CheckMates at CheckThat! 2026: Combining Ranking Models and LLM Judges for Numerical Reasoning Trace Verification

CheckThat! at CLEF 2026

Hana Ghanem[*,†], Maria Bassem[*,†], Marwan Torki and Nagwa El-Makky

Faculty of Engineering, Alexandria University

Abstract

This paper presents our approach for multilingual numerical fact-checking in the CheckThat! 2026 Task 2. The task involves ranking candidate reasoning traces for English, Arabic, and Spanish claims to predict the final claim verdict. We explored two main directions: a trained re-ranker model that scores reasoning traces, and a rubric-based LLM evaluator that assesses the quality and consistency of these traces. Based on our experiments, we propose a hybrid method that combines both signals through score fusion for final trace re-ranking and verdict extraction. On the official leaderboard, our system achieved strong multilingual performance, ranking first in both Arabic and Spanish and sixth in English. The system obtained average scores of 0.6043 for Arabic, 0.4847 for Spanish, and 0.4200 for English, showing that combining the re-ranker with rubric-guided LLM evaluation is effective for numerical claim verification across multiple languages.

Keywords

Numerical Fact-Checking, Reasoning Trace Ranking, Multilingual NLP, LLM-as-a-Judge,

1. Introduction

Numerical fact-checking aims to verify claims involving quantities, comparisons, statistics, and temporal information. Unlike general fact verification, it requires models not only to identify relevant evidence, but also to understand numerical relationships correctly and determine whether the reasoning leading to a verdict is reliable. The challenge becomes more difficult in multilingual settings, where differences in data distribution, linguistic expression, and annotation quality can affect both reasoning and final prediction. [1]

In this work, we study multilingual numerical fact-checking through the task of reasoning-trace ranking. This trace re-ranking task introduces additional challenges, as candidate traces may contain incorrect reasoning, inconsistent verdicts, or signals that do not fully agree with the gold claim label.

Our initial data analysis shows important differences in the available multilingual data. The label distribution is imbalanced across languages, with the *False* class occurring more frequently, while the *Conflicting* class is absent from the Arabic data. We also observe cases where trace verdicts are inconsistent with the gold labels, including semantically similar traces associated with different verdict outcomes. Such patterns can introduce noisy supervision and make ranking in reasoning sequences more difficult.

To address this problem, we investigate supervised re-ranking models trained under different learning objectives, including pointwise and pairwise ranking [2, 3]. We further study the effects of the input context, verdict supervision, evidence inclusion, and multilingual training. We observed that trained re-rankers still struggled to recognize complex reasoning errors, particularly those involving numerical data and evidence grounding, so we also explore large language models as rubric-based judges for evaluating

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

[*]Corresponding author.

[†]These authors contributed equally.

es-hanawalid2025@alexu.edu.eg (H. Ghanem); es-MariaBasem2025@alexu.edu.eg (M. Bassem); mtorki@alexu.edu.eg (M. Torki); nagwamakky@alexu.edu.eg (N. El-Makky)

trace quality. Finally, we examine whether combining learned ranking signals with reasoning-oriented evaluation can provide a stronger basis for trace selection and verdict prediction.

The main contributions of this work are:

- Analyzing multilingual reasoning-trace data with respect to label imbalance and inconsistencies.
- A comparison between supervised reasoning-trace ranking approaches and a study of the effects of learning objective, context, evidence, and multilingual training.
- Exploring large language models (LLMs) as rubric-guided evaluators of reasoning quality and evidence grounding.
- Investigating a hybrid approach that combines supervised re-ranking with LLM-based evaluation for multilingual numerical fact checking.

2. Related Work

2.1. Numerical Fact-Checking and Multilingual Verification

Numerical fact-checking requires models to reason over quantities, comparisons, and temporal information rather than rely only on textual relevance. QuanTemp established this problem through a real-world numerical claim verification benchmark [1]. The CLEF 2025 CheckThat! Task 3 extended this setting to English, Spanish, and Arabic, highlighting challenges in evidence selection, multilingual transfer, and label imbalance [4]. The CLEF 2026 CheckThat! Task 2 further introduces reasoning-trace ranking, where systems must select useful generated traces before predicting the final verdict [5, 3].

Previous CLEF 2025 systems explored fine-tuned open-source LLMs, evidence processing, and hybrid methods for numerical claim verification. LIS used question-guided retrieval followed by LoRA-based verdict prediction, while TIFIN investigated evidence handling, multilingual imbalance, and context configuration [6, 7]. CornellNLP and NGU_Research further explored hybrid LLM pipelines for final claim classification [8, 9]. In contrast, our work focuses on the additional challenge of evaluating and ranking candidate reasoning traces.

2.2. Reasoning Trace Generation and Selection

Generating multiple reasoning paths can improve the final prediction, as shown by self-consistency [10]. More closely related to our task, *Think Right, Not More* applies verifier-based Best-of-N selection to numerical claim verification and identifies reasoning drift as an important failure mode [2].

This is especially relevant because generated reasoning traces may appear plausible while still being incorrect. Prior work has shown that chain-of-thought reasoning can be unfaithful, rationalize incorrect outputs, or become unnecessarily long, causing the model to drift away from the evidence [11]. These observations motivate the need to evaluate and filter the reasoning traces before making a final prediction.

2.3. Learning-to-Rank and LLM-based Evaluation

Our approach combines supervised trace re-ranking with rubric-based LLM evaluation. Learning-to-rank studies motivate our comparison of pointwise and pairwise scoring strategies, while rubric-based evaluators motivate assessing reasoning quality using explicit criteria [12].

We use these directions together: the supervised re-ranker learns task-aligned trace preferences, while the LLM judge provides an additional signal for logical consistency and evidence grounding in multilingual numerical verification.

3. Methodology

To address these challenges, we propose a hybrid approach for multilingual reasoning-trace ranking and verdict prediction. Given a claim, its evidence, and a set of reasoning traces, our system assigns a

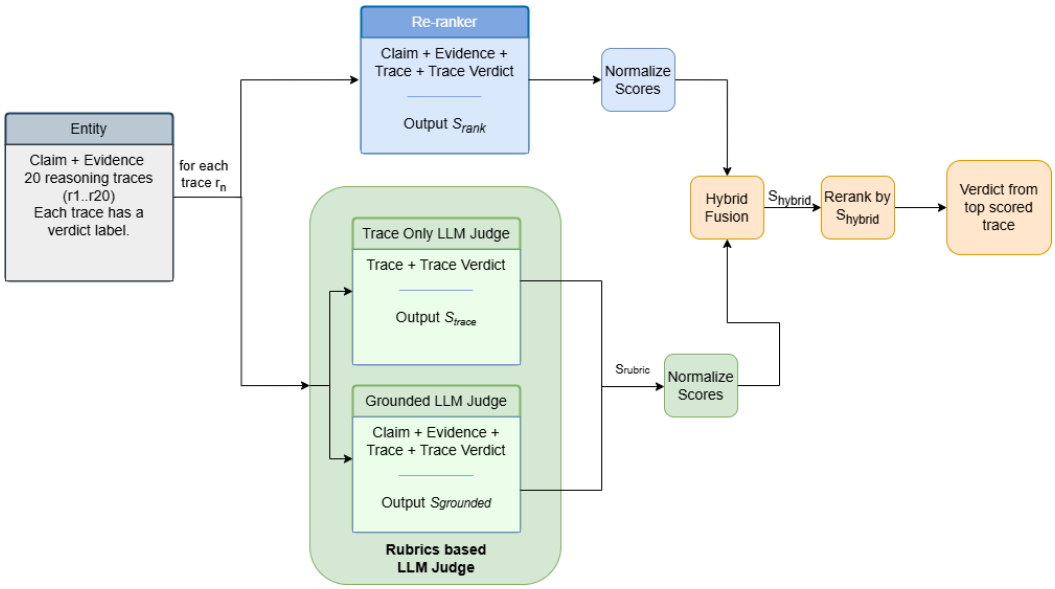

Figure 1: Overview of the proposed hybrid reasoning-trace ranking system. Traces are independently scored by a re-ranker and a rubric-guided LLM evaluator. Their normalized scores are fused to produce the final ranking, from which the predicted verdict is selected.

score to each reasoning trace and ranks the traces according to their usefulness for reaching the correct verdict, as shown in Figure 1.

Our system consists of multiple components. The first component is a supervised **Qwen-based pointwise re-ranker**, which learns from the provided training dataset to assign a utility score to each candidate reasoning trace. The second component is a **rubric-guided large language model evaluator**, which assesses the quality of each trace according to explicit criteria related to reasoning quality and evidence grounding. Finally, these two scoring signals are combined in a **hybrid ranking framework** to produce the final ordering of the reasoning traces and select the predicted verdict.

In the following subsections, we describe each scoring component independently, including the rubric design used by the LLM evaluator. We then explain how their scores are normalized and combined in the final hybrid approach, followed by the final re-ranking and verdict-selection procedure.

3.1. Pointwise Qwen-Based Re-ranker

The first component is a supervised pointwise re-ranker based on **Qwen/Qwen2.5-3B-Instruct**[13]. The model was adapted using Low-Rank Adaptation (LoRA) and a custom scoring head[14]. In the pointwise setting, each candidate reasoning trace is treated independently and assigned a scalar score estimating whether it is useful for obtaining the gold verdict.

For each reasoning trace, the model input is constructed from the claim, the available evidence, the candidate trace, and its associated verdict.

The input sequence is tokenized and truncated to a maximum length of 1024 tokens. The encoded representation is passed through a scoring head to produce a single unbounded scalar logit:

$$z_i = f_\theta(c, e, r_i, v_i). \tag{1}$$

The pointwise model is trained using binary cross-entropy with logits:

$$\mathcal{L}_{\text{pointwise}} = -\frac{1}{M}\sum_{i=1}^{M}\left[y_i \log \sigma(z_i) + (1-y_i)\log(1-\sigma(z_i))\right], \tag{2}$$

where M is the number of traces in the training batch and $\sigma(\cdot)$ denotes the sigmoid function. In implementation, this objective is computed using **BCEWithLogitsLoss**, which applies the sigmoid transformation internally in a numerically stable manner.

At inference time, the raw model's logits are used as ranking scores. For each claim, all candidate reasoning traces are independently scored and then sorted in descending order. This formulation provides a simple and stable supervised ranking signal. In our experiments, the pointwise objective was compared with alternative ranking objectives and produced the strongest overall results among the tested supervised re-ranking configurations. The initial results obtained by the pointwise re-ranker with maximum sequence length of 1024, evidence and verdict are reported in Table 1. In all results tables, the Average score is computed as the mean of Macro-F1 and Recall@5.

Table 1
Re-ranker results reported on the Test Dataset under pointwise reranking and 1024 maximum sequence length.

Language	Avg. Score	Macro F1	Recall@5
Arabic	0.385	0.510	0.260
English	0.387	0.576	0.201
Spanish	0.392	0.560	0.224

3.2. Rubric-Guided LLM Evaluation

Although the trained re-ranker provides a task-aligned scoring signal, numerical fact verification may require recognizing errors that are difficult to learn from noisy or limited supervision alone. To provide an additional reasoning-oriented signal, we evaluate candidate traces using an LLM prompted as a rubric-guided judge.

We evaluated multiple candidate LLMs for this role and selected `Qwen/Qwen3.5-27B` based on validation performance[15]. Unlike the supervised re-ranker, the judge is not trained on the trace-ranking labels. Instead, it receives the candidate trace together with the relevant context and scores it according to reasoning-quality criteria.

This evaluation approach uses two rubric groups. The first is a *trace-only rubric*, which evaluates the quality of the reasoning trace by itself. The second is a *grounded rubric*, which evaluates the relationship between the claim, the evidence, and the reasoning trace.

Let q_i^{trace} denote the trace-only rubric score and q_i^{ground} denote the claim–evidence–trace grounding score for candidate trace r_i. The final rubric score is computed as:

$$s_i^{\text{rubric}} = \alpha q_i^{\text{trace}} + (1 - \alpha) q_i^{\text{ground}}, \tag{3}$$

where α specifies the weight assigned to intrinsic trace quality. In our selected configuration, after multiple tests, the trace-only component receives a weight of 0.4 and the grounded component receives a weight of 0.6.

The rubric-based judge provides a different signal from the trained re-ranker. The results are reported in Table 2. It was observed that, while the supervised model is trained to match the available trace supervision, the LLM judge explicitly assesses whether a trace is logically sound and grounded in the evidence. We therefore use its output as a complementary score in a hybrid system rather than as a standalone replacement for supervised re-ranking.

Table 2
Rubric-Guided LLM results reported on the Test Set.

Language	Avg. Score	Macro F1	Recall@5
Arabic	0.336	0.452	0.220
English	0.403	0.570	0.236
Spanish	0.364	0.511	0.217

3.3. Rubric Design

The first rubric is a *trace-only rubric*, which evaluates the internal quality of the reasoning trace, such as logical coherence, clarity, consistency, and whether the trace verdict follows from the reasoning. The second is a *grounded rubric*, which evaluates the relationship between the claim, evidence, and reasoning trace, including whether the trace is supported by evidence and correctly interprets numerical or temporal information. These rubrics are organized into two sets of evaluation questions, shown in Table 3. The prompts used to generate the scores can be found in Appendix A.

3.4. Hybrid Reasoning-Trace Ranking Framework

After describing the two scoring components independently, we combine them in the final hybrid reasoning-trace ranking framework shown in Figure 1. For each instance, let c denote the claim, e denote the available evidence, and $\mathcal{R} = \{r_1, r_2, \ldots, r_N\}$ denote the set of candidate reasoning traces, where each trace r_i is associated with a candidate verdict v_i. In this setting, $N = 20$. Our goal is to produce a ranking over the candidate traces and then select the final verdict from the highest-ranked trace.

Each trace is evaluated by the two scoring components described above. First, the re-ranker produces a learned utility score:

$$s_i^{\text{rank}} = f_\theta(c, e, r_i, v_i), \tag{4}$$

where f_θ is the trained Qwen-based scoring model. Second, the rubric-based LLM evaluator produces a reasoning-quality score:

$$s_i^{\text{rubric}} = g(c, e, r_i, v_i), \tag{5}$$

where g denotes the prompted LLM judge.

Since the two components produce scores on different scales, the re-ranker outputs are normalized between 0 and 10 before fusion. Let $\widetilde{s}_i^{\text{rank}}$ and $\widetilde{s}_i^{\text{rubric}}$ denote the normalized scores. The final hybrid score is computed as:

$$s_i^{\text{hybrid}} = \lambda \widetilde{s}_i^{\text{rank}} + (1 - \lambda)\widetilde{s}_i^{\text{rubric}}, \tag{6}$$

where λ controls the contribution of each of the re-ranker and the rubric-based evaluator, and $\lambda = 0.4$ in our setup. The candidate traces are sorted in descending order according to s_i^{hybrid}. The final claim verdict is selected as the verdict associated with the top-ranked reasoning trace.

3.5. Final Re-ranking and Verdict Selection

For each claim, the supervised re-ranker and the rubric-based judge independently score all candidate reasoning traces. We normalize the two sets of scores and combine them using the weighted fusion function defined in the previous subsection. The traces are then reordered according to the resulting hybrid scores, while maintaining alignment between each trace and its associated verdict. Our final

Rubric Group	Evaluation Criteria
Trace-only	1. Trace Understanding
	2. Numbers and calculations
	3. Internal Consistency
	4. Overthinking
	5. Specificity
Grounded	1. Claim Understanding
	2. Evidence Use
	3. Claim Evidence Consistency
	4. Verdict Justification
	5. Unsupported Assumptions

Table 3
Rubric dimensions used by the LLM judge to evaluate candidate reasoning traces.

verdict-selection strategy follows a Best-of-N setup. Given the ranked candidate set, the verdict associated with the highest-ranked trace is selected as the prediction for the claim.

4. Experiments and Results

4.1. Datasets and Metrics

We evaluated our system on the multilingual numerical fact-checking datasets provided in the CLEF CheckThat! Task 2 shared task. The experiments covered English, Spanish, and Arabic claim-evidence pairs together with generated reasoning traces and associated verdict labels as shown in Table 4.

Table 4
Dataset statistics for Task 2.

Language	Number of Claims
English	8,000
Spanish	2,808
Arabic	3,260

Due to the multilingual nature of the benchmark, the datasets exhibit varying levels of label imbalance and reasoning-trace quality across languages.

For ranking evaluation, we report Recall@k, which measures whether useful reasoning traces are ranked within the top-k positions. For classification evaluation, we report Precision, Recall, and Macro F1 across the `True`, `False`, and `Conflicting` labels. In addition, we report the average score used by the organizers for the leaderboard. It is defined as:

$$\text{Average Score} = \frac{\text{Macro F1} + \text{Recall@5}}{2} \tag{7}$$

For the Overall row results in the upcoming tables, predictions from all languages are concatenated and the official metrics are computed on the pooled set. Therefore, the Overall Macro F1 is not the arithmetic mean of the per-language Macro F1 values.

We evaluated several ranking configurations for multilingual numerical fact-checking across English, Spanish, and Arabic datasets.

Our experiments were organized into two main groups:

- Impact analysis of evidence.
- Data and language configuration analysis.

4.2. Training Setup for Pointwise Ranking Objective

The primary experiments used the `Qwen2.5-3B-Instruct` model as the supervised reasoning-trace re-ranker. The model was trained using LoRA-based parameter-efficient fine-tuning. For the main pointwise ranking experiments, we used a maximum sequence length of 1024 tokens. The training configuration is summarized below:

- Base model: `Qwen2.5-3B-Instruct`
- Maximum sequence length: 1024
- Number of epochs: 2
- Learning rate: 5×10^{-5}
- Effective batch size: 256

For parameter-efficient fine-tuning, we used LoRA with:

- LoRA rank (r): 8

- LoRA alpha (α) : 16

The best checkpoint was selected using validation loss.

4.3. Rubric-Guided LLM Configuration

For LLM-based evaluation, we used `Qwen/Qwen3.5-27B` with a temperature of `0.7` and a top-p value of `0.95`. The model was prompted to return a numerical score for each reasoning trace. We parsed the numerical value from the model output; if no valid number was returned, the prompt was retried until a valid score was obtained.

4.4. Data and Language Configuration Analysis

We compared monolingual and multilingual training settings in order to analyze the effect of multilingual supervision on reasoning-trace ranking performance. Those experiments were done on a pairwise ranking setting and maximum sequence length of 512.

Table 5
Comparison between monolingual and multilingual training settings.

Training Setting	Lang.	Macro F1	Recall@5	Avg. Score
Monolingual	EN	0.4901	0.2807	0.3854
Multilingual	EN	**0.5225**	**0.2969**	**0.4097**
Monolingual	ES	**0.3629**	**0.2023**	**0.2826**
Multilingual	ES	**0.3629**	0.2021	0.2825
Monolingual	AR	0.5075	0.2582	0.3829
Multilingual	AR	**0.5319**	**0.2755**	**0.4037**

Note. The experiments reported above were done on pairwise re-ranking setting with evidence, with verdict and with maximum sequence length of 512 tokens.

The results in Table 5 show that multilingual training improved performance for English and Arabic compared to their monolingual counterparts. English improved from an average score of 0.3854 to 0.4097, while Arabic improved from 0.3829 to 0.4037.

For Spanish, the monolingual and multilingual settings produced nearly identical results. This suggests that multilingual supervision did not provide a clear benefit for Spanish in this configuration.

Overall, multilingual training provided better robustness, where cross-lingual supervision may help compensate for label imbalance. That's why we will use this multilingual setup for the final submissions.

4.5. Evidence Impact Analysis

We analyzed the effects of evidence-aware ranking and verdict supervision on reasoning-trace quality. Specifically, we evaluated how including or excluding the evidence text as part of the re-ranker input affects performance. This experiment was conducted using a pairwise ranking setup with a maximum sequence length of 512.

Interestingly, removing evidence did not reduce the overall score as shown in Table 6. The no-evidence setting achieved a slightly higher overall average score, mainly because of higher Recall@5. However, the evidence-aware model achieved higher Macro F1 and stronger classification performance in English and Arabic. This suggests that including evidence helps the model make more stable final predictions.

At the same time, the no-evidence setting remained competitive in ranking performance and achieved slightly higher Recall@5 overall, indicating that the model may sometimes rank traces based on patterns in the claim, trace, or predicted verdict rather than on direct evidence grounding. From this observation, we concluded that evidence alone is not sufficient to explain trace usefulness in this task. The quality of the reasoning trace itself plays an important role, which shows the effect of the rubric-based evaluation approach that explicitly assesses reasoning quality and evidence grounding.

Table 6
Effect of evidence-aware inputs on pairwise ranking performance reported on validation dataset.

Configuration	Lang.	Macro F1	Recall@5	Avg. Score
With Evidence	EN	**0.5225**	0.2969	**0.4097**
Without Evidence	EN	0.5063	**0.3072**	0.4068
With Evidence	ES	**0.3629**	0.2021	0.2825
Without Evidence	ES	0.3620	**0.2688**	**0.3154**
With Evidence	AR	**0.5319**	**0.2755**	**0.4037**
Without Evidence	AR	0.5236	0.2639	0.3938
With Evidence	Overall	**0.5619**	0.2730	0.4175
Without Evidence	Overall	0.5537	**0.2892**	**0.4215**

Note. Bold values indicate the best result for each configuration.

4.6. Official Test Results

Our official submission used a hybrid fusion approach that combines a pointwise supervised re-ranker with a rubric-guided LLM judge, with fusion weight $\lambda = 0.4$. Table 7 reports our official leaderboard results obtained by hybrid approach on the final test set for each language.

Table 7
Official Leaderboard results per language using hybrid approach.

Language	Avg. Score	Macro F1	Recall@5	True F1	Conflicting F1	False F1	Rank
Arabic	**0.6043**	**0.8679**	**0.3406**	0.8426	0.0000	0.8933	**1**
English	0.4200	0.5948	0.2452	0.7114	0.1866	0.8864	6
Spanish	**0.4847**	**0.6749**	**0.2946**	0.6154	0.4962	0.9131	**1**

The official results show that our system achieved strong multilingual performance across the three evaluation tracks. It ranked first on both the Arabic and Spanish tasks. For English, the system ranked sixth.

The strongest overall performance was obtained on Arabic, mainly due to strong True and False class performance. Spanish also showed strong results, especially for the False and Conflicting classes. English remained more challenging, mainly due to its lower Conflicting F1.

5. Ablation Studies

5.1. Ranking Objective Comparison

We compared pointwise and pairwise ranking objectives under the same experimental setup. Both models were trained using evidence-aware inputs and a maximum sequence length of 1024 tokens. For each language and ranking configuration, the best-performing checkpoint was selected based on the validation average score.

We initially explored pairwise ranking extensively because ranking-based supervision appeared promising. However, detailed ablation analysis revealed pairwise objectives to be highly sensitive to noisy multilingual supervision and contradictory verdict annotations. The pointwise ranking ultimately provided more stable and robust performance.

We also attempted to explore listwise ranking, where the model would learn from the ordering of all candidate reasoning traces for each claim. However, this experiment was aborted because the training data did not provide reliable ground-truth rankings over the full list of traces. Constructing listwise supervision from verdict agreement alone introduced noisy training targets, especially in cases where

Table 8
Comparison between pointwise and pairwise ranking objectives using the best checkpoint for each configuration reported on validation dataset.

Ranking Type	Lang.	Macro F1	Recall@5	Avg. Score
Pointwise	EN	**0.5163**	**0.3034**	**0.4099**
Pairwise	EN	0.5160	0.3013	0.4087
Pointwise	ES	**0.3854**	**0.2177**	**0.3016**
Pairwise	ES	0.3844	0.2166	0.3005
Pointwise	AR	**0.5395**	**0.2772**	**0.4084**
Pairwise	AR	0.5269	0.2701	0.3985
Pointwise	Overall	**0.5575**	**0.2802**	**0.4189**
Pairwise	Overall	0.5560	0.2771	0.4166

traces had contradictory or inconsistent verdict annotations. Therefore, we did not continue with listwise training and focused the final supervised analysis on pointwise and pairwise settings.

The results in Table 8 show that pointwise ranking slightly outperformed pairwise ranking overall.

5.2. Context Length Analysis

We evaluated the impact of increasing the context length on reasoning-trace ranking performance with pairwise ranking objective.

Table 9
Effect of maximum sequence length on pairwise ranking performance reported on validation dataset.

Max Len	Lang.	Macro F1	Recall@5	Avg. Score
512	EN	**0.5225**	0.2969	**0.4097**
1024	EN	0.5112	**0.3076**	0.4094
512	ES	0.3629	0.2021	0.2825
1024	ES	**0.3900**	**0.2177**	**0.3039**
512	AR	**0.5319**	**0.2755**	**0.4037**
1024	AR	0.5267	0.2722	0.3995
512	Overall	**0.5619**	0.2730	**0.4175**
1024	Overall	0.5516	**0.2814**	0.4165

The results in Table 9 show that increasing the maximum sequence length from 512 to 1024 did not consistently improve pairwise ranking performance. The 1024-token setting achieved slightly higher Recall@5 overall, suggesting that longer inputs may help retrieve better-ranked traces. However, the 512-token setting achieved a higher overall Macro F1 and average score.

This indicates that longer context does not always lead to better performance in noisy reasoning-trace ranking. For pairwise training, longer inputs may introduce additional irrelevant or contradictory information, making the model more sensitive to noisy supervision. Therefore, the 512-token setting provided a slightly better balance between classification performance and ranking quality overall.

5.3. Verdict Supervision Ablation

We evaluated the effect of verdict supervision while keeping the evidence input fixed. Both settings use evidence-aware inputs and pairwise ranking with maximum sequence length of 512.

The results in Table 10 show that verdict supervision improved the overall Macro F1, Recall@5, and average score. The improvement was especially clear in English and Arabic, while Spanish remained almost unchanged. This suggests that verdict information provides a useful supervision signal for pairwise reasoning-trace ranking.

Table 10
Effect of verdict supervision on pairwise ranking performance.

Configuration	Lang.	Macro F1	Recall@5	Avg. Score
With Verdict	EN	**0.5225**	**0.2969**	**0.4097**
Without Verdict	EN	0.5000	0.2834	0.3917
With Verdict	ES	**0.3629**	0.2021	0.2825
Without Verdict	ES	**0.3629**	**0.2023**	**0.2826**
With Verdict	AR	**0.5319**	**0.2755**	**0.4037**
Without Verdict	AR	0.4975	0.2069	0.3522
With Verdict	Overall	**0.5619**	**0.2730**	**0.4175**
Without Verdict	Overall	0.5251	0.2495	0.3873

6. Conclusion and Future Work

In this work, we explored multilingual reasoning-trace ranking for numerical fact-checking in the CheckThat! 2026 shared task. We investigated supervised re-ranking approaches under multiple training configurations, including pointwise and pairwise ranking objectives, multilingual versus monolingual training, evidence-aware inputs, verdict supervision, and different context lengths.

Our experiments showed that reasoning-trace ranking is highly sensitive to noisy supervision and inconsistent verdict annotations. Although pairwise ranking produced competitive results, pointwise ranking provided more stable and robust performance across multilingual settings.

We further demonstrated that verdict supervision and multilingual training generally improved overall performance, while evidence-aware ranking improved reasoning grounding and classification stability.

To address limitations of purely supervised ranking, we additionally explored rubric-guided LLM evaluation for reasoning quality and evidence grounding. Combining supervised re-ranking with rubric-based evaluation provided a complementary hybrid framework for multilingual numerical fact-checking.

Our final system achieved strong multilingual performance, ranking first on both the Arabic and Spanish evaluation tracks in the official leaderboard.

For future work, further analysis of the English track is recommended in order to better understand why it remained more challenging than Arabic and Spanish. In particular, the characteristics of English reasoning traces and the presence of noisy or contradictory annotation patterns should be investigated more thoroughly. Additionally, smaller and more computationally efficient LLM judge models than `Qwen-27B` could be explored to reduce inference cost while maintaining reasoning-evaluation quality.

Declaration on Generative AI

During the preparation of this work, we used ChatGPT and Grammarly for grammar and spelling checks, as well as for paraphrasing and rewording. After using these tools, we carefully reviewed and edited the content as needed and take full responsibility for the final version of this publication.

References

[1] A. Anand, A. Anand, V. Setty, et al., Quantemp: A real-world open-domain benchmark for fact-checking numerical claims, arXiv preprint arXiv:2403.17169 (2024).

[2] P. Chungkham, V. Viswanathan, V. Setty, A. Anand, Think right, not more: Test-time scaling for numerical claim verification, in: Empirical Methods in Natural Language Processing (EMNLP) 2025, Association for Computational Linguistics, 2025, pp. 24345–24363.

[3] J. M. Struß, S. Schellhammer, S. Dietze, V. V., V. Setty, T. Chakraborty, P. Nakov, A. Anand, P. Chungkham, S. Hafid, D. Sahnan, K. Todorov, The clef-2026 checkthat! lab: Advancing multilingual fact-checking, in: R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, Z. Ren, S. Verberne, N. Bai, M. Mansoury (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2026, pp. 325–335.

[4] H. Bouamor, G. Iturra-Bocaz, P. Galuščáková, F. Alam, Overview of the clef-2025 checkthat! lab task 3 on fact-checking numerical claims (2025).

[5] V. V, V. Setty, P. Chungkham, A. Anand, F. Alam, Overview of the CLEF-2026 CheckThat! lab task 2 on fact-checking numerical claims, CLEF 2026, Jena, Germany, 2026.

[6] Q. T. Le, I. Badache, A. Yacoub, M. E. A. Hamri, Lis at checkthat! 2025: multi-stage open-source large language models for fact-checking numerical claims, in: Notebook for the CheckThat! Lab at the 16th Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038, 2025.

[7] K. Gurumurthy, A. Shrivastava, P. K. Rajpoot, P. Devadiga, B. Hazarika, M. Jain, M. Sharma, A. Suneesh, A. B. Suresh, A. U. Baliga, Tifin at checkthat! 2025: cross-lingual subjectivity classification in news through monolingual, multilingual, and zero-shot learning, Faggioli et al (2025).

[8] L. Duesterwald, A. Arora, C. Cardie, Cornellnlp at checkthat! 2025: hybrid llama–gpt-4 ensembles with confidence filtering for numerical claim verification, Faggioli et al (2025).

[9] M. A. Abdallah, R. M. Fekry, S. R. El-Beltagy, Ngu_research at checkthat! 2025: an llm based hybrid fact-checking pipeline for numerical claims, Faggioli et al (2025).

[10] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, arXiv preprint arXiv:2203.11171 (2022).

[11] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, C. Callison-Burch, Faithful chain-of-thought reasoning, in: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 305–329.

[12] H. Hashemi, J. Eisner, C. Rosset, B. Van Durme, C. Kedzie, Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 13806–13834.

[13] Q. Team, Qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[15] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.

Appendix A: LLM Judge Prompt

The complete evaluation prompt used for the LLM judge is included below for reproducibility. The prompt requests structured numeric scores so that its output can be parsed and combined with the supervised re-ranker score.

Trace-only LLM evaluation Prompt

You are judging the QUALITY of a reasoning trace for a fact-checking task.
Important: You do NOT see the evidence. Judge only whether the trace is internally good.
Score from 0 to 10 using this rubric:

1. **Claim understanding:** Does the trace correctly understand the claim and identify the key thing being checked?
2. **Numbers/dates:** If the claim involves numbers, quantities, dates, years, rankings, percentages, counts, or time expressions, does the trace notice and reason about them?
3. **Internal consistency:** Does the conclusion follow from the reasoning? Penalize contradictions such as saying there is not enough information but then concluding that the claim is true.
4. **No overclaiming:** If the trace says evidence is missing or insufficient, it should not confidently say True or False.
5. **Specificity:** Is the trace specific to this claim, or is it generic boilerplate?

Return JSON only:

```
{
  "score": number from 0 to 10,
  "reason": "brief reason, one or two sentences"
}
```

Figure 2: Prompt template used for rubric-guided trace-only LLM evaluation of candidate reasoning traces.

Grounded LLM Evaluation Prompt

You are judging the QUALITY of a reasoning trace for a fact-checking task.
You see the claim, the evidence, and the reasoning trace. Judge whether the trace is faithful to the evidence and useful for reaching the correct verdict.
Score from 0 to 10 using this rubric:

1. **Claim understanding:** Does the trace correctly understand the claim and identify the key numbers, dates, or conditions?
2. **Evidence use:** Does the trace accurately describe what the evidence says? Penalize invented evidence, ignored relevant evidence, or misread evidence.
3. **Claim-evidence comparison:** Does the trace directly compare the claim against the evidence instead of discussing them separately?
4. **Verdict justification:** Does the final conclusion follow from the evidence? Penalize cases where the evidence is missing or irrelevant but the trace says "True" confidently.
5. **No unsupported assumptions:** Does the trace avoid adding outside facts or guesses that are not in the provided evidence?

Return JSON only:

```
{
  "score": number from 0 to 10,
  "reason": "brief reason, one or two sentences"
}
```

Figure 3: Prompt template used for rubric-guided grounded LLM evaluation of candidate reasoning traces.