Overview of the CLEF-2025 CheckThat! Lab Task 3 on Fact-Checking Numerical Claims

Notebook for the CheckThat! Lab at CLEF 2025

Venktesh V[1], Vinay Setty[2], Avishek Anand[1], Boushra Bendou[3,*], Maram Hasanain[4], Houda Bouamor[3], Gabriel Iturra-Bocaz[2], Petra Galuščáková[2] and Firoj Alam[3]

[1]*Delft University of Technology, Netherlands*
[2]*University of Stavanger, Norway*
[3]*Carnegie Mellon University in Qatar, Qatar*
[4]*Qatar Computing Research Institute, Qatar*

Abstract

We present an overview of the CheckThat! Lab 2025 Task 3, part of CLEF 2025. The task focuses on verifying claims with numerical quantities and temporal expressions. Numerical claims are defined as those requiring validation of explicit or implicit quantitative or temporal details. It is conducted in three languages: Arabic, Spanish, and English. A total of 258 valid runs were submitted by 13 unique teams across languages, with 4 participants in Spanish and Arabic. 10 teams participated in fact-checking English numerical claims. Among these teams, the use of transformer pre-trained language models (PLMs) was the most frequent. A few teams also employed Large Language Models (LLMs). We provide a description of the dataset, the task setup, including evaluation settings, and a brief overview of the participating systems. As is customary in the CheckThat! Lab, we release all the datasets as well as the evaluation scripts to the research community. This will enable further research on identifying challenges with fact-checking numerical claims that can assist various stakeholders, such as fact-checkers, financial research analysts, and policymakers.

1. Introduction

There has been growing interest in developing tools [1], methods [2], and benchmarks [3, 4] to enhance the fact-checking process. Automating fact-checking is challenging, as many claims are complex and require sophisticated reasoning for accurate validation, especially those involving numerical data. Numerical claims often appear more credible due to the *Numeric-Truth effect* [5], leading to uncritical acceptance. Recent studies show verifying numerical claims is more difficult than non-numerical ones [6, 7]. For example, the social media claim that "CDC quietly deletes 6,000 COVID vaccine deaths from its website" exaggerates a clerical correction, causing unnecessary panic. This demonstrates the need for automated verification of such misleading claims.

This task focuses on verifying claims with numerical quantities and temporal expressions. Numerical claims are defined as those requiring validation of explicit or implicit quantitative or temporal details. Participants must classify each claim as *True, False, or Conflicting* based on a short list of evidence. Each claim is accompanied by the top-100 pieces of evidence retrieved using BM25 from our collection. The collection was carefully curated through pooling evidences from retrieval using different query understanding mechanisms [8, 9]. These evidences can be used after re-ranking to perform claim verification with a classification or generative model that can perform the task of Natural Language Inference (NLI). The objective here is to also evaluate the numerical reasoning capabilities of the claim verification model. The task is available in English, Spanish, and Arabic.

The CheckThat! 2025 lab was held in the framework of CLEF 2025 [10, 11].[1] Figure 1 shows the

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*This work was done during the author's internship at QCRI.
v.viswanathan-1@tudelft.nl (V. V); vsetty@acm.org (V. Setty); avishek.anand@tudelft.nl (A. Anand); bbendou@andrew.cmu.edu (B. Bendou); mhasanain@hbku.edu.qa (M. Hasanain); hbouamor@cmu.edu (H. Bouamor); gabriel.e.iturrabocaz@uis.no (. G. Iturra-Bocaz); petra.galuscakova@uis.no (P. Galuščáková); fialam@hbku.edu.qa (F. Alam)

[1]https://checkthat.gitlab.io

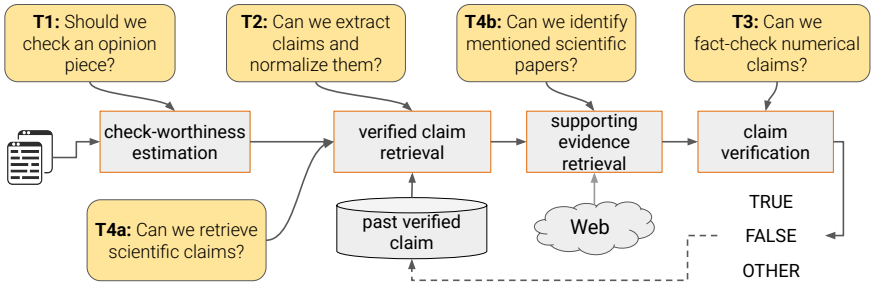

Figure 1: The CheckThat! verification pipeline, featuring the four tasks along with the CheckThat! 2025 tasks. This paper mainly focused on *T3:* fact-checking on numerical factual claims

full CheckThat! identification and verification pipeline, highlighting the four tasks targeted in this seventh edition of the lab: Task 1 on subjectivity, Task 2 on claims extraction & normalization, Task 3 on numerical claims (this paper), and Task 4 on scientific web discourse.

The remainder of the paper is organized as follows: Section 3 describes the datasets released with the task. We present the evaluation setup in Section 4. Section 5 discusses the system submissions and the official results. Section 2 presents some related work. Finally, we provide some concluding remarks in Section 6.

2. Related Work

Automated claim-verification is key to mitigating growing misinformation [12, 13, 2]. Existing works on automated fact-checking primarily focus on synthetic claims collected from Wikipedia [12, 14, 13] that are not representative of real-world claims. More efforts have been made to build systems for real-world claims in domains like politics [8, 15, 16, 17], science [18, 19, 20], health [21] and climate change [22].

To combat real-world misinformation, verifying claims containing numerical information is especially important. Such claims, citing statistics, figures, or time spans are often perceived as more credible, a phenomenon known as the *numeric truth effect* [5]. While real-world fact-checking benchmarks have been proposed, they do not particularly focus on numerical claims [23, 8, 3]. There are synthetic datasets that require tabular data to verify the claims [24, 25, 26], but these claims and tables do not necessarily contain numerical quantities. Recent efforts by [27] aim to create more realistic claims from Wikipedia by identifying cited statements, but these do not reflect the typical distribution of claims verified by fact-checkers.

Among those that focus on simple statistical claims, [28, 29], the authors propose a weak supervision approach using freebase. These claims are not only synthetic in nature, but they can be answered with simple KB facts such as Freebase and does not require numerical understanding or reasoning. Similarly, [30] explore the extraction of formulae for checking numerical consistency in financial statements by also relying on Wikidata.

QuanTemp [31] was the first real-world open-domain benchmark for fact-checking numerical claims. It comprises real-world claim containing quantitative or temporal information and require verification of these numerical information to verify the claim which in turn requires numerical reasoning. The goal of the benchmark was to also test numerical contextualization and numerical reasoning capabilities of transformer based NLI models [32] and LLMs [33, 34]. The benchmark was also constructed in an open-domain setting to improve retrieval and ranking capabilities for fact-checking. In the original work, we demonstrate that LLMs fall short and are ill-suited for verifying such numerical claims when compared to focused fine-tuning of smaller NLI models pre-trained to interpret numerical quantities. This is further extended to other languages like Arabic and Spanish for Task 3 in this iteration of CheckThat!.

The focus on claim verification has been a focus previous editions of the CLEF CheckThat! lab in

2018-2023 [35, 36, 37, 38]. The initial edition [39] proposed Task 2 which dealt with verifying claims made by politicians as part of debates or speeches and was offered in English and Arabic. The data was collected from 2016 US presidential campaign and participants were asked to classify the claims to one of true, false or half-true categories. Subsequent iterations offered a task (Task 3) verification of claims in news articles with associated topics [36] with veracity prediction being on a four point scale: true, false, partially true or other. CLEF CheckThat! 2022 particularly focused on verifying the central claims in news articles [40].

Following this tradition, we offer a claim veracity prediction task but particularly focused on numerical claims across three languages: English, Spanish and Arabic.

Example: Claim decomposition example

Claim: Discretionary spending has increased over 20-some percent in two years if you don't include the stimulus. If you put in the stimulus, it's over 80 percent.

[Decomposition]: [Q1]: Has discretionary spending increased in the past two years?
[Q2]: Does the increase in discretionary spending exclude the stimulus?
[Q3]: Is there evidence to support the claim that

Figure 2
Example for claim decomposition

Example: Claim decomposition example

Claim: A claim is designated as numeric if the numeric aspect is one of the crucial aspects to be verified in the claim. While there culd be other important aspects that may determine if a claim is correct or not, if numerical claim is one of the aspects it is still a numerical claim.
Examples: For instance in the claim **Example 1:** "A man wearing justice for breonna taylor shot and killed 3 men in a retired cops bar."
Here there are several aspects to be verified: that the man was wearing the said t-shirt linking him to BLM protests, he killed 3 men and the act was carried out at retired cops bar.
While there are 3 aspects one crucial aspect is verification of fact that if he killed 3 men. because if this number say was misrepresented it would cause more panic due to spread of such misinformation.
Example 2: "The chattsigarh police conveyed that naxal terrorists were involved in blast on January 22, 1794 " Here while the important aspect to be verified seems to be the terrorist group part, it is also a temporal claim as it is crucial to verify if blasts occured on said date.

Figure 3
Guidelines for identifying numerical claims - manual annotation

3. Datasets

The dataset contains multigenre content in Arabic, English, and Spanish. The dataset is collected from various fact-checking domains through Google Fact-check Explorer API[2], complete with detailed metadata and an evidence corpus sourced from the web. Our pipeline filters out numerical claims for the task. An overview of dataset statistics is shown in Table 1.

[2]https://toolbox.google.com/factcheck/apis

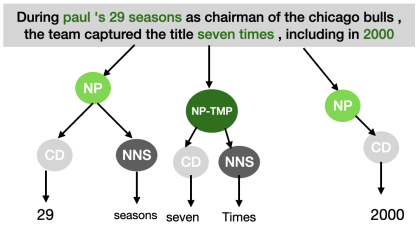

Figure 4: Example of identification of quantitative segments from the claim. NNS is noun plural form, NP-TMP is temporal noun phrase. For Spanish, "NUM" is used as identifier to detect numerical spans instead of "CD".

Extraction of quantitative segments: To extract numerical claims and filter out non-numerical ones, we employ the constituency parse of the original claim as shown in Figure 4. Figure 4 demonstrates how quantitative segments are identified using constituency parse of an English claim.

The nodes with the cardinal number POS tag "CD" are identified from the constituency parse. To avoid false positives (for example: "The one and only"), we then parse these nodes' ancestors and extract noun phrases from their least common ancestors. Using these noun phrases as root nodes, we perform a prefix traversal of their subtrees. We then filter from set of all claims, those with at least one quantitative segment, as numerical claims for our dataset.

This approach still has one limitation, as it may include claims with non-quantitative terms like "Covid-19" or "F-35". To remedy this, we require more than one quantitative segment, excluding any nouns like "Covid-19" mentions, to qualify as a numerical claim. Our self-assessment of 1000 sample claims from English from the dataset indicates a 95% accuracy of our approach.

We follow the same process for Arabic. For extraction of Spanish numerical claims, the process is similar with exception of POS tag used for identifying numerical spans. The POS tag corresponds to "NUM" instead of "CD". We follow a similar manual evaluation process for Spanish and observe an accuracy of **97%** demonstrating that our approach to identify quantitative segments and extract numerical claims is robust across languages.

Data characteristics: We use the train and validation sets from the English dataset released in [6], and also curate Arabic and Spanish claims.

English: For the English test set we collect new real-world English numerical claims additionally to the evaluation set released in [6] to avoid label leakage. We also analyze the distribution of different categories of numerical claims namely: statistical, comparative, interval and temporal with examples shown in Table 2.

Arabic: The Arabic dataset only consists of claims belonging to the categories *True* and *False* for verification, as real-world distribution of conflicting claims for Arabic is too low. The claims in Arabic dataset collected from diverse online sources, including news outlets, social media, and professional fact-checking websites. Our goal was to curate a dataset that captures a wide range of topics such as politics, health, and economics, where numerical misinformation is common. We also incorporated verified claims from AraFacts [41], a large dataset of professionally fact-checked Arabic claims. A distribution of different categories of numerical claims along with the examples in as shown in Table 4.

Spanish: We employ a similar approach adopted for English to filter numerical claims with the exception of change in POS tag used. The guidelines shown in Figure 3 were employed by an expert in the language to verify the correctness of the curated numerical claims. A distribution of different categories of numerical claims along with the examples in as shown in Table 3.

Evidence Collection: Evidence for claims in all languages were obtained from search engines by excluding fact-checking websites to avoid leakage of fact-checker justification and verdict. For each claim, we decompose them to yes/no type sub-questions as shown in Figure 2, and issue the original claim and generated sub-questions as queries to the search engines. Additionally, for English, evidences are also obtained by other decomposition approaches like subclaim generation to increase diversity of evidence pool. All evidences are pooled to form the collection.

Table 1
Dataset statistics for different languages, such as English, Spanish and Arabic.

Split	English				Spanish				Arabic			
	T	F	C	Total	T	F	C	Total	T	F	C	Total
Train	1,824	5,770	2,341	9,935	127	1,200	179	1,506	975	1,216	-	2,191
Dev	617	1,795	672	3,084	30	299	48	377	274	313	-	587
Test	717	2,275	664	3,656	115	1,539	152	1,806	206	276	-	482

Table 2
A broad overview of different categories of claims in QuanTemp (English, full training set) with examples

Category	Examples	#of claims
Statistical	We've got 7.2% unemployment (in Ohio), but when you include the folks who have stopped looking for work, it's actually over 10%.	7302 (47.07%)
Temporal	The 1974 comedy young frankenstein directly inspired the title for rock band aerosmiths song walk this way	4193 (27.03%)
Interval	In Austin, Texas, the average homeowner is paying about $1,300 to $1,400 just for recapture, meaning funds spent in non-Austin school districts	2357 (15.19%)
Comparison	A vaccine safety body has recorded 20 times more COVID jab adverse reactions than the government's Therapeutic Goods Administration.	1645 (10.60%)

Table 3
A broad overview of different categories of claims in QuanTemp (Spanish full training set) with examples

Category	Examples	#of claims
Statistical	Una pareja senegalesa inventarse 18 hijos y defraud más de 360.000 euros en ayudas sociales en España.	790 (52.45%)
Temporal	Dina Boluarte está promocionando una plataforma de inversión desarrollada por Elon Musk a S/ 950 en abril de 2024	334 (22.17%)
Interval	Es peligroso vacunar contra el covid 19 a adolescentes entre 12 y 15 años	232 (15.40%)
Comparison	Homicidios de ex-FARC se redujeron 10,8 % en 2020 frente a 2019.	150 (9.96%)

4. Evaluation Settings

The numerical claim verification task primarily involved classifying the claim given the evidence pool to one of the three classes True, False or Conflicting by verifying aspects of the claim against evidence. While it was posed as a three-way classification task for English and Spanish, it was a two-way /True or False) classification task for Arabic.

Table 4
A broad overview of different categories of claims in QuanTemp (Arabic) for a subset of training set (776 claims) with examples

Category	Examples	#of claims
Statistical	الرئيس التركي أردوغان يتبرع بـ 500 ألف اختبار سريع لفيروس كورونا لتونس.	259 (33.37%)
Temporal	استقر سعر صرف العملات العربية اليوماستقر سعر صرف العملات العربية اليوم الخميس 13 أكتوبر 2022 ، في البنك المركزي المصري، وفقا لأخر تحديث.	337 (43.42%)
Interval	أكد رئيس حكومة الوحدة الوطنية عبد الحميد الدبيبة أن سعر أضحية عيد الأضحى المبارك سيكون ما بين 400 إلى 600 دينار.	160 (20.62%)
Comparison	انخفض معدل سوء التغذية في الأردن إلى النصف بين عامي ٢٠٠٠ و٢٠١٥، ليصل إلى ٤	20 (2.58%)

4.1. Metrics

We ranked the participants based on the Macro-F1 scores and not accuracy to account for class imbalance. Additionally, participants were also asked to report per-class F1 scores to measure if the systems performed reasonable well across different veracity classes.

4.2. Evidence and Retrieval Setup

Apart from the evidence collection mentioned in Section 3, we also provide the evidences from first-stage retrieval. For each claim we retrieve top-100 evidences using BM25 based on Elasticsearch implementation. Top-100 evidences were independently provided for each decomposition approach to offer flexibility for participants to use any decomposition approach.

Participants were encouraged to apply any re-ranking method to the provided evidence. They were also allowed to use their own retrieval approaches over the full document collection, as the BM25 results were offered solely for convenience. For claim verification, participants were free to use any model of their choice.

5. Results and Overview of the Systems

5.1. Results

Table 5 shows the performance of the official submissions on the test set. The official run was the last valid blind submission by each team. The table shows the runs ranked on the basis of the macro-F1 and

includes all three languages.

A total of 258 valid runs were submitted by 13 unique teams across all languages. Among them, 10 teams participated in fact-checking English numerical claims, while 4 teams submitted runs for Spanish and Arabic. The final leaderboard in Table 5 includes 12 teams across languages, as one of the submissions was withdrawn from the leaderboard.

English: A total of 10 teams participated in fact-checking numerical claims in English. Most of the teams employed BM25 evidence followed by re-ranking using cross-encoder. Several teams employed creative approaches like data augmentation by translation from claims in Arabic and Spanish. At least 6 teams employ LLM based approach to reason over the evidence and provide claim veracity predictions. Only one of the teams managed to outperform the baseline reported in [31]. Only the system from LIS outperformed the strongest baseline from the original English benchmark [31] though this result does not seem to be statistically significant.

Arabic: Four teams participated in Arabic with LIS being the top performing system with surprisingly high macro-F1 of 96.15 %. This system is the same as their system employed for English claims. Since they employ LLMs with ability to handle multiple languages, their solution generalizes beyond English. The high gains in Arabic can also be explained by the task not having a Conflicting task, and it's reduction to binary classification of True or False compared to English and Spanish. The main advantage of system from LIS over other systems which underperform in Arabic seems to be improvement of retrieval using dense retrieval approaches like **Linq-Embed-Mistral** and fine-tuning LLMs instead of prompting for veracity prediction / NLI.

Spanish: We observe a trend in Spanish similar to that of English. Spanish proves to be much harder due to the three-way classification task, where conflicting is hard to detect due to multi-aspect nature of the claim and granularity of reasoning required to identify half-true aspects of the original claim. While LIS team outperforms other teams, they only manage to attain a macro-F1 of **50.34**. This could also be explained by lack of MTEB evaluations in Spanish for Linq-Embed-Mistral. Hence, better alternatives to this retrieval model could exist. It also highlights the need for improving numerical reasoning and parsing capabilities of LLMs when employed for claim verification. As very few benchmarks exist for Spanish fact-checking and almost none for numerical fact-checking, the drop in performance could be explained by lack of sufficient training data for fine-tuning the LLM employed for the veracity prediction part.

5.2. Overview of the Systems

Among all participating teams, LIS was the top performer across all languages. TIFIN, NGU_Research, DS@GT-CheckThat! performed well in the respective languages they participated. Most teams employed generative models like gpt-4o-mini or Qwen LLMs to decompose claims, followed by BM25 based retrieval for retrieving evidence and transformer based cross-encoder models for re-ranking the evidences. For claim verification, fine-tuned transformer based NLI models were employed by some teams where transformers were trained as discriminative models on the training sets provided. Some teams employed prompting based approaches to leverage LLMs like gpt-4o-mini or reasoning models like deepseek-r1 to perform claim verification.

Team **LIS** [42] used QwQ-32B to generate question followed by Linq-Embed-Mistral to retrieve evidence from the corpus by combining the questions and claims. Mistral-Small-24B-Instruct-2501 was fine-tuned to obtain the final veracity labels. The Qwen model seem to overcome certain limitations associated with GPT-3.5 and GPT-4 series models used in baselines [6]. Particularly LIS demonstrated that LLMs can be improved on task of claim verification through fine-tuning and by employing reasoning based LLMs to improve claim decomposition.

Team **DS@GT-CheckThat!** [43] performed pre-processing to normalize the number and dates of the claims and decomposed questions from these claims. They employed GPT-4o-mini to decompose the claims. BM25 was employed for first stage retrieval to prioritize documents relevant to the claim and sub-questions. This is followed by re-ranking the documents using cross-encoder/ms-marco-MiniLM-L-12-v2 or mixedbread-ai/mxbai-rerank-large-v1. The main workhorse model for the veracity

Table 5
Task 3: Overview of the approaches for fact-checking numerical claims.

Team	Rank			Model												Macro-F1		
	Arabic	Spanish	English	BM25	cross-encoder	gpt-4o-mini	Qwen	Llama	DeepSeek	ModernBERT	Math-Roberta	RoBERTa-base	QWQ-32B	Qwen-8B	Deberta-Large-MNLI	mxbai-rerank-large-v1	granite-3.3-8b-instruct	Arabic	Spanish	English
LIS [42]	1	1	1										☑					**96.15**	**50.34**	**59.54**
DS@GT-CheckThat! [43]			3															-	-	52.10
TIFIN [44]	3		2	☑											☑		☑	55.36		55.70
ClaimIQ [45]			9					☑										-	-	42.43
FraunhoferSIT [46]			4		☑			☑				☑						-	-	51.00
NGU_Research [47]	2	3		☑	☑				☑									63.52	24.41	-
JU_NLP	4		5	☑	☑													36.38	-	48.83
CornellNLP [48]			6	☑		☑		☑										-	-	48.57
UGLPN [49]			7	☑								☑						-	-	45.53
UCOM_UNAM_PLN [50]		2		☑		☑												-	35.95	-
News-polygraph*			8	☑	☑				☑									-	-	42.86
KSU			10	☑	☑		☑											-	-	25.57

classification was ModernBERT - an optimized model based on the BERT architecture, that can natively support longer sequence length.

Team **TIFIN** [44] employed inverse class weighting in the claim verification step to address class imbalance and give greater importance to minority classes. They also used strategies such as oversampling to balance training examples and label smoothing to prevent the model from becoming overconfident in its predictions. Additionally, they incorporated Focal Loss to fine-tune the verification model *microsoft/deberta-large-mnli* using LoRA, allowing the model to focus on harder examples. To further enhance performance, they used the *ibm-granite/granite-3.3-8b-instruct* model to summarize contexts before feeding them to the verification model. An interesting insight presented by the authors is that performance does not scale linearly with model size – smaller LLMs outperformed 70B-scale LLaMA models, challenging assumptions based on scaling laws. Their approach also demonstrated the benefits of multilingual data augmentation in improving claim verification performance.

Team **NGU_Research** [47] employed a hybrid retrieval approach, experimenting with various pretrained encoder-based models – including BGE, E5, and Gemini – as embedding models. They ultimately selected pretrained embeddings from OpenAI's **text-embedding-3-large** model, combined with BM25 filtering using Qdrant database collections for each language. For claim verification, they used DeepSeek and GPT-4o-mini on the retrieved evidence.

Team **ClaimIQ** [45] presented their core approach of fine-tuning LLMs using Low-Rank Adaptation (LoRA) for the task of claim verification, combined with existing retrieval and ranking strategies to procure evidence. The authors observed that this approach outperformed prompting-based methods and other NLI models on the validation set. However, these performance gains did not generalize to the evaluation set, highlighting that fine-tuning larger LLMs can lead to overfitting. This further underscores the challenging nature of the task, which requires models not only to classify but also to perform numerical contextualization and reasoning.

Team **Fraunhofer_SIT** [46] follows a three-stage architecture: (1) evidence candidate retrieval using dense vectors, (2) re-ranking using a fine-tuned cross-encoder, and (3) final claim classification

using a large NLI model (Roberta-large-MNLI). The authors first pre-compute document embeddings using sentence-transformers/all-MiniLM-L6-v2 and store them in a FAISS index to enable efficient nearest-neighbor retrieval. At inference time, the top-100 candidate evidence snippets are retrieved for each claim. These are then re-ranked using a custom cross-encoder based on cross-encoder/ms-marco-MiniLM-L-6-v2, fine-tuned in a contrastive learning setup detailed as follows.

To improve re-ranking, the authors opted not to use the vanilla MS MARCO model employed in the baseline. Instead, they leveraged weak supervision by using gold-labeled evidence snippets for training and validation claims. These gold snippets were summarized using LLaMA 3.1-8B to remove irrelevant or noisy content, resulting in cleaner positive examples. The cross-encoder was fine-tuned using a contrastive approach, where the claim served as the anchor, the summarized gold evidence as the positive, and the top-100 BM25-retrieved documents as negatives. Although they ranked fifth on the leaderboard, their approach demonstrates that **enhancing retrieval** can significantly boost downstream NLI performance without relying on expensive LLM-based methods for claim verification.

The **KSU** team employed BM25 evidence with cross-encoder based re-ranking. Additionally, since evidence snippets are long, they employed a *unsloth/Qwen3-8B-bnb-4bit* model to filter out most important snippets. This was followed by a fine-tuned NLI model over the snippets for veracity prediction.

However, even the top performing system from LIS did not reach the best possible performance attainable using the provided evidence collection, as outlined in [31]. This reinforces the observation in [31] that LLMs struggle to contextualize and accurately interpret numerical information in claims and evidence. It highlights the challenging nature of the task, which requires reasoning over mixed modalities of numerical and textual data, the ability to contextualize and compare numerical values, and performing numerical reasoning for claim verification. These findings also demonstrate that the task is **far from being solved**.

6. Conclusion and Future Work

We presented an overview of Task 3 of the CLEF 2025 `CheckThat!` lab, which focused on fact-checking numerical claims. Approaches ranged from prompting Large Language Models (LLMs) to fine-tuning open-source LLMs of smaller parameter scales. Some participants also focused on improving evidence re-ranking and demonstrated this can improve claim verification performance with smaller transformers models. Some participants also employed creative data enrichment techniques by translating claims from other languages to train the NLI model on a larger, well-balanced augmented dataset. While some effort was made on improving loss functions used to train the NLI model this did not yield any statistically significant improvements. Only the top-1 result in English leaderboard outperformed the strong baseline reported by organizers in [31]. However, the gains of this system over the baseline was not statistically significant and also falls short of the upper bound by a large margin reported in [31]. This demonstrates that numerical fact-checking is not yet solved and might require exploration of improving LLMs or NLI models capabilities to perform numerical contextualization and reasoning. It also demonstrates a need to scale and improve reasoning capabilities of LLMs which could be the focus on future iterations of CLEF `CheckThat!`.

7. Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check and rewording some text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content. No other generative AI or model was used in preparation of this paper.

Acknowledgments

The work of F. Alam and M. Hasanain is partially supported by NPRP 14C-0916-210015 from the Qatar National Research Fund, part of Qatar Research Development and Innovation Council (QRDI).

References

[1] V. Setty, Factcheck editor: Multilingual text editor with end-to-end fact-checking, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2744–2748.

[2] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[3] 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: 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.

[4] M. Schlichtkrull, Z. Guo, A. Vlachos, Averitec: A dataset for real-world claim verification with evidence from the web, in: A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 65128–65167. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/cd86a30526cd1aff61d6f89f107634e4-Paper-Datasets_and_Benchmarks.pdf.

[5] N. Sagara, Consumer understanding and use of numeric information in product claims, University of Oregon, 2009.

[6] V. Venktesh, A. Anand, A. Anand, V. Setty, Quantemp: A real-world open-domain benchmark for fact-checking numerical claims, in: 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Association for Computing Machinery (ACM), 2024, pp. 650–660.

[7] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, FEVEROUS: fact extraction and verification over unstructured and structured information, in: J. Vanschoren, S. Yeung (Eds.), Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021.

[8] J. Chen, A. Sriram, E. Choi, G. Durrett, Generating literal and implied subquestions to fact-check complex claims, 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. 3495–3516. URL: https://aclanthology.org/2022.emnlp-main.229/. doi:10.18653/v1/2022.emnlp-main.229.

[9] L. Pan, X. Wu, X. Lu, A. T. Luu, W. Y. Wang, M.-Y. Kan, P. Nakov, Fact-checking complex claims with program-guided reasoning, arXiv preprint arXiv:2305.12744 (2023). URL: https://arxiv.org/abs/2305.12744. arXiv:2305.12744.

[10] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. V., The clef-2025 checkthat! lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 467–478.

[11] F. Alam, J. M. Struß, T. Chakraborty, S. Dietze, S. Hafid, K. Korre, A. Muti, P. Nakov, F. Ruggeri, S. Schellhammer, V. Setty, M. Sundriyal, K. Todorov, V. Venktesh, Overview of the CLEF-2025 CheckThat! Lab: Subjectivity, fact-checking, claim normalization, and retrieval, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina,

G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.
[12] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.
[13] 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: R. Aly, C. Christodoulopoulos, O. Cocarascu, Z. Guo, A. Mittal, M. Schlichtkrull, J. Thorne, A. Vlachos (Eds.), Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Dominican Republic, 2021, pp. 1–13. URL: https://aclanthology.org/2021.fever-1.1/. doi:10.18653/v1/2021.fever-1.1.
[14] Y. Jiang, S. Bordia, Z. Zhong, C. Dognin, M. Singh, M. Bansal, HoVer: A dataset for many-hop fact extraction and claim verification, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 3441–3460. URL: https://aclanthology.org/2020.findings-emnlp.309. doi:10.18653/v1/2020.findings-emnlp.309.
[15] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. URL: https://aclanthology.org/P17-2067. doi:10.18653/v1/P17-2067.
[16] T. Alhindi, S. Petridis, S. Muresan, Where is your evidence: Improving fact-checking by justification modeling, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 85–90. URL: https://aclanthology.org/W18-5513. doi:10.18653/v1/W18-5513.
[17] W. Ostrowski, A. Arora, P. Atanasova, I. Augenstein, Multi-hop fact checking of political claims, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 3892–3898. URL: https://doi.org/10.24963/ijcai.2021/536. doi:10.24963/ijcai.2021/536, main Track.
[18] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, H. Hajishirzi, Fact or fiction: Verifying scientific claims, in: 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.
[19] J. Vladika, F. Matthes, Scientific fact-checking: A survey of resources and approaches, 2023. arXiv:2305.16859.
[20] D. Wright, D. Wadden, K. Lo, B. Kuehl, A. Cohan, I. Augenstein, L. L. Wang, Generating scientific claims for zero-shot scientific fact checking, 2022. arXiv:2203.12990.
[21] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, 2020. arXiv:2010.09926.
[22] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, M. Leippold, Climate-fever: A dataset for verification of real-world climate claims, 2021. arXiv:2012.00614.
[23] M. Schlichtkrull, Z. Guo, A. Vlachos, Averitec: A dataset for real-world claim verification with evidence from the web, 2023. URL: https://arxiv.org/abs/2305.13117. arXiv:2305.13117.
[24] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, W. Y. Wang, Tabfact: A large-scale dataset for table-based fact verification, 2020. arXiv:1909.02164.
[25] R. Aly, Z. Guo, M. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, Feverous: Fact extraction and verification over unstructured and structured information,

2021. arXiv:2106.05707.
[26] X. Lu, L. Pan, Q. Liu, P. Nakov, M.-Y. Kan, SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 7787–7813. URL: https://aclanthology.org/2023.emnlp-main.483. doi:10.18653/v1/2023.emnlp-main.483.
[27] R. Kamoi, T. Goyal, J. D. Rodriguez, G. Durrett, Wice: Real-world entailment for claims in wikipedia, 2023. arXiv:2303.01432.
[28] A. Vlachos, S. Riedel, Identification and verification of simple claims about statistical properties, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 2596–2601. URL: https://aclanthology.org/D15-1312. doi:10.18653/v1/D15-1312.
[29] J. Thorne, A. Vlachos, An extensible framework for verification of numerical claims, in: Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 37–40. URL: https://aclanthology.org/E17-3010.
[30] Y. Cao, H. Li, P. Luo, J. Yao, Towards automatic numerical cross-checking: Extracting formulas from text, in: Proceedings of the 2018 World Wide Web Conference, WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018, p. 1795–1804. URL: https://doi.org/10.1145/3178876.3186166. doi:10.1145/3178876.3186166.
[31] V. V, A. Anand, A. Anand, V. Setty, Quantemp: A real-world open-domain benchmark for fact-checking numerical claims, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 650–660. URL: https://doi.org/10.1145/3626772.3657874. doi:10.1145/3626772.3657874.
[32] J. Zhang, Y. Moshfeghi, Elastic: Numerical reasoning with adaptive symbolic compiler, 2022. URL: https://arxiv.org/abs/2210.10105. arXiv:2210.10105.
[33] M. Akhtar, A. Shankarampeta, V. Gupta, A. Patil, O. Cocarascu, E. Simperl, Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data, 2023. URL: https://arxiv.org/abs/2311.02216. arXiv:2311.02216.
[34] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[35] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeño, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, P. Nakov, Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media, CEUR Workshop Proceedings, 2020.
[36] P. Nakov, G. D. S. 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, volume 12657 of *ECIR '21*, 2021, pp. 639–649.
[37] P. Nakov, G. Da San Martino, F. Alam, S. Shaar, H. Mubarak, N. Babulkov, Overview of the CLEF-2022 CheckThat! lab task 2 on detecting previously fact-checked claims, in: N. Faggioli, Guglielmo andd Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022, Bologna, Italy, 2022.
[38] P. Nakov, F. Alam, G. Da San Martino, M. Hasanain, R. N. Nandi, D. Azizov, P. Panayotov, Overview of the CLEF-2023 CheckThat! lab task 4 on factuality of reporting of news media, in: M. Aliannejadi, G. Faggioli, N. Ferro, Vlachos, Michalis (Eds.), Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF 2023, Thessaloniki, Greece, 2023.
[39] P. Nakov, A. Barrón-Cedeño, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, G. Martino, Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims: 9th International Conference of the CLEF Association, CLEF

2018, Avignon, France, September 10-14, 2018, Proceedings, 2018, pp. 372–387. doi:10.1007/978-3-319-98932-7_32.

[40] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouani, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvåg, V. Setty (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2022, pp. 416–428.

[41] Z. Sheikh Ali, W. Mansour, T. Elsayed, A. Al-Ali, AraFacts: the first large arabic dataset of naturally occurring claims, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021, pp. 231–236.

[42] 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: [51], 2025.

[43] M. Heil, A. Pramov, Ds@gt at checkthat! 2025: Evaluating context and tokenization strategies for numerical fact verification, in: [51], 2025.

[44] B. Hazarika, P. Devadiga, P. K. Rajpoot, A. U. Baliga, K. Gurumurthy, M. Jain, M. Sharma, A. Shrivastava, A. Suneesh, A. B. Suresh, Tifin at checkthat! 2025: X-verify - multi-lingual nli-based fact checking with condensed evidence, in: [51], 2025.

[45] A. S. Anik, M. F. K. Chowdhury, A. Wyckoff, S. R. Choudhury, ClaimIQ at CheckThat! 2025: Comparing prompted and fine-tuned language models for verifying numerical claims, in: [51], 2025.

[46] A. Runewicz, P. M. Ranly, I. Vogel, M. Steinebach, Fraunhofer SIT at CheckThat! 2025: Multi-instance evidence pooling for numerical claim verification, in: [51], 2025.

[47] 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, in: [51], 2025.

[48] L. Duesterwald, A. Arora, C. Cardie, CornellNLP at CheckThat! 2025:hybrid llama–gpt-4 ensembles with confidence filtering for numerical claim verification, in: [51], 2025.

[49] M. d. C. Toapanta Bernabé, M. A. García Cumbreras, L. A. Ureña López, D. Mora, UGPLN at CheckThat! 2025: Meta-ensemble transformers for numerical claim verification in spanish, in: [51], 2025.

[50] G. Acosta, E. Morales, H. Gómez-Adorno, UCOM_UNAM_PLN @ Checkthat 2025: Evaluating llms in a two-step architecture for numerical fact checking, in: [51], 2025.

[51] G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.