CoDraft: Taxonomy-Grounded Enrichment for Benchmarking Ordinal Product Similarity in Intellectual Property Litigation

Anonymous ACL submission

Abstract

Evaluating product similarity in Intellectual Property (IP) litigation extends far beyond lexical matching; it requires grounding isolated, context-sparse terms within the strict NICE classification taxonomy. To address the lack of specialized benchmarks, we introduce a novel dataset of real-world EUIPO opposition proceedings, uniquely annotated on a 5-level ordinal scale of legal collision severity. To overcome context sparsity, we propose *CoDraft*, an LLM-based CoT pipeline that dynamically injects taxonomy-grounded attributes (*Nature* and *Purpose*) into bare text. Finally, we establish a robust Multi-Task Cross-Encoder baseline, jointly optimized for ordinal similarity and auxiliary NICE class identification. Experiments show that *CoDraft* effectively resolves lexical ambiguity across architectures, with our baseline achieving an 81.85% Macro-F1.

1 Introduction

Trademark similarity evaluation is fundamental in Intellectual Property (IP) litigation (World Intellectual Property Organization, 2023), encompassing visual, phonetic, and conceptual comparisons to detect legal collisions. A critical sub-task within the domain is product similarity. When comparing two products, legal examiners assess numerous facets—their physical nature and intended purposes, distribution channels, methods of use, competitive or complementary dynamics, etc. (Court of Justice of the European Communities, 1998). Synthesizing these facets, they determine the overall severity of a legal collision along a continuous ordinal spectrum.

Unlike standard Semantic Textual Similarity (STS) (Reimers and Gurevych, 2019), legal product comparison demands deep semantic congruence over mere lexical matching (Kondrak, 2005). This congruence is governed by the NICE Classification (World Intellectual Property Organization, 1957), and reflected in real-world litigation as well as official examiner resources (European Union Intellectual Property Office (EUIPO), 2026b), such as opposition proceedings at the European Union Intellectual Property Office (EUIPO). Our study leverages both, thus adhering to IP laws while improving the automated evaluation of product similarity.

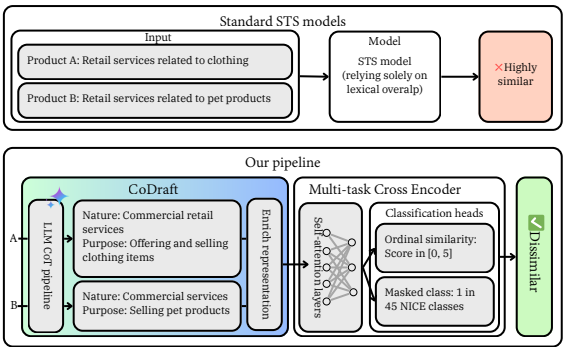

Figure 1: Lexical deception: Standard models fail on boilerplate overlaps. Through multi-task training, *CoDraft*'s taxonomy-grounded enrichment drives resolution by injecting explicit *Nature* and *Purpose* facets.

The process faces two barriers. First, existing datasets oversimplify IP litigation into binary classification (Trappey et al., 2021), ignoring the ordinal spectrum. Second, product names suffer from extreme context sparsity (e.g., "retail services for clothing" vs. "retail services for pet products", Figure 1). Standard dense retrievers, therefore, often misclassify based on boilerplate overlap rather than actual commercial targets.

To bridge these gaps, this paper establishes a new benchmark for ordinal product similarity. Our main contributions are: **(1)** We introduce a novel 5-level EUIPO *product name* similarity dataset derived from official rulings. **(2)** We focus on *Nature* and *Purpose*— foundational, static facets (Court of Justice of the European Communities, 1998) that, unlike dynamic, market-dependent counterparts, are strictly governed by the NICE taxonomy (World In-

tellectual Property Organization, 1957)—and propose **CoDraft** (Collaborative Drafting), an LLM-driven Chain-of-Thought (CoT) pipeline (Wei et al., 2022) that mitigates context sparsity by injecting these facets into isolated names. **(3)** We establish a strong Multi-Task Cross-Encoder baseline, optimized with a custom Rank-Aware Loss based on Focal Loss (Lin et al., 2017) and an auxiliary masked head for NICE-domain regularization, demonstrating the impact of token-level cross-attention on resolving IP disputes. [1]

2 Related Work

Application of NLP to the legal domain is advancing rapidly (Chalkidis et al., 2022). While some progress has been made in general Legal Judgment Prediction (LJP) (Cui et al., 2022), IP litigation remains underrepresented. Existing datasets are often limited to binary classification (Trappey et al., 2021; Nguyen et al., 2021), failing to capture the continuous ordinal spectrum of legal confusion. Moreover, prior works often neglect the semantic alignment of underlying *product names* constrained by the NICE taxonomy. We directly address this gap by introducing a novel ordinal dataset derived from official EUIPO rulings.

Evaluating product similarity is a highly specialized Semantic Textual Similarity (STS) task. Standard dense retrievers (Reimers and Gurevych, 2019) and early-interaction Cross-Encoders (Nogueira and Cho, 2019) struggle with extreme context sparsity. Static knowledge bases (Miller, 1995) are too rigid for evolving commerce. Zero-shot LLM expansion risks legally inadmissible hallucinations (Ji et al., 2023). Diverging from open-ended generation, *CoDraft* uses LLMs strictly as schema-constrained (Lu et al., 2022) reasoning engines, safely utilizing LLM world knowledge within IP law boundaries.

3 Dataset

Our dataset consists of **21,340 product name pairs**, derived directly from official EUIPO opposition proceedings under permitted research exemptions (European Union Intellectual Property Office (EUIPO), 2026a).. **Context Sparsity:** Names averaging at $\sim$**7 words** confirms the severe context sparsity inherent in IP disputes, where critical details are compressed into boilerplate text. This undermines standard dense retrievers, requiring our knowledge injection pipeline. **NICE Taxonomy Grounding:** Products are categorized under the 45-class WIPO NICE system. The data reflects the complex reality of IP law. For instance, sharing the same NICE class does not guarantee similarity (e.g., *"Gas cleaners"* and *"Refrigerating apparatus"* are both in Class 11, but legally *Dissimilar*), while cross-class oppositions are frequent. Thus, NICE acts as a legal domain constraint, rather than a trivial categorical shortcut. **5-Level Ordinal Formulation:** The 5-level ordinal scale derived directly from rulings demands that predictive models respect topological relationships across the severity spectrum, rather than merely separating bins.

Similarity Level	Pairs	Percentage
0: *Dissimilar*	11,799	55.29%
1: *Low Similar*	807	3.78%
2: *Similar*	2,718	12.74%
3: *High Similar*	388	1.82%
4: *Identical*	5,628	26.37%
Total	**21,340**	**100.00%**

Table 1: Class distribution of 21,340 pairs, maintained across 80:10:10 stratified splits.

Class Imbalance: As detailed in Table 1, the label distribution reflects real-world legal proceedings by exhibiting a severe, bipolar skew. The dataset is heavily dominated by clear-cut cases (*Dissimilar* at 55.29% and *Identical* at 26.37%), while transitional boundaries form minority classes (e.g., *High Similar* at 1.82%).

3.1 LLM-Assisted Knowledge Injection

We propose Collaborative Drafting (*CoDraft*), which replaces unconstrained zero-shot generation with a schema-constrained CoT framework (Wei et al., 2022). By prompting an LLM to perform an analytical trace before extracting structured attributes, *CoDraft* minimizes hallucinations and ensures legally accurate semantic enrichment.

Crucially, this enrichment is grounded in the NICE Classification (World Intellectual Property Organization, 1957), a globally recognized IP taxonomy structurally divided into 45 distinct classes (1–34 for goods, 35–45 for services). Rather than relying on the LLM's memory alone, we explicitly inject the official NICE **Class Headings** into the prompt context C. These headings function as rigid legal boundaries, defining the permitted material characteristics for goods or the operational domains for services. By structuring C strictly around these headings, we force the LLM to disambiguate

[1] **Reproducibility:** Code, data splits, and prompt templates will be publicly released upon acceptance.

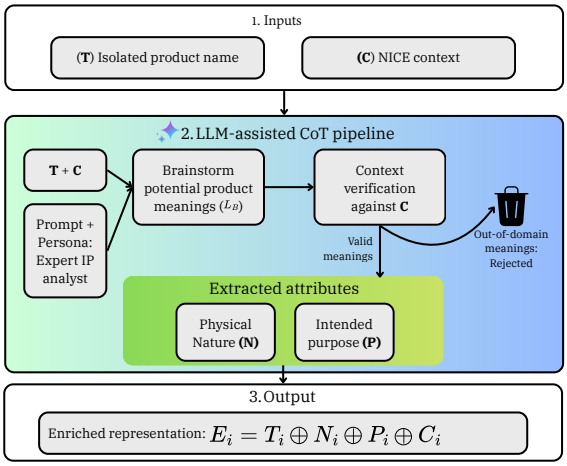

Figure 2: The CoDraft pipeline. Intermediate CoT reasoning trace allows the model to reliably extract taxonomy-grounded *Nature* and *Purpose* attributes from isolated product names.

isolated names within their class, ensuring the extracted *Nature* and *Purpose* are legally accurate and contextually bounded (the complete prompt design and schema are provided in **Appendix A**).

Casting the LLM as an expert IP Analyst, we formulate the extraction for an isolated name T_i and its NICE taxonomic context C_i as an autoregressive function:

$$(L_{B,i}, L_{V,i}, N_i, P_i) = f_{\text{LLM}}(T_i, C_i)$$

This enforces a rigid reasoning trace: **(1) Brainstorming** ($L_{B,i}$) generates unconstrained potential meanings for T_i, **(2) Context Verification** ($L_{V,i}$) cross-references $L_{B,i}$ against the specific Class Headings in C_i to explicitly reject out-of-domain meanings. Based on this boundary, the LLM extracts the *Nature* (N_i) and intended *Purpose* (P_i).

To connect to the downstream Cross-Encoder, the enriched representation E_i for name T_i is concatenated ($E_i = T_i \oplus N_i \oplus P_i \oplus C_i$). For a product pair (A, B), these strings are tokenized into the joint input sequence X:

$$X = \texttt{[CLS]} \oplus \text{Tok}(E_A) \oplus \texttt{[SEP]} \oplus \text{Tok}(E_B) \oplus \texttt{[SEP]}$$

The formulation ensures that early token-level cross-attention operates on augmented, legally bounded contexts for both product names.

3.2 Multi-Task Cross-Encoder Architecture

To capture deep semantic interactions, our Multi-Task Cross-Encoder processes the concatenated sequence X via a pre-trained multilingual Transformer. This enables token-level cross-attention across both product names and LLM-injected taxonomies. Given the contextualized hidden states $\mathbf{H} \in \mathbb{R}^{B \times L \times d}$ (batch size B, sequence length L, hidden dimension d), the network branches into two specialized classification heads:

Ordinal Similarity Head: The `[CLS]` representation ($\mathbf{h}_{\text{CLS}} \in \mathbb{R}^d$) is projected to logits $\mathbf{z}_{\text{sim}} \in \mathbb{R}^5$, yielding class probabilities $\mathbf{p} = \text{softmax}(\mathbf{z}_{\text{sim}})$. To penalize topological errors and address class imbalance, we optimize this head using a **Rank-Aware Focal Loss** ($\mathcal{L}_{\text{sim}}$). It combines a standard Focal Loss ($\mathcal{L}_{\text{focal}} = -(1-p_y)^\gamma \log p_y$) and Expected Rank penalty. Defining the expected rank as $\mathbb{E}[\hat{y}] = \sum_{c=0}^{4} c \cdot p_c$, the rank penalty is the Mean Squared Error against the ground-truth rank y: $\mathcal{L}_{\text{rank}} = (\mathbb{E}[\hat{y}] - y)^2$. The primary objective is $\mathcal{L}_{\text{sim}} = \mathcal{L}_{\text{focal}} + \alpha \mathcal{L}_{\text{rank}}$.

Masked Class Identification Head: To force the model to internalize the NICE classification rather than exploit lexical overlap, we introduce a masked domain prediction task. A random textual span of the NICE facets is replaced with `[MASK]`. Its hidden state ($\mathbf{h}_{\text{MASK}} \in \mathbb{R}^d$) is projected to logits $\mathbf{z}_{\text{class}} \in \mathbb{R}^K$ ($K = 45$, the number of NICE classes) to predict the category via Cross-Entropy loss ($\mathcal{L}_{\text{class}}$).

The joint training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sim}} + \lambda \mathcal{L}_{\text{class}}$$

where the scaling factor λ acts as a regularizer. During inference, this auxiliary head is bypassed. The final severity rank is derived directly via $\arg\max(\mathbf{z}_{\text{sim}})$, eliminating any dependency on ground-truth NICE labels.

4 Experiments

Baselines: Our Multi-Task Cross-Encoder, built upon `BAAI/bge-reranker-v2-m3` (Chen et al., 2024) is evaluated against two baselines: (1) **Lexical ML** (XGBoost via TF-IDF) to establish a lexical performance floor, and (2) **Standard Cross-Encoders** (`microsoft/deberta-v3-base` (He et al., 2021)) as a direct ablation baseline, without knowledge injection or multi-task objectives.

Evaluation Metrics: We employ **Macro-F1** to ensure equitable assessment across similarity levels, penalizing poor performance on critical but underrepresented boundary classes.

Implementation Details: For *CoDraft* extraction, we query `Gemini-2.5-Flash` (Reid et al., 2024) (temperature 0.0) to ensure deterministic formatting (see **Appendix A**). All models were

Model	Per-Class F1 Score					Macro Averages			Input Features
	0	1	2	3	4	Precision	Recall	F1	
ML Baselines									
XGBoost via TF-IDF (*vanilla*)	85.95	46.05	55.60	40.82	77.72	57.98	66.88	61.23	Vanilla names
XGBoost via TF-IDF (w/ *CoDraft*)	92.85	62.97	66.85	49.66	88.03	71.85	72.48	72.07	LLM-expanded
Standard Cross-Encoders									
Cross-Encoder (*vanilla*)	95.40	54.55	70.22	29.91	90.96	74.08	65.95	68.21	Vanilla names
Cross-Encoder (w/ *CoDraft*)	96.42	58.39	76.43	36.67	94.51	75.74	70.71	72.48	LLM-expanded
Ours									
Multi-Task Cross-Encoder (w/ *CoDraft*)	**96.60**	**72.39**	**80.40**	**65.82**	**94.02**	**81.45**	**82.27**	**81.85**	**LLM-expanded**

Table 2: Performance on the EUIPO dataset. The *CoDraft* pipeline resolves context sparsity across all baselines. Our Multi-Task Cross-Encoder sets a strong ordinal benchmark, proving highly effective for IP evaluation.

trained on a single RTX A4000 GPU. Hyperparameter configurations are in **Appendix E**.

5 Results

5.1 Main Results

Table 2 evaluates our models on the EUIPO dataset, yielding three key findings:

Universal Impact of *CoDraft*: *CoDraft* consistently boosts performance across all baselines. Strikingly, XGBoost w/ *CoDraft* (72.07% Macro F1) directly outperforms the vanilla Cross-Encoder (68.21%). It further elevates neural models, improving the standard Cross-Encoder to 72.48%, proving that taxonomy-grounded semantics are necessary for accurate IP evaluation.

State-of-the-Art Benchmark: Leveraging the *CoDraft* pipeline, our Multi-Task Cross-Encoder significantly outperforms all baselines, achieving a benchmark Macro F1 of **81.85%**. By combining ordinal ranking with end-to-end BGE features, it establishes a robust SOTA model with balanced Precision (81.45%) and Recall (82.27%).

Resolution of Hard Classes: Across all baselines, Class 3 is consistently the most challenging to predict, dropping to as low as 29.91% F1 in the vanilla Cross-Encoder. Our Multi-Task approach mitigates this bottleneck effectively, doubling the performance on this specific tier to 65.82% (supported by error analyses in **Appendix B**, qualitative case studies in **Appendix C**, and t-SNE embedding visualizations in **Appendix D**).

5.2 Ablation Studies

Table 3 quantifies the contribution of each component via systematic ablation.

CoDraft Enrichment: Removing LLM-extracted attributes drops F1 by 5.05%. This proves isolated names lack semantic density; without *Nature* and *Purpose*, the model regresses to deceptive lexical overlaps (see **Appendix D** for visualization).

Configuration	P (↑)	R (↑)	F1 (↑)
Full Model (Ours)	**81.45**	**82.27**	**81.85**
w/o CoDraft Enrichment	75.64	78.42	76.80
w/o Multi-Task Head	73.75	81.46	76.76
w/o Rank-Aware Loss	75.71	81.45	77.88

Table 3: Ablation results on the EUIPO test set. All metrics are reported as percentages (%).

Multi-Task Learning: Disabling the auxiliary masked head ($\lambda = 0$) degrades F1 by 5.09%. Forcing the model to reconstruct the masked NICE layer prevents computational shortcuts, ensuring internal representations comply with the NICE taxonomy.

Rank-Aware Optimization: Replacing our ordinal loss with standard Cross-Entropy drops F1 to 77.88%. Standard losses treat similarity tiers as isolated groups, causing severe topological misclassifications. Our loss ensures alignment with the continuous IP severity spectrum by penalizing rank deviations (see error clusters in **Appendix B**).

6 Conclusion

We addressed data scarcity and context sparsity in IP collision detection. We introduced a novel 5-level ordinal dataset derived directly from official EUIPO rulings; proposed *CoDraft*, a schema-constrained LLM pipeline that enriches isolated names with taxonomy-grounded, static attributes; and established a strong Multi-Task Cross-Encoder baseline optimized with a Rank-Aware Focal Loss, achieving an 81.85% Macro F1. Ultimately, we have built a robust, domain-regulated foundation for product similarity, and set the stage for tackling further challenges in extracting dynamic, market-dependent attributes—a necessary step in the realization of true legal reasoning.

Limitations

Inference Latency: CoDraft relies on autoregressive LLM generation, introducing latency unsuitable for real-time retrieval across massive databases. Currently, our Multi-Task Cross-Encoder operates optimally as a second-stage reranker. Future work will explore Knowledge Distillation (KD) to transfer this reasoning into a lightweight, real-time Bi-Encoder architecture.

Attribute Extraction: The extracted *Nature* and *Purpose* attributes lack a human-annotated gold standard. Consequently, we cannot strictly isolate the LLM's extraction accuracy from the downstream classification performance. While the downstream Macro-F1 improvements strongly suggest the extractions are legally useful, the LLM may still introduce subtle semantic misalignments. Future research should involve expert legal annotation to construct a dedicated evaluation set for these intermediate reasoning steps.

Legal Facets: While CoDraft successfully extracts static attributes (*Nature* and *Purpose*), real-world litigation also considers dynamic, market-dependent facets (e.g., distribution channels, competitive/complementary dynamics). Future works expanding our framework to extract these facets, however, must consider their dynamic nature, which exacerbates LLM hallucination. To that end, some similarity assessments in the raw EUIPO proceedings contain rich "Comments" written by legal examiners to justify their decision; future works can leverage these expert rationales via Retrieval-Augmented Generation (RAG) to ground LLM reasoning in real-world contexts.

Jurisdictional Bias: Our novel dataset is derived exclusively from EUIPO proceedings, inherently reflecting European legal frameworks and thresholds. Given that product similarity assessments can vary significantly across jurisdictions, future research should investigate the cross-lingual generalization of CoDraft and its adaptability to other international IP offices (e.g., USPTO, CNIPA).

Ethical Considerations

This research complies with the Legal Notices and General Terms and Conditions by the European Union Intellectual Property Office (EUIPO) regarding database access and use. The extraction and processing of the product term similarity records were conducted strictly under the permitted Text and Data Mining (TDM) exemptions for non-commercial, scientific research purposes. In adherence to the EUIPO's requirements, all dataset splits, annotations, and models derived during this study are intended strictly for open academic use, with full credit attributed to the EUIPO as the original authoritative source.

References

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Court of Justice of the European Communities. 1998. Canon kabushiki kaisha v metro-goldwyn-mayer inc. (case c-39/97). European Court Reports I-5507.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models. *arXiv preprint arXiv:2204.04859*.

European Union Intellectual Property Office (EUIPO). 2026a. Legal notices.

European Union Intellectual Property Office (EUIPO). 2026b. Similarity tool.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Grzegorz Kondrak. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Hwang, Lianhui Qin, Yejin Hwang, and Yejin Choi. 2022. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of NAACL*.

George A Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Thuy Nguyen and 1 others. 2021. Trademark similarity search using deep learning. In *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Amy JC Trappey, Charles V Trappey, Udhaya Kumar Govindarajan, Jyun-Jie Sun, and Chien-Chih Ku. 2021. An intelligent trademark similarity analysis system based on natural language processing and machine learning. *Advanced Engineering Informatics*, 48:101294.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

World Intellectual Property Organization. 1957. Nice agreement concerning the international classification of goods and services for the purposes of the registration of marks. As revised and amended.

World Intellectual Property Organization. 2023. *World Intellectual Property Indicators 2023*. WIPO.

A CoDraft Prompt Design and Schema

To enforce CoT reasoning and prevent LLM hallucination, we utilize a few-shot prompting strategy combined with a constrained JSON output schema (via Pydantic). The few-shot examples (e.g., disambiguating "Apple" and "Mercury") explicitly teach the model to reject out-of-domain meanings based on the WIPO NICE descriptions.

1. Prompt Template with Few-Shot Examples

System Instruction:
You are an expert Intellectual Property Analyst specializing in the NICE Classification system. Use 'Chain of Draft' reasoning to disambiguate product names.

User Instruction:
Analyze the following LIST of products. For each item, strictly follow the CoDraft steps:
1. Brainstorm: List potential meanings.
2. Critique: Compare with the provided NICE Class ID & Description to select the correct context.
3. Define: Specify the Nature, Purpose, and create an Expanded Name.

EXAMPLE 1 (Good Reasoning):
Input: Product Name: "Apple", Class: 31, Description: "Agricultural, aquacultural, horticultural..."
Step 1 (Brainstorm): Could be a fruit, a tech company, or a record label.
Step 2 (Critique): Class 31 covers horticultural products. Tech companies (Class 9) and labels (Class 41) are incorrect here. Correct meaning: The edible fruit.
Nature: Fresh fruit | **Purpose:** Human consumption | **Expanded Name:** Fresh apples

EXAMPLE 2 (Good Reasoning):
Input: Product Name: "Mercury", Class: 5, Description: "Pharmaceuticals, medical and veterinary..."
Step 1 (Brainstorm): Could be a planet, a chemical element (metal), or a pharma ingredient.
Step 2 (Critique): Class 5 covers medical preparations. Element mercury (Class 1) or planets (N/A) are incorrect. Correct meaning: A substance used in medicine.
Nature: Pharmaceutical preparation | **Purpose:** Medical treatment | **Expanded Name:** Mercury-based pharmaceutical preparations

---

INPUT DATA TO PROCESS: `{batch_json}`

2. Constrained JSON Output Schema

- `product_name` (String): The exact product name from the input list.
- `step1_brainstorming` (List[String]): List 3 potential meanings of the Product Name in general contexts.
- `step2_context_verification` (String): Compare the drafted meanings with the NICE Description. Explicitly reject mismatched meanings and select the correct one.
- `nature` (String): The specific physical nature of the product (Noun). Generics like 'Product', 'Item' are forbidden. (e.g., 'Chemical cleaning solution').
- `purpose` (String): The primary specific function or usage (Verb-ing phrase). (e.g., 'Removing grease from surfaces').
- `expanded_name` (String): The fully expanded, professional product name.

B Error Analysis

To understand the boundary limitations of our proposal, we conducted an error analysis based on the confusion matrix (Figure 3). The predictions reveal critical insights into both model power and the

Product Name Pair	Baseline	*CoDraft* Enrichment (Nature & Purpose)	Ours (CoDraft)
A: Gas cleaners *(Class 11)* **B:** Refrigerating apparatus *(Class 11)*	High Similar *(False)*	**A:** [Nature: Gas purifier] [Purpose: Removing impurities from air] **B:** [Nature: Cooling device] [Purpose: Lowering temperature]	Dissimilar ***(True)***
A: Depilatory preparations *(Class 3)* **B:** Creams for leather *(Class 3)*	Similar *(False)*	**A:** [Nature: Cosmetic cream] [Purpose: Removing human body hair] **B:** [Nature: Maintenance chemical] [Purpose: Conditioning leather goods]	Dissimilar ***(True)***

Table 4: Qualitative analysis of challenging EUIPO cases. Baselines fail due to shared NICE classes, whereas *CoDraft* successfully isolates the true commercial intent.

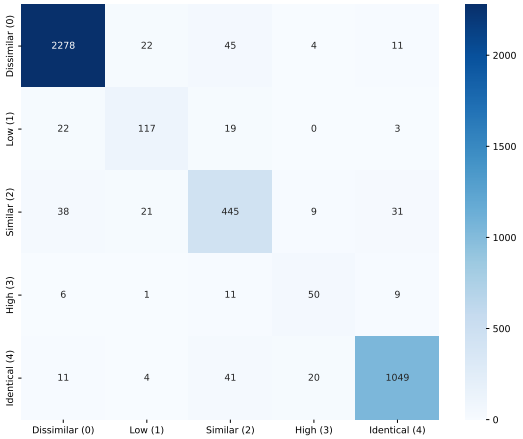

Figure 3: Confusion matrix of our LLM-Assisted Multi-Task Cross-Encoder on the test set. Errors are tightly clustered along the diagonal, validating the ordinal constraints.

inherent complexities of product similarity:

Success of Ordinal Constraints: The most prominent observation is the absence of serious ordinal errors. Out of over a thousand true *Identical* (Class 4) pairs, only 11 were misclassified as *Dissimilar* (Class 0), and vice versa (10 pairs). This proves that our Rank-Aware Focal Loss successfully penalizes topological bias, forcing the model to respect the continuous severity spectrum rather than treating classes as isolated taxonomy groups.

Adjacent Boundary Ambiguity: The vast majority of misclassifications are tightly clustered immediately adjacent to the diagonal. A representative error occurs when evaluating the pair: *"flavourings, other than essential oils, for use in electronic cigarettes"* versus *"flavourings, other than essential oils, for tobacco"*. While the ground-truth legal ruling distinguishes them as *High Similar* (Class 3) due to their distinct target markets (e-cigarettes vs. traditional tobacco), our model overestimates the semantic overlap of the core product nature, predicting them as *Identical* (Class 4). This confusion highlights the subjective nuances in IP law; the threshold between "highly similar" and "identical" is exceptionally small, emphasizing the inherent struggle in perfectly modeling human legal thresholds.

C Qualitative Case Studies

To further illustrate how the *CoDraft* pipeline resolves intra-class ambiguities, Table 4 presents specific examples from the EUIPO dataset. Standard cross-encoders often struggle when distinct products share the exact same NICE Class. By explicitly surfacing the underlying *Nature* and *Purpose*, *CoDraft* successfully disentangles these semantically diverse commercial targets.

D Embedding Space Visualization

To qualitatively validate the substantial performance leap achieved by our framework, Figure 4 visualizes the t-SNE projections of the output representation space, comparing the baseline against our proposed approach.

As illustrated in Figure 4A, the vanilla cross-encoder exhibits severe representation entanglement. Without explicit semantic boundaries, the intermediate similarity tiers (Classes 1, 2, and 3) are heavily fragmented and entangled with the extreme classes (0 and 4) due to lexical overlaps in product names. Notably, the challenging Class 3 (Highly Similar, purple points) is entirely scattered and assigned to the surrounding clusters.

However, Figure 4B demonstrates the representation space produced by our Multi-Task Cross-Encoder operating on *CoDraft* enriched inputs. The injection of explicit *Nature* and *Purpose* attributes forces the model to disentangle the vector space. The embeddings now form tightly cohesive clusters that strictly respect the ordinal legal spectrum. In particular, the previously fragmented Class 3 is now condensed into a distinct cluster, perfectly

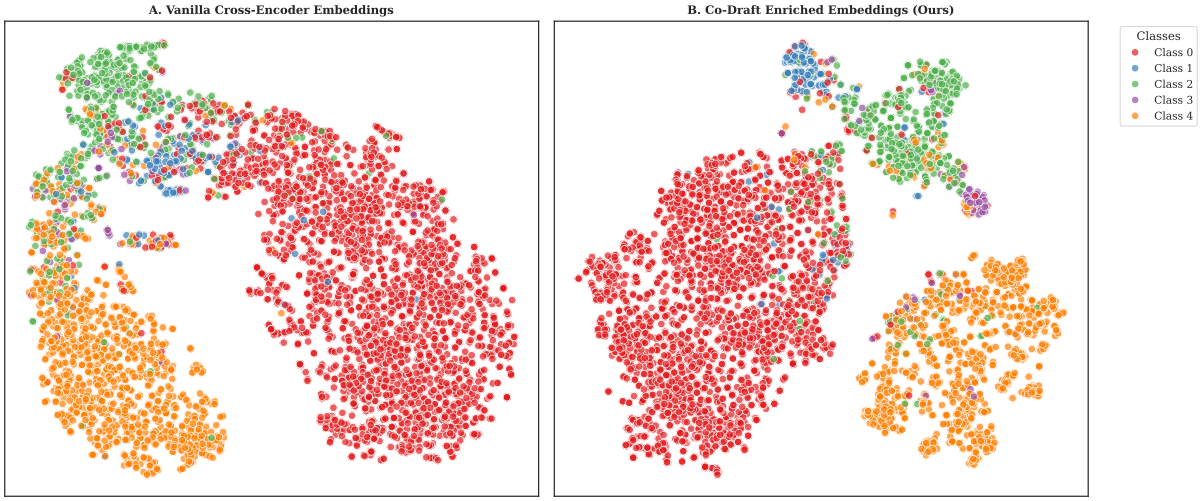

Figure 4: t-SNE projection of the representation space. (A) The baseline vanilla cross-encoder struggles with severe entanglement, failing to isolate intermediate classes. (B) Our proposed *CoDraft*-enriched Multi-Task model learns a highly structured, ordinal space, resolving the ambiguities of complex IP taxonomy.

separating ambiguous intermediate terms from the similar and identical extremes.

E Implementation Details

To ensure full reproducibility of our experiments, Table 5 outlines the complete set of hyperparameters, architectural configurations, and hardware specifications used to train our proposed multi-task cross-encoder and the traditional machine learning baselines.

Hyperparameter	**Value**
General & Text Processing	
Base Pre-trained Model	`bge-reranker-v2-m3`
Max Sequence Length	256 tokens
TF-IDF Max Features	5000 (for baselines)
Hardware	1× NVIDIA RTX A4000
Multi-Task Cross-Encoder (Ours)	
Optimizer	AdamW
Base Learning Rate	2×10^{-5}
Weight Decay	0.02
Training Epochs	10
Effective Batch Size	32 (8×4 grad. acc.)
LR Scheduler	Cosine (warmup 10%)
Loss Function	Rank-Aware Focal Loss
Multi-Task Weights	$\gamma = 2, \alpha = 0.47, \lambda = 0.26$
Mixed Precision	FP16 (Grad. Checkpointing)
Traditional ML Baselines	
XGBoost Parameters	`max_depth=6`, `lr=0.1`, `n_estimators=500`

Table 5: Detailed training configurations and hyperparameters.