Cascading Circle Loss : Order-Preserving Contrastive Learning for Graded Legal Retrieval

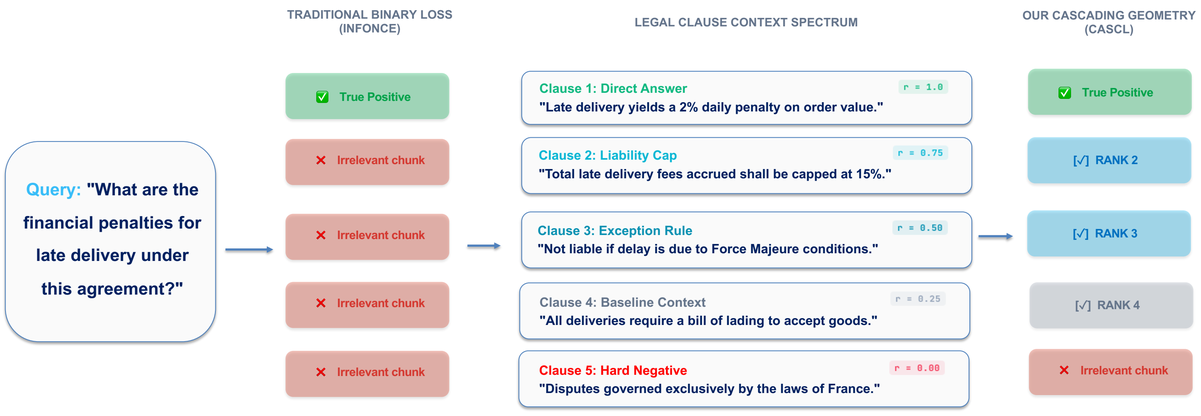
Rayane Kimouche[1] , Ahmed Touila[2]

[1] Machine Learning Research Engineer – Dilitrust , [2] Lead Machine Learning – Dilitrust

International Conference on Machine Learning (ICML) 2026 - AI4Law Workshop - Seoul, South Korea

rayane.kimouche@dilitrust.com, ahmed.touila@dilitrust.com

Introduction



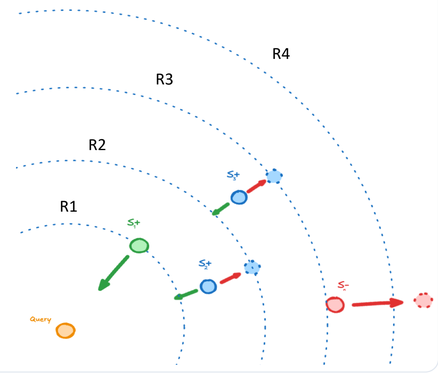
Problem Statement

Legal corpora exhibit three structural properties that make fine-grained ranking essential:

- **High lexical overlap —** statutes and contractual clauses use highly standardised language, so neighbouring paragraphs often share near-identical phrasing while encoding distinct legal conditions. Binary supervision cannot distinguish them **[9].**
- **Hierarchical relevance —** a legal topic (e.g liability, confidentiality) is rarely confined to a single paragraph but distributed across definitions, exceptions, and remedies, inducing a continuous relevance spectrum.
- **Sensitivity to misranking —** when the top of the ranked slate mixes directly applicable clauses with merely related ones, ordering errors propagate to downstream extraction, summarization and reasoning as factual errors.

Existing contrastive losses are not designed for this setting :

- **Binary losses [1,2] treat all positives identically** — they learn what is relevant but discard how relevant.
- **Ordering-aware losses [3] distribute gradient uniformly,** wasting capacity on already-satisfied constraints.
- **Multi-stage methods [4] require separate stabilisation phases** to prevent margin collapse.
- **Discrete-level approaches [5]** group different relevance degrees under the same label, with no cross-level ordering guarantee and quadratic cost.
- **Soft-target variants [6,7]** control how hard each positive is pulled but lack a stopping mechanism — the softmax drives every positive toward the query indefinitely.



Method: Cascading Circle Loss

What would be a good loss for training a legal RAG system ?

1. **Robust positive/negative separation** — excel at binary retrieval
2. **Intra-positive ordering** — higher relevance → higher similarity, with proportional gaps
3. **Top-of-ranking focus** — swapping positions 1–2 is catastrophic, swapping 10–11 is less impactful.
4. **Noise robustness** — low-relevance labels are unreliable; don't amplify them

Dataset

Corpus : Multilingual legal clauses in 6 languages (FR, PT, EN, IT, ES, DE). 150K train / 30K test examples, each filtered to contain ≥ 3 distinct relevance levels.

Labels : Human experts select the positive (r=1.0). Remaining soft positive clauses scored by 3 LLM judges (majority vote) **[11]** following a structured prompt . **Three safeguards :** human anchor, ensemble vote, label diversity filter.

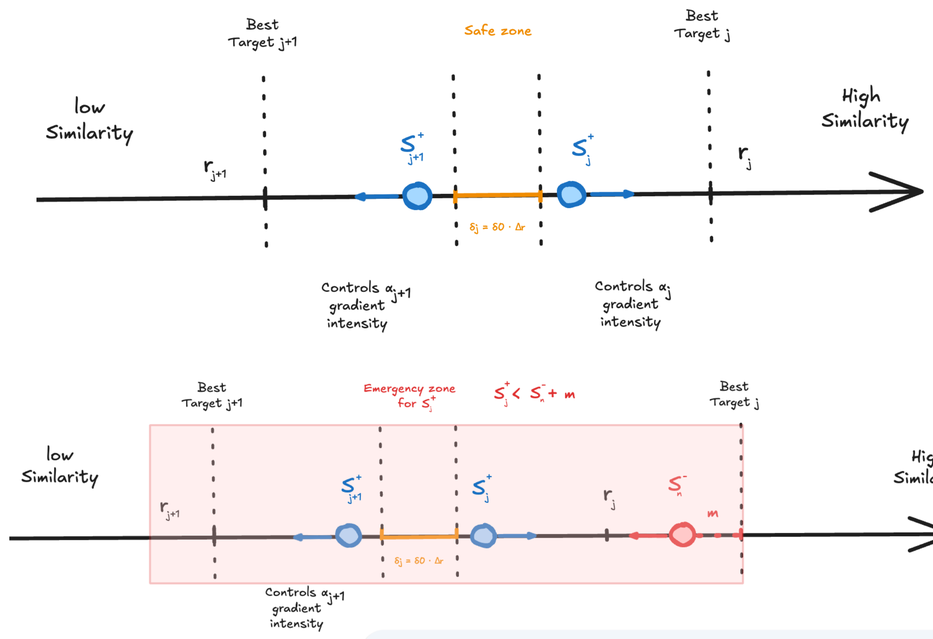
Construction : InfoNCE: positive + hard negatives (r=0). Cascade: positive + soft positives (r ∈ {0.2–0.8}) + hardest negative (highest similarity), ensuring transitivity to the rest of negatives.

$$\mathcal{L}_{CasCL} = \mathcal{L}_{InfoNCE}(s_1; \mathcal{N}, \tau) + \mu \cdot \mathcal{L}_{cascade} \qquad O(K+L)$$

$$-\log \frac{e^{s_1/\tau}}{e^{s_1/\tau} + \sum_{k\in\mathcal{N}} e^{s_k/\tau}} \qquad \sum_{j=1}^{K} w_j \log\left[1 + e^{\gamma(\alpha_{j+1}s_{j+1} - \alpha_j s_j + \delta_j)}\right]$$

Term 1 — InfoNCE (Scale-Setting) — Term 2 — Cascade (Order-Preserving)

- **Relevance Gain Weights (w_j):** Concentrates computational capacity on highly critical top-tier sorting decisions and acts as an automatic noise filter on fluid lower-tier labels:
$$w_j = 2^{r_j} - 1$$
- **Proportional Margin (δ_j):** Analytically adapts the geometric safety buffer based on the exact distance between neighboring tags, requiring zero manual grid searches:
$$\delta_j = \delta_0 \cdot (r_j - r_{j+1}) \quad \text{with} \quad \delta_0 = C$$
- **Two-Regime Self-Paced weight (α_j):** Prevents highly qualified passages from becoming stranded behind hard negatives during self-paced refinement:
$$\alpha_j = max\left([r_j - s_j]_+, [s_{neg} + m_{safe} - s_j]_+\right)$$
- **Absolute Lower Self-Paced Weight (α_{j+1}):** Penalizes low-tier snippets that drift aggressively upward towards the query:
$$\alpha_{j+1} = [s_{j+1} - r_{j+1}]_+$$



$$\frac{\partial \mathcal{L}_{cascade}}{\partial s_j} = \underbrace{+w_{j-1} * \gamma * [s_j - r_j]_+ \sigma_{j-1}}_{G_j^{\downarrow}\text{ (link above, repulses)}} \underbrace{- w_j * \gamma * \alpha_j \sigma_j}_{G_j^{\uparrow}\text{ (link below, attracts)}}$$

Experimental Results & Ablations

Model	Recall@1	MAP@10	NDCG@5	NDCG@10	NDCG@20
InfoNCE only	0.3672	0.5186	0.9402	0.9266	0.9220
CasCL (w/o InfoNCE)	0.2856	0.4464	0.9248	0.9179	0.9171
CasCL (no relevance weight)	0.3499	0.5063	0.9367	0.9292	0.9289
CasCL (γ = 15)	0.3782	0.5287	0.9428	0.9307	0.9278
CasCL (γ = 32)	0.3671	0.5205	0.9414	0.9313	0.9289
CasCL (γ = 64)	0.3577	0.5122	0.9377	0.9287	0.9270
CasCL (γ = 120)	0.3433	0.5012	0.9341	0.9255	0.9245
CasCL + SupCon **[10]**	0.3795	0.5292	0.9451	0.9351	0.9324
CasCL + neg-relative α	0.3707	0.5258	0.9444	0.9355	0.9342
CasCL + two-regime (no rel. gain)	0.3886	0.5386	0.9494	0.9378	0.9339
CasCL (no selfpaced)	0.4110	0.5567	0.9322	0.9182	0.9108
CasCL Two regime loss	**0.4114**	**0.5563**	**0.9511**	**0.9387**	**0.9360**

Figure— **Performance Comparison vs. Baselines**
CasCL vs existing contrastive losses · Qwen3-Embedding-0.6B

Figure — **Ablation : Component Impact Across All Metrics**
Key variants vs full CasCL Two-regime loss

CasCL achieves the highest score on every metric:

- **Recall@1 (0.411)** — **The system places the correct clause at position 1** more often than any baseline.
- **MAP@10 (0.556)** — **This captures whether all relevant clauses are ranked above irrelevant ones, not just the first**. A high MAP means the system doesn't just find one good result — it pushes the entire relevant set above the noise.
- **nDCG@5/10/20 (0.951 / 0.939 / 0.936)**: This is the only metric that is relevance-weighted — it rewards placing a directly applicable liability clause above a tangentially related definition, not just relevant above irrelevant. **The consistently high nDCG across all cutoffs means the cascade successfully enforces fine-grained ordering throughout the ranked slate, not just at the very top.**

No baseline manages both metric families:

- Circle Loss [2] / MNRL — strong binary separation but lag on nDCG; they know what is relevant but are blind to how relevant.
- InfoNCE [1] — pulls all positives toward uniformly high similarity, crowding the binary boundary.
- Margin MSE [12] — captures graded signal but lacks contrastive push; collapses on Recall@1 (0.364).

Every component is load-bearing:

- **Cascade (w/o InfoNCE)** — retrieves relevant clauses but in wrong order; nDCG degrades.
- **Relevance gain** — treats costly top-rank misorderings with the same urgency as invisible low-rank noise.
- **No self-pacing (α=1)** — MAP peaks (0.557) but nDCG drops; uniform gradients push everything above negatives but fail to fine-tune ordering.
- **Two-regime resolves this** : emergency gradients rescue stranded items (MAP), self-paced shutoff concentrates on violated links (nDCG).
- **γ sensitivity** — soft (γ=15) ≫ rigid (γ=120); legal text has fluid semantic boundaries.

Conclusion

- **No more binary/graded trade-off :** CasCL achieves SOTA on both metric families simultaneously without the need for two phases, preserving fine-grained semantic boundaries between relevance levels rather than collapsing them into binary decisions or forcing them into rigid discrete bins.
- **Graded positives get the right treatment :** Invisible to InfoNCE, positioned by the cascade — each converging to a similarity proportional to its relevance, with full ordering guaranteed by transitivity.
- **Gradient goes where it matters :** Relevance gain weighting focuses on costly top-rank decisions, two-regime self-pacing rescues stranded items and shuts off on solved links. No gradient wasted, none missing.

References

[1] van den Oord et al. Contrastive predictive coding. arXiv:1807.03748, 2018.
[2] Sun et al. Circle Loss. CVPR, 2020.
[3] Zhou et al. Ladder Loss. AAAI, 2020.
[4] Pitawela et al. CLOC: Contrastive learning for ordinal classification. CVPR, 2025.
[5] Xi et al. Mine and Refine. arXiv:2602.17654, 2026.
[6] Zhu et al. Generalized contrastive learning. TheWebConf, 2025.
[7] Denize et al. Similarity contrastive estimation. arXiv:2212.11187, 2022.
[8] Järvelin & Kekäläinen. Cumulated gain-based evaluation. ACM TOIS, 2002.
[9] Kim et al. Deep metric learning beyond binary supervision. CVPR, 2019.
[10] Khosla et al. Supervised contrastive learning. NeurIPS, 2020.
[11] Zheng et al. Judging LLM-as-a-Judge. NeurIPS, 2023.
[12] Hofstätter et al. Cross-architecture knowledge distillation. arXiv:2010.02666, 2020.