology-domain structure. Dropout alone performs poorly, and classical augmentations provide only modest gains, with cropping combined with sentence shuffling again the strongest among them. Section-based augmentation substantially improves classification, performing on par with or better than the high-dimensional gte-Qwen2-7B-instruct model, despite using far fewer parameters and a smaller embedding dimension. The **+IPC match** variant achieves the best classification scores overall, which is expected since its training signal is directly tied to IPC subgroups overlap—essentially a form of supervised clustering. Our section-based method, though fully self-

[7] https://www.elastic.co/docs/reference/query-dsl/mlt-query

[8] Due to space constraints, in the main text we report only the best-performing configuration of section-based augmentation for each task. Comprehensive results across all section combinations are included in Appendix C.

Table 2: Performance on Prior Art Retrieval (%) and IPC Subclass Classification (%) using KNN ($K = 10$). "Ours" denotes contrastive training with different augmentation policies. Retrieval is reported on Abstract→Abstract and Claims→All settings; classification is reported as P@1/3/5. **Best** and second-best scores are highlighted.

Model	Retrieval: Abs → Abs			Retrieval: Clm → All			Dim	Classification (KNN)		
	R@20	R@50	R@100	R@20	R@50	R@100		P@1	P@3	P@5
BM25	27.23	40.40	51.08	31.06	42.56	56.35	–	–	–	–
PatentBERT	13.43	19.21	27.58	13.74	19.09	25.59	768	49.91	25.02	16.82
SPECTER 2.0	27.01	35.44	44.27	23.18	34.86	46.10	768	52.57	26.13	17.62
Pat-SPECTER	32.03	46.44	58.98	30.84	49.64	64.59	768	55.27	27.83	18.52
BERT-for-patents	26.82	39.96	49.56	26.60	40.21	49.09	1024	57.43	28.08	18.60
PaECTER	39.64	**58.29**	67.08	44.53	62.88	**76.60**	1024	60.00	29.88	19.77
gte-Qwen2-7B-instruct	41.24	54.93	65.97	43.16	61.28	72.99	3584	61.10	29.66	19.57
Dropout (TA→TA)	29.70	45.14	56.21	18.22	29.29	39.83	1024	48.33	24.43	16.50
Ours — Dropout + Classical Data Augmentation										
+ shuffle	31.66	46.75	56.10	27.13	38.65	48.67		48.54	24.83	16.61
+ crop	39.32	55.01	65.43	34.10	50.35	61.51		56.19	27.77	18.41
+ shuffle/crop	41.54	56.25	66.45	35.64	51.80	64.55		56.62	28.01	18.55
+ paraphrase	36.96	54.08	66.13	35.22	52.81	63.05		54.43	27.26	18.25
+ IPC match	33.66	48.46	59.25	33.52	49.25	65.73		**63.53**	**30.37**	**20.00**
Ours — Dropout + Section Augmentation										
+ claim	**44.00**	58.04	**71.22**	**44.85**	**63.40**	75.70		61.10	29.88	19.77
+ summary	42.54	57.95	70.79	42.38	60.55	71.02		61.35	30.23	19.83
+ background	40.93	55.90	66.75	39.38	58.20	69.41		62.11	30.14	19.63
+ claim/summary	42.83	58.22	70.72	44.06	61.54	73.72		60.54	30.04	19.77
+ claim/background	42.38	57.73	69.46	41.95	59.65	72.62		62.27	30.24	19.83
+ summary/background	42.03	56.26	68.24	41.21	58.11	71.24		62.20	30.11	19.73

supervised, comes close to this upper bound. However, as Table 2 shows, the IPC-matched model suffers a sharp drop in retrieval performance, indicating that enforcing strict IPC-driven clustering encourages coarse-grained grouping by technology domains but undermines the ability to capture fine-grained semantic continuity across patents. By contrast, section-based augmentation strikes a better balance between global domain structure and local semantic relations.

We also observe that different sections contribute differently across tasks. *Claims* and *summaries* are most effective for retrieval, likely because they provide concise, semantically focused descriptions of the inventive concept, which align well with prior-art matching. By contrast, the *background* section contributes more to classification, as it explicitly situates the invention within a broader technical field and emphasizes domain-level context.

6.3 Embedding Diagnostics

Beyond task performance, we analyze the geometry of the learned embedding spaces. We follow Wang and Isola (2020) and report three complementary metrics: (i) **alignment** (lower is better), computed over *citing–cited* pairs from the prior-art benchmark (not training positives), measuring how well related patents are pulled together; (ii) **uniformity** (lower is better), quantifying how evenly embeddings spread on the unit hypersphere; and (iii) **Singular Spectrum Divergence (SSD)** (lower is better), defined as the KL divergence between the normalized singular value spectrum and the uniform distribution. Equivalently, it corresponds (up to an additive constant) to the negative Shannon entropy of the spectrum, and is closely related to effective rank and spectral entropy measures used in prior work on anisotropy (Roy and Vetterli, 2007; Godey et al., 2024b). Because SSD is bounded above by $\log d$ (embedding dimensionality), we normalize it by $\log d$ to allow fair comparison across models with different hidden sizes.

Formal mathematical definitions of the three metrics are deferred to Appendix D.

Figure 3 situates all models in the alignment–uniformity plane, with SSD shown by point size and retrieval performance by color. Three clear patterns emerge:

- **Citation-trained baselines** (PaECTER, SPECTER 2.0, Pat-SPECTER) exhibit excellent alignment but weaker uniformity and higher SSD, indicating compact clusters around supervision signals but limited isotropy. This favors tight lexical/semantic matches but hinders generalization.

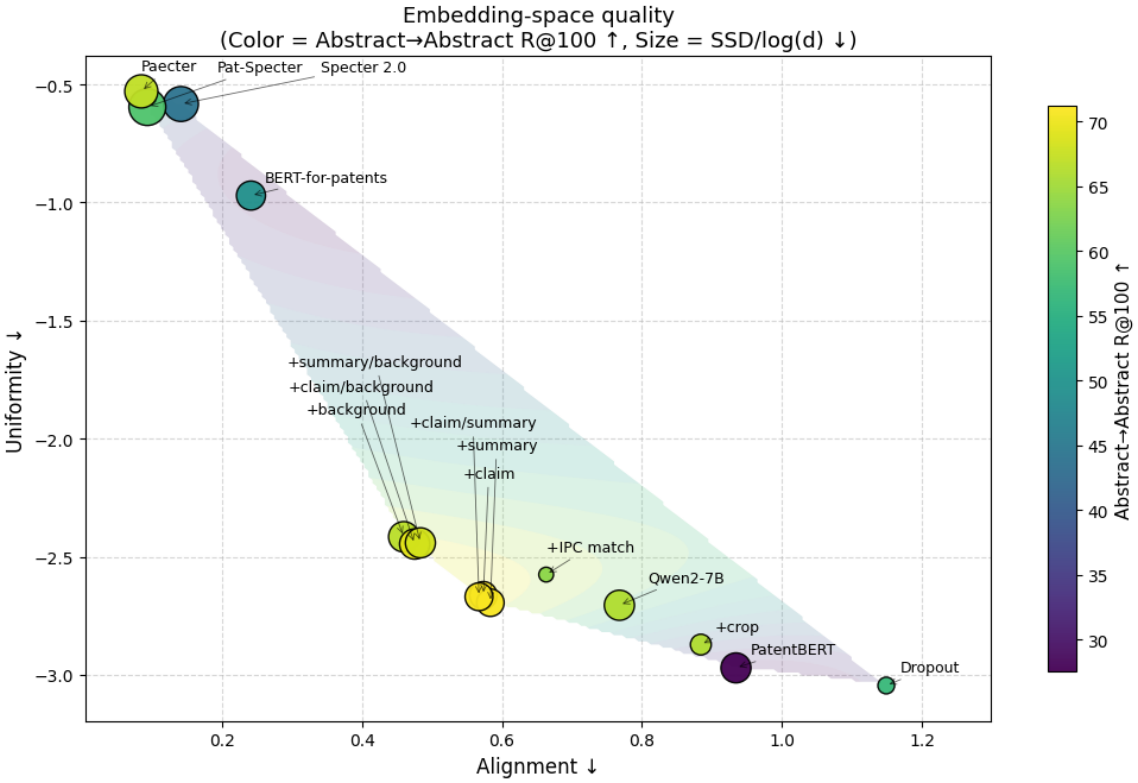

Figure 3: Embedding-space diagnostics. Each point shows alignment (x-axis) and uniformity (y-axis), dot size encodes normalized SSD, and color encodes prior-art retrieval performance (R@100, Abstract→Abstract). Contours indicate an RBF-smoothed performance field within the convex hull of observed points.

- **Dropout-only contrastive** produces the opposite profile: lower uniformity and SSD (more isotropic embeddings) but much worse alignment, meaning semantic neighbors are no longer preserved. This "over-dispersion" reflects a curse-of-dimensionality effect, where embeddings spread evenly yet lose local cohesion, leading to degraded retrieval and IPC KNN classification.

- **Section-based augmentation** strikes a balance. Models using claims, summaries, or their combinations achieve both low alignment (good local cohesion) and improved uniformity/SSD compared to citation-trained baselines. This balanced geometry explains their superior retrieval and stable classification performance.

In short, the diagnostics reinforce our central claim: *leveraging intra-document sections produces augmentations that maintain semantic alignment while avoiding over-dispersion, yielding embedding spaces that are both structured and well-distributed.*

7 Conclusion

We proposed a contrastive framework for patent representation learning that exploits the natural section structure of patent documents to construct semantically diverse positive pairs. Different sections—abstracts, claims, summaries, and descriptions—capture the same invention from complementary perspectives, providing richer self-supervised signals than standard dropout-based augmentations.

Experiments demonstrate that section-based models not only outperform classical augmentation strategies but also rival or surpass citation- and IPC-supervised baselines, achieving strong results on both retrieval and classification. These findings suggest that structure-aware self-supervision yields embedding spaces that are both locally coherent and globally well organized. We hope this work encourages the patent-research community to look beyond noisy supervision and explore self-supervision as a scalable and robust path for future patent representation learning.

Limitations

While our method demonstrates strong performance across multiple tasks, several limitations remain. First, we rely on a fixed set of section pairs (at most two per example), treating all combinations equally without dynamically selecting views based on their semantic complementarity or informativeness. Second, we use only a single positive pair per training instance, which may underexploit the rich multi-view structure of patent documents. Third, our dataset has not been filtered for patent families, so related filings from the same invention family may occasionally appear as negatives in the contrastive pool. This can introduce mild label noise and underestimate the model's potential performance. Fourth, our pretraining is limited to English USPTO applications. Although we evaluate cross-jurisdictionally on EPO search-report pairs, we have not yet trained on non-US or multilingual corpora; extending both training and evaluation to additional jurisdictions and languages would better assess view alignment and robustness. Finally, our method is designed for well-structured documents like patents; applying it to less structured domains (e.g., scientific papers or legal decisions) may require more sophisticated view mining or filtering strategies.

References

Sophia Althammer, Mark Buckley, Sebastian Hofstätter, and Allan Hanbury. 2021. Linguistically informed masking for representation learning in the patent domain. *arXiv preprint arXiv:2106.05768*.

Adrien Bardes, Jean Ponce, and Yann LeCun. 2021. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*.

Hamid Bekamiri, Daniel S Hain, and Roman Jurowetzki. 2024. Patentsberta: A deep nlp based hybrid model for patent distance and classification using augmented sbert. *Technological Forecasting and Social Change*, 206:123536.

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

Sebastian Björkqvist and Juho Kallio. 2023. Building a graph-based patent search engine. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3300–3304.

Matthias Blume, Ghobad Heidari, and Christoph Hewel. 2024. Comparing complex concepts with transformers: Matching patent claims against natural language text. *arXiv preprint arXiv:2407.10351*.

Ömer Çağatan. 2024. Unsee: Unsupervised non-contrastive sentence embeddings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 384–393.

Nuo Chen, Linjun Shou, Ming Gong, Jian Pei, Bowen Cao, Jianhui Chang, Daxin Jiang, and Jia Li. 2023. Alleviating over-smoothing for unsupervised sentence representation. *arXiv preprint arXiv:2305.06154*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. *arXiv preprint arXiv:2204.10298*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

Jinghao Deng, Fanqi Wan, Tao Yang, Xiaojun Quan, and Rui Wang. 2023. Clustering-aware negative sampling for unsupervised sentence representation. *arXiv preprint arXiv:2305.09892*.

David Ellis. 2025. Samsung takes top spot in u.s. patents for third year running while tsmc rises into second place; after four-year falloff, grants increase nearly 4%. Press-release on IFI CLAIMS 2024 patent data rankings.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Mainak Ghosh, Sebastian Erhardt, Michael E Rose, Erik Buunk, and Dietmar Harhoff. 2024. Paecter: Patent-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2402.19411*.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024a. Anisotropy is inherent to self-attention in transformers. *arXiv preprint arXiv:2401.12143*.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024b. Why do small language models underperform. *Studying language model saturation via the softmax bottleneck*, 2404.

Ivan Haščič, Jérôme Silva, and Nick Johnstone. 2015. The use of patent statistics for international comparisons and analysis of narrow technological fields.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent classification by fine-tuning bert language model. *World Patent Information*, 61:101965.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Huahang Li, Shuangyin Li, Yuncheng Jiang, and Gansen Zhao. 2022. Copate: a novel contrastive learning framework for patent embeddings. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1104–1113.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Anemily Machina and Robert Mercer. 2024. Anisotropy is not inherent to transformers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4892–4907.

Duccio Pappadopulo and Marco Farina. 2024. Non-contrastive sentence representations via self-supervision. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4274–4284.

Julian Risch, Nicolas Alder, Christoph Hewel, and Ralf Krestel. 2020. PatentMatch: A dataset for matching patent claims with prior art. *ArXiv e-prints*.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

Rob Srebrovic and Jay Yonamine. 2020. Leveraging the bert algorithm for patents with tensorflow and bigquery. Technical report, Global Patents, Google, https://services.google.com/fh/files/blogs/bert_for_patents_white_paper.pdf.

Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K. Sarkar, Scott Duke Kominers, and Stuart M. Shieber. 2022. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Adam van Hoewijk and Henrik Holmström. 2022. Exploring supervision levels for patent classification.

Konrad Vowinckel and Volker D Hähnke. 2023. Searchformer: Semantic patent embeddings by siamese transformers for prior art search. *World Patent Information*, 73:102192.

Tianduo Wang and Wei Lu. 2022. Differentiable data augmentation for contrastive sentence representation learning. *arXiv preprint arXiv:2210.16536*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR.

Xing Wu, Chaochen Gao, Yipeng Su, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2021. Smoothed contrastive learning for unsupervised sentence embedding. *arXiv preprint arXiv:2109.04321*.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Qingfa Xiao, Shuangyin Li, and Lei Chen. 2023. Topicdpr: Topic-based prompts for dense passage retrieval. *arXiv preprint arXiv:2310.06626*.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. Simcse++: Improving contrastive learning for sentence embeddings from two perspectives. *arXiv preprint arXiv:2305.13192*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. *arXiv preprint arXiv:2105.11741*.

Amirhossein Yousefiramandi and Ciaran Cooney. 2025. Patent language model pretraining with modernbert. *arXiv preprint arXiv:2509.14926*.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR.

Yan Zhang, Ruidan He, Zuozhu Liu, Lidong Bing, and Haizhou Li. 2021. Bootstrapped unsupervised sentence representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5168–5180.

Yanzhao Zhang, Richong Zhang, Samuel Mensah, Xudong Liu, and Yongyi Mao. 2022a. Unsupervised sentence representation via contrastive learning with mixing negatives. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11730–11738.

Yuhao Zhang, Hongji Zhu, Yongliang Wang, Nan Xu, Xiaobo Li, and Binqiang Zhao. 2022b. A contrastive framework for learning sentence representations from pairwise and triple-wise perspective in angular space. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4892–4903.

Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. *arXiv preprint arXiv:2205.00656*.

A Data Preprocessing Details

Corpus selection. We use the Harvard USPTO Patent Dataset (HUPD) (Suzgun et al., 2022), which contains all English-language USPTO utility patent applications between 2004–2018. For training, we select patent applications filed between 2010–2018. This yields 2,780,124 documents after filtering (see below).

Section extraction. From each patent we extract six sections in addition to the title: *abstract*, *claims*, *summary*, *background*, *figure description*, and *detailed description*. The *title* is concatenated with the *abstract*, separated by `[SEP]`. Extraction rules:

- **Abstract**: directly from the `abstract` field; patents without an abstract are discarded.
- **Claims**: from the `claims` field. Canceled claims (e.g., "1–16. (cancelled)") are removed, valid numbering is preserved.
- **Summary** and **Background**: taken directly from the corresponding HUPD fields. We remove section headings that appear between `<SOH>` and `<EOH>` tags, keeping only the body text. For *summary*, we additionally remove entries misclassified as figure descriptions, identified via regular expressions (e.g., beginning with "BRIEF DESCRIPTION OF THE DRAWINGS").
- **Figure description** and **Detailed description**: extracted from the `full_description` field using heading-based segmentation. We remove non-technical prefatory text (e.g., "CROSS-REFERENCE TO RELATED APPLICATIONS").

Filtering. We apply the following filters to ensure content quality:

- Non-technical prefatory text (e.g., cross-reference statements, government interest statements, or generic legal disclaimers) is removed.
- Patents with empty abstracts are dropped.
- Sections shorter than 15 words are discarded (titles are exempted from this rule).

Special tokens. Following BERT-for-Patents (Srebrovic and Yonamine, 2020), we prepend special section tokens to mark different parts of the document: `[abstract]`, `[claim]`, `[summary]`, and `[invention]` (for background). These four tokens are part of the original BERT-for-Patents vocabulary. In addition, we introduce two new tokens, `[drawing]` and `[description]`, to represent the *figure description* and *detailed description* sections, respectively. Both are initialized with the same embedding as `[invention]`.

Statistics. Table 1 (main paper) reports average word counts per section. In general, abstracts are short (∼100 words), while claims and descriptions are long (up to 1,500+ words). This natural length diversity reinforces the motivation for section-based augmentation.

B Baseline Model Descriptions

1. **PatentBERT** (Lee and Hsiang, 2020): A `bert-base-uncased` model fine-tuned on large-scale patent claim data for CPC subclass classification. For our evaluation, we remove the classification head and use the encoder to extract sentence embeddings.
2. **Bert-for-patents** (Srebrovic and Yonamine, 2020): A large-scale BERT-based model pretrained on over 100 million patent documents. It is trained on all components of a patent (including abstract, claims, and description) and serves not only as a competitive baseline

but also as the initialization checkpoint for our encoder.

3. **SPECTER 2.0** (Singh et al., 2022): An advancement over the original Specter (Cohan et al., 2020), this model is based on SciBERT—a BERT variant trained from scratch on a large corpus of scientific literature—and fine-tuned using a triplet loss on scientific paper citation graphs. SPECTER 2.0 benefits from training on an order of magnitude more data across 23 scientific fields, making it a robust model for scientific document embedding.

4. **PaECTER and Pat-SPECTER** (Ghosh et al., 2024): These two models are fine-tuned on patent citation graphs. PaECTER is derived from Bert-for-patents, whereas Pat-SPECTER 2.0 extends the Specter approach to the patent domain. In these models, the document representation is obtained by concatenating the title and abstract of the patent, capturing the most salient information.

5. **gte-Qwen2-7B-instruct**[9] (Li et al., 2023): Beyond domain-specific models, we include a state-of-the-art general text embedding model. gte-Qwen2-7B-instruct, a member of the General Text Embedding (GTE) family, ranked first in both English and Chinese on the Massive Text Embedding Benchmark (MTEB) as of June 16, 2024. Its performance on MTEB demonstrates its strong cross-domain and multilingual representation capabilities.

6. **BM25**: As a classical baseline from the field of information retrieval, BM25 provides a non-neural, lexical matching approach that is widely used for document ranking. Despite its simplicity, BM25 remains a competitive baseline in many retrieval scenarios.

C Full Results and Additional Analysis

This appendix provides complete numbers and analyses complementing the main text. We first present *IPC classification* (Linear Probe & KNN), then *prior-art retrieval* (Abs→Abs and Clm→All), and finally a controlled ablation of *positive-only* vs. *positive+negative* augmentation.

C.1 IPC Subclass Classification

Linear Probe setup (complementary to KNN). As a complementary evaluation, we also perform IPC subclass classification using a supervised linear probe. While KNN evaluates the clustering quality of raw embeddings, linear probing tests whether semantic category information can be linearly separated.

We use the same held-out HUPD slice as for KNN (30k docs from 2005–2009). A frozen-encoder 2-layer MLP (hidden 1024, ReLU, sigmoid, BCE) is trained with 85/15 train/dev split. While probing introduces additional tuning, it offers a complementary view to KNN (linearly separable vs. geometry-preserving).

Table 3 presents a comprehensive view of IPC subclass classification results using both Linear Probe and KNN. Several observations emerge:

Model Dimensionality and Representation Efficiency. Despite having a significantly larger embedding size (3584), the embedding model **gte-Qwen2-7B-instruct** fails to outperform our 1024-dimensional representations on either Linear Probe or KNN. In contrast, many of our section-augmented models—especially those using summary or background—achieve better precision scores at a much lower dimensional cost. This highlights the efficiency of our contrastively trained encoders, which leverage structural inductive biases from the patent document itself rather than relying on sheer model size.

Linear vs. KNN Classification. We observe that many of our models, especially those trained with section-augmented views (e.g., +claim, +summary), exhibit stronger performance in the Linear Probe setting than in KNN, suggesting that *the learned representations are highly linearly separable.* In contrast, most baseline models such as BERT-for-patents and PaECTER, while competitive in KNN, show smaller gains under a learned classifier, suggesting our training framework shapes the embedding space with sharper inter-class boundaries. This effect may be attributed in part to the use of an MLP head during training, which—although discarded during inference—may encourage the base encoder to produce embeddings with improved linear decision surfaces.

[9] https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct

Table 3: IPC Subclass Classification Performances (%) using Linear Probe and KNN ($K = 10$). All models use TA as input. "Ours" denotes contrastive training with different augmentation strategies. Best results are **bolded**; second-best are underlined.

Model	Dim	Linear Probe			KNN		
		P@1	P@3	P@5	P@1	P@3	P@5
PatentBERT	768	54.14	27.85	19.01	49.91	25.02	16.82
SPECTER 2.0	768	50.68	26.92	18.51	52.57	26.13	17.62
Pat-SPECTER	768	53.35	27.95	19.14	55.27	27.83	18.52
BERT-for-patents	1024	55.99	29.22	19.71	57.43	28.08	18.60
PaECTER	1024	57.61	30.06	20.42	60.00	29.88	19.77
gte-Qwen2-7B-instruct	3584	61.55	30.91	20.64	61.10	29.66	19.57
Dropout Only — SimCSE-style training							
Dropout (TA → TA)	1024	55.45	28.85	19.89	48.33	24.43	16.50
Ours — Dropout + Classical Data Augmentation							
+ shuffle		55.97	29.20	19.93	48.54	24.83	16.61
+ crop		58.87	30.16	20.57	56.19	27.77	18.41
+ shuffle/crop		59.37	30.32	20.68	56.62	28.01	18.55
+ paraphrase		59.28	30.20	20.47	54.43	27.26	18.25
+ IPC match		63.17	**31.69**	**21.29**	**63.53**	**30.37**	**20.00**
Ours — Dropout + Section Augmentation							
+ claim		62.31	31.46	21.18	61.10	29.88	19.77
+ summary		62.99	31.52	21.17	61.35	30.23	19.83
+ background		62.94	31.56	21.22	62.11	30.14	19.63
+ drawing		62.29	31.27	21.15	60.78	29.84	19.56
+ description		62.58	31.45	21.18	61.57	29.88	19.55
+ claim/summary		62.76	31.53	21.27	60.54	30.04	19.77
+ claim/background		**63.39**	**31.69**	**21.29**	62.27	30.24	19.83
+ claim/drawing		62.56	31.31	21.18	60.76	29.90	19.68
+ claim/description		62.67	31.35	21.22	61.50	29.98	19.82
+ summary/background		62.76	31.62	21.25	62.20	30.11	19.73
+ summary/drawing		62.61	31.28	21.15	61.28	29.88	19.62
+ summary/description		62.97	31.37	21.20	61.41	30.08	19.73
+ background/drawing		63.28	31.43	21.21	62.11	30.22	19.78
+ background/description		62.94	31.53	21.25	61.86	29.95	19.67
+ drawing/description		62.47	31.43	21.18	61.39	29.88	19.63

C.2 Prior-art Retrieval

C.2.1 Evaluation Set and Metric

Evaluating prior-art search on the full patent corpus is impractical; we therefore construct a focused benchmark of 200 queries with positives and challenging hard negatives.

We start from the European Patent Office (EPO) full-text collection, which includes XML-formatted patent texts and metadata dating back to 1978. Notably, since 2012, the dataset includes search reports containing citations issued by examiners with paragraph-level references. Following PatentMatch[10] (Risch et al., 2020), we extract cited pairs labeled "X" (novelty-destroying), "Y" (obvious in combination), and "A" (background) as positive examples, adapting their pipeline in Elasticsearch.

Query selection. We sample 200 EPO applications filed in 2021–2022 using a two-step strategy to ensure both diversity and citation coverage:

- **Maximizing second-level coverage**: We iteratively select patents that introduce new second-level citations (i.e., patents cited by cited patents), maximizing the overall diversity of the citation graph.
- **IPC diversity enforcement**: We ensure coverage across all eight IPC sections (A–H) by prioritizing documents with underrepresented classifications during early sampling rounds.

Hard negative sampling. To simulate realistic retrieval conditions, we include challenging negative examples for each query:

1. **MLT Negatives**: Retrieved using Elasticsearch's "more-like-this" queries, based on a query's title, abstract, and citing claim across four fields (title, abstract, claims, description).
2. **IPC-Matching Negatives**: Patents from the same IPC subclass as the query but not cited.

[10] https://github.com/julian-risch/PatentMatch

3. **Cited-of-Cited Negatives**: Patents referenced by cited documents but not directly linked to the query.

Negatives are pooled globally so the test set mirrors open-world retrieval. This yields a compact yet diverse evaluation benchmark summarized in Table 4.

Table 4: Statistics for the 200-query evaluation set. Each test document includes citations (positive examples) and hard negatives derived using various strategies.

Size	Mean Cited	Cited-of-Cited	MLT	IPC
	Positive	Hard Negatives		
48,110	2.5	4.4	553.2	159.5

We report Recall@K (K=20/50/100), i.e., the fraction of cited documents ranked in the top-K.

C.2.2 Full Results

Table 5 extends the main-text comparison by reporting all augmentation variants. Overall patterns are consistent with Table 2 in the main paper. Full per-policy numbers are reported here for transparency and reproducibility.

C.2.3 Which Sections Are Retrieved?

To better understand how different models exploit the multi-section structure of patents, we analyze the composition of their top-100 retrieved documents in the *Claims → All Sections* setting. Figure 4 shows, for each method, the proportion of retrieved documents originating from the *abstract*, *claims*, or *description*, alongside retrieval performance (R@100).

Baseline models. Citation-tuned systems (PatSPECTER, PaECTER) largely return claim→claim neighbors (often 70–95%), exhibiting strong *section homophily*. This aligns with good recall, but narrows the retrieved context.

Classical augmentations. Dropout-only is highly claim-centric (>90% claims) with low recall. Shuffling/cropping modestly raise description share (and recall), while paraphrasing brings more descriptions (∼28%), indicating semantic variation improves diversity.

Section-based augmentation. Our policies retrieve a markedly more balanced mix—30–47% descriptions at comparable or better recall (e.g., `+claim` and `+summary` near ∼75 R@100). Combining sections (e.g., `+claim/summary`, `+claim/drawing`) yields the most even profiles (roughly 40% claims, 40–45% descriptions, 15% abstracts).

This reduced section homophily explains the robustness of section-based training in heterogeneous discourse settings.

C.3 Ablation: Positive-only vs. Positive+Negative Augmentation

Setup. We compare **pos-only** (augment positives; negatives fixed as TA) with **pos+neg** (apply the same policy to both) to decouple the impact of diversifying positives from diversifying negatives.

Findings. Table 6 summarizes results (reporting pos scores and Δ=pos+neg−pos). Two trends are consistent across policies: (1) enriching *positives alone* already surpasses dropout-only, confirming that semantic/structural diversity in views is the primary driver; (2) extending augmentation to *negatives* yields further gains—especially on *Clm→All* and KNN—by diversifying negative neighborhoods and reducing section homophily (cf. Fig. 4). These observations motivate adopting the **pos+neg** configuration as our main setting.

D Embedding Diagnostics

To evaluate embedding-space quality, we report three complementary metrics: **alignment**, **uniformity** (both from Wang and Isola, 2020), and **singular spectrum divergence (SSD)**. SSD complements the pairwise perspective of alignment/uniformity with a spectral view of anisotropy. Together, these metrics provide a holistic assessment of local similarity structure and global geometric spread.

Alignment. Alignment quantifies how close positive pairs (e.g., citing–cited patents) are in the embedding space:

$$\mathcal{A} = \mathbb{E}_{(i,j)\sim P_{\text{pos}}}\left[\|\mathbf{z}_i - \mathbf{z}_j\|^2\right].$$

Lower alignment means semantically related documents are embedded closer.

Uniformity. Uniformity measures how well the embeddings are spread across the unit hypersphere:

$$\mathcal{U} = \log \mathbb{E}_{(i,j)\sim P_{\text{data}}}\left[e^{-2\|\mathbf{z}_i - \mathbf{z}_j\|^2}\right].$$

Lower values indicate more uniform coverage (less concentration or collapse).

Table 5: Prior Art Retrieval Performances (%). "Ours" denotes contrastive training with different augmentation strategies. Best results are **bolded**; second-best are underlined.

Model	Abstract → Abstract			Claims → All Sections		
	R@20	R@50	R@100	R@20	R@50	R@100
BM25	27.23	40.40	51.08	31.06	42.56	56.35
PatentBERT	13.43	19.21	27.58	13.74	19.09	25.59
SPECTER 2.0	27.01	35.44	44.27	23.18	34.86	46.10
Pat-SPECTER	32.03	46.44	58.98	30.84	49.64	64.59
BERT-for-patents	26.82	39.96	49.56	26.60	40.21	49.09
PaECTER	39.64	**58.29**	67.08	44.53	62.88	**76.60**
gte-Qwen2-7B-instruct	41.24	54.93	65.97	43.16	61.28	72.99
Dropout Only — SimCSE-style training						
Dropout (TA → TA)	29.70	45.14	56.21	18.22	29.29	39.83
Ours — Dropout + Classical Data Augmentation						
+ shuffle	31.66	46.75	56.10	27.13	38.65	48.67
+ crop	39.32	55.01	65.43	34.10	50.35	61.51
+ shuffle/crop	41.54	56.25	66.45	35.64	51.80	64.55
+ paraphrase	36.96	54.08	66.13	35.22	52.81	63.05
+ IPC match	33.66	48.46	59.25	33.52	49.25	65.73
Ours — Dropout + Section Augmentation						
+ claim	**44.00**	58.04	**71.22**	**44.85**	**63.40**	75.70
+ summary	42.54	57.95	70.79	42.38	60.55	71.02
+ background	40.93	55.90	66.75	39.38	58.20	69.41
+ drawing	42.67	55.64	66.54	42.10	59.59	72.24
+ description	41.35	53.75	64.89	40.32	58.16	69.27
+ claim/summary	42.83	58.22	70.72	44.06	61.54	73.72
+ claim/background	42.38	57.73	69.46	41.95	59.65	72.62
+ claim/drawing	43.05	56.97	68.32	43.10	61.79	73.57
+ claim/description	42.18	55.65	66.39	42.85	60.37	72.12
+ summary/background	42.03	56.26	68.24	41.21	58.11	71.24
+ summary/drawing	42.08	57.46	67.25	43.59	59.69	71.41
+ summary/description	41.05	55.27	65.33	41.85	58.30	70.87
+ background/drawing	40.89	56.82	68.12	42.30	59.12	71.92
+ background/description	42.08	55.83	65.43	42.03	57.37	68.57
+ drawing/description	41.77	55.84	64.99	42.34	57.27	70.22

Singular Spectrum Divergence (SSD). Let $M \in \mathbb{R}^{n \times d}$ be the column-centered embedding matrix, with normalized singular values $\mathbf{s} = (s_1, \ldots, s_d)$ such that $\sum_i s_i = 1$. We define:

$$SSD = D_{\mathrm{KL}}(\mathbf{s} \,\|\, \mathcal{U}_d) = \sum_{i=1}^{d} s_i \log \frac{s_i}{1/d},$$

where $\mathcal{U}_d$ is the uniform distribution over d dimensions. Equivalently, $SSD = \log d - H(\mathbf{s})$, i.e., the gap between maximum possible entropy and the entropy of the spectrum. Lower values imply more isotropic, information-rich representations; higher values mean variance is concentrated in a few dominant directions. Because $SSD \leq \log d$, we normalize by $\log d$ for comparability across models of different embedding dimension.

We report section-wise diagnostics (alignment, uniformity, SSD) in heatmap Fig. 5. We summarize the main observations below.

1. **Alignment.** Baselines generally show the lowest alignment (best grouping) for *claims*, likely reflecting their highly formulaic style, followed by *descriptions*, while *TA* is weakest—except for citation-trained models (Pat-SPECTER, PaECTER) where TA is explicitly optimized. Under our *section-based* augmentation, the ordering is less consistent: in many cases TA still shows the weakest alignment, but in others (e.g., background- or description-based views) claims alignment deteriorates more. This suggests that cross-section positives reduce overfitting to claim-specific templates while also redistributing alignment difficulty across sections.

2. **Uniformity & SSD (spread vs. isotropy).** Dropout-only and other non–section-based methods typically make *TA* the most isotropic section (lowest uniformity and SSD). By contrast, under most of our *section-based* augmentations the pattern reverses: *claims* become the most isotropic, often surpassing descriptions as well. This shift suggests that cross-section positives redistribute variance across sec-

Table 6: Positive-only vs. Positive+Negative augmentation across all policies. Retrieval is Recall@100 on EPO prior-art search (Abs→Abs and Clm→All); classification is IPC subclass KNN P@1. "pos" = positive-only; Δ = (pos+neg) − (pos). Positive Δ means gains from also augmenting negatives.

Augmentation	Abs→Abs (R@100)		Clm→All (R@100)		IPC KNN P@1	
	pos	Δ	pos	Δ	pos	Δ
Dropout (TA→TA)	56.75	—	38.78	—	48.22	—
Shuffling	57.20	-1.10	49.73	-1.06	48.74	-0.20
Cropping	65.97	-0.54	60.95	+0.56	55.61	+0.58
Shuffle+Crop	67.43	-0.98	62.66	+1.89	56.69	-0.07
Paraphrasing	66.03	+0.10	62.80	+0.25	54.14	+0.29
IPC match	63.69	-4.44	65.10	+0.63	64.14	-0.61
Section: claim	69.97	+1.25	73.09	+2.61	61.10	+0.00
Section: summary	70.26	+0.53	70.24	+0.78	61.35	+0.57
Section: background	68.80	-2.05	67.09	+2.32	62.20	-0.09
Section: drawing	66.76	-0.22	66.70	+5.54	60.76	-1.20
Section: description	65.41	-0.52	66.17	+3.10	61.14	+0.43
Section: claim+summary	71.27	-0.55	71.40	+2.32	59.95	+0.59
Section: claim+background	69.47	-0.01	67.64	+4.98	61.48	+0.79
Section: claim+drawing	68.16	+0.16	67.49	+6.08	60.92	-0.16
Section: claim+description	65.87	+0.52	67.87	+4.25	61.26	+0.24
Section: summary+background	69.02	-0.78	68.40	+2.84	61.64	+0.09
Section: summary+drawing	67.49	-0.24	67.40	+4.01	60.92	+0.36
Section: summary+description	65.76	-0.43	66.89	+3.98	60.69	+0.72
Section: background+drawing	69.40	-1.28	67.09	+4.83	61.93	+0.18
Section: background+description	67.02	-1.59	67.02	+1.55	61.44	+0.42
Section: drawing+description	66.62	-1.63	66.10	+4.12	60.72	+0.91

tions: TA no longer dominates isotropy, while claims—which are syntactically rigid and semantically dense—become the primary anchor for enforcing isotropic structure.

3. **Impact of the training signal.** Classical augmentations (shuffle, crop, paraphrase) operate only within the Title+Abstract view and introduce mainly surface-level perturbations; accordingly, section-wise metrics remain relatively close across TA/claims/description. Label-driven cross-document pairing (IPC-MATCH) shows a different pattern: it tightly regularizes the TA view—yielding low SSD/log d for TA—while *claims* and *descriptions*, which are not directly aligned by the label, exhibit noticeably higher SSD/log d and weaker uniformity (variance concentrated in a few directions). Our *section-based augmentation* (self-supervised cross-discourse positives) displays the opposite trend: it reduces over-dispersion and redistributes isotropy more evenly across sections—claims often become the most isotropic (lowest SSD/log d), and the gap among TA/claims/description narrows. These observations suggest hybrid strategies worth exploring, e.g., injecting a small fraction of cross-section positives into IPC-MATCH or using section-aware temperatures/weights to harmonize constraints across views.

E Document Topology Analysis (Intra-Document Alignment)

To complement the global embedding diagnostics (§D), we analyze the internal geometric consistency of each document—how closely the embeddings of its sections (*Abstract*, *Claims*, *Description*) are aligned in the representation space. This provides a local, document-level view of how well the encoder preserves intra-document semantic structure.

Definition. For a document d with section embeddings $\{\mathbf{z}_{d,s}\}_{s\in S_d}$, we define its *intra-document alignment* as the average cosine distance among all section pairs:

$$A_{\text{intra}}(d) = \mathbb{E}_{(s_i,s_j)\in S_d}\left[1-\cos(\mathbf{z}_{d,s_i},\mathbf{z}_{d,s_j})\right].$$

To account for the model's overall embedding dispersion, we normalize this by the expected random-pair distance:

$$\text{IDA-Ratio} = \frac{\mathbb{E}_d[A_{\text{intra}}(d)]}{\mathbb{E}_{(i,j)\sim P_{\text{rand}}}[1-\cos(\mathbf{z}_i,\mathbf{z}_j)]}.$$

where P_{rand} denotes pairs of sections randomly sampled from different documents in the corpus.

Lower values indicate stronger intra-document coherence relative to global spread.

Interpretation. Intra-document alignment complements global diagnostics (alignment, uniformity, SSD) by assessing whether different sections of the same patent occupy a compact subspace. A lower IDA-Ratio thus reflects *localized semantic cohesion*—sections that differ stylistically yet encode the same invention remain close in embedding space even when global isotropy improves.

As shown in Figure 6, **section-based contrastive learning produces the strongest intra-document coherence** across all section pairs. Compared to dropout or citation-supervised baselines, our models yield markedly lower normalized distances between *Abstract–Claims*, *Abstract–Description*, and *Claims–Description*, indicating tighter internal topology and smoother transitions across patent discourse levels.

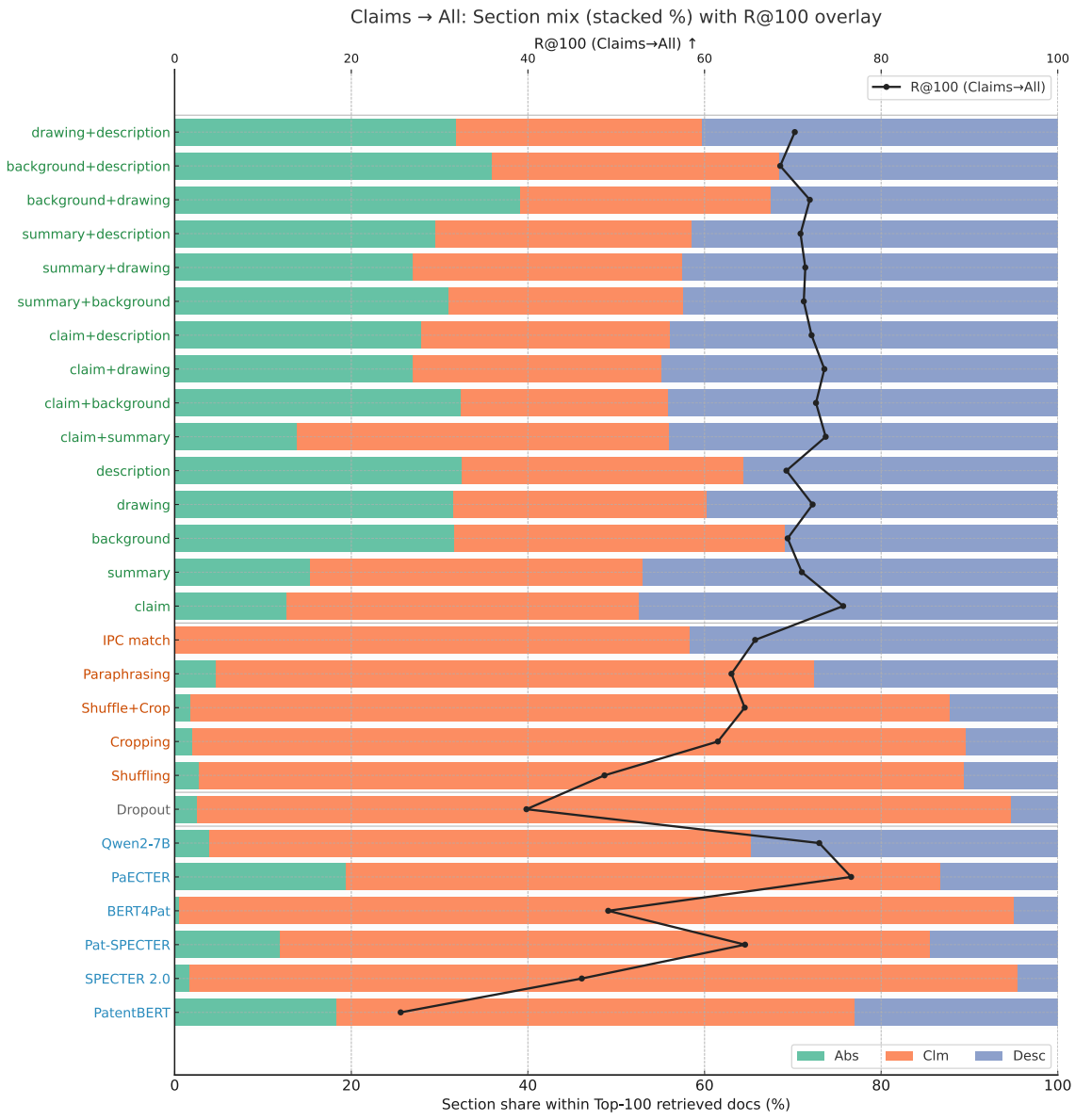

Figure 4: Section distribution of top–100 retrieved documents in *Claims→All*. Our section-augmented model retrieves a more balanced mix beyond *claims*, increasing the share of *summary* and *background*, which provide complementary discourse cues.

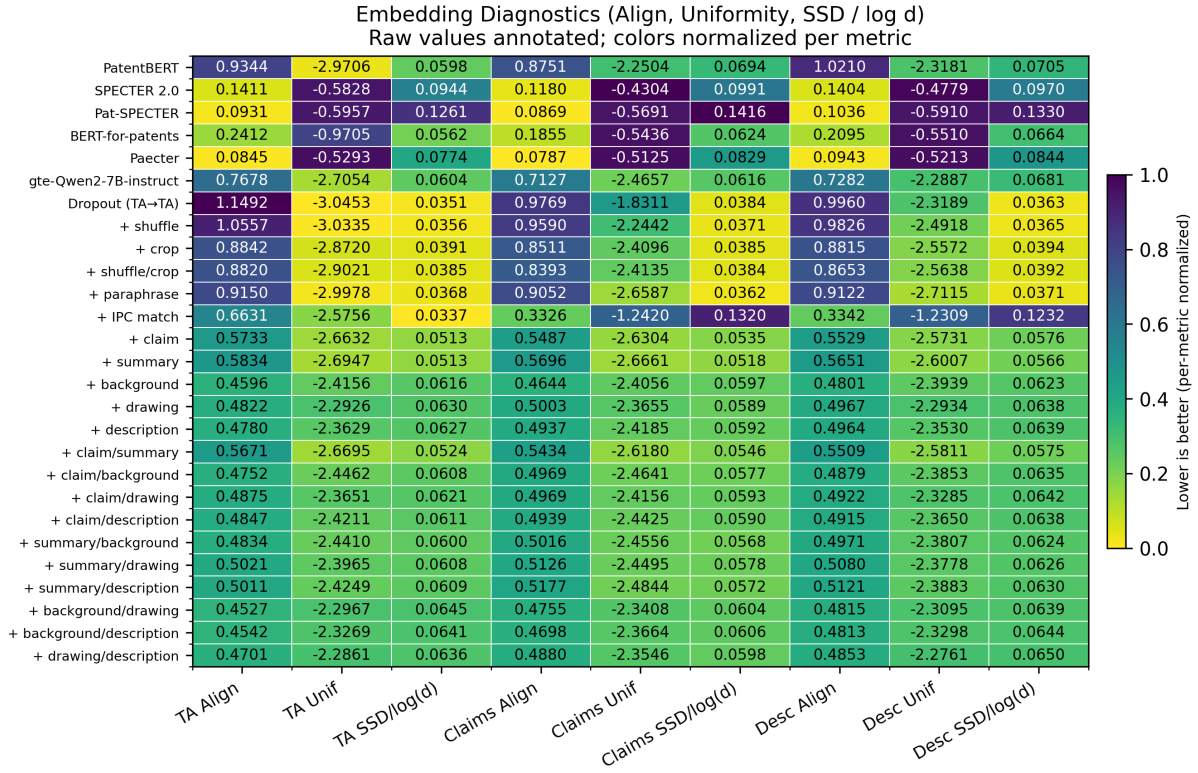

Figure 5: Embedding Space Diagnostics across three patent sections: Title+Abstract (TA), Claims, and Description. We report Alignment ↓, Uniformity ↓, and Singular Spectrum Divergence (SSD/log d) ↓. Lower values indicate better geometry.

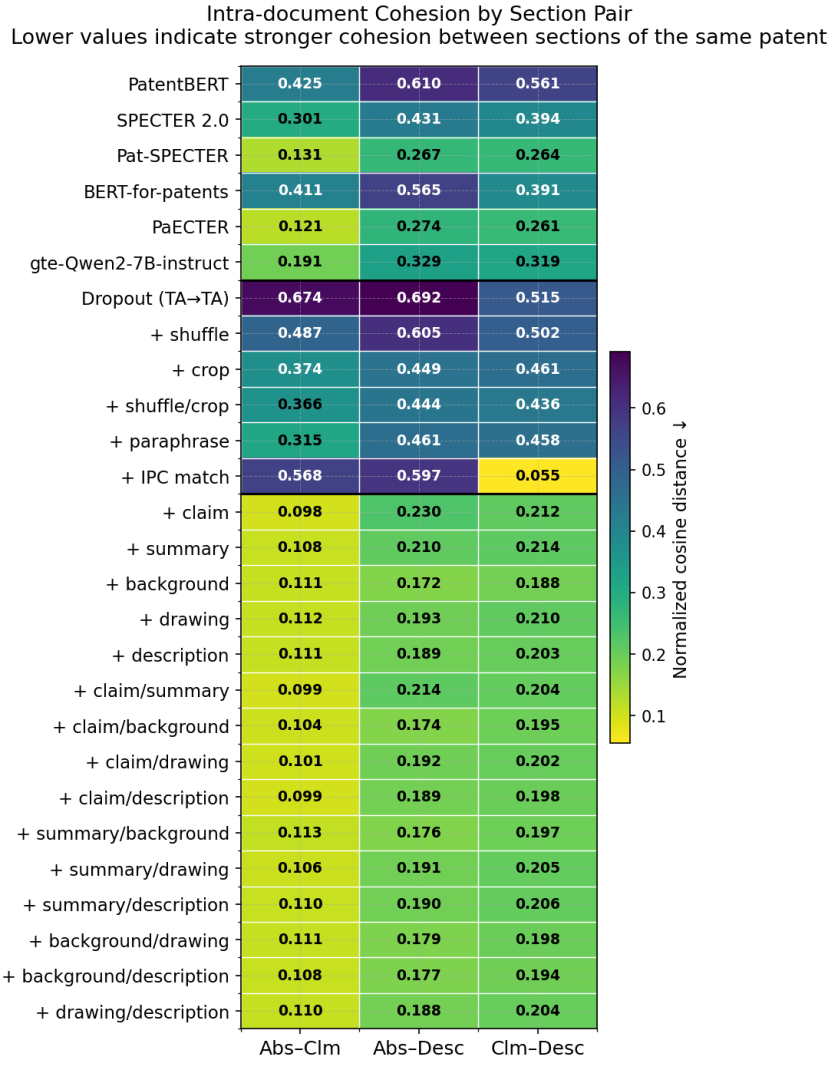

Figure 6: **Intra-document alignment ratio** across models. Lower values (↓) indicate stronger semantic cohesion between sections of the same patent.