Vicomtech at GRACE 2026: From Fine-Tuning to Multi-Strategy LLMs and Ensembles for Spanish Clinical Argument Mining

Alexander Platas[1,2,*,†], Alvaro Ruiz[1,†], Adriana Rodríguez[1,†], Paula Reyero[1,†], Elena Zotova[1,†], Pablo Turón[1,†], Iker Pérez[1,†] and Montse Cuadros[1,†]

[1]*Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)*
[2]*Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU), Paseo Manuel de Lardizabal, 1, 20018, Donostia/San-Sebastián (Spain)*

Abstract

This paper describes the participation of the **VicomNLP** team in Track 2 of the GRACE 2026 shared task, which focuses on clinical case reasoning through Argument Mining in Spanish. Track 2 frames the validation of medical diagnoses and treatments as a hierarchical pipeline across three subtasks: evidence sentence detection, fine-grained evidence span extraction (premises and claims), and directed relation classification (support or attack). We present a comprehensive comparative study evaluating encoder-only, encoder-decoder, and decoder-only architectures under both supervised fine-tuning and In-Context Learning paradigms. To address data scarcity, we develop a source unification and cross-lingual projection pipeline to augment the training data using the CasiMedicos-Arg dataset. Additionally, we implement task-specific ensembling strategies based on weighted voting and token-level clustering to mitigate error propagation across the pipeline. Our best configuration—an ensemble combining open-source domain-specific models and commercial large language models—secured 1st place in the official competition with an overall F1-score of 59.84%, with our submissions capturing the top four positions on the leaderboard. To support reproducibility, the code, system descriptions, and resources developed for this work are publicly available at https://github.com/Vicomtech/grace2026 repository.

Keywords

Clinical NLP, Large Language Models (LLM), Argument Mining (AM), Evidence-Based Medicine

1. Introduction

In modern healthcare, evidence-based medicine demands that clinical decisions be backed by solid, verifiable rationale. As Artificial Intelligence (AI) becomes increasingly integrated into the medical domain, the transition from black-box predictive models toward explainable AI has become a paramount priority [1]. Argument Mining (AM) offers a promising paradigm to bridge this gap by automatically identifying the structure of human reasoning in text, isolating evidence components, and mapping their logical connections. While AM has advanced significantly in general domains, its application to clinical narratives remains scarce, particularly for non-English languages.

To address this limitation, the Granular Recognition of Argumentative Clinical Evidence (GRACE) 2026 shared task [2, 3] introduces the first AM challenge tailored specifically to Spanish clinical texts. Specifically, Track 2 focuses on clinical case reasoning based on questions from the Spanish Resident Medical Intern (MIR) national examination. The task models the cognitive process required to validate a correct diagnosis or treatment. This is structured as a three-stage hierarchical pipeline where models

IberLEF 2026, September 2026, León, Spain
*Corresponding author.
†These authors contributed equally.
aplatas@vicomtech.org (A. Platas); aruiz@vicomtech.org (A. Ruiz); arodriguezf@vicomtech.org (A. Rodríguez); preyero@vicomtech.org (P. Reyero); ezotova@vicomtech.org (E. Zotova); pturon@vicomtech.org (P. Turón); iperezd@vicomtech.org (I. Pérez); mcuadros@vicomtech.org (M. Cuadros)
0009-0002-5501-017X (A. Platas); 0009-0008-8282-7955 (A. Ruiz); 0009-0005-3069-0449 (A. Rodríguez); 0000-0001-5238-4550 (P. Reyero); 0000-0002-8350-1331 (E. Zotova); 0000-0002-5563-1120 (P. Turón); 0009-0008-4582-2311 (I. Pérez); 0000-0002-3620-1053 (M. Cuadros)

must: (i) identify relevant evidence sentences, (ii) extract fine-grained sub-sentence spans corresponding to argumentative entities (*Premises* and *Claims*), and (iii) classify the directed logical relations (*Support* or *Attack*) between them. This setup introduces severe computational challenges, most notably the cascading of errors across sequential tasks and an acute class imbalance during relation detection.

In this paper, we describe the competitive systems developed by the **VicomNLP** team for GRACE Track 2. We conduct a rigorous empirical exploration of diverse language model architectures—spanning encoder-only, encoder-decoder, and state-of-the-art decoder-only Large Language Models (LLMs). Our methodologies encompass task-specific supervised fine-tuning, Parameter-Efficient Fine-Tuning via Low-Rank Adaptation (LoRA) [4], In-Context Learning (ICL) [5] setups using both independent and global multi-task prompting strategies.

Furthermore, we conduct a cross-resource technical exploration by standardizing and aligning the external CasiMedicos-Arg dataset [6] to match the GRACE annotation schema, manually resolving localized translation artifacts. This allows us to evaluate the generalizability of our techniques across different clinical data sources prior to their final deployment on the GRACE dataset.

Our experimental results demonstrate that while specialized encoder-based fine-tuning achieves optimal performance for sentence classification, prompting large generative decoders is more effective for structured span extraction and relation mapping. Our developed ensembles successfully mitigated individual model limitations. Ultimately, the **VicomNLP** systems dominated the shared task overall ranking leaderboard, capturing the top four positions. Our best-performing configuration (RUN5) achieved the highest overall F1-score of 59.84%, outperforming the baseline by 11.94 points and establishing a substantial margin over competing systems.

The main contributions of our work are summarized as follows:

- We perform a comprehensive benchmark evaluating how encoder, encoder-decoder, and decoder-only models handle complex, multi-stage clinical AM under various training paradigms.
- We engineer an automated alignment and injection pipeline to project argumentative relations onto translated text, successfully generating an extended, unified version of the CasiMedicos-Arg dataset for data augmentation in clinical reasoning.
- We implement a robust, hyperparameter-controlled ensemble mechanism tailored to the unique textual and relational outputs required by the GRACE schema.
- We deliver state-of-the-art results in Spanish clinical reasoning, achieving 1st place overall in the GRACE 2026 Track 2 competition.

The implementation details of our systems, the experimental results, and the updated resource related to the CasiMedicos-Arg dataset is publicly available in a shared repository [7].

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature in clinical AM. Section 3 details the shared task framework, the datasets utilized and data unification pipeline, while Section 4 outlines our proposed architectures and task-specific ensembling strategies. Our experimental configurations and official competition results are reported in Section 5. Finally, Section 6 provides a comprehensive analysis of our findings, and Section 7 concludes the paper with directions for future work.

2. Related Work

AM focuses on the pragmatic-level analysis of discourse, applying argumentation theory to model and automatically analyze textual data [8]. It consists of two main stages: argument extraction and argument relation prediction. First, AM systems extracted argumentative structures from free text with BM25, TF-IDF, Word2Vec [9, 10, 11], BiLSTMs [12, 13]. Recently, AM with LLMs has led to performance improvements in argument extraction [14, 15].

In the medical domain, [1, 16, 17] proposed an end-to-end pipeline for argumentative outcome analysis in clinical trials leveraging the MEDLINE database and deep bidirectional transformers combined with

neural architectures such as LSTM, GRU, and CRF. In Spanish NLP, automatic translation combined with annotation projection from English to Spanish showed its effectiveness for constructing annotated data [18]. [19] explored AM in data-scarce settings using Natural Language Inference (NLI)-based techniques on clinical texts derived from MIR exam explanations, demonstrating that token classification combined with NLI outperforms direct text classification approaches for extracting argumentative structures that support or refute candidate diagnoses.

Antidote CasiMedicos dataset [20] comprises MIR exam questions enriched with gold explanatory arguments authored by volunteer medical doctors, constituting one of the first resources for argumentation-driven explainable AI in Spanish. Extending this line, [6] presented CasiMedicos-Arg, the first multilingual dataset for Medical Question Answering annotated with gold argumentative structures across four languages (EN, ES, FR, IT). CasiMedicos-Arg makes it possible to study not only AM but also generative argumentation, and serves as a key auxiliary resource in our work. [21] presented MedExpQA, the first multilingual benchmark (EN, ES, FR, IT) for evaluating LLMs on Medical Question Answering with reference gold explanations, and showed through RAG experiments that performance for non-English languages still lags considerably behind English.

Pre-trained transformer models have become the backbone for clinical NLP tasks, particularly for sequence labelling and relation extraction. Encoder-only models in the BERT family, including domain-adapted variants such as `BioBERT`, `ClinicalBERT`, and the multilingual `XLM-RoBERTa`, have demonstrated consistent state-of-the-art results on biomedical named entity recognition and relation extraction. For Spanish clinical text, multilingual models such as `EriBERTa` [22]. Encoder-decoder models, such as `Medical-mT5` [23] and `mT5` [24] extend the paradigm to generative formulations of extraction tasks, enabling span extraction and structured output generation in a text-to-text framework. LLMs showed their competitive results in AM task in medical texts in English [25]. Evaluating OpenAI's `GPT-4o` and `o1` on the MIR 2024 exam report in Spanish [26] achieved accuracies of 89.8% and 92.6% respectively, substantially above the average candidate score, confirming that state-of-the-art LLMs possess strong domain knowledge for the MIR setting.

3. Task Description and Datasets

3.1. Subtask Descriptions

Track 2 models the clinical reasoning required to validate a correct diagnosis or treatment while rejecting competing alternatives. The track is structured into 3 hierarchical subtasks designed to be solved as a pipeline, with one's error cascading onto the next, as illustrated in Figure 1.

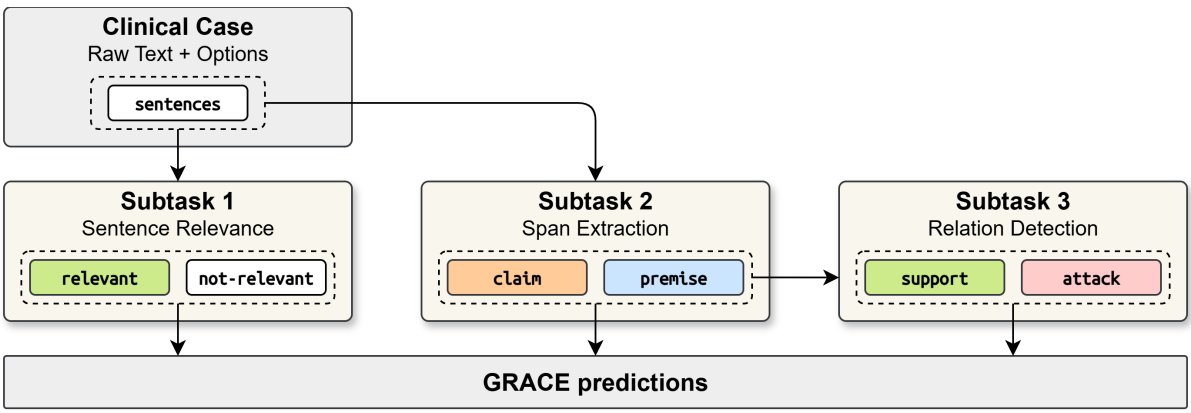

Figure 1: Hierarchical pipeline for the GRACE Track 2 subtasks.

- **Subtask 1 — Evidence sentence detection:** A binary sentence-level classification task. Systems must classify each sentence in the clinical narrative as positive (*Relevant*) if it contains information required to support or refute at least one of the candidate options, or negative (*Not-Relevant*) otherwise.
- **Subtask 2 — Evidence span detection:** A fine-grained extraction task requiring deep contextual understanding. Systems must identify the exact textual boundaries of specific clinical evidence (terms, phrases, or complex clauses) within the text, especially the extracted relevance sentences. These extracted sub-sentence spans act as *Premises*, while the five candidate answer options typically function as competing *claims*.
- **Subtask 3 — Relation detection:** A directed relation classification task. Given the detected evidence spans and the candidate claims, the system must predict whether there is a logical relation linking them. The relation is classified as *Support* if the evidence suggests the option is clinically appropriate, or *Attack* if the evidence suggests the option is incorrect, contraindicated, or insufficient.

3.2. Datasets

Official GRACE Track 2 dataset. The primary dataset is the official GRACE 2026 Track 2 dataset, composed of Spanish clinical cases derived from MIR medical examination material. Each case contains the full clinical text, sentence boundaries, five answer options, and gold annotations aligned with the three subtasks: sentence-level relevance labels, fine-grained argumentative entities, and directed *Support*/*Attack* relations. The dataset comprises 254 cases, split into 128 training, 24 development, and 102 test instances.

The official records are distributed in a GRACE JSON format. Each instance includes the raw clinical text, metadata describing the clinical context and answer choices, and an annotation block containing sentence relevancy labels, argumentative entities, and relations. This format serves as the target representation for all training, inference, and evaluation components in our system.

CasiMedicos-Arg dataset. To increase the amount and diversity of available argumentative supervision, we also used CasiMedicos-Arg [6], a multilingual medical question-answering dataset with explanatory argumentative structures. The dataset contains 558 MIR-style clinical cases with doctor-authored argument annotations. It is available in Spanish and in automatically translated English, French, and Italian versions, with annotated argumentative components and *Support*/*Attack* relations.

Before using this resource in our experiments, we produced an extended version of CasiMedicos-Arg prepared for GRACE-style argument mining. The processing normalizes the original heterogeneous files into JSONL, preserves the original train/dev/test organization, projects English relation annotations to the Spanish, French, and Italian records, and applies manually checked fixes for ambiguous or skipped relation alignments. The main statistical differences between the original dataset and the extended version are displayed in Table 1. Detailed preprocessing reports, alignment examples, and further statistics are provided in the project repository [7].

Dataset	File formats	Languages w/ Text Data	Languages w/ Relations	Clinical cases	Support Relations	Attack Relations
Original	*jsonl, json, tsv*	*en, es, fr, it*	*en*	558	2431	1106
Extended (our)	*jsonl*	*en, es, fr, it*	*en, es, fr, it*	558	∼2420 *	∼1085 *

*Average by language. Minor reductions reflect the resolution of underlying duplicate pairs.

Table 1
Comparison of the CasiMedicos-Arg dataset structure and statistics before and after extension.

Unified GRACE-style representation. Since the official GRACE data and CasiMedicos-Arg follow different annotation conventions, we converted the processed CasiMedicos-Arg records into the GRACE JSON schema before training and experimentation. This produced a homogeneous data representation in which original GRACE instances and converted CasiMedicos-Arg instances can be handled by the same preprocessing, modeling, ensembling, and evaluation code.

The source unification pipeline reconstructs the raw clinical text from tokenized CasiMedicos-Arg records, recomputes sentence, answer-choice, and entity offsets, and maps BIO argument labels to GRACE-style *Premise* and *Claim* entities. Relation endpoints are then aligned with the converted entities, preserving their original *Support* or *Attack* labels when both endpoints can be mapped consistently. Because CasiMedicos-Arg does not provide GRACE-style sentence relevance annotations, these labels are derived from the converted argumentative structure: a sentence is marked as *Relevant* when it contains an argumentative component participating in a valid relation, and as *Not-Relevant* otherwise.

Finally, all converted records are validated for offset consistency, sentence-label alignment, and relation integrity. Each instance receives an `origin` field indicating whether it comes from the official GRACE dataset or from CasiMedicos-Arg. The resulting unified files allow us to compare models trained only on GRACE with models trained or prompted using additional CasiMedicos-derived supervision while preserving a single GRACE-compatible format throughout the pipeline. Detailed unification procedures, code and execution examples are also provided in the project repository [7].

4. Proposed Approach

For the experimental evaluation, three distinct model architectures were explored. For each architecture, task-specific techniques were applied to address each subtask independently. This section describes the language models employed, their corresponding learning methodologies, and the proposed approaches. Each subtask was addressed independently to allow for a systematic evaluation of various models through both ICL and fine-tuning strategies.

4.1. Encoder-only Models

Table 2 displays encoder-only transformer architectures we adopted as the underlying framework for all three subtasks. Encoder-based language models generate contextualized representations for each input token by leveraging bidirectional self-attention mechanisms, allowing the model to capture both local and global semantic dependencies within the clinical narrative. These contextual embeddings can then be adapted to different downstream tasks through lightweight task-specific prediction heads.

Model	Open Source	Medical Domain	Size	Release Date
mBERT [27]	✓	✗	$<$ 1B	May, 2019
XLM-RoBERTa [28]	✓	✗	$<$ 1B	Apr, 2020
EriBERTa [22]	✓	✓	$<$ 1B	Jun, 2023
BioMedRoBERTa [29]	✓	✓	$<$ 1B	Sep, 2021
IXAmBERT [30]	✓	✓	$<$ 1B	May, 2020

Table 2
Encoder-only models used.

Subtask 1. The first subtask is formulated as a binary sentence classification problem. Each sentence extracted from the clinical narrative is independently encoded by the transformer model, and the representation of the reference classification token is used to obtain a sentence-level embedding. A linear classification head is subsequently applied to predict whether the sentence is *Relevant* or *Not-Relevant*.

Subtask 2. The second subtask is addressed as a Named Entity Recognition (NER)-style sequence labeling problem. Instead of assigning a single label to the entire sentence, the model predicts a label for each token in the input sequence. To identify the exact boundaries of evidence spans, we employ a BIOES tagging scheme, which explicitly marks the beginning (B), inside (I), outside (O), end (E), and single-token (S) entities. Each token is classified into tags representing *Premise*, *Claim*, or *Outside*, enabling the extraction of fine-grained argumentative evidence from the clinical text.

Subtask 3. The third subtask is formulated as a multi-class relation classification problem. Given an evidence span and a candidate claim, both elements are encoded jointly and processed by the transformer encoder. The contextual representation of the reference classification token is then used as a holistic representation of the evidence–claim pair. A classification layer predicts the semantic relation between both components, assigning one of three possible labels: *Support*, *Attack*, or *Nothing*. The *Nothing* label is used when no meaningful argumentative relation exists between the evidence span and the candidate claim.

4.2. Encoder-Decoder Models

As summarized in Table 3, the implementation is fully based on a text-to-text formulation and is realized with multilingual T5 fine-tuning. Rather than treating the task as a standard classifier with a task-specific head, the system uses an encoder-decoder architecture that generates short label strings.

Model	Open Source	Medical Domain	Size	Release Date
mT5-base [24]	✓	✗	$<$ 1B	Mar, 2021
Medical mT5-large [23]	✓	✓	$<$ 1B	Apr, 2024

Table 3
Encoder-decoder models used.

Subtask 1. For each dataset example, the prompt is constructed so that the sentence under evaluation appears in highlighted form inside the case context, and answer choices are appended to preserve decision context. This input design ensures that the notion of relevance is grounded in the alternatives the model must distinguish, which is essential in medical reasoning tasks where evidential value is conditional on competing options. Label handling is designed to preserve compatibility with external evaluation while keeping generation simple. Gold labels from the dataset are represented externally as *Relevant* and *Not-Relevant*, but they are mapped internally to *yes* and *no* for generation targets. Constrained decoding builds a prefix trie over the valid generated label sequences and injects a prefix-allowed token function at prediction time. As a result, generated outputs are restricted to valid labels followed by end-of-sequence termination, preventing malformed generations. Any unexpected decoded form is deterministically mapped to *Not-Relevant*.

Subtask 2. The preprocessing builds four complementary sentence-centered sample types. (1) *sentence* samples are created for each context sentence and include entities whose offsets fall inside that sentence. (2) *claims* samples capture *claim* entities appearing in answer options. (3) *choice_premise* samples supervise *Premise* entities inside options. (4) *sentence_window* samples recover entities spanning multiple sentences through short multi-sentence fragments. Targets are serialized deterministically as typed spans sorted by offset; samples without valid entities use `none`.

The model receives a Spanish instruction prompt plus contextual text and generates typed span lists instead of token-level tags. A constrained parser extracts predictions by recognizing the prefixes `Premise:` and `Claim:`. Predicted spans are then mapped back to character offsets in the original text by searching within sample-specific regions, such as sentence boundaries, option ranges, or localized fragments. Type constraints are applied per sample type to avoid structurally invalid labels. This decomposition provides denser and more localized supervision than document-level generation while preserving compatibility with offset-based evaluation through postprocessing and aggregation. It addresses the trade-off between efficient generative training and faithful task scoring.

Subtask 3. The dataset for subtask 3 is prepared in two stages. First, case-level annotations are converted into pair-level examples by generating all premise–claim combinations. Annotated relations are labeled as *Support* or *Attack*, while unannotated pairs receive the label *Nothing*, providing explicit negative supervision and training abstention behavior. Second, records are grouped by `case_id` and `arg1_id`, where `arg1` corresponds to the evidence span. Each grouped instance contains one evidence fragment and all candidate options. The input prompt is formulated as a Spanish two-step reasoning instruction asking the model first to assess relevance between the evidence and each option, and then to determine whether the evidence supports or contradicts relevant options. The prompt includes the evidence fragment, all options, and the full clinical context. Targets are generated as structured multi-label sequences with positional indices, e.g., 1:*ataca* 2:*apoya* 3:*ninguna*. During postprocessing, Spanish labels are normalized to canonical English labels.

4.3. Decoder-only Models

Table 4 provides a comprehensive summary of decoder-only models selected for this study. Our selection criteria prioritize a diverse range of architectures, encompassing both open-source and commercial models, covering general-purpose and domain-specific (medical) variants. Ranging in size from 2B to 27B parameters, these models represent the latest iterations within their respective families, ensuring that our comparative analysis captures state-of-the-art performance trends.

Model	Open Source	Medical Domain	Size	Release Date
GPT 5.4 mini [31]	✗	✗	–	Mar, 2026
Gemini 3 Flash [32]	✗	✗	–	Dec, 2025
Qwen 3.5 [33]	✓	✗	2B, 4B	Mar, 2026
Gemma 4 [34]	✓	✗	5B (E2B)	Apr, 2026
MedGemma [35]	✓	✓	4B, 27B (text-only)	May, 2025
MedGemma 1.5 [36]	✓	✓	4B	Jan, 2026
MediPhi [37]	✓	✓	4B	Jul, 2025

Table 4
LLMs used. "–" indicates unknown data.

All LLMs were tested under both zero-shot and few-shot ICL paradigms. Furthermore, we performed fine-tuning specifically on domain-adapted models, namely `MedGemma` and `MedGemma-1.5`. To manage the computational demands of the 27B parameter models while maintaining performance integrity, we utilized LoRA, which facilitated parameter-efficient training and optimized resource allocation.

Subtask 1. Sentence relevancy detection was formulated as a binary classification problem. For each clinical case, the model received the complete clinical context, the candidate answer options, and a sentence to classify. The model was constrained to generate one of two labels: *Relevant* or *Not-Relevant*. This formulation was applied consistently across both ICL-based inference and fine-tuning settings. In the fine-tuning setup, each training instance was converted into a chat-style prompt-response pair, where the assistant response corresponded to the gold relevancy label. During inference, zero-shot and few-shot prompts followed the same task definition, and the generated outputs were normalized through deterministic post-processing, assigning malformed or ambiguous generations to the *Not-Relevant* class.

Subtask 2. Entity extraction was performed at the sentence level using the full clinical case as contextual input. The task targets argumentative *Premise* and *Claim* entities. Two extraction strategies were evaluated. The first treats the task as joint extraction, where both *Premise* and *Claim* spans are predicted by the model. The second uses a hybrid deterministic approach: answer options are directly inserted as *Claim* entities with preserved offsets, while the model extracts only *Premise* spans from the clinical text. This design improves premise recall while ensuring high claim precision through deterministic claim generation. In both settings, the model processes one sentence at a time with full-case context.

Subtask 3. Relation classification was formulated as a pairwise argumentative relation detection task. Given the full clinical case and a premise–claim pair, the model classified their relation as *Support*, *Attack*, or *Nothing*. In order to mitigate the class imbalance arising from the generation of all possible premise-claim combinations, a subset of these pairs was randomly discarded during training. Finally, only predictions labeled as *Support* or *Attack* were retained in the final output, while *Nothing* predictions were filtered out as they do not correspond to explicit relation annotations.

Global Approach. Additionally, we implemented a "global" experimental approach, where the models were tasked with resolving all three subtasks simultaneously within a single inference request under strict prompting restrictions which are specified in the project repository [7]. This setup aims to assess the models' capacity to handle multi-task dependencies and maintain coherence when generating comprehensive outputs.

4.4. Ensembling approaches

We explored ensemble methods as a task-specific aggregation mechanism over multiple GRACE-format prediction files. Given that the three subtasks produce structurally different outputs, predictions are ensembled independently for sentence relevance, entity extraction, and relation extraction, while preserving a shared JSON structure compatible with the official scorer.

The aggregation process preserves the original case information and replaces only the prediction block with the aggregated output. Model identifiers are inferred from input files and may optionally be assigned task-specific weights computed using the official evaluation scorer. These weights increase the influence of higher-performing models in weighted aggregation schemes.

Table 5 summarizes the ensemble strategies explored for each subtask.

Subtask	Strategy	Description
1	Majority voting	Most frequent sentence relevance label across models.
	Weighted voting	Model votes weighted by task-specific validation scores.
2	Best model	Entity predictions taken from the highest-scoring model.
	Exact vote	Entities matched by identical start offset, end offset, and type.
	Cluster vote	Clusters same-type spans via token IoU to tolerate boundary variation.
3	Exact vote	Keeps relations whose endpoints exactly match aggregated entity spans.
	Entity-aligned vote	Relation endpoints mapped to aggregated entities via token-level IoU.
	Cluster vote	Relations clustered by argument-span overlap and relation type.

Table 5
Task-specific ensemble strategies implemented for the GRACE subtasks.

Subtask 1. Sentence relevance is treated as a sequence labelling problem, where each sentence receives a final label through majority or weighted voting. In case of ties, the system applies a deterministic policy favouring *Relevant* labels, in order to avoid discarding potentially useful argumentative evidence.

Subtask 2. Entity extraction requires aggregating span-based predictions. Exact voting only combines entities with identical boundaries and types, whereas cluster voting relaxes this requirement by grouping spans of the same type according to token-level Intersection over Union (IoU). This makes the ensemble more tolerant to small boundary differences across models.

Subtask 3. Relation extraction is aggregated over the final entity space, ensuring that relation arguments point to valid entities in the aggregated prediction block. Exact voting requires strict endpoint agreement, while entity-aligned voting and cluster voting use token-level IoU to tolerate minor span differences. As with entity aggregation, relations are retained only when supported by a minimum number of distinct models.

The behaviour of the ensemble pipeline is controlled by three main hyperparameters:

- **Minimum votes**: minimum number of distinct models required to retain an entity or relation.
- **Entity IoU**: token-level overlap threshold used for entity clustering and entity alignment.
- **Relation IoU**: token-level overlap threshold used for relation clustering and endpoint alignment.

These parameters regulate the precision–recall trade-off: higher thresholds and stricter IoU values yield more conservative ensembles, while lower thresholds increase recall by allowing weaker agreement across models. The framework thus combines complementary model behaviours while maintaining scorer-compatible GRACE outputs.

5. Experimentation and Results

This section describes the comprehensive experimental evaluation conducted for each of the three hierarchical subtasks within Track 2. Our evaluation highlights the performance discrepancies across encoder-only, encoder-decoder, and decoder-only architectures, analyzing the impact of training paradigms (zero-shot, few-shot, and fine-tuning) and the integration of auxiliary domain data. It is important to note that all performance metrics reported in this section were obtained by evaluating the models on the official validation dataset.

5.1. Experiments Overview

5.1.1. Subtask 1 - Evidence Sentence Detection

Evidence sentence detection, is a binary sentence-level classification task, which had the primary objective of determining whether a given sentence within the clinical narrative contained crucial information to support or refute any candidate options. To assess performance, we have employed the F_1-score of the positive (*Relevant*) class across all systems.

The empirical data shown in Figure 2 allow us to draw several crucial conclusions regarding sentence relevance detection in clinical texts: First, task-specific fine-tuning demonstrated clear superiority, as fine-tuned models systematically outperformed broad-scale generative prompting. This trend is closely tied to the impact of domain pre-training; specialized medical models such as `EriBERTa` and `MedGemma` substantially outperformed their corresponding general-domain baselines. This superior performance confirms that a pre-existing understanding of clinical terminology (such as symptoms, patient history, and diagnostic values) is crucial for determining if a statement holds evidential value.

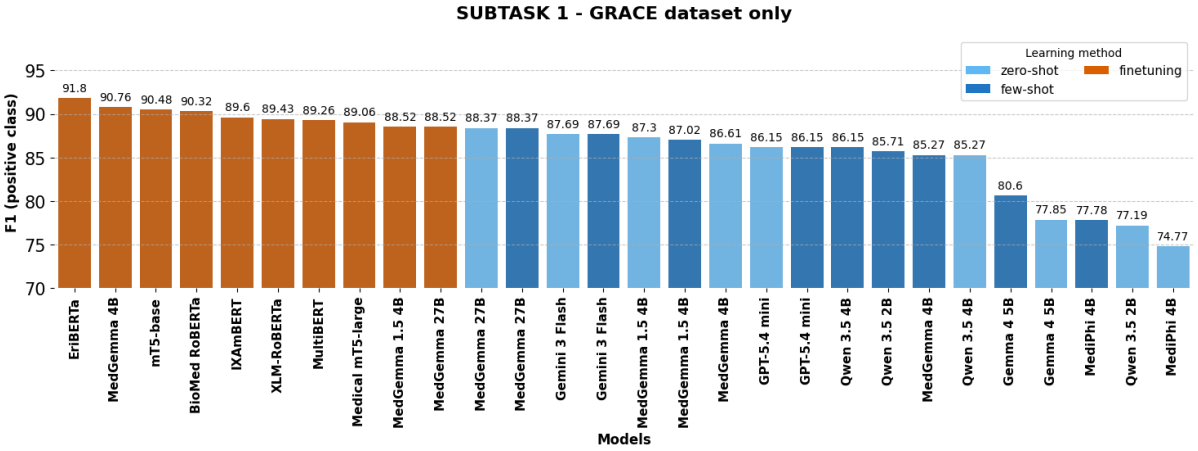

Figure 2: Language model performance on subtask 1 using only the GRACE dataset

Conversely, an ICL paradox was observed during experimentation. Few-shot prompting yielded only minor accuracy enhancements when applied to smaller-scale models, whereas stacking conversational examples within the prompts of larger generative models actually triggered a degradation in results.

Finally, cross-domain data augmentation proved ineffective for this task, as shown in Figure 3. Combining the GRACE dataset with translated data from the CasiMedicos-Arg dataset (specifically introducing the Spanish, Spanish+English. Spanish+English+Italian+French subsets) failed to deliver tangible performance gains for subtask 1, leaving the performance ceiling stagnant at 91.8%. This strongly suggests that the core linguistic patterns defining sentence relevance are already fully captured by the native GRACE training set.

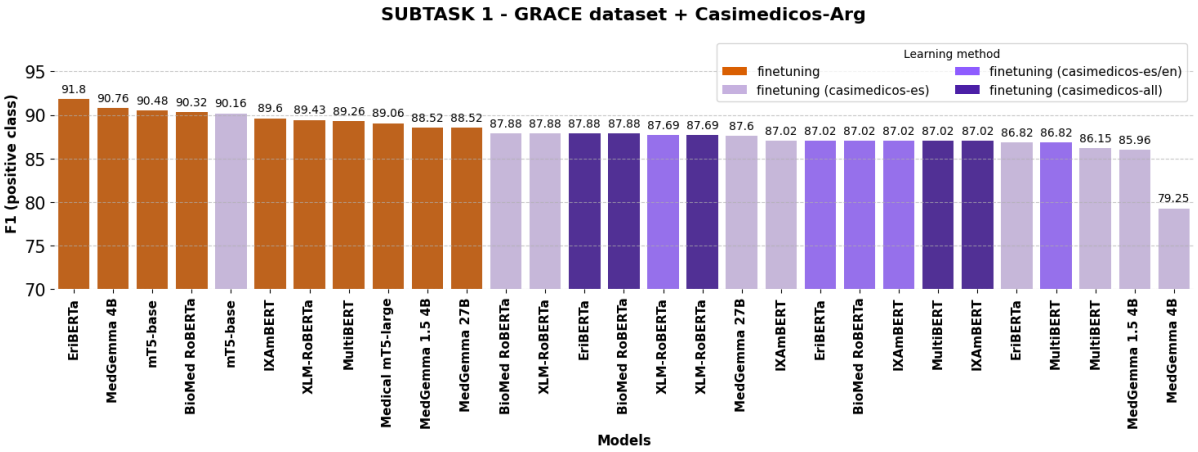

Figure 3: Language models performance on subtask 1 using Grace and Casimedicos-Arg datasets.

5.1.2. Subtask 2 - Evidence Span Detection

Subtask 2 presents a significantly higher level of linguistic complexity, as it requires models to pinpoint the exact textual boundaries of premises and claims at a fine-grained, sub-sentence level. Figure 4 presents the results obtained with both ICL and fine-tuning approaches when using only the data provided by the shared task.

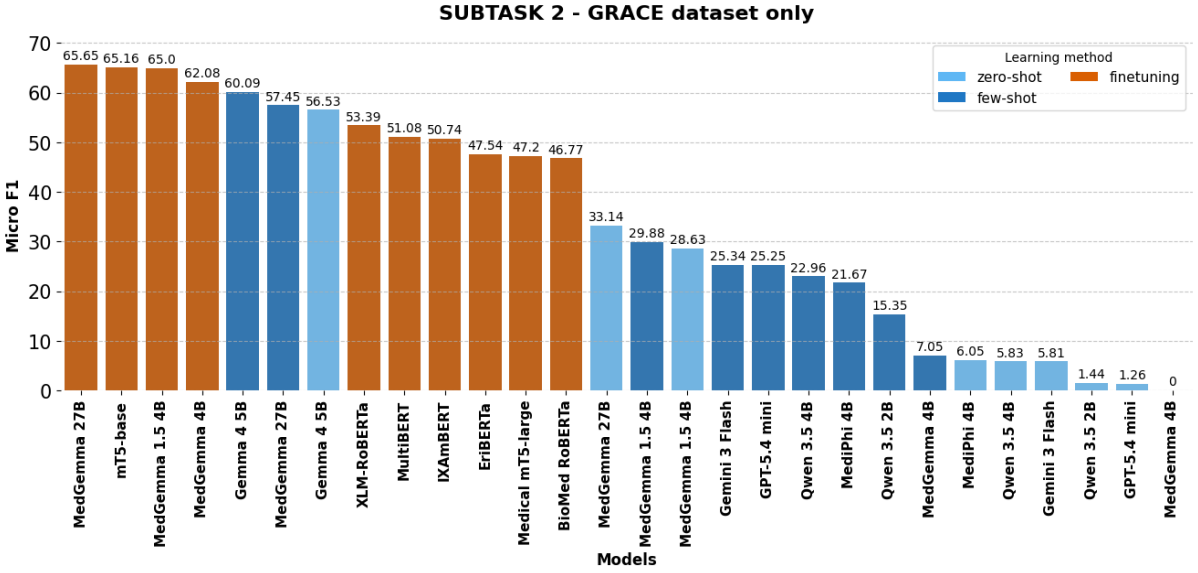

Figure 4: Language models performance on subtask 2 using only the GRACE dataset.

The results clearly show that, given the complexity of the task, ICL substantially underperforms compared to fine-tuned models. Among the evaluated ICL configurations, only the `MedGemma-27B` model in the few-shot setting achieves performance levels that approach those of the fine-tuned models.

Overall, these findings highlight the limitations of ICL for this task and the importance of task-specific adaptation through fine-tuning. Interestingly, as shown in Figure 5, incorporating the CasiMedicos-Arg dataset, particularly its Spanish subset, leads to a slight improvement over fine-tuning using only the GRACE training data. However, the performance gain is marginal, suggesting that the additional domain-specific data provides only limited benefits for this task.

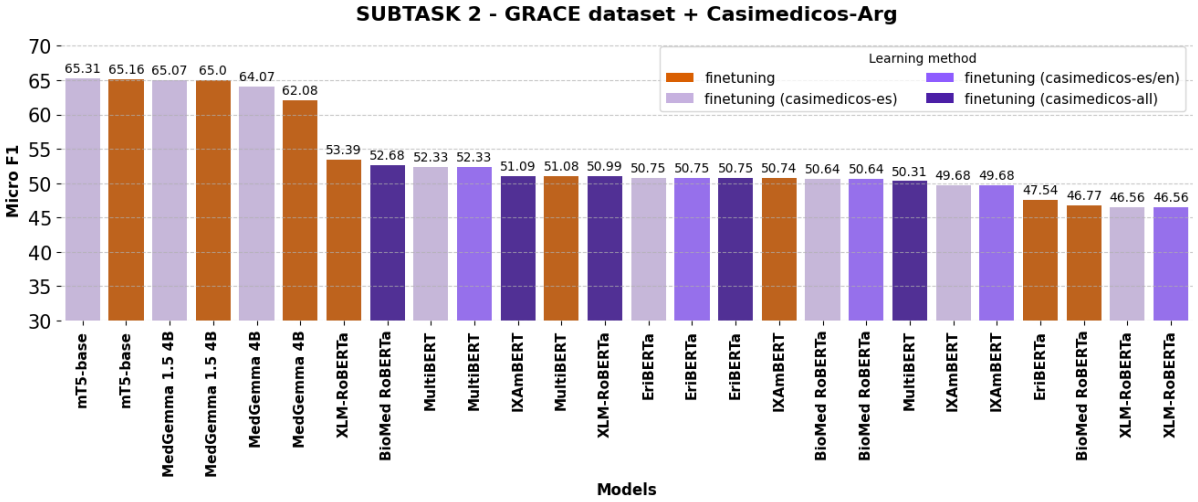

Figure 5: Language models performance on subtask 2 using GRACE and Casimedicos-Arg datasets.

5.1.3. Subtask 3 - Relation Detection

Subtask 3 maps directed argumentative edges linking extracted evidence spans to candidate option claims. For this evaluation, and to ensure a fair comparison across approaches, the gold-standard annotations from subtask 2 were used as the starting point for relation classification.

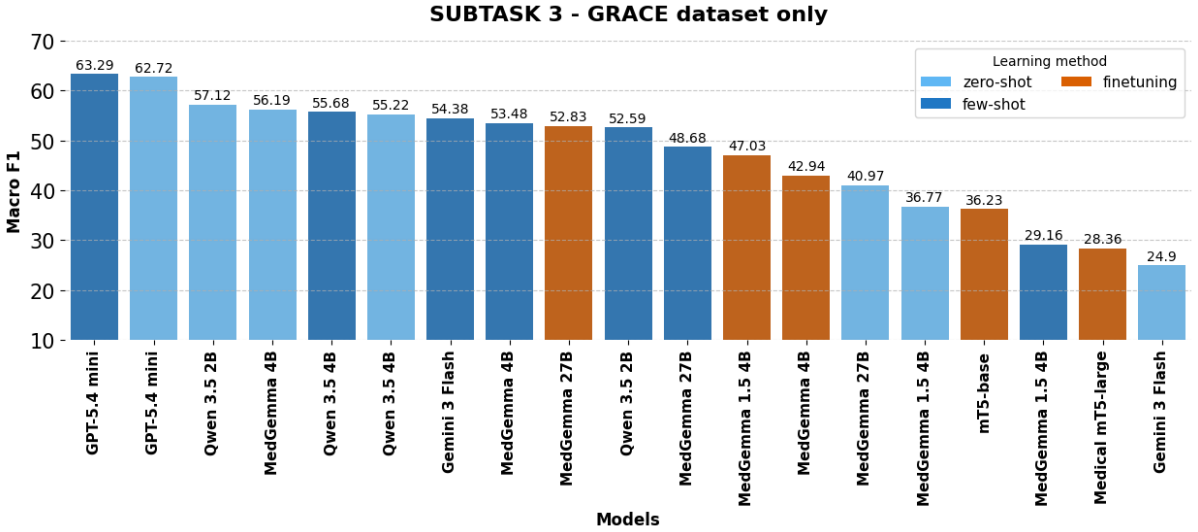

Figure 6: Language models performance on subtask 3 using only the GRACE dataset.

Since the LLM-based approach classified relations independently by evaluating every possible premise–claim pair, and because most premise–claim combinations do not correspond to a valid relation, a third class, *Nothing*, was introduced to indicate the absence of a relationship between a premise and a claim.

During inference, all premise–claim pairs were assigned to one of three classes: *Support*, *Attack*, or *Nothing*. Pairs predicted as *Nothing* were subsequently removed from the final prediction file, such that the official evaluation was performed exclusively on the original relation types.

In contrast to the previous tasks, a notable performance deficit was observed with fine-tuning (see Figure 6), as fine-tuned models systematically underperformed compared to zero-shot and few-shot approaches. Additionally, it is critical to highlight the total absence of encoder-only baselines, as encoder-only architectures were not included or successfully evaluated within this specific relation detection subtask.

5.1.4. Global Approach

Under the global approach, models predict all three subtasks in a single inference step to directly yield the complete JSON output. Table 6 outlines the performance metrics for each model per subtask, along with the score differentials relative to independent task execution.

Model	Zero-shot			Few-shot		
	Subtask 1	Subtask 2	Subtask 3	Subtask 1	Subtask 2	Subtask 3
MedGemma 4B	82.09 ↓ 4.52	61.75 ↑ 61.75	6.70 ↓ 49.49	67.83 ↓ 17.44	67.92 ↑ 60.87	13.17 ↓ 40.31
MedGemma 1.5 4B	39.02 ↓ 48.28	65.02 ↑ 36.39	10.71 ↓ 26.06	77.78 ↓ 9.24	57.00 ↑ 27.12	2.28 ↓ 26.88
MedGemma 27B	83.33 ↓ 5.04	64.79 ↑ 31.65	17.57 ↓ 23.40	78.18 ↓ 10.19	66.09 ↑ 8.64	20.26 ↓ 28.42
Qwen 3.5 2B	82.54 ↑ 5.35	62.19 ↑ 60.75	4.64 ↓ 52.48	78.95 ↓ 6.76	61.83 ↑ 46.48	10.51 ↓ 42.08
Qwen 3.5 4B	81.67 ↓ 3.60	62.60 ↑ 56.77	11.30 ↓ 43.92	81.60 ↓ 4.55	69.20 ↑ 46.24	21.57 ↓ 34.11
Gemini 3 Flash	74.55 ↓ 13.14	66.95 ↑ 61.14	25.17 ↑ 0.27	66.67 ↓ 21.02	69.06 ↑ 43.72	28.60 ↓ 25.78
GPT 5.4 mini	88.19 ↓ 2.04	62.76 ↑ 61.50	23.69 ↓ 39.03	88.89 ↑ 2.74	64.59 ↑ 39.34	22.25 ↓ 41.04

Table 6
Metrics obtained through the global resolution of the subtasks. The smaller values represent the difference compared to the individual resolution of the subtasks. The values in green indicate an increase (↑) in the metrics, while red values indicate a decrease (↓).

Notably, all tested models demonstrate a substantial improvement in subtask 2 using this strategy. In contrast, performance degrades in subtasks 1 and 3. The drop in subtask 3 is anticipated, as it relies on subtask 2 predictions; despite the surprisingly high metrics in subtask 2, error propagation occurs compared to using the gold standard.

Additionally, no clear pattern or significant variance emerges between zero-shot and few-shot settings. Processing all tasks simultaneously means that providing large prompt examples can unpredictably affect model performance based on the specific architecture.

5.2. Submitted Approaches

Each submission requires predictions for all three subtasks. Consequently, our selection process focused not only on identifying the top-performing individual models for each task but also on determining the optimal model combinations. To achieve this, the ensembling pipeline was automated, enabling the evaluation of multiple model combinations and strategies to identify the highest-performing ensembles.

5.2.1. Subtask 1

Regarding subtask 1, Table 7 lists the top-performing individual models, specifically those achieving an F_1-score above 0.90 on the dev set. For this task, two ensemble configurations yielded the best overall results. The non-commercial (NC) version combines the top four individual models (`EriBERTa`, `MedGemma-4B`, `BioMed-RoBERTa`, and `mT5`) via weighted voting. The commercial (C) version incorporates these same models alongside predictions from `GPT-5.4-mini` under both zero-shot and few-shot settings. Ultimately, `EriBERTa`, `MedGemma-4B`, and `mT5` were selected for individual submission (in addition to the two ensembles) due to their high scores and architectural diversity.

Subtask 1 - Best runs				
Run	**Model**	**Method**	**Dataset**	**Score ↓**
RUN5	Ensemble (C)	Weighted Voting	GRACE + CasiMedicos	92.80
RUN4	Ensemble (NC)	Weighted Voting	GRACE	92.68
RUN1	EriBERTa	Finetuning	GRACE	91.80
RUN2	MedGemma 4B	Finetuning	GRACE	90.76
RUN3	mT5-base	Finetuning	GRACE	90.48
-	BioMed RoBERTa	Finetuning	GRACE	90.32
-	mT5-base	Finetuning	GRACE + CasiMedicos	90.16

Table 7
Top results for subtask 1. "–" denotes that the model was not chosen for any submitted run.

5.2.2. Subtask 2

For subtask 2, two ensemble configurations also surpassed the performance of the best individual models. The NC ensemble comprises `Qwen-3.5-4B` (few-shot, global), `MedGemma-4B` (few-shot, global), `MedGemma-27B` (fine-tuned), and `MedGemma-1.5-4B` (fine-tuned). The commercial (C) ensemble includes these same architectures complemented by `Gemini-3-Flash` in a few-shot setting.

Table 8 lists the top-performing models for the subtask 2. The five highest-scoring individual models (all utilizing the global approach) were selected for submission. As subtask 2 is independent of subtask 1, the run ordering was kept consistent to ensure that the best-performing models for both tasks were aligned within the same submission runs.

Subtask 2 - Best runs				
Run	**Model**	**Method**	**Dataset**	**Score ↓**
RUN5	Ensemble (C)	Weighted Cluster Vote + Restricted IoU	GRACE	73.14
RUN4	Ensemble (NC)	Weighted Cluster Vote + Restricted IoU	GRACE	72.92
RUN1	Qwen 3.5 4B	ICL few-shot (global)	GRACE	69.20
RUN2	Gemini 3 Flash	ICL few-shot (global)	GRACE	69.06
RUN3	MedGemma 4B	ICL few-shot (global)	GRACE	67.92
-	MedGemma 27B	ICL few-shot (global)	GRACE	66.09
-	MedGemma 27B	Finetuning (LoRA)	GRACE	65.65
-	mT5-base	Finetuning	GRACE + CasiMedicos	65.31

Table 8
Top results for subtask 2. "–" denotes that the model was not chosen for any submitted run.

5.2.3. Subtask 3

Although ensemble techniques were explored for subtask 3, they were ultimately excluded from the final submissions. Because this subtask depends heavily on the predictions from subtask 2, modifying the outputs at this stage introduced inconsistencies with earlier results, which decreased overall performance. Table 9 presents the highest-scoring models for subtask 3.

To identify the optimal combinations, downstream predictions for subtask 3 were generated using the outputs from the five selected runs in subtask 2. Although certain models, such as `GPT-5.4-mini`, significantly outperformed the rest, we implemented a sampling-without-replacement strategy, ensuring that no subtask 3 model configuration was assigned to more than one run. However, due to financial and latency constraints, `Gemini-3-Flash` was excluded from this subtask despite being the top performer. Instead, the second-best model, `GPT-5.4-mini` (few-shot), was selected and deployed across two distinct runs (RUN4 and RUN5).

Subtask 3 - Best runs				Subtask 2 runs				
Model	**Method**	**Dataset**	**Score ↓**	**RUN1**	**RUN2**	**RUN3**	**RUN4**	**RUN5**
GPT 5.4 mini	ICL few-shot	GRACE	63.29	25.65	27.83	23.52	**29.63**	30.18
GPT 5.4 mini	ICL zero-shot	-	62.72	25.87	**27.75**	23.79	29.15	30.15
Qwen 3.5 2B	ICL zero-shot	-	57.12	16.72	18.18	12.47	17.85	20.41
MedGemma 4B	ICL zero-shot	-	56.19	14.87	15.50	**13.01**	16.47	16.97
Qwen 3.5 4B	ICL few-shot	GRACE	55.68	3.91	2.68	2.61	4.45	4.32
Gemini 3 Flash	ICL few-shot	GRACE	54.38	28.22	28.42	23.94	30.37	**30.91**
MedGemma 27B	Finetuning (LoRA)	GRACE	52.83	**23.93**	22.20	16.64	24.05	26.65

Table 9
Top results for subtask 3. "–" denotes that no dataset was used. The right side shows the same metrics based on subtask 2 run predictions instead of the gold standard. Best combinations without repetition are highlighted in **bold**. Selected combinations for submission are highlighted in green.

5.3. Final Results

Finally, Table 10 displays the official final results obtained on the test set. Notably, the submitted runs maintained the exact performance ranking observed during the development phase; Run 5 emerged as the top-performing configuration in the overall ranking (securing first place), whereas Run 3 was the lowest-scoring (ranking sixth). Our submissions successfully captured the top four positions, establishing a substantial margin of 3 to 6 points ahead of the fifth-place competitor.

Team	**Pipeline**	**Overall ↓**	**Subtask1**	**Subtask2**	**Subtask3**
VicomNLP	RUN5	**59.84**	**78.98**	**72.55**	**27.99**
VicomNLP	RUN4	58.59	78.68	71.30	25.79
VicomNLP	RUN1	56.97	79.67	70.14	21.11
VicomNLP	RUN2	56.40	73.54	69.66	25.99
FranRodrigo	GPT5.4	**53.67**	**80.14**	**64.75**	**16.13**
VicomNLP	RUN3	50.69	78.32	62.10	11.64
Baseline	Encoder + Classifier	47.90	73.48	62.63	7.58
UMUTeam	Pretrained medical Qwen-4b	**44.29**	76.92	**54.86**	**1.10**
UMUTeam	Medical pretrained + pretrained LLM	44.25	**77.77**	54.02	0.95
UMUTeam	Pretrained Qwen-4b	44.15	**77.77**	53.94	0.75
FranRodrigo	XLM-R-base + sentence-aligned splitter v3	25.93	69.37	8.42	0.00

Table 10
Final F1-Score ordered by overall. Our team is denoted as VicomNLP. Best scores for each team are highlighted in **bold**, and the absolute best score is underlined.

On the other hand, analyzing the performance of each subtask individually reveals that, with the exception of subtask 1 (where scores fell significantly below expectations, dropping from 0.90 to 0.78) our approaches achieved the top four results in both subtasks 2 and 3. Furthermore, the metrics obtained for these two subtasks remained remarkably consistent with those observed during the dev set evaluation. Overall, our results demonstrate that all submitted systems outperform the baseline established by the organizers, both in terms of the overall evaluation score and across each individual subtask.

6. Discussion

The automatic analysis of argumentative evidence and logical relations in clinical texts is critical for the adoption of medical NLP systems [1, 17]. The GRACE 2026 shared task presents a challenging benchmark that combines sentence relevance detection, argument component extraction, and relation prediction in Spanish medical reasoning.

Our results show that clinical case reasoning through AM benefits from careful alignment between model architecture, learning paradigm, and task formulation. In sentence-level relevance classification, fine-tuned encoder-only and encoder-decoder models achieved the strongest results. In contrast,

decoder-only models became more competitive in extraction-oriented subtasks, where the output structure is variable and more sensitive to prompt design. This finding is consistent with prior work showing that relation prediction and structured extraction can be reframed effectively as generation or sequence-to-sequence problems [14, 15], but that performance depends heavily on how the task is defined.

Data augmentation was beneficial when used for supervised fine-tuning of smaller or domain-adapted models. In contrast, in-context learning was more sensitive to the exact choice of demonstrations, and adding more examples did not necessarily improve performance when the examples came from a related but not identical distribution.

Our findings also suggest that translated annotations can be a practical way to increase training data for Spanish clinical NLP. In this shared task, however, the Spanish test did not provide clear evidence of cross-lingual transfer, since any multilingual benefit should be evaluated under the same target-language conditions [18].

Notably, prompting strategy had a strong effect across decoder-only models. Strict prompts with clear task instructions and complete examples outperformed simpler few-shot settings, and in entity extraction they even surpassed the fine-tuned baseline. However, relation prediction remained difficult even under the strongest prompting configuration. In this case, pairwise relation prediction from explicit candidate generation and structured inference was more effective than global generation alone.

Post-processing and ensembling were specially valuable in the entity extraction and relation classification subtasks. In particular, refining candidate entities before relation prediction improved the quality of the downstream inputs and led to significant gains in relation accuracy. This supports the view that end-to-end performance in AM is often determined not only by the base model, but also by how outputs are filtered, paired, and aggregated across stages. For high-stakes medical applications, this staged design aids inspection and aligns better with explainable AI goals in clinical settings.

Overall, the results indicate that smaller open-source models can remain competitive when they are combined with task-specific data preparation, strong prompting, and careful aggregation of outputs.

7. Conclusions and Future Work

This work presented the participation of the **VicomNLP** team in Track 2 of the GRACE 2026 shared task, addressing clinical argument mining through a combination of discriminative and generative language models. Across the three subtasks, our results show that no single modeling paradigm consistently dominates all stages of the pipeline.

Our findings suggest that clinical argument mining should be approached as a multi-stage problem in which different subtasks benefit from different model architectures and learning paradigms. Rather than relying on a single model family, hybrid pipelines combining discriminative and generative approaches constitute a promising direction for robust clinical reasoning. Furthermore, ensembling and post-processing strategies contributed to mitigating error propagation across the pipeline and improving downstream performance.

Beyond the competitive results obtained in the shared task, this work also introduces an extended multilingual version of the CasiMedicos-Arg dataset, providing an additional resource for future research on clinical argument mining. The resulting dataset enabled the assessment of generalization across clinical data sources and offers a foundation for further investigations into multilingual transfer and low-resource clinical NLP.

Future work will focus on addressing the main limitations identified in this study. First, we plan to investigate end-to-end and multi-task learning approaches that jointly model sentence relevance, argument component extraction, and relation prediction. Such architectures could alleviate the error propagation effects observed in the current pipeline-based framework and improve global consistency across subtasks.

Acknowledgements

This work is funded by the Basque Government under the HAZITEK 2024 programme (grant number ZE- 2024/00030)

Declaration on Generative AI

Generative AI tools (ChatGPT and Gemini) were used to improve the writing style, assist with abstract drafting, and check grammar and spelling. These tools were not used to generate scientific conclusions, interpret results, or make research decisions. After using these tools, the authors reviewed and edited the content as necessary and take full responsibility for the content of this publication.

References

[1] T. Mayer, E. Cabrio, S. Villata, Evidence type classification in randomized controlled trials, in: Proceedings of the 5th Workshop on Argument Mining, 2018, pp. 29–34.

[2] I. De la Iglesia, Overview of GRACE at IberLEF 2026: Granular Recognition of Argumentative Clinical Evidence, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2026), CEUR Workshop Proceedings, 2026.

[3] A. Bonet-Jover, L. Chiruzzo, J. González Barba, 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 Workshop Proceedings, 2026.

[4] 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.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[6] E. Sviridova, A. Yeginbergen, A. Estarrona, E. Cabrio, S. Villata, R. Agerri, CasiMedicos-Arg: A Medical Question Answering Dataset Annotated with Explanatory Argumentative Structures, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 18463–18475. URL: https://aclanthology.org/2024.emnlp-main.1026.

[7] Vicomtech, VicomNLP team in Track 2 of the GRACE 2026 shared task: Technical resources and information., 2026. URL: https://github.com/Vicomtech/grace2026, accessed: 2026-06-22.

[8] I. Habernal, I. Gurevych, Argumentation mining in user-generated web discourse, Computational Linguistics 43 (2017) 125–179. URL: https://aclanthology.org/J17-1004/. doi:10.1162/COLI_a_00276.

[9] S. Teufel, A. Siddharthan, C. Batchelor, Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics, in: P. Koehn, R. Mihalcea (Eds.), Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2009, pp. 1493–1502. URL: https://aclanthology.org/D09-1155/.

[10] R. Rinott, L. Dankin, C. Alzate Perez, M. M. Khapra, E. Aharoni, N. Slonim, Show me your evidence - an automatic method for context dependent evidence detection, 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. 440–450. URL: https://aclanthology.org/D15-1050/. doi:10.18653/v1/D15-1050.

[11] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, N. Slonim, Stance classification of context-dependent claims, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 251–261. URL: https://aclanthology.org/E17-1024/.

[12] 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.

[13] H. V. Nguyen, D. J. Litman, Argument mining for improving the automated scoring of persuasive essays, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[14] A. de Wynter, T. Yuan, "i'd like to have an argument, please": Argumentative reasoning in large language models, in: Computational Models of Argument, IOS Press, 2024, pp. 73–84.

[15] M. Guida, J. Otmakhova, E. Hovy, L. Frermann, Llms for argument mining: Detection, extraction, and relationship classification of pre-defined arguments in online comments, in: Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association, 2025, pp. 176–191.

[16] 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. URL: https://www.sciencedirect.com/science/article/pii/S0933365721000919. doi:https://doi.org/10.1016/j.artmed.2021.102098.

[17] T. Mayer, E. Cabrio, S. Villata, Transformer-based argument mining for healthcare applications, in: ECAI 2020-24th European Conference on Artificial Intelligence, 2020.

[18] A. Yeginbergen, R. Agerri, Crosslingual argument mining in the medical domain, Procesamiento del Lenguaje Natural (2024) 297–312.

[19] M. Urruela, S. Martín, I. De la Iglesia, A. Barrena, Medical argument mining: Exploitation of scarce data using nli systems, Procesamiento del Lenguaje Natural 75 (2025) 251–262.

[20] R. Agerri, I. Alonso, A. Atutxa, A. Berrondo, A. Estarrona, I. García-Ferrero, I. Goenaga, K. Gojenola, M. Oronoz, I. Perez-Tejedor, G. Rigau, A. Yeginbergenova, Hitz@antidote: Argumentation-driven explainable artificial intelligence for digital medicine, in: SEPLN 2023: 39th International Conference of the Spanish Society for Natural Language Processing., 2023.

[21] I. Alonso, M. Oronoz, R. Agerri, Medexpqa: Multilingual benchmarking of large language models for medical question answering, Artificial Intelligence in Medicine (2024) 102938. URL: https://www.sciencedirect.com/science/article/pii/S0933365724001805. doi:https://doi.org/10.1016/j.artmed.2024.102938.

[22] I. de la Iglesia, A. Atutxa, K. Gojenola, A. Barrena, Eriberta: A bilingual pre-trained language model for clinical natural language processing, 2023. URL: https://arxiv.org/abs/2306.07373. arXiv:2306.07373.

[23] I. García-Ferrero, R. Agerri, A. Atutxa Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, A. Zaninello, MedMT5: An open-source multilingual text-to-text LLM for the medical domain, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 11165–11177. URL: https://aclanthology.org/2024.lrec-main.974/.

[24] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, CoRR abs/2010.11934 (2020). URL: https://arxiv.org/abs/2010.11934. arXiv:2010.11934.

[25] J. Cabessa, H. Hernault, U. Mushtaq, Argument mining in BioMedicine: Zero-shot, in-context learning and fine-tuning with LLMs, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 122–131. URL: https://aclanthology.org/2024.clicit-1.16/.

[26] P. Benito, M. Isla-Jover, P. González-Castro, P. J. Fernández Esparcia, M. Carpio, I. Blay-Simón, P. Gutiérrez-Bedia, M. J. Lapastora, B. Carratalá, C. Carazo-Casas, Gpt-4o and openai o1 perfor-

mance on the 2024 spanish competitive medical specialty access examination: Cross-sectional quantitative evaluation study, JMIR Med Educ 12 (2026) e75452. URL: https://mededu.jmir.org/2026/1/e75452. doi:10.2196/75452.
[27] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[28] 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, 2020. URL: https://arxiv.org/abs/1911.02116. arXiv:1911.02116.
[29] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021. arXiv:2109.03570.
[30] A. Otegi, A. Agirre, J. A. Campos, A. Soroa, E. Agirre, Conversational question answering in low resource scenarios: A dataset and case study for Basque, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 436–442. URL: https://aclanthology.org/2020.lrec-1.55/.
[31] OpenAI, GPT-5.4 System Card, Technical Report, OpenAI, 2026. URL: https://arxiv.org/abs/2601.03267, technical report.
[32] Google, Gemini 3 pro large language model, 2026. URL: https://gemini.google.com.
[33] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.
[34] Google DeepMind, Gemma: Google deepmind open models, 2026. URL: https://deepmind.google/models/gemma/.
[35] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, et al., Medgemma technical report, arXiv preprint arXiv:2507.05201 (2025).
[36] A. Sellergren, C. Gao, F. Mahvar, T. Kohlberger, F. Jamil, M. Traverse, A. Tono, B. Sadjad, L. Yang, C. Lau, et al., Medgemma 1.5 technical report, arXiv preprint arXiv:2604.05081 (2026).
[37] J.-P. Corbeil, A. Dada, J.-M. Attendu, A. Ben Abacha, A. Sordoni, L. Caccia, F. Beaulieu, T. Lin, J. Kleesiek, P. Vozila, A modular approach for clinical SLMs driven by synthetic data with pre-instruction tuning, model merging, and clinical-tasks alignment, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 19352–19374. URL: https://aclanthology.org/2025.acl-long.950/. doi:10.18653/v1/2025.acl-long.950.

A. Online Resources

All detailed technical information—including all prompts used with all LLM approaches, the code used for dataset formatting and ensembling, and the complete CasiMedicos-Arg dataset with all relations in all available languages—is available in our GitHub repository: https://github.com/Vicomtech/grace2026. As shown below, for each of the subtasks, two different prompts have been developed: “Sentence-by-Sentence Annotation Prompt” and “All-in-One Inference Annotation Prompt”. Both prompts were evaluated for every model and task, but only the best-performing results are reported in our study.

Subtask 1

Sentence-by-Sentence Annotation Prompt

Eres un experto médico. Tu tarea es la **Detección de Oraciones de Evidencia**.
Analiza la siguiente oración dentro del caso clínico y determina si es `"relevant"` o `"not-relevant"` para apoyar o refutar las opciones de respuesta.

Caso Clínico:
`{context}`

Opciones:
`{choices_str}`

Oración a evaluar:
`{sentence}`

Responde únicamente con `"relevant"` o `"not-relevant"`.

All-in-One Inference Annotation Prompt

Eres un experto clínico. Tu tarea es evaluar una lista numerada de oraciones de un caso clínico. Debes determinar si cada oración contiene evidencia médica **RELEVANTE** (síntomas, historial, pruebas) o si es **IRRELEVANTE** (texto de relleno, saludos, la pregunta final).

Restricciones obligatorias:

- `"sentence_relevancy"` debe tener exactamente una etiqueta por cada oración recibida, y usa solo `"relevant"` o `"not-relevant"`.
- Devuelve únicamente JSON válido.

Formato obligatorio de salida:

```
{
    "sentence_relevancy": [
        "relevant",
        "not-relevant"
    ]
}
```

Subtask 2

Sentence-by-Sentence Annotation Prompt

Eres un experto en razonamiento clínico y extracción de información. Tu tarea es identificar y extraer fragmentos de texto exactos que representen 'Premises' o 'Claims' dentro de una oración específica, utilizando el caso clínico completo solo como contexto de fondo.

Definiciones:

- **Premise:** Evidencia clínica objetiva (hechos, mediciones, síntomas, observaciones o antecedentes médicos del paciente).
- **Claim:** Opciones de respuesta o hipótesis (diagnósticos candidatos, propuestas de tratamiento o pronósticos).

Reglas de extracción:

1. Evalúa **ÚNICAMENTE** la oración a analizar.
2. El fragmento extraído debe ser una copia **EXACTA** (respetando mayúsculas, puntuación y espacios) de cómo aparece en la oración. Una 'Claim' puede abarcar la oración completa.
3. Si la oración no contiene ninguna 'Premise' ni 'Claim' (ej. texto de relleno o preguntas genéricas), debes devolver un array vacío: `[]`
4. Responde **ÚNICAMENTE** con un array JSON válido, sin introducciones ni explicaciones previas.
5. Formato de salida: `[{"text": fragmento de texto, "type": Premise/Claim}]`

Contexto:
`{raw_text}` **Oración a analizar:**

`{segment_text}`

All-in-One Inference Annotation Prompt

Eres un experto en razonamiento clínico y extracción de información. Tu tarea es identificar y extraer fragmentos de texto exactos que representen 'Premises' o 'Claims' dentro del caso clínico proporcionado.

Definiciones:

- **Premise:** Evidencia clínica objetiva (hechos, mediciones, síntomas, observaciones).
- **Claim:** Opciones de respuesta o hipótesis.

Reglas de extracción:

1. El fragmento extraído debe ser una copia **EXACTA** del texto original.
2. Asigna a cada Premise un `'local_id'` correlativo (`p1`, `p2`...) y el `'source_index'` de la oración donde aparece.
3. Extrae las Claims con su respectivo ID.

Formato obligatorio de salida:

```
{
    "premises": [
        {
            "local_id": "p1",
            "source_index": 0,
            "text": "fragmento exacto minimo"
        }
    ],
    "claims": [
        {
            "id": "1",
            "text": "texto exacto de la opcion"
        }
    ]
}
```

Subtask 3

Sentence-by-Sentence Annotation Prompt

Eres un experto clínico. Tu tarea es evaluar la relación argumentativa entre una evidencia (Premise) y una opción candidata (Claim) basándote en el caso clínico proporcionado.

Las posibles relaciones entre 'premise' y 'claim' son:

- **Support:** Si la premise apoya, confirma o es consistente con la claim.
- **Attack:** Si la premise contradice, refuta, descarta o hace improbable la claim.
- **Nothing:** Si la premise y la claim no tienen relación.

Responde únicamente con `'Support'`, `'Attack'` o `'Nothing'`.

Caso Clínico:
`{raw_text}`

Evidencia (Premise):
`{premise_text}`

Opción (Claim):
`{claim_text}`

All-in-One Inference Annotation Prompt

Eres un razonador clínico. Se te dará una PREMISA (un hecho del paciente) y un CLAIM (una posible respuesta/diagnóstico). Tu tarea es determinar la relación argumentativa entre ellos:

- `'Support'`: La premisa apoya, confirma o es consistente con el claim.
- `'Attack'`: La premisa contradice, descarta o hace improbable el claim.

Restricciones obligatorias:

- Cada `"premise_id"` debe ser el ID proporcionado para la Premise analizada.
- Cada `"claim_id"` debe corresponder a la opción recibida.
- Usa solo `"Support"` o `"Attack"`.
- Devuelve únicamente JSON válido.

Formato obligatorio de salida:

```
{
    "relations": [
        {
            "premise_id": "p1",
            "claim_id": "3",
            "relation_type": "Support"
        }
    ]
}
```

Global (all subtask in one inference)

Inferencia Unificada (End-to-End)

All-in-One Inference Annotation Prompt

Eres un experto médico en razonamiento clínico MIR y extracción de argumentos clínicos.
Tu tarea es resolver tres subtareas de razonamiento clínico en una única inferencia.

Recibirás:

1. Un caso clínico completo.
2. Una lista de oraciones del contexto clínico, cada una con un índice.
3. Una lista de opciones de respuesta, cada una con un `id`. A las cuales denominaremos *Claims*.

IMPORTANTE - Tu tarea consiste en:

1. Clasificar cada oración del contexto como `"relevant"` o `"not-relevant"`.
2. Extraer únicamente *Premises* mínimas desde las oraciones relevantes.
3. Relacionar las *Premises* extraídas con las *Claims*/opciones usando el `id` de la opción.

Subtarea 1 — Evidence Sentence Detection:
Clasifica cada oración del contexto clínico como:

- `"relevant"`: contiene evidencia clínica útil para apoyar o refutar alguna opción.
- `"not-relevant"`: no aporta evidencia clínica útil para decidir entre las opciones.

Reglas para relevancia:

- Las oraciones con síntomas, signos, antecedentes, resultados de pruebas, evolución temporal, factores de riesgo, negaciones clínicas o hallazgos clínicos suelen ser `"relevant"`.
- Las preguntas genéricas como "¿Cuál es el diagnóstico más probable?", "¿Cuál es el tratamiento indicado?" o frases que solo introducen la pregunta suelen ser `"not-relevant"`.
- Una oración no debe ser `"relevant"` solo por introducir la pregunta.

Subtarea 2 — Minimal Premise Span Detection:
Extrae fragmentos exactos de texto que sean *Premises*. Una *Premise* es una unidad mínima de evidencia clínica objetiva del caso:

- síntoma, signo, antecedente
- edad o sexo si son clínicamente relevantes
- duración o evolución temporal
- resultado de prueba, hallazgo de exploración
- factor de riesgo, ausencia o negación clínica relevante
- dato que apoye o refute una opción

Reglas estrictas para Premises:

1. Extrae *Premises* solo desde oraciones clasificadas como `"relevant"`.
2. El texto debe ser una copia **EXACTA** de un fragmento de la oración original.
3. No reformules. No inventes.
4. No incluyas offsets.
5. No extraigas texto de las opciones como *Premise*.
6. No extraigas la pregunta final como *Premise*.
7. No devuelvas una oración completa si contiene varias unidades clínicas.
8. Si una oración contiene varias evidencias clínicas, divide la oración en varias *Premises* mínimas.
9. Cada *Premise* debe expresar una sola unidad clínica.
10. Prefiere el span más corto que conserve el significado clínico.
11. Mantén modificadores clínicamente importantes, por ejemplo duración, localización, severidad o resultado de prueba.
12. No incluyas introducciones, conectores ni relleno, solo la evidencia clínica concreta.

Subtarea 3 — Argumentative Relation Detection:
Relaciona cada *Premise* con una o varias *Claims* cuando exista una relación clara. Las *Claims* son las opciones de respuesta recibidas. Debes referirte a ellas usando su `id`.

Tipos de relación:

- `"Support"`: la *Premise* apoya, favorece, confirma o es consistente con esa opción.
- `"Attack"`: la *Premise* contradice, descarta, debilita o hace improbable esa opción.

No incluyas relaciones dudosas.

Formato obligatorio de salida:

```
{
    "sentence_relevancy": [
        "relevant",
        "not-relevant"
    ],
    "premises": [
        {
            "local_id": "p1",
            "source_index": 0,
            "text": "fragmento exacto minimo"
        }
    ],
    "relations": [
        {
            "premise_id": "p1",
            "claim_id": "3",
            "relation_type": "Support"
        }
    ]
}
```

Restricciones obligatorias:

- `"sentence_relevancy"` debe tener exactamente una etiqueta por cada oración recibida.
- Usa solo `"relevant"` o `"not-relevant"`.
- Cada `"source_index"` debe ser el índice de la oración de la que sale la *Premise*.
- Cada `"text"` debe aparecer literalmente dentro de la oración indicada.
- Cada `"text"` debe ser el menor fragmento clínicamente suficiente, no la oración completa.
- Cada `"premise_id"` debe existir en `"premises"`.
- Cada `"claim_id"` debe corresponder a una opción recibida.
- Usa solo `"Support"` o `"Attack"`.
- Devuelve únicamente JSON válido.
- Limita tu bloque de pensamiento a un máximo de 3 párrafos cortos antes de dar la respuesta final.