UMUTeam at GRACE@IberLEF 2026: Instruction Tuning of LLMs and Medical Continual Pretraining for Clinical Argument Mining

Ronghao Pan[1], Pedro José Vivancos-Vicente[2], Jorge Gómez-Navalón[1], Tomás Bernal-Beltrán[1] and Rafael Valencia-García[1,*]

[1]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*
[2]*VÓCALI SISTEMAS INTELIGENTES S.L. Parque Científico de Murcia, Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, Spain*

Abstract

This paper describes the participation of UMUTeam in the GRACE shared task at IberLEF 2026. We participated in both Track 1, focused on argumentative evidence in clinical trial texts, and Track 2, focused on clinical case reasoning. Our approach formulates the subtasks as supervised instruction-tuning problems and fine-tunes Qwen3-4B models using parameter-efficient adaptation with LoRA and 4-bit quantization. The main objective of our work is to compare an instruction-tuned pretrained Qwen3-4B model with an instruction-tuned medically pretrained Qwen3-4B model. The medical variant was further adapted using clinical practice guidelines, the CoWeSe dataset, and the SMC dataset. In Track 1, the medically pretrained model achieved the best overall UMUTeam score and improved argumentative component detection, although the general model performed better in relation detection. In Track 2, the medical model also obtained the best overall score, with gains in evidence and relation extraction, while the general model was slightly better in sentence relevance detection. Overall, the results show that medical pretraining provides modest and task-dependent improvements.

Keywords

Continual Pretraining, Large Language Models, Instruction Tuning, Clinical Argument Mining

1. Introduction

Clinical narratives often contain complex argumentative structures that are essential for understanding diagnostic reasoning, therapeutic decisions, and evidence-based medical conclusions. In clinical case descriptions, relevant arguments are not always expressed as complete sentences. They may appear as short medical terms, descriptive phrases, clauses, or longer textual fragments whose interpretation depends on the surrounding context. This makes clinical argument mining a challenging task, since systems must identify not only which parts of the text are argumentative, but also the exact boundaries of each argumentative component and the relations that connect them.

In this working note, we describe the participation of UMUTeam in the GRACE [1] shared task at IberLEF 2026 [2]. The shared task is composed of several disticts datasets. The dataset of Track 1 is composed of randomized controlled trial (RCT) abstracts, segmented and annotated for argumentative structure from [3] and [4], and the dataset of Track 2 is composed of clinical cases from MIR examinations [1]. Track 1 addresses argument mining in randomized clinical trial evidence, where systems must identify argumentative roles and relations in clinical trial texts. Track 2 focuses on clinical case reasoning, where the goal is to extract fine-grained argumentative components with exact textual boundaries and identify directed argumentative relations between them. Our team participated in both tracks, allowing

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
ronghao.pan@um.es (R. Pan); pedro.vivancos@vocali.net (P. J. Vivancos-Vicente); jorge.gomezn@um.es (J. Gómez-Navalón); tomas.bernalb@um.es (T. Bernal-Beltrán); valencia@um.es (R. Valencia-García)
https://www.vocali.net/ (P. J. Vivancos-Vicente)
0009-0008-7317-7145 (R. Pan); 0009-0005-7185-6054 (J. Gómez-Navalón); 0009-0006-6971-1435 (T. Bernal-Beltrán); 0000-0003-2457-1791 (R. Valencia-García)

us to evaluate a common instruction-tuning methodology across two different clinical argument mining scenarios.

For Track 1, our approach models the task as a combination of argumentative role classification and relation classification. In the first stage, the system assigns each relevant textual segment one of the target argumentative roles, including Premise, Claim, MajorClaim, and Non-argumentative. In the second stage, the system predicts directed argumentative relations between components, considering labels such as Support, Attack, Partial-Attack, and No-relation. This formulation allows the system to process clinical trial evidence in a structured way, while also accounting for the presence of non-argumentative material and relation absence between candidate component pairs.

For Track 2, we formulate the task as a structured generation problem over clinical case narratives. The system first extracts argumentative entities from the full clinical context and generates a valid JSON list containing their start and end offsets, textual content, entity type, and identifier. Each entity is labeled as either Premise or Claim. Then, using the extracted or gold argumentative entities, the relation extraction module predicts directed links between them, distinguishing Support and Attack relations. This track requires a finer level of granularity than sentence-level classification, since the model must recover exact textual boundaries and preserve consistency between predicted spans, identifiers, and relations.

A central objective of our participation is to evaluate the performance of different open-source Large Language Models (LLMs) in clinical argument mining. In particular, we compare general-domain base models with corresponding variants adapted through continual medical training with different medical dataset such as SMC [5] and CoWeSe [6]. This comparison is motivated by the hypothesis that additional exposure to biomedical and clinical data may improve the ability of language models to recognize medically relevant argumentative evidence and to reason over relations between clinical statements. By evaluating both base and medically adapted models across Track 1 and Track 2, we aim to analyze whether domain specialization provides measurable benefits in both sentence-level and fine-grained structured prediction settings.

Our system relies on supervised instruction tuning with parameter-efficient fine-tuning. Specifically, we use Low-Rank Adaptation (LoRA) [7] together with quantized model loading, which enables us to fine-tune several model families under limited computational resources. Across both tracks, the tasks are converted into instruction-following examples where the model receives a clinical context and task-specific instructions, and is trained to generate either a label or a structured JSON output. The generated predictions are then post-processed to normalize labels, validate offsets, ensure consistent identifiers, and produce submission-compatible JSON files.

2. Related Work

Argument mining aims to identify argumentative components and the relations between them [8]. Traditional approaches commonly address component detection and relation classification as separate tasks [9]. Although this modular design simplifies system development, it also introduces error propagation, since a relation cannot be correctly predicted when one of its components is missing or has incorrect boundaries.

More recent studies have explored joint, multi-task, and transformer-based architectures that exploit dependencies between component types and argumentative relations [10]. Nevertheless, relation detection remains more difficult than component identification because it requires discourse-level reasoning, polarity interpretation, and the classification of a large and imbalanced set of candidate component pairs.

Biomedical argument mining focuses on claims and evidence expressed in clinical and scientific documents. Previous work has addressed argumentative structures in randomized clinical trials and has shown that biomedical texts require both general discourse understanding and domain-specific knowledge [11, 12]. Other studies have connected clinical argument mining with evidence-based medicine and explored multi-task architectures for modelling components and relations [13]. These

approaches also confirm that relation prediction is strongly affected by errors introduced during component detection.

The GRACE [1] shared task combines several of these challenges. Track 1 addresses argumentative structures in clinical trial evidence, while Track 2 includes sentence relevance classification, fine-grained evidence identification, and relation prediction in multiple-choice clinical cases.

Domain-specific language models such as BioBERT [14] and PubMedBERT [15] have shown that biomedical pretraining can improve tasks including entity recognition, relation extraction, and question answering. However, medical adaptation does not necessarily improve every downstream task, particularly those that depend on general reasoning or instruction-following capabilities. This trade-off is especially relevant to clinical argument mining, where fine-grained evidence detection may benefit from medical terminology, while relevance and relation classification also require broader semantic and discourse-level knowledge.

Instruction tuning provides a common framework for adapting generative models to classification, extraction, and question answering. Biomedical resources such as BioInstruct [16] and AlpaCare [17] have shown that medical instruction tuning can improve performance when the training instructions are closely related to the target task.

Our work formulates the GRACE subtasks as instruction-following problems for a causal language model. Separate models are trained for sentence relevance, component classification, and relation classification. We compare a general Qwen3-4B [18] model with a version further pretrained on Spanish medical and clinical texts, including resources such as CoWeSe and SMC [6, 5]. This comparison allows us to examine whether medical pretraining provides consistent benefits across subtasks with different linguistic and reasoning requirements.

3. Methodology

This section describes the methodology followed by UMUTeam for the GRACE shared task at IberLEF 2026. Our system is based on supervised instruction tuning of generative language models for clinical argument mining. We participated in both Track 1 and Track 2. Track 1 focuses on argumentative evidence in randomized clinical trial texts, whereas Track 2 focuses on argumentative reasoning in clinical case narratives.

Across both tracks, the subtasks were formulated as instruction-following classification problems, as shown in Figure 1. Each training instance contains a task-specific instruction, the complete clinical context, and the corresponding gold response. Depending on the subtask, the expected response is an argumentative role, a relevance label, an entity type, or an argumentative relation.

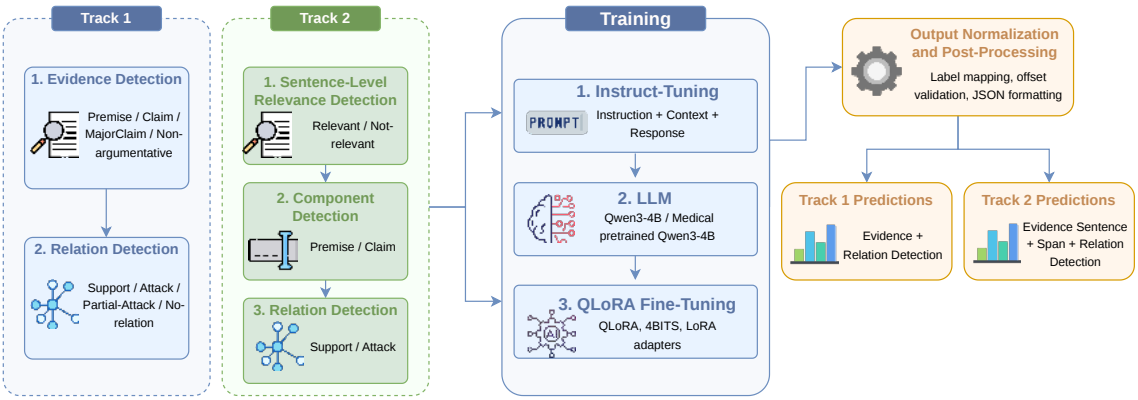

Figure 1: Overview of the UMUTeam approach for Track 1 and Track 2 of GRACE

3.1. Models

In this work, we evaluated instruction tuning of a medically pretrained Qwen3-4B model and compared it with a continual medical training variant. The continual-training model was further adapted using clinical practice guidelines, the CoWeSe dataset [6], and the SMC dataset [5].

The objective of this comparison was to determine whether additional exposure to domain-specific clinical and biomedical resources improves performance in clinical argument mining. Both variants were fine-tuned independently for each subtask using the same instruction-tuning framework.

3.2. Instruction-Tuning Formulation

All subtasks were transformed into supervised instruction-tuning examples following the same general structure: a task-specific instruction, the complete clinical context, and the expected gold response. The instruction specifies the task, the valid output labels, and the relevant input unit, such as a sentence, span, or pair of argumentative components. During training, the response contains the gold label. During inference, the response field is left empty and the generated continuation is interpreted as the prediction. The exact prompts used for reproducibility are reported in Appendix A.

3.3. Track 1: Clinical Trial Evidence

Track 1 was divided into two subtasks:

1. **Evidence Detection**. The first stage assigns an argumentative role to each textual segment. The valid labels are `Premise`, `Claim`, `MajorClaim`, and `Non-argumentative`. Each example contains the complete randomized clinical trial text as context and the target sentence in the instruction. Annotated argumentative components were assigned their original labels. Sentences or textual fragments that were not covered by argumentative annotations were incorporated as `Non-argumentative` examples.
2. **Relation Detection**. The second stage predicts the directed argumentative relation from a source component, `arg1`, to a target component, `arg2`. The valid relation labels are `Support`, `Attack`, `Partial-Attack`, and `No-relation`. The instruction includes the source and target component identifiers and texts, and explicitly asks the model to consider negation markers because they may reverse the polarity of a relation. Positive examples were extracted from the gold relations. Directed entity pairs without an annotated relation were used as `No-relation` examples.

3.4. Track 2: Clinical Case Reasoning

For Track 2, the input files contain the complete clinical case in `raw_text`, sentence-level metadata, answer choices, gold annotations, and an empty prediction structure. The three subtasks were trained independently using the official training and development splits. Therefore, this track is organized into three subtasks:

1. **Sentence-Level Relevance Detection**. The first subtask determines whether a sentence from the clinical context contains information that is relevant for supporting or refuting at least one answer option. It was formulated as a binary classification task with the labels `relevant` and `not-relevant`. For each document, the sentences in `metadata.context_sentences` were aligned with the labels in `annotations.sentence_relevancy`. One instruction-tuning example was created for each sentence. The target sentence was included in the instruction, while the complete clinical case was supplied as context. The preprocessing stage generated 500 training instances and 102 development instances. The relevance classifier was trained for five epochs. Predictions were normalized to the two valid labels before evaluation.

Parameter	T1 Role	T1 Relation	T2 Relevance	T2 Entity	T2 Relation
LoRA rank	16	16	16	16	16
LoRA alpha	32	32	32	32	32
LoRA dropout	0.05	0.05	0.05	0.05	0.05
Target modules	q_proj, k_proj, v_proj, o_proj				
Epochs	3	3	5	3	3
Learning rate	2e-5	2e-4	2e-4	1e-4	1e-4
Train batch size	8	8	8	2	2
Evaluation batch size	8	8	8	2	2
Gradient accumulation	2	2	2	4	4
Effective batch size	16	16	16	8	8
Weight decay	0.05	0.05	0.05	0.05	0.05
Warmup ratio	0.03	0.03	0.03	0.03	0.03
Maximum sequence length	1024	2048	2048	2048	2048

Table 1
Training hyperparameters used for Track 1 and Track 2.

2. **Argumentative Entity Classification**. The second subtask assigns an argumentative type to each annotated textual span. The valid labels are `Premise` and `Claim`. For each entity in `annotations.entities`, the text was extracted directly from the original document using `raw_text[start:end]`. The entity span was included in the instruction, while the complete clinical case was provided as context. This formulation evaluates whether the model can distinguish evidence-like spans from answer-option claims when the candidate span is known.
3. **Argumentative Relation Classification**. The third subtask predicts the directed relation between two annotated argumentative entities. The valid labels are `Support` and `Attack`. For each document, a dictionary was created to map entity identifiers to their text, type, and offsets. Each gold relation in `annotations.relations` was then converted into an independent instruction-tuning instance containing the two entity identifiers, their types, and their text.

3.5. Parameter-Efficient Fine-Tuning

All classifiers were fine-tuned using LoRA. The models were loaded using 4-bit NF4 quantization with double quantization. LoRA adapters were applied to the attention projection modules `q_proj`, `k_proj`, `v_proj`, and `o_proj`.

The LoRA configuration used a rank of 16, an alpha value of 32, and a dropout value of 0.05. Gradient checkpointing was enabled to reduce GPU memory consumption. Specific training hyperparameters used for Track 1 and Track 2, are shown in Table 1.

3.6. Evaluation and Output Normalization

The official training split was used for supervised fine-tuning, while the development split was used for checkpoint selection. Macro-F1 was used as the main model-selection metric because it gives equal weight to all classes and is more appropriate for imbalanced label distributions.

For all subtasks, only the first non-empty generated line was considered. Quotation marks, surrounding whitespace, and minor capitalization differences were removed. The output was then mapped to the valid label set of each subtask.

In addition to the metrics computed during the trainer evaluation loop, we performed an explicit autoregressive evaluation using greedy decoding. This evaluation reproduces the inference procedure used for the final predictions and provides a more reliable estimate of exact-label classification performance.

Model	Overall	Component Detection		Relation Detection	
		Strict	Relaxed	Strict	Relaxed
General Qwen3-4B	13.83	25.65	36.92	**2.01**	**3.09**
Medical Qwen3-4B	**14.64**	**27.78**	**39.02**	1.49	1.84

Table 2
Official test-set results for the two UMUTeam systems in Track 1. All values are macro-F1 scores expressed as percentages.

4. Results and Discussion

This section reports the official test-set results obtained by UMUTeam in the GRACE shared task at IberLEF 2026. We compare two Qwen3-4B configurations: a general-purpose pretrained model and a model with additional medical-domain pretraining. Both models were adapted to the task through instruction tuning.

The objective of this comparison is to assess whether medical-domain pretraining improves the identification of argumentative components and relations in clinical texts. Unless otherwise stated, all reported values are macro-F1 scores expressed as percentages.

4.1. Track 1 Results

Table 2 presents the official results for Track 1. The Medical Qwen3-4B achieved the highest overall score, with 14.64 points, compared with 13.83 for the General Qwen3-4B. However, the effect of medical pretraining differed across the two subtasks.

4.1.1. Evidence Detection

The Medical Qwen3-4B outperformed the general model under both evaluation settings. It obtained a strict macro-F1 of 27.78, compared with 25.65 for the General Qwen3-4B. Under relaxed matching, the corresponding scores were 39.02 and 36.92. Medical pretraining therefore produced absolute improvements of 2.13 and 2.10 points under strict and relaxed evaluation, respectively.

These results suggest that domain-specific pretraining was beneficial for identifying argumentative components in randomized clinical trial reports. Previous exposure to biomedical terminology and clinical discourse may have helped the model identify claims and supporting evidence more accurately.

For both systems, the relaxed scores were approximately 11 points higher than the strict scores. This difference indicates that many predicted components overlapped with the reference annotations but did not reproduce their exact textual boundaries. Span-boundary detection therefore remains an important source of error.

4.1.2. Relation Detection

The opposite trend was observed for relation detection. The General Qwen3-4B achieved a strict macro-F1 of 2.01 and a relaxed macro-F1 of 3.09, whereas the Medical Qwen3-4B obtained 1.49 and 1.84, respectively. Thus, the general model outperformed the medical model by 0.52 points under strict evaluation and by 1.25 points under relaxed evaluation.

This result shows that the benefits of medical pretraining did not extend uniformly to all Track 1 subtasks. Relation detection requires identifying the direction and argumentative function of links between components, which may depend more strongly on discourse-level representations than on specialized terminological knowledge.

The absolute scores were low for both systems, making relation detection the main limitation of the Track 1 pipeline. In addition to the difficulty of distinguishing relations such as `Support`, `Attack`, and `Partial-Attack`, this task is affected by error propagation. A relation can only be evaluated correctly when both its source and target components have also been identified successfully.

Metric	Medical − General
Overall score	+0.81
Component detection, strict	+2.13
Component detection, relaxed	+2.10
Relation detection, strict	−0.52
Relation detection, relaxed	−1.25

Table 3
Absolute score differences between the Medical and General Qwen3-4B systems in Track 1.

Model	Overall	Sentence Relevance	Component Detection		Relation Detection	
		Macro-F1	Strict	Relaxed	Strict	Relaxed
General Qwen3-4B	44.15	**77.77**	53.94	57.85	0.75	2.20
Medical Qwen3-4B	**44.29**	76.92	**54.86**	**59.23**	**1.10**	**2.44**

Table 4
Official test-set results for the two UMUTeam systems in Track 2. Relevance denotes sentence-level relevance classification.

Table 3 summarizes the absolute differences between the Medical and General Qwen3-4B systems. Overall, medical pretraining improved the Track 1 score by 0.81 points, from 13.83 to 14.64. This gain was driven by component detection, where the medical model improved both strict and relaxed scores. However, the general model remained more effective for relation prediction. These findings indicate that the effect of medical pretraining is subtask-dependent rather than uniformly beneficial across the full Track 1 pipeline.

4.2. Track 2 Results

Table 4 reports the official results for Track 2. The Medical Qwen3-4B obtained the highest overall UMUTeam score, reaching 44.29, compared with 44.15 for the General Qwen3-4B. Although the difference was small, the two systems showed distinct strengths across the three subtasks.

4.2.1. Sentence-Level Relevance Detection

The General Qwen3-4B achieved the best result for sentence-level relevance detection, with a macro-F1 of 77.77, compared with 76.92 for the Medical Qwen3-4B. It also obtained a higher accuracy, reaching 78.63 compared with 77.35. The differences were 0.85 macro-F1 points and 1.28 accuracy points, respectively.

The absence of an improvement from medical pretraining suggests that this binary classification task may depend primarily on general semantic matching and discourse-level information. Compared with span and relation extraction, determining whether a complete sentence is relevant may require less fine-grained medical knowledge.

4.2.2. Component Detection

The Medical Qwen3-4B performed better in component detection. It obtained a strict macro-F1 of 54.86, compared with 53.94 for the general model, and a relaxed macro-F1 of 59.23, compared with 57.85. These differences correspond to absolute improvements of 0.92 and 1.38 points, respectively.

The larger gain under relaxed matching suggests that the medical model more frequently identified semantically appropriate evidence, even when its predicted boundaries did not match the gold annotations exactly.

The class-level results of the Medical Qwen3-4B reveal a considerable difference between the two component types. Under strict evaluation, the model obtained an F1 score of 10.86 for `Premise` and 98.86 for `Claim`. Under relaxed evaluation, the corresponding scores were 19.61 and 98.86.

Metric	Medical − General
Overall score	$+0.14$
Relevance detection, accuracy	-1.28
Relevance detection, macro-F1	-0.85
Component detection, strict	$+0.92$
Component detection, relaxed	$+1.38$
Relation detection, strict	$+0.35$
Relation detection, relaxed	$+0.24$

Table 5
Absolute score differences between the Medical and General Qwen3-4B systems in Track 2.

Claims generally correspond to complete answer options with clearly defined boundaries. Premises, by contrast, may consist of short symptoms, measurements, diagnoses, clinical findings, or clauses embedded within longer sentences. This structural difference likely contributed to the substantially lower premise-detection performance.

4.2.3. Relation Detection

Relation detection was the most difficult Track 2 subtask for both systems. The General Qwen3-4B obtained strict and relaxed macro-F1 scores of 0.75 and 2.20, respectively. The Medical Qwen3-4B improved these values to 1.10 and 2.44.

Although the absolute improvements were limited to 0.35 points under strict matching and 0.24 points under relaxed matching, they correspond to relative increases of approximately 46.7% and 10.9%, respectively. These relative figures should be interpreted cautiously because they are calculated from very low baseline scores.

Several factors may explain the difficulty of this subtask. First, relation prediction depends on the successful identification of both source and target components, resulting in error propagation from the component-detection stage. Second, the number of candidate pairs increases rapidly with the number of detected components, creating a highly imbalanced classification problem. Third, distinguishing whether evidence supports or attacks an answer requires combining clinical knowledge with polarity, directionality, and discourse-level context.

Table 5 presents the absolute score differences between the Medical and General Qwen3-4B systems in Track 2. Positive values favor the medical model. As shown in Table 5, medical pretraining did not improve sentence-level relevance detection, with decreases of 1.28 accuracy points and 0.85 macro-F1 points. However, it produced modest gains in component detection and relation extraction. The overall difference between the two systems was only 0.14 points, indicating that medical pretraining did not provide a broad improvement across Track 2. Instead, its benefits were concentrated in tasks requiring fine-grained interpretation of clinical evidence.

5. Discussion

Across the two tracks, medical-domain pretraining consistently benefited component detection. It improved both strict and relaxed component scores in Track 1 and Track 2, although the magnitude of the improvements was moderate.

The effect on the remaining subtasks was less consistent. The General Qwen3-4B performed better for Track 1 relation detection and Track 2 relevance classification, whereas the Medical Qwen3-4B performed better for Track 2 relation detection. Domain-specific pretraining should therefore not be regarded as a uniform improvement across the complete argument-mining pipeline.

A second consistent finding is the gap between strict and relaxed component evaluation. This pattern indicates that the systems frequently identified the relevant textual region but failed to reproduce the exact annotation boundaries. Finally, the low relation-detection scores show that a sequential pipeline is highly sensitive to upstream errors.

6. Conclusions and Future Work

This paper described the participation of UMUTeam in the GRACE shared task at IberLEF 2026. We addressed both Track 1 and Track 2 using instruction-tuned Qwen3-4B models with parameter-efficient fine-tuning.

Our main objective was to compare a pretrained Qwen3-4B model with a medically pretrained variant. The results show that medical pretraining provided modest but task-dependent improvements. In both tracks, the medical model achieved the best overall UMUTeam score and improved component detection. However, the general model performed better in some subtasks, particularly sentence relevance and relation detection.

The results also show that relation prediction was the most challenging part of the task. Its performance was considerably lower than that of component and relevance classification, mainly due to error propagation, class imbalance, and the complexity of relation direction and polarity. The gap between strict and relaxed evaluation further indicates that exact span boundary detection remains difficult.

Future work will focus on the main weaknesses observed in this study: span extraction and relation detection. We plan to explore token-level or span-aware models, joint component–relation architectures, improved negative sampling, class balancing, constrained decoding, and a more detailed error analysis by entity type, relation type, span length, and medical specialty.

Acknowledgments

This research is part of the research project MedicaLLM (CPP2024-011574) funded by MICIU/AEI /10.13039/501100011033 and the European Regional Development Fund (ERDF EU/FEDER UE)-a way of making Europe.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL in order to: Grammar and spelling check. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. De la Iglesia, A. Atutxa, A. Barrena, K. Gojenola, R. Martínez, S. Montalvo, M. Á. Rodríguez, S. Zakhir Puig, V. Gómez Martínez, 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. Á. González-Barba, L. Chiruzzo, Overview of IberLEF 2026: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), co-located with the 42nd Conference of the Spanish Society for Natural Language Processing (SEPLN 2026), CEUR-WS.org, 2026.

[3] A. 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, RaquelMartínez, ClinArgES: Clinical argumentation spanish corpus, in: Computational Models of Argument - Proceedings of COMMA 2026, Frontiers in Artifi cial Intelligence and Applications, IOS Press, 2026.

[5] L. Dionis, G. Alvaro, M. Dylan, B. Daniel, Spanish Medica LLM, https://huggingface.co/datasets/somosnlp/SMC, 2024. Dataset on Hugging Face.

[6] C. P. Carrino, J. Armengol-Estapé, O. d. G. Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, M. Villegas, Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models, arXiv preprint arXiv:2109.07765 (2021).
[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., LoRA: Low-rank adaptation of large language models, ICLR 1 (2022) 3.
[8] A. Peldszus, M. Stede, Joint prediction in MST-style discourse parsing for argumentation mining, in: L. Màrquez, C. Callison-Burch, J. Su (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 938–948. URL: https://aclanthology.org/D15-1110/. doi:10.18653/v1/D15-1110.
[9] V. Niculae, J. Park, C. Cardie, Argument Mining with Structured SVMs and RNNs, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 985–995. URL: https://aclanthology.org/P17-1091/. doi:10.18653/v1/P17-1091.
[10] S. Eger, J. Daxenberger, I. Gurevych, Neural End-to-End Learning for Computational Argumentation Mining, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 11–22. URL: https://aclanthology.org/P17-1002/. doi:10.18653/v1/P17-1002.
[11] T. Mayer, Argument mining on clinical trials, Ph.D. thesis, Université Côte d'Azur, 2020.
[12] T. Mayer, S. Marro, E. Cabrio, S. Villata, Enhancing evidence-based medicine with natural language argumentative analysis of clinical trials, Artificial intelligence in medicine 118 (2021) 102098.
[13] T. Mayer, E. Cabrio, S. Villata, Transformer-based argument mining for healthcare applications, in: ECAI 2020-24th European Conference on Artificial Intelligence, 2020.
[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.
[15] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.
[16] H. Tran, Z. Yang, Z. Yao, H. Yu, BioInstruct: instruction tuning of large language models for biomedical natural language processing, Journal of the American Medical Informatics Association 31 (2024) 1821–1832.
[17] X. Zhang, C. Tian, X. Yang, L. Chen, Z. Li, L. R. Petzold, Alpacare: Instruction-tuned large language models for medical application, arXiv preprint arXiv:2310.14558 (2023).
[18] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025).

A. Prompt Templates

This appendix reports the exact prompt templates used in the instruction-tuning experiments.

Prompt A.1: General instruction-tuning template

```
### Instruction:
{task-specific instruction}

### Context:
{clinical text}

### Response:
{gold label}
```

Prompt A.2: Track 1 argumentative role classification prompt

```
Dada la siguiente frase de un resumen de ensayo clinico aleatorizado (RCT),
    clasificala en exactamente un rol argumentativo. Las unicas etiquetas validas son:
     Premise, Claim, MajorClaim, Non-argumentative. Devuelve solo la etiqueta.

Sentence: "{sentence}"
```

Prompt A.3: Track 1 relation classification prompt

```
Dado el conjunto de componentes argumentativos identificados, determina la relacion
    dirigida desde el componente fuente (arg1) hacia el componente objetivo (arg2).

La relacion debe clasificarse en exactamente una de las siguientes etiquetas:
- Support
- Attack
- Partial-Attack
- No-relation

Definiciones de etiquetas:
- Support: el componente fuente apoya, justifica o aporta evidencia para el componente
     objetivo.
- Attack: el componente fuente cuestiona, contradice o se opone al componente objetivo
    .
- Partial-Attack: el componente fuente debilita o cuestiona parcialmente el componente
     objetivo, sin contradecirlo por completo.
- No-relation: no hay una relacion argumentativa dirigida explicita entre arg1 y arg2.

Importante: la negacion es critica en esta tarea. Presta especial atencion a
    marcadores de negacion como no, nunca, sin, falta de y expresiones similares,
    porque omitir una negacion puede invertir completamente la polaridad de la
    relacion.

Usa el contexto del documento cuando sea necesario.
Devuelve solo una etiqueta.

Source component (arg1_id={arg1_id}): "{source_text}"
Target component (arg2_id={arg2_id}): "{target_text}"
```

Prompt A.4: Track 2 Subtask 1 sentence relevance prompt

```
Dada una oracion del caso clinico, clasificala en una tarea binaria de relevancia.

La oracion debe clasificarse en exactamente una de las siguientes etiquetas:
- relevant
- not-relevant

Definiciones de etiquetas:
- relevant: la oracion aporta informacion necesaria para apoyar o refutar al menos una
     opcion.
- not-relevant: la oracion no aporta informacion util para apoyar o refutar ninguna
    opcion.

Usa el contexto del documento cuando sea necesario.
Devuelve solo una etiqueta.

Sentence: "{sentence}"
```

Prompt A.5: Track 2 Subtask 1 training instance

```
### Instruction:
Dada una oracion del caso clinico, clasificala en una tarea binaria de relevancia.

La oracion debe clasificarse en exactamente una de las siguientes etiquetas:
- relevant
- not-relevant

Definiciones de etiquetas:
- relevant: la oracion aporta informacion necesaria para apoyar o refutar al menos una
     opcion.
- not-relevant: la oracion no aporta informacion util para apoyar o refutar ninguna
    opcion.

Usa el contexto del documento cuando sea necesario.
Devuelve solo una etiqueta.

Sentence: "{sentence}"

### Context:
{complete clinical case}

### Response:
{relevant|not-relevant}
```

Prompt A.6: Track 2 Subtask 2 entity classification prompt

```
Dada una oracion del caso clinico, clasificala en una tarea binaria de los limites
    textuales exactos de los componentes argumentativos (Premise y Claim), ya sean
    terminos individuales, frases descriptivas o clausulas complejas.
La oracion debe clasificarse en exactamente una de las siguientes etiquetas:
- Premise
- Claim

Definiciones de etiquetas:
- Premise: la oracion aporta informacion necesaria para apoyar o refutar al menos una
    opcion.
- Claim: la oracion expresa una afirmacion que puede ser apoyada o refutada.

Usa el contexto del documento cuando sea necesario.
Devuelve solo una etiqueta.

Sentence: "{entity_text}"
```

Prompt A.7: Track 2 Subtask 2 training instance

```
### Instruction:
Dada una oracion del caso clinico, clasificala en una tarea binaria de los limites
    textuales exactos de los componentes argumentativos (Premise y Claim), ya sean
    terminos individuales, frases descriptivas o clausulas complejas.
La oracion debe clasificarse en exactamente una de las siguientes etiquetas:
- Premise
- Claim

Definiciones de etiquetas:
- Premise: la oracion aporta informacion necesaria para apoyar o refutar al menos una
    opcion.
```

```
- Claim: la oracion expresa una afirmacion que puede ser apoyada o refutada.

Usa el contexto del documento cuando sea necesario.
Devuelve solo una etiqueta.

Sentence: "{entity_text}"

### Context:
{complete clinical case}

### Response:
{Premise|Claim}
```

Prompt A.8: Track 2 Subtask 3 relation classification prompt

```
Dado un par de entidades argumentativas de un caso clinico, clasifica la relacion
    entre ambas en una tarea binaria.
arg1_id y arg2_id son IDs de entidades existentes en annotations.entities.
La salida debe ser exactamente una etiqueta entre:
- Support
- Attack

Definiciones:
- Support: arg1 apoya el contenido argumentativo de arg2.
- Attack: arg1 contradice, debilita o refuta el contenido argumentativo de arg2.

Usa el contexto completo cuando sea necesario.
Devuelve solo la etiqueta.

arg1_id: {arg1_id}
arg2_id: {arg2_id}
arg1_type: {arg1_type}
arg2_type: {arg2_type}
arg1_text: "{arg1_text}"
arg2_text: "{arg2_text}"
```

Prompt A.9: Track 2 Subtask 3 training instance

```
### Instruction:
Dado un par de entidades argumentativas de un caso clinico, clasifica la relacion
    entre ambas en una tarea binaria.
arg1_id y arg2_id son IDs de entidades existentes en annotations.entities.
La salida debe ser exactamente una etiqueta entre:
- Support
- Attack

Definiciones:
- Support: arg1 apoya el contenido argumentativo de arg2.
- Attack: arg1 contradice, debilita o refuta el contenido argumentativo de arg2.

Usa el contexto completo cuando sea necesario.
Devuelve solo la etiqueta.

arg1_id: {arg1_id}
arg2_id: {arg2_id}
arg1_type: {arg1_type}
arg2_type: {arg2_type}
arg1_text: "{arg1_text}"
```

```
arg2_text: "{arg2_text}"

### Context:
{complete clinical case}

### Response:
{Support|Attack}
```