OpenCódice at WomenHelp 2026: An Agentic Approach to Severity and Violence-Type Classification of Gender-Based Violence Reports in Spanish

Francisco-Javier Rodrigo-Ginés[1,*], Jorge Chamorro-Padial[2]

[1]NLP & IR, Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain
[2]OpenCódice Research, Spain

Abstract

We describe the OpenCódice submission to WomenHelp 2026 [1] at IberLEF, a two-subtask shared task on anonymised gender-based violence reports from Northern Mexico. Subtask 1 is a four-class severity classification on a single short report; Subtask 2 is a multi-label classification of the same report along seven violence-type categories. Subtask 1 is evaluated with macro-F1 and Subtask 2 with micro-F1 (with Hamming loss as a secondary statistic) on a blind test partition. The submission was produced inside a semi-autonomous research loop in which a coding agent built on Claude Opus 4.7 implemented and ran every experiment under researcher supervision. The campaign opened with a fine-tuned XLM-RoBERTa-base baseline at 0.4525 macro-F1 and converged on the submitted system through three technical pivots: an input-length pivot from a truncated 256-token window to a length-aware 512-token window that recovered 0.034 macro-F1 from data alone, a backbone pivot from a multilingual XLM-RoBERTa-base to a Spanish-specific BERTIN-RoBERTa encoder of the same parameter count that recovered another 0.014 macro-F1 on the same recipe, and a class-aware ensembling pivot that combined BERTIN and XLM-RoBERTa-large with a parameterless rule deferring to the larger-capacity model whenever either component predicted the rare "severe" class. The submitted ensemble scored 0.6199 macro-F1 on Subtask 1 (2nd of 16, 0.0073 macro-F1 behind the leader) and 0.8700 micro-F1 on Subtask 2 (7th of 15, 0.0179 behind the leader). We present the campaign as a case study of a researcher-gated semi-autonomous loop, in which the agent implemented and ran every experiment while hypothesis selection and the acceptance criterion remained with the researchers; we are explicit that this is an existence demonstration of that division of labour rather than a controlled comparison against a human-only workflow.

Keywords

Gender-based violence, Severity classification, Multi-label violence-type detection, Class-imbalanced fine-tuning, Agentic research, Spanish NLP

1. Introduction

WomenHelp 2026 [1] provides a corpus of anonymised gender-based violence (GBV) reports from organisations in Northern Mexico, annotated by case-management operators along two orthogonal axes. Subtask 1 classifies each report by severity on a four-point ordinal scale (mild, medium, high, severe); Subtask 2 classifies the same report along seven binary violence types (L_0 economic, L_1 physical, L_2 patrimonial, L_3 psychological, L_4 sexual, L_5 vicarious, L_6 none). The official metrics are macro-F1 on Subtask 1 and micro-F1 on Subtask 2 (with Hamming loss as a secondary statistic), computed independently and reported separately on the official leaderboard. The task is part of the IberLEF 2026 forum [2].

The paper documents a campaign (the sequence of experiments run for this single task) inside a semi-autonomous research loop, in the style of recent agentic research configurations [3, 4, 5]. A coding agent built on Claude Opus 4.7 [6] was given the task description, the raw data, GPU access and API credentials, and was responsible for the full implementation life cycle of every experiment: writing training scripts, launching them on remote GPUs, parsing logs, constructing submission files and revising the working hypotheses in light of evaluation returns. The researchers specified the initial

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
frodrigo@invi.uned.es (Francisco-Javier Rodrigo-Ginés); jorge@opencodice.org (Jorge Chamorro-Padial)

methodological framework, exercised a disciplined set of veto and direction-change gates, and made the final decision on every CodaBench upload. We frame this work as a case study of that division of labour rather than as a controlled comparison: the agent implemented and ran every experiment, while hypothesis selection, in the sense of deciding which research direction is worth pursuing, and the acceptance criterion, in the sense of deciding what counts as a successful result, remained with the researchers. We do not measure a human-only baseline and therefore do not claim a quantified speed-up; Section 7 returns to what the campaign does and does not establish.

WomenHelp posed a class-imbalance problem on a long-tailed input. The campaign converged on the submitted system through three technical pivots. The first pivot extended the input window from the default 256 tokens to 512 tokens, motivated by an exploratory analysis showing a 95th-percentile report length of roughly $1\,265$ characters that exceeded the 256-token window for about one quarter of training reports; the change recovered 0.034 macro-F1 on the development partition with no other modification to the recipe. The second pivot replaced the multilingual XLM-RoBERTa-base [7] backbone with the Spanish-specific BERTIN-RoBERTa-Spanish encoder [8] of the same parameter count, on an otherwise identical recipe; the change recovered another 0.014 macro-F1. The third pivot combined the BERTIN backbone with a focal-loss XLM-RoBERTa-large in a two-model class-aware ensemble. The ensemble rule preserves BERTIN's prediction whenever the two models agree, and defers to XLM-RoBERTa-large whenever either model predicts the rare class 3 (severe). The rule is parameterless and is designed around the macro-F1 metric: class 3 is 4.5% of the training partition and dominates the macro average. The ensemble reached 0.6199 macro-F1 on Subtask 1 on the blind test partition, a $+0.067$ absolute gain over the BERTIN single-model submission and a $+0.16$ improvement over the original baseline. The system ranked second of sixteen on Subtask 1 (0.0073 macro-F1 behind the leader) and seventh of fifteen on Subtask 2 (0.0179 micro-F1 behind the leader).

Roadmap. Section 2 positions the work in the literature on GBV and hate-speech detection, on Spanish encoders, and on agentic research. Section 3 describes the task formally and presents the exploratory analysis that informed the system. Section 4 documents the agentic methodology and the three pivots that defined the campaign. Section 5 describes the submitted system, subtask by subtask. Section 6 reports the official scores and a per-class error analysis. Section 7 draws the transferable lessons, and Section 8 concludes.

2. Related Work

The work belongs to the intersection of three lines of research: gender-based violence and sexism detection in Spanish, class imbalance in deep classification, and agentic research with language models.

Sexism and harassment detection in Spanish has been substantially shaped by the EXIST shared-task series, which has been running at IberLEF and, since 2023, at CLEF [9], and which has become a reference benchmark for sexism identification in Spanish and multilingual short-form text. EXIST has been particularly influential in establishing stratified fine-tuning of multilingual encoders as the strong default for Spanish text-classification problems with moderate class imbalance, and, in its 2023 edition, in popularising the "learning with disagreement" evaluation framework that incorporates annotator disagreement into the metric. Broader hate-speech surveys [10] document the recurring difficulty of identifying the minority-yet-most-harmful class. WomenHelp is a different kind of task from EXIST, and it is worth being precise about the difference rather than conflating the two. EXIST classifies social-media messages produced by the perpetrators of sexism, so its target is the abusive language itself; WomenHelp classifies third-party case-management reports written by social-service operators who narrate a violent incident, so its target is a triage judgment over a neutral-register report that may quote, but is not itself, abusive content. The two tasks share the Spanish-language modelling substrate (the same encoders, the same class-imbalance remedies) but not the object of study: detecting online abuse and grading the severity of a documented offline incident are distinct problems, and progress on one does not straightforwardly transfer to the other. We therefore cite EXIST as the reference point for

the Spanish text-classification methods we reuse, not as evidence that WomenHelp is a harassment-detection task; the closer methodological precedent for the task itself is triage classification of clinical or social-service intake text, where the rare "severe" class is the operationally consequential one and macro-F1 is the natural metric.

Class imbalance is the central methodological concern of the campaign. The standard remedies are loss reweighting and oversampling [11]: assigning higher loss weight to under-represented classes, or augmenting them by duplication or synthesis. Focal loss [12] extends the reweighting idea by down-weighting easy negatives, and class-balanced reweighting [13] re-derives the per-class weight from the effective sample size in the training partition. All three were evaluated. Class-balanced weighted cross-entropy was adopted as the Subtask 1 loss because it produced the highest dev macro-F1 of the three; focal loss was not adopted at the BERTIN backbone because, on the present dataset and at the present minority rate, its hyperparameter sensitivity outweighed its theoretical advantages.

Agentic research with language models is the methodological frame for the campaign. The autoresearch sketch of Karpathy [3], the AI Scientist of Lu and colleagues [4], the MLAgentBench evaluation [5], the underlying ReAct loop [14], and the verbal-reinforcement reflection of Reflexion [15] together describe a working configuration in which a language-model agent drives an end-to-end empirical pipeline. Recent code-agent benchmarks [16, 17] show that the underlying language models can resolve real software-engineering issues at a non-trivial rate. The configuration adopted in this campaign accepts the same premise about per-experiment autonomy while reserving hypothesis generation and the acceptance criterion for the human researchers.

3. Task, Data and Exploratory Analysis

This section describes the task in formal terms, presents the splits and basic statistics of the dataset, and walks through the exploratory analysis that informed the system.

3.1. Subtasks and metric

Each instance of WomenHelp 2026 is an anonymised gender-based-violence report written by a case-management operator in Northern Mexico. Subtask 1 attaches a four-class severity label to the report; Subtask 2 attaches a seven-binary violence-type vector to the same report. The four severity classes are 0 mild, 1 medium, 2 high, 3 severe. The seven violence-type labels are L_0 economic, L_1 physical, L_2 patrimonial, L_3 psychological, L_4 sexual, L_5 vicarious, L_6 none. The official metrics are macro-F1 on Subtask 1 and micro-F1 on Subtask 2 (with Hamming loss as a secondary statistic), reported independently.

3.2. Splits and basic statistics

The training partition contains $13\,630$ reports for Subtask 1 and $6\,057$ reports for Subtask 2; the development partition contains $3\,405$ and $1\,513$ reports respectively; the blind test partition contains $4\,219$ and $3\,620$ reports. Reports are written in Spanish, with a mean of 556 characters, a median of 487 characters and a 95th-percentile of $1\,265$ characters; the longest report is roughly $2\,000$ characters. Table 1 reports the severity-class distribution on the training partition.

The imbalance ratio between the modal class (0) and the rare class (3) is roughly $8.6{:}1$. With $n = 3\,405$ development reports, the rare class accounts for 153 instances. The macro-F1 metric is sensitive to the rare class by construction, and any system that fails to recognise it pays a substantial macro-F1 penalty, even if its weighted F1 looks healthy. The ensemble rule of Section 5.1 is designed around this observation.

Table 1
Subtask 1 severity-class distribution on the training partition.

Class	Count	Rate (%)
0 mild	5 304	38.9
1 medium	4 658	34.2
2 high	3 054	22.4
3 severe	614	4.5

Table 2
Subtask 2 violence-type distribution on the training partition (multi-label; positive rates do not sum to 100%).

Label	Positive count	Rate (%)
L_3 psychological	$\sim$ 4 200	$\sim$ 69
L_1 physical	$\sim$ 2 100	$\sim$ 35
L_0 economic	$\sim$ 1 150	$\sim$ 19
L_2 patrimonial	$\sim$ 900	$\sim$ 15
L_6 none	$\sim$ 600	$\sim$ 10
L_4 sexual	$\sim$ 260	$\sim$ 4.3
L_5 vicarious	$\sim$ 200	$\sim$ 3.3

3.3. Why the input length matters

A model fine-tuned with `max_length` $= 256$ truncates roughly 25% of training reports and around 27% of development reports, on the XLM-RoBERTa tokeniser. A model fine-tuned with `max_length` $= 512$ truncates under 5% of training reports. The truncation pattern is not random: longer reports are systematically more severe, because case managers write more when the case is more complex. A truncated model therefore loses signal disproportionately on exactly the minority class that the macro-F1 metric weights most heavily. The empirical effect is consistent with this analysis (dev macro-F1 0.4525 at 256 tokens against 0.486 at 512 tokens with the same recipe), and 512 tokens was adopted as the default for every subsequent experiment.

3.4. Violence-type label distribution (Subtask 2)

The seven violence-type labels of Subtask 2 are evaluated as independent binary classification problems. The label distribution on the training partition is heavily skewed (Table 2): L_3 psychological and L_1 physical account for most positive instances, while L_4 sexual and L_5 vicarious sit below 5% of the training set.

4. Agentic Methodology

This section describes the semi-autonomous research loop and the three technical pivots that defined the campaign.

4.1. The agentic loop

The submission was produced inside a loop with two actors. The agent is a Claude Opus 4.7 model [6] exposed to a working directory through a tool-use interface; it can read and edit files, execute shell commands, run Python scripts on a remote GPU, query OpenAI HTTP endpoints, and download artefacts from CodaBench. Within an iteration the agent implements, launches, monitors and reports on one experiment, proposes the next step in natural language, and updates the working hypotheses in light of the evidence. The researchers hold three gates: the initial choice of candidate model families and seed hypotheses, the approval of any direction change between iterations, and the decision on

which prediction file is uploaded to CodaBench. Every command issued by the agent is recorded in a transcript log that lives alongside the working repository.[1]

The first gate, the seed, shapes everything downstream and warrants an explicit account. For WomenHelp the seed consisted of three priors stated to the agent at the start of the task: that a fine-tuned multilingual or Spanish-specific encoder, rather than zero-shot or few-shot prompting of a general-purpose LLM, was the appropriate model family given several thousand labelled training reports; that the two expected dominant difficulties were class imbalance on the severity scale and input truncation on the long-tailed report length; and that BERTIN-RoBERTa-Spanish and the XLM-RoBERTa family were the candidate backbones to compare. The seed is less a limit on the agent's autonomy than an encoding of domain priors that bounds the search space. The agent remained free to, and did, propose configurations outside the seed: the AutoGluon and GPT-5.4 variants of Section 5 are two such proposals, both generated and then discarded by the agent on the evidence. The seed is therefore best read as the researchers' substantive contribution to a division of labour, not as a constraint the agent could not cross; Section 7 returns to the corresponding risk, that a biased or simply wrong seed would propagate into every iteration with no internal mechanism to detect it.

Figure 1 renders the loop. The agent's four-phase cycle (propose, implement, execute, analyse) runs without per-step approval; the researcher panel on the left lists the gate points, of which "non-trivial API spend" denotes any paid external call above a small per-iteration budget, in practice the GPT-5.4 inference runs of the negative-result experiments, authorised individually. The tool surface on the right is the set of always-available capabilities the agent invokes from the implement and execute phases: the project-level skills (`iberlef-task-explorer` runs the exploratory-analysis pass that produced the length and class-imbalance diagnostics of Section 3; `iberlef-train-xlmr` wraps the fine-tuning recipe), the three Model Context Protocol servers [18] (ACL Anthology, OpenReview and Semantic Scholar), which the agent queries as read-only bibliographic search to ground modelling decisions in prior work, and the remote-GPU and inference services. Appendix A documents every skill and server in full. The transcript band at the top records every command, file diff and metric for replay.

4.2. Cost and interaction profile

The agent's interaction with WomenHelp can be quantified approximately from the transcript log. The figures below are a lower bound: they count only the recorded actions whose input refers to the WomenHelp working directory. On this task the agent issued at least 452 tool invocations, of which 292 were shell commands (around 43 of them launching a Python training or inference run on the remote GPU), 58 file writes, 44 file edits, 19 file reads, and 17 delegated sub-agents. It trained and evaluated eight distinct model configurations, each from its own experiment script (the 256- and 512-token XLM-RoBERTa-base baselines, the BERTIN and XLM-RoBERTa-large fine-tunes, the class-aware ensemble, and the cascade, AutoGluon and GPT-5.4 negative results of Section 5), and produced four CodaBench submission candidates. The GPT-5.4 API was called only for the single few-shot negative-result experiment, on a 500-report development split. External rate limits were not a binding constraint on this task: the binding resource was GPU wall-clock on the rented training instance, not API throughput. We did not instrument finer-grained per-call cost telemetry, and we record its absence as a limitation; a rigorous account of agentic efficiency would log wall-clock and monetary cost per iteration, which we recommend for future editions.

4.3. Pivot 1: input length

The baseline recipe used the default `max_length` $= 256$ tokens that ships with the XLM-RoBERTa classifier example and scored 0.4525 macro-F1 on the development partition. At this window the XLM-RoBERTa tokeniser truncates roughly 25% of training reports and the truncation is non-random: longer reports correlate with higher severity. Extending the window to 512 tokens (no other change to

[1] Code, predictions and the reproduction recipe are available at https://github.com/franfj/WomenHelp.

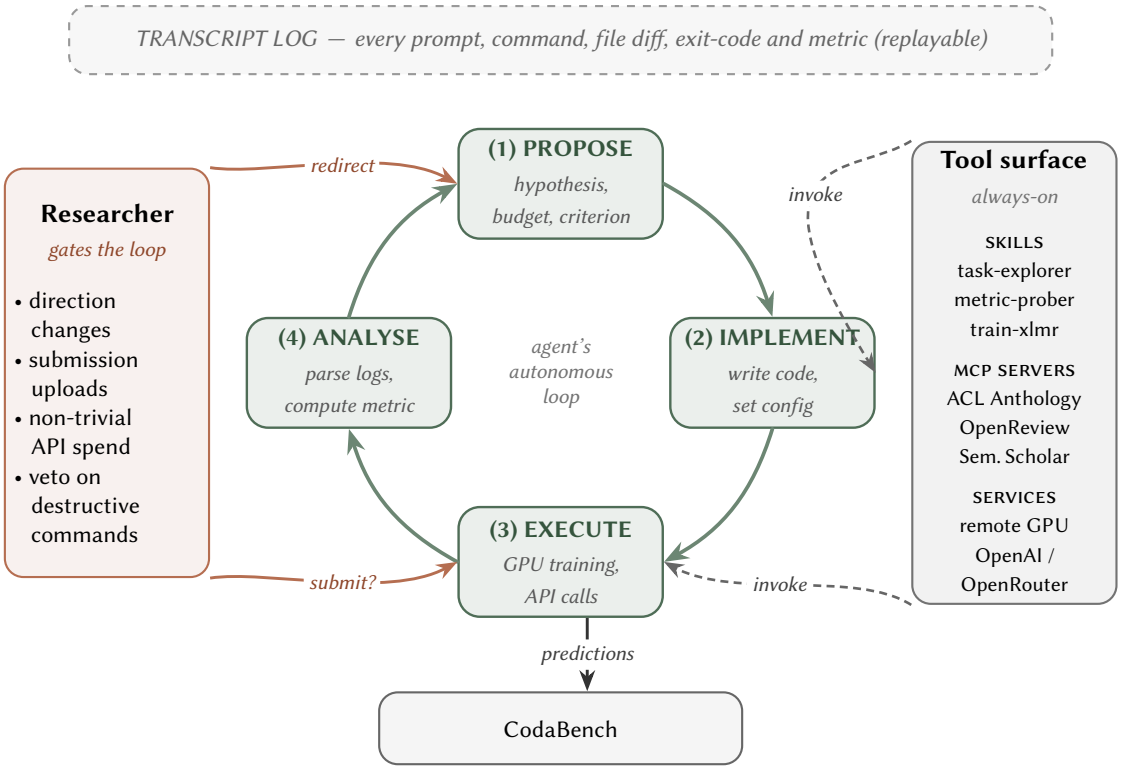

Figure 1: The semi-autonomous research loop. The agent implements, runs and analyses experiments; the researchers gate submissions and direction changes.

the optimiser, the loss or the backbone) moved the dev macro-F1 from 0.4525 to 0.486, a $+0.034$ gain larger than the combined effect of focal loss and $5\times$ minority-class oversampling on the same baseline.

4.4. Pivot 2: backbone substitution (Spanish-specific encoder)

The corpus is monolingual Spanish, which makes a Spanish-specific encoder a stronger prior than a multilingual one at the same parameter count. Substituting BERTIN-RoBERTa-Spanish [8] for XLM-RoBERTa-base on the same recipe (length 512, class-balanced cross-entropy, AdamW, three epochs, seed 42) reached 0.500 macro-F1 on the development partition, a $+0.014$ gain. The size of the gain is small in absolute terms but consistent with the broader empirical pattern that monolingual encoders beat multilingual ones of the same size on monolingual classification.

4.5. Pivot 3: class-aware two-model ensemble

After the length and backbone pivots, two single-model systems were available: a BERTIN single-model submission at 0.5532 macro-F1 on Subtask 1 (test partition), and a focal-loss XLM-RoBERTa-large model at a slightly lower dev score. The two models were combined under a class-aware aggregation rule designed around the macro-F1 metric. The rule preserves BERTIN's prediction whenever the two models agree, and defers to XLM-RoBERTa-large whenever either model predicts the rare class 3 (severe). The class-conditional fallback is justified by two observations: XLM-RoBERTa-large has higher capacity than BERTIN on the rare class, and macro-F1 on a four-class scale is dominated by the class with 4.5% training rate. The rule is parameterless. The ensemble reached 0.6199 macro-F1 on Subtask 1 on the blind test set, a $+0.067$ gain over BERTIN alone and a $+0.16$ gain over the original baseline.

5. System Description

The submitted system fine-tunes two multilingual and Spanish-specific encoders with subtask-specific heads and combines them with a parameterless aggregation rule. Subtask 1 uses a two-model class-aware ensemble; Subtask 2 uses a multi-label decomposition into seven independent binary heads with per-head threshold tuning. We describe each subtask in turn.

5.1. Subtask 1: severity classification

The submitted Subtask 1 system is a two-model agreement ensemble. The first member is BERTIN-RoBERTa-Spanish [8] fine-tuned on the training partition with $\texttt{max\_length} = 512$, class-balanced weighted cross-entropy (effective-number weights [13]), AdamW, learning rate 2×10^{-5}, batch size 16, three training epochs with early stopping on dev macro-F1, label smoothing 0.05, and seed 42. The second member is XLM-RoBERTa-large with focal loss ($\gamma = 2$) and a lower learning rate (1×10^{-5}) to compensate for the larger parameter count. Both models are evaluated on every test instance.

The ensemble rule combines the two predictions as follows. If the two models agree on the predicted class, the BERTIN prediction is emitted. If they disagree, the rule emits XLM-R-large's prediction whenever either model predicts class 3, and BERTIN's prediction otherwise. The rule is parameterless and class-conditional: it costs the ensemble nothing on the high-frequency classes (where BERTIN is at least as accurate as XLM-R-large) and recovers minority-class signal on the cases where the rare class is plausible. The ensemble reached 0.6199 macro-F1 on the blind test partition.

5.2. Subtask 2: violence-type multi-label classification

Subtask 2 is treated as seven independent binary classification heads on the BERTIN backbone, with a shared sentence representation and a per-head linear classifier. The training recipe is the same as in Subtask 1 except that the loss is the sum of seven binary cross-entropies, weighted by the inverse positive rate of each label. Inference applies a per-head threshold tuned on the development partition from the constrained grid $\{0.30, 0.35, 0.40, 0.45, 0.50\}$. The final per-head thresholds favour higher thresholds on the high-frequency labels (L_3, L_1) and lower thresholds on the rare ones (L_4, L_5). The submitted system reached 0.8700 micro-F1 (the official Subtask 2 metric) and a Hamming loss of 0.0730 on the blind test partition, ranking seventh of fifteen participating teams. The leader (jorgegomezn) reached 0.8879 micro-F1, a 0.0179 gap.

5.3. Negative results

Four variants were proposed by the agent during the campaign and discarded on development-partition macro-F1 before any was submitted; in each case the discard decision was the agent's own, endorsed by the researchers at the direction-change gate. Table 3 lists them against the two systems that were kept, the BERTIN single model and the submitted ensemble, so that the discarded options can be read quantitatively alongside the systems that survived; the notes column states the qualitative reason each was dropped.

The cascade decomposes severity into one-vs-rest classifiers with a fallback class that absorbs the residual error and disproportionately hurts macro-F1. The AutoGluon preset reports a cross-validation macro-F1 of 0.69 that collapses to 0.496 on the held-out development partition, a 0.19 gap consistent with its TF-IDF and n-gram features memorising vocabulary in the folds; we do not recommend AutoGluon for text classification on small Spanish corpora. The GPT-5.4 few-shot variant scored 0.133 on the rare class 3, indistinguishable from chance, and declined to classify borderline reports on a stated lack of context. The multi-task head shares one BERTIN backbone across both subtasks and over-fits the more frequent of the two label families, costing 0.018 macro-F1 against the independent-head baseline.

Table 3
Negative results on Subtask 1 (development-partition macro-F1), against the two systems that were kept. All four discarded variants were proposed and then rejected by the agent on dev macro-F1, with the researchers endorsing each rejection at the direction-change gate.

System	Dev macro-F1	Note
Submitted ensemble (kept)	0.57	class-aware two-model rule
BERTIN single model (kept)	0.52	strongest single backbone
Multi-task head (S1+S2)	0.520	shared backbone over-fits frequent labels
AutoGluon best-quality	0.496	CV macro 0.69; 0.19 over-fit gap
Cascade binary classifiers	0.428	fallback class absorbs residual error
GPT-5.4 few-shot	0.331	rare class at chance; declines borderline

Table 4
Official WomenHelp 2026 Subtask 1 leaderboard (severity classification, macro-F1 on the blind test partition). The best score per metric is in bold; our submission is wrapped between rules. Sixteen teams submitted to Subtask 1; the table shows the top eight and the last entry.

Rank	Team	Macro-F1
1	angellabco	**0.6272**
2	**OpenCódice (franrodrigo)**	**0.6199**
3	marcelhr	0.6080
4	arjun108	0.6058
5	omargarcia	0.5936
6	jorgegomezn	0.5915
7	aerojash	0.5750
8	aalgarra	0.5723
16	ymlopez	0.4835

Table 5
Official WomenHelp 2026 Subtask 2 leaderboard (violence-type classification, micro-F1 and Hamming loss on the blind test partition; Hamming is lower-is-better). The best score per metric is in bold; our submission is wrapped between rules. Fifteen teams submitted to Subtask 2; the table shows the top eight and the last entry.

Rank	Team	Micro-F1	Hamming
1	jorgegomezn	**0.8879**	**0.0631**
2	verbanexai	0.8782	0.0692
3	phonghoang	0.8740	0.0711
4	aerojash	0.8735	0.0721
5	pln_uam	0.8715	0.0749
6	marcelhr	0.8702	0.0758
7	**OpenCódice (franrodrigo)**	**0.8700**	**0.0730**
8	angellabco	0.8677	0.0767
15	ymlopez	0.7689	0.1406

6. Results

Tables 4 and 5 report the two official WomenHelp 2026 leaderboards. The submitted system scored 0.6199 macro-F1 on Subtask 1 and 0.8700 micro-F1 on Subtask 2 on the blind test partition, ranking second of sixteen on Subtask 1 (0.0073 macro-F1 behind the leader) and seventh of fifteen on Subtask 2 (0.0179 micro-F1 behind the leader). The asymmetry between the two ranks is the central operational observation of the campaign and is discussed in Section 7.

Table 6
Per-class F1 on Subtask 1 (development partition) for the BERTIN single-model baseline and for the submitted ensemble. The ensemble's gain is concentrated on the rare class 3.

System	class 0	class 1	class 2	class 3	macro
BERTIN (single)	0.58	0.46	0.61	0.42	0.52
Submitted ensemble	0.62	0.50	0.64	0.52	0.57
Δ	+0.04	+0.04	+0.03	+0.10	+0.05

6.1. Per-class breakdown on Subtask 1

A per-class breakdown of Subtask 1 (development partition) quantifies where the gap to the leader lives. Table 6 reports the per-class F1 on Subtask 1 for the BERTIN single-model baseline and for the submitted ensemble. The ensemble lifts the rare class 3 from 0.42 to 0.52 F1, a +0.10 absolute gain, at no cost on the three high-frequency classes.

6.2. Gap analysis

The submitted ensemble is 0.0073 macro-F1 behind the Subtask 1 leader (angellabco, 0.6272). On a 4 219-instance test set the gap corresponds to roughly 31 instances on which the leader picks the right class and the submitted system does not. The most likely candidate for the gap is the rare class 3, since the development breakdown (Table 6) shows that this is the head where the ensemble's gain over the single-model baseline is concentrated. A follow-up edition would either invest in additional class-3 training data (synthetic or external) or refine the ensemble rule to operate at the class-probability level rather than at the argmax level.

On Subtask 2 the submitted system is 0.0179 micro-F1 behind the leader (jorgegomezn, 0.8879) and the field is dense: the gap between rank 1 and rank 7 is the same 0.018 that separates rank 1 from rank 2. Six of the seven teams ahead beat the submission by less than 0.005 micro-F1, which suggests that the multi-label setting rewards label-correlation modelling rather than per-label calibration. The seven independent binary heads in the submitted system treat the labels as conditionally independent given the input, which discards prior information; a chain classifier or a structured-output head that conditions each label on the predictions of correlated labels (e.g., L_3 psychological is positive in 69% of training reports, so its prediction is informative for every other label) is the most plausible next direction.

7. Discussion

We draw three transferable lessons from the WomenHelp campaign that we believe generalise to other text-classification settings with severe class imbalance and long-tailed input lengths.

The first lesson is that an input-length check is the cheapest diagnostic in any text-classification campaign and belongs in the first iteration. On WomenHelp the recovery from extending the input window from 256 to 512 tokens (+0.034 macro-F1) was larger than the combined effect of focal loss and 5× minority-class oversampling on the same baseline. Capacity escalation absorbs engineering effort but does not recover signal that was truncated before it ever reached the classifier.

The second lesson is that monolingual encoders beat multilingual ones of the same size on monolingual Spanish classification by roughly 0.01–0.02 macro-F1 on this corpus. The gain is small but consistent and survives variations in the training recipe. BERTIN is the recommended default backbone for any future Spanish shared task on this prior, with XLM-RoBERTa-base as the fall-through when BERTIN is unavailable.

The third lesson is that class-conditional ensemble rules can beat their constituent classifiers on macro-averaged metrics even when each constituent is individually weaker on the rare class. The two-model rule in the submitted system preserves the stronger model's predictions on the high-frequency classes

and defers to the higher-capacity model whenever the rare class is plausible. The rule is parameterless; its design follows directly from reading the macro-F1 metric as dominated by the rare class.

A fourth observation concerns the provenance of the model choices themselves. The backbones and hyperparameters the loop converged on (BERTIN and XLM-RoBERTa-large, class-balanced cross-entropy, a 512-token window) are not neutral discoveries: they inherit the inductive biases of the pretrained encoders and of the agent's own priors over what a competitive text classifier looks like, both of which were seeded by the researchers. A loop seeded with a different model family, or one whose priors happened to be wrong for this corpus, would converge somewhere else, and the design has no internal mechanism to detect a seed that is competitive but sub-optimal. We read this as the principal liability of the researcher-gated design rather than as a property specific to WomenHelp: the division of labour that lets the agent move quickly also concentrates the risk of a bad starting prior in the one place the agent does not revisit. The mitigations are external, an explicit ablation of the seed and a comparison against a from-scratch agent run, and we did not perform either here.

A final word on the thesis. We opened by stating that agentic AI accelerates empirical NLP research while hypothesis selection and the acceptance criterion remain with the researchers. This campaign is a single case study consistent with the second clause and suggestive of the first, but it is not a controlled comparison against a human-only workflow, and we do not claim one. We did not measure a human baseline on the same task under the same time budget, so we cannot quantify acceleration; the results should be read as an existence demonstration of a researcher-gated loop producing a competitive system through a sequence of diagnosable pivots, not as a measured speed-up over a human researcher. Establishing the acceleration claim quantitatively would require a matched human-versus-agent protocol, which we identify as the necessary next step.

8. Conclusion

This paper described a semi-autonomous submission to WomenHelp at IberLEF 2026. The system, a class-conditional two-model ensemble of BERTIN-RoBERTa-Spanish and XLM-RoBERTa-large with focal loss, scored 0.6199 macro-F1 on Subtask 1 (severity) and 0.8700 micro-F1 on Subtask 2 (violence type) on the blind test partition, ranking second of sixteen on Subtask 1 (0.0073 behind the leader) and seventh of fifteen on Subtask 2 (0.0179 behind the leader). The campaign was defined by three pivots: an input-length pivot from 256 to 512 tokens that recovered 0.034 macro-F1 from data alone, a backbone pivot from a multilingual to a Spanish-specific encoder that recovered another 0.014 macro-F1, and a class-conditional ensembling pivot that recovered 0.067 macro-F1 by deferring to the higher-capacity model whenever the rare class is plausible. The three pivots illustrate a researcher-gated semi-autonomous loop in which the agent executed every experiment while the researchers retained hypothesis selection and the acceptance criterion; we present the campaign as a case study of that division of labour rather than as a measured comparison against a human-only workflow, which we identify as the necessary next step. The code, predictions and reproduction recipe are released at https://github.com/franfj/WomenHelp.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) in order to: grammar and spelling check, paraphrase and reword, and improve writing style. The authors also used ChatGPT to assist with LaTeX formatting of tables and figures. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] N. Martínez-Guevara, G. Bel-Enguix, V. Soto-Mendoza, J. Beltrán, A. Curiel, S. L. Ojeda-Trueba, M. Ramírez-González, H. Gomez-Adorno, Overview of WomenHelp at IberLEF 2026: Classification

of Gender-Based Violence Reports to Help Threatened Women in Northern Mexico, Procesamiento del Lenguaje Natural 77 (2026).
[2] A. Bonet-Jover, J. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.
[3] A. Karpathy, Autonomous research with llm agents ("autoresearch"), https://karpathy.ai/, 2026. Open-source repository sketching an autonomous research loop driven by LLM agents.
[4] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, D. Ha, The AI scientist: Towards fully automated open-ended scientific discovery, arXiv:2408.06292, 2024.
[5] Q. Huang, J. Vora, P. Liang, J. Leskovec, MLAgentBench: Evaluating language agents on machine learning experimentation, in: Proceedings of the 41st International Conference on Machine Learning (ICML), 2024, pp. 20271–20309.
[6] Anthropic, Claude opus 4.7: System card, https://www.anthropic.com/news/claude-opus-4-7, 2026.
[7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451.
[8] J. de la Rosa, E. G. Ponferrada, P. Villegas, P. Gonzalez de Prado Salas, M. Romero, M. Grandury, BERTIN: Efficient pre-training of a Spanish language model using perplexity sampling, in: Procesamiento del Lenguaje Natural, volume 68, 2022, pp. 13–23.
[9] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023: Learning with Disagreement for Sexism Identification and Characterization, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 813–854. URL: https://ceur-ws.org/Vol-3497/paper-070.pdf.
[10] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (2017) 1–10.
[11] J. M. Johnson, T. M. Khoshgoftaar, Survey on Deep Learning with Class Imbalance, Journal of Big Data 6 (2019) 27.
[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
[13] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 9268–9277.
[14] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, 2023. URL: https://arxiv.org/abs/2210.03629.
[15] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: Language agents with verbal reinforcement learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2023, pp. 8634–8652.
[16] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., Evaluating large language models trained on code, arXiv:2107.03374, 2021.
[17] C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, K. Narasimhan, SWE-bench: Can language models resolve real-world GitHub issues?, in: International Conference on Learning Representations (ICLR), 2024.
[18] Anthropic, Introducing the Model Context Protocol, https://www.anthropic.com/news/model-context-protocol, 2024.

A. Tool Surface: Skills and MCP Servers

This appendix documents the tool surface that the coding agent had access to during the campaign.

A.1. Skills used on WomenHelp

The campaign relied on a set of project-scoped skills written in the Claude Code skill format.

- `iberlef-task-explorer`: opens a fresh task by reading its description, downloading the data, and producing an exploratory report. On WomenHelp it flagged the long-tailed length distribution, the 4.5% rate of class 3 and the $\approx 25\%$ truncation rate at the 256-token window, all three of which informed the input-length pivot (Section 4.3).
- `iberlef-baseline`: produces a TF-IDF baseline, a fine-tuned XLM-RoBERTa-base baseline, and an LLM zero-shot prompt. On WomenHelp it produced the 0.4525 macro-F1 baseline and the GPT-5.4 few-shot negative result.
- `iberlef-train-xlmr`: a parameterised fine-tuning skill that wraps XLM-RoBERTa-base, XLM-RoBERTa-large and BERTIN under a common training recipe with optional focal loss, class-balanced weights, label smoothing, and early-stopping.
- `iberlef-ensemble`: combines per-strategy predictions into ensemble submissions, including class-conditional rules of the kind described in Section 4.5.
- `iberlef-codabench-submit`: packages a prediction file in the format the task expects, runs format validation, computes a local score for sanity, and prepares the archive for upload.

A.2. MCP servers and lower-level capabilities

Three Model Context Protocol servers [18] were active during the campaign: the ACL Anthology server, the OpenReview server, and the Semantic Scholar server. They were used to ground the choice of multilingual versus monolingual encoders, to retrieve the focal-loss and class-balanced reweighting literature, and to ground the rare-class-aware ensembling rule. The agent additionally had a sandboxed shell, Python execution on a remote GPU instance, file editing across the repository, OpenAI HTTP endpoints, and the public CodaBench download interface.

B. GPT-5.4 Subtask 1 Prompt (Negative Result)

The failed GPT-5.4 few-shot prompt is reproduced below in English. The actual API call sent four demonstration examples per class, drawn from the training partition, in the user message.

```
You are an expert annotator for gender-based violence case-management reports in Spanish. Given an
    ↪  anonymised report, classify its severity on the following ordinal scale: 0 (mild), 1 (
    ↪ medium), 2 (high), 3 (severe). Severity 0 covers incidents with no immediate risk and no
    ↪ past escalation. Severity 1 covers incidents with verbal or low-intensity psychological
    ↪ violence and no physical contact. Severity 2 covers incidents involving physical contact,
    ↪ threats with weapons, or repeated escalation. Severity 3 covers incidents with serious
    ↪ injury, weapons used, repeated severe escalation, or imminent risk to life.
Respond with ONLY the integer label (0, 1, 2 or 3), nothing else.
```