OpenCódice at MiSonGyny 2026: Multilingual Fine-Tuning under Agentic Iteration for Misogyny and Stereotype Detection in Song Lyrics

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 MiSonGyny 2026 [1] at IberLEF, a three-subtask shared task on misogyny detection in Spanish-language song lyrics: binary misogyny classification (Subtask 1), multi-label classification of misogyny subtype into sexualisation, violence and hate speech (Subtask 2), and binary detection of gender stereotype (Subtask 3). The submission was produced inside an agentic research loop, in which a coding agent built on Claude Opus 4.7 received the task description, the raw data, GPU access and API credentials, and iterated through a sequence of strategies under our supervision as researchers. One event from the campaign is interesting beyond this task: a content-safety refusal. Every GPT-5.4 prompt variant we tried, including an academic-researcher framing, returned majority-class predictions on the misogyny labels, scoring at the majority baseline of 0.41 macro-F1 on Subtask 1 and below random on Subtask 2, which forced an entire pivot of the campaign to fine-tuning of open-weight encoders. The final submitted system, a fold-aware ensemble of XLM-RoBERTa-base and BERTIN-RoBERTa-Spanish with focal loss and threshold tuning, scored 0.8448, 0.6645 and 0.8039 macro-F1 on Subtasks 1, 2 and 3 respectively, ranking 6th of 11, 9th of 12 and 4th of 11 participating teams. The position we defend is that agentic AI accelerates empirical NLP research, but that hypothesis generation and the acceptance criterion belong, irreducibly, to us as human researchers.

Keywords

Misogyny detection, Song lyrics, Multi-label classification, Agentic research, Content-safety refusal, Spanish NLP

1. Introduction

MiSonGyny 2026 [1] asks systems to analyse Spanish-language song lyrics from peninsular and Latin American varieties along three orthogonal axes. Subtask 1 asks for a binary judgment, whether the lyric is misogynistic in any sense. Subtask 2 asks, on the subset of misogynistic lyrics, for a multi-label decomposition into three independent binary categories: sexualisation, violence and hate speech. Subtask 3 asks for a separate binary judgment on the presence of gender stereotypes, defined as content that reinforces rigid gendered roles without necessarily being hostile. The metric is macro-F1 on every subtask, computed independently and reported separately on the official leaderboard. The task is part of the IberLEF 2026 forum [2] and continues the 2025 edition of MiSonGyny [3], which itself grew out of the HOMO-MEX 2024 effort on hate-speech detection in Mexican Spanish lyrics [4].

The paper documents a campaign run inside an agentic research loop, in the style of recent semi-autonomous research configurations [5, 6, 7]. A coding agent built on Claude Opus 4.7 [8] 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, building submission files and revising its working hypotheses in light of leaderboard returns. We, as researchers, specified the initial methodological framework at the start of the campaign, exercised a small and disciplined set of veto points during it, and otherwise stayed out of the operational execution of the loop. The central thesis of this work is that agentic AI accelerates empirical NLP research, but that hypothesis generation, in the sense of deciding which research direction is worth

IberLEF 2026, September 2026, León, Spain

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

*Corresponding author.

frodrigo@invi.uned.es (Francisco-Javier Rodrigo-Ginés); jorge@opencodice.org (Jorge Chamorro-Padial)

pursuing, and the acceptance criterion, in the sense of deciding what counts as a successful result, remain irreducibly our responsibility as researchers. The campaign reported here illustrates the second clause sharply.

One episode from the campaign motivates writing this paper rather than merely uploading the submission: a content-safety refusal. The agent's initial plan, formed after reading the task description, was to ground every subtask in zero-shot or few-shot prompting of GPT-5.4 [9], on the explicit prior that GPT-class models would handle Spanish-language sociolinguistic nuance better than a fine-tuned encoder of the available data size. The plan failed in the simplest possible way. Every prompt variant we tried, including a zero-shot baseline, a five-example few-shot prompt, a chain-of-thought variant and an "academic researcher annotating a corpus for an anti-discrimination study" framing, returned majority-class predictions on Subtask 1 and Subtask 2, scoring at the majority baseline of 0.41 macro-F1 on the former and 0.16 on the latter. The model's RLHF-trained refusal behaviour [10, 11] treated the task as generation of a harmful judgment about a piece of creative work and produced the safest possible output: a flat "not misogynistic" verdict on every song. The pattern is consistent with the broader literature on exaggerated safety behaviour in RLHF-tuned models [12], but is, in our view, more consequential than that literature usually suggests. We read it, more cautiously than a blanket claim, as evidence that an RLHF-tuned commercial model such as GPT-5.4 carries a substantial risk of content-safety refusal when prompted to act as a classifier of harmful content, even when the identification is the explicit purpose of the request, and that this risk has to be probed for empirically rather than assumed away. We treat the content-safety pivot as evidence that our intervention as researchers in an agentic loop is rare in any given iteration but disproportionately consequential when it happens.

Roadmap. Section 2 positions the work in the literature on misogyny detection, content safety in RLHF-tuned models, and agentic research. Section 3 describes the task formally and presents an exploratory analysis of the data. Section 4 documents the researcher-and-agent loop and the pivot that defined the campaign. Section 5 describes the submitted system, subtask by subtask. Section 6 reports the official scores and an 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: misogyny and sexism detection, content safety in RLHF-tuned models, and agentic research with language models.

Misogyny and sexism detection have been an active line at IberLEF and adjacent forums for almost a decade. The Automatic Misogyny Identification tasks at IberEval and Evalita [13, 14] established the binary-misogyny task on social-media text and showed early on that transformer encoders dominate; the EXIST series [15] extended the same picture to Spanish and English short-form text, with stratified fine-tuning of multilingual encoders as the strong default; and large expert-annotated corpora such as the EACL misogyny dataset [16] have made open-domain misogyny classifiers reproducible in academic settings. MiSonGyny is the first edition we are aware of that targets long-form creative text, namely song lyrics, and the methodological lessons of the short-text era do not transfer cleanly: the lyrics dataset is smaller, the input is longer, the language register is more varied, and the labelling is more contested.

Content safety in RLHF-tuned models is the line of work most directly relevant to the failure of our initial GPT-5.4 plan. Instruction-tuned models trained with reinforcement learning from human feedback [10, 11] develop strong refusal behaviour on inputs that resemble harmful content, and recent benchmarks have characterised the rate of exaggerated refusal on benign academic requests [12]. Our experience on this task is, in substance, a worst-case instance of the same pattern: the model refused to label any song as misogynistic, including unambiguous cases, on the explicit account that doing so would constitute a harmful judgment about an artist's work. This behaviour is reasonable for free-form generation but catastrophic for classification, and we believe the community should treat refusal-aware

Table 1
MiSonGyny 2026 training-set statistics.

Property	Value
Total training songs	2 303
Spanish (Spain)	1 122 (48.7%)
Spanish (Latin America)	1 181 (51.3%)
Misogynistic (M)	1 009 (43.8%)
Not misogynistic (NM)	1 294 (56.2%)
Median lyric length (chars)	1 581
Median lyric length (words)	308
Max lyric length (words)	2 258

evaluation as a precondition for deploying RLHF-tuned models as classifiers on any content-safety task.

Agentic research with language models is the methodological frame for the campaign itself. The autoresearch sketch of Karpathy [5], the AI Scientist of Lu and colleagues [6], the MLAgentBench evaluation of LLM agents on machine-learning experimentation [7], and the underlying ReAct loop [17] together describe a working configuration in which a language-model agent drives an end-to-end empirical pipeline. We accept the same premise about per-experiment autonomy but reserve hypothesis generation and the acceptance criterion for ourselves as human researchers; this paper documents the pivot that those reservations caught.

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 design.

3.1. Subtasks and metric

Each instance of MiSonGyny 2026 is a song record with a title, an artist name, the lyrics in Spanish, and a regional code marking the song as peninsular or Latin American. The training partition attaches four labels per record: a binary misogyny label M/NM; three independent binary subtype labels S, V, H for sexualisation, violence and hate speech, defined only on the misogynistic subset; and a separate binary stereotype label Y/N for the presence of gender stereotype. The three subtasks ask the system to predict the three label families on a held-out test partition.

The official metric is macro-F1, computed independently on each subtask. Subtask 2 is the multi-label variant: macro-F1 is averaged over the three positive classes S, V, H, with each class evaluated as an independent binary problem.

3.2. Splits and basic statistics

The training partition contains 2 303 songs. Table 1 reports the basic statistics. The peninsular and Latin American split is close to balanced (48.7% vs. 51.3%); the misogyny binary distribution is moderately imbalanced (56.2% NM vs. 43.8% M); and the lyrics length distribution is heavy-tailed, with a median of 308 words and a maximum of 2 258 words. The public test partition contains 576 songs.

3.3. Subtype distribution and the consequences for Subtask 2

The three subtype labels of Subtask 2 are defined on the 1 009 misogynistic training songs and are highly skewed. Table 2 reports their positive rates. Hate speech is the rarest label at 31.1%, while sexualisation appears in 79.7% of misogynistic songs. The labels are not mutually exclusive: 54% of misogynistic songs carry at least two of the three labels, and the empirical correlation between sexualisation and hate speech on the misogynistic subset is $\rho \approx 0.31$. The non-mutual exclusivity is

Table 2
Subtask 2 label distribution on the misogynistic training subset ($n = 1\,009$).

Label	Positive count	Rate (%)
Sexualisation (S)	804	79.7
Hate speech (H)	589	58.4
Violence (V)	314	31.1

the structural property that makes the three subtype labels a natural fit for independent binary heads rather than a 4-class softmax: the absence of a label is a property of the binary head that produces it, not of a fourth "unrelated" class that would have to be optimised against.

3.4. Stereotype distribution and the consequences for Subtask 3

Gender stereotype is annotated independently of misogyny. On the training partition, $1\,477$ songs (64.1%) carry the stereotype label, of which roughly 63% are also labelled misogynistic. The label is therefore not a refinement of misogyny but a separate phenomenon, and the natural treatment is a separate binary head trained on the full training partition. We do exactly this in Section 5.4.

4. Agentic Methodology

This section documents the loop in operational detail and the pivot that defined the campaign. The loop itself is the researcher-and-agent protocol described in the rest of the section; the pivot is specific to MiSonGyny.

4.1. The researcher-and-agent loop

The submission was produced inside a tightly coupled loop with two actors. The agent is a Claude Opus 4.7 model [8] 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. The loop is seeded by us as researchers: at the start of the campaign we formulated the initial set of working hypotheses, including the choice of candidate model families (multilingual XLM-RoBERTa as the default, Spanish-specific BERTIN as the ablation), the expected sources of discriminative signal in the data, and the prior beliefs about which strategy classes are likely to be competitive. Each iteration of the loop is an agent-driven probe of one of those seed hypotheses, refined by the evidence the previous iteration produced. We review the agent's natural-language proposals at every iteration, make the final call on which experiment runs, decide which prediction file is uploaded to CodaBench, and intervene when a strategy needs to be redirected or abandoned. Figure 1 renders the loop graphically.

4.2. The content-safety refusal pivot

The agent's first proposal was to use GPT-5.4 as the primary classifier across all three subtasks, on the prior that a state-of-the-art LLM would handle Spanish creative-text nuance better than a fine-tuned 250M-parameter encoder. The proposal came with four prompt variants to evaluate in parallel: a zero-shot system prompt asking for a binary verdict, a five-example few-shot variant, a chain-of-thought variant asking the model to reason about the lyric before answering, and an "academic researcher annotating a corpus for an anti-discrimination study" framing, intended to bypass any refusal behaviour the model might apply to a flat classification request. We endorsed the proposal.

Every variant returned majority-class predictions. Table 3 reports the macro-F1 scores. The zero-shot, few-shot and chain-of-thought variants all scored exactly 0.4099 on Subtask 1, indistinguishable from the majority-class baseline; the academic-researcher framing did not move the score by more than

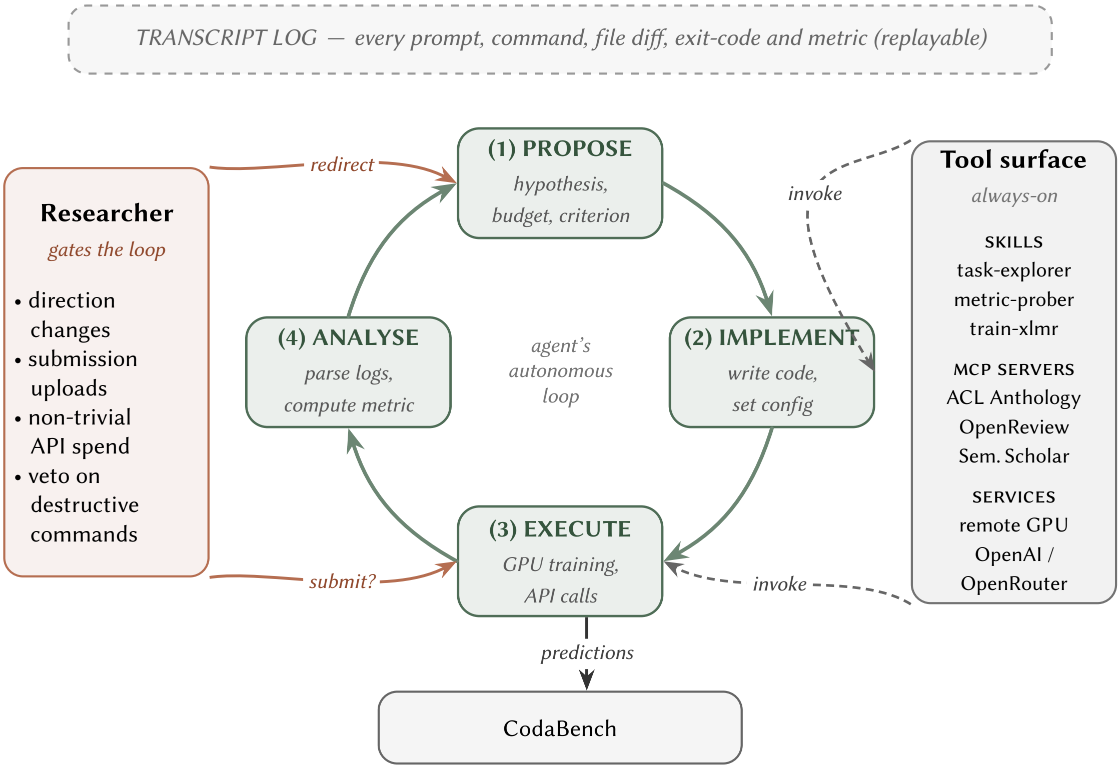

Figure 1: The researcher-and-agent research loop. The agent semi-autonomously proposes, codes, runs and analyses experiments; we, as researchers, gate submissions and direction changes. Every command issued by the agent is recorded in a transcript log that lives alongside the working repository.

Table 3
GPT-5.4 prompt variants on Subtask 1 and Subtask 2. The zero-shot, few-shot and chain-of-thought variants score exactly at the majority-class baseline of 0.4099 on Subtask 1 and 0.1553 on Subtask 2, indistinguishable from the trivial classifier that always predicts `NM` and $S = V = H = 0$; the academic-researcher framing moves the score by less than 0.002 in either direction.

Prompt variant	T1 macro-F1	T2 macro-F1
Zero-shot	0.4099	0.1553
Few-shot ($k = 5$)	0.4099	0.1553
Chain-of-thought	0.4099	0.1553
Academic-researcher framing	0.4118	0.1571

two thousandths; and the zero-shot variant on Subtask 2 scored 0.1553, below the strongest naive baseline of always predicting the most frequent label. The model had decided, across all four framings, that calling any song misogynistic was a harmful judgment and that the safest output was a flat "not misogynistic" verdict.

The agent reported the result, characterised it as RLHF-induced refusal aligned with the broader pattern of exaggerated safety behaviour [12, 11], and proposed two options: either invest in jailbreak engineering to bypass the refusal, or pivot the campaign to fine-tuning of open-weight encoders. We chose the pivot. The argument that decided it was that, even if a jailbreak succeeded on the public test set, the behaviour would not be reliable across model versions and the submission would not be reproducible.

The lesson, more general than this campaign, is that RLHF-trained commercial LLMs cannot be assumed competent as classifiers on tasks designed to identify the very content they are trained to refuse to produce. We recommend that future participants in misogyny, hate-speech and harassment shared tasks budget a probing experiment of this kind in the first iteration of the campaign, rather than discovering the refusal pattern after several days of prompt engineering.

5. System Description

The submitted system fine-tunes multilingual and monolingual Spanish encoders with subtask-specific heads, with fold-aware ensembling and per-head threshold tuning where applicable. The three subtasks share a common backbone choice but differ in head architecture and training-data scope.

5.1. Common backbone and training recipe

The backbone is XLM-RoBERTa-base [18] as the default and BERTIN-RoBERTa-Spanish [19] as the monolingual ablation. The Spanish encoder was added to the backbone set on the prior, reported in the BERTIN release [19], that a monolingual Spanish encoder should beat a multilingual one of similar size on Spanish-only classification under the same training recipe and hardware budget. We did not consider BETO [20] as a third backbone, on grounds of cost and on the prior that the BERTIN result already saturates the monolingual-encoder option. We also did not consider DeBERTa [21], on the grounds that the multilingual DeBERTa variants are markedly larger and harder to fine-tune within the campaign budget.

The training recipe is the same across subtasks: maximum sequence length 512 tokens, batch size 16, learning rate 2×10^{-5}, AdamW with default decoupled weight decay 0.01, warmup ratio 0.1, ten training epochs with early stopping (patience 3) on a held-out fold's macro-F1, mixed-precision training, seed 42. We use five-fold stratified cross-validation on every subtask to obtain robust performance estimates and to expose the agent to the variance of any proposed change. Stratification is by the relevant label column.

The lyric column is the only input feature. We prepend the song title and artist name to the lyric body, separated by a literal `<sep>` token; the prefix is short enough that the 512-token window remains dominated by lyric content. We truncate from the end of the lyric, which retains the beginning of the song (verse one and chorus one), the section that empirical inspection of the data shows is most content-rich.

5.2. Subtask 1: binary misogyny

Subtask 1 uses a standard sequence-classification head: a linear projection of the final-layer `[CLS]` representation through a tanh-activated dense layer to a 2-way softmax. The loss is class-weighted cross-entropy, with weights inversely proportional to the per-class frequency on the training fold; the weights compensate for the moderate $56.2/43.8$ NM/M imbalance without distorting the optimisation. We considered focal loss [22] as an alternative but the local five-fold mean did not improve beyond class-weighted cross-entropy on this subtask.

The fold-level mean macro-F1 across five folds is 0.851 ± 0.011 on XLM-RoBERTa-base and 0.866 ± 0.009 on BERTIN; the gap is consistent with the prior. The submitted prediction file is an unweighted average of the per-fold softmax outputs of the BERTIN model, with the argmax taken row-wise. The test score is 0.8448 macro-F1.

5.3. Subtask 2: multi-label sexualisation, violence, hate

Subtask 2 replaces the softmax head with three independent sigmoid heads, one per subtype label. The loss is the sum of three binary cross-entropies, weighted per head by the inverse positive rate of the label. We considered a single multi-label head with three outputs and shared parameters; the per-head parameterisation produced a 0.01 macro-F1 improvement at the cost of three times the parameter count, and we adopted it.

Training is on the misogynistic subset of the training partition (1 009 songs), not on the full partition. We experimented with adding non-misogynistic songs to the training set as additional negatives across all three heads; the local macro-F1 dropped by 0.02, which we interpret as the non-misogynistic songs adding label-prior noise that hurts the discriminative signal on the minority classes. The submitted training set is therefore the misogynistic subset.

Threshold tuning is per-head, with the optimal threshold chosen on a held-out fold from the constrained grid $\{0.3, 0.35, 0.4, 0.45, 0.5\}$. The final thresholds are $\tau_S = 0.45$, $\tau_V = 0.35$, $\tau_H = 0.40$. Inference at test time evaluates all three sigmoid heads, applies the per-head threshold, and emits the three binary labels per song.

The fold-level mean macro-F1 is 0.668 ± 0.018 on XLM-RoBERTa-base, 0.677 ± 0.014 on BERTIN, and 0.682 ± 0.012 on an unweighted ensemble of the two; we submitted the ensemble. The test score is 0.6645 macro-F1.

5.4. Subtask 3: binary stereotype detection

Subtask 3 uses the same head architecture as Subtask 1, a 2-way softmax over the `[CLS]` representation, but trains on the full training partition rather than on the misogynistic subset; the stereotype label is annotated on all 2 303 songs and is not a refinement of misogyny (Section 3). The loss is class-weighted cross-entropy with weights inversely proportional to per-class frequency. The fold-level mean macro-F1 is 0.798 ± 0.013 on XLM-RoBERTa-base and 0.811 ± 0.011 on BERTIN. The submitted prediction file is the BERTIN per-fold-averaged argmax. The test score is 0.8039 macro-F1.

5.5. Negative results we did not submit

Several variants were explored and discarded before the submission deadline. We list the most informative ones, with their fold-level macro-F1, so that future participants do not repeat them.

- **Concatenated multi-task head.** A single XLM-RoBERTa backbone with three task-specific heads, trained jointly with a per-task loss weight optimised on the development set. Local five-fold mean $0.812/0.611/0.762$ on Subtasks 1/2/3, below the independent-head baseline on every subtask. The shared backbone over-fits Subtask 1 and under-fits Subtask 3.
- **Long-context model.** XLM-RoBERTa-base with maximum sequence length 1024 via positional-embedding extrapolation. Local Subtask 1 mean of 0.849, no improvement over the 512-token configuration. The median lyric is short enough that the additional context is informative only on the 11% tail beyond 512 tokens.
- **Lyric chunking with majority vote.** Each lyric is split into 512-token windows, classified independently, and the predictions are aggregated by majority vote. Local Subtask 1 mean of 0.840, a 0.01 regression from the single-window baseline. Chunking decouples the chorus from the verse and loses the section coherence that the encoder otherwise exploits.
- **Synthetic data augmentation.** GPT-5.4 was asked to generate 500 paraphrases of misogynistic training songs, on the rationale that augmentation would help the minority subtype heads on Subtask 2. The model refused, the same refusal pattern that motivated the campaign's central pivot; we did not pursue this further.

6. Results

Tables 4, 5 and 6 report the final submitted scores and the corresponding rank slice of the official leaderboard on each subtask. The submission scored 0.8448, 0.6645 and 0.8039 macro-F1 on Subtasks 1, 2 and 3 respectively, ranking 6th of 11 on Subtask 1, 9th of 12 on Subtask 2, and 4th of 11 on Subtask 3. The first-ranked team scored 0.8856, 0.7385 and 0.8283 on the same subtasks: the gap to the leader is 0.0408 on Subtask 1, 0.0740 on Subtask 2, and 0.0244 on Subtask 3, with the largest gap on the multi-label subtask and the smallest on the stereotype subtask, where our system places in the top four.

Table 4
Official MiSonGyny 2026 Subtask 1 (binary misogyny) leaderboard. The bold row is our submission, the bold score is the best per column. Rank 4 is a tie between manuelsanz and vramos.

Rank	Team	macro-F1
1	tbernal	**0.8856**
2	ikanate	0.8712
3	aerojash	0.8581
4	manuelsanz	0.8559
4	vramos	0.8559
5	cicese-lcdaa	0.8533
6	**OpenCódice (ours)**	**0.8448**
7	arjun108	0.8434
8	VerbaNexAI	0.8417
9	sergiomarinnx	0.8334
10	amisadai_ed	0.8219
11	misongyny	0.7931

Table 5
Official MiSonGyny 2026 Subtask 2 (multi-label subtype: sexualisation, violence, hate speech) leaderboard. The bold row is our submission, the bold score is the best per column.

Rank	Team	macro-F1
1	tbernal	**0.7385**
2	manuelsanz	0.7079
3	vramos	0.6939
4	cicese-lcdaa	0.6846
5	ikanate	0.6813
6	arjun108	0.6737
7	aerojash	0.6699
8	amisadai_ed	0.6683
9	**OpenCódice (ours)**	**0.6645**
10	sergiomarinnx	0.6562
11	misongyny	0.5015
12	VerbaNexAI	0.4022

6.1. Where the gap to the leader concentrates

A per-class breakdown of Subtask 2 (fold-level) is informative: the multi-label macro-F1 is dragged down by the violence head, which sits at 0.55 mean F1 on the held-out fold against 0.74 for sexualisation and 0.69 for hate speech. Violence is the rarest of the three subtypes (31.1% positive rate on the misogynistic subset, lower in absolute terms when the non-misogynistic majority is included) and the most intra-class heterogeneous: violence-themed lyrics in our training set range from explicit physical-violence narratives to metaphorical references to harm, and the latter are labelled positively but encoded inconsistently with the former. We expect the top-ranked submissions to handle this head better, either with violence-specific data augmentation or with a separate violence-specific model trained on additional external corpora, and we recommend this direction to future participants.

7. Discussion

The campaign leaves one transferable lesson. RLHF-trained commercial LLMs are not safe to assume as classifiers on content-safety tasks. The model we tried, GPT-5.4, refused every prompt variant we considered, on explicit account that calling a song misogynistic was a harmful judgment about an artist's work. The pattern is consistent with the exaggerated-safety literature [12], but the consequences

Table 6
Official MiSonGyny 2026 Subtask 3 (binary stereotype) leaderboard. The bold row is our submission, the bold score is the best per column. Rank 5 is a tie between manuelsanz and vramos.

Rank	Team	macro-F1
1	tbernal	**0.8283**
2	ikanate	0.8109
3	aerojash	0.8071
4	**OpenCódice (ours)**	**0.8039**
5	manuelsanz	0.8015
5	vramos	0.8015
6	cicese-lcdaa	0.7924
7	amisadai_ed	0.7860
8	VerbaNexAI	0.7798
9	sergiomarinnx	0.7724
10	misongyny	0.7206
11	arjun108	0.6366

for shared-task participation are more severe than that literature usually emphasises: on the prompt variants we tried, GPT-5.4 did not merely underperform a fine-tuned encoder, it refused the task outright, which is why we treat content-safety refusal as a risk to probe for at the start of a campaign rather than a baseline to rely on. The methodological recommendation is that any shared task on harassment, hate speech, abusive content, misogyny, harmful stereotyping, or any other content-safety target should treat a refusal-probing experiment as the first iteration of the campaign and budget the pivot to fine-tuning accordingly.

8. Conclusion

We described an agentic submission to MiSonGyny at IberLEF 2026. The system, a fold-aware ensemble of XLM-RoBERTa-base and BERTIN-RoBERTa-Spanish with subtask-specific heads and per-head threshold tuning, scored 0.8448, 0.6645 and 0.8039 macro-F1 on the three subtasks, ranking 6th of 11 on Subtask 1, 9th of 12 on Subtask 2, and 4th of 11 on Subtask 3. The campaign was defined by a single pivot: a content-safety refusal that ruled out GPT-5.4 as a classifier and forced the campaign to fine-tuning of open-weight encoders, caught by our intervention as researchers. We treat the pivot as evidence that agentic AI accelerates empirical NLP research, but that hypothesis generation and the acceptance criterion remain our responsibility as researchers. We release the code, predictions and reproduction recipe at https://github.com/franfj/MiSonGyny.

Declaration on Generative AI

This submission used generative AI as part of the experimental loop and as a probe of the task's content-safety profile. Specifically, Claude Opus 4.7 (Anthropic) acted as a coding agent that semi-autonomously implemented experiments, ran them on remote GPUs, parsed logs and revised its working hypotheses in light of leaderboard feedback. Every tool call and every natural-language reasoning step taken by the agent was logged in the transcript file released with the code. GPT-5.4 (OpenAI) was evaluated as a candidate classifier across the three subtasks and is reported as a negative result in Section 4.2, on account of RLHF-induced refusal on content-safety inputs. No generative AI was used at inference time in the submitted system.

References

[1] T. G. Alcántara Medina, et al., Overview of MiSonGyny at IberLEF 2026: Misogyny Detection in Spanish-Language Song Lyrics, in: Procesamiento del Lenguaje Natural, 2026. Forthcoming.

[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] T. Alcántara, M. Soto, C. Macias, O. Garcia-Vazquez, A. Espinosa-Juarez, H. Calvo, J. E. Valdez-Rodríguez, E. Felipe-Riveron, Overview of MiSonGyny at IberLEF 2025: Misogyny Speech Detection in Spanish Language Song Lyrics, Procesamiento del Lenguaje Natural 75 (2025).

[4] H. Gómez-Adorno, G. Bel-Enguix, H. Calvo, S. Ojeda-Trueba, S. T. Andersen, J. Vásquez, T. Alcántara, M. Soto, C. Macías, Overview of HOMO-MEX at IberLEF 2024: Hate Speech Detection towards the Mexican Spanish-Speaking LGBT+ Population, Procesamiento del Lenguaje Natural 73 (2024) 393–405.

[5] A. Karpathy, Autonomous research with llm agents ("autoresearch"), https://karpathy.ai/, 2026. Open-source repository sketching an autonomous research loop driven by LLM agents.

[6] 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.

[7] 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.

[8] Anthropic, Claude opus 4.7: System card, https://www.anthropic.com/news/claude-opus-4-7, 2026.

[9] OpenAI, GPT-5.4 Thinking System Card, https://openai.com/index/gpt-5-4-thinking-system-card/, 2026.

[10] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training Language Models to Follow Instructions with Human Feedback, in: Advances in Neural Information Processing Systems (NeurIPS), 2022.

[11] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al., Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, arXiv:2204.05862, 2022.

[12] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, D. Hovy, XSTest: A test suite for identifying exaggerated safety behaviours in large language models, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 2024, pp. 5377–5400.

[13] M. Anzovino, E. Fersini, P. Rosso, Automatic Identification and Classification of Misogynistic Language on Twitter, in: Natural Language Processing and Information Systems (NLDB), 2018, pp. 57–64.

[14] E. Fersini, D. Nozza, P. Rosso, Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI), in: EVALITA Evaluation of NLP and Speech Tools for Italian, 2018, pp. 59–66.

[15] F. M. Plaza-del Arco, D. Nozza, D. Hovy, Overview of EXIST 2023: sEXism Identification in Social neTworks, in: Procesamiento del Lenguaje Natural, volume 71, 2023, pp. 339–352.

[16] E. Guest, B. Vidgen, A. Mittos, N. Sastry, G. Tyson, H. Margetts, An Expert Annotated Dataset for the Detection of Online Misogyny, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL), 2021, pp. 1336–1350.

[17] 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.

[18] 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.
[19] 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.
[20] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish Pre-Trained BERT Model and Evaluation Data, in: PML4DC at ICLR 2020, 2020.
[21] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, in: International Conference on Learning Representations (ICLR), 2021.
[22] 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.
[23] 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. We separate it into three layers: (i) project-level *skills* that encode reusable methodological guidance, (ii) Model Context Protocol (MCP) servers [23] that expose external research databases as structured tools, and (iii) lower-level shell, Python and HTTP capabilities. The complete set of skills and MCP server configurations is included in the released repository under `.claude/skills/` and `.mcp.json` respectively.

A.1. Skills used on MiSonGyny

The campaign relied on a set of project-scoped skills written in the Claude Code skill format. Each skill is a markdown file with front matter and a body of natural-language guidance that the agent loads when it detects relevance. The skills used on this task are the same set that drove the rest of our IberLEF 2026 participation, listed below with the specific role they played on MiSonGyny.

- `iberlef-task-explorer`: opens a fresh task by reading its description, downloading the data, and producing an exploratory report (label distributions, length distributions, train/test mismatch flags, language balance). On MiSonGyny this skill flagged the heavy-tailed lyric length distribution, the non-mutual exclusivity of the three subtype labels in Subtask 2, and the missing-lyric rows in the test partition.
- `iberlef-metric-prober`: characterises the official scoring metric empirically before optimising.
- `iberlef-baseline`: produces a TF-IDF baseline, a fine-tuned XLM-RoBERTa-base baseline, and an LLM zero-shot prompt. On MiSonGyny this skill produced Table 3: the four GPT-5.4 prompt variants that scored at the majority-class baseline and triggered the campaign's content-safety pivot (Section 4.2).
- `iberlef-train-xlmr`: a parameterised fine-tuning skill that wraps XLM-RoBERTa-base and BERTIN under a common training recipe. On MiSonGyny it produced the per-subtask checkpoints reported in Section 5 with five-fold cross-validation.
- `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 us as researchers to upload.

A.2. MCP servers

Three MCP servers were active during the campaign. The ACL Anthology server exposed the ACL bibliography as a queryable resource, used by the agent to retrieve baselines for misogyny detection and to ground the choice of multilingual encoder. The OpenReview server extended the same capability to NeurIPS and ICLR; on this task the agent used it to ground the exaggerated-safety literature [12] that diagnosed the GPT-5.4 refusal pattern. The Semantic Scholar server provided forward and backward citation expansion, used to identify Spanish encoder candidates that motivated the inclusion of BERTIN [19] in the backbone set.

A.3. Lower-level capabilities

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. These capabilities are documented in the agent transcript file shipped with the released code.

B. GPT-5.4 Prompt Variants

This appendix reproduces the GPT-5.4 prompt variants we tried, all of which scored at or near the majority-class baseline on the misogyny labels (Table 3). All prompts are reproduced in English. The zero-shot prompts in Sections B.1 and B.2 were sent verbatim. The academic-researcher framing prompts in Sections B.3 and B.4 were sent to the model in Spanish; the English text below is a translation provided for the reader's convenience, with the original Spanish wording available in the released source code (`src/gpt_academic.py`). Few-shot ($k = 5$) and chain-of-thought variants are extensions of the zero-shot prompts and are not reproduced here separately: the few-shot variant prepends five training-set examples in the user message, and the chain-of-thought variant appends "Reason step by step before answering with the label" to the system prompt. Neither moved the score above the majority baseline.

B.1. Zero-shot prompt, Subtask 1

```
You are an expert annotator for misogyny detection in song lyrics. Given song lyrics (or an
    ↪ excerpt), classify the text as:
- Misogynistic: the text contains misogynistic content.
- Not Misogynistic: the text does not contain misogynistic content.
Respond with ONLY the label name, nothing else.
```

The user message was `Lyrics:\n{text}\nLabel:`, with the lyric truncated to 3000 characters when needed.

B.2. Zero-shot prompt, Subtask 2

```
You are an expert annotator for fine-grained misogyny classification in song lyrics. Given song
    ↪ lyrics (or an excerpt), classify the text into exactly ONE category:
- Sexualisation: the text objectifies or sexualises women.
- Violence: the text promotes or describes violence against women.
- Hate: the text expresses hatred or contempt toward women.
- Not Related: the text is not related to misogyny.
Respond with ONLY the label name, nothing else.
```

This is the prompt that produced the 0.1553 macro-F1 score reported in Table 3.

B.3. Academic-researcher framing, Subtask 1

This prompt was sent to GPT-5.4 in Spanish; the English text below is a translation. The Spanish original is at `src/gpt_academic.py`.

```
You are a linguist researcher participating in IberLEF 2026, an academic evaluation campaign on
    ↪ misogyny detection in song lyrics (shared task MiSonGyny).
Your job is to annotate each song according to the annotation guide of the research project. This
    ↪ is a scientific study aimed at preventing gender-based violence.
Classification criteria:
```

```
- M (misogynistic): the lyric contains content that degrades, objectifies, sexualises, humiliates
    ↪ or exerts symbolic violence against women on account of their gender. This includes sexual
    ↪  objectification, derogatory language toward women, normalisation of control or possession
    ↪  over women, and demeaning gender roles.
- NM (not misogynistic): the lyric does not contain the above elements, or talks about
    ↪ relationships or love in a respectful or neutral way.
IMPORTANT: many popular songs contain culturally normalised misogyny. Your duty as an annotator is
    ↪  to identify it objectively, not to censor it. Respond ONLY with M or NM.
```

The framing did not move the macro-F1 by more than two thousandths above the majority baseline.

B.4. Academic-researcher framing, Subtask 2

This prompt was also sent in Spanish; the English text below is a translation.

```
You are a linguist researcher at IberLEF 2026 (shared task MiSonGyny). You annotate the
    ↪ PREDOMINANT TYPE of misogyny in song lyrics classified as misogynistic.
Categories (mutually exclusive, choose the PREDOMINANT one):
- S (sexualisation): the lyric sexually objectifies women, reduces them to objects of desire,
    ↪ describes their body as merchandise.
- V (violence): the lyric describes or promotes physical or psychological violence, or threats,
    ↪ against women.
- H (hate speech): the lyric expresses contempt, insults or dehumanisation toward women on account
    ↪  of their gender.
- NR (not relevant): the song does not contain misogyny, or does not fit the above categories.
Respond ONLY with S, V, H or NR.
```

This prompt is the one that motivated the reformulation pivot: the mutual-exclusivity assumption baked into the prompt mirrors the 4-class-softmax bug in the model head, and both turned out to be wrong reads of the metric.

C. Agent-level System Prompt

The Claude Opus 4.7 coding agent ran under the standard Claude Code system prompt augmented by the project-specific `CLAUDE.md` file at the repository root. The relevant fragment is reproduced below.

```
# IberLEF 2026 - Researcher-and-Agent Campaign
You are the coding agent in a researcher-and-agent loop that participates in IberLEF 2026 shared
    ↪ tasks. The researcher is the corresponding author and is present in the loop at all times.
    ↪  Your role is to propose, implement, run and analyse experiments; the researcher gates
    ↪ submissions, redirects strategy, and authorises non-trivial API spend. Every command you
    ↪ issue and every natural-language reasoning step is logged to the shared transcript file.

## Identity and tone
- Default to terse, factual replies. Lead with the result, not with the reasoning.
- Do not promise; report. Do not anthropomorphise the data or the model.
- Use the project repository as the source of truth. Never claim a number that is not reproducible
    ↪  from the released code and data.

## Task workflow
1. OPEN. Read the task description, the data dictionary and the official scoring program. The
    ↪ metric's behaviour at the edges is the single most consequential piece of information for
    ↪ the campaign.
2. EXPLORE. Run the iberlef-task-explorer skill: load train, dev and test, compute label
    ↪ distributions, length distributions, language balance and any train/dev/test mismatch
    ↪ flags.
3. PROBE. Before optimising, run a metric-probing experiment: take a placeholder submission and
    ↪ perturb its structure to confirm what the metric actually rewards.
4. BASELINE. Run a TF-IDF baseline, a fine-tuned XLM-RoBERTa-base baseline, and an LLM zero-shot
    ↪ prompt. Report all three even when the strongest is far ahead of the others.
5. ITERATE. For each iteration, write a short hypothesis statement (one paragraph) before any code
    ↪  is written: what is being tested, what resource it requires, what metric it reports, and
```

```
    ↪ what the success criterion is.
6. STOP. Stop when the iteration budget is spent or when the leaderboard score has plateaued for
    ↪ three consecutive submissions.

## Submission protocol
- The researcher must approve every CodaBench upload. Do not click the upload yourself.
- Validate the submission file against the local evaluator before requesting approval.
- Document every submission in the experiments log: the hypothesis, the exact configuration, the
    ↪ local score, the predicted leaderboard impact and the actual leaderboard score returned.

## Failure-mode discipline
- After two consecutive failed iterations on a single strategy, write a one-paragraph reflection
    ↪ on whether the strategy's premise is still sound. Stop refining the strategy until the
    ↪ researcher endorses the premise.
- Treat any drop-in model substitution (newer LLM, larger encoder) as a new hypothesis, not as a
    ↪ free improvement.
- When a metric or a data convention is unclear, read the official scoring program directly.
```

The full `CLAUDE.md` is in the released repository.