LIMBO@GT ARC at CheckThat! 2026: Ensemble Models of Evidence-aware and Verdict Candidate Detector for Numerical Claim Verification

Le Tao[1,*], Sirius Zhong[1]

[1]Georgia Institute of Technology, North Ave NW, Atlanta, GA 30332

Abstract

Automated claim verification system is critical for mitigating the rapid spread of disinformation and misinformation across platforms and languages. Verdict predictions and ranking of reasoning traces for real-world numerical claims Verification presents substantially greater challenges than synthetic claim verification on Wikipedia. Moreover, The theoretical upper bound of the required evaluation metrics imposed by the provided reasoning traces and corresponding verdict lists further exacerbates the difficulty. The LIMBO@GT ARC submission to CheckThat! Task 2 of CLEF 2026 uses the Llama family as the base transformer model and trains the reward verifier model by explicitly incorporating evidence justification into the input representation. This approach further proposes to incorporate a no-gold detector to model the absence of the gold label from the candidate verdict set. The detector learns only features available at training and inference time, such as the verdict distribution. The selected submission achieved an average Recall@5 and Macro Average F1 score of 0.4234, ranking third among 22 effective submissions. Ongoing post-competition work should reconcile conflicting F1 results, systematically compare models trained solely on the provided dataset against those leveraging data augmentation or external knowledge, and explore multilingual claim verification and test-time scaling on large-scale datasets.

Keywords

Numerical Claim Verification, Verdict Distribution, Evidence-aware, Base Transformer Model, No-gold Detector, LIMBO@GT ARC, CheckThat!, CheckThat! Task 2, CLEF 2026

1. Introduction

The rapid spread of disinformation and misinformation across platforms and languages severely undermines trust in online public discourse. Automated fact-checking systems aim to address this issue. While many approaches rely on synthetic claims from Wikipedia [1, 2], recent work has concentrated on numerical claims that naturally involve quantitative and temporal expressions [3, 4, 5]. Automated fact-checking of real-world numerical claims is substantially more challenging than verifying synthetic claims derived from Wikipedia. Task 2 of CheckThat! 2026 addresses this setting by requiring the participants to rank reasoning traces and predict the final verdict based on the provided reasoning traces [6, 7].

In this study, we thoroughly investigate the data distribution and analyze the theoretical upper bound of the required evaluation metrics. We use the Llama family as the base transformer model and trained the reward verifier by explicitly incorporating retrieved evidence into the input representation. We further introduced a no-gold detector as a conservative fallback mechanism. During inference, the detector operates without access to the ground-truth label and is instead trained in a supervised manner to identify whether the gold label is absent from the observed candidate verdict set.

2. Related Work

Learning-to-rank framework [8] formalized ranking as a supervised learning problem for ordering documents by relevance and categorized ranking methods into pointwise, pairwise, and listwise ap-

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

*Corresponding author.

tao8@gatech.edu (L. Tao); zlbeidou@gmail.com (S. Zhong)

proaches.

In the domain of automated fact-checking, FEVER established a large-scale benchmark in which claims are classified as supported, refuted, or not enough information based on retrieved Wikipedia evidence [1]. More recently, QuanTemp [5] extended this setting to real-world numerical claim verification. QuanTemp contains 15,514 fact-checker-validated claims and 423,320 evidence snippets spanning comparative, statistical, interval, and temporal claim types. Its experimental results showed that claim decomposition and numerically specialized natural language inference (NLI) models improve verification performance, while also highlighting the persistent challenges of evidence retrieval and numerical reasoning in automated fact-checking.

Selective Classification for Deep Neural Networks [9] introduced a framework that allows neural networks to abstain on uncertain predictions, thereby trading coverage for lower error rates and enabling users to target a desired risk level with probabilistic guarantees.

"Think Right, Not More" further investigated adaptive test-time scaling for numerical claim verification by dynamically allocating computational resources during inference [10].

3. Data Overview

There are three languages in this competition's datasets, English, Spanish and Arabic. The training and validation sets are 8000 with English, 2808 with Spanish and 3260 with Arabic. It's worth noting that in development phase the training set and validation set are already been splitted fixedly. It was set that way perhaps because it allows the participants to gain some insights in the dev phase submission. In fact, the participants could split the training dataset in a whole. Just be careful with the grid search when splitting.

The training and validation datasets of English are part from QuanTemp, and 8000 in total. Each sample item in training and validation datasets comprises a claim, 15 to 20 reasoning traces list along with corresponding verdict list. Also, the training is provided with annotated ground truth label. And the test dataset contains 2558 sample item without ground truth label.

4. Methodology

As discussed in the Abstract and Introduction, the underlying data distribution is critical to the proposed approach. Accordingly, the Methodology section is divided into two parts: (1) Data Analysis and Preprocessing and (2) the Ensemble Pipeline for Evidence-Aware and Verdict Candidate Detection.

4.1. Data Analysis and Preprocessing

4.1.1. Deduplication

Three claims in the training set appear multiple times, while the validation set contains no duplicated claims. For these duplicate claims, the corresponding reasoning traces and verdict lists all differ. We therefore select and retain the samples that are closer to the ground truth. Regarding evidence, only one training instance contains nearly identical evidence strings that differ solely in date format. As this reflects a difference in formatting conventions, we retain this evidence as is.

Regarding reasoning traces, exact string matching identified 120 duplicates out of 6,400 instances in the training set and 6 out of 1,600 in the validation set. After normalizing whitespace and date formats, the numbers increased to 128 and 7 respectively. Furthermore, the total numbers of superfluous duplicate sentences beyond the first occurrence within each record were 186 in the training set and 7 in the validation set under exact string matching, and 172 and 2 respectively after whitespace and date-format normalization. We retain the reasoning traces unchanged, as this study focuses primarily on the pipeline itself. The data augmentation and external knowledge learning will be discussed in future work section.

In summary, these duplicate instances constitute a negligible portion of the overall corpus. And though the competition submission don't check the query_id, It's worth noting that the query_id is not consistent starting from 0. So the last sample in provide test dataset is marked as 2604.

Table 1
Duplicate reasoning traces within individual records (exact and normalized matching)

Dataset	Exact		Norm.	
	Rec.	Extra	Rec.	Extra
Train	120 / 6400	186	128 / 6400	207
Val	6 / 1600	7	7 / 1600	8

4.1.2. Verdict Distribution and Upper Bound

The distribution of verdict lists by presence of gold candidate shows that approximately 30% of claims do not contain any gold candidate within the corresponding verdict lists of reasoning traces. This indicates a theoretical upper bound on the achievable accuracy when using majority voting or selecting the highest-scoring candidate. Moreover, this issue substantially degrades Macro Average F1 performance.

Table 2
Distribution of verdict lists by presence of gold candidate

Split	Claims	Average candidates	At Least One Gold Candidate	None of Gold Candidate
Train	6400	19.85	70.05%	29.95%
Val	1600	15.00	70.56%	29.44%

Table 3
No-ground-truth rate per reasoning path for training and validation claims

Split	Label	At Least One Gold Candidate Rate	No–gold candidate Rate	Path Positive Rate
train	True	60.94%	39.06%	46.42%
	False	73.66%	26.34%	51.71%
	Conflicting	68.24%	31.76%	39.50%
validation	True	66.12%	33.88%	52.76%
	False	71.52%	28.48%	50.12%
	Conflicting	71.80%	28.20%	43.99%

4.2. Proposed Pipeline of Ensemble Models of Evidence-aware and Verdict Candidates Detector

Using the Llama family as the base transformer architecture, we trained the reward verifier by explicitly incorporating retrieved evidence into the input representation. We further introduced a no-gold detector as a conservative fallback mechanism. During inference, the detector operates without access to the ground-truth label and is instead trained in a supervised manner to identify whether the gold label is absent from the observed candidate verdict set. The overall pipeline, presented in Figure 3, consists of claim complexity analysis, data preprocessing, training, and inference stages and we only experiment on English set in this study. The complexity verifier is inspired by recent advances in adaptive test-time scaling for numerical fact verification. Further details can be found in the paper Think Right, Not More: Test-Time Scaling for Numerical Claim Verification [10].

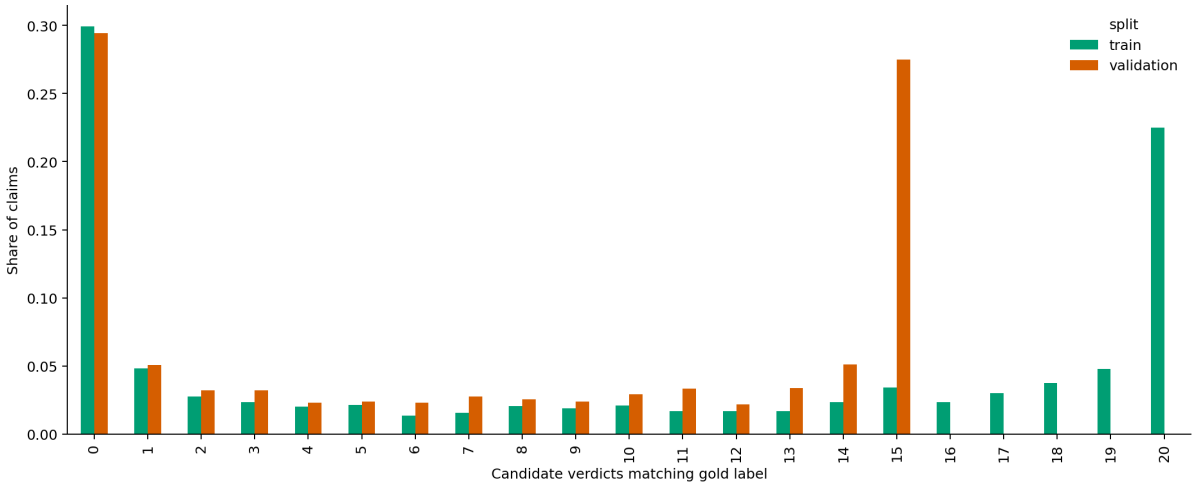

Figure 1: Distribution of Number of Gold Verdict Candidate Paths per Claim

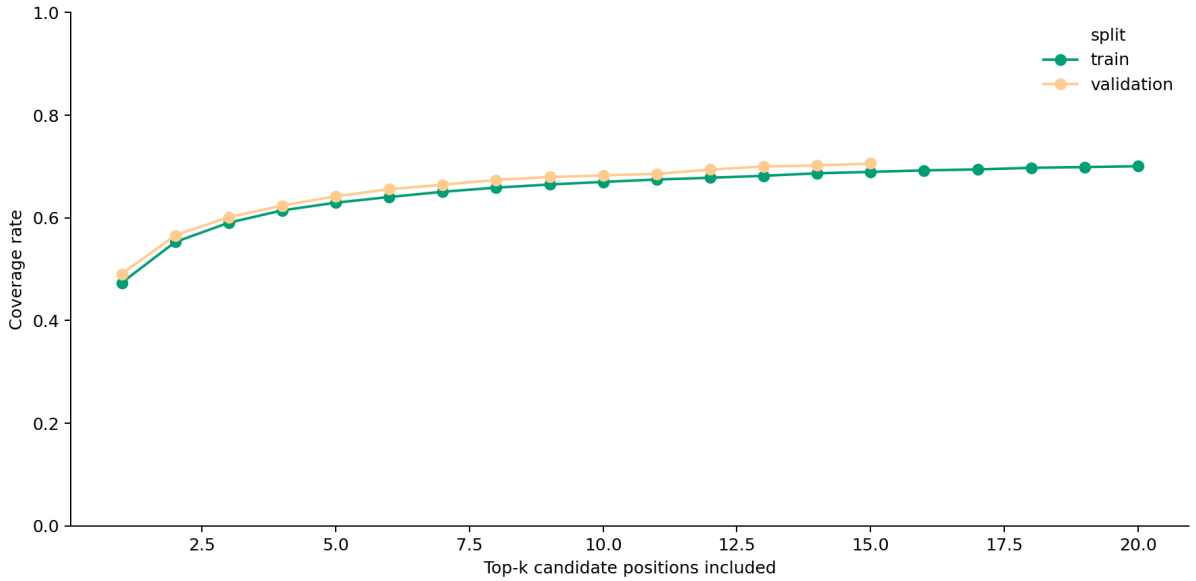

Figure 2: Cumulative Gold Coverage by Candidate Position

4.2.1. Evidence-aware Verification

We constructed training instances by simply using the combination of claim, verdict, and evidence as the input. This allows the model to additionally observes selected evidence snippets associated with the claim which enables direct comparison between the reasoning path and the supporting evidence. The model was then trained on complete reasoning paths for each claim.

4.2.2. No-gold Detector

In the baseline pipeline, the verifier assumes that at least one correct candidate exists and applies the Best-of-(N) (BoN) strategy to select the highest-scoring candidate. However, our analysis reveals that approximately 30% of claims contain no-gold candidate whose verdict matches the provided ground-truth label. In such cases, even a perfect ranker cannot produce the correct verdict prediction because the candidate verdict set itself is defective.

To address this issue, we introduce a no-gold detector as a conservative fallback mechanism. The detector is trained as a supervised binary classifier. During training, each claim is labeled Yes if its

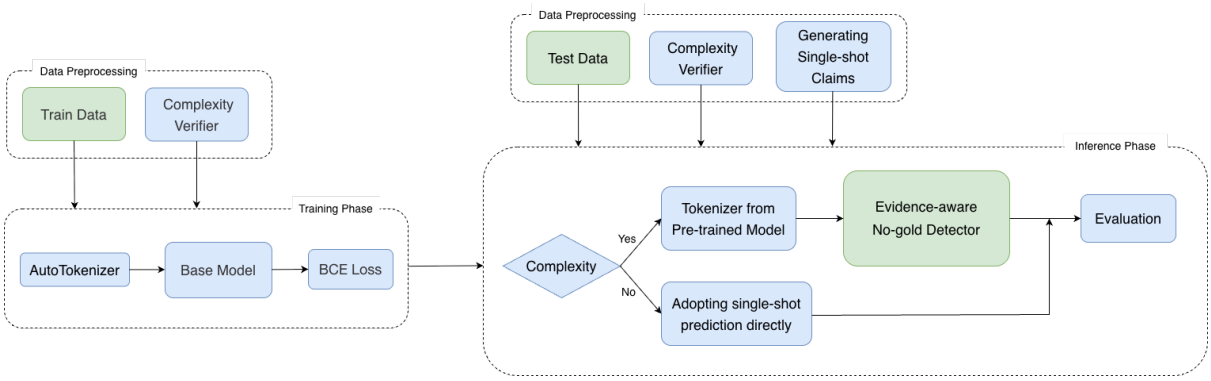

Figure 3: Overview of the Proposed Pipeline for Numerical Fact-checking Pipeline

candidate set contains at least one verdict matching the ground-truth label; otherwise, it is labeled No. Unlike the verifier, which evaluates reasoning traces individually, the detector operates on aggregate information from the entire candidate set of reasoning traces. Specifically, it learns discriminative signals derived from verifier score statistics (e.g., maximum score, score variance, entropy, and the margin between the top-1 and top-2 scores) as well as the distribution of predicted verdicts across candidates.

If the detector labels it as No, all candidate reasoning traces and verdicts are discarded, and the verifier instead predicts the final prediction among False, True, Conflicting directly and also focuses more on the evidences in the input. During inference, the detector requires no access to ground-truth labels. It relies solely on the statistical patterns learned during training to estimate whether the candidate set is likely to contain a correct verdict.

Notably, this mechanism can improve overall accuracy by mitigating failures caused by defective candidate sets. However, it does not improve Recall@5, since it does not alter the candidate generation process or introduce additional correct candidates of reasoning traces.

5. Results and Discussion

5.1. Main Competition Results

The best competition-phase result in this study achieved an average Recall@5 and Macro Average F1 score of 0.4234, ranking third among 22 effective submissions. The Macro Average F1 and Recall@5 were 0.6002 and 0.2466 respectively. And the corresponding F1 scores for True, Conflicting, and False were 0.74, 0.1702, and 0.8903, respectively.

5.2. Experimental Parameters in Evidence-aware Verification

Our primary experiment on evidence-aware verification was conducted using Llama-3.1-8B as the base model. The corresponding experimental settings are summarized in Table 4. The evidence-aware verifier yielded a modest but consistent improvement in Macro Average F1 over the baseline verifier. One plausible explanation is that the additional evidence does not introduce fundamentally new information, but instead provides a stronger grounding signal that helps the verifier assess whether a reasoning path faithfully reflects the original evidence, particularly when the verdict assignments within the reasoning traces are incorrect.

We also experimented on a smaller model, Llama-3.2-1B, but observed inferior performance. This suggests that a larger base model provides better representation capacity and improves downstream verification performance.

In addition, we performed a grid search over key hyperparameters, including the number of training epochs (3, 5, 10, 20), learning rate (5×10^{-4}, 5×10^{-5} to 1×10^{-5}), effective batch size (256, 512, 1024),

Table 4
the Hyperparameter of model training

Parameter	Value
Number of Training Epochs	5
Learning Rate	5×10^{-5}
Effective Batch Size	512
Maximum Sequence Length	1024
LoRA Configurations (Rank/Alpha Pair)	16/32

maximum sequence length (256, 512, 1024, 2048), and LoRA configurations (rank/alpha pairs of 8/16, 16/32, 32/64, and 64/128). The experiment results in figure 4 show substantial degradation in test set, indicating overfitting introduced by extensive hyperparameter tuning. One possible explanation is that the training data contain relatively consistent claim lengths and reasoning patterns, whereas the test set exhibits much greater variability in both claim length and reasoning traces, resulting in reduced generalization.

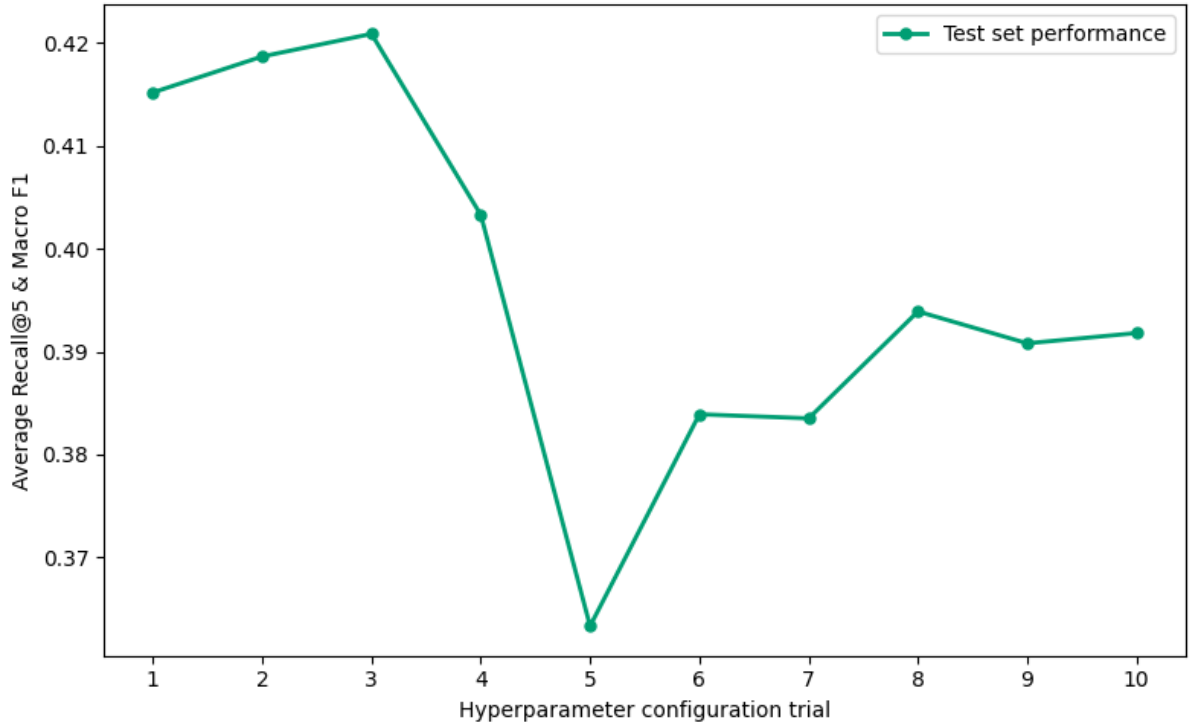

Figure 4: Test Performance of Evidence-aware Verification Across Grid Search Configurations

5.3. No-gold Detector

We further evaluated the no-gold detector using the fixed hyperparameter shown in Table 4. It achieved a macro average F1 score of 0.6002, with a substantial improvement in the F1 score for the true class, reaching 0.74. This result is broadly consistent with our data analysis that approximately 30% of claims in the training set have no gold candidate in the corresponding reasoning-trace verdict lists, and about 39% of these instances are labeled true. However, performance on the conflicting class remained relatively weak, suggesting that additional data patterns may not have been captured.

6. Future Work

The highest-priority direction for future work is to improve the conflicting F1 score. Experiments on data augmentation and external knowledge learning could be conducted in the post-competition analysis. Additional possible directions could be evaluating adaptive test-time scaling for numerical claim verification in larger datasets, as well as integrating claim-complexity prediction with uncertainty- and consistency-based routing [11]. The verifier could also be extended beyond binary BoN re-ranking toward rubric-guided process evaluation by incorporating evidence relevance, arithmetic consistency, decomposition quality, and self-critique signals [12]. Future evaluation settings could be further expanded to stage-wise [13] and multilingual benchmarks.

7. Conclusions

The LIMBO@GT ARC at Task 2 of CheckThat! 2026 trained the reward verifier by explicitly incorporating evidence into the input representation and introduced a no-gold detector as a conservative fallback mechanism. The best competition-phase result in this study achieved a Macro Average F1 of 0.6002 and a Recall@5 of 0.2466. The proposed pipeline yielded modest improvements in Recall@5 and it needs further improvement in metrics of Recall@5 and Conflicting F1.

Acknowledgements

We thank the Data Science at Georgia Tech (DS@GT) CLEF competition group for their support. This research was supported in part through research cyberinfrastructure resources and services provided through the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA [14].

Declaration on Generative AI

During the preparation of this work, the author(s) used X-GPT-5 and Gramby in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: A Large-scale Dataset for Fact Extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 809–819. doi:10.18653/v1/N18-1074.

[2] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task, in: Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), 2021, pp. 1–13. doi:10.18653/v1/2021.fever-1.1.

[3] J. Chen, A. Sriram, E. Choi, G. Durrett, Generating Literal and Implied Subquestions to Fact-check Complex Claims, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 3495–3516. doi:10.18653/v1/2022.emnlp-main.229.

[4] M. S. Schlichtkrull, Z. Guo, A. Vlachos, AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023.

[5] 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.

[6] 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.

[7] V. V, V. Setty, P. Chungkham, A. Anand, F. Alam, Overview of the CLEF-2026 CheckThat! Lab Task 2 on Fact-Checking Numerical Claims, 2026.

[8] T.-Y. Liu, Learning to Rank for Information Retrieval, Foundations and Trends in Information Retrieval 3 (2009) 225–331. doi:10.1561/1500000016.

[9] Y. Geifman, R. El-Yaniv, Selective Classification for Deep Neural Networks, in: Advances in Neural Information Processing Systems 30 (NeurIPS), 2017, pp. 4878–4887.

[10] P. Chungkham, V. V, V. Setty, A. Anand, Think Right, Not More: Test-Time Scaling for Numerical Claim Verification, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025, 2025, pp. 24345–24363. doi:10.18653/v1/2025.findings-emnlp.1322.

[11] H. Wang, M. Khalid, Q. Wu, J. Gao, C. Cao, Fact-Checking with Large Language Models via Probabilistic Certainty and Consistency, 2026. doi:10.48550/arXiv.2601.02574.

[12] Y. Wan, T. Fang, Z. Li, Y. Huo, W. Wang, H. Mi, D. Yu, M. R. Lyu, Inference-Time Scaling of Verification: Self-Evolving Deep Research Agents via Test-Time Rubric-Guided Verification, 2026. doi:10.48550/arXiv.2601.15808. arXiv:2601.15808.

[13] H. Lin, Z. Chen, Z. Shen, Z. Luo, Z. Ye, J. Ma, T.-S. Chua, G. Xu, Towards Comprehensive Stage-wise Benchmarking of Large Language Models in Fact-Checking, 2026. doi:10.48550/arXiv.2601.02669. arXiv:2601.02669.

[14] PACE, Partnership for an Advanced Computing Environment (PACE), 2017. URL: http://www.pace.gatech.edu.