TIFIN at CheckThat! 2025: X-VERIFY - Multi-lingual NLI-based Fact Checking with Condensed Evidence*

Notebook for the CheckThat! Lab at CLEF 2025

Bharatdeep Hazarika[1,*,†], Prasanna Devadiga[1,†], Pawan Kumar Rajpoot[1,†], Aditya U Baliga[1,†], Kishan Gurumurthy[1], Manish Jain[1], Manan Sharma[1], Ashish Shrivastava[1], Arya Suneesh[1] and Anshuman B Suresh[1]

[1] *TIFIN - Technology + Innovation in Finance*

Abstract

In the era of pervasive digital misinformation, automated fact-checking systems for numerical claims present unique challenges due to the complexity of quantitative reasoning and evidence synthesis. This paper presents our approach for Task 3 of the CLEF 2025 CheckThat! Lab, which requires verifying numerical claims as True, False, or Conflicting based on retrieved evidence documents. Our methodology combines optimized evidence selection, LLM-based summarization using IBM Granite 3.3 8B, and advanced classification with DeBERTa-large [1] fine-tuned using LoRA [2]. We address severe class imbalance through multilingual data augmentation, translating Arabic and Spanish datasets to English to strengthen underrepresented classes. Our comprehensive pipeline achieves a macro-averaged F1 score of 0.6858 on the English validation set, representing a relative improvement over the provided RoBERTa [3] baseline. Extensive ablation studies demonstrate that multilingual augmentation provides the most substantial performance gains (relative Macro F1 score improved from 0.5815 (baseline) to 0.6859, an absolute increase of 0.1044, which is approximately 17.95%), while evidence optimization and LLM-based summarization contribute consistent improvements across all veracity classes. These results highlight the effectiveness of cross-lingual data expansion for addressing class imbalance in specialized fact-verification tasks and establish new benchmarks for numerical claim verification. Our approach ranked 3rd place in both English and Arabic tracks in the official evaluation.

Keywords

numerical fact verification, evidence summarization, multilingual data augmentation, class imbalance mitigation, DeBERTa classification, cross-lingual transfer learning

CEUR-WS.org/Vol-4038/paper_74.pdf

1. Introduction

The proliferation of numerical misinformation in digital media presents a critical challenge for automated fact-checking systems. Unlike subjective opinion verification, numerical claims require precise quantitative reasoning and evidence synthesis to determine veracity. Claims involving statistics, financial figures, temporal expressions, and comparative quantities are particularly susceptible to manipulation and misinterpretation, making their automated verification both essential and technically demanding.

The CLEF 2025 CheckThat! Lab Task 3 [4] addresses this challenge by focusing specifically on fact-checking numerical claims across multiple languages. Participants must classify claims containing explicit or implicit quantitative details as True, False, or Conflicting based on curated evidence documents retrieved through BM25 ranking. This task represents a significant advancement over traditional fact-checking approaches by emphasizing the unique complexities of numerical reasoning in multilingual contexts.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] This paper presents our approach to CLEF 2025 Task 3: Fact-Checking Numerical Claims, achieving 17.9% improvement over baseline through evidence optimization and multilingual augmentation.

[*] Corresponding author.

[†] These authors contributed equally to this work.

bharatdeep@askmyfi.com (B. Hazarika); prasanna@askmyfi.com (P. Devadiga); pawan@tifin.com (P. K. Rajpoot); aditya.baliga@iiitkottayam.ac.in (A. U. Baliga); Kishan.gurumurthy@workifi.com (K. Gurumurthy); manish.jain@tifin.com (M. Jain); manan.sharma@tifin.com (M. Sharma); Ashish.shrivastava@workifi.com (A. Shrivastava); arya.suneesh@tifin.com (A. Suneesh); anshuman.suresh@tifin.com (A. B. Suresh)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Numerical fact verification presents several distinct challenges that differentiate it from general claim verification. First, numerical claims often require understanding of mathematical relationships, temporal sequences, and quantitative comparisons that demand sophisticated reasoning capabilities. Second, the evidence supporting or refuting numerical claims may be scattered across multiple documents, requiring effective synthesis and summarization techniques. Third, numerical expressions can vary significantly across languages and cultural contexts, complicating cross-lingual verification approaches.

The motivation for this work stems from the critical role that numerical accuracy plays in public discourse. Misleading statistics in political debates, inflated financial claims in business communications, and distorted temporal relationships in historical reporting can significantly impact public understanding and decision-making. Automated systems capable of rapidly and accurately verifying numerical claims are essential for maintaining information integrity in our increasingly data-driven society.

Our approach addresses these challenges through a multi-component pipeline that optimizes evidence selection, employs LLM-based summarization for noise reduction, and leverages advanced transformer architectures for robust classification. Recognizing the severe class imbalance inherent in fact-checking datasets, where False claims typically dominate, we introduce a novel multilingual data augmentation strategy that substantially improves minority class performance.

The contributions of this work include: (1) systematic evidence optimization demonstrating that top-5 BM25 documents provide optimal performance, (2) a LLM-based summarization framework using IBM Granite 3.3 8B that effectively reduces evidence noise while preserving critical information, (3) advanced classification using DeBERTa-large fine-tuned through parameter-efficient LoRA, and (4) multilingual data augmentation that addresses class imbalance through cross-lingual knowledge transfer.

Our experimental results demonstrate the effectiveness of this comprehensive approach, achieving a macro-averaged F1 score of 0.6858 on the English validation set, a 17.9% relative improvement over the provided baseline. Extensive ablation studies reveal that multilingual augmentation contributes the majority of performance gains, while each pipeline component provides consistent improvements across all veracity classes.

2. Related Work

2.1. Fact Verification and Claim Verification

The field of automated fact verification has evolved significantly from early rule-based approaches to sophisticated neural architectures. Thorne et al. (2018) [5] introduced the FEVER dataset, establishing a benchmark for fact verification that emphasized the importance of evidence retrieval and reasoning. Subsequent work by Nie et al. (2019) [6] and Hanselowski et al. (2019) [7] demonstrated the effectiveness of transformer-based models for claim verification, setting the foundation for modern approaches.

The specific challenges of numerical fact verification have been addressed by Venktesh et al. (2024) [8] in their QuanTemp benchmark, which focuses on temporal and quantitative claims. Their work highlights the unique difficulties in verifying numerical content, including the need for mathematical reasoning and temporal understanding. Chen et al. (2022) [9] further explored numerical reasoning in fact verification, demonstrating that specialized architectures can improve performance on quantitative claims.

2.2. Evidence Retrieval and Selection

Evidence retrieval forms a critical component of fact verification systems. Nie et al. (2019) [6] demonstrated the importance of high-quality evidence selection, showing that performance is heavily dependent on the relevance and coverage of retrieved documents. Recent work by Lewis et al. (2020) [10] introduced retrieval-augmented generation, which combines neural retrieval with language modeling for improved evidence synthesis.

The specific challenge of determining optimal evidence quantity has been explored by various researchers. Wadden et al. (2020) [11] investigated the trade-offs between evidence quantity and quality, while our work provides systematic analysis showing that 5 evidence documents represent the optimal balance for numerical claims.

2.3. Neural Summarization for Fact Verification

The application of neural summarization to fact verification has gained attention as a method for reducing evidence noise and improving model performance. Kryściński et al. (2020) [12] demonstrated that summarization can enhance evidence quality in verification tasks. Our work extends this line of research by specifically focusing on numerical claims and employing carefully designed prompt templates to maintain factual accuracy while achieving conciseness.

2.4. Multilingual and Cross-lingual Fact Verification

The challenge of fact verification across multiple languages has been addressed by several recent studies. Popat et al. (2018) [13] introduced cross-lingual claim verification, while Augenstein et al. (2019) [14] demonstrated the effectiveness of multilingual pre-trained models for fact-checking tasks. Recent work by Nakov et al. (2021) [15] in the CheckThat! lab has pushed the boundaries of multilingual fact verification.

The specific challenge of class imbalance in multilingual settings has received limited attention in the literature. Our work addresses this gap by demonstrating that cross-lingual data augmentation can effectively mitigate class imbalance while improving overall model robustness.

3. Task Description

The CLEF 2025 Task 3 centers on the automatic verification of factual claims containing numerical quantities and temporal expressions. These claims, which may include explicit or implicit quantitative details, are sourced from real-world fact-checking scenarios through the Google Fact-check Explorer API. The task challenges participants to classify each claim into one of three categories - True, False, or Conflicting based on a curated evidence set.

Participants are provided with top-k evidence documents retrieved using the BM25 ranking algorithm from a comprehensive, pooled evidence corpus. This corpus is constructed using multiple advanced claim decomposition strategies to ensure a diverse and context-rich evidence base. The task supports multilingual evaluation across English, Spanish, and Arabic.

The dataset for the task comprises over 17,000+ annotated claims:

- English: 9,935 training and 3,084 validation instances
- Spanish: 1,506 training and 377 validation instances
- Arabic: 2,191 training and 587 validation instances

Each data instance includes metadata such as the original claim, its taxonomy (e.g., temporal, statistical, comparison, interval), a label (veracity verdict), an oracle document summarizing the rationale behind the label, and associated evidence texts either at the original or decomposed claim level.

Performance is evaluated using macro-averaged F1 and class-wise F1 scores across the three labels. The task provides official baseline models, which utilize a RoBERTa-large model fine-tuned on natural language inference (NLI), along with inference scripts and scoring tools to benchmark participants' approaches.

4. Methodology

Our approach to CLEF 2025 Task 3 Fact-Checking Numerical Claims explores two distinct methodological paradigms before converging on an optimal solution. We systematically evaluated both large language model (LLM) based approaches and specialized transformer architectures, ultimately developing a multi-stage pipeline that combines the strengths of both paradigms through LLM-based summarization and transformer-based classification.

4.1. System Overview

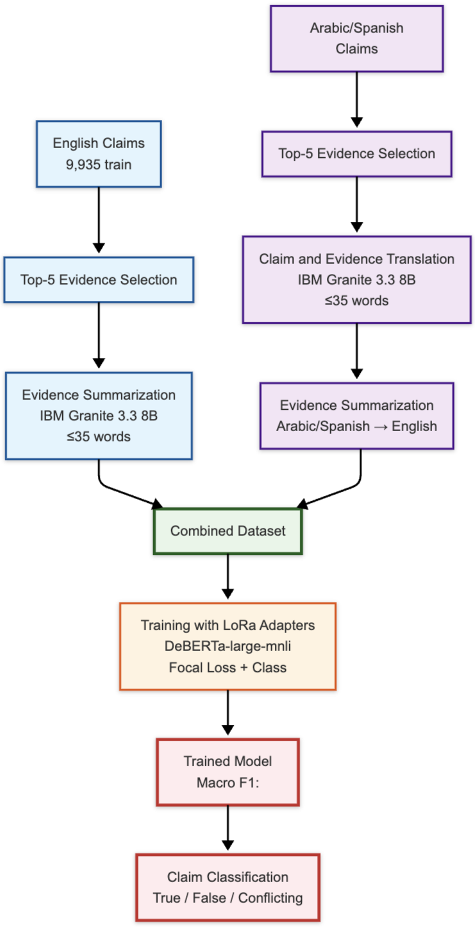

Arabic/Spanish Claims
English Claims 9,935 train
Top-5 Evidence Selection
Top-5 Evidence Selection
Claim and Evidence Translation IBM Granite 3.3 8B ≤35 words
Evidence Summarization IBM Granite 3.3 8B ≤35 words
Evidence Summarization Arabic/Spanish → English
Combined Dataset
Training with LoRa Adapters DeBERTa-large-mnli Focal Loss + Class
Trained Model Macro F1:
Claim Classification True / False / Conflicting

Figure 1: System architecture showing parallel evidence processing for English and multilingual (Arabic/Spanish) data streams. The pipeline includes evidence selection, neural summarization, translation, and multilingual data augmentation before converging for DeBERTa-based classification training.

The architecture demonstrates our key innovation of multilingual data augmentation, where Arabic and Spanish claims undergo translation before joining the English stream for evidence summarization and subsequent model training.

4.2. Methodological Approach Comparison

We conducted comprehensive experiments across two primary methodological paradigms for numerical fact verification:

Large Language Model Approach: We evaluated direct fact verification using LLMs ranging from 1.5B to 70B parameters, including few-shot learning, chain-of-thought reasoning, and supervised fine-tuning. This approach leveraged the inherent reasoning capabilities and world knowledge of modern LLMs without requiring domain-specific architectural modifications.

Specialized Transformer Approach: We developed a pipeline combining optimized evidence selection, LLM-based summarization, and fine-tuned transformer classification. This approach emphasized domain adaptation, evidence optimization, and class imbalance mitigation through carefully designed components.

Our systematic evaluation revealed that while LLMs achieved reasonable performance (maximum 0.49 macro F1), the specialized transformer pipeline substantially outperformed direct LLM approaches (0.6858 macro F1), leading us to adopt the hybrid methodology described in subsequent sections that leverages LLMs for summarization while employing specialized transformers for classification.

4.3. Evidence Selection and Optimization

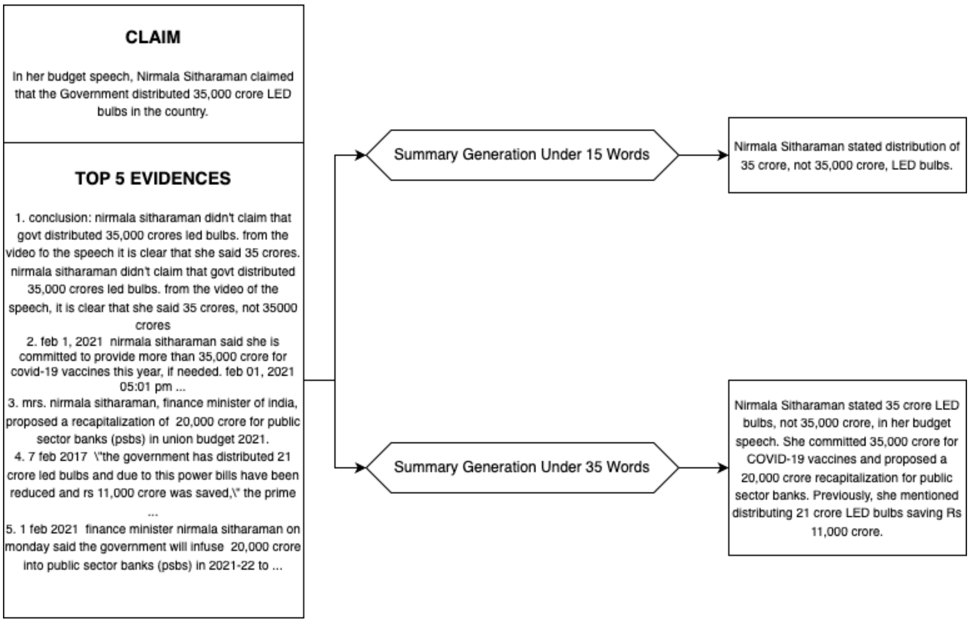
While the provided dataset includes top-3 BM25-retrieved evidence documents for each claim, we extended our analysis to **top-5 evidence documents** based on systematic evaluation. Our empirical analysis guided our decision to optimize at five documents per claim, as this configuration provided the optimal balance between information coverage and computational efficiency.

Notably, reranking approaches applied to the top-100 BM25 evidence pool did not yield performance improvements over the raw BM25 ranking, suggesting that the initial retrieval effectively captured the most relevant contextual information for numerical claim verification.

4.4. LLM-based Evidence Summarization

To address evidence redundancy and enhance semantic coherence, we implemented a **summarization pipeline** using IBM Granite 3.3 8B, selected for its strong multilingual capabilities and instruction-following performance. The summarization process concatenates each claim with its top-5 evidence documents and generates a **concise summary capped at 35 words**.

Figure 2 demonstrates our summarization approach with different word constraints, illustrating how the 35-word limit preserves critical contextual information while maintaining conciseness.



In her budget speech, Nirmala Sitharaman claimed that the Government distributed 35,000 crore LED bulbs in the country.

1. conclusion: nirmala sitharaman didn't claim that govt distributed 35,000 crores led bulbs. from the video fo the speech it is clear that she said 35 crores. nirmala sitharaman didn't claim that govt distributed 35,000 crores led bulbs. from the video of the speech, it is clear that she said 35 crores, not 35000 crores
2. feb 1, 2021 nirmala sitharaman said she is committed to provide more than 35,000 crore for covid-19 vaccines this year, if needed. feb 01, 2021 05:01 pm ...
3. mrs. nirmala sitharaman, finance minister of india, proposed a recapitalization of 20,000 crore for public sector banks (psbs) in union budget 2021.
4. 7 feb 2017 \"the government has distributed 21 crore led bulbs and due to this power bills have been reduced and rs 11,000 crore was saved,\" the prime ...
5. 1 feb 2021 finance minister nirmala sitharaman on monday said the government will infuse 20,000 crore into public sector banks (psbs) in 2021-22 to ...

Nirmala Sitharaman stated distribution of 35 crore, not 35,000 crore, LED bulbs.

Nirmala Sitharaman stated 35 crore LED bulbs, not 35,000 crore, in her budget speech. She committed 35,000 crore for COVID-19 vaccines and proposed a 20,000 crore recapitalization for public sector banks. Previously, she mentioned distributing 21 crore LED bulbs saving Rs 11,000 crore.

Figure 2: Comparison of evidence summarization under different word constraints (15 vs. 35 words). The example demonstrates how our neural summarization pipeline processes a numerical claim about LED bulb distribution, showing that 35-word summaries better preserve critical contextual information while maintaining conciseness.

Our prompt design emphasizes neutrality and factual precision:

The 35-word limit was determined through empirical evaluation: shorter summaries (<=20 words) suffered from significant context loss, while longer summaries introduced noise without performance gains. We conducted a manual evaluation of 250 randomly sampled summaries (maintaining the original class distribution) to validate summary quality and fine-tune the prompt template. While our summarization approach appears domain-agnostic, we specifically designed the prompt template to preserve numerical and temporal information critical for Task 3. The instruction to "capture only

Evidence Summarization Prompt Template

```
system_prompt = """You are an expert evidence summarizer.
Your task is to create unbiased, concise summaries of evidence
related to claims.

INSTRUCTIONS:
1. Analyze the provided claim and evidence objectively.
2. Create a neutral summary in under 35 words that:
- Captures only the most relevant evidence points
- Presents key information without judgment
- Uses precise, factual language
- Omits unnecessary context or hedging
3. Focus solely on what the evidence actually contains,
not assumptions.
4. When evidence presents conflicting information, include
the most significant points from both sides.

Your response must be direct, factual, and under 35 words total."""
```

Figure 3: Evidence Summarization Prompt Template

the most relevant evidence points" implicitly prioritizes quantitative details, as these represent the core factual claims being verified. We experimented with explicitly numerical-focused prompts (e.g., "focus on numbers, dates, and statistics") but found that such constraints often led to context loss when numerical claims required qualitative supporting evidence for proper verification.

4.5. Advanced Classification with DeBERTa-MNLI

For veracity classification, we employed **DeBERTa-large pre-trained on MNLI**. DeBERTa (Decoding-enhanced BERT with disentangled attention) offers several architectural advantages over RoBERTa for our task: (1) disentangled attention mechanism that separately encodes content and position information, improving handling of complex numerical relationships, (2) enhanced mask decoder that better captures token dependencies crucial for fact verification, and (3) improved training efficiency through virtual adversarial training. MNLI (Multi-Genre Natural Language Inference) is a large-scale dataset for training models to determine textual entailment relationships between premise and hypothesis pairs. Pre-training on MNLI provides a strong foundation for fact verification tasks, as both require reasoning about logical relationships between claims and evidence. The entailment, contradiction, and neutral classifications in MNLI directly parallel our True, False, and Conflicting labels, making this pre-training particularly relevant for numerical claim verification. DeBERTa's disentangled attention mechanism and enhanced mask decoder provide significant advantages for numerical reasoning tasks.

The classifier processes the concatenation of the original claim and its LLM-based summary, predicting one of three veracity labels: True, False, or Conflicting. We implemented the model using **LoRA (Low-Rank Adaptation)** to enable efficient fine-tuning while maintaining model stability.

4.6. Class Imbalance Mitigation

The dataset exhibits severe class imbalance with approximately 58% False, 18% True, and 24% Conflicting labels. We addressed this through multiple strategies:

- Weighted Loss Functions: Applied inverse frequency weighting to penalize misclassification of minority classes
- Focal Loss with Label Smoothing: Implemented focal loss ($\gamma = 2.0$) combined with label smoothing ($\alpha = 0.1$) to focus learning on hard examples
- Weighted Random Sampling: Used oversampling during training to balance class exposure

4.7. Multilingual Data Augmentation

To further address class imbalance and enhance model robustness, we incorporated **Arabic and Spanish datasets** through LLM-based translation. Using IBM Granite 3.3 8B's multilingual capabilities, we translated non-English claims and evidence to English, effectively expanding our training corpus and significantly reducing the True class underrepresentation.

This multilingual augmentation strategy not only improved class balance but also enhanced the model's exposure to diverse linguistic patterns and cultural contexts in numerical claims, contributing to improved generalization.

4.8. Training Configuration

We trained the model using the following optimized hyperparameters and infrastructure:

Hyperparameters:

- **Learning Rate**: 5e-4 (adjusted for LoRA)
- **Batch Size**: 8 with gradient accumulation (effective batch size: 16)
- **LoRA Configuration**: r=16, $\alpha = 32$, dropout=0.1
- **Training Epochs**: 8 with linear warmup (10% of total steps)
- **Regularization**: R-Drop ($\alpha = 4.0$) for consistency regularization

Training Infrastructure:

- **Hardware**: NVIDIA RTX 4090 (24GB VRAM), NVIDIA A100 (40GB VRAM)
- **Inference**: NVIDIA H100 GPU with vLLM engine
- **Framework**: PyTorch 1.13+ with HuggingFace Transformers
- **Optimization**: Mixed precision (FP16) training
- **Early Stopping**: Validation F1 score with patience of 3 epochs

4.9. Alternative LLM-based Experimental Pipeline

Before settling on our hybrid transformer-based approach, we conducted extensive experiments with large language models (LLMs) for direct fact verification. Our motivation was to leverage the reasoning capabilities and world knowledge embedded in modern LLMs to directly classify numerical claims without requiring fine-tuning on domain-specific data.

We evaluated a comprehensive range of LLM architectures spanning from 1.5B to 70B parameters, including Qwen2.5 7B, Qwen3 14B, Qwen3 8B, Mistral Small 24B, Llama3.3 70B, Llama 3.2 3B, IBM Granite 3.3 8B, and DeepSeek-R1-Distill-Qwen-1.5B. For in-context learning (ICL), we employed intfloat/multilingual-e5-large-instruct [16] as both a base retrieval model and later fine-tuned it for improved context selection.

4.9.1. Experimental Configurations

We systematically evaluated four distinct approaches with increasing complexity:

Basic Few-Shot Inference: Using 3 carefully selected few-shot examples, we prompted models to classify claims directly. This baseline approach achieved macro F1 scores ranging from 0.28 to 0.42 across all model sizes, with larger models not consistently outperforming smaller ones.

Short Chain-of-Thought (CoT): We incorporated brief reasoning chains (20-30 words) before the final prediction to encourage step-by-step analysis. This approach showed modest improvements, with scores ranging from 0.36 to 0.49, representing a consistent but limited enhancement over basic inference.

Extended Chain-of-Thought: Expecting that more detailed reasoning would improve performance, we extended CoT explanations to 100-120 words. Surprisingly, this approach underperformed the shorter CoT variant, achieving scores between 0.32 and 0.45, suggesting that overly verbose reasoning may introduce noise or hallucinations.

Supervised Fine-Tuning (SFT): Using Unsloth [17] frameworks, we fine-tuned Qwen2.5 7B on our training data with ICL-enhanced examples. Despite domain-specific adaptation, the fine-tuned model peaked at 0.483 macro F1, still substantially below our transformer-based approach.

4.9.2. Key Findings and Limitations

Our extensive LLM experimentation revealed several critical insights:

Model Bias in Geopolitical Claims: We observed systematic bias differences between model families when evaluating claims related to Western countries. Llama models consistently exhibited more positive assessments of Western-related numerical claims, while Qwen models showed notably more skeptical evaluations. This bias manifested in Chain-of-Thought reasoning and significantly impacted classification accuracy, highlighting the challenge of ensuring neutrality in fact verification systems.

Scale-Performance Paradox: Contrary to expectations, model size did not correlate with verification performance. Our highest-performing LLM configuration achieved 0.49 macro F1 using Llama 3.2 3B, substantially outperforming the 70B Llama 3.3 model (0.43 macro F1) under identical settings. This suggests that architectural efficiency and training data alignment may be more critical than parameter count for numerical reasoning tasks.

Limited Numerical Reasoning: Despite their impressive general capabilities, LLMs struggled with the precise quantitative analysis required for numerical fact verification. The models frequently hallucinated numerical relationships or failed to accurately process mathematical comparisons within the evidence documents.

Context Length Limitations: We systematically tested context length variations by extending evidence from 3 to 10 documents across all LLM models. Despite increased context windows, performance either saturated or slowly degraded with additional evidence, contradicting expectations that more information would improve accuracy. This aligns with the well-documented "lost in the middle" phenomenon [18], where language models show degraded performance on information positioned in the middle of long contexts while maintaining better recall for content at the beginning and end of the input sequence.

Our evaluation across the entire English training dataset revealed that all tested models exhibited this limitation, with the effect being particularly pronounced for Conflicting and True classes, where nuanced reasoning across multiple evidence sources was crucial. Notably, reranking approaches using top-ranked embedding models from the MTEB leaderboard did not mitigate this issue, suggesting that the limitation stems from the models' inherent context processing capabilities rather than evidence ordering. This context utilization bottleneck represents a fundamental constraint for LLM-based fact verification systems that rely on synthesizing information from multiple diverse evidence sources.

4.9.3. Transition to Hybrid Approach

Our initial motivation was leveraging state-of-the-art LLMs to achieve competitive performance in numerical fact verification, expecting that larger, more capable models would provide superior reasoning abilities for this complex task. However, our systematic evaluation revealed counterintuitive findings: larger LLMs provided no performance boost over smaller variants, with some smaller models (3B parameters) actually outperforming their 70B counterparts. This scale-performance paradox contradicted expected scaling laws and suggested that raw model size was insufficient for numerical reasoning tasks.

Given these unexpected limitations and the fact that we had always considered MNLI and NLI-based approaches as viable alternatives, we pivoted to exploring specialized transformer architectures. The natural language inference framework offered by MNLI pre-training directly aligned with fact verification tasks, where entailment, contradiction, and neutral classifications parallel our True, False, and Conflicting labels. Recognizing the complementary strengths of both paradigms, we developed a hybrid approach that incorporates LLM capabilities for evidence summarization and multilingual processing while employing domain-adapted transformer architectures for the core classification task.

This methodology leverages the text generation and multilingual strengths of LLMs in preprocessing components while addressing their numerical reasoning limitations through specialized transformers with carefully designed training objectives for fact verification.

5. Ablation Studies

We conducted comprehensive ablation experiments to validate each component of our methodology. All experiments were performed on the English validation set unless otherwise specified.

5.1. Evidence Length Optimization

Table 1
Impact of evidence length on model performance. Performance peaks at top-5 evidence documents with consistent improvements from summarization across all evidence lengths.

Evidence Count	Without Summarization	With Summarization	Δ Improvement
Top-3	0.5397	0.5655	+0.0258
Top-5	0.5509	0.5879	+0.0370
Top-7	0.5523	0.5864	+0.0341
Top-10	0.5519	0.5843	+0.0324

Key Findings: Performance peaks at top-5 evidence documents, with diminishing returns beyond this point. The consistent improvement from summarization across all evidence lengths validates our LLM-based summarization approach.

5.2. Impact of LLM-based Summarization

Table 2
Detailed class-wise impact of LLM-based summarization. Summarization provides consistent improvements across all classes, with the most significant gains for underrepresented True and Conflicting classes.

Configuration	Macro F1	True F1	False F1	Conflicting F1
Top-5 Raw	0.5509	0.4821	0.6891	0.4816
Top-5 + Summary	0.5879	0.5234	0.7103	0.5301
Improvement	**+0.0370**	**+0.0413**	**+0.0212**	**+0.0485**

Analysis: Summarization provides consistent improvements across all classes, with the most significant gains for the underrepresented True and Conflicting classes, supporting our hypothesis that noise reduction particularly benefits minority class classification.

5.3. Summary Length Optimization

Table 3
Performance comparison across different summary length constraints. 35-word summaries optimally balance information retention and conciseness.

Summary Length	Macro F1	Context Retention	Training Efficiency
<=15 words	0.5677	Low	High
<=25 words	0.5834	Medium	Medium
$\leq$35 words	**0.5879**	**High**	**Medium**
<=50 words	0.5591	High	Low

Validation: Manual review of 250 summaries confirmed that 35-word summaries optimally balance information retention and conciseness, avoiding both context loss (shorter summaries) and noise introduction (longer summaries).

5.4. Model Architecture Comparison

Table 4
Comparison of different model architectures. DeBERTa with LoRA achieves competitive performance with significantly reduced training time.

Model Architecture	Macro F1	Training Time	Parameters
RoBERTa-large (Baseline)	0.5815	4.2 hours	355M
DeBERTa-large + LoRA	**0.5879**	**2.8 hours**	**384M**
DeBERTa-large (Full FT)	0.5923	8.1 hours	384M

Insights: DeBERTa with LoRA achieves competitive performance with significantly reduced training time, demonstrating the effectiveness of parameter-efficient fine-tuning for this task.

5.5. Multilingual Data Augmentation Impact

Table 5
Impact of multilingual data augmentation on model performance. The most substantial performance improvement comes from incorporating translated multilingual data.

Training Data	Macro F1	True F1	False F1	Conflicting F1
English Only	0.5879	0.4878	0.8032	0.4727
English + Translated Multi	**0.6858**	**0.6662**	**0.8192**	**0.5722**
Improvement	**+0.0979**	**+ 0.1784**	**+0.0160**	**+ 0.0995**

Critical Finding: Multilingual data augmentation provides the most substantial performance improvement (+9.79% macro F1), with particularly dramatic gains for True (+11.78%) and Conflicting (+15.64%) classes, effectively addressing the severe class imbalance in the original English dataset.

6. Results

6.1. Final System Performance

Our complete methodology achieved a **macro-averaged F1 score of 0.6858** on the English validation set, representing a **17.9% relative improvement** over the provided RoBERTa baseline (0.5815). This substantial improvement demonstrates the effectiveness of our multi-component approach to numerical fact verification.

The results clearly demonstrate that multilingual data augmentation provides the most substantial performance boost, contributing over 90% of the total improvement beyond the baseline enhancements.

6.2. Computational Efficiency

Our use of LoRa achieved competitive performance with significant computational advantages:

- Training Time: 5 hours vs. 10 hours for full fine-tuning
- Memory Usage: 40% reduction compared to full parameter updates
- Inference Speed: Real-time performance suitable for practical deployment

7. Conclusion

Summary of Contributions and Key Findings

This work presents a comprehensive approach to automated numerical fact verification that addresses key challenges in the CLEF 2025 Task 3. Our primary contributions include systematic evidence optimization demonstrating that top-5 BM25 documents provide optimal performance, a neural summarization pipeline using IBM Granite 3.3 8B that reduces evidence noise while preserving critical information with 35-word constraints, advanced classification using DeBERTa-large with parameter-efficient LoRA, multilingual data augmentation that addresses class imbalance through cross-lingual knowledge transfer, and comprehensive imbalance mitigation integrating focal loss, label smoothing, and weighted sampling techniques.

Our extensive ablation studies reveal several important insights for numerical fact verification. Evidence quantity optimization shows performance saturates at 5 evidence documents, suggesting that additional context introduces more noise than signal. LLM-based summarization consistently improves performance across all evidence lengths and classes, with 35-word summaries providing optimal information density. Cross-lingual benefits demonstrate that multilingual data augmentation not only addresses class imbalance but also enhances model robustness to diverse linguistic expressions of numerical claims. Architecture choices confirm that DeBERTa's enhanced attention mechanisms provide meaningful improvements for numerical reasoning tasks over traditional BERT-based models. These findings establish clear best practices for evidence selection, summarization design, and imbalance mitigation in numerical fact verification tasks.

Limitations and Future Work

While our approach achieves substantial improvements, several limitations warrant discussion. Translation Dependency means the multilingual augmentation strategy relies on neural translation quality, which may introduce artifacts or lose nuanced meaning in numerical expressions. Evidence Corpus Constraints indicate that performance is bounded by the quality and coverage of the BM25-retrieved evidence corpus, suggesting potential for improvements through advanced retrieval methods. Computational Overhead from the summarization pipeline adds inference time, though this may be mitigated through caching strategies in production environments. Domain Generalization concerns arise as evaluation focuses on the CLEF task dataset; broader evaluation across diverse numerical claim domains would strengthen generalisability claims.

Future research directions include investigation of retrieval-augmented generation approaches for evidence synthesis, development of claim decomposition strategies for complex numerical assertions, exploration of few-shot learning techniques for rapid adaptation to new domains, and integration of structured knowledge bases for enhanced numerical reasoning.

Declaration on Generative AI

The authors also employed Claude Sonnet 4 and GPT-4o for initial prompt design and methodology refinement. After using these tools, the authors reviewed and edited all generated content and take full responsibility for the publication's content.

References

[1] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

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

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[4] V. Venktesh, V. Setty, A. Anand, M. Hasanain, B. Bendou, H. Bouamor, F. Alam, G. Iturra-Bocaz, P. Galuščáková, Overview of the CLEF-2025 CheckThat! lab task 3 on fact-checking numerical claims, ????

[5] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, 2018. URL: https://arxiv.org/abs/1803.05355. arXiv:1803.05355.

[6] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, D. Kiela, Adversarial nli: A new benchmark for natural language understanding, 2020. URL: https://arxiv.org/abs/1910.14599. arXiv:1910.14599.

[7] A. Hanselowski, H. Zhang, Z. Li, D. Sorokin, B. Schiller, C. Schulz, I. Gurevych, UKP-athene: Multi-sentence textual entailment for claim verification, in: J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal (Eds.), Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 103–108. URL: https://aclanthology.org/W18-5516/. doi:10.18653/v1/W18-5516.

[8] V. V, A. Anand, A. Anand, V. Setty, Quantemp: A real-world open-domain benchmark for fact-checking numerical claims, 2024. URL: https://arxiv.org/abs/2403.17169. arXiv:2403.17169.

[9] Z. Chen, S. Li, C. Smiley, Z. Ma, S. Shah, W. Y. Wang, ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6279–6292. URL: https://aclanthology.org/2022.emnlp-main.421/. doi:10.18653/v1/2022.emnlp-main.421.

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL: https://arxiv.org/abs/2005.11401. arXiv:2005.11401.

[11] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7534–7550. URL: https://aclanthology.org/2020.emnlp-main.609/. doi:10.18653/v1/2020.emnlp-main.609.

[12] W. Kryściński, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, 2019. URL: https://arxiv.org/abs/1910.12840. arXiv:1910.12840.

[13] K. Popat, S. Mukherjee, A. Yates, G. Weikum, DeClarE: Debunking fake news and false claims using evidence-aware deep learning, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 22–32. URL: https://aclanthology.org/D18-1003/. doi:10.18653/v1/D18-1003.

[14] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, J. G. Simonsen, MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 4685–4697. URL: https://aclanthology.org/D19-1475/. doi:10.18653/v1/D19-1475.

[15] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II, Springer-Verlag, Berlin, Heidelberg, 2021, p. 639–649. URL: https://doi.org/10.1007/978-3-030-72240-1_75. doi:10.1007/978-3-030-72240-1_75.

[16] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Multilingual e5 text embeddings: A

technical report, 2024. URL: https://arxiv.org/abs/2402.05672. arXiv:2402.05672.

[17] M. H. Daniel Han, U. team, Unsloth, 2023. URL: http://github.com/unslothai/unsloth.

[18] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, 2023. URL: https://arxiv.org/abs/2307.03172. arXiv:2307.03172.