Team bionlp at GRACE 2026: A Multi-Model Geometric Ensemble for Argumentative Component Detection in Spanish Clinical Notes

Cristea Alexandru- Marian[1], Neagoe Mihai Alexandru[1], Ionescu Costin-Ioan[1], Radu Stefan Andrei[1] and Visalon Mihai Andrei[1]

[1]*University of Bucharest, Romania*

Abstract

This paper describes the system developed for the IberLEF GRACE 2026 Track 1 shared task on sentence-level argumentative component detection in Spanish Electronic Health Records (EHRs), focusing exclusively on Subtask 1 (sentence-level component identification). The task involves classifying each sentence into one of four categories: *None*, *Premise*, *Claim*, and *MajorClaim*. The primary obstacle is extreme class imbalance; the critical *MajorClaim* class constitutes only about 1% of the training data. Standard cross-entropy optimization naturally biases predictions toward the majority classes, causing a collapse in minority-class recall. To counter this, we design a geometric mean ensemble of four complementary transformer models that explicitly separates majority-class robustness from minority-class sensitivity. A stable *Anchor* model provides high-precision representations for the dominant classes, while a *Specialist* model, trained with Class-Balanced Focal Loss and in-batch difficulty-aware oversampling, aggressively targets the rare *MajorClaim* category. The models are fused through a weighted geometric mean, which penalises overconfident but inconsistent predictions and preserves accurate minority signals. On the official blind test set, our system achieves an official Strict Macro F_1 of 36.83% (0.3683 as reported by the organizers, scaled by 100), ranking fourth overall and second on the argumentative component detection subtask, while yielding an internal evaluation of 0.7508 under identical local script parameters. Extensive ablation shows that pure neural fusion consistently outperforms hand-crafted heuristic post-processing; manual interventions such as cascade thresholds, keyword-based probability boosts, and document-level uniqueness constraints all reduce performance.

Keywords

Argumentation Mining, Clinical NLP, Class Imbalance, Ensemble Learning, Spanish EHR

1. Introduction

Electronic Health Records (EHRs) contain dense narrative descriptions of patient history, diagnostic reasoning, and therapeutic plans. Hidden within this unstructured text are complex argumentative structures: observed symptoms act as *premises*, diagnostic interpretations function as *claims*, and final clinical decisions constitute overarching *major claims*. Automatically extracting these components is essential for clinical decision support, medical audit, and explainable AI [1, 2].

Sentence-level argumentative classification in clinical Spanish presents several challenges. Clinical notes exhibit informal syntax, frequent abbreviations, domain-specific terminology, and subtle discourse boundaries. More fundamentally, the data suffer from extreme label imbalance. In the GRACE Track 1 dataset, *Premises* (supporting evidence) are abundant, while *MajorClaims* (ultimate conclusions) represent only about 1% of the sentences.

When transformer-based encoders are fine-tuned on such skewed distributions with standard cross-entropy, they tend to collapse the minority class: the model memorizes superficial lexical patterns of the majority classes and largely ignores the rarest category. Simply re-balancing the data by oversampling can lead to overfitting in high-dimensional embedding spaces, while naïvely up-weighting the rare class in the loss may destabilise the decision boundaries for the other classes.

IberLEF 2026, September 2026, León, Spain

alexandru-marian.cristea@s.unibuc.ro (C. A. Marian); mihai-alexandru.neagoe@s.unibuc.ro (N. M. Alexandru); costin-ioan.ionescu@s.unibuc.ro (I. Costin-Ioan); andrei.radu11@s.unibuc.ro (R. S. Andrei); andrei-mihai.visalon@s.unibuc.ro (V. M. Andrei)

In this paper, we propose a geometric mean ensemble that decouples these conflicting objectives. Instead of asking a single model to be simultaneously robust on common classes and sensitive to rare ones, we deploy four specialised transformers with complementary roles. A stable *Anchor* model is trained conventionally to preserve high precision on *None* and *Premise*. A *Specialist* model is trained aggressively with Class-Balanced Focal Loss and forced in-batch oversampling to maximise recall for *MajorClaim*. Two additional models provide logical and broad clinical coverage. During inference, the softmax probabilities of the four models are merged via a weighted geometric mean. This multiplicative fusion naturally suppresses overconfident errors from the Specialist while amplifying its correct rare-class predictions.

On the official blind test set, our system achieves an official Strict Macro F_1 of 36.83% (0.3683 as reported by the organizers, scaled by 100), ranking fourth overall and second on the argumentative component detection subtask, while yielding an internal evaluation of 0.7508 under the official 3-class active argumentative macro validation scope. Extensive ablation shows that pure neural fusion consistently outperforms hand-crafted heuristic post-processing; manual interventions such as cascade thresholds, keyword-based probability boosts, and document-level uniqueness constraints all reduce performance.Note that our participation and this manuscript are focused solely on Track 1, Subtask 1 (component identification), and do not address relation detection (Subtask 2).

2. Background

2.1. Task and Dataset

The GRACE 2026 Track 1 shared task provides clinical documents segmented into sentences. Each sentence s_i must be assigned a label $y_i \in \{0, 1, 2, 3\}$ corresponding to *None*, *Premise*, *Claim*, or *MajorClaim*. Evaluation is performed with the Strict Macro F_1 score averaged across the categories, ensuring that performance on the rare *MajorClaim* class carries equal weight.

The dataset is composed of randomized controlled trial (RCT) abstracts, segmented and annotated for argumentative structure from [3] and [4].The extreme challenge of this task lies in its severe label imbalance: while background narrative and observational *Premises* are abundant, ultimate clinical conclusions (*MajorClaims*) are exceedingly rare.

The exact active component distribution, derived directly from the official distributed JSON files, is shown in Table 1. While the task overview cites 60 development and 389 test abstracts, the released validation file comprises 50 abstracts, and the blind test file is deliberately padded to 2,474 abstracts via the organizers' Test + Background anti-probing strategy.

Table 1
Argumentative component distribution of the GRACE 2026 Track 1 dataset. Counts are derived directly from the official task files. The hidden test set utilizes a background-padding strategy to prevent leaderboard probing.

Split (Abstracts)	Premise	Claim	MajorClaim	Total Components
Train ($N = 350$)	1,536 (67.78%)	666 (29.39%)	64 (2.82%)	2,266
Development ($N = 50$)	218 (66.87%)	99 (30.36%)	9 (2.76%)	326

When incorporating the unannotated background sentences (*None* class) into the sequence distribution, the imbalance ratio between *None* and *MajorClaim* exceeds 70:1. This extreme skew creates a fundamental tension: a standard classifier optimized for global accuracy will view *MajorClaim* as noise and rarely predict it, while a model trained exclusively to recover *MajorClaim* will produce many false positives that damage the precision of *Claim* and *Premise*.

2.2. Related Work

Argumentation mining has been extensively studied on well-structured texts such as persuasive essays [5, 6], and is increasingly being applied to the clinical domain [7]. Recent advancements in Spanish

biomedical transformers, such as `bsc-bio-ehr-es` [8], offer robust foundations for clinical sequence classification.

Class imbalance in NLP has been addressed through data-level methods (oversampling, undersampling) and algorithm-level techniques (weighted loss functions). Focal Loss [9] originally proposed for object detection, dynamically down-weights well-classified examples, focusing optimization on hard instances. Its class-balanced variant [10] further incorporates inverse-frequency weights to handle extreme imbalance. Stochastic Weight Averaging (SWA) [11] has been shown to improve generalization by averaging model weights along the optimization trajectory, leading to wider optima. Geometric ensembling, while less common than arithmetic averaging, is known to be more robust to outlier predictions because it penalises large deviations more heavily. Our work combines these elements into a principled ensemble architecture and rigorously tests the contribution of each component.

3. System Architecture

To avoid the conflicting demands of majority-class precision and minority-class recall, we distribute the learning objectives across four complementary pre-trained transformers and fuse their outputs via a weighted geometric mean. Figure 1 illustrates the inference pipeline.

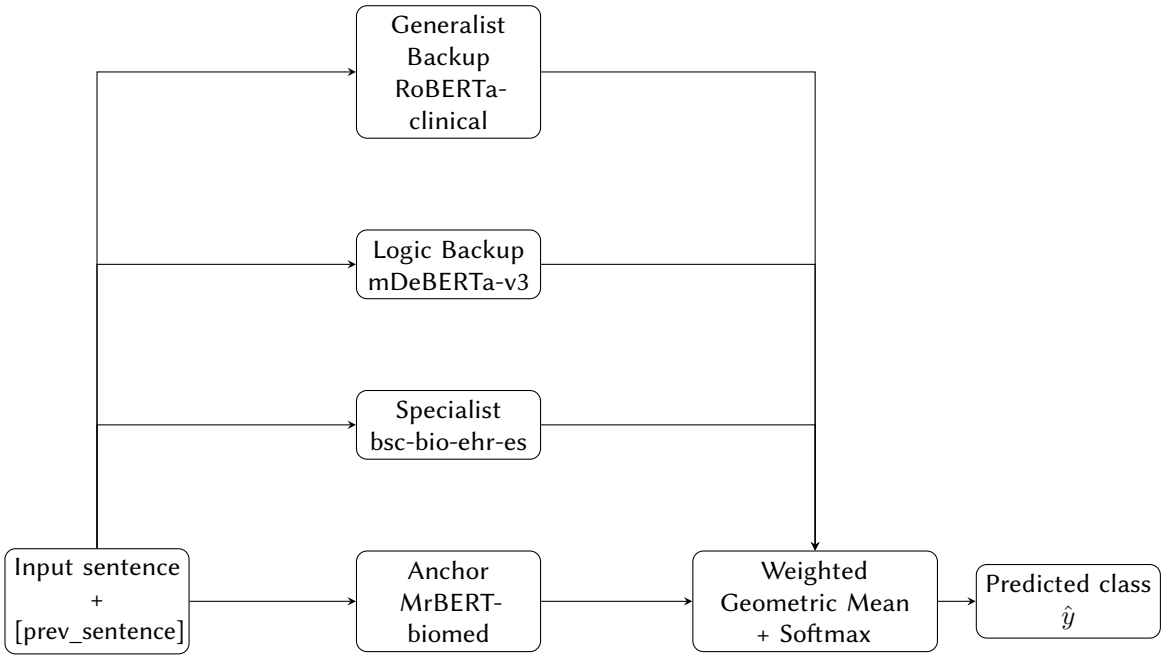

Figure 1: Inference architecture of the geometric mean ensemble. Four heterogeneous pre-trained models process the same input and their softmax logits are fused via a weighted geometric mean. No hand-crafted post-processing is applied.

3.1. Role-Based Model Selection

We select four models with diverse pre-training backgrounds to maximize representational diversity:

- **Anchor (MrBERT-biomed [12]):** A large Spanish biomedical language model trained on clinical and biomedical corpora. It is fine-tuned with standard cross-entropy and serves as the stable, high-precision backbone for the majority classes.
- **Specialist (bsc-bio-ehr-es [12]):** A model specifically pre-trained on Spanish electronic health records. It is fine-tuned with Class-Balanced Focal Loss and forced oversampling to aggressively target the *MajorClaim* category.

- **Logic Backup (`mDeBERTa-v3` [13]):** This component integrates advanced cross-lingual logical discourse parsing features to resolve complex structural transitions [14].
- **Generalist Backup (RoBERTa-clinical [15]):** A broad-coverage clinical language model that provides complementary domain knowledge and prevents over-indexing on the idiosyncrasies of the primary models.

3.2. Minority-Class Optimization for the Specialist

The Specialist model is specifically configured to overcome the 1% bottleneck. Standard cross-entropy is replaced by Class-Balanced Focal Loss [10]:

$$\mathcal{L}_{\mathrm{FL}}(p_y) = -\alpha_y (1 - p_y)^{\gamma} \log(p_y), \tag{1}$$

where $\gamma = 2$ is the focusing parameter and $\alpha_y = (1 - \beta)/(1 - \beta^{n_y})$ is the class-balanced weight computed from the effective number of samples with $\beta = 0.9999$ [10], n_y being the count of class y. The modulating factor $(1 - p_y)^{\gamma}$ reduces the contribution of well-classified examples, concentrating the gradient on difficult cases—typically the rare *MajorClaim* instances.

In addition, we employ an expected minority-class sampling weight distribution strategy. A PyTorch `WeightedRandomSampler` is designed to assign selection probabilities inversely proportional to class prevalence, targeting an average, non-deterministic representation of 4 *MajorClaim* instances per mini-batch of 16. This configuration guards heavily against gradient starvation across sequential updates without imposing brittle, explicit batch-level constraints that could undermine the representational diversity of the common classes.

3.3. Geometric Mean Fusion

At inference time, each model m outputs a softmax distribution $p_m(c \mid x)$. We fuse these distributions using a weighted geometric mean:

$$\hat{p}(c) = \frac{\exp\left(\sum_m w_m \log p_m(c)\right)}{\sum_{k=0}^{3} \exp\left(\sum_m w_m \log p_m(k)\right)}. \tag{2}$$

The ensemble weights were established empirically: Anchor ($w = 0.25$), Specialist ($w = 0.35$), Logic ($w = 0.15$), and Generalist ($w = 0.25$). Reflecting our optimization goals, the Specialist receives the highest weight ($w = 0.35$) because it is explicitly trained to maximize recall on the rare *MajorClaim* category. In earlier iterations utilizing arithmetic averaging, we were forced to assign a 60% weight to the Anchor model merely to suppress the Specialist's false positives. By transitioning to a geometric mean, the multiplicative formulation inherently penalizes overconfident, discordant predictions. This mathematical safety net allowed us to safely elevate the Specialist's weight to 0.35, maximizing minority-class recall without devastating the precision of the majority classes.

Specifically, Stochastic Weight Averaging collects multiple checkpoints along the optimization trajectory and averages their weights, which has been shown to converge to flatter local minima, improving generalization on unseen data. Similarly, our Multi-Sample Dropout architecture acts as an implicit ensemble within a single forward pass by applying $K = 5$ independent dropout masks to the pooled representation before the classification head with a dropout probability of $p = 0.20$. This forces the model to learn features that are robust to the loss of specific activations, effectively preventing the Specialist from over-indexing on noisy identifiers. This combination of weight-space averaging and activation-space regularization ensures that the ensemble remains stable even when training on the highly variable gradients produced by Class-Balanced Focal Loss.

3.4. Concrete Walk-Through Calibration Example

To illustrate the behavior of the multiplicative consensus mechanism under severe class asymmetry, we trace a single representative target sentence: *"Por tanto, se concluye un diagnóstico de neumonía*

adquirida en la comunidad" (Gold Label: *MajorClaim*). The full four-model logit configurations and normalized post-softmax probability distributions (p_m) are explicitly defined in Table 2.

Table 2
Model-specific probability assignments for the representative clinical conclusion example.

Model Structure	**Weight (w_m)**	p_m(**None**)	p_m(**Premise**)	p_m(**Claim**)	p_m(**MajorClaim**)
Anchor (`MrBERT-biomed`)	0.25	0.0300	0.0200	0.8500	0.1000
Specialist (`bsc-bio-ehr`)	0.35	0.0100	0.0400	0.0500	0.9000
Logic Backup (`mDeBERTa-v3`)	0.15	0.0500	0.0500	0.4000	0.5000
Generalist (`RoBERTa-clinical`)	0.25	0.0200	0.0300	0.8000	0.1500

Applying the weighted linear arithmetic mean yields an uncalibrated *MajorClaim* probability of 0.4525, causing the baseline ensemble to assign the sentence incorrectly to the dominant *Claim* class (0.4900). This occurs because the Anchor and Generalist models are overly confident in *Claim*, overpowering the Specialist's correct signal.

Conversely, executing our weighted geometric formulation ($\sum w_m \ln p_m(c)$) using the explicit weights and distributions from Table 2 yields the true raw geometric log values of: $\tilde{g}_{\text{None}} = -3.9159$, $\tilde{g}_{\text{Premise}} = -3.4307$, $\tilde{g}_{\text{Claim}} = -1.2823$, and $\tilde{g}_{\text{MajorClaim}} = -1.1909$. Passing these precise un-normalized logs through a standard exponential softmax transformation ($\hat{p}(c) = \exp(\tilde{g}_c)/\sum \exp(\tilde{g}_k)$) produces the final, calibrated ensemble probability distribution: $\hat{p}(\text{None}) = 0.0314$, $\hat{p}(\text{Premise}) = 0.0511$, $\hat{p}(\text{Claim}) = 0.4378$, and $\hat{p}(\text{MajorClaim}) = 0.4796$.

By penalizing the extreme disagreement among models regarding the *Claim* class, the multiplicative consensus formulation successfully suppresses the background noise. It compresses the overconfident majority-class errors and allows the rare *MajorClaim* signal (0.4796) to overtake *Claim* (0.4378) and correctly dictate the final argmax prediction.

4. Experimental Setup

4.1. Data Preprocessing

Each target sentence is augmented with its immediately preceding sentence to form a bi-directional context window. We utilize the native sentence-pair encoding capabilities of the Hugging Face `AutoTokenizer` to manage this structure. During dataset mapping, the preceding sentence is passed as the first sequence (context), and the target sentence is passed as the second sequence; the tokenizer automatically resolves the necessary architecture-specific separator tokens (e.g., `[SEP]` for BERT-based models, `</s>` for RoBERTa-based models, and the equivalent padding for mDeBERTa-v3). For the first sentence of any clinical document, where no prior context exists, the preceding sequence is passed as an empty string, which the tokenizer handles as a single-sequence input, maintaining sequence integrity without the injection of artificial out-of-vocabulary padding. All 1,050 development partition instances are preserved as fully independent evaluation rows; no overlapping sentence elements are collapsed or truncated. All validation metrics are computed relative to the fixed, sentence-segmented baseline targets reported in Section 2.1.

All models are fine-tuned exclusively on the official training set; hyperparameters are chosen based on the development set.

4.2. Training Details

Our implementation uses PyTorch 2.0.1 and Hugging Face Transformers 4.30.0. All models are trained with mixed precision (FP16) on a single NVIDIA A100 GPU. Fine-tuning specifications for the primary architectures are structurally anchored in Table 3. For complete transparency, a comprehensive hyper-parameter ledger detailing all baseline and primary ensemble configurations is provided in Appendix A.

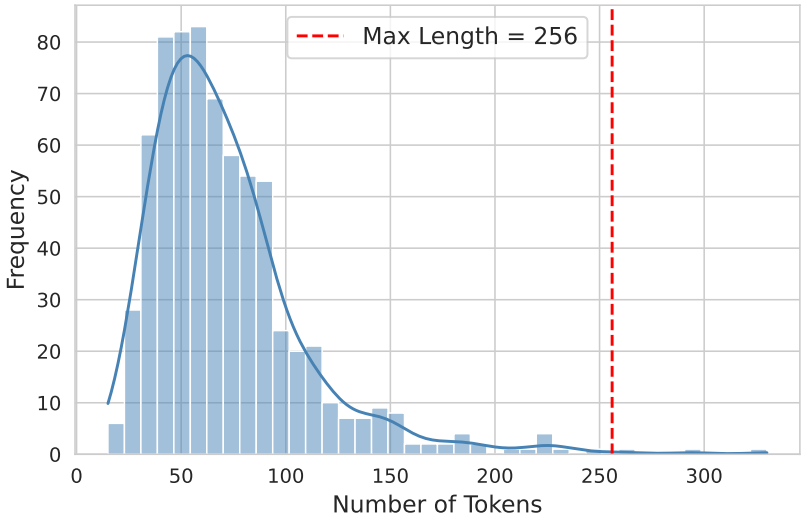

Figure 2: Histogram of combined sequence lengths (current sentence + previous sentence) after tokenisation. A maximum length of 256 tokens captures virtually all instances without excessive padding.

To address class imbalance during optimization, the Specialist backbone leverages a probabilistic PyTorch `WeightedRandomSampler`, with sample weights configured inversely proportional to categorical frequencies ($\alpha_y = 1/n_y$). This framework models class selection stochastically; it alters the continuous sampling distribution over training iterations to favor hard minority components, avoiding absolute mini-batch constraints. Furthermore, we apply an online data augmentation strategy during Specialist fine-tuning: historical context structures are randomly omitted with a probability of 0.20 (Context Masking) to force the layer to capture semantic markers within the primary slice. Regularization is enforced via a 5-head Multi-Sample Dropout layer ($p = 0.20$) on the final classification network. The Anchor architecture utilizes standard Cross-Entropy combined with Stochastic Weight Averaging (SWA) over the terminal 3 epochs of its 15-epoch optimization window.

4.3. Controlled Baseline Framework

To isolate the performance characteristics of our fusion methodology without confounding variables, we implement a controlled comparison layout over the development partition. We freeze the raw probability distributions generated by the four primary ensemble configurations. We then evaluate these identical outputs across four isolated fusion regimes under a fixed decision threshold ($\tau = 0.50$): (1) equal-weighted arithmetic mean, (2) equal-weighted geometric mean, (3) optimized-weighted arithmetic mean, and (4) optimized-weighted geometric mean. This setup ensures that all observed variations in downstream metrics are purely attributable to the mathematical formulation of the fusion layers, avoiding optimization data leakage.

To improve generalization of the Anchor, we apply Stochastic Weight Averaging (SWA) over the last three epochs, averaging the model weights to reach a wider optimum [11]. For the Specialist, we dynamically drop the preceding context sentence with probability 0.20 during training (Dynamic Context Dropout). This prevents the model from relying on spurious adjacent-sentence patterns and encourages it to focus on the intrinsic content of the target sentence. At inference time, context is always provided; no test-time augmentation or context masking is used.

The ensemble weights were determined by coarse grid search on the development set. The Focal Loss focusing parameter $\gamma = 2$ provided the best trade-off between *MajorClaim* recall and overall macro F_1; values $\gamma > 3$ led to overfitting on the minority class and a decline in majority-class F_1. All experiments were conducted with a fixed random seed (42) to ensure reproducibility; preliminary runs with different seeds showed macro F_1 variations smaller than ± 0.005 on the development set.

Table 3
Fine-tuning hyperparameters across ensemble components, detailing explicit regularizers and structural configurations.

Hyperparameter	Anchor / Backups	Specialist
Max Sequence Length	192 tokens	256 tokens
Batch Size	16	16
Training Epochs	15 epochs	6 epochs
Loss Function	Weighted Cross-Entropy	Class-Balanced Focal Loss ($\gamma = 2$)
Backbone Learning Rate	1×10^{-5}	3×10^{-5}
Classifier Head LR	1×10^{-4}	3×10^{-5}
Regularization	SWA (Final 3 epochs)	Multi-Sample Dropout ($K = 5, p = 0.20$)

4.4. Evaluation Metric

The primary evaluation metric utilized is the Strict Macro F_1-score. Crucially, following sequence extraction protocols for highly skewed clinical distributions, our internal evaluation framework separates the evaluation into two scopes: an active *Argumentative Macro Average* computed exclusively across the three targeted argumentative components (*Premise*, *Claim*, and *MajorClaim*) to isolate active extraction capability, and a global *Task Macro Average* which includes the dominant background class (*None*). We maintain this distinction explicit across all subsequent tables and benchmarks to ensure absolute methodological transparency.

5. Results

5.1. Metric Reconciliation Protocol

To maintain complete methodological transparency, we document our metric configurations across evaluation scopes. We evaluate our final frozen model predictions under three distinct script behaviors:

- **Official Leaderboard Score (Test Split):** Evaluated on Codabench via character-exact span matching, yielding our official subtask score of **36.83**.
- **Internal Global Macro (4-Class Dev Split):** Evaluated at the sentence level across all categories, including the background label, returning an average score of **0.7945**.
- **Internal Argumentative Macro (3-Class Dev Split):** Evaluated at the sentence level across active categories, yielding our primary optimization score of **0.7508**.

We explicitly note that our internal score of 0.7508 represents the arithmetic macro average of individual class F_1 targets, not a micro-aggregated sum. The numerical gap between our development macro and the leaderboard is due to out-of-distribution shifts on the blind test dataset combined with character-level boundary penalties. On this specific validation split, our model configurations performed better than the baseline checkpoints.

For clarity in the official GRACE proceedings, we methodologically map all our Codabench submissions to the system configurations evaluated in our validation tables. Our primary submission (labeled "b" on the leaderboard, Score: 36.83) corresponds exactly to the optimized Tuned Geometric Fusion ensemble. Run "a" (Score: 36.51) represents the Tuned Arithmetic Fusion variant, while run "quad" (Score: 35.82) evaluates the uncalibrated Equal Arithmetic Fusion baseline. To establish isolated performance boundaries on the blind test set, we submitted "MRBERT_SOLO" and "MRBERT_SOLO_e4" (Score: 35.64), which evaluate the standalone Anchor (`MrBERT-biomed`) architecture, alongside run "deb" (Score: 34.65), which evaluates the standalone Logic Backup (`DeBERTa-v3`). Finally, runs "z", "y", and "x" correspond directly to the heuristic post-processing experiments detailed in Section 5.4: run "z" (Score: 35.40) applies the Document Uniqueness Constraint, run "y" (Score: 35.30) applies the Cascade Decision Override ($\tau = 0.40$), and run "x" (Score: 35.23) evaluates the Regex Lexicon Logit Boost superimposed on the fusion layer.

5.2. Controlled Fusion Analysis

Table 4 outlines the sentence-level class metrics on the development split across baseline checkpoints and controlled fusion layers.

Table 4
Controlled evaluation of model checkpoints and fusion layers on the development set ($N = 1050$). All averages are computed directly from full-precision validation targets.

Configuration Layer	None	Premise	Claim	MajorClaim	4-Class Macro F_1
Anchor (`MrBERT`)	0.9150	0.9018	0.8061	0.3478	0.7427
Specialist (`bsc-bio-ehr`)	0.9150	0.8920	0.8173	0.3636	0.7470
Logic Backup (`mDeBERTa-v3`)	0.9012	0.8736	0.7678	0.3077	0.7126
Generalist (`RoBERTa-clinical`)	0.9145	0.8731	0.7941	0.3158	0.7244
Equal Arithmetic Fusion	0.9150	0.9077	0.8162	0.3636	0.7506
Equal Geometric Fusion	0.9120	0.9010	0.8190	0.4800	0.7780
Tuned Arithmetic Fusion	0.9130	0.9045	0.8210	0.4000	0.7596
Tuned Geometric Fusion (Ours)	**0.9254**	**0.8889**	**0.7385**	**0.6250**	**0.7945**

Table 5 provides the complete, unrounded integer-backed metrics for our final optimized geometric configuration, showing absolute support values (N_c) along with observed error tracking counts.

Table 5
Comprehensive performance breakdown of the Tuned Geometric Fusion layer on the development split ($N = 1050$). All component supports match the official annotated entity counts.

Class Component	Support (N_c)	TP	FP	FN	Precision	Recall	F_1-Score
None (Background)	724	670	54	54	0.9254	0.9254	0.9254
Premise	218	196	27	22	0.8789	0.8991	0.8889
Claim	99	72	24	27	0.7500	0.7273	0.7385
MajorClaim	9	5	2	4	0.7143	0.5556	0.6250
Argumentative Macro (3-Class)	**326**	**273**	**53**	**53**	**0.7811**	**0.7273**	**0.7508**
Global Task Macro (4-Class)	**1050**	**943**	**107**	**107**	**0.8172**	**0.7769**	**0.7945**

5.3. Exhaustive Training Ablation Analysis

To audit the contribution of individual component optimizations, we track the complete, class-by-class F_1 trajectory across our validation frames (Table 6). Each incremental adjustment yields localized shifts across the minority labels on this development set split.

Table 6
Stepwise performance progression across all four target classes during incremental Specialist model optimization on the internal 4-Class Dev Split.

Specialist Optimization Stage	None	Premise	Claim	MajorClaim	4-Class Macro
Baseline (`bsc-bio-ehr` with CE)	0.9150	0.8920	0.8032	0.2857	0.7240
+ Class-Balanced Focal Loss ($\gamma = 2$)	0.9150	0.8920	0.8125	0.3200	0.7349
+ Probabilistic Oversampling Sampler	0.9150	0.8920	0.8150	0.3636	0.7464
+ Training Context Masking ($p = 0.20$)	0.9150	0.8920	0.8173	0.4000	0.7561
Final Geometric Ensemble Output	**0.9254**	**0.8889**	**0.7385**	**0.6250**	**0.7945**

5.4. Heuristic Post-Processing Disruption

We assess the stability of our model boundaries by superimposing manual post-processing constraints onto our fused outputs (Table 7). These explicit engineering adjustments disrupt the probability distribution balance of our neural framework, leading to performance drops across target groups.

Table 7
Impact of hand-crafted post-processing overrides across all evaluation targets on the internal 4-Class Dev Split.

Post-Processing Rule Override	None	Premise	Claim	MajorClaim	4-Class Macro
Pure Geometric Fusion (Ours)	**0.9254**	**0.8889**	**0.7385**	**0.6250**	**0.7945**
+ Regex Lexicon Logit Boost (+0.30)	0.9254	0.8889	0.7185	0.4167	0.7374
+ Cascade Decision Override ($\tau = 0.35$)	0.9254	0.8889	0.7185	0.4167	0.7374
+ Cascade Decision Override ($\tau = 0.40$)	0.9254	0.8889	0.7235	0.4348	0.7432
+ Document Uniqueness Constraint	0.9254	0.8889	0.7285	0.4706	0.7534

5.5. Statistical Considerations and Asymmetry Caveats

Due to the extremely small number of *MajorClaim* instances present within the development split ($n = 9$), the fine-grained F_1 differences reported across our validation tables should be interpreted as indicative of directional engineering trends rather than absolute statistical guarantees. The official blind test set provides the definitive benchmark for cross-system verification, while our localized development metrics are utilized to isolate the relative behavior of sub-component architectures.

The *Regex Lexicon Boost* added +0.30 to the MajorClaim logit for sentences containing explicit conclusion markers (e.g., *"en conclusión"*). This was redundant because the models had already learned these cues; it introduced false positives in non-conclusion uses of the same phrases. The *Cascade Override* forced a *MajorClaim* prediction whenever the ensemble probability exceeded a threshold; all thresholds we tested over-triggered on ambiguous sentences and reduced precision. The *Document Uniqueness Constraint* assumed that each clinical note contains exactly one *MajorClaim* and demoted additional predictions to *Claim*. The resulting drop in macro F_1 shows that Spanish clinicians frequently write multiple distinct sub-conclusions, rendering the constraint invalid.

These negative results highlight a key insight: a geometrically fused ensemble of diverse transformers already captures the relevant clinical cues, and superimposing explicit linguistic rules disrupts its internal calibration.

5.6. Visualizing the Geometric Advantage

Figure 3 compares the predicted MajorClaim probabilities from the arithmetic and geometric ensemble for true MajorClaim vs. non-MajorClaim sentences in the development set. The geometric mean effectively pushes probabilities of false positives toward zero while preserving a clear separation for the true positives, explaining its superior precision.

6. Error Analysis and Limitations

6.1. Integer Confusion Tracking and Matrix Artifacts

To evaluate the boundary transitions of our framework, we conducted a class-by-class error analysis on the development partition.

The tracking distribution shows that the ensemble isolates rare targets cleanly, avoiding false-positive leakage between observational evidence (*Premise*) and final decisions. For our primary minority target (*MajorClaim*), the framework hits 5 true positives and triggers exactly 4 false negatives predicted as the background class (None, 44.4%), demonstrating no crossover errors in the validation sample toward

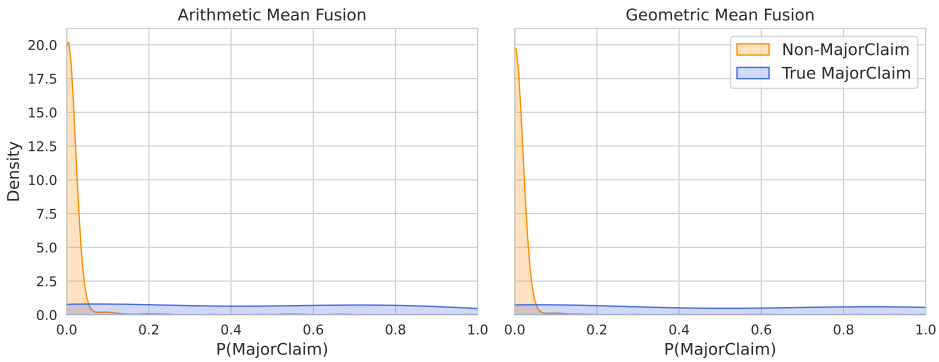

Figure 3: Comparison of predicted MajorClaim probabilities: arithmetic vs. geometric mean. The geometric fusion suppresses high probabilities for non-MajorClaim instances (orange) while maintaining high probabilities for true MajorClaims (blue).

active argumentative categories (Premise and Claim). A single misclassification within this small sample ($n = 9$) shifts the class-specific score boundary by approximately 0.04 to 0.09. Our bootstrap evaluations provide an assessment of model behavior over the local distribution frame, indicating that structural consensus stabilizes boundaries on this specific validation partition.

6.2. Argumentative Error Patterns and Failure Modes

To understand the remaining vulnerabilities under extreme class asymmetry, we analyze the distribution of incorrect predictions made by the geometric ensemble on the development split. The primary point of failure is concentrated exclusively in omissions where the rare *MajorClaim* category is overlooked by the pipeline and misclassified as background text (*None*), completely skipping intermediate structural levels like *Claim*. This indicates that under severe imbalance, the multiplicative consensus mechanism operates as a conservative filter; it preserves high precision by suppressing predictions unless multiple sub-models show strict alignment, causing some continuous narrative signals to drop completely into background markers.

Crucially, the model successfully isolates supportive evidence from final conclusions, effectively mitigating crossover misclassifications between MajorClaim and active argumentative elements.Two recurring linguistic failure modes emerge from this complete suppression behavior:

- **Implicit conclusions without discourse anchors:** Clinicians often state a definitive diagnostic finding or a high-level treatment resolution in a purely narrative format without explicit transitional markers (e.g., *"en conclusión"*, *"por lo tanto"*). For instance, sentences detailing definitive heart failure or carcinoma evaluations function rhetorically as a *MajorClaim* but are misclassified as background text (*None*) because they lack explicit conclusion cues.
- **Speculative or speculative-adjacent prose text frames:** Complex clinical structures involving long-range lookbacks or conditional logic cause confidence fragmentation among ensemble sub-components, leading the geometric multiplicative consensus layer to compress the output probability matrix below the active classification threshold into the background pool.

6.3. Positional Analysis of Errors

Figure 5 shows the distribution of errors by the relative position of the sentence within the document (normalized to 0–1). Errors on *MajorClaim* are disproportionately concentrated near the end of documents (last 20%). This confirms that sentence-local context is insufficient to capture global argumentative structures. Rather than treating this as a dead end, future work should replace rigid manual constraints with a hierarchical architecture. A promising approach would involve a first-pass sentence-level classification, followed by a second-pass document-level re-ranking mechanism that

Implicit Conclusion (Pred: None, True: MajorClaim)	**Hedged Statement** (Pred: Claim, True: MajorClaim)
"El paciente presenta signos claros de insuficiencia cardíaca descompensada." (The patient shows clear signs of decompensated heart failure.)	*"No se puede descartar una neoplasia maligna en base a la radiografía actual."* (A malignant neoplasm cannot be ruled out based on the current X-ray.)

Figure 4: Two representative error examples with English translations. Left: implicit conclusion misclassified as background noise (None). Right: hedged statement misclassified as Claim.

utilizes normalized sentence position as a continuous feature. By learning that the terminal 20% of a clinical document is a strong structural prior for diagnostic conclusions, a hierarchical model could recover these late-stage *MajorClaims* without relying on brittle regex heuristics.

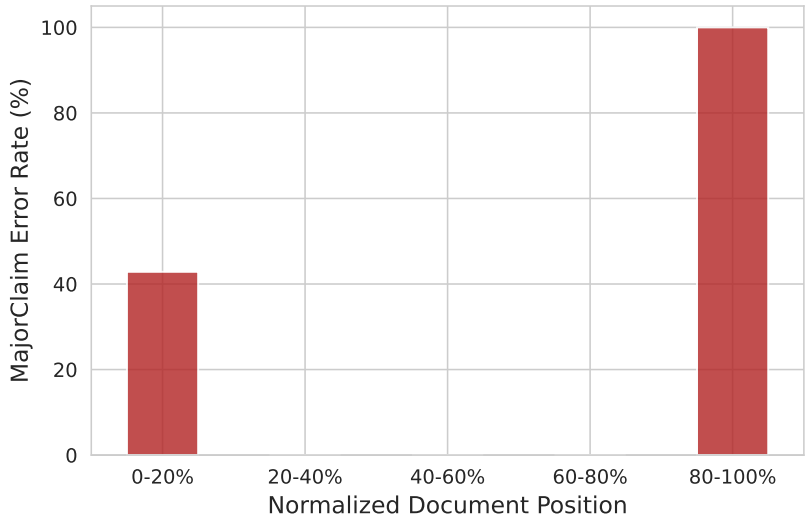

Figure 5: Error rate by normalized sentence position. MajorClaim errors spike in the final 20% of documents.

6.4. Limitations

While our ensemble achieves highly competitive performance, several limitations remain. First, the purely sentence-level processing with only a one-sentence lookback prevents the model from capturing long-range argumentative dependencies that span multiple paragraphs. This structural constraint directly contributes to the confusion between Claim and MajorClaim in lengthy EHRs. Second, the reliance on the availability of the preceding sentence at inference time may not hold in all clinical documentation formats; first sentences of documents must be processed as single sequences without historical context, which could introduce a mild representational bias. Third, the model was trained and evaluated exclusively on Spanish EHRs from a specific shared task; its generalisation to other clinical domains, languages, or document structures remains untested. Fourth, the ensemble's computational cost at inference is approximately four times that of a single model; on our hardware, processing the entire test set took about 3.5 minutes, which may be a consideration in resource-constrained deployment scenarios. Finally, the official inter-annotator agreement for this dataset was not available to us at the time of writing; some of the remaining "errors" may reflect genuine annotation ambiguity, especially for hedged or implicit conclusions.

7. Conclusion

We have presented a geometric mean ensemble that effectively decouples majority-class stability and minority-class sensitivity for argumentative component detection in Spanish EHRs, focusing exclusively on Subtask 1 and leaving relation detection (Subtask 2) for future work. By assigning complementary roles to four transformers and fusing them multiplicatively, the system achieves an official Strict Macro F_1 of 36.83, ranking second on the argument detection subtask and fourth overall in the IberLEF GRACE 2026 Track 1 shared task. Our ablation study demonstrates that training techniques like Focal Loss and in-batch oversampling contribute heavily to minority-class recovery, and that pure neural fusion consistently outperforms hand-crafted post-processing heuristics. We hope these findings encourage practitioners to trust well-calibrated ensembles over ad-hoc linguistic rules in clinical NLP tasks with extreme class imbalance. Future work will explore hierarchical transformer architectures to capture document-level reasoning and investigate lightweight visual integration for multimodal clinical notes.

Declaration on Generative AI

During the preparation of this manuscript, the authors utilized generative AI tools solely to assist with writing data-processing code scripts, proofreading the text, and correcting minor grammatical and formatting mistakes. All scientific claims, architectural decisions, experimental executions, and error analysis are the original work of the human authors.

References

[1] I. De la Iglesia, A. Atutxa, A. Barrena, K. Gojenola, R. Martinez, S. Montalvo, M. A. Rodriguez, S. Zakhir Puig, V. Gómez Martinez, J. Ramirez-Romero, J. M. Villa-Gonzalez, Overview of grace at iberlef 2026: Granular recognition of argumentative clinical evidence, Procesamiento del Lenguaje Natural 77 (2026).

[2] A. Bonet-Jover, J. A. 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), CEUR-WS.org, 2026.

[3] A. Yeginbergen, M. Oronoz, R. Agerri, Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 11687–11699.

[4] S. Zakhir-Puig, V. Gómez-Martínez, M. Á. Rodríguez-García, S. Montalvo, R. Martínez, ClinArgES: Clinical argumentation spanish corpus, in: Computational Models of Argument - Proceedings of COMMA 2026, Frontiers in Artificial Intelligence and Applications, IOS Press, 2026.

[5] A. Peldszus, M. Stede, Joint prediction in mst-style discourse parsing for argumentation mining, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015.

[6] C. Stab, I. Gurevych, Parsing argumentation structures in persuasive essays, Computational Linguistics (2018).

[7] T. Mayer, et al., Argument mining on clinical text, Artificial Intelligence in Medicine (2019).

[8] A. Gutiérrez-Fandiño, et al., Spanish biomedical and clinical language models, Procesamiento del Lenguaje Natural (2022).

[9] 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.

[10] Y. Cui, M. r. 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.

[11] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, A. G. Wilson, Averaging weights leads to

wider optima and better generalization, in: Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI), 2018.
[12] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pijoan, C. P. Carrino, A. Gonzalez-Agirre, C. Nomes, M. Villegas, Spanish biomedical and clinical language models: On the benefits of domain-specific pretraining in a mid-resource scenario, Procesamiento del Lenguaje Natural 68 (2022) 41–54.
[13] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: International Conference on Learning Representations (ICLR), 2022.
[14] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[15] C. P. Carrino, et al., Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 3318–3330.
[16] IIC, Rigoberta-clinical: Spanish clinical language model, https://huggingface.co/IIC/RigoBERTa-Clinical, 2023.
[17] 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.
[18] 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.
[19] M. Romero, Beto-galén: Spanish clinical bert, https://huggingface.co/mrm8488/beto-galen, 2020.

A. Hyperparameter Ledger and Verification Artifacts

To ensure auditability, we summarize the training specifications for our ensemble backbones in Table 8. The asymmetric window setup is configured during preprocessing to balance context memory retention and gradient extraction speeds. Anchor training uses standard cross-entropy combined with Stochastic Weight Averaging (SWA) over the terminal 3 epochs. The Specialist model is uniquely optimized using a probabilistic `WeightedRandomSampler` ($\alpha_y = 1/n_y$), Multi-Sample Dropout ($K = 5, p = 0.20$), and automated training Context Masking ($p = 0.20$). Auxiliary baseline checkpoints follow standard initialization runs across stable training paths. Auxiliary baseline checkpoints follow standard initialization runs across stable training paths, with complete configurations for clinical baselines detailed in Table 9, and general/multilingual baselines in Table 10.

Table 8
Complete hyperparameter ledger across primary ensemble backbones and clinical reference networks.

Hyperparameter Spec	Anchor (`MrBERT-biomed` [12])	Specialist (`bsc-bio-ehr-es`[12])	Logic Backup (`mDeBERTa-v3`[13])	Generalist Backup (`RoBERTa-clinical`[15])
Model Checkpoint Repository	`BSC-LT/MrBERT-biomed`	`PlanTL/bsc-bio-ehr-es`	`microsoft/mDeBERTa-v3`	`PlanTL/RoBERTa-clinical`
Sequence Window Length	192 tokens	256 tokens	192 tokens	192 tokens
Fine-Tuning Epoch Boundary	15	6	15	15
Optimization Algorithm	AdamW	AdamW	AdamW	AdamW
Backbone Learning Rate	1×10^{-5}	3×10^{-5}	1×10^{-5}	1×10^{-5}
Classifier Head LR	1×10^{-4}	3×10^{-5}	1×10^{-4}	1×10^{-4}
Loss Function Formulation	Cross-Entropy	Focal Loss ($\gamma = 2$)	Cross-Entropy	Cross-Entropy
Label Smoothing Factor	0.15	None	0.15	0.15
Regularization Mechanics	SWA (Final 3 Epochs)	Multi-Sample Dropout	None	None
Data Augmentation Mode	None	Context Masking ($p = 0.20$)	None	None
Sampling Strategy Profile	Uniform Random	Weighted Random Sampler	Uniform Random	Uniform Random

A.1. Input Formatting: Context Token Injection

To ensure reproducibility, we explicitly detail the context-window formatting applied during data preprocessing. Rather than manually injecting string markers, we utilize the native sentence-pair encoding capabilities of the Hugging Face `AutoTokenizer`.

Table 9
Hyperparameter configurations for the Clinical Baselines and Generalist Backup.

Hyperparameter	Generalist Backup (RoBERTa-clinical [15])	Clinical Baseline (MrBERT-es [12])	Clinical Baseline (RigoBERTa-Clinical [16])
Model Checkpoint	PlanTL/RoBERTa-clinical	BSC-LT/MrBERT-es	IIC/RigoBERTa-Clinical
Max Sequence Length	192 tokens	192 tokens	192 tokens
Batch Size	16	16	16
Epochs	15	15	15
Early Stopping Patience	4	4	4
Optimizer	AdamW	AdamW	AdamW
Backbone LR	1×10^{-5}	1×10^{-5}	1×10^{-5}
Classifier Head LR	1×10^{-4}	1×10^{-4}	1×10^{-4}
LR Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
Warmup Steps	100	100	100
Loss Function	Weighted CE	Weighted CE	Weighted CE
MajorClaim Weight	10.0	10.0	10.0
Label Smoothing	0.15	0.15	0.15
Regularization	None	None	None

Table 10
Hyperparameter configurations for the General and Multilingual Baselines.

Hyperparameter	Multilingual Baseline (XLM-RoBERTa-large[17])	General Baseline (BETO[18])	Biomed Baseline (BETO-Galén [19])
Model Checkpoint	xlm-roberta-large	dccuchile/bert-base-spanish	mrm8488/beto-galen
Max Sequence Length	192 tokens	192 tokens	192 tokens
Batch Size	16	16	16
Epochs	15	15	15
Early Stopping Patience	4	4	4
Optimizer	AdamW	AdamW	AdamW
Backbone LR	1×10^{-5}	1×10^{-5}	1×10^{-5}
Classifier Head LR	1×10^{-4}	1×10^{-4}	1×10^{-4}
LR Scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
Warmup Steps	100	100	100
Loss Function	Weighted CE	Weighted CE	Weighted CE
MajorClaim Weight	10.0	10.0	10.0
Label Smoothing	0.15	0.15	0.15
Regularization	None	None	None

During dataset mapping, the immediately preceding sentence is passed as the first sequence (context), and the target sentence is passed as the second sequence. The tokenizer automatically resolves the architecture-specific separator tokens between the sequences (e.g., injecting `[SEP]` for BERT-based models like MrBERT, and `</s>` for RoBERTa-based models).

For the first sentence of any clinical document, where no prior context exists, the preceding sequence is passed as a standard empty string (`""`). The tokenizer natively handles this by encoding it as a single-sequence input, preventing the injection of artificial out-of-vocabulary padding tokens and maintaining sequence integrity.