

Weak Supervision under Annotation Scarcity for Fine-Grained Legal Clause Classification

Christoffer Löffler^{a,*}, Andrea Martínez Freile^b

^a*School of Computer Engineering, Pontificia Universidad Católica de Valparaíso, Brasil 2950, Valparaíso, 2340025, Valparaíso, Chile*

^b*Faculty of Law, Universidad Adolfo Ibáñez, Av. Padre Hurtado 750, Viña del Mar, 2581793, Valparaíso, Chile*

ARTICLE INFO

Keywords:

Weak Supervision
Legal NLP
Multilabel Text Classification
Class Imbalance
Low-resource NLP
Terms of Service

ABSTRACT

This paper presents a leakage-safe dataset and evaluation framework for weak supervision in fine-grained abusive clause classification of Chilean Terms of Service under annotation scarcity. The corpus pairs 8,692 expert-annotated clauses from 100 companies with a decontaminated pool of 11,104 unlabeled text chunks from 191 companies. We compare strict supervision, flat self-training, sequential hard pseudo-labeling, soft teacher–student distillation, and robust-loss refinements across monolingual Spanish, multilingual, and legal-domain transformer backbones. Evaluated on the categorization stage, i.e., on clauses known to be abusive, the best of eight weakly supervised configurations improves Macro-F1 over strict supervision in all 15 backbone–category settings, by 9.99% on average and up to 17.15%. Fixed in advance, soft sequential distillation gains 7.08%. Micro-F1 declines slightly (−0.30 and −1.29 points, respectively), so the benefit concentrates on rare subtypes rather than aggregate accuracy. Soft sequential distillation is the most reliable overall default, while focal-style and generalized cross-entropy losses yield selective gains that a Friedman–Nemenyi analysis cannot distinguish from it. Weak supervision is most effective for the “DARK” and “GRAY” clause families, and weak-pool construction materially affects downstream performance. An end-to-end evaluation of the complete detect-then-categorize pipeline shows, however, that these gains are confined to the categorization stage: once detector errors are propagated, the advantage of soft distillation is no longer visible and deployed performance is bounded by detector recall. Overall, uncertainty-aware weak supervision substantially improves the fine-grained categorization of abusive clauses under annotation scarcity, while end-to-end gains depend on improving upstream detection.

1. Introduction

In the burgeoning digital economy, Terms of Service (ToS) agreements govern the complex relationships between online platforms and their users. In practice, however, these documents are often characterized by excessive length, opaque legal language, and a profound power asymmetry between providers and consumers (Lippi et al., 2019; de la Maza Gazmuri, 2003). Consequently, potentially abusive clauses, i.e., those that unfairly limit consumer rights or grant disproportionate power to the provider, frequently remain undetected by the average user (Lippi et al., 2020).

The challenge of automated detection is compounded in the context of fine-grained abusive clause analysis. Abusive contractual mechanisms are rarely articulated through simple, explicit formulations. Instead, they are often subtle, context-dependent, and linguistically heterogeneous.

These difficulties are exacerbated in low-resource and region-specific settings, such as Chilean Spanish ToS (Löffler et al., 2025, 2026). In this domain, the scarcity of expert-annotated data presents a significant bottleneck. Labeling requires specialized legal expertise, making the process both costly and slow. Currently, the available Spanish-language corpora are limited to 100 annotated ToS documents, where the distribution of clause classes is highly imbalanced (Löffler et al., 2026).

While labeled data is scarce, a vast amount of contractual text remains accessible online without annotations, suggesting a clear opportunity for weakly supervised approaches. This paper studies whether weak supervision can alleviate data scarcity in the legal domain (Freitas et al., 2026). The core premise is that labels can be propagated from models trained on a small set of expert-annotated clauses to a larger, unlabeled ToS pool, thereby expanding the training signal for fine-grained recognition. However, whether such a strategy is genuinely beneficial for legal multi-label classification at clause level remains an open question. Pseudo-label noise interacts poorly with extreme

*Corresponding author

 christoffer.loffler@pucv.cl (C. Löffler); andrea.martinez@uai.cl (A. Martínez Freile)
ORCID(s): 0000-0003-1834-8323 (C. Löffler); 0000-0003-1275-2763 (A. Martínez Freile)

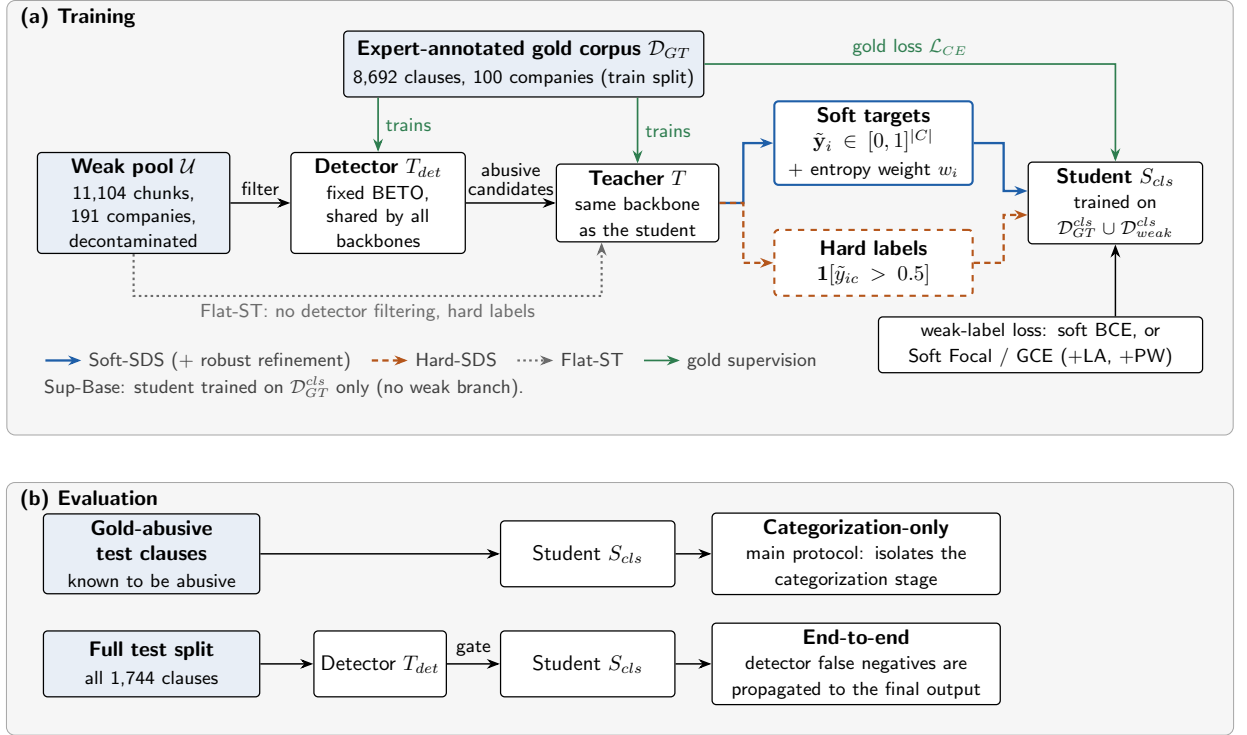


Figure 1: Overview of the weak-supervision framework. (a) Training: a fixed abusiveness detector T_{det} filters the unlabeled weak pool $\mathcal{U}$, and a backbone-matched teacher T , trained on the gold corpus, labels the retained candidates. Soft-SDS passes the teacher's probability vector together with an entropy-based confidence weight, Hard-SDS passes thresholded labels, and Flat-ST bypasses the detector and pseudo-labels the whole pool with hard labels. The student S_{cls} is trained on the union of gold and weak data. Robust-loss refinements change only the weak-label loss. Sup-Base omits the weak branch entirely. (b) Evaluation: the categorization-only protocol applies the student to test clauses known to be abusive, whereas the end-to-end setting gates the full test split with the detector: its false negatives propagate to the final output.

class imbalance (Kim et al., 2020; Wei et al., 2021; Arazo et al., 2020), overlapping subtype definitions, and the contextual nature of legal abusiveness (Duarte and Berton, 2023; Löffler et al., 2025). Moreover, observed gains may reflect genuine representation improvements, changes in calibration, or threshold shifts that favor rare labels (Kang et al., 2020).

We provide a systematic study of when and how weak supervision improves the automated classification of potentially abusive clauses in Chilean ToS. To support this investigation, we curated a clean weak-supervision pool by collecting online Terms of Service from 191 companies operating in Chile. This resulted in a final processed dataset of 11,104 unlabeled text chunks, with structural statistics that closely mirror the existing expert-annotated corpus (Löffler et al., 2026).

Through an extensive empirical evaluation, we analyze pseudo-label propagation across diverse transformer backbones, including monolingual Spanish (Cañete et al., 2023; Gutiérrez-Fandiño et al., 2022), multilingual (Devlin et al., 2019; Conneau et al., 2020), and specialized legal-domain models (Niklaus et al., 2024). Our comparison is organized as a methodological progression from a strictly supervised baseline and harder sequential pseudo-label variants to soft-label weak supervision, followed by robust-loss refinement on top of the soft-label setup. Crucially, all experiments are conducted under strict split hygiene, using a leakage-safe weak-supervision pool and held-out ground-truth evaluation to ensure no contamination occurs. Figure 1 summarizes the framework, its main variants and the two evaluation settings.

The main contributions of this paper are as follows:

- We present the first leakage-safe, cross-backbone benchmark of weak supervision for fine-grained abusive clause classification in Chilean Spanish, covering monolingual, multilingual, and legal-domain transformer models.

- We conduct a systematic comparison of weak-supervision strategies, from sequential hard pseudo-labeling to soft-label distillation and robust-loss refinement.
- We show that, within the evaluated benchmark, soft sequential distillation is the most reliable overall default strategy, whereas robust-loss methods provide selective, category- and architecture-dependent improvements rather than a universal replacement.
- We provide an ablation analysis of weak-data construction, examining how text chunking methodology affects downstream performance.

Overall, the study offers a controlled benchmark of weak-supervision design choices for low-resource, multilabel legal classification, and yields practical guidance on soft-label transfer, robust-loss refinement, and weak-data granularity.

The remainder of this paper is organized as follows. Section 2 reviews recent advances in Legal Natural Language Processing (NLP) and weak supervision for imbalanced data. Section 3 describes the ground-truth corpus and the newly curated weak-supervision pool. Section 4 details the proposed supervised and weakly supervised architectures. Section 5 outlines the training protocol and metrics. Section 6 reports our empirical findings and ablations, followed by a discussion in Section 7. Finally, Section 8 addresses the study’s limitations, and Section 9 concludes the paper.

1.1. Research Objectives

The primary objective of this study is to determine whether weak supervision can effectively alleviate the scarcity of expert-annotated data for fine-grained, multi-label classification of potentially abusive clauses in Chilean Spanish Terms of Service (ToS). In particular, the study examines whether pseudo-label propagation from models trained on a limited annotated corpus can provide a reliable additional training signal for clause-level legal classification in a low-resource setting characterized by strong class imbalance, overlapping label definitions, and context-sensitive forms of abusiveness. A second objective is to develop and validate a leakage-safe weak-supervision corpus for this task. To this end, the study constructs an unlabeled collection of ToS documents from companies operating in Chile, performs chunking to clause-level text, and assesses whether the resulting unlabeled pool is suitable for pseudo-labeling. A third objective is to systematically compare alternative weak-supervision strategies under a unified experimental protocol. Specifically, the study evaluates sequential hard pseudo-labeling, soft-label distillation, and robust-loss refinement. Then, it analyzes their relative effectiveness for improving multi-label abusive clause recognition under held-out ground-truth evaluation. A fourth objective is to examine the robustness of weak-supervision gains across different pretrained transformer families. For this purpose, the study compares monolingual Spanish, multilingual, and legal-domain models to identify whether the benefits of pseudo-label-based training remain consistent across architectures with different linguistic coverage and domain specialization. A fifth objective is to analyze when and how weak supervision improves performance in this legal NLP setting. The study investigates whether observed gains are associated with improved representations, calibration changes, or decision-threshold effects, particularly for rare labels and difficult clause categories. Finally, the study aims to identify practical design choices for weakly supervised legal text classification. In particular, it examines how the construction of weak data influences downstream performance and seeks to derive empirically grounded guidance for building reliable weak-supervision pipelines for low-resource legal NLP.

2. Related Work

This work lies at the intersection of legal NLP, weak supervision, and learning under class imbalance and label noise. We review each in turn and then position the study with respect to prior work.

2.1. Legal NLP for Unfair Clause Analysis

Transformer-based models now support a wide range of legal NLP tasks, including judgment prediction, contract analysis, and clause classification (Greco and Tagarelli, 2024). Progress has concentrated on high-resource languages and established benchmarks, leaving region-specific settings comparatively underexplored (Löffler et al., 2025). This matters for consumer-protection work in particular, because the definition of an unfair clause is fixed by national statute and case law, so a model’s target concept does not transfer across jurisdictions even when the contractual language superficially does.

Foundational work by Lippi et al. (2019) established the feasibility of automated unfair-clause detection, framing it as sentence-level classification over Terms of Service and demonstrating that supervised models can approach expert agreement on the binary task. Subsequent work has extended this along two axes. One concerns explanation: Ruggeri et al. (2022) pair clause labels with expert-written rationales, improving the interpretability of predictions at the cost of a substantially heavier annotation burden. The other concerns task formulation: Dadas et al. (2024) recast detection as retrieval-based ranking against a repository of known abusive clauses, which sidesteps some annotation cost but does not address low-resource multilabel subtype prediction. In the Chilean context, Löffler et al. (2025) formalized abusive-clause analysis as a two-stage problem, separating detection of problematic clauses from assignment of fine-grained legal subtypes, and later work grounded subtype predictions in retrieved domain evidence (Löffler et al., 2026).

The resources available to the field have broadened considerably. Recent contributions include harmful-terms datasets (Tsai et al., 2025) and multilingual privacy-policy and Terms of Service collections (Bernhard et al., 2025; Galassi et al., 2025), which begin to correct the English-centric bias of earlier corpora and make cross-lingual comparison possible. Complementary directions attack the problem from the knowledge side rather than the data side. Ontology-based reasoning represents rights and obligations explicitly so that unfairness can be derived rather than only classified (Molino-Peña et al., 2025; Confalonieri et al., 2024), while domain-specific pretraining adapts general language models to legal register and terminology (Colombo et al., 2024; Juttu et al., 2025).

For Spanish specifically, several pretrained families differ along the two dimensions that plausibly matter here, language coverage and domain specialization, and the trade-off between them is an empirical question rather than a settled one. BETO (Cañete et al., 2023) and RoBERTa-BNE (Gutiérrez-Fandiño et al., 2022) are monolingual models trained on large Spanish corpora and provide language-matched baselines. Multilingual models such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) test whether broader cross-lingual pretraining, and the greater capacity that usually accompanies it, compensates for the dilution of Spanish-specific signal. Legal-domain models such as Legal-XLM-RoBERTa (Niklaus et al., 2024) target specialized terminology directly. Prior results suggest that which family is preferable is task-dependent (Löffler et al., 2025), and our benchmark is constructed to answer this question for the present task rather than to assume it.

2.2. Weak Supervision and Pseudo-Labeling

Weak supervision covers learning from imperfect signals in place of exhaustive manual annotation: heuristic rules, auxiliary models, and pseudo-labels over unlabeled data. Distant supervision, in its original formulation, transfers labels through an external alignment or knowledge source (Mintz et al., 2009). In self-training, the supervision instead comes from the model’s own predictions. A model trained on labeled data annotates unlabeled examples and is retrained on the union. Early work established the basic procedure (Lee, 2013), and later teacher–student formulations showed that it remains effective at modern scale when combined with noise injection and capacity asymmetry (Xie et al., 2020). The central failure mode of this family is confirmation bias, in which the model’s own overconfident errors are recycled as targets and progressively reinforced (Arazo et al., 2020). This risk is heightened precisely where supervision is scarcest and where labels are ambiguous, which is the regime we study.

Knowledge distillation (Hinton et al., 2015) offers a different treatment of the same signal. Rather than committing to a thresholded decision, the student learns from the teacher’s full predictive distribution, so that the relative probability the teacher assigns to competing labels is preserved as supervision. In legally ambiguous settings, subtype assignments overlap, expert judgments differ on borderline clauses, and a teacher posterior may carry information that a hard negative destroys. Appendix A makes this argument precise for the multilabel case.

A large body of work seeks to limit the influence of unreliable weak labels by other means. Confidence-aware filtering, curriculum-style staged exposure and semi-supervised sample selection all restrict which pseudo-labels are used and when (Cascante-Bonilla et al., 2021; Zhang et al., 2021; Li et al., 2019), while complementary-label and gradient-aware objectives regularize how strongly each contributes (Ahmed et al., 2024; Zhang et al., 2024). A further line co-adapts the labeling process and the downstream model, either through meta-learning or by casting the problem as bilevel optimization (Chen et al., 2025; Li et al., 2024), and programmatic frameworks automate the construction and aggregation of labeling functions themselves (Akdog and Cicekli, 2024; Su, 2025; Peng et al., 2024). The choice of divergence between teacher and student is also consequential, since forward- and reverse-KL objectives behave differently when the teacher is unreliable, the former spreading student mass across the teacher’s support and the latter concentrating it (Gu et al., 2024; Yao et al., 2025).

A final line replaces a single pseudo-label source with agreement among several. Mixture-of-Agents annotation (Onan et al., 2025), Bayesian weak-to-strong generalization (Cui et al., 2025) and representation-level consensus (Hou et al., 2026) all pursue robustness through agreement rather than reliance on one model. The appeal is clear when annotation targets are subjective or overlapping, as they are here, but the requirement of multiple trained models or auxiliary views is a real cost in a low-resource setting. We evaluate a consensus baseline of this kind directly in Section 6.6.

2.3. Learning under Noise and Class Imbalance

Noisy and imbalanced supervision has motivated a range of robust training objectives. Generalized Cross Entropy interpolates between mean absolute error, which is robust to label noise but slow to converge, and cross-entropy, which is the reverse (Zhang and Sabuncu, 2018). Focal Loss down-weights well-classified examples so that optimization concentrates on hard or minority cases (Lin et al., 2017). Logit adjustment (Menon et al., 2020) and positive-class weighting (Cui et al., 2019) instead correct for label prevalence directly, the former by shifting decision boundaries according to estimated priors and the latter by reweighting the loss. We evaluate all four as refinements layered on soft-label supervision, since they address a source of difficulty, i.e., extreme class imbalance, that is orthogonal to the pseudo-label noise addressed by soft targets.

Beyond static reweighting, semi-supervised methods adapt pseudo-label thresholds dynamically so that rare classes are not starved of supervision by a globally calibrated cutoff (Sirbu et al., 2025; Joy and Rahman, 2026), and auxiliary balanced classifiers have been proposed to mimic a more even label distribution while still training on the full dataset (Yang et al., 2025b). Augmentation-based approaches enrich minority representations through synthetic or interpolative examples (Song et al., 2024; Yan et al., 2024; Ghosh et al., 2023), while multilabel methods exploit label structure directly through co-occurrence reranking, imbalance-aware resampling, or contrastive prototypes (Yan et al., 2025; Suyal et al., 2025; Audibert et al., 2024). Most of this literature assumes hard supervision on a fixed labeled set. In weakly supervised multilabel classification the problem is different in kind. The learner must extract what information the pseudo-labels contain without overfitting their errors, which makes robust losses a complement to soft-label transfer rather than a substitute for it. We test this expectation directly by evaluating each refinement on top of, rather than instead of, the soft-label pipeline.

2.4. Positioning

Our setting differs from most weak-supervision studies in three respects. First, the task is fine-grained abusive-clause analysis in Chilean Terms of Service, where the label space is legally specialized, severely long-tailed, and defined by statute rather than by annotator intuition. Second, the task decomposes naturally into binary detection followed by family-specific multilabel categorization, and this structure is preserved by the methods we evaluate rather than collapsed into a single flat problem. Whether that decomposition is worth preserving is one of the questions the benchmark answers. Third, weak data construction is itself part of the problem rather than a preprocessing detail, since the unlabeled pool must be decontaminated against the gold benchmark and segmented at a granularity that matches it, and we show that this choice affects downstream performance more than the loss function does. Section 7.2 returns to how these differences shape the results.

3. Data and Task Setup

This section introduces the legal and empirical setting of our study. Section 3.1 first describes the consumer-protection context in which abusive clauses arise in online ToS. We then present the expert-labeled corpus in Section 3.2 and its label taxonomy in Section 3.3, including the three clause families used throughout this work. Next, Section 3.4 formalizes the detection and categorization tasks derived from this benchmark. Finally, Section 3.5 describes the construction of the weak-supervision pool used to augment the limited expert-annotated data.

3.1. Problem Context

The analysis of online consumer contracts, such as Terms of Service, is an important legal and societal problem. These contracts regulate the relationship between digital platforms and users, yet they are often written in lengthy and highly technical language that is difficult for consumers to understand (Lippi et al., 2019; SERNAC, 2021; Löffler et al., 2025). As a result, users may agree to clauses that restrict their rights, impose disproportionate burdens, or conflict with applicable law without being fully aware of their implications (Jara Amigo, 1999; de la Maza Gazmuri, 2003).

Table 1

Class Statistics for the detection tasks of the Chilean Abusive ToS Extended dataset.

Dataset	Label	Train	Val	Test	Total
Joint	ok	4,999	719	1,441	7,159
	abusive	1,066	164	303	1,533
DARK	ok	5,642	818	1,617	8,077
	abusive	423	65	127	615
GRAY	ok	5,655	819	1,631	8,105
	abusive	410	64	113	587
ILLEGAL	ok	5,515	803	1,587	7,905
	abusive	550	80	157	787
All clauses		6,065	883	1,744	8,692

In the Chilean context, these concerns are governed primarily by Law No. 19.496 on the Protection of Consumer Rights (“Ley de Protección de los Derechos de los Consumidores”, LPC). The problem is further exacerbated by the cross-border deployment of standardized ToS, which are frequently drafted for other jurisdictions and then applied in Chile without adequate legal adaptation (SERNAC, 2021; Löffler et al., 2025). This can give rise to clauses that are incompatible with Chilean consumer law, including, for example, mandatory arbitration provisions or references to foreign legal regimes that may improperly limit consumers’ access to local protections and remedies.

Automated analysis of such contracts is therefore valuable both for large-scale legal monitoring and for supporting expert review (Dadas et al., 2024). At the same time, the task is challenging because legally problematic clauses are heterogeneous, often subtle, and not always reducible to explicit rule violations (Löffler et al., 2025). This motivates the need for datasets and models that can distinguish between directly unlawful clauses, clearly abusive but not explicitly codified practices, and legally ambiguous cases.

3.2. Expert-Labeled Corpus

We use the extended supervised ground-truth (GT) corpus introduced in prior work on abusive clause analysis in Chilean online consumer contracts (Löffler et al., 2025, 2026). The corpus consists of manually curated clauses extracted from the Terms of Service of 100 companies, with a complete company list provided in the supplemental materials. Legal experts reviewed these contracts in detail and annotated problematic clauses at the clause level.

The corpus is organized as a multi-label benchmark. Clauses may be labeled as non-abusive in the detection setting, or assigned one or more fine-grained subtype labels in the categorization setting. The benchmark groups problematic clauses into three top-level families: “ILLEGAL”, “DARK”, and “GRAY”. These families serve both as a legal taxonomy and as the basis for the evaluation setup adopted in this paper.

Tables 1 and 2 summarize the train/validation/test distributions for the detection and fine-grained categorization tasks, respectively.

3.3. Clause Families and Dataset Characteristics

The benchmark distinguishes three families of problematic contractual clauses. The ILLEGAL family contains clauses that directly conflict with Chilean legal requirements or are explicitly prohibited. The DARK family contains clauses that are clearly one-sided or abusive in practice, even when they are not unambiguously codified as illegal, such as unilateral modification or unjustified termination provisions. The GRAY family contains clauses whose legality is more context-dependent and open to interpretation, making them the most difficult to assess consistently. For a thorough legal explanation of each class we refer to Löffler et al. (2025, 2026).

These families differ not only in legal character, but also in their empirical structure, see Table 2. ILLEGAL consists of nine clause types and is highly imbalanced: two classes (“NA” and “LPC PRO”) account for roughly 60% of all labels, while the rarest classes (“ng” and “COT”) each represent only about 1% of the dataset. This makes reliable performance on minority labels especially difficult, particularly under Macro-F1 evaluation.

DARK consists of six clause types and is comparatively more balanced than the other two families, but exhibits stronger label co-occurrence. Clauses in this category often combine multiple forms of one-sided contractual behavior, leading to more complex multi-label patterns. One possible interpretation is that this family occupies an intermediate position between clearly unlawful and highly ambiguous clauses: it is less constrained by extreme rarity than ILLEGAL, yet less dependent on highly context-sensitive legal interpretation than GRAY.

Table 2

Class statistics for the classification task of the Chilean Abusive ToS Extended dataset.

Dataset	Label	Train	Val	Test	Total
ILLEGAL	NA	234	34	67	335
	LPC PRO	118	17	34	169
	RC	68	10	19	97
	LPC	56	8	16	80
	LPC INT	54	8	15	77
	LPC JUS	25	4	7	36
	acp	13	2	3	18
	ng	7	1	2	10
	COT	6	1	2	9
	Total	581	85	165	831
DARK	ltd	197	28	57	282
	cr	157	23	45	225
	nod	47	7	13	67
	ter	29	4	8	41
	er	27	4	8	39
	ch	23	3	7	33
	Total	480	69	138	687
GRAY	des risk	132	19	37	188
	des uni	130	19	35	184
	des reser	127	18	36	181
	bfe	53	8	15	76
	des def	32	5	9	46
	des det	27	4	8	39
	des inf	21	3	7	31
	des lic	15	2	5	22
	des us	10	1	3	14
	Total	547	79	155	781

GRAY consists of nine clause types and is also highly imbalanced. Five of the nine classes account individually for only between 1.8% and 5.9% of the dataset, while the remaining four majority classes contain between 9.7% and 24.1% of samples. In addition to this imbalance, GRAY is the most open to legal interpretation, as many such clauses depend heavily on contractual context and may not be uniformly assessed even by experts.

Overall, the dataset presents three central challenges for automated analysis. First, class frequencies are highly imbalanced, with a small number of frequent labels and several rare ones. Second, labels often co-occur within the same clause, especially in the DARK and GRAY families, where problematic language frequently combines several overlapping concerns. Third, because expert legal annotation is costly and time-intensive, the amount of fully labeled data remains limited relative to the complexity of the task. Figure 2 illustrates the label distribution and co-occurrence structure of the dataset.

Together, these characteristics motivate the use of multi-label modeling to capture overlap between labels, and they also motivate weakly supervised learning as a practical way to exploit additional imperfect supervision when large-scale expert annotation is not feasible.

3.4. Task Formulation

We study abusive clause analysis in Chilean Terms of Service under a two-stage setting. The first task is binary abusive clause detection, in which each clause is classified as either “Okay” or “Abusive”. The second task is fine-grained abusive clause categorization, in which only abusive clauses are considered and may receive one or more subtype labels.

Following the organization of the benchmark, the categorization task is analyzed separately for the three clause families ILLEGAL, DARK, and GRAY. This decomposition is consistent with prior work on unfair-clause analysis, where the screening of potentially unfair clauses is separated from the assignment of more specific legal

Weak Supervision under Annotation Scarcity

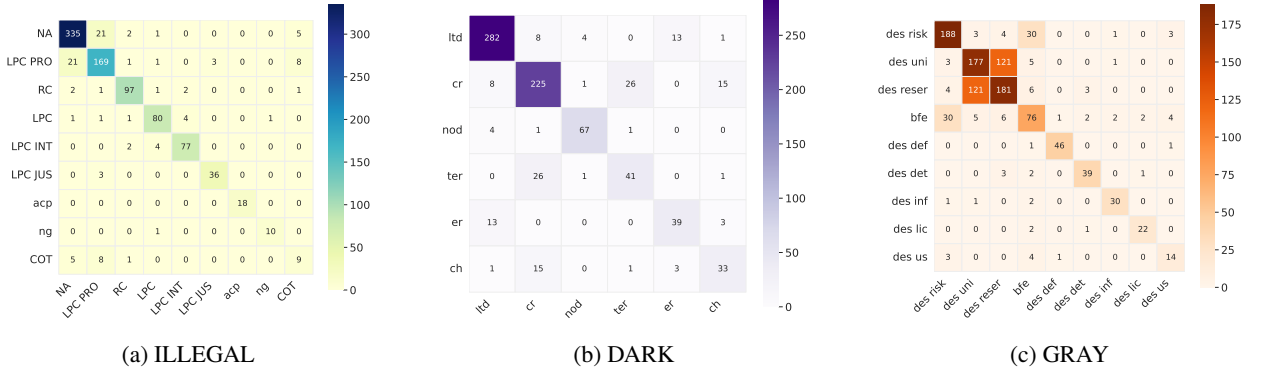


Figure 2: Label co-occurrence matrices for the multi-label classification tasks. Diagonal elements represent the total count of each label, while off-diagonal elements show the frequency of label pairs appearing in the same sample.

Table 3

Weak-supervision pool filtering breakdown.

Filter Stage	Documents	Chunks	% of Total
Initial Raw Pool	251	15,785	100.0%
Manually filtered	18	3,276	20.8%
Noise Classifier	42	853	5.4%
Exact GT Overlap	—	108	0.7%
Substring GT Overlap	—	444	2.8%
Final Weak pool	191	11,104	70.3%

categories (Lippi et al., 2019; Löffler et al., 2025). It also reflects the fact that these families differ substantially in legal interpretation difficulty, label imbalance, and co-occurrence structure.

Our main weak-supervision pipeline preserves this two-stage formulation by using clause detection as an upstream filtering step for subtype categorization. In addition, we evaluate flat self-training variants that do not preserve this structure, allowing us to assess whether task decomposition is beneficial in the weak-supervision setting.

3.5. Weak-Supervision Pool Construction

Because expert legal annotation is expensive and the GT corpus is relatively small, we construct a weak-supervision pool from additional Chilean ToS documents. This pool serves as the unlabeled corpus for pseudo-label propagation and self-training. Our goal is to provide additional in-domain supervision while minimizing distribution mismatch and preventing leakage into the gold benchmark.

3.5.1. Document Acquisition and Quality Filtering

The collection process began with a target set of 251 companies maintaining a commercial presence in Chile. Following the initial automated collection, we applied a three-stage filtering process that resulted in 191 usable ToS documents, see Table 3.

The steps are as follows:

1. Manual filtering: we removed 18 empty, corrupt, or malformed PDF/HTML files, as well as obviously irrelevant documents such as landing pages and job advertisements, leaving 233 documents.
2. Automated noise filtering: to further prune low-quality data, we implemented a lightweight quality-control classifier. Using five manually identified noise documents, such as landing pages that contain no ToS content, as a seed set, we trained a logistic regression model with TF-IDF features to detect and remove an additional 42 documents containing fragmented or non-contractual text.
3. Chunk-level overlap filtering: To prevent data leakage, we performed decontamination steps using literal, Unicode-normalized, and substring-level overlap matching against the GT chunks. This process identified and

Table 4

Statistics of the ground-truth (GT) labeled corpus and the curated weak-supervision pool.

Dataset / Split	Companies	Chunks	Mean	Median
Ground Truth	100	8,692	64.9	50
Train	—	6,065	65.8	51
Val	—	883	62.1	47
Test	—	1,744	63.3	49
Weak Pool	191	11,104	46.6	42

removed 108 exact matches and 444 overlapping chunks, resulting in a final evaluation-safe pool with zero remaining literal text intersections with the GT corpus.

The final weak-supervision pool consists of 11,104 high-quality, legally relevant text chunks drawn from 191 companies. We refer to the supplemental materials for details on the company list. This corpus provides a realistic basis for studying pseudo-label propagation in a low-resource legal setting.

3.5.2. Distribution-Aligned Chunking Strategy

Because pseudo-label quality is sensitive to distribution mismatch and noisy predictions (Arazo et al., 2020), an important challenge in our setting is ensuring that the teacher model is applied to text segments comparable in scope and complexity to those seen during supervised training. To reduce noise from granularity mismatch, we designed a two-stage, distribution-aligned chunking strategy.

1. Structural extraction: we used the `unstructured`¹ library with Spanish language support to partition documents into logical units. To preserve the semantic integrity of legal headers and sections, we employed a chunk-by-title strategy with permissive thresholds of up to 1,500 characters, avoiding arbitrary character-level splits that might sever clause context.
2. Semantic refinement: Because structurally extracted paragraphs may remain excessively long, we applied a sentence-aware re-splitting pass using `spaCy`’s `es_core_news_sm` model². Any chunk exceeding 100 words was recursively split at sentence boundaries and re-aggregated to target the GT median length of approximately 50 words.

This alignment is verified by the resulting corpus statistics. The weak-supervision pool features a mean length of 46.6 words (median 42), closely approximating the ground truth dataset’s mean of 64.9 words (median 50). This corresponds to a Kolmogorov–Smirnov (KS) (Massey, 1951) statistic of 0.15. A purely structural paragraph-level segmentation without re-splitting resulted in chunks 2.3x longer than the GT, or a KS statistic of 0.54, significantly degrading classification performance as our ablations show in Section 6.4.

4. Methodology

The central challenge in Chilean ToS analysis is the scarcity–imbalance bottleneck: expert-annotated labels for fine-grained abusive subtypes are rare and highly imbalanced, yet unlabeled text is abundant. We address this by integrating distant supervision, teacher–student learning, and noise-aware optimization into a common weak-supervision framework. Our methodological focus is not on a single monotonic sequence of architectures, but on a progression from stricter and harder pseudo-label formulations toward soft-label weak supervision, followed by selective robust-loss refinement. Figure 1 gives an overview of the framework, showing the training pipeline with its main variants and the two evaluation settings used in this paper.

4.1. Problem Formalization and Notation

Let $\mathcal{D}_{GT}^{det}$ denote the ground-truth detection set and $\mathcal{D}_{GT}^{cls}$ the ground-truth categorical set for abusive subtypes. We further assume access to a large, decontaminated unlabeled pool $\mathcal{U}$ of Chilean ToS fragments. Our objective is to

¹<https://unstructured.io/>, version 0.22.18

²<https://spacy.io/>, version 3.8.14

optimize a student classifier S using an expanded categorical training set

$$\tilde{D}^{cls} = D_{GT}^{cls} \cup D_{weak}^{cls},$$

where the weak-label set D_{weak}^{cls} is generated by a teacher model T applied to $\mathcal{U}$.

We begin from a strict supervised reference point, denoted Supervised Baseline (Sup-Base). This configuration follows the two-stage task structure of abusive clause analysis: a binary detector T_{det} identifies abusive clauses, and a categorical classifier S_{cls} , trained only on D_{GT}^{cls} , predicts subtype labels. Sup-Base provides the expert-only reference corresponding to Chilean annotated data alone.

To distinguish the effects of weak data from those of task decomposition, we also define a harder single-stage control, denoted Flat Self-Training (Flat-ST). This configuration collapses the pipeline into a direct pseudo-labeling setup (Lee, 2013; Xie et al., 2020) by propagating labels across $\mathcal{U}$ without detector-based filtering. Flat-ST therefore tests whether gains stem merely from weak-data exposure or from preserving the detection–classification hierarchy.

A second harder variant is Hard Sequential Distant Supervision (Hard-SDS). Hard-SDS preserves the sequential structure of the task, but propagates discrete pseudo-labels from the teacher to the student. Comparing Hard-SDS against Flat-ST isolates the effect of task decomposition under weak supervision, while comparing Hard-SDS against soft-label methods isolates the effect of label form. Because Hard-SDS reduces teacher outputs to thresholded decisions, it remains vulnerable to confirmation bias and pseudo-label error reinforcement (Arazo et al., 2020). For this reason, we treat it as an informative sequential baseline rather than the preferred default.

4.2. Soft Sequential Distant Supervision

Our main weak-supervision method is Soft Sequential Distant Supervision (Soft-SDS). Soft-SDS retains the sequential teacher–student structure, but replaces hard pseudo-labels with soft probability targets under a knowledge-distillation view (Hinton et al., 2015). Instead of training on thresholded teacher decisions, the student learns from the teacher’s full multilabel probability vector over the retained weak instances. This is particularly appropriate in legal clause analysis, where subtype assignments may be ambiguous, overlapping, or weakly expressed.

The optimization objective is

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{soft} \mathcal{L}_{soft}, \quad (1)$$

where $\mathcal{L}_{CE}$ denotes the supervised loss on ground-truth instances from D_{GT}^{cls} and $\mathcal{L}_{soft}$ denotes the confidence-weighted loss on weak soft targets from D_{weak}^{cls} .

For each weak example $x_i \in D_{weak}^{cls}$, the teacher provides a soft multilabel target $\tilde{y}_i \in [0, 1]^{|C|}$, where C is the set of subtype labels in the corresponding task. We additionally assign each weak example a scalar confidence weight $w_i \in [0, 1]$ derived from the uncertainty of $\tilde{y}_i$, so that sharper teacher predictions contribute more strongly and uncertain pseudo-labels exert less influence (Kendall and Gal, 2017; Sohn et al., 2020). We therefore define

$$\mathcal{L}_{soft} = \frac{\sum_{x_i \in D_{weak}^{cls}} w_i \text{BCE}(S_{cls}(x_i), \tilde{y}_i)}{\sum_{x_i \in D_{weak}^{cls}} w_i}, \quad (2)$$

where $\text{BCE}(\cdot, \cdot)$ denotes binary cross-entropy with soft targets. Because the task is multilabel rather than multiclass, we model each label as an independent soft binary decision and therefore use BCE with soft targets as the primary distillation objective. This can be interpreted as an independent-label analogue of distribution matching, while avoiding the simplex constraints of multiclass KL-based distillation.

The confidence score is computed from the average binary entropy of the teacher probabilities:

$$w_i = 1 - \frac{1}{|C| \log 2} \sum_{c \in C} \left[-\tilde{y}_{ic} \log \tilde{y}_{ic} - (1 - \tilde{y}_{ic}) \log(1 - \tilde{y}_{ic}) \right]. \quad (3)$$

Thus, weak examples whose teacher probabilities are close to 0 or 1 receive larger weights, whereas examples with many near-uncertain label probabilities are down-weighted. In this way, Soft-SDS combines structure-preserving weak supervision with uncertainty-aware soft-label learning. Throughout the paper, we treat Soft-SDS as the main default weak-supervision method against which harder variants and refinement strategies are compared.

The choice of soft over hard targets is not only an empirical preference. Because the gradient of binary cross-entropy with respect to a logit is the signed discrepancy between prediction and target, thresholding a teacher posterior $\hat{y}_c$ perturbs that gradient by $\min(\hat{y}_c, 1 - \hat{y}_c)$, i.e., the teacher's own estimate of its per-class error probability. This offset is largest where the teacher is least certain, and it is systematic rather than stochastic, so it accumulates instead of averaging out. Thresholding is also not neutral with respect to class frequency, since a diffuse posterior over a rare subtype rarely clears the threshold at all. Appendix A derives both properties, and Section 6.4.1 reports the measurements that confirm them.

4.3. Robust-Loss Refinements

Although Soft-SDS reduces some of the brittleness of hard pseudo-label propagation, pseudo-label noise and class imbalance remain central challenges. We therefore study a family of robust-loss refinements that preserve the same sequential soft-label pipeline while modifying the weak-label objective.

These refinements keep the teacher–student structure and weak-data construction of Soft-SDS fixed, but replace or augment the weak-label loss with objectives designed to improve robustness under noisy and imbalanced supervision. They are not introduced as universal replacements for Soft-SDS. Instead, they are treated as refinement mechanisms that may provide additional gains in particular clause families or backbone architectures.

Because abusive clause categorization is a multilabel problem, we model each class $c \in C$ as an independent binary decision. Let $\hat{y}_c = \sigma(z_c)$ denote the student's predicted probability for class c , derived from logit z_c , and let $p_c \in [0, 1]$ denote the corresponding target, which may be either a hard ground-truth label or a soft teacher probability. We define the target-aligned probability

$$p_t = p_c \hat{y}_c + (1 - p_c)(1 - \hat{y}_c), \quad (4)$$

which is the probability that the student assigns to the outcome indicated by the target, taken in expectation over the target when p_c is soft. For a hard target, $p_c \in \{0, 1\}$, p_t equals $\hat{y}_c$ or $1 - \hat{y}_c$ and approaches 1 as the prediction agrees with the label. For a soft target with $0 < p_c < 1$, p_t is affine in $\hat{y}_c$ and bounded above by $\max(p_c, 1 - p_c) < 1$. It cannot reach 1 for any prediction. We discuss the consequences below. For reference, the per-class BCE term is

$$\mathcal{L}_{BCE}^{(c)} = -[p_c \log(\hat{y}_c) + (1 - p_c) \log(1 - \hat{y}_c)]. \quad (5)$$

Focal Loss (FL) (Lin et al., 2017) modifies BCE by down-weighting easy examples and concentrating learning on harder or minority cases. To adapt it to soft pseudo-labels, we use

$$\mathcal{L}_{SoftFocal} = \sum_{c \in C} (1 - p_t)^\gamma \cdot \mathcal{L}_{BCE}^{(c)}, \quad (6)$$

where γ is the focusing parameter. In our experiments, we evaluate $\gamma \in \{1.0, 1.5\}$. For hard targets this recovers the standard focal behavior. As the student prediction approaches the label $p_t \rightarrow 1$, the modulating factor $(1 - p_t)^\gamma$ suppresses the loss, hence, harder mismatches receive larger updates.

For soft targets the modulation behaves differently, because $p_t \leq \max(p_c, 1 - p_c) < 1$ whenever $0 < p_c < 1$. The modulating factor then never vanishes and the loss is not driven to zero, not even when the student reproduces the target exactly. At $\hat{y}_c = p_c$, that is, the minimizer of the BCE term, $p_t = p_c^2 + (1 - p_c)^2$, which equals 0.5 for $p_c = 0.5$ and approaches 1 only as p_c approaches 0 or 1. In the limiting case $p_c = 0.5$ of a maximally uncertain target, the factor is $(1 - p_t)^\gamma = 2^{-\gamma}$ for every prediction. This means that the focal term reduces to a constant down-weighting of the BCE for that class.

More generally, for a soft target the modulating factor ranges over $[\min(p_c, 1 - p_c)^\gamma, \max(p_c, 1 - p_c)^\gamma]$ as $\hat{y}_c$ sweeps from one extreme to the other, and this range narrows to a point as $p_c \rightarrow 0.5$. Soft Focal Loss therefore concentrates the learning signal on mismatches only where the teacher is confident, and acts as a bounded, target-dependent down-weighting where the teacher is uncertain. This is complementary to the instance-level weight w_i of Eq. (3), which down-weights uncertain weak instances as a whole, whereas the focal factor adds a per-class modulation that is weakest exactly where the soft target carries the least information.

The same upper bound on p_t applies to the Generalized Cross Entropy (GCE) (Zhang and Sabuncu, 2018) term, whose per-class loss is likewise bounded away from zero for uncertain soft targets. GCE is designed to reduce the influence of extreme outliers and mislabeled examples. In our multilabel soft-target setting, it is defined as

$$\mathcal{L}_{GCE} = \sum_{c \in C} \frac{1 - p_t^q}{q}, \quad (7)$$

Table 5

Main method configurations evaluated in this study.

Method	Labels	View Sep.	Design Goal
Sup-Base	None	Yes	Supervised baseline
Flat-ST	Hard	No	Hard-label control
Hard-SDS	Hard	Yes	Sequential hard-label baseline
Soft-SDS	Soft	Yes	Soft-label weak supervision
Soft-SDS + Robust Loss	Soft	Yes	Loss refinement

where $q \in (0, 1]$ controls the degree of noise tolerance. We evaluate GCE at $q = 0.7$.

Optionally, we refine these losses via two alternative mechanisms. First, Logit Adjustment (LA) (Menon et al., 2020) addresses class imbalance by shifting logits according to class prevalence before computing prediction probabilities. Let π_c denote the empirical prevalence of class c in the training set. We define

$$z'_c = z_c - \tau \log(\pi_c), \quad (8)$$

where $\tau = 1.0$ in our experiments. Predictions are then computed as $\hat{y}_c = \sigma(z'_c)$, after which BCE or Soft Focal Loss is applied.

Second, we implement imbalance-oriented refinement via positive class weighting (PW) (Cui et al., 2019). Let $\pi_c = \frac{N_c^+}{N_c}$ denote the empirical positive prevalence of class c , and define

$$w_c = \frac{1 - \pi_c}{\pi_c}. \quad (9)$$

The corresponding weighted BCE becomes

$$\mathcal{L}_{PW} = - \sum_{c \in C} [w_c \cdot p_c \log(\hat{y}_c) + (1 - p_c) \log(1 - \hat{y}_c)]. \quad (10)$$

In combined formulations such as Soft Focal + PW, this weighting is applied within the BCE component before the focal modulation.

Taken together, these refinements define a family of optimization-level modifications layered on top of the Soft-SDS pipeline. Their role is to test whether category-specific refinement can improve over the strongest overall default. Table 5 summarizes the main configurations evaluated in this study. The principal comparison is between the supervised baseline, harder pseudo-label variants, and the Soft-SDS default, with robust-loss methods evaluated as refinements. Figure 1(a) marks the path that each configuration takes through the training pipeline.

5. Experimental Design

This section details the experimental framework used to evaluate the methodological choices introduced in Section 4. Our experiments are designed to test whether weak-supervision gains generalize across language specialization, model scale, and legal-domain pretraining, and to determine which form of pseudo-label integration provides the strongest overall default for low-resource legal clause classification. Accordingly, we organize the evaluation around four primary methodological axes:

1. **Supervised vs. Weakly Supervised Learning:** Comparing the expert-only supervised baseline (Sup-Base) against sequential and non-sequential weak-supervision variants.
2. **View Separation vs. Direct Self-Training:** Assessing the value of preserving the detection–classification hierarchy by comparing sequential methods against the flat self-training control (Flat-ST).
3. **Harder Variants vs. Soft-Label Weak Supervision:** Evaluating whether soft-label sequential distillation (Soft-SDS) improves over harder pseudo-label formulations such as Flat-ST and Hard-SDS.
4. **Soft Supervision vs. Robust-Loss Refinement:** Measuring whether robust-loss objectives layered on top of Soft-SDS provide additional category-specific or architecture-dependent gains.

5.1. Backbone Models

We select a diverse set of transformer encoders to assess the robustness of our weak-supervision strategies across different pretraining regimes. The main text focuses on five representative backbones:

- BETO (Cañete et al., 2023) is a monolingual Chilean Spanish BERT (Devlin et al., 2019) model serving as our primary language-matched reference.
- RoBERTa-BNE base (Gutiérrez-Fandiño et al., 2022) is a monolingual Spanish model from a different pretraining family than BETO, used to verify that findings generalize across Spanish corpora.
- RoBERTa-BNE large is its higher-capacity variant, included to separate the effect of pretraining family from that of model scale.
- XLM-RoBERTa Large (Conneau et al., 2020) is a high-capacity multilingual model used to test if scale can compensate for language specialization.
- Legal-XLM-RoBERTa (Niklaus et al., 2024) is a multilingual model pretrained on legal corpora, targeting domain-specific lexical patterns.

For all experiments we use a single fixed BETO detector, chosen for its strong and reliable performance on this task (Löffler et al., 2025, 2026). Because the detector is shared, every backbone receives an identical set of weak-pool candidates, such that differences between backbones cannot be attributed to differences in which unlabeled clauses were selected. For the categorical stage, however, each backbone serves as its own teacher. Teacher and student share an architecture and differ only in their training data.

This deviates from the common practice of distilling from a larger teacher, and the choice is deliberate. The independent variable in this study is the backbone, and the quantity of interest is the effect of the form of the weak label (hard versus soft, sequential versus flat). A single strong teacher shared across all students would confound these two factors, because observed differences would then reflect the capacity gap between that fixed teacher and each student rather than the properties of the weak-supervision strategy. Backbone-matched pairing holds the teacher’s inductive bias, tokenization and pretraining distribution fixed relative to the student, which isolates the label-form effect we set out to measure.

A second consideration is specific to the low-resource regime. The teacher is trained on 550, 423 and 410 gold clauses for ILLEGAL, DARK and GRAY respectively. At this scale, teacher quality is bounded by the available supervision rather than by parameter count, so a larger teacher does not straightforwardly imply a better teacher. We tested this directly with a cross-architecture control in which an XLM-RoBERTa large teacher supervises a BETO student under otherwise identical conditions. Averaged over three categories and five seeds, the stronger teacher changes Macro-F1 by +0.32 points (95% CI $[-1.88, +2.66]$, paired t -test $p = 0.79$, winning in 8 of 15 runs), which is indistinguishable from no effect. Backbone-matched pairing therefore costs nothing measurable here, and we retain it for the experimental-control reason given above.

The additional robustness backbones mDeBERTa-v3 (He et al., 2022) and RoBERTalex (Gutiérrez-Fandiño et al., 2021) are reported in the supplemental materials.

5.2. Evaluation Protocol and Hyperparameters

To ensure the rigorous assessment of model performance and the reproducibility of our findings, we adopt a standardized evaluation and training protocol. All categorical results are computed exclusively on the ground-truth test splits of the categorical tasks (categorization-only setting, Figure 1(b)). By evaluating only on known abusive clauses, we ensure that fine-grained subtype errors are not masked by the high frequency of non-abusive samples. We utilize three primary metrics to capture different facets of model performance. First, Detection F_1 measures the binary effectiveness of the initial detection stage. For categorical performance, we report Micro- F_1 and Macro- F_1 . Let TP_c , FP_c , and FN_c denote the true positives, false positives, and false negatives for a specific abusive subtype $c \in C$. The metrics are defined as:

$$\text{Micro-}F_1 = \frac{\sum_{c \in C} 2 \cdot TP_c}{\sum_{c \in C} (2 \cdot TP_c + FP_c + FN_c)} \quad (11)$$

$$\text{Macro-}F_1 = \frac{1}{|C|} \sum_{c \in C} \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \quad (12)$$

While Micro- F_1 provides an aggregate measure of instance-level reliability, Macro- F_1 is our primary metric of interest, as it assigns equal weight to rare but legally critical subtypes, such as the rarest ILLEGAL clause types.

Regarding training, all models are trained under a unified regime to facilitate a fair comparison across backbone architectures. We employ a learning rate of 3×10^{-5} and an effective batch size of 32, realized with gradient accumulation. Training is conducted for a maximum of 50 epochs, using early stopping with a patience of 20 epochs based on validation Macro- F_1 to prevent overfitting while ensuring convergence on rare labels. We report mean $\pm$ sample standard deviation across five independent runs initialized with different random seeds. Every comparison on which we base a claim is additionally accompanied by a paired significance test, a multiplicity correction and an effect size, as set out in Section 6.5.

All weak supervision experiments maintain strict split hygiene. Teacher models are selected based on validation performance within the ground-truth splits, and the test set remained strictly untouched until final evaluation. The unlabeled pool $\mathcal{U}$ is decontaminated to remove overlaps with any ground-truth data before propagation. In the sequential SDS variants, the teacher identifies candidate abusive clauses in the weak pool, which are then labeled as either hard pseudo-labels or soft probability targets before being merged into the student’s training set.

5.3. Ablation Studies

A critical design decision in distant supervision involves the structural granularity, i.e., the chunking strategy, of the unlabeled instances in $\mathcal{U}$. Legal contracts typically consist of long, nested paragraphs, and the chunking strategy may strongly affect weak-supervision performance. To ensure that the performance gains are not negated by a structural distribution mismatch or shift, we ablate two distinct chunking strategies for the unlabeled pool.

We compare our distribution-aligned chunking with paragraph-level segmentation without the semantic refinement in Section 6.4. As part of our experimental validation, we test the sensitivity of Soft-SDS variants to this alignment. This comparison allows us to determine whether distributional alignment reduces the inherent noise mismatch and improves the stability of pseudo-label propagation.

6. Results

We organize the results around the methodological questions introduced in Section 5. First, Section 6.1 compares our weakly supervised training regime to strong baselines. Next, Section 6.2 examines whether weak supervision improves over strict supervision. Section 6.3 then shows the effects of noise robust loss functions and their extensions on different tasks and backbones, and Section 6.4 reports the ablations on the architecture components, i.e., task decomposition and label form, as well as on the weak-data construction. Section 6.5 tests the statistical significance of the main comparisons, Section 6.6 compares Soft-SDS against external weak-supervision baselines under a common protocol, Section 6.7 evaluates the complete two-stage pipeline with error propagation, and Section 6.8 analyzes pseudo-label quality and teacher sensitivity.

6.1. Competitive Performance Analysis

This section tests whether weakly supervised training outperforms not only standard fine-tuning, but also stronger LLM-based baselines that are prompted via few-shot learning or Retrieval-Augmented Generation (RAG). To demonstrate this, we compare models trained using our framework against strong baselines in Table 6, such as the open-weight or proprietary LLMs Qwen3 (14b/235b) (Yang et al., 2025a), DeepSeek-v3.2 (671b) (DeepSeek-AI et al., 2025), Meta Llama 3 (70b) (Dubey et al., 2024), and OpenAI GPT 5.2 (Singh et al., 2025) and OSS (120b) (OpenAI et al., 2025). The comparison also includes simpler baselines such as SVM (Cortes and Vapnik, 1995) with TF-IDF features and Majority Vote based on multilingual E5 (Wang et al., 2024) embeddings. For more specific details on these strong baselines we refer to Löffler et al. (2026).

The results show that our training framework is especially competitive on the DARK and GRAY categories, where it often matches or surpasses the strong LLM baselines. For instance, BETO with our training method achieves the best DARK Macro-F1 (0.65) and DARK Micro-F1 (0.82) in the entire table, exceeding the strongest RAG competitors such as Qwen3 14b (0.62/0.79) and GPT 5.2 (0.61/0.79), as well as all few-shot GPT 5.2 variants, while BETO has only 110

Method	ILLEGAL		DARK		GRAY	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
<i>Baselines</i>						
SVM (TF-IDF)	0.36±0.02	0.64±0.01	0.43±0.02	0.77±0.01	0.59±0.02	0.70±0.02
Majority Vote (M-E5 embeddings)	0.46±0.00	0.71±0.00	0.51±0.00	0.76±0.00	0.57±0.00	0.67±0.00
<i>RAG methods</i>						
GPT 5.2	0.55 ±0.02	0.71±0.02	0.61±0.01	0.79±0.00	0.68 ±0.01	0.73±0.01
Qwen3 235b	0.42±0.02	0.66±0.00	0.61±0.04	0.79±0.03	0.65±0.02	0.73±0.00
Qwen3 14b	0.49±0.01	0.72±0.01	0.62±0.00	0.79±0.00	0.70 ±0.02	0.77 ±0.01
DeepSeek-v3.2	0.45±0.02	0.63±0.03	0.56±0.06	0.74±0.03	0.63±0.01	0.72±0.01
Meta Llama3 70b	0.44±0.02	0.64±0.01	0.50±0.14	0.60±0.12	0.66±0.02	0.73±0.01
GPT OSS 120b	0.46±0.02	0.68±0.02	0.55±0.02	0.76±0.02	0.65±0.01	0.71±0.00
<i>Few-shot methods</i>						
GPT 5.2 (5-shot)	0.35±0.01	0.44±0.02	0.62±0.01	0.70±0.01	0.63±0.01	0.68±0.00
GPT 5.2 (3-shot)	0.44±0.02	0.51±0.01	0.58±0.01	0.64±0.01	0.62±0.01	0.61±0.00
GPT 5.2 (1-shot)	0.34±0.01	0.41±0.01	0.42±0.02	0.55±0.02	0.54±0.01	0.57±0.01
<i>Our training framework</i>						
BETO	0.47±0.02	0.71±0.02	0.65 ±0.02	0.82 ±0.01	0.67±0.02	0.74±0.01
Legal XLM-R	0.50±0.03	0.73 ±0.01	0.62±0.01	0.80±0.01	0.63±0.02	0.71±0.01
BNE-R base	0.47±0.02	0.70±0.00	0.60±0.02	0.80±0.01	0.63±0.01	0.75 ±0.02
BNE-R large	0.50±0.01	0.72±0.02	0.62±0.04	0.79±0.01	0.63±0.02	0.73±0.01
XLM-R large	0.54 ±0.03	0.73 ±0.02	0.63 ±0.03	0.81 ±0.01	0.64±0.05	0.74±0.03

Table 6

Comparison across dense RAG methods, few-shot methods, and simple baselines (first reported by Löffler et al. (2026)) with our training framework. Columns report Macro-F1 and Micro-F1 with standard deviation for ILLEGAL, DARK, and GRAY categories over five random seeds. We highlight the top-2 results in bold. The configuration of the LLM baselines is given in Appendix B.

million parameters, whereas Qwen3 has 14 billion and GPT 5.2 is estimated to have about 1.3 trillion parameters (Li, 2026). One possible interpretation is that the DARK category is particularly well aligned with our training framework, since it is more balanced than ILLEGAL, yet less reliant on context than GRAY, while still involving substantial label co-occurrence that may benefit from uncertainty-aware soft-label transfer.

On GRAY, our method also remains highly competitive. BETO reaches 0.67 Macro-F1 and 0.74 Micro-F1, while BNE-R base reaches GRAY Micro-F1 of 0.75, outperforming GPT 5.2 RAG and Qwen3 235b on Micro-F1 and clearly improving over the few-shot baselines, whose Micro-F1 ranges only from 0.57 to 0.68. Even where a RAG method attains the highest GRAY Macro-F1, such as Qwen3 14b with 0.70, our method stays close while offering stronger or tied Micro-F1 and more stable performance overall. This pattern is consistent with the idea that weak supervision helps on GRAY by expanding the training signal for a highly context-dependent category, although the greater legal ambiguity of GRAY likely limits how far any method can separate the classes.

On ILLEGAL, the picture is more mixed: GPT 5.2 RAG obtains the top Macro-F1 (0.55), while our best models, such as XLM-R large at 0.54 and BNE-R large or Legal XLM-R at 0.50, remain competitive and still outperform most few-shot settings and several RAG alternatives at considerably lower computational cost. A likely reason is that ILLEGAL is the most constrained by extreme class imbalance, with a few labels dominating the dataset, such as “NA” and “LPC PRO”, and several rare classes contributing disproportionately to Macro-F1 difficulty, such as “COT” or “ng”, which makes consistent gains harder to obtain even with improved training.

Overall, Table 6 shows that our method leverages the weak supervision training signals to compete strongly with much larger LLM-based approaches, with particularly clear advantages on the more nuanced DARK and GRAY categories. These findings support our claim that the proposed training strategy is an effective and robust alternative to LLMs.

Backbone	Category	Sup-Base	Best Method	Best Method Score	Rel. Δ (Ma/Mi)
BETO	ILLEGAL	44.34 $\pm$ 2.35 / 71.52 $\pm$ 1.42	SF ($\gamma = 1.5$)	47.27 $\pm$ 1.56 / 70.92 $\pm$ 1.75	+6.61% / -0.84%
	DARK	55.17 $\pm$ 4.10 / 80.96 $\pm$ 0.75	SF ($\gamma = 1.5$)	64.63 $\pm$ 1.95 / 81.53 $\pm$ 0.91	+17.15% / +0.70%
	GRAY	59.13 $\pm$ 2.01 / 72.46 $\pm$ 0.73	GCE ($q = 0.7$)	67.31 $\pm$ 2.38 / 74.20 $\pm$ 1.23	+13.83% / +2.40%
Legal XLM-R	ILLEGAL	47.47 $\pm$ 4.17 / 73.80 $\pm$ 1.53	SF ($\gamma = 1.0$)	49.87 $\pm$ 2.90 / 73.33 $\pm$ 1.14	+5.06% / -0.64%
	DARK	53.33 $\pm$ 4.15 / 78.90 $\pm$ 2.45	Flat-ST	62.28 $\pm$ 1.00 / 80.29 $\pm$ 0.62	+16.78% / +1.76%
	GRAY	58.90 $\pm$ 2.05 / 72.69 $\pm$ 1.96	GCE ($q = 0.7$)	62.51 $\pm$ 2.13 / 70.52 $\pm$ 1.29	+6.13% / -2.99%
BNE-R base	ILLEGAL	40.55 $\pm$ 3.37 / 70.99 $\pm$ 1.74	Soft-SDS	47.12 $\pm$ 2.35 / 70.41 $\pm$ 0.37	+16.20% / -0.82%
	DARK	52.48 $\pm$ 1.87 / 80.44 $\pm$ 1.43	SF ($\gamma = 1.5$) + PW	59.78 $\pm$ 2.22 / 79.58 $\pm$ 0.83	+13.91% / -1.07%
	GRAY	55.93 $\pm$ 1.16 / 74.21 $\pm$ 2.23	Soft-SDS	63.34 $\pm$ 1.28 / 74.79 $\pm$ 1.53	+13.25% / +0.78%
BNE-R large	ILLEGAL	43.17 $\pm$ 4.15 / 70.93 $\pm$ 0.89	SF ($\gamma = 1.5$)	50.46 $\pm$ 1.47 / 71.81 $\pm$ 1.89	+16.89% / +1.24%
	DARK	58.80 $\pm$ 1.39 / 79.76 $\pm$ 1.24	GCE ($q = 0.7$)	61.74 $\pm$ 3.67 / 78.72 $\pm$ 0.80	+5.00% / -1.30%
	GRAY	59.18 $\pm$ 1.97 / 73.43 $\pm$ 1.76	SF ($\gamma = 1.5$) + LA	63.30 $\pm$ 2.06 / 73.06 $\pm$ 0.76	+6.96% / -0.50%
XLM-R large	ILLEGAL	53.75 $\pm$ 1.79 / 74.99 $\pm$ 1.41	SF ($\gamma = 1.5$)	54.01 $\pm$ 3.31 / 73.05 $\pm$ 1.93	+0.48% / -2.59%
	DARK	57.38 $\pm$ 3.18 / 81.12 $\pm$ 1.68	Soft-SDS	63.27 $\pm$ 2.62 / 81.33 $\pm$ 0.91	+10.26% / +0.26%
	GRAY	63.27 $\pm$ 2.17 / 75.62 $\pm$ 2.08	SF ($\gamma = 1.0$)	64.15 $\pm$ 5.22 / 73.75 $\pm$ 2.66	+1.39% / -2.47%

Table 7

Comparison of supervised baseline vs. best weak supervision across architectures and categories. The relative improvement column reports the percentage change in (Macro-F1 / Micro-F1) with respect to Sup-Base.

6.2. Impact of Weakly Supervised Learning

This section tests whether weak supervision yields a real advantage over a strict supervised baseline, and whether that advantage generalizes across backbones and clause families rather than appearing only in isolated cases.

To answer this, we compare the supervised reference (Sup-Base) against the best weakly supervised configuration observed for each backbone–category pair. Table 7 summarizes this comparison and reports both categorical Macro-F1 and Micro-F1, together with the relative change with respect to Sup-Base. This view is intentionally agnostic about which specific weakly supervised variant wins in each case: the purpose here is to establish whether weak supervision as a family provides a meaningful advantage before turning to the more fine-grained question of which method should be preferred by default.

The broad pattern is positive. In all 15 backbone–category combinations shown in Table 7, the best weakly supervised method improves Macro-F1 over Sup-Base, often by a substantial margin. Section 6.5 quantifies significance. The largest gains appear in settings where the supervised baseline is comparatively data-limited. BETO improves by +17.15% on DARK and +13.83% on GRAY, Legal XLM-RoBERTa base improves by +16.78% on DARK, BNE RoBERTa base improves by +16.20% on ILLEGAL and +13.25% on GRAY, and BNE RoBERTa large improves by +16.89% on ILLEGAL. XLM-RoBERTa large is more mixed, but still shows a clear gain on DARK (+10.26%). Taken together, these results indicate that weak supervision can materially strengthen clause-category recognition across very different encoder families, including monolingual Spanish and multilingual backbones.

At the same time, the improvements are not uniform in character. The strongest and most consistent benefits are in Macro-F1, which is the metric most sensitive to rare subtype performance and therefore the most relevant summary for this long-tailed legal classification setting. Micro-F1 declines slightly across the 15 tasks, with Soft-SDS losing 1.29 Micro-F1 points on average against Sup-Base (95% CI $[-2.00, -0.63]$, Wilcoxon $p = 0.002$, Hedges' $g = -0.86$), winning in only 3 of 15 settings. The decline is small but systematic, and we state it as a trade-off. This can be seen, for example, in BETO ILLEGAL (+6.61% macro, -0.84% micro), Legal XLM-RoBERTa base GRAY (+6.13% macro, -2.99% micro), and XLM-RoBERTa large ILLEGAL (+0.48% macro, -2.59% micro). Appendix C resolves this to the level of individual classes. Splitting each family at its median gold prevalence, weak supervision improves the eleven tail classes by 5.00 F1 on average against 1.60 for the thirteen head classes. The interpretation is that weak supervision is particularly effective at improving the balance of subtype recognition, but that these gains do not always translate into higher aggregate accuracy over the majority label mass. This pattern is expected in a severely imbalanced multilabel task, and it reinforces the value of emphasizing Macro-F1 when assessing whether weak data helps category resolution rather than only dominant-class prediction.

A second important observation is that the winning weakly supervised method is not the same everywhere. Soft-SDS is itself the best method in some prominent cases, including BNE RoBERTa base on ILLEGAL and GRAY

Method	Mean Rank ↓	Median Rank ↓	Top-3 Rate ↑	Bottom-2 Rate ↓	Rank SD ↓
SF ($\gamma = 1.5$)	3.13	2.0	60.0%	6.7%	2.10
SF ($\gamma = 1.0$)	3.33	3.0	73.3%	6.7%	1.84
Soft-SDS	3.47	3.0	53.3%	0.0%	1.64
GCE ($q = 0.7$)	3.87	4.0	40.0%	6.7%	2.17
SF ($\gamma = 1.5$) + PW	5.07	6.0	26.7%	26.7%	2.22
SF ($\gamma = 1.5$) + LA	5.20	6.0	26.7%	40.0%	2.40
Flat-ST	5.60	7.0	20.0%	53.3%	2.53
Hard-SDS	6.33	7.0	0.0%	60.0%	1.40

Table 8

Default-method stability across all tasks based on Competition Ranking.

and XLM-RoBERTa large on DARK, which supports its role as the strongest overall default. However, several of the numerically best results are achieved by robust-loss refinements layered on top of the soft-label pipeline, especially focal-style variants and, in some cases, GCE. For example, BETO attains its best scores with Soft Focal on ILLEGAL and DARK and with GCE on GRAY. This confirms the paper’s intended distinction between a best overall default and best category-specific refinement. Weak supervision is broadly beneficial, Soft-SDS is a strong general-purpose choice, and robust-loss modifications can provide additional upside in selected architecture–category settings.

Overall, Table 7 shows that weak supervision is not merely viable in this task, but often materially better than strict supervision. At the same time, the table also makes clear that these gains are architecture- and category-dependent, and that the best-performing weakly supervised variant varies across settings. This motivates the next level of analysis that examines which form of weak supervision provides the strongest overall default, and when robust refinements improve upon it.

6.3. Refinement Method Selection

Our next goal is to determine which weak-supervision method should be preferred in practice. Because performance varies across backbones and clause families, this question has two parts: which method is the strongest overall default, and which methods are best treated as category-specific refinements rather than universal replacements.

To answer this, we summarize the benchmark in two complementary ways. Table 8 reports task-wise rank statistics across all backbone–category combinations, using competition ranking over the full method set. This view is designed to capture consistency across the benchmark rather than only peak wins. Table 9 then evaluates each alternative relative to Soft-SDS, which serves as the reference soft-label default, and reports both upside and downside using Macro-F1 gains, win rates, and tolerance-based non-inferiority (Walker and Nowacki, 2011), with 1-point tolerance. The full result grid is provided in the supplemental materials.

The default-method comparison in Table 8 shows that the strongest region of the method space is concentrated in the soft-label and robust-loss families. Soft Focal ($\gamma = 1.5$) attains the best mean rank (3.13) and best median rank (2.0), while Soft Focal ($\gamma = 1.0$) and Soft-SDS follow closely with mean ranks of 3.33 and 3.47, respectively. At first glance, this could suggest that a focal-style loss should replace Soft-SDS as the main recommendation. However, the remaining columns point to a more nuanced interpretation. Soft-SDS is the only top-three method with a Bottom-2 rate of 0.0%, whereas both focal variants incur at least some low placements. Moreover, Soft Focal ($\gamma = 1.0$) has the highest Top-3 rate (73.3%), but Soft Focal ($\gamma = 1.5$) combines stronger central rank statistics with higher dispersion. By contrast, the harder pseudo-label baselines are clearly weaker defaults: Flat-ST and Hard-SDS have the worst mean and median ranks, the highest Bottom-2 rates, and little evidence of broad competitiveness. These rank differences should not be over-read. A Friedman test over the eight methods and 15 tasks rejects the null of equal average ranks ($\chi^2 = 25.20$, $df = 7$, $p < 0.001$), but the Nemenyi post-hoc critical difference at $\alpha = 0.05$ is 2.71 rank positions, and the gap between the top-ranked Soft Focal ($\gamma = 1.5$) and Soft-SDS is only about one third of a rank position. Hard-SDS is the sole method the test separates from the top. The apparent ordering among the leading four methods is therefore not statistically resolvable at this sample size, and thus, we do not claim it is.

The benchmark therefore supports a family-level conclusion rather than a single unconditional winner, with the most reliable default choices being the Soft-SDS and soft robust-loss region, while harder pseudo-label methods are

Refinement Method	Avg Δ Mac $\uparrow$	Med Δ Mac $\uparrow$	Avg Δ Mic	Win Rate $\uparrow$	Non-Inf Rate $\uparrow$	Large-Loss $\downarrow$
SF ($\gamma = 1.5$)	+0.15	+0.40	+0.31	66.7%	80.0%	20.0%
SF ($\gamma = 1.0$)	-0.11	+0.21	+0.24	53.3%	66.7%	33.3%
GCE ($q = 0.7$)	-0.82	+0.09	-0.24	53.3%	60.0%	40.0%
SF ($\gamma = 1.5$) + PW	-1.03	-1.28	-0.38	20.0%	40.0%	60.0%
SF ($\gamma = 1.5$) + LA	-1.42	-1.22	-0.73	26.7%	40.0%	60.0%
Hard-SDS	-2.18	-2.20	+1.44	6.7%	40.0%	60.0%
Flat-ST	-2.57	-2.08	+0.03	20.0%	33.3%	66.7%

Table 9

Refinement value of Soft Focal and GCE variants relative to Soft-SDS (Tolerance $\tau = 1.0$). The differences between Soft-SDS and the robust-loss refinements are not statistically significant after Holm correction, with only Hard-SDS separated from Soft-SDS (Table 13).

not preferred as general strategies. Within that region, the choice can be made on secondary grounds: Soft-SDS is the only top-three method that never lands in the bottom two, which is what recommends it as a default.

Table 9 clarifies how these methods should be interpreted relative to Soft-SDS. Among all alternatives, Soft Focal ($\gamma = 1.5$) is the only refinement that is positive on average in both Macro-F1 and Micro-F1 relative to Soft-SDS, with an average gain of +0.15 Macro-F1 and +0.31 Micro-F1. It also achieves the highest win rate against Soft-SDS (66.7%) and the strongest non-inferiority rate within the 1-point tolerance (80.0%), while keeping its large-loss rate to 20.0%. This makes it the clearest first refinement to try when additional tuning is feasible. Soft Focal ($\gamma = 1.0$) is more mixed: although its median macro gain remains positive (+0.21) and its average micro gain is also positive (+0.24), its mean macro gain is slightly negative and its downside is materially larger. GCE is even more selective. Its median macro gain is close to neutral (+0.09), indicating that it can match or slightly exceed Soft-SDS in many settings, but its average macro result is negative (-0.82) and its large-loss rate rises to 40.0%, implying greater volatility. The LA and PW combinations are weaker still, with negative mean and median macro gains and substantial downside. These results support treating robust-loss variants as selective refinements rather than as blanket replacements for the soft-label default.

The practical recommendation is therefore two-level. Soft-SDS remains the strongest overall default because it combines consistently strong ranking performance with the best downside profile among the leading methods. If additional method tuning is possible, Soft Focal ($\gamma = 1.5$) is the first refinement to try, since it offers the clearest positive upside over Soft-SDS with limited risk. GCE remains a reasonable secondary alternative in settings where focal-style refinement is less effective, but its use should be treated as architecture- and category-dependent. The full grid reported in the supplemental materials reinforces this interpretation by showing that the strongest individual results are distributed across Soft-SDS and robust-loss variants, even though the broadest default recommendation should prioritize consistency across the full benchmark rather than isolated wins.

6.4. Framework Ablation Studies

This section focuses on the architectural ablations in Section 6.4.1, followed by the ablations of the weak pool's construction algorithm in Section 6.4.2.

6.4.1. Architecture Ablation

This section ablates the architectural decisions of weak supervision and evaluates what components improve over expert-only supervised learning.

In a first test, we compare Sup-Base against any weakly supervised variant, i.e., Flat-ST, Hard-SDS, and Soft-SDS. As shown in Table 10, weak supervision frequently improves Macro-F1 relative to Sup-Base, especially for DARK and GRAY. For example, Soft-SDS increases Macro-F1 over Sup-Base for all five backbones on DARK, and it also yields clear gains on GRAY for BETO, BNE RoBERTa base, and BNE RoBERTa large. However, the corresponding Micro-F1 gains are smaller and less consistent, and ILLEGAL remains comparatively difficult across methods. These results show that weak supervision is beneficial overall, but mainly by improving subtype coverage under class imbalance rather than by uniformly increasing instance-level accuracy.

Second, we assess whether preserving the detection–classification hierarchy is preferable to direct flat self-training. We examine this by comparing the sequential methods Hard-SDS and Soft-SDS against Flat-ST. Table 10 indicates that

Backbone	Category	Sup-Base	Flat-ST	Hard-SDS	Soft-SDS
BETO	ILLEGAL	44.34 $\pm$ 2.35 / 71.52 $\pm$ 1.42	45.31 $\pm$ 1.59 / 70.38 $\pm$ 1.53	46.58 $\pm$ 3.21 / 72.18 $\pm$ 0.57	46.69 $\pm$ 1.62 / 70.80 $\pm$ 2.29
	DARK	55.17 $\pm$ 4.10 / 80.96 $\pm$ 0.75	55.48 $\pm$ 2.81 / 79.08 $\pm$ 1.42	59.40 $\pm$ 3.15 / 81.00 $\pm$ 1.23	62.20 $\pm$ 1.59 / 81.19 $\pm$ 1.03
	GRAY	59.13 $\pm$ 2.01 / 72.46 $\pm$ 0.73	62.63 $\pm$ 1.84 / 72.57 $\pm$ 1.53	61.24 $\pm$ 2.26 / 72.79 $\pm$ 1.30	65.04 $\pm$ 3.14 / 71.97 $\pm$ 1.70
Legal XLM-R base	ILLEGAL	47.47 $\pm$ 4.17 / 73.80 $\pm$ 1.53	47.13 $\pm$ 2.13 / 73.09 $\pm$ 1.53	46.96 $\pm$ 3.37 / 72.89 $\pm$ 2.10	47.33 $\pm$ 2.26 / 71.49 $\pm$ 1.72
	DARK	53.33 $\pm$ 4.15 / 78.90 $\pm$ 2.45	62.28 $\pm$ 1.00 / 80.29 $\pm$ 0.62	60.07 $\pm$ 2.63 / 80.20 $\pm$ 2.17	60.31 $\pm$ 2.99 / 78.62 $\pm$ 1.03
	GRAY	58.90 $\pm$ 2.05 / 72.69 $\pm$ 1.96	62.21 $\pm$ 1.62 / 73.15 $\pm$ 1.62	61.53 $\pm$ 2.82 / 73.29 $\pm$ 2.71	60.74 $\pm$ 3.75 / 69.38 $\pm$ 2.97
BNE-R base	ILLEGAL	40.55 $\pm$ 3.37 / 70.99 $\pm$ 1.74	44.80 $\pm$ 6.05 / 71.38 $\pm$ 2.43	43.32 $\pm$ 1.80 / 71.58 $\pm$ 1.45	47.12 $\pm$ 2.35 / 70.41 $\pm$ 0.37
	DARK	52.48 $\pm$ 1.87 / 80.44 $\pm$ 1.43	52.30 $\pm$ 2.70 / 78.60 $\pm$ 0.81	53.28 $\pm$ 3.70 / 80.31 $\pm$ 1.18	58.94 $\pm$ 3.62 / 78.12 $\pm$ 0.72
	GRAY	55.93 $\pm$ 1.16 / 74.21 $\pm$ 2.23	61.82 $\pm$ 1.06 / 75.44 $\pm$ 0.54	58.02 $\pm$ 1.06 / 75.96 $\pm$ 0.53	63.34 $\pm$ 1.28 / 74.79 $\pm$ 1.53
BNE-R large	ILLEGAL	43.17 $\pm$ 4.15 / 70.93 $\pm$ 0.89	44.55 $\pm$ 1.18 / 68.11 $\pm$ 1.88	43.03 $\pm$ 4.26 / 70.89 $\pm$ 2.18	45.14 $\pm$ 2.94 / 69.94 $\pm$ 2.36
	DARK	58.80 $\pm$ 1.39 / 79.76 $\pm$ 1.24	58.35 $\pm$ 1.29 / 79.06 $\pm$ 1.42	60.62 $\pm$ 1.92 / 80.40 $\pm$ 2.08	60.72 $\pm$ 1.03 / 79.34 $\pm$ 1.23
	GRAY	59.18 $\pm$ 1.97 / 73.43 $\pm$ 1.76	59.45 $\pm$ 1.69 / 71.79 $\pm$ 0.84	59.33 $\pm$ 1.90 / 71.81 $\pm$ 1.03	61.53 $\pm$ 2.20 / 70.99 $\pm$ 0.98
XLM-R large	ILLEGAL	53.75 $\pm$ 1.79 / 74.99 $\pm$ 1.41	53.87 $\pm$ 4.61 / 73.00 $\pm$ 0.93	52.14 $\pm$ 1.87 / 74.07 $\pm$ 1.41	52.80 $\pm$ 3.83 / 70.94 $\pm$ 0.95
	DARK	57.38 $\pm$ 3.18 / 81.12 $\pm$ 1.68	49.22 $\pm$ 21.75 / 73.67 $\pm$ 17.14	59.51 $\pm$ 3.05 / 81.89 $\pm$ 1.26	63.27 $\pm$ 2.62 / 81.33 $\pm$ 0.91
	GRAY	63.27 $\pm$ 2.17 / 75.62 $\pm$ 2.08	60.61 $\pm$ 2.22 / 73.25 $\pm$ 1.43	60.87 $\pm$ 3.11 / 74.71 $\pm$ 1.47	63.40 $\pm$ 2.54 / 73.10 $\pm$ 0.78

Table 10

Empirical progression across the first three experimental design axes. Scores represent Macro-F1 / Micro-F1 with standard deviation.

sequential weak supervision is generally more reliable than flat self-training. This is most visible on DARK, where Soft-SDS or Hard-SDS usually outperform Flat-ST, and where Flat-ST exhibits clear instability for XLM-R large. Even in settings where Flat-ST is competitive, such as Legal XLM-RoBERTa base on DARK and GRAY, the sequential methods remain at least comparable in most cases and more stable overall. This supports the value of maintaining view separation between abusive-clause detection and subtype classification, rather than collapsing the task into a single flat pseudo-labeling step.

Lastly, we test whether soft-label sequential distillation improves over harder pseudo-label formulations. To do this, we compare Soft-SDS directly against both Flat-ST and Hard-SDS. The pattern in Table 10 favors Soft-SDS as the strongest overall method among the three. It achieves the highest Macro-F1 on BETO for DARK and GRAY, on BNE RoBERTa base for all three categories, on BNE RoBERTa large for DARK and GRAY, and on XLM-R large for DARK and GRAY. In addition, Soft-SDS avoids some of the pronounced failures observed for Flat-ST, while generally improving on or matching Hard-SDS. The main exception is ILLEGAL, where improvements are smaller and sometimes mixed. Overall, these findings support soft-label sequential distillation as the most reliable default among the weak-supervision strategies evaluated in the benchmark.

The mechanism behind the Soft-SDS advantage over Hard-SDS is measurable. Appendix A shows that thresholding a teacher posterior discards supervision in a way that is biased towards frequent classes, and Table 16 quantifies this over the weak pool. Across the tail classes of all three families, thresholding removes roughly two-thirds of the probability mass the teacher assigns, and for three of them (*acp* and *ng* in ILLEGAL and *des us* in GRAY) it removes the weak signal entirely, leaving no above-threshold instance while the soft mass remains strictly positive. Head classes retain comparable supervision under both schemes. The effect is not specific to one teacher: Table 17 shows the same collapse for all five backbones, with 15 of 55 tail class–backbone combinations receiving no above-threshold label at all and soft-to-hard mass ratios of 1.4–3.5 $\times$. Hard-SDS therefore concentrates weak supervision on the head of the distribution, not because the teacher is silent about rare subtypes, but because what it says never clears the decision threshold. The prediction that follows is directional and testable. Soft targets should improve Macro-F1, which weights all subtypes equally, while conceding ground on Micro-F1, which is dominated by frequent labels. Soft-SDS improves Macro-F1 over Hard-SDS on 14 of 15 backbone–category settings (+2.18 points, $p = 0.006$ after Holm correction), while Hard-SDS is the better of the two on Micro-F1 by 1.44 points. The two metrics move in opposite directions because of this training-signal mechanism.

6.4.2. Data Construction

To assess whether the main conclusions depend on specific data-construction choices, we compare our distribution-aligned weak pool with the paragraph-level chunking strategy.

To analyze this, we report the BETO results from the paragraph-level ablation pool alongside those from the aligned weak pool, focusing on Soft-SDS as our principal soft-label approach. Crucially, to ensure a fair comparison and avoid ground-truth leakage, both pools were subjected to the same strict filtering. Table 11 summarizes this comparison.

Table 11

Chunking ablation on BETO comparing the distribution-aligned weak pool with the paragraph-level weak pool. Scores are reported as Macro-F1 / Micro-F1.

Category	Method	Paragraph Pool	Aligned Pool
ILLEGAL	Soft-SDS	50.17 $\pm$ 2.86 / 69.99 $\pm$ 2.52	46.69 $\pm$ 1.62 / 70.80 $\pm$ 2.29
DARK	Soft-SDS	56.55 $\pm$ 2.35 / 78.88 $\pm$ 1.85	62.20 $\pm$ 1.59 / 81.19 $\pm$ 1.03
GRAY	Soft-SDS	62.94 $\pm$ 1.60 / 71.38 $\pm$ 1.20	65.04 $\pm$ 3.14 / 71.97 $\pm$ 1.70

Table 12

Aggregate paired comparisons across the 15 backbone–category tasks. Δ is the mean difference in percentage points; the interval is a percentile bootstrap 95% CI (20,000 resamples); p is a two-sided Wilcoxon signed-rank test over the 15 paired tasks; g is Hedges’ g . The best-of-8 row is descriptive, as the method is selected on the same test scores used for the comparison, so its p -value is anti-conservative.

Comparison	Δ	95% CI	Wins	p	g
<i>Macro-F1</i>					
Soft-SDS vs. Sup-Base	+3.71	[+2.26, +5.16]	13/15	0.0006***	+1.18
Soft-SDS vs. Hard-SDS	+2.18	[+1.20, +3.18]	14/15	0.0009***	+1.02
Soft-SDS vs. Flat-ST	+2.57	[+0.80, +4.68]	12/15	0.0084**	+0.60
best-of-8 vs. Sup-Base [†]	+5.21	[+3.80, +6.65]	15/15	0.0001***	+1.67
<i>Micro-F1</i>					
Soft-SDS vs. Sup-Base	-1.29	[-2.00, -0.63]	3/15	0.0020**	-0.86

[†] descriptive; selection-biased.

The effect of weak-pool construction is category-dependent but informative. On DARK, the aligned pool significantly improves performance: Macro-F1 rises from 56.55 to 62.20, and Micro-F1 from 78.88 to 81.19. On GRAY, the aligned variant also benefits, increasing Macro-F1 from 62.94 to 65.04 and Micro-F1 from 71.38 to 71.97. Interestingly, on ILLEGAL, the paragraph-level pool outperforms the aligned pool in terms of Macro-F1 (50.17 vs. 46.69), although Micro-F1 is roughly comparable (69.99 vs. 70.80). This suggests that the longer paragraph context may preserve cross-clause dependencies that are particularly informative for the most heterogeneous label family.

This result ties back to the broader claim of the paper: weak supervision is not only a matter of choosing the right training architecture. It also depends on constructing a weak pool whose granularity matches the target distribution. Distribution-aligned chunking is highly beneficial for DARK and GRAY but not universally superior, reinforcing the category-dependent nature of the weak-supervision gains reported throughout this study.

6.5. Statistical Significance of the Main Comparisons

In this section, we test the paper’s claims on two levels. First, we report results within a task in Appendix D, where the low power of $n = 5$ means they are included for completeness. Across tasks, $n = 15$ paired observations support a Wilcoxon signed-rank test, and this is the level at which the aggregate claims live. Confidence intervals are percentile bootstrap intervals over 20,000 resamples, and effect sizes are Hedges’ g .

One methodological point deserves emphasis because it affects how the headline number should be read. Section 6.2 compares Sup-Base against the best weakly supervised method in each cell. A test on that quantity inherits the winner’s curse of a maximum over eight correlated candidates. We therefore designate Soft-SDS as the primary pre-specified contrast, and report the best-of-eight contrast separately and explicitly as descriptive.

Table 12 gives the result. Note that this table reports differences in percentage points, whereas Table 7 reports relative change. Both are given below where they are first used, so that the two can be reconciled. Following the pre-specified comparison, Soft-SDS improves Macro-F1 over Sup-Base by 3.71 absolute points on average, equivalently a relative gain of 7.08%, with a bootstrap interval excluding zero and a large effect size. The selection-biased best-of-eight figure is 5.21 points, or 9.99% relative, roughly 1.5 points higher. Soft-SDS also beats both harder pseudo-label variants. Micro-F1 moves the other way, by -1.29 points, and this too is statistically reliable. This is exactly the trade-off predicted in Appendix A.1 and not a failure of the method.

Table 13

Method ranking across the 15 backbone–category tasks. Friedman omnibus: $\chi^2(7) = 25.20$, $p = 7.0e - 04$. The Nemenyi critical difference at $\alpha = 0.05$ is 2.71. The final columns give the pairwise comparison against Soft-SDS (Wilcoxon signed-rank over the 15 tasks, Holm-corrected over the seven comparisons). A positive Δ favors Soft-SDS.

Method	Avg. rank	Δ vs. Soft-SDS	Wins	p_{Holm}
SF ($\gamma = 1.5$)	3.13	-0.15	5/15	1.0000
SF ($\gamma = 1.0$)	3.33	+0.11	7/15	1.0000
Soft-SDS	3.47	—	—	—
GCE ($q = 0.7$)	3.87	+0.82	7/15	1.0000
SF ($\gamma = 1.5$) + PW	5.07	+1.03	12/15	0.0513
SF ($\gamma = 1.5$) + LA	5.20	+1.42	11/15	0.0723
Flat-ST	5.60	+2.57	12/15	0.0502
Hard-SDS	6.33	+2.18	14/15	0.0060**

The method ranking of Section 6.3 requires more caution. A Friedman test over eight methods and 15 tasks rejects the null of equal average ranks, but the Nemenyi critical difference at $\alpha = 0.05$ is 2.71 rank positions, which is wider than the spread among the four leading methods. Table 13 reports the ranking together with pairwise Holm-corrected comparisons against Soft-SDS. Only Hard-SDS is separated from the top of the ranking. The apparent superiority of Soft Focal ($\gamma = 1.5$) over Soft-SDS amounts to 0.15 points and is not distinguishable from zero, and neither are the differences involving GCE or the second focal variant. We conclude that, at the family level, soft-label methods and soft robust-loss refinements occupy the strong region of the method space, hard pseudo-labeling does not, and finer discrimination within the strong region would require more tasks than this benchmark provides.

6.6. Comparison with External Weak-Supervision Baselines

The comparisons above are internal to our own method family. To test whether Soft-SDS is competitive against the wider weak-supervision literature, we additionally evaluate five external baselines. Four represent distinct method families: (i) consensus aggregation over multiple teachers using the Dawid–Skene annotator model (Dawid and Skene, 1979), which estimates each teacher’s reliability by expectation maximization rather than weighting all teachers equally; (ii) confidence-thresholded consistency training in the style of FixMatch (Sohn et al., 2020); (iii) class-adaptive pseudo-label thresholds in the style of FlexMatch (Zhang et al., 2021); and (iv) bilevel meta-learned reweighting of weak labels. The fifth, reverse-KL distillation (Gu et al., 2024), tests an asymmetric divergence and addresses a question raised in our own Limitations.

This is a common-protocol comparison rather than an evaluation of each method under its original default configuration. All methods use the same two backbones, the same splits, the same weak instances, and a fixed 0.5 decision threshold, following the protocol of the main results. To achieve this, two settings that the external implementations apply by default, and that Soft-SDS does not use, were disabled: oversampling of the gold set to balance it against the pseudo-label pool (Soft-SDS simply concatenates the two), and validation-tuned per-label decision thresholds. The baselines reported in Table 14 are therefore standardized variants that share the weak data and the decision rule of Soft-SDS. Their scores isolate the contribution of the learning objective and label-aggregation strategy, and should not be read as the performance of each method under its authors’ recommended settings, which may differ from the figures reported here. Two differences cannot be removed and are stated explicitly. First, FixMatch consumes co-labeled pseudo-labels by construction, and second, the bilevel method reserves 20% of the gold training set as a meta-validation split.

Table 14 shows that, under this common protocol, no external baseline significantly outperforms Soft-SDS. The largest mean gain is obtained by Dawid–Skene consensus at +0.87 Macro-F1 points, but the effect is not significant and its uncertainty interval includes zero. This suggests that the additional reliability modeling has limited value in this setting. The five teachers are already broadly consistent, with estimated reliabilities between 0.72 and 0.96, leaving little exploitable heterogeneity for per-teacher reweighting.

The relevant conclusion is therefore not that Soft-SDS is more accurate than the external baselines, which the data do not show, but that it offers a more favorable cost–robustness trade-off. It remains statistically competitive with all five alternatives while requiring fewer assumptions and less machinery. Dawid–Skene consensus presupposes an ensemble of five trained teachers and costs multiple times the teacher-training time of Soft-SDS, for a non-significant gain. Bilevel

Table 14

Common-protocol comparison of external weak-supervision baselines against Soft-SDS, over six backbone–category tasks (BETO and BNE-RoBERTa large $\times$ three clause families, five seeds each). All methods use identical splits, identical weak instances and a fixed 0.5 decision threshold; baseline-specific gold-set oversampling and validation-tuned per-label thresholds were disabled, so the rows are standardized variants rather than each method's original default configuration. A positive Δ means the baseline outperforms Soft-SDS, p is a Wilcoxon signed-rank test over the six tasks, Holm-corrected across the baselines. No baseline differs significantly from Soft-SDS.

Baseline	Family	Δ Macro-F1	Wins	p_{Holm}
Consensus (Dawid–Skene)	consensus	+0.87	5/6	1.000
Reverse KL	asymmetric divergence	+0.51	4/6	1.000
FixMatch	confidence thresholding	+0.48	3/6	1.000
Adaptive threshold	adaptive pseudo-labeling	−0.13	2/6	1.000
Bilevel reweighting	meta-learning	−0.73	4/6	1.000

Table 15

End-to-end evaluation on the full test split, Macro-F1. The gold-abusive setting corresponds to the main protocol used throughout the paper, where categorization is evaluated only on clauses known to be abusive. The end-to-end setting evaluates the full two-stage pipeline by applying the abusiveness detector before categorization. The detection recall column reports the share of gold-abusive clauses that are passed to the categorizer.

Family	Gold-abusive		End-to-end		Det. recall
	Sup-Base	Soft-SDS	Sup-Base	Soft-SDS	
ILLEGAL	45.28	46.02	27.20	28.17	0.65
DARK	56.39	60.40	23.98	24.02	0.56
GRAY	59.86	64.19	27.72	26.56	0.44

reweighting reserves 20% of an already small gold training set as a meta-validation split, which is unattractive under annotation scarcity. Threshold-based methods introduce additional pseudo-label selection rules, whereas Soft-SDS uses the teacher probability vector directly and therefore preserves weak tail-class signal that hard thresholding may discard. On BETO, Soft-SDS trains in 274 s compared with 344–355 s for FixMatch, Bilevel and Reverse KL, before counting any additional teacher requirements. Thus, Soft-SDS is best understood as a conservative default for low-resource legal classification. It is not statistically inferior to standardized variants of more elaborate weak-supervision baselines, but is simpler, cheaper, and less dependent on extra teachers, validation data, or threshold tuning. Whether the baseline-specific settings removed for this comparison, such as gold-set oversampling or per-label threshold tuning, would change the ranking is a separate question that the present experiment does not address.

6.7. End-to-End Evaluation with Error Propagation

The preceding experiments evaluate the categorization stage under the protocol described in Section 5, where the input clauses are already known to be abusive. This setting isolates the component studied in this paper, but it does not measure the performance of the full two-stage pipeline. In deployment, only clauses selected by the abusiveness detector are passed to the downstream categorizer. We therefore additionally evaluate the complete pipeline on the full test split (end-to-end setting, Figure 1(b)). For this experiment, categorization predictions are gated by the detector selected in the published pipeline of Löffler et al. (2025), so that detector false negatives are propagated to the final multilabel output.

Table 15 shows that end-to-end performance is substantially lower than the isolated categorization scores. Relative Macro-F1 decreases by approximately 39–60% once detector errors are propagated. This drop is consistent with the detector recall values: the detector misses 55 of 157 ILLEGAL, 56 of 127 DARK, and 63 of 113 GRAY abusive clauses, which removes 56, 64, and 79 gold subtype labels, respectively, before categorization is applied.

The comparison also shows that the advantage of Soft-SDS observed under the gold-abusive protocol is no longer visible in the full pipeline. Averaged across the three families, Soft-SDS improves over Sup-Base by +3.03 Macro-F1 points in the gold-abusive setting, but changes by only −0.05 Macro-F1 points end-to-end. This should not be interpreted as evidence that the categorizer itself fails under deployment conditions. Rather, under a hard-gated

architecture, detector false negatives remove abusive clauses from the categorization input, and downstream classifiers cannot assign subtype labels to clauses they never receive.

We report this experiment to make explicit how the isolated categorization results translate to the complete pipeline. For the evaluated two-stage architecture, deployed performance is detector-limited, especially for GRAY clauses, where detector recall is only 0.44. Improving abusiveness detection is therefore likely to yield larger end-to-end gains than further refinements of the categorization model alone.

6.8. Pseudo-Label Quality and Sensitivity

This subsection examines the estimated reliability of the teachers that generate the pseudo-labels, and the sensitivity of the method to teacher choice and weak-pool construction.

We first analyze teacher reliability. We fit a Dawid–Skene model to the predictions of five classifiers in order to estimate per-teacher, per-class reliability, measured as the balanced accuracy of the estimated confusion matrix. These estimates are inferred using only the agreement structure among teachers, without access to gold labels. Averaged over classes, estimated reliability ranges from 0.72 to 0.96 and tracks backbone capacity, i.e., XLM-RoBERTa large is the most reliable teacher in every family, with estimated reliabilities of 0.890, 0.956, and 0.962 on ILLEGAL, DARK, and GRAY, respectively. RoBERTa-BNE base is the least reliable teacher, with corresponding values of 0.723, 0.824, and 0.764. Reliability is lowest on ILLEGAL for every teacher. This is also the family with the most and the sparsest classes, and it is the family for which weak supervision yields the smallest gains in our main results.

Table 16 reports the per-class supervision mass contributed by the teacher under thresholding and under soft weighting. The relevant observation for pseudo-label quality is that thresholding removes a large share of the probability mass assigned to tail classes. On average, approximately two-thirds of the teacher-assigned mass for tail classes is discarded, and for three tail classes no mass remains after thresholding. This provides a plausible explanation for why soft pseudo-labels are more effective than hard pseudo-labels in our experiments. Soft targets preserve graded teacher information for rare labels that would otherwise be lost under thresholding.

We next test whether the method depends on a particular teacher. Substituting a stronger cross-architecture teacher yields a mean gain of +0.32 Macro-F1 points relative to the main teacher choice, while aggregating all five teachers with Dawid–Skene consensus yields a mean gain of +0.87 Macro-F1 points. Neither difference is statistically significant at the sample size used here. We therefore interpret these comparisons conservatively. The method does not appear to depend strongly on a specific teacher, but the present experiments cannot rule out smaller teacher-choice effects. The larger nominal gain from Dawid–Skene aggregation is consistent with the idea that combining teachers can add information beyond selecting a single stronger teacher, but this pattern should be treated as suggestive rather than conclusive.

7. Discussion

The results show that weak supervision is a viable strategy for fine-grained abusive clause classification in a low-resource legal setting, but its benefits are not uniform across methods, backbones, or clause families. Overall, the main gains appear in Macro-F1, indicating that weak supervision is especially useful for improving performance on rare and underrepresented subtypes. By contrast, Micro-F1 improvements are smaller and less consistent, suggesting that the additional weak signal mainly helps balance subtype recognition rather than uniformly improving instance-level accuracy. These conclusions concern the categorization stage evaluated on gold-abusive clauses. As Section 6.7 shows, in the hard-gated two-stage pipeline the categorization gains are currently masked by detector errors. The claims below should be read as claims about subtype categorization rather than about end-to-end abusive clause analysis.

A central finding of the benchmark is that soft-label sequential distillation (Soft-SDS) is the most reliable overall default within the evaluated setting. Although it is not always the top-performing method on every backbone–category pair, it offers the most stable performance across the benchmark and avoids the larger downside observed for harder pseudo-label approaches. This supports the use of soft teacher distributions in the present legal setting, where subtype overlap, ambiguity, and class imbalance make hard pseudo-label propagation comparatively brittle.

At the same time, the results do not support a single universal refinement beyond Soft-SDS. Among the alternatives, Soft Focal is the clearest candidate for further tuning, as it is the only refinement that improves both Macro-F1 and Micro-F1 on average relative to Soft-SDS. However, its gains are selective rather than universal, and its greater variability suggests that it should be treated as a targeted refinement rather than as a replacement for the soft-label default. GCE shows a similar but more volatile pattern.

The category-level differences are also informative. DARK is the most receptive to weak supervision, whereas ILLEGAL remains the most difficult setting, likely due to its larger label space and more extreme sparsity. GRAY shows intermediate behavior, with improvements that depend more strongly on backbone and method choice. These patterns indicate that the effectiveness of weak supervision is shaped not only by model design, but also by the internal structure of the target label family.

The chunking ablation further shows that weak-pool construction is an important part of the overall method. Distribution-aligned chunking generally leads to better results than paragraph-level segmentation, particularly for DARK and GRAY, although ILLEGAL suggests that longer contextual spans may still be beneficial in some cases. This implies that weak-supervision quality depends not only on pseudo-labeling strategy, but also on how unlabeled legal text is segmented before training.

7.1. Theoretical and Practical Implications

Theoretically, the findings support the view that weak supervision is more effective in legally ambiguous and highly imbalanced settings when uncertainty is preserved through soft-label transfer rather than collapsed into hard pseudo-labels. They also suggest that weak supervision in legal NLP should be understood as a structured transfer problem, i.e., preserving the decomposition between detection and subtype classification can improve pseudo-label reliability.

Practically, the results support a simple recommendation for low-resource legal NLP. Soft-SDS should be used as the default weak-supervision method, while focal-style robust losses can be tested as optional refinements when additional tuning is feasible. The results also show that weak-data construction matters in practice. In particular, chunking strategy should be treated as part of the method rather than as a neutral preprocessing step.

7.2. How This Study Differs from Existing Work

Unlike most prior work, which focuses either on supervised abusive-clause detection or on generic weak-supervision benchmarks, this study evaluates weak supervision specifically for fine-grained abusive clause classification in a low-resource legal setting in Spanish. Rather than proposing a new general-purpose semi-supervised optimizer, the paper contributes a controlled, leakage-safe, cross-backbone benchmark. We take this to be the more useful contribution in a setting where the practical question is not which optimizer is best in principle, but which of several established strategies survives contact with a few hundred expert-annotated clauses.

The benchmark is designed to separate three questions that are usually entangled: (i) whether weak supervision helps at all, answered against a strict supervised baseline; (ii) whether preserving the detection–categorization decomposition helps, answered against flat self-training; and (iii) whether soft transfer beats hard, answered by holding everything except the form of the pseudo-label constant. Section 6.5 shows that the first and third of these separate cleanly, while the finer ordering among the leading methods does not. Section 6.6 extends the comparison beyond our own method family to standardized variants of five external baselines under a common protocol, and finds that none of them significantly outperforms the simplest method in the set.

A further distinguishing feature is that weak-pool construction is treated as a methodological variable rather than as preprocessing. The chunking ablation moves DARK by more than five Macro-F1 points, a larger effect than any loss-function refinement or teacher-side choice we evaluate, which suggests that in this regime the composition of the unlabeled pool deserves at least as much attention as the learning objective applied to it.

Finally, Section 6.7 evaluates the complete pipeline rather than the categorization stage in isolation. We are not aware of comparable work in this area that reports the propagated cost of detection errors alongside stage-level results. In our case it points towards future work, i.e., the categorization gains this paper measures are real, but they are currently limited by detector performance.

8. Limitations

Several limitations should be acknowledged. First, the study focuses on Chilean Terms of Service in Spanish, and the findings may not transfer directly to other legal domains, jurisdictions, or languages. Differences in legal drafting style, doctrinal structure, and class prevalence may affect both pseudo-label quality and the relative advantage of soft-label weak supervision.

Second, although the weak-supervision pool was collected under a leakage-safe protocol, its properties still depend on the specific crawling, filtering, and chunking pipeline used in this study. The chunking ablation shows that weak-data construction can materially influence downstream performance. Accordingly, the present conclusions apply most directly to the evaluated clause-level weak-supervision regime.

Third, the supervised benchmark remains small and highly imbalanced. This is inherent to the task, but it also means that performance estimates are sensitive to subtype sparsity, annotation granularity, and seed variation. For this reason, the results should be interpreted in terms of consistent cross-backbone patterns rather than as definitive point estimates for individual backbone–category pairs.

Fourth, end-to-end evaluation over the full test split shows that the isolated categorization gains do not directly translate into full-pipeline gains. Macro-F1 falls substantially relative to the gold-abusive protocol used throughout the paper, and the advantage of Soft-SDS over Sup-Base is no longer observed. This is a limitation of the evaluated hard-gated pipeline. Since detector recall ranges from 0.44 to 0.65, many abusive clauses are filtered out before subtype classification can occur. Accordingly, the categorization improvements reported here should be interpreted as conditional on successful upstream detection, while deployed performance remains strongly affected by the abusiveness detector.

Finally, the study evaluates a selected set of backbones and weak-supervision strategies rather than the full space of possible teacher models, loss functions, or semi-supervised designs. Future work should test whether the present conclusions extend to larger legal corpora, alternative teacher calibration schemes, and other forms of uncertainty-aware pseudo-labeling.

9. Conclusion

This paper presented a cross-backbone benchmark of weak supervision for fine-grained abusive clause classification in Chilean Terms of Service. Across monolingual Spanish, multilingual, and legal-domain transformers, and with categorization evaluated on clauses known to be abusive, weak supervision improves Macro-F1 in every one of the 15 backbone–category settings evaluated, by 9.99% on average and by up to 17.15%, with the gains concentrated on the rare subtypes that matter most in this long-tailed legal setting.

The main empirical conclusion is that, across the evaluated benchmark, soft-label sequential distillation (Soft-SDS) is the most reliable overall default. The advantage of soft over hard pseudo-labels is statistically robust, while the finer ordering within the leading group is not. Robust-loss variants such as Soft Focal and GCE are statistically indistinguishable from Soft-SDS, and we recommend the latter on the grounds of consistency and simplicity rather than measured superiority. Harder pseudo-label approaches are the one family the benchmark separates clearly, and they are less reliable as general strategies. A comparison against standardized variants of five external weak-supervision methods under a common protocol found none that significantly outperforms Soft-SDS, though several match it at substantially greater cost.

The experiments also show that weak-supervision effectiveness is shaped by the structure of the target category and by the construction of the unlabeled pool. DARK benefits most consistently, ILLEGAL remains the most difficult family, and the chunking ablation confirms that weak-data granularity materially affects downstream behavior. These findings indicate that successful weak supervision in legal NLP depends not only on model choice, but also on uncertainty-aware label transfer and careful weak-pool design.

The scope of these conclusions is set by the end-to-end evaluation. When the categorizer is placed behind the abusiveness detector of the deployed two-stage pipeline, detector recall of 0.44–0.65 removes a large share of abusive clauses before subtype categorization can act on them. The improvements demonstrated in this paper therefore concern the downstream categorization stage and are conditional on successful upstream detection. Under the evaluated architecture, deployed performance is detector-limited, and improving detection is the most effective path toward end-to-end gains.

Overall, the study supports weak supervision as a practical tool for the fine-grained categorization stage of low-resource legal text classification. Within that stage, it suggests that the most effective recipe is a soft-label teacher–student pipeline, optionally followed by selective robust-loss refinement when additional tuning is feasible.

Funding

This research has been supported by the following grant: ANID/FONDECYT/INICIACION/N.11250673.

Data availability

The source Terms of Service documents analyzed in this study are not publicly redistributed due to legal and copyright restrictions. Derived data, annotations, and code supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of competing interest

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

A. Gradient-Level Analysis of Soft versus Hard Pseudo-Labels

This appendix derives the two properties asserted in Section 4 and measured in Section 6.4.1, namely that hard thresholding introduces a systematic gradient bias concentrated where the teacher is least reliable, and that it withdraws supervision from rare classes in particular. Section A.3 below closes with the empirical prediction that follows.

A.1. The gradient offset introduced by thresholding

For a multilabel model with sigmoid outputs $\hat{y}_c = \sigma(z_c)$ and binary cross-entropy loss, the gradient with respect to the logit of class c is

$$\frac{\partial \mathcal{L}_{BCE}^{(c)}}{\partial z_c} = \hat{y}_c - p_c, \quad (13)$$

that is, the signed discrepancy between the prediction and the target. The two weak-supervision variants differ only in p_c . Soft-SDS uses the teacher posterior $\tilde{y}_c \in [0, 1]$, whereas Hard-SDS uses the thresholded decision $\mathbb{1}[\tilde{y}_c > \theta]$ with $\theta = 0.5$. Their gradients therefore differ by

$$g_c^{\text{hard}} - g_c^{\text{soft}} = \tilde{y}_c - \mathbb{1}[\tilde{y}_c > 0.5] \implies |g_c^{\text{hard}} - g_c^{\text{soft}}| = \min(\tilde{y}_c, 1 - \tilde{y}_c). \quad (14)$$

Equation (14) states that the excess gradient introduced by hard thresholding is exactly the teacher's own estimate of its per-class error probability on that instance. Two properties of this term matter. It is maximal precisely where the teacher is least reliable, reaching 0.5 at $\tilde{y}_c = 0.5$ and vanishing only when the teacher is already certain. And it is systematic rather than stochastic. For a fixed weak instance the offset has the same sign at every epoch, so it accumulates rather than averaging out. This is the optimization-level statement of confirmation bias (Arazo et al., 2020). Hard pseudo-labels do not merely add noise, but instead add a persistent bias whose magnitude is largest on exactly the ambiguous clauses that characterize legal subtype assignment.

Soft targets remove this bias term by construction. Since the minimizer of $\mathbb{E}[\text{BCE}(\hat{y}, p)]$ in $\hat{y}$ is $\hat{y}^* = p$, training on soft targets drives the student towards the teacher's posterior, whereas training on hard targets drives it towards a maximally confident point estimate that the teacher's own evidence does not support.

The entropy weight w_i of Eq. (3) acts on the same quantity from the other side. It is monotonically decreasing in the mean binary entropy of $\tilde{y}_i$, while the residual term $\min(\tilde{y}_c, 1 - \tilde{y}_c)$ in Eq. (14) is monotonically increasing in it. The two components of Soft-SDS therefore address one source of error jointly. Soft targets eliminate the systematic component of the pseudo-label bias, and the entropy weighting shrinks the remaining variance on precisely those instances where the teacher is least informative.

A.2. Why this does not cause minority-class overfitting

Two structural properties bound the influence of the weak branch. First, $\mathcal{L}_{soft}$ is normalized by $\sum_i w_i$, so it is a weighted average rather than a sum. Its scale is independent of the size of the weak pool, and the weak signal cannot swamp the gold term $\mathcal{L}_{CE}$ however many weak instances are added.

Second, and more importantly for rare labels, thresholding is not neutral with respect to class frequency. For a minority class the teacher posterior is typically diffuse and rarely exceeds θ , so the expected number of positive supervision signals contributed by the weak pool is

$$N_c^{\text{hard}} = \sum_i \mathbb{1}[\tilde{y}_{ic} > 0.5] \approx 0, \quad N_c^{\text{soft}} = \sum_i w_i \tilde{y}_{ic} > 0. \quad (15)$$

Table 16

Weak-label mass contributed by the BETO teacher over the detector-filtered weak pool, by class. N_{hard} counts instances whose teacher probability exceeds the 0.5 threshold, N_{soft} is the confidence-weighted probability mass $\sum_i w_i \tilde{y}_{ic}$ that soft targets retain. Head/tail follows a median gold-prevalence split within each family. Thresholding removes most or all supervision for tail classes, while soft targets preserve a strictly positive signal.

Family	Class	Gold %	N_{hard}	N_{soft}	Head/tail
ILLEGAL	COT	1.1	2	6.4	tail
	LPC	10.2	4	9.5	head
	LPC INT	9.8	26	24.4	head
	LPC JUS	4.5	4	11.5	tail
	LPC PRO	21.3	111	86.4	head
	NA	42.5	135	122.9	head
	RC	12.4	76	62.0	head
	acp	2.4	0	4.3	tail
	ng	1.3	0	4.3	tail
DARK	ch	5.4	9	11.8	tail
	cr	37.1	193	159.3	head
	er	6.4	10	13.8	tail
	ltd	46.3	312	268.8	head
	nod	11.1	7	12.9	head
	ter	6.9	2	7.3	tail
GRAY	bfe	12.9	36	49.8	head
	des def	7.8	32	30.2	head
	des det	6.6	3	16.0	tail
	des inf	5.1	5	16.2	tail
	des lic	3.7	2	12.5	tail
	des reser	31.0	150	116.6	head
	des risk	32.0	190	130.3	head
	des uni	31.7	133	98.6	head
	des us	2.4	0	15.1	tail

Rather than over-supplying rare subtypes, soft targets supply a graded signal for every instance, while hard thresholding reduces that signal to the subset of cases in which a rare subtype happens to clear the decision threshold. Hence, hard labels concentrate weak supervision on the head of the distribution. Table 16 reports both quantities per class for the BETO teacher over the detector-filtered weak pool, and Table 17 summarizes the tail for all five teachers. Section 6.4.1 discusses the result.

A.3. Empirical prediction

Because Macro-F1 weights all subtypes equally while Micro-F1 is dominated by frequent labels, the mechanism predicts that soft targets should improve Macro-F1 while maintaining or losing slightly on Micro-F1. Both halves are borne out, in both directions: Soft-SDS improves Macro-F1 over Hard-SDS on 14 of 15 backbone–category settings (+2.18 points, $p = 0.006$ after Holm correction), while Hard-SDS is the better of the two on Micro-F1 by 1.44 points. The same asymmetry appears against strict supervision, where Soft-SDS gains 3.71 Macro-F1 points and loses 1.29 Micro-F1 points (Section 6.5).

B. Configuration of the LLM Baselines

The LLM results in Table 6 are cited from Löffler et al. (2026) and were produced with the following configuration.

Commercial models were queried through the OpenAI API using pinned snapshots gpt-5.2-2025-12-11, gpt-5-mini-2025-08-07 and gpt-5-nano-2025-08-07. Open-weight models of 70b parameters and above were served through OpenRouter (openai/gpt-oss-120b, deepseek-v3.2, qwen3-235b, llama-3-70b). Smaller models were run locally with Ollama 0.14.3 at 8-bit quantization (qwen3:14b-q8_0, llama3.1:8b-instruct-q8_0, gemma3:27b). Of these, Table 6

Table 17

Tail-class weak-label mass across all five teachers, summarizing the per-class breakdown of Table 16 for the remaining backbones. Tail classes are those below the median gold prevalence within their family (11 of the 24 classes). N_{hard} and N_{soft} are summed over the tail. The zero column counts tail classes for which thresholding leaves no above-threshold instance at all. Across the five teachers this happens in 15 of 55 class–backbone combinations.

Teacher	N_{hard}	N_{soft}	Ratio	Tail classes with $N_{\text{hard}} = 0$
BETO	37	118.9	3.2×	3/11
Legal XLM-R	41	135.4	3.3×	3/11
BNE-R base	28	99.0	3.5×	5/11
BNE-R large	47	112.3	2.4×	2/11
XLM-R large	77	109.9	1.4×	2/11

reports the configurations used in the original study’s main comparison; the remaining models are documented in Löffler et al. (2026). Pipelines were implemented in Haystack.

Reasoning effort was disabled for all models. No temperature or nucleus-sampling parameter was set, so each provider’s default sampling behavior applied throughout. Each configuration was run five times with seeds 1–5.

All reported LLM results use dense retrieval with “intfloat/multilingual-e5-large” (1024 dimensions) over a Qdrant store built exclusively from the gold training split, returning the top $k = 5$ passages. No reranking stage was applied.

The RAG prompt instructs the model to act as an automated classifier and to emit only a parsable list of label strings. It then supplies the permitted label set, the retrieved clauses with their gold annotations, and the query clause. The few-shot prompts are static per shot count (1, 3 and 5 shots per class), constructed by spaced sampling over clause length so that examples span the length distribution, and shuffled to mitigate recency bias.

C. Per-Class Performance Breakdown

Table 18 reports test F1 for every class under strict supervision and under soft sequential distillation, together with the support and gold prevalence of each class. It supports the interpretation offered in Section 6.2, that the Macro-F1 gains reflect improved recognition of rare subtypes rather than a uniform improvement.

Splitting each family at its median gold prevalence, weak supervision improves the eleven tail classes by 5.00 F1 on average against 1.60 for the thirteen head classes, a threefold difference. The largest gains fall on rare types. GRAY des det improves by 40.8 points and DARK ch by 19.4.

The pattern is not uniform improvement. Four classes remain at zero F1 under both methods: ILLEGAL cot and ng, DARK er and GRAY des us. Three of these have only two or three test instances, so no method can be meaningfully evaluated on them here. Some classes also regress, most notably GRAY des inf at -11.5 . The gain is therefore tail-weighted rather than universal, which is what the mechanism in Appendix A predicts. Soft targets redistribute supervision toward rare subtypes, raising the equally weighted Macro-F1 at a modest cost to the frequent labels that dominate Micro-F1.

D. Per-Task Significance Tests

Table 19 reports the Soft-SDS against Sup-Base comparison separately for each of the 15 backbone–category tasks, with $n = 5$ seeds per test and Holm correction applied across the 15 comparisons. We include these for completeness but do not base claims on them, because the tests have little power, and instead refer to the aggregate analysis of Section 6.5 for evidence.

CRedit authorship contribution statement

Christoffer Löffler: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Andrea Martínez Freile:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – review & editing, Funding acquisition.

Table 18

Per-class test F1 for BETO, strict supervision against soft sequential distillation, averaged over five seeds. Support is the number of test instances carrying the label. Prevalence is measured on the abusive portion of the training split, and head/tail splits each family at its median prevalence, matching Appendix A. Per-class metrics were recorded by the revision runs rather than the original ones, so column aggregates differ from Table 7 by seed-level noise.

Family	Class	Sup.	Gold %	Sup-Base	Soft-SDS	Δ
ILLEGAL	COT	2	1.1	0.0	0.0	+0.0
	LPC	16	10.2	51.5	56.5	+4.9
	LPC INT	15	9.8	62.9	62.0	-0.9
	LPC JUS	7	4.5	29.7	29.9	+0.2
	LPC PRO	34	21.3	70.0	69.2	-0.8
	NA	67	42.5	82.7	82.4	-0.4
	RC	19	12.4	82.6	87.9	+5.3
	acp	3	2.4	28.0	23.3	-4.7
	ng	2	1.3	0.0	0.0	+0.0
DARK	ch	7	5.4	40.7	60.1	+19.4
	cr	45	37.1	85.1	86.4	+1.4
	er	8	6.4	0.0	0.0	+0.0
	ltd	57	46.3	89.3	89.5	+0.2
	nod	13	11.1	92.7	91.4	-1.3
	ter	8	6.9	30.7	41.4	+10.7
GRAY	bfe	15	12.9	16.4	32.1	+15.7
	des def	9	7.8	93.1	91.0	-2.1
	des det	8	6.6	16.2	57.0	+40.8
	des inf	7	5.1	78.5	67.1	-11.5
	des lic	5	3.7	100.0	100.0	+0.0
	des reser	36	31.0	74.5	77.9	+3.4
	des risk	37	32.0	77.9	79.8	+1.9
	des uni	35	31.7	82.1	75.6	-6.5
	des us	3	2.4	0.0	0.0	+0.0
<i>Mean over tail classes (11)</i>						+5.00
<i>Mean over head classes (13)</i>						+1.60

References

- Ahmed, M., Wen, B., Ao, L., Pan, S., Su, J., Cao, X., Liu, Y., 2024. Towards Robust Learning with Noisy and Pseudo Labels for Text Classification. *Information Sciences* 661, 120160. doi:10.1016/j.ins.2024.120160.
- Akdag, S.H., Cicekli, N.K., 2024. Early detection of fake news on emerging topics through weak supervision. *Journal of Intelligent Information Systems* 62, 1263–1284. doi:10.1007/s10844-024-00852-1.
- Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K., 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning, in: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. doi:10.1109/IJCNN48605.2020.9207304.
- Audibert, A., Gauffre, A., Amini, M.R., 2024. Exploring Contrastive Learning for Long-Tailed Multi-label Text Classification, in: Bifet, A., Davis, J., Krilavičius, T., Kull, M., Ntoutsis, E., Žliobaitė, I. (Eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track*. Springer Nature Switzerland, Cham. volume 14947, pp. 245–261. doi:10.1007/978-3-031-70368-3_15.
- Bernhard, D., Nenadic, L., Bechtold, S., Kubicek, K., 2025. Multilingual Scraper of Privacy Policies and Terms of Service, in: *Proceedings of the Symposium on Computer Science and Law on ZZZ*, ACM, Munich Germany. pp. 55–63. doi:10.1145/3709025.3712215.
- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J., 2023. Spanish Pre-trained BERT Model and Evaluation Data. *arXiv:2308.02976*.
- Cascante-Bonilla, P., Tan, F., Qi, Y., Ordóñez, V., 2021. Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 6912–6920. doi:10.1609/aaai.v35i8.16852.
- Chen, S., Xu, H., Zhang, R., Li, M., Zhang, Y., Lu, Y., Han, B., Cheung, Y.m., Wang, H., 2025. Unlocker: Disentangle the Deadlock of Learning from Label-noisy and Long-tailed Data, in: 39th Conference on Neural Information Processing Systems (NeurIPS 2025), Sydney, Australia.
- Colombo, P., Pires, T.P., Boudiaf, M., Culver, D., Melo, R., Corro, C., Martins, A.F.T., Esposito, F., Raposo, V.L., Morgado, S., Desa, M., 2024. SaulLM-7B: A pioneering Large Language Model for Law. doi:10.48550/arXiv.2403.03883, arXiv:2403.03883.
- Confalonieri, R., Kutz, O., Calvanese, D., Alonso-Moral, J.M., Zhou, S.M., 2024. The role of ontologies and knowledge in Explainable AI: Editorial. *Semantic Web* 15, 933–936. doi:10.3233/SW-243529.

Table 19

Per-task paired comparison of Soft-SDS against the supervised baseline (Macro-F1, percentage points, $n = 5$ seeds). Confidence intervals are t -based on the paired differences and g is Hedges' correction of Cohen's d_z . p_{Holm} is corrected across the 15 comparisons. Significance: $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Backbone	Category	Δ	95% CI	p	p_{Holm}	g
BETO	ILLEGAL	+2.35	[-1.85, +6.54]	0.1949	1.0000	+0.56
	DARK	+7.04	[+1.29, +12.78]	0.0272	0.2995	+1.22
	GRAY	+5.91	[+1.80, +10.01]	0.0162	0.2265	+1.43
Legal XLM-R	ILLEGAL	-0.14	[-7.48, +7.20]	0.9603	1.0000	-0.02
	DARK	+6.99	[+0.33, +13.65]	0.0436	0.4359	+1.04
	GRAY	+1.84	[-5.17, +8.86]	0.5065	1.0000	+0.26
BNE-R base	ILLEGAL	+6.56	[+1.87, +11.26]	0.0178	0.2310	+1.39
	DARK	+6.46	[+1.56, +11.36]	0.0215	0.2584	+1.31
	GRAY	+7.41	[+5.69, +9.14]	0.0003	0.0042**	+4.27
BNE-R large	ILLEGAL	+1.97	[-2.38, +6.32]	0.2767	1.0000	+0.45
	DARK	+1.92	[-0.63, +4.47]	0.1051	0.8412	+0.75
	GRAY	+2.34	[-2.24, +6.93]	0.2292	1.0000	+0.51
XLM-R large	ILLEGAL	-0.94	[-6.31, +4.42]	0.6507	1.0000	-0.17
	DARK	+5.89	[-0.42, +12.19]	0.0606	0.5456	+0.93
	GRAY	+0.12	[-5.37, +5.62]	0.9534	1.0000	+0.02

- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised Cross-lingual Representation Learning at Scale, in: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297. doi:10.1007/BF00994018.
- Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S., 2019. Class-Balanced Loss Based on Effective Number of Samples, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9260–9269. doi:10.1109/CVPR.2019.00949.
- Cui, Z., Zhang, Z., Sun, G., Wu, W., Zhang, C., 2025. Bayesian WeakS-to-Strong from Text Classification to Generation. International Conference on Learning Representations 2025, 29777–29794.
- Dadas, S., Kozłowski, M., Poświata, R., Pereńkiewicz, M., Białas, M., Grębowiec, M., 2024. A support system for the detection of abusive clauses in B2C contracts. Artificial Intelligence and Law doi:10.1007/s10506-024-09408-8.
- Dawid, A.P., Skene, A.M., 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Applied Statistics 28, 20. doi:10.2307/2346806, arXiv:10.2307/2346806.
- de la Maza Gazmuri, I., 2003. Contratos por adhesión y cláusulas abusivas ¿Por qué el Estado y no solamente el mercado? Revista chilena de derecho privado, 109–148.
- DeepSeek-AI, Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., Lu, C., Zhao, C., Deng, C., Xu, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Li, E., Zhou, F., Lin, F., Dai, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Li, H., Liang, H., Wei, H., Zhang, H., Luo, H., Ji, H., Ding, H., Tang, H., Cao, H., Gao, H., Qu, H., Zeng, H., Huang, J., Li, J., Xu, J., Hu, J., Chen, J., Xiang, J., Yuan, J., Cheng, J., Zhu, J., Ran, J., Jiang, J., Qiu, J., Li, J., Song, J., Dong, K., Gao, K., Guan, K., Huang, K., Zhou, K., Huang, K., Yu, K., Wang, L., Zhang, L., Wang, L., Zhao, L., Yin, L., Guo, L., Luo, L., Ma, L., Wang, L., Zhang, L., Di, M.S., Xu, M.Y., Zhang, M., Zhang, M., Tang, M., Zhou, M., Huang, P., Cong, P., Wang, P., Wang, Q., Zhu, Q., Li, Q., Chen, Q., Du, Q., Xu, R., Ge, R., Zhang, R., Pan, R., Wang, R., Yin, R., Xu, R., Shen, R., Zhang, R., Liu, S.H., Lu, S., Zhou, S., Chen, S., Cai, S., Chen, S., Hu, S., Liu, S., Hu, S., Ma, S., Wang, S., Yu, S., Zhou, S., Pan, S., Zhou, S., Ni, T., Yun, T., Pei, T., Ye, T., Yue, T., Zeng, W., Liu, W., Liang, W., Pang, W., Luo, W., Gao, W., Zhang, W., Gao, X., Wang, X., Bi, X., Liu, X., Wang, X., Chen, X., Zhang, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Li, X., Yang, X., Li, X., Chen, X., Su, X., Pan, X., Lin, X., Fu, X., Wang, Y.Q., Zhang, Y., Xu, Y., Ma, Y., Li, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Qian, Y., Yu, Y., Zhang, Y., Ding, Y., Shi, Y., Xiong, Y., He, Y., Zhou, Y., Zhong, Y., Piao, Y., Wang, Y., Chen, Y., Tan, Y., Wei, Y., Ma, Y., Liu, Y., Yang, Y., Guo, Y., Wu, Y., Wu, Y., Cheng, Y., Ou, Y., Xu, Y., Wang, Y., Gong, Y., Wu, Y., Zou, Y., Li, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z.F., Ren, Z.Z., Zhao, Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Huang, Z., Wu, Z., Li, Z., Zhang, Z., Xu, Z., Wang, Z., Gu, Z., Zhu, Z., Li, Z., Zhang, Z., Xie, Z., Gao, Z., Pan, Z., Yao, Z., Feng, B., Li, H., Cai, J.L., Ni, J., Xu, L., Li, M., Tian, N., Chen, R.J., Jin, R.L., Li, S.S., Zhou, S., Sun, T., Li, X.Q., Jin, X., Shen, X., Chen, X., Song, X., Zhou, X., Zhu, Y.X., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Huang, Z., Xu, Z., Zhang, Z., Ji, D., Liang, J., Guo, J., Chen, J., Xia, L., Wang, M., Li, M., Zhang, P., Chen, R., Sun, S., Wu, S., Ye, S., Wang, T., Xiao, W.L., An, W., Wang, X., Sun, X., Wang, X., Tang, Y., Zha, Y., Zhang, Z., Ju, Z., Zhang, Z., Qu, Z., 2025. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. doi:10.48550/arXiv.2512.02556, arXiv:2512.02556.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, Minneapolis, MN.

- Duarte, J.M., Berton, L., 2023. A review of semi-supervised learning for text classification. *Artificial Intelligence Review* 56, 9401–9469. doi:10.1007/s10462-023-10393-8.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnston, J., Saxe, J., Jia, J., Alwala, K.V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Rantala-Year, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardaś, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenheide, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Tan, X.E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., De Paola, B., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymmer, D., Li, D., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzmán, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Damlaj, I., Molybog, I., Tufanov, I., Veliche, I.E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhota, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimploukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N.P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Albiero, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., 2024. The Llama 3 Herd of Models. *arXiv:2407.21783*.
- Freitas, L., Rodrigues, T., Rodrigues, G., Edokawa, P., Farias, A., 2026. Pseudo-labeling for multi-label legal text classification, in: de Freitas, R., Furtado, D. (Eds.), *Intelligent Systems*, Springer Nature Switzerland, Cham. pp. 363–377.
- Galassi, A., Lagioia, F., Jablonowska, A., Lippi, M., 2025. Unfair clause detection in terms of service across multiple languages. *Artificial Intelligence and Law* 33, 641–689. doi:10.1007/s10506-024-09398-7.
- Ghosh, S., Evuru, C.K.R., Kumar, S., Ramaneswaran, S., Sakshi, S., Tyagi, U., Manocha, D., 2023. DALE: Generative Data Augmentation for Low-Resource Legal NLP, in: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore. pp. 8511–8565. doi:10.18653/v1/2023.emnlp-main.528.
- Greco, C.M., Tagarelli, A., 2024. Bringing order into the realm of Transformer-based language models for artificial intelligence and law. *Artificial Intelligence and Law* 32, 863–1010. doi:10.1007/s10506-023-09374-7.
- Gu, Y., Dong, L., Wei, F., Huang, M., 2024. MiniLLM: Knowledge Distillation of Large Language Models, in: *International Conference on Learning Representations (ICLR 2024)*, Vienna, Austria.
- Gutiérrez-Fandiño, A., Armengol-Estapé, J., Gonzalez-Agirre, A., Villegas, M., 2021. Spanish Legalese Language Model and Corpora. *arXiv:2110.12201*.

- Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C.P., Gonzalez-Agirre, A., Armentano-Oller, C., Rodriguez-Penagos, C., Villegas, M., 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural* , 39–60doi:10.26342/2022-68-3, arXiv:2107.07253.
- He, P., Gao, J., Chen, W., 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, in: The Eleventh International Conference on Learning Representations, Kigali, Rwanda.
- Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the Knowledge in a Neural Network. doi:10.48550/arXiv.1503.02531, arXiv:1503.02531.
- Hou, Y., Ma, J., Li, H., Han, B., Yu, J., Jia, Y., 2026. Beyond Distribution Estimation: Simplex Anchored Structural Inference Towards Universal Semi-Supervised Learning, in: Proceedings of the 43 Rd International Conference on Machine Learning, PMLR, Seoul, South Korea.
- Jara Amigo, R., 1999. Ámbito de aplicación de la ley chilena de protección al consumidor: inclusiones y exclusiones, in: Corral Talciani, H. (Ed.), *Derecho del consumo y protección al consumidor: estudios sobre la Ley N° 19.496 y las principales tendencias extranjeras*. Universidad de los Andes, Santiago.
- Joy, U.G., Rahman, M.M., 2026. AdaptiveSSL: A unified semi-supervised learning framework for robust classification via adaptive uncertainty calibration and dynamic labeling. *Knowledge and Information Systems* 68, 41. doi:10.1007/s10115-025-02634-9.
- Juttu, N.P.P., Singireddy, S., Gona, S., Timilsina, S., 2025. Text to Trust: Evaluating Fine-Tuning and LoRA Trade-offs in Language Models for Unfair Terms of Service Detection. doi:10.48550/arXiv.2510.22531, arXiv:2510.22531.
- Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y., 2020. Decoupling Representaitons and Classifier for Long-tailed Recognition, in: International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia.
- Kendall, A., Gal, Y., 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, in: 31st Conference on Neural Information Processing Systems, Long Beach, CA.
- Kim, J., Hur, Y., Park, S., Yang, E., Hwang, S.J., Shin, J., 2020. Distribution Aligning Refinery of Pseudo-label for Imbalanced Semi-supervised Learning, in: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.
- Lee, D.H., 2013. Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, in: ICML 2013 Workshop : Challenges in Representation Learning (WREPL), Atlanta, GA.
- Li, B., 2026. Incompressible Knowledge Probes: Estimating Black-Box LLM Parameter Counts via Factual Capacity. doi:10.48550/ARXIV.2604.24827.
- Li, J., Socher, R., Hoi, S.C.H., 2019. DivideMix: Learning with Noisy Labels as Semi-supervised Learning, in: International Conference on Learning Representations.
- Li, S., Yuan, G., Yang, M., Shen, Y., Li, C., Xu, R., Zhao, X., 2024. Improving Semi-Supervised Text Classification with Dual Meta-Learning. *ACM Transactions on Information Systems* 42, 1–28. doi:10.1145/3648612.
- Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy. pp. 2980–2988.
- Lippi, M., Contissa, G., Jablonowska, A., Lagioia, F., Micklitz, H.W., Palka, P., Sartor, G., Torroni, P., 2020. The Force Awakens: Artificial Intelligence for Consumer Law. *Journal of Artificial Intelligence Research* 67, 169–190. doi:10.1613/jair.1.11519.
- Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H.W., Sartor, G., Torroni, P., 2019. CLAUDETTE: An Automated Detector of Potentially Unfair Clauses in Online Terms of Service. *Artificial Intelligence and Law* 27, 117–139. doi:10.1007/s10506-019-09243-2, arXiv:1805.01217.
- Löffler, C., Martínez Freile, A., Rey Pizarro, T., 2025. Predicting potentially abusive clauses in Chilean terms of services with natural language processing. *Artificial Intelligence and Law* doi:10.1007/s10506-025-09462-w.
- Löffler, C., Rey Pizarro, T., Miranda Vásquez, D.I., Martínez Freile, A., 2026. Retrieval-augmented detection of potentially abusive clauses in chilean terms of service. URL: <https://arxiv.org/abs/2605.26019>, arXiv:2605.26019.
- Massey, F.J., 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* 46, 68–78. doi:10.2307/2280095, arXiv:2280095.
- Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S., 2020. Long-tail learning via logit adjustment, in: International Conference on Learning Representations, Addis Ababa, Ethiopia.
- Mintz, M., Bills, S., Snow, R., Jurafsky, D., 2009. Distant supervision for relation extraction without labeled data, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - ACL-IJCNLP '09, Association for Computational Linguistics, Suntec, Singapore. p. 1003. doi:10.3115/1690219.1690287.
- Molino-Peña, E., Cruz-Lorite, J.M., García, J.M., Ruiz-Cortés, A., 2025. TOSL: An Ontology to Detect Abusive Services, in: OPAL'25: ODRL and beyond: Practical Applications and Challenges for Policy-Base Access and Usage Control, Portorož, Slovenia.
- Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., Ho, D., 2024. MultiLegalPile: A 689GB Multilingual Legal Corpus, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 15077–15094. doi:10.18653/v1/2024.acl-long.805.
- Onan, A., Nasution, A.H., Celikten, T., 2025. Toward Reliable Annotation in Low-Resource NLP: A Mixture of Agents Framework and Multi-LLM Benchmarking. *IEEE Access* 13, 211620–211644. doi:10.1109/ACCESS.2025.3643829.
- OpenAI, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R.K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, C., Fomenko, V., Gariopov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K.G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E.Y., Leclerc, G., Lennon, J.P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O.,

- Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., Zhao, S., 2025. Gpt-oss-120b & gpt-oss-20b Model Card. doi:10.48550/arXiv.2508.10925, arXiv:2508.10925.
- Peng, L., Gu, Y., Dong, C., Wang, Z., Shang, J., 2024. Text Grafting: Near-Distribution Weak Supervision for Minority Classes in Text Classification, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA. pp. 3741–3752. doi:10.18653/v1/2024.emnlp-main.219.
- Ruggeri, F., Lagioia, F., Lippi, M., Torroni, P., 2022. Detecting and explaining unfairness in consumer contracts through memory networks. Artificial Intelligence and Law 30, 59–92. doi:10.1007/s10506-021-09288-2.
- SERNAC, 2021. Resolución Exenta N°931. Aprueba circular interpretativa sobre criterios de equidad en las estipulaciones contenidas en contratos de adhesión de consumo.
- Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A.J., Ananthram, A., Nathan, A., Luo, A., Helyar, A., Madry, A., Efremov, A., Spyra, A., Baker-Whitcomb, A., Beutel, A., Karpenko, A., Makelov, A., Neitz, A., Wei, A., Barr, A., Kirchmeyer, A., Ivanov, A., Christakis, A., Gillespie, A., Tam, A., Bennett, A., Wan, A., Huang, A., Sandjideh, A.M., Yang, A., Kumar, A., Saraiva, A., Vallone, A., Gheorghe, A., Garcia, A.G., Braunstein, A., Liu, A., Schmidt, A., Mereskin, A., Mishchenko, A., Applebaum, A., Rogerson, A., Rajan, A., Wei, A., Kotha, A., Srivastava, A., Agrawal, A., Vijayvergiya, A., Tyra, A., Nair, A., Nayak, A., Eggers, B., Ji, B., Hoover, B., Chen, B., Chen, B., Barak, B., Minaiev, B., Hao, B., Baker, B., Lightcap, B., McKinzie, B., Wang, B., Quinn, B., Fioca, B., Hsu, B., Yang, B., Yu, B., Zhang, B., Brenner, B., Zetino, C.R., Raymond, C., Lugaresi, C., Paz, C., Hudson, C., Whitney, C., Li, C., Chen, C., Cole, C., Voss, C., Ding, C., Shen, C., Huang, C., Colby, C., Hallacy, C., Koch, C., Lu, C., Kaplan, C., Kim, C., Minott-Henriques, C.J., Frey, C., Yu, C., Czarnecki, C., Reid, C., Wei, C., Decareaux, C., Scheau, C., Zhang, C., Forbes, C., Tang, D., Goldberg, D., Roberts, D., Palmie, D., Kappler, D., Levine, D., Wright, D., Leo, D., Lin, D., Robinson, D., Grabb, D., Chen, D., Lim, D., Salama, D., Bhattacharjee, D., Tsipras, D., Li, D., Yu, D., Strouse, D.J., Williams, D., Hunn, D., Bayes, E., Arbus, E., Akyurek, E., Le, E.Y., Widmann, E., Yani, E., Proehl, E., Sert, E., Cheung, E., Schwartz, E., Han, E., Jiang, E., Mitchell, E., Sigler, E., Wallace, E., Ritter, E., Kavanaugh, E., Mays, E., Nikishin, E., Li, F., Such, F.P., Peres, F.d.A.B., Raso, F., Bekerman, F., Tsimplouras, F., Chantzis, F., Song, F., Zhang, F., Raila, G., McGrath, G., Briggs, G., Yang, G., Parascandolo, G., Chabot, G., Kim, G., Zhao, G., Valiant, G., Leclerc, G., Salman, H., Wang, H., Sheng, H., Jiang, H., Wang, H., Jin, H., Sikchi, H., Schmidt, H., Aspegren, H., Chen, H., Qiu, H., Lightman, H., Covert, I., Kivlichan, I., Silber, I., Sohl, I., Hammoud, I., Clavera, I., Lan, I., Akkaya, I., Kostrikov, I., Kofman, I., Etinger, I., Singal, I., Hehir, J., Huh, J., Pan, J., Wilczynski, J., Pachocki, J., Lee, J., Quinn, J., Kiros, J., Kalra, J., Samaroo, J., Wang, J., Wolfe, J., Chen, J., Wang, J., Harb, J., Han, J., Wang, J., Zhao, J., Chen, J., Yang, J., Tworek, J., Chand, J., Landon, J., Liang, J., Lin, J., Liu, J., Wang, J., Tang, J., Yin, J., Jang, J., Morris, J., Flynn, J., Ferstad, J., Heidecke, J., Fishbein, J., Hallman, J., Grant, J., Chien, J., Gordon, J., Park, J., Liss, J., Kraaijeveld, J., Guay, J., Mo, J., Lawson, J., McGrath, J., Vendrow, J., Jiao, J., Lee, J., Steele, J., Wang, J., Mao, J., Chen, K., Hayashi, K., Xiao, K., Salahi, K., Wu, K., Sekhri, K., Sharma, K., Singhal, K., Li, K., Nguyen, K., Gu-Lemberg, K., King, K., Liu, K., Stone, K., Yu, K., Ying, K., Georgiev, K., Lim, K., Tirumala, K., Miller, K., Ahmad, L., Lv, L., Clare, L., Fauconnet, L., Itow, L., Yang, L., Romaniuk, L., Anise, L., Byron, L., Pathak, L., Maksin, L., Lo, L., Ho, L., Jing, L., Wu, L., Xiong, L., Mamitsuka, L., Yang, L., McCallum, L., Held, L., Bourgeois, L., Engstrom, L., Kuhn, L., Feuvrier, L., Zhang, L., Switzer, L., Kondraciuk, L., Kaiser, L., Joglekar, M., Singh, M., Shah, M., Stratta, M., Williams, M., Chen, M., Sun, M., Cayton, M., Li, M., Zhang, M., Aljubei, M., Nichols, M., Haines, M., Schwarzer, M., Gupta, M., Shah, M., Huang, M., Dong, M., Wang, M., Glaese, M., Carroll, M., Lampe, M., Malek, M., Sharman, M., Zhang, M., Wang, M., Pokrass, M., Florian, M., Pavlov, M., Wang, M., Chen, M., Wang, M., Feng, M., Bavarian, M., Lin, M., Abdool, M., Rohaninejad, M., Soto, N., Staudacher, N., LaFontaine, N., Marwell, N., Liu, N., Preston, N., Turley, N., Ansman, N., Blades, N., Pancha, N., Mikhaylin, N., Felix, N., Handa, N., Rai, N., Keskar, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Gleeson, O., Mishkin, P., Lesiewicz, P., Baltescu, P., Belov, P., Zhokhov, P., Pronin, P., Guo, P., Thacker, P., Liu, Q., Yuan, Q., Liu, Q., Dias, R., Puckett, R., Arora, R., Mullapudi, R.T., Gaon, R., Miyara, R., Song, R., Aggarwal, R., Marsan, R.J., Yemirur, R., Xiong, R., Kshirsagar, R., Nuttall, R., Tsiupa, R., Eldan, R., Wang, R., James, R., Ziv, R., Shu, R., Nigmatullin, R., Jain, S., Talaie, S., Altman, S., Arnesen, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Yoo, S., Heon, S., Ethersmith, S., Grove, S., Taylor, S., Bubeck, S., Banerjee, S., Amdo, S., Zhao, S., Wu, S., Santurkar, S., Zhao, S., Chaudhuri, S.R., Krishnaswamy, S., Shuaiqi, Xia, Cheng, S., Anadkat, S., Fishman, S.P., Tobin, S., Fu, S., Jain, S., Mei, S., Egoian, S., Kim, S., Golden, S., Mah, S.Q., Lin, S., Imm, S., Sharpe, S., Yadowsky, S., Choudhry, S., Eum, S., Sanjeev, S., Khan, T., Stramer, T., Wang, T., Xin, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Degry, T., Shadwell, T., Fu, T., Gao, T., Garipov, T., Sriskandarajah, T., Sherbakov, T., Kaftan, T., Hiratsuka, T., Wang, T., Song, T., Zhao, T., Peterson, T., Kharitonov, V., Chernova, V., Kosaraju, V., Kuo, V., Pong, V., Verma, V., Petrov, V., Jiang, W., Zhang, W., Zhou, W., Xie, W., Zhan, W., McCabe, W., DePue, W., Ellsworth, W., Bain, W., Thompson, W., Chen, X., Qi, X., Xiang, X., Shi, X., Dubois, Y., Yu, Y., Khakbaz, Y., Wu, Y., Qian, Y., Lee, Y.T., Chen, Y., Zhang, Y., Xiong, Y., Tian, Y., Cha, Y., Bai, Y., Yang, Y., Yuan, Y., Li, Y., Zhang, Y., Yang, Y., Jin, Y., Jiang, Y., Wang, Y., Wang, Y., Liu, Y., Stubenvoll, Z., Dou, Z., Wu, Z., Wang, Z., 2025. OpenAI GPT-5 System Card. doi:10.48550/arXiv.2601.03267, arXiv:2601.03267.
- Sirbu, I., Popovici, R.A., Caragea, C., Trăuşan-Matu, Ş., Rebedea, T., 2025. MultiMatch: Multihead Consistency Regularization Matching for Semi-Supervised Text Classification, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China. pp. 2792–2808.
- Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C., Cubuk, E.D., Kurakin, A., Li, C.L., 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems 33, 596–608.
- Song, H., Kim, M., Lee, J.G., 2024. Toward Robustness in Multi-Label Classification: A Data Augmentation Strategy against Imbalance and Noise. Proceedings of the AAAI Conference on Artificial Intelligence 38, 21592–21601. doi:10.1609/aaai.v38i19.30157.
- Su, J., 2025. Category-aware self-training for extremely weakly supervised text classification. Expert Systems with Applications 269, 126431. doi:10.1016/j.eswa.2025.126431.
- Suyal, H., Naresh Shrivhare, S., Shrivastava, G., Singh, R., Singhal, A., 2025. IA-KNNR: A Novel Imbalance-Aware Approach for Handling Multi-Label Class Imbalance Problem. IEEE Access 13, 119999–120017. doi:10.1109/ACCESS.2025.3586146.
- Tsai, E., Mangaokar, N., Zheng, B., Zheng, H., Prakash, A., 2025. Harmful Terms and Where to Find Them: Measuring and Modeling Unfavorable Financial Terms and Conditions in Shopping Websites at Scale, in: Proceedings of the ACM on Web Conference 2025, ACM, Sydney NSW Australia. pp. 990–1003. doi:10.1145/3696410.3714573.

- Walker, E., Nowacki, A.S., 2011. Understanding Equivalence and Noninferiority Testing. *Journal of General Internal Medicine* 26, 192–196. doi:10.1007/s11606-010-1513-8.
- Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F., 2024. Multilingual E5 Text Embeddings: A Technical Report. doi:10.48550/arXiv.2402.05672, arXiv:2402.05672.
- Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F., 2021. CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA. pp. 10852–10861. doi:10.1109/CVPR46437.2021.01071.
- Xie, Q., Luong, M.T., Hovy, E., Le, Q.V., 2020. Self-Training With Noisy Student Improves ImageNet Classification, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA. pp. 10684–10695. doi:10.1109/CVPR42600.2020.01070.
- Yan, T., Zhang, X., Luo, Z., 2024. LTACL: Long-tail awareness contrastive learning for distantly supervised relation extraction. *Complex & Intelligent Systems* 10, 1551–1563. doi:10.1007/s40747-023-01226-w.
- Yan, Y., Liu, J., Zhang, B.W., 2025. LabelCoRank: Revolutionizing Long Tail Multi-Label Classification with Co-Occurrence Reranking. *Journal of Artificial Intelligence Research* 83. doi:10.1613/jair.1.16968.
- Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z., 2025a. Qwen3 Technical Report. doi:10.48550/arXiv.2505.09388, arXiv:2505.09388.
- Yang, R., Tian, T., Tian, J., 2025b. Versatile Teacher: A class-aware teacher–student framework for cross-domain adaptation. *Pattern Recognition* 158, 111024. doi:10.1016/j.patcog.2024.111024.
- Yao, W., Yang, W., Wang, Z., Lin, Y., Liu, Y., 2025. Revisiting Weak-to-Strong Generalization in Theory and Practice: Reverse KL vs. Forward KL, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria. pp. 2860–2888. doi:10.18653/v1/2025.findings-acl.148.
- Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., Shinozaki, T., 2021. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling, in: 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Online Conference, Canada.
- Zhang, S., Zhu, C., Li, H., Cai, J., Yang, L., 2024. Gradient-aware learning for joint biases: Label noise and class imbalance. *Neural Networks* 171, 374–382. doi:10.1016/j.neunet.2023.12.028.
- Zhang, Z., Sabuncu, M., 2018. Generalized cross entropy loss for training deep neural networks with noisy labels, in: Advances in Neural Information Processing Systems, Montréal, Canada.