# Overview of the CLEF-2025 CheckThat! Lab Task 2 on Claim Normalization

Notebook for the CheckThat! Lab at CLEF 2025

Megha Sundriyal[1], Tanmoy Chakraborty[2] and Preslav Nakov[3]

[1]Indraprastha Institute of Information Technology Delhi, India
[2]Indian Institute of Technology Delhi, India
[3]Mohamed bin Zayed University of Artificial Intelligence, UAE

**Abstract**

We present an overview of Task 2 from CheckThat! at CLEF 2025, which focuses on claim normalization. The tasks asks systems to transform informal and often noisy social media posts into clear, concise, and verifiable statements known as normalized claims, which capture the core factual assertion of a post, which makes them much easier to verify and fact-check. The task is especially relevant in multilingual and low-resource contexts, where the diversity of languages and limited labelled data pose serious challenges. Task 2 was conducted in two distinct settings: (i) monolingual, where systems were trained and tested on the same language, and (ii) zero-shot, where models had to normalize claims in a new target language without any in-language training data. The monolingual track covered thirteen languages, including English, German, French, Spanish, Portugese, Hindi, Marathi, Punjabi, Tamil, Arabic, Thai, Indonesian, and Polish. While the zero-shot setting introduced seven more languages, such as Dutch, Romanian, Bengali, Telugu, Korean, Greek, and Czech. This structure allowed us to evaluate both language-specific performance and cross-lingual generalization. In total, 18 teams participated in Task 2, submitting 1,226 valid runs across the two settings. The submissions were evaluated using the METEOR score. Many teams leveraged transformer-based models, multilingual embeddings, and retrieval-augmented strategies. In this paper, we outline the task setup, give details about the datasets, and provide a detailed summary of the diverse approaches adopted by the participating teams.

**Keywords**

Claim Normalization, Social Media Posts, Multilinguality, Claims



## 1. Introduction

Social media have revolutionized global communication, removing geographical barriers and allowing global knowledge exchange. However, it has also become a breeding ground for misinformation, spreading false claims quickly across languages and cultures [1]. These false claims jeopardize the integrity of online discourse and public trust. For instance, they have affected various critical events, including the 45th US Presidential Election [2], the COVID-19 pandemic [3, 4], the Russia–Ukraine conflict [5], etc. While journalists and fact-checkers work tirelessly to ensure the accuracy of online content, the sheer volume and the linguistic diversity of social media posts make it difficult to identify and debunk every single claim across diverse languages effectively [6]. In recent years, several studies have examined the needs of fact-checkers and have identified tasks that could be automated to reduce their manual efforts and to improve the effectiveness of their work [7, 8, 9, 10]. These tasks include looking for the source of evidence for verification [11], exploring other versions of misinformation [12], and searching within existing fact-checking datasets [13].

Social media posts are often written in vague, informal language, frequently mixing opinions, using rhetorical questions, and incomplete thoughts. This makes it difficult to extract clear, check-worthy claims, defined as factual statements that can be verified or disproven [10]. Recently, Sundriyal et al. [14] introduced the task of claim normalization, which aims to simplify a given text containing a claim, such as a long, noisy social media post, into a concise and precise statement.

*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*
meghas@iiitd.ac.in (M. Sundriyal); Preslav.Nakov@mbzuai.ae (P. Nakov)
0000-0002-2268-0137 (M. Sundriyal); 0000-0002-0210-0369 (T. Chakraborty); 0000-0002-3600-1510 (P. Nakov)

The task is a precursor to fact-checking, distilling the essence of the claim and removing any unnecessary information, thereby increasing the efficiency and reliability of the fact-checking process.

Despite efforts to combat misinformation across a variety of languages [15, 16, 17], research into claim normalization has been predominantly English-centred. The Task 2 from CheckThat! at CLEF 2025 aims to bridge this gap by offering the task in a multi-lingual setting.

The CheckThat! Lab aims to accelerate the development of tools and datasets that enable different phases of the fact-checking pipeline. Since its beginning, the lab has organized several shared tasks that represent real-world issues in misinformation detection and verification, with a focus on multilingual, cross-domain, and practical applications. The 2025 edition of the lab included four tasks in monolingual, multilingual, and cross-lingual settings, covering over 20 languages across these tasks [18]. This paper presents Task 2 on Claim Normalization, which addresses the problem of converting informal, noisy social media posts into clear, concise, and verifiable claims. The task plays a vital role in bridging unstructured content with structured fact-checking workflows, especially in multilingual and low-resource settings.

**Task Description.** In this year's CheckThat! Lab, Task 2 addressed the growing need to extract verified claims from the informal language found on social media. Unlike standard fact-checking pipelines that rely on well-formed input, our task aimed to rewrite user-generated content—often imprecise, opinionated, or fragmented—into clear, concise, and factual statements, the way that human fact-checkers formulate the claims they are checking.

The task is especially timely and relevant in multilingual and low-resource settings. To simulate realistic fact-checking scenarios, Task 2 was conducted in two settings:

- *Monolingual*: In the monolingual setting, training, development datasets are provided for the language used for testing. The model is trained, validated, and tested on the same language, allowing it to learn language-specific structures and patterns. The languages included in this setup are English, German, French, Spanish, Portugese, Hindi, Marathi, Punjabi, Tamil, Arabic, Thai, Indonesian, and Polish.
- *Zero-shot*: The zero-shot setting provides only the test data for the target language, without any corresponding training or development data for it. The participants may train their models using data from other languages or conduct zero-shot experiments with large language models (LLMs), evaluating the performance in the target language without prior exposure. This setup tests the model's ability to generalize to unseen languages. The languages in this setting are Dutch, Romanian, Bengali, Telugu, Korean, Greek, and Czech.

In the following sections, we give details about our dataset, a detailed overview of the participating systems, and discussion of the approaches.

## 2. Related Work

Fact-checking is critical for combating the spread of false claims. As fully automating manual fact-checking is very time-consuming, researchers have worked on specific subtasks that can help human fact-checkers. This encompasses a spectrum of tasks, including claim detection [10, 19], claim check-worthiness assessment [20, 21], claim span identification [8, 17], claim verification [22, 23], etc.

The proliferation of false claims on social media platforms has led to the development of specialized systems tailored for handling informal texts from these platforms [24, 25, 26]. These systems are designed to quickly identify and debunk potentially misleading information, allowing for timely intervention by human fact-checkers. Within the larger context of fact-checking, claim normalization has recently emerged as an important novel research direction. Sundriyal et al. [14] introduced this task of claim normalization, which distils the key claim from long noisy social media posts.

Most existing methods aimed at combating misinformation have primarily focused on English [14, 26, 25]. However, there has been a recent surge in interest regarding the advancement of fact-checking techniques for various languages. Jaradat et al. [15] developed ClaimRank, an online system to identify sentences with credible claims in Arabic and English. Gupta and Srikumar [27] developed X-FACT, a multilingual dataset for factual verification of real-world claims across 25 languages. Mittal et al. [17] released X-CLAIM, a multilingual dataset for claim span identification, consisting of 7,000 real-world claims collected from various social media platforms in five Indian languages and English. Pikuliak et al. [28] introduced MultiClaim, a multilingual dataset for detecting previously checked claim retrieval. They gathered 28k social media posts in 27 languages, 206k professional fact-checks in 39 languages, and 31k connections between these two groups. Chang et al. [29] introduced a multilingual version of the FEVER dataset. Over the past seven years, the CheckThat! Lab organized several multilingual claim-related tasks as part of CLEF, gradually expanding language support and attracting an increasing number of submissions [30, 31, 32, 9, 33, 34]. The most recent edition of the CheckThat! lab included six tasks in fifteen languages, including Arabic, Bulgarian, English, Dutch, French, Georgian, German, Greek, Italian, Polish, Portuguese, Russian, Slovene, Spanish, and code-mixed Hindi-English [34].

Despite the growing interest in fact-checking across multiple languages, the task of claim normalization has been largely unexplored beyond English [14]. This narrow focus presents challenges as multilingual social media platforms host content in multiple languages, and thus claims originate in many languages. Moreover, linguistic nuances and cultural contexts complicate the task, emphasizing the need for multilingual approaches. This motivated our multilingual claim normalization task this year.

## 3. Dataset

Below, we describe the dataset for our tasks, which we call mCLAN.

### 3.1. Data Compilation

Inspired by the principle of *dataset recycling* Koch et al. [35], we identified and reused four datasets, which we repurposed for the task of claim normalization. This reduces annotation effort as well as subjective annotation biases. Below, we describe each dataset in detail:

*(a) CLAN* [14]: It contains 6,388 social media posts, each with normalized claims from various fact-checking websites. Notably, every example in the dataset is in English. We use all the pairs of a post and its corresponding normalized claim.

*(b) MultiClaim* [28]: It contains multilingual fact-checking pairs obtained from 142 fact-checking sites, making it the largest dataset of fact-checks released to date, encompassing 39 languages. Each fact-checking article is represented in the dataset by its claim, title, publication date, and URL. However, the entire text of the articles has not been published. In addition, the dataset includes relevant social media posts with text, OCR of attached images (if any), publication date, social media platform, and fact-checker rating for each post. We used this dataset to collect claims from fact-checking websites and corresponding social media posts for our study. This allowed us to extract 21k in post-claim pairs. It is worth noting that we only use monolingual pairs from this dataset in our work.

*(c) X-Claim* [17]: This is a multilingual dataset labeled for claim spans and includes six languages, primarily focusing on low-resource languages. The authors collected social media posts and corresponding claims from several fact-checking websites. They used a variety of filtering rules to eliminate posts containing videos, Instagram reels, or excessively short or long text. Using AWESOME-ALIGN [36], they found word tokens in the post-sentence that matched those in the normalized claim. The claim span was then calculated as a sequence of word tokens that began with the first aligned word token and ended with the last aligned word token in the sentence. Given that each example in this dataset included social media posts and the corresponding claims obtained from the fact-check sites, we used all the examples in the dataset: 5,840 post-claim pairs in six languages.

**Table 1**
Examples of social media posts and their corresponding normalized claims from mCLAN.

| | Social Media Post | Normalized Claim |
|---|---|---|
| English | Something to #consider don't you #think ? Something to #consider, don't you #think? Something to #consider, don't you #think? 40 years worth of research...*no vaccine for HIV *At least 100 years of research...no vaccine for cancer Ongoing research... no vaccine for the common cold Less than a year for a Covid vaccine? | Vaccines for HIV, cold, and cancer should deter you from getting the Covid-19 vaccine. |
| German | Das reiche Deutschland, wir haben das geringste Durchschnittseinkommen, die geringsten Renten und die dümmsten Wähler. *(Translation: Rich Germany, we have the lowest average income, the lowest pensions and the stupidest voters.)* | Deutschland hat geringste Durchschnittseinkommen und Renten. *(Translation: Germany has the lowest average incomes and pensions.)* |
| French | Regardez les merveilles et miracles de DIEU. Un bébé né lors de l'éboulement de Bafoussam. Son coordon onbilical est encore là relié à sa maman décédée Regardez les merveilles et miracles. *(Translation: Look at the wonders and miracles of GOD. A baby born during the Bafoussam landslide. His onbilical coordinate is still there connected to his deceased mother Look at the wonders and miracles.* | Né pendant l'éboulement à Bafoussam. *(Translation: Born during the landslide in Bafoussam.)* |

**Table 2**
Dataset statistics for all 20 languages.

| Split | Arabic | Bengali | Czech | German | Greek | English | French | Hindi | Korean | Marathi |
|---|---|---|---|---|---|---|---|---|---|---|
| **Train** | 470 | 0 | 0 | 386 | 0 | 11,374 | 1,174 | 1,081 | 0 | 137 |
| **Dev** | 118 | 0 | 0 | 101 | 0 | 1,171 | 147 | 50 | 0 | 50 |
| **Test** | 100 | 81 | 123 | 100 | 156 | 1,285 | 148 | 100 | 274 | 100 |
| **Split** | **Indonesian** | **Dutch** | **Punjabi** | **Polish** | **Portugese** | **Romanian** | **Spanish** | **Tamil** | **Telugu** | **Thai** |
| **Train** | 540 | 0 | 445 | 163 | 1,735 | 0 | 3,458 | 102 | 0 | 244 |
| **Dev** | 137 | 0 | 50 | 41 | 223 | 0 | 439 | 50 | 0 | 61 |
| **Test** | 100 | 177 | 100 | 100 | 225 | 141 | 439 | 100 | 116 | 100 |

*(d) Twitter Dataset* [37]: The authors proposed an abstractive text summarization dataset consisting of noisy claims from Twitter and their gold summaries for efficiently detecting previously fact-checked claims that use abstractive summaries to generate crisp queries. They crawled Twitter for URLs from fact-checking organizations like Snopes, PolitiFact, The Quint, etc., resulting in a preliminary collection of Tweet and Claim Review[1] pairs. Pairs with tweets in languages other than English were discarded, as were such with only image or video content. They also ensured that each tweet included a claim and could be textually summarized to match the corresponding Claim Review. The final dataset only included <*Social Media Content, Claim Review*> pairs with both components in English. We used all the 567 pairs provided in this dataset.

To ensure the data quality of the final compiled corpus, we randomly selected 50 examples from each language and asked native speakers to verify the post and the corresponding normalized claims. For languages where we could not find native speakers, we used the Google Translate API to translate them into English and cross-checked the quality of the examples. Table 1 shows a few examples from our mCLAN dataset in different languages. We consolidated all examples from these datasets and performed a combined analysis. Table 1 shows a few examples from mCLAN dataset in different languages.

## 3.2. Data Statistics and Analysis

Through data compilation, we obtained a total of 28,012 instances in twenty languages from all datasets. To maintain uniformity, we used the train/dev/test splits from the original datasets. For languages with a small number of instances, e.g., around 100, we only kept the test sets with no training data. Table 2 gives details about the final dataset and the train/dev/test splits.

[1]Short summary of the claim written by the fact-checker.

To better apprehend the distribution of languages, we analysed the dataset linguistically. With its diverse vocabulary and flexible syntax, English is the primary language used on several social media platforms [38]. Thus, our dataset is also primarily composed of English examples. While German and Dutch are less dominant, they still benefit from a shared Latin script and similar grammatical structure. The Indic languages in the dataset encompass Hindi, Marathi, Punjabi, and Bengali. Hindi uses the Devanagari script. Marathi also uses the Devanagari script, albeit with some differences in the characters. The Gurmukhi script is used for Punjabi, and Bengali is written using the Bengali script. Due to their diverse scripts and extensive use of diacritics, Indic languages pose unique computational challenges. The dataset also includes two languages from the Dravidian languages: Tamil and Telugu. Both are important representatives of the Dravidian language family, with scripts derived from the ancient Brahmic script.

## 4. Submissions

We received submissions from 18 teams, totalling 1,226 valid runs across all the languages; 12 of these teams submitted their working notes. Table 3 lists all teams and their ranking for each language.

**Table 3**
List of the participating teams and their rankings. Teams marked with a + did not submit working notes.

| Team | English | Arabic | German | French | Hindi | Marathi | Indonesian | Punjabi | Polish | Portugese | Spanish | Tamil | Thai | Bengali | Telugu | Dutch | Czech | Greek | Romanian | Korean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| dfkinit2b [39] | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| DS@GT [40] | 2 | 2 | 1 | 1 | 2 | 4 | 1 | 5 | 1 | 1 | 1 | 3 | 1 | 4 | 5 | 5 | 3 | 4 | 4 | 3 |
| TIFIN [41] | 3 | 5 | 5 | 6 | 7 | 6 | | 4 | 5 | | 5 | 5 | | 5 | 6 | 4 | | | | |
| AKCIT-FN [42] | 4 | 6 | 3 | 3 | 5 | 5 | 3 | 2 | 3 | 3 | 3 | 2 | 2 | 3 | 2 | 2 | 4 | 2 | 2 | 2 |
| Factiverse and IAI [43] | 5 | 7 | 4 | 4 | 8 | 9 | 4 | 7 | 6 | 6 | 6 | 8 | 4 | 6 | 7 | | 5 | 5 | 5 | |
| rohan_shankar+ | 6 | | | | | | | | | | | | | | | | | | | |
| manan-tifin+ | 7 | | | 7 | 9 | 7 | | | 5 | | | | | 5 | 6 | | | | | |
| MMA [44] | 8 | 3 | 7 | 8 | 6 | 3 | 5 | 6 | 7 | 4 | 4 | 6 | | | | | | | | |
| UNH [45] | 9 | | | | | | | | | | | | | | | | | | | |
| Investigators [46] | 10 | | | | | | | | | | 8 | | | | | | | | | 5 |
| OpenFact [47] | 11 | 4 | 6 | 5 | 4 | 2 | 6 | 3 | 4 | 5 | 7 | 4 | 5 | 2 | 3 | 3 | 2 | 3 | 3 | 4 |
| Nikhil_Kadapala+ | 12 | | | | | | | | | | | | | | | | | | | |
| aryasuneesh+ | 13 | | 5 | 6 | 7 | 6 | | 4 | | | 5 | 5 | 6 | | | | | | | |
| JU_NLP@M&S [48] | 14 | | | | | | | | | | | | | | | | | | | |
| uhh_dem4ai+ | 15 | | | | | | | | | | | | | | | | | | | |
| UmuTeam [49] | 16 | 8 | 8 | 9 | 10 | 8 | 7 | 8 | 8 | 7 | 9 | 7 | 7 | 7 | 8 | 6 | 6 | 6 | 6 | 6 |
| VSE+ | 17 | | | | | | | | | | | | | | | | | | | |
| saivineetha [50] | | | | | 3 | | | | | | | | | | 4 | | | | | |

**Baseline and Evaluation Metric.** We used mT5-large as our baseline. For the monolingual setting, we fine-tuned the model using language-specific training data, translating the instruction *"Identify the central claim in the given post: <input post>"* into the language of the test claim. This allowed the model to operate directly in the target language. We used METEOR as an evaluation measure.

Table 6 presents the results for the monolingual setup, while Table 4 reports the scores for the zero-shot setup. Most of the teams outperformed the baseline, while dfkinit2b [39], DS@GT [40], TIFIN [41], and AKCIT-FN [42] consistently ranked among the top-performers across most languages. Team dfkinit2b [39] was ranked first in 6 out of 13 languages in the monolingual setting. In the zero-shot setting, they were first across all seven unseen languages.

**Table 4**
**Zero-shot Results**. METEOR scores for languages without training data.

| # | Team Name | Score |
|---|---|---|
| | **Telugu** | |
| 1 | dfkinit2b | 0.5257 |
| 2 | AKCIT-FN | 0.5176 |
| 3 | OpenFact | 0.4559 |
| 4 | saivineetha | 0.3774 |
| 5 | DS@GT | 0.3171 |
| 6 | TIFIN | 0.2502 |
| 6 | manan-tifin | 0.2502 |
| – | *Baseline* | *0.2005* |
| 7 | Factiverse and IAI | 0.0802 |
| 8 | UmuTeam | 0.0269 |
| | **Dutch** | |
| 1 | dfkinit2b | 0.2001 |
| 2 | AKCIT-FN | 0.1922 |
| 3 | OpenFact | 0.1866 |
| 4 | TIFIN | 0.1720 |
| 5 | DS@GT | 0.1608 |
| 6 | UmuTeam | 0.0817 |
| – | *Baseline* | *0.0751* |

| # | Team Name | Score |
|---|---|---|
| | **Romanian** | |
| 1 | dfkinit2b | 0.2950 |
| 2 | AKCIT-FN | 0.2516 |
| 3 | OpenFact | 0.2350 |
| 4 | DS@GT | 0.2220 |
| 5 | Factiverse and IAI | 0.2097 |
| – | *Baseline* | *0.0915* |
| 6 | UmuTeam | 0.0779 |
| | **Bengali** | |
| 1 | dfkinit2b | 0.3777 |
| 2 | OpenFact | 0.2959 |
| 3 | AKCIT-FN | 0.2916 |
| 4 | DS@GT | 0.2435 |
| 5 | TIFIN | 0.2030 |
| 5 | manan-tifin | 0.2030 |
| – | *Baseline* | *0.1333* |
| 6 | Factiverse and IAI | 0.1068 |
| 7 | UmuTeam | 0.0451 |

| # | Team Name | Score |
|---|---|---|
| | **Korean** | |
| 1 | dfkinit2b | 0.1339 |
| 2 | AKCIT-FN | 0.1209 |
| 3 | DS@GT | 0.1156 |
| 4 | OpenFact | 0.1050 |
| – | *Baseline* | *0.0231* |
| 5 | Investigators | 0.0149 |
| 6 | UmuTeam | 0.0014 |
| | **Greek** | |
| 1 | dfkinit2b | 0.2619 |
| 2 | AKCIT-FN | 0.2567 |
| 3 | OpenFact | 0.2333 |
| 4 | DS@GT | 0.2250 |
| 5 | Factiverse and IAI | 0.1455 |
| – | *Baseline* | *0.0830* |
| 6 | UmuTeam | 0.0062 |
| | **Czech** | |
| 1 | dfkinit2b | 0.2519 |
| 2 | OpenFact | 0.2144 |
| 3 | DS@GT | 0.1959 |
| 4 | AKCIT-FN | 0.1734 |
| 5 | Factiverse and IAI | 0.1571 |
| – | *Baseline* | *0.0602* |
| 6 | UmuTeam | 0.0544 |

## 4.1. Overview of the Systems

Most teams used sequence-to-sequence generation strategies for claim normalization, typically relying on transformer-based models. The most prevalent approach involved fine-tuning pretrained models such as BART, T5, mBART, and LLaMA on monolingual data.

Team **dfkinit2b** [39] participated in both settings, testing zero- and few-shot prompting with models such as Gemma-3, Qwen-3, Qwen-2.5, Llama-3.3, and Mistral. They explored various prompts and used cosine similarity to select demonstrations for few-shot learning. They also included adapter fine-tuning, data pre-processing with language checks and emoji removal, and data augmentation via translation. For the final submission, they ensembled top-performing model outputs by computing embedding centroids with multilingual SentenceTransformers and selecting claims closest to these centroids.

Team **DS@GT** [40] embedded the unnormalized claims from the pooled train and development datasets, as well as from the test set, using state-of-the-art embeddings for each language. For testing, a GPT-4o mini model was prompted following the approach discussed in [14], using the top-3 most similar examples from the train and development sets as in-context examples. The final response for the monolingual task was derived by combining the best-matching answer from the train and development sets, based on cosine similarity, and the output of the GPT-4 model. For zero-shot, they used a modified version of CACN [14], essentially using the prompting method with standard examples.

Team **TIFIN** [41] fine-tuned Qwen-14B using LoRA with 4-bit precision for efficiency. They pre-processed data by filtering meaningful post-claim pairs, removing duplicates, and creating a unified multilingual dataset. Instruction-based fine-tuning incorporated Chain-of-Thought prompting with 5W1H questions to guide claim extraction. During inference, context resolution replaced partial posts with complete ones, and few-shot prompting with similar examples improved claim structure. This approach aimed to boost claim extraction accuracy and multilingual performance.

Team **AKCIT-FN** [42] adopted a dual-strategy approach tailored to data availability. For the 13 supervised languages, they fine-tuned various language-specific and multilingual Small Language Models (SLMs) such as PTT5, AraT5, and Varta T5. For the seven zero-shot languages, they used prompting with Large Language Models (LLMs) such as the GPT series, Gemini, and Qwen 2.5. Their methodology also included a data cleaning algorithm to remove repetitive content and trailing *None* placeholders, as well as cross-split deduplication. Few-shot prompting experiments for monolingual settings involved selecting examples randomly, based on difficulty (METEOR score), or using HDBSCAN cluster prototypes for semantic diversity.

Team **Factiverse and IAI** [43] focused on the monolingual setting, comparing four main approaches: zero-shot prompting, fine-tuning, Fixed In-Context Learning (FICL), and Adaptive In-Context Learning (AICL). For the ICL methods, they used a ChromaDB vector store with all-MiniLM-L6-v2 embeddings to retrieve semantically similar examples from the training data based on cosine distance. While FICL used a fixed number of top-K examples, the team's novel AICL approach dynamically selected examples by applying a cosine distance threshold, eliminating the need to pre-determine the number of shots. They also explored data augmentation via machine translation for low-resource languages.

The **MMA** team [44] focused on the monolingual setting, exploring several model architectures and training strategies. Their approaches included fine-tuning a unified multilingual umt5 model on all languages, as well as training separate umt5 models for each language. They also tested zero-shot prompting with Qwen2.5 models and employed a parameter-efficient fine-tuning (PEFT) method using LoRA, which involved a two-stage process of first extracting key points and then generating a claim from those points. For Arabic, they conducted specific experiments by fine-tuning ara-t5 and augmenting the training data with scraped post-claim pairs from the Google Fact Check Tools API.

The **UmuTeam** [49] used a generative approach based on the Flan-T5-Base model for the Claim Extraction and Normalization task. Their strategy varied based on the data setting: for the monolingual scenarios, they fine-tuned a separate instance of Flan-T5-Base for each language, using only that language's specific training data to allow the models to specialize. For the zero-shot languages, they fine-tuned a single Flan-T5-Base model on the concatenated training data from all other languages, aiming to leverage cross-lingual transfer for generalization.

The **UNH** team [45] only experimented with the English language. Their fine-tuning experiments included fully fine-tuning a Flan-T5 Large model, using LoRA for a Flan-T5 Base model, and fine-tuning a DeepSeek-Llama-8b model. Their prompting strategies involved few-shot prompting with keyword-based example selection, iterative self-refinement to improve claim quality, and a Max Multi-Prompt method that simulated choosing the best output from several targeted prompts.

**Saivineetha** [50] focused on Hindi and Telugu. For Hindi, which was in the monolingual setting, they performed Parameter-Efficient Fine-Tuning (PEFT) using QLoRA with 4-bit quantization on the Gemma 2 9B instruct model. The model was instruction fine-tuned on the provided Hindi dataset of posts and normalized claims. For Telugu, which was in the zero-shot setting, they used zero-shot prompting with the Gemma 3 12B instruct model, using a prompt template designed to convert unstructured Telugu posts into normalized claims.

The **JU_NLP@M&S** team [48] framed the claim normalization task as a monolingual sequence-to-sequence generation problem, centered on fine-tuning a BART-Large transformer model. Their methodology included a preprocessing module for tokenization using byte-level BPE, padding inputs to a fixed length, and truncating where necessary. Model training was conducted for 5 epochs using Hugging Face's Seq2SeqTrainer, employing mixed-precision (FP16) to optimize memory usage and a learning rate of 3e-5. For inference, they used beam search with four beams to enhance the quality of the generated claims.

Team **Investigators** [46] focused on the claim normalization task by fine-tuning several models, including LLaMA-3.2, BART, and T5, with a particular focus on the flan-t5-base model for the final submission. Their methodology was primarily monolingual, with extensive experiments on the English and Spanish datasets. Before training, they implemented a pre-processing pipeline to filter out records that were not in the target language. For the zero-shot setting, they experimented with cross-lingual transfer by training a model on the Spanish dataset and then evaluating it on the Korean test data.

Team **OpenFact** [47] experimented with several decoder-only LLMs, including LLaMA 3.1, DeepSeek-R1, and GPT-4.1-mini. Their methodology had three steps: (1) generating up to three initial claim candidates, (2) iteratively refining each candidate using a self-reflection technique where the model provides feedback on its output, and (3) using an LLM as a judge to select the best among the refined candidates. They also performed supervised fine-tuning on the GPT-4.1-mini model using the cleaned training data.

**Table 5**
Detailed overview of the approaches used by the participating teams. FT stands for Fine-Tuning, ICL for In-Context Learning, and PEFT for Parameter-Efficient Fine-Tuning.

| Team | Setting | | Data | | | Approach | | | | | Model Family | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Monolingual | Zero-Shot | Content Filtering | Deduplication | Data Augmentation | Fine-Tuning (Full/PEFT) | Zero-Shot Prompting | Retrieval-Augmented ICL | Self-Reflection/Reasoning | Ensemble/Hybrid System | T5-family (T5, Flan-T5) | BART | GPT-family | Llama-family | Qwen | Gemma |
| dfkinit2b [39] | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | ✓ | | | | ✓ | ✓ | ✓ |
| DS@GT [40] | ✓ | ✓ | | ✓ | | | ✓ | ✓ | | ✓ | | | ✓ | | | |
| TIFIN [41] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | ✓ | |
| AKCIT-FN [42] | ✓ | ✓ | | ✓ | | ✓ | ✓ | ✓ | | | ✓ | ✓ | ✓ | | ✓ | |
| Factiverse and IAI [43] | ✓ | | | | ✓ | ✓ | ✓ | ✓ | | | ✓ | | | ✓ | | |
| MMA [44] | ✓ | | | | ✓ | ✓ | ✓ | | | | ✓ | | | | ✓ | |
| UNH [45] | ✓ | | | | | ✓ | ✓ | ✓ | ✓ | | ✓ | | ✓ | ✓ | | |
| Investigators [46] | ✓ | ✓ | ✓ | | | ✓ | | | | | ✓ | ✓ | | ✓ | | |
| OpenFact [47] | ✓ | ✓ | ✓ | | | ✓ | | | ✓ | | | | ✓ | ✓ | | |
| JU_NLP@M&S [48] | ✓ | | | ✓ | | ✓ | | | | | | ✓ | | | | |
| Saivineetha [50] | ✓ | ✓ | | | | ✓ | ✓ | | | | | | | | | ✓ |
| UmuTeam [49] | ✓ | ✓ | | | | ✓ | | | | | ✓ | | | | | |

## 5. Discussion of Approaches

The participating teams in CheckThat! 2025 Task 2 tried several strategies for multilingual claim normalization. These approaches can be analyzed through four major dimensions: model architecture, fine-tuning vs. in-context learning paradigms, data handling, and performance across monolingual and zero-shot settings. An overview of the approaches is given in Table 5.

### 5.1. Model Architectures

The primary area of divergence among the teams was their selection of model architecture. Some teams handled the task as a typical sequence-to-sequence problem, using encoder-decoder models that excel at summarization. For instance, the JU_NLP@M&S team fine-tuned BART-Large for monolingual text-to-text generation. UmuTeam, Investigators, and MMA explored variants of T5, including multilingual models such as Flan-T5 and UMT5. In contrast, other teams used decoder-only large language models to improve their in-context learning and reasoning abilities. OpenFact evaluated models such as LLaMA 3.1, DeepSeek-R1, and GPT-4.1-mini. Similarly, TIFIN and dfkinit2b chose Qwen for their multilingual performance and efficiency in fine-tuning. This distinction highlights the trade-off between the recognised strengths of encoder-decoder in generation tasks and the growing potential of decoder-only models for flexible reasoning.

### 5.2. Adaptation Strategies

Fine-tuning was a common choice among the participating teams. Several teams employed parameter-efficient fine-tuning (PEFT) approaches, such as LoRA or QLoRA. For example, Saivineetha fine-tuned Gemma 2 for Hindi, while TIFIN and dfkinit2b applied LoRA to Qwen models. OpenFact's supervised fine-tuning of GPT-4.1-mini was reported to be their most effective configuration. In contrast, teams using decoder-only models emphasized in-context learning (ICL). DS@GT used retrieval-based ICL, pulling top-3 similar examples from the training set as dynamic prompts. dfkinit2b also adopted semantic similarity-driven selection for few-shot prompts. Factiverse used Adaptive In-Context Learning (AICL), which dynamically modifies the number of in-context examples depending on similarity thresholds. TIFIN implemented a 5W1H prompting strategy, structuring claim-related information into six categories (Who, What, Where, When, Why, and How) to guide model reasoning. OpenFact and UNH also used self-refinement, where an LLM iteratively critiques and improves its outputs.

### 5.3. Data Handling and Hybrid Methods

Due to the noisy nature of social media data, data preprocessing becomes crucial. OpenFact used GPT-4.1-mini to filter out training instances with some mismatches with the ground truth. To augment the training data, MMA scraped additional Arabic samples using Google's Fact Check Tools API, while Investigators used the Gemini API to generate synthetic examples. DS@GT created a retrieval-first pipeline that reused existing normalizations when similar posts were found. dfkinit2b employed an ensemble method to generate claims based on five different approaches. The output closest to the centroid of all created embeddings was then chosen. This strategy worked well in both monolingual and zero-shot settings.

### 5.4. Adapting to Monolingual and Zero-Shot Scenarios

In the monolingual setting, where training data for 13 languages was available, the participating teams either trained language-specific models or used the data to retrieve information for ICL. Saivineetha, for example, trained a dedicated Hindi model, while DS@GT and Factiverse retrieved similar examples to construct prompts dynamically. In the zero-shot setting, the teams had to rely on cross-lingual generalization. UmuTeam and MMA developed multilingual models based on merged monolingual data and applied them to zero-shot languages. Other teams, such as TIFIN and DS@GT, used English-centric prompts and relied on the inherent multilingual capacity of the LLMs to handle the target languages. Among the most effective zero-shot strategies were dfkinit2b's ensemble approach and OpenFact's fine-tuned GPT-4.1-mini, both of which performed consistently well across languages without labeled data.

## 6. Conclusion

We presented a detailed overview of Task 2 from the CheckThat! Lab at CLEF 2025. It focused on claim normalization, the task of transforming informal and noisy social media content into clear, concise, and verifiable statements. In total, 18 teams participated in the task. Most of the participants used Transformer-based models, with a clear trend towards leveraging large language models from the T5, Qwen, and Llama families. Common and effective strategies included parameter-efficient fine-tuning, retrieval-augmented in-context learning, and sophisticated data preprocessing. The dual setting for monolingual and zero-shot evaluation provided a valuable framework for assessing both language-specific adaptation and cross-lingual generalization.

## Declaration on Generative AI

In this study, we employed mT5-large as the baseline system. All experiments were carried out under controlled conditions. To help with spell check suggestions, OpenAI GPT-4o was accessed through a plugin on Overleaf. The authors thoroughly evaluated and edited all of the tool's suggestions. No generative AI tools were used to generate the content of the main manuscript. The authors take full responsibility for the final content of the publication.

## References


[1] S. Muhammed T, S. K. Mathew, The Disaster of Misinformation: A Review of Research in Social Media, International Journal of Data Science and Analytics 13 (2022) 271–285.

[2] H. Allcott, M. Gentzkow, Social Media and Fake News in the 2016 Election, Journal of Economic Perspectives 31 (2017) 211–236.

[3] F. Alam, S. Shaar, F. Dalvi, H. Sajjad, A. Nikolov, H. Mubarak, G. Da San Martino, A. Abdelali, N. Durrani, K. Darwish, A. Al-Homaid, W. Zaghouani, T. Caselli, G. Danoe, F. Stolk, B. Bruntink, P. Nakov, Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 611–649.

[4] 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: Proceedings of the 44th European Conference on IR Research: Advances in Information Retrieval, ECIR '22, Springer-Verlag, Berlin, Heidelberg, 2022, pp. 416–428.

[5] I. Khaldarova, M. Pantti, Fake news, Journalism Practice 10 (2016) 891–901.

[6] N. L. Tsang, M. Feng, F. L. Lee, How Fact-Checkers Delimit Their Scope of Practices and Use Sources: Comparing Professional and Partisan Practitioners, Journalism 24 (2023) 2232–2251.

[7] A. Barrón-Cedeño, F. Alam, T. Chakraborty, T. Elsayed, P. Nakov, P. Przybyła, J. M. Struß, F. Haouari, M. Hasanain, F. Ruggeri, et al., The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness, in: European Conference on Information Retrieval, Springer, 2024, pp. 449–458.

[8] M. Sundriyal, A. Kulkarni, V. Pulastya, M. S. Akhtar, T. Chakraborty, Empowering the Fact-checkers! Automatic Identification of Claim Spans on Twitter, 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. 7701–7715.

[9] 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, Overview of the CLEF-2022 CheckThat! Lab on Fighting the COVID-19 Infodemic and Fake News Detection, in: Proceedings of the 13th International Conference of the CLEF Association: Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF '2022, Bologna, Italy, 2022.

[10] S. Gupta, P. Singh, M. Sundriyal, M. S. Akhtar, T. Chakraborty, LESA: Linguistic Encapsulation and Semantic Amalgamation Based Generalised Claim Detection from Online Content, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 3178–3188.

[11] 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.

[12] A. Kazemi, K. Garimella, D. Gaffney, S. Hale, Claim Matching Beyond English to Scale Global Fact-Checking, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4504–4517.

[13] S. Shaar, N. Babulkov, G. Da San Martino, P. Nakov, That is a Known Lie: Detecting Previously Fact-Checked Claims, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 3607–3618.

[14] M. Sundriyal, T. Chakraborty, P. Nakov, From Chaos to Clarity: Claim Normalization to Empower Fact-Checking, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 6594–6609.

[15] I. Jaradat, P. Gencheva, A. Barrón-Cedeño, L. Màrquez, P. Nakov, ClaimRank: Detecting Check-Worthy Claims in Arabic and English, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 26–30.

[16] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, et al., The CLEF-2023 CheckThat! Lab: Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority, in: European Conference on Information Retrieval, Springer, 2023, pp. 506–517.

[17] S. Mittal, M. Sundriyal, P. Nakov, Lost in Translation, Found in Spans: Identifying Claims in Multilingual Social Media, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 3887–3902.

[18] 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.

[19] M. Sundriyal, P. Singh, M. S. Akhtar, S. Sengupta, T. Chakraborty, DESYR: Definition and Syntactic Representation Based Claim Detection on the Web, in: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 1764–1773.

[20] M. Sundriyal, M. S. Akhtar, T. Chakraborty, Leveraging Rationality Labels for Explainable Claim Check-Worthiness, IEEE Transactions on Artificial Intelligence (2025).

[21] P. Gencheva, P. Nakov, L. Màrquez, A. Barrón-Cedeño, I. Koychev, A context-aware approach for detecting worth-checking claims in political debates, in: Proc. of RANLP, 2017, pp. 267–276.

[22] M. Sundriyal, G. Malhotra, M. S. Akhtar, S. Sengupta, A. Fano, T. Chakraborty, Document Retrieval and Claim Verification to Mitigate COVID-19 Misinformation, in: Proc. of workshop on CONSTRAINT, ACL, 2022, pp. 66–74.

[23] M. Glockner, I. Staliūnaitė, J. Thorne, G. Vallejo, A. Vlachos, I. Gurevych, AmbiFC: Fact-checking Ambiguous Claims with Evidence, Transactions of the Association for Computational Linguistics 12 (2024) 1–18.

[24] M. Hardalov, A. Chernyavskiy, I. Koychev, D. Ilvovsky, P. Nakov, CrowdChecked: Detecting Previously Fact-Checked Claims in Social Media, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online only, 2022, pp. 266–285.

[25] E. C. Choi, E. Ferrara, FACT-GPT: Fact-Checking Augmentation via Claim Matching with LLMs, in: Companion Proceedings of the ACM Web Conference 2024, 2024, pp. 883–886.

[26] C. P. Drolsbach, K. Solovev, N. Pröllochs, Community Notes Increase Trust in Fact-Checking on

Social Media, PNAS nexus 3 (2024) pgae217.
[27] A. Gupta, V. Srikumar, X-Fact: A New Benchmark Dataset for Multilingual Fact Checking, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 675–682.
[28] M. Pikuliak, I. Srba, R. Moro, T. Hromadka, T. Smoleň, M. Melišek, I. Vykopal, J. Simko, J. Podroužek, M. Bielikova, Multilingual Previously Fact-Checked Claim Retrieval, 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. 16477–16500.
[29] Y.-C. Chang, C. Kruengkrai, J. Yamagishi, XFEVER: Exploring Fact Verification across Languages, in: Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023), 2023, pp. 1–11.
[30] P. Nakov, A. Barrón-Cedeno, T. Elsayed, R. Suwaileh, L. Màrquez, W. Zaghouani, P. Atanasova, S. Kyuchukov, G. Da San Martino, Overview of the CLEF-2018 CheckThat! Lab on Automatic Identification and Verification of Political Claims, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9, Springer, 2018, pp. 372–387.
[31] S. Shaar, A. Nikolov, N. Babulkov, F. Alam, A. Barrón-Cedeno, T. Elsayed, M. Hasanain, R. Suwaileh, F. Haouari, G. Da San Martino, et al., Overview of CheckThat! 2020 English: Automatic Identification and Verification of Claims in Social Media., CLEF (Working Notes) 2696 (2020).
[32] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, et al., Overview of the CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News, in: Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization, CLEF '2021, Bucharest, Romania (online), 2021, pp. 264–291.
[33] A. Barrón-Cedeño, F. Alam, A. Galassi, G. Da San Martino, P. Nakov, T. Elsayed, D. Azizov, T. Caselli, G. S. Cheema, F. Haouari, et al., Overview of the CLEF–2023 CheckThat! Lab on Checkworthiness, Subjectivity, Political Bias, Factuality, and Authority of News Articles and their Source, in: International conference of the cross-language evaluation forum for European languages, Springer, 2023, pp. 251–275.
[34] A. Barrón-Cedeño, F. Alam, J. M. Struß, P. Nakov, T. Chakraborty, T. Elsayed, P. Przybyła, T. Caselli, G. Da San Martino, F. Haouari, et al., Overview of the CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 28–52.
[35] B. Koch, E. Denton, A. Hanna, J. G. Foster, Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
[36] Z.-Y. Dou, G. Neubig, Word Alignment by Fine-tuning Embeddings on Parallel Corpora, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 2112–2128.
[37] V. Bhatnagar, D. Kanojia, K. Chebrolu, Harnessing Abstractive Summarization for Fact-Checked Claim Detection, in: Proceedings of the 29th International Conference on Computational Linguistics, 2022, pp. 2934–2945.
[38] A. Petrosyan, Most used languages online by share of websites 2024, 2024. Accessed: 01 June 2024.
[39] T. Anikina, I. Vykopal, S. Kula, R. K. Chikkala, N. Skachkova, J. Yang, V. Solopova, V. Schmitt, S. Ostermann, dfkinit2b at CheckThat! 2025: Leveraging LLMs and Ensemble of Methods for Multilingual Claim Normalization, in: [51], 2025.
[40] A. Pramov, J. Ma, B. Patel, DS@GT at CheckThat! 2025: A Simple Retrieval-First, LLM-Backed Framework for Claim Normalization, in: [51], 2025.
[41] M. Sharma, A. Suneesh, M. Jain, P. K. Rajpoot, P. Devadiga, B. Hazarika, A. Shrivastava, K. Gu-

rumurthy, A. B. Suresh, A. U. Baliga, TIFIN at CheckThat! 2025: Reasoning-Guided Claim Normalization for Noisy Multilingual Social Media Posts, in: [51], 2025.
[42] F. L. N. Almada, K. D. P. Mariano, M. A. Dutra, V. E. d. S. Monteiro, J. R. S. Gomes, A. R. Galvão Filho, A. d. S. Soares, Akcit-FN at CheckThat!2025: Switching Fine-Tuned SLMs and LLM Prompting for Multilingual Claim Normalization, in: [51], 2025.
[43] P. Amatya, V. Setty, Factiverse and IAI at CheckThat! 2025: Adaptive ICL for Claim Extraction, in: [51], 2025.
[44] M. Saeed, M. Yasser, M. Torki, N. Elmakky, MMA at CheckThat! 2025: Multilingual Claim Normalization of Social-Media Posts, in: [51], 2025.
[45] J. Wilder, N. Kadapala, Y. Xu, M. Alsaadi, M. Rogers, P. Agrawal, A. Hassick, L. Dietz, UNH at Check That! 2025: Fine-tuning Vs Prompting, in: [51], 2025.
[46] S. M. A. Hashmi, S. Aamir, M. Anas, T. Usmani, F. Alvi, A. Samad, Investigators at CheckThat! 2025: Using LLMs to Improve Fact-Checking, in: [51], 2025.
[47] M. Sawiński, K. Węcel, E. Księżniak, OpenFact at CheckThat! 2025: Application of self-reflecting and reasoning LLMs for fact-checking claim normalization, in: [51], 2025.
[48] M. Mondal, S. Saha, D. Saha, D. Das, JU_NLP@M&S at CheckThat! 2025: Automated Claim Extraction and Normalization for Misinformation Detection in Social Media Content, in: [51], 2025.
[49] T. B. Beltrán, R. Pan, J. A. García Díaz, R. Valencia García, UmuTeam at CheckThat! 2025: Language-specific versus multilingual models for Fact-Checking, in: [51], 2025.
[50] S. V. Baddepudi Venkata Naga Sri, Saivineetha at CheckThat! 2025: Exploring Fine-Tuning and Zero-Shot Approaches for Claim Normalization, 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.

**Table 6**
**Monolingual Results**. METEOR scores for languages with training data.

| | English | |
|---|---|---|
| 1 | dfkinit2b | 0.4569 |
| 2 | DS@GT | 0.4521 |
| 3 | TIFIN | 0.4114 |
| 4 | AKCIT-FN | 0.4058 |
| 5 | Factiverse | 0.4049 |
| 6 | rohan_shankar | 0.3920 |
| 7 | manan-tifin | 0.3881 |
| 8 | MMA | 0.3841 |
| 9 | UNH | 0.3737 |
| 10 | Investigators | 0.3565 |
| 11 | OpenFact | 0.3370 |
| 12 | Nikhil_Kadapala | 0.3321 |
| 13 | aryasuneesh | 0.3153 |
| 14 | JU_NLP@M&S | 0.3098 |
| – | *Baseline* | *0.2865* |
| 15 | uhh_dem4ai | 0.2612 |
| 16 | UmuTeam | 0.1660 |
| 17 | VSE | 0.0070 |

| | Marathi | |
|---|---|---|
| 1 | dfkinit2b | 0.3888 |
| 2 | OpenFact | 0.3048 |
| 3 | MMA | 0.2793 |
| 4 | DS@GT | 0.2608 |
| 5 | AKCIT-FN | 0.2181 |
| – | *Baseline* | *0.2159* |
| 6 | aryasuneesh | 0.1521 |
| 6 | TIFIN | 0.1521 |
| 7 | manan-tifin | 0.1230 |
| 8 | UmuTeam | 0.0877 |
| 9 | Factiverse and IAI | 0.0847 |

| | Portuguese | |
|---|---|---|
| 1 | DS@GT | 0.5770 |
| 2 | dfkinit2b | 0.5744 |
| 3 | AKCIT-FN | 0.5290 |
| 4 | MMA | 0.4719 |
| 5 | OpenFact | 0.3779 |
| 6 | Factiverse and IAI | 0.3381 |
| – | *Baseline* | *0.3011* |
| 7 | UmuTeam | 0.1898 |

| | Thai | |
|---|---|---|
| 1 | DS@GT | 0.5859 |
| 2 | AKCIT-FN | 0.3179 |
| 3 | dfkinit2b | 0.2999 |
| – | *Baseline* | *0.2015* |
| 4 | Factiverse and IAI | 0.0965 |
| 5 | OpenFact | 0.0872 |
| 6 | aryasuneesh | 0.0464 |
| 7 | UmuTeam | 0.0147 |

| | Hindi | |
|---|---|---|
| 1 | dfkinit2b | 0.3275 |
| 2 | DS@GT | 0.3001 |
| 3 | saivineetha | 0.2996 |
| 4 | OpenFact | 0.2722 |
| 5 | AKCIT-FN | 0.2706 |
| 6 | MMA | 0.2641 |
| 7 | aryasuneesh | 0.2604 |
| 7 | TIFIN | 0.2604 |
| – | *Baseline* | *0.2283* |
| 8 | Factiverse and IAI | 0.2125 |
| 9 | manan-tifin | 0.2080 |
| 10 | UmuTeam | 0.0132 |

| | French | |
|---|---|---|
| 1 | DS@GT | 0.5273 |
| 2 | dfkinit2b | 0.4703 |
| 3 | AKCIT-FN | 0.3811 |
| 4 | Factiverse and IAI | 0.3750 |
| 5 | OpenFact | 0.3605 |
| 6 | aryasuneesh | 0.3441 |
| 6 | TIFIN | 0.3441 |
| – | *Baseline* | *0.2833* |
| 7 | manan-tifin | 0.2768 |
| 8 | MMA | 0.2469 |
| 9 | UmuTeam | 0.1649 |

| | Punjabi | |
|---|---|---|
| 1 | dfkinit2b | 0.3307 |
| 2 | AKCIT-FN | 0.3038 |
| 3 | OpenFact | 0.2696 |
| 4 | aryasuneesh | 0.2685 |
| 4 | TIFIN | 0.2685 |
| 5 | DS@GT | 0.2567 |
| 6 | MMA | 0.1834 |
| – | *Baseline* | *0.1594* |
| 7 | Factiverse and IAI | 0.1251 |
| 8 | UmuTeam | 0.0097 |

| | Indonesian | |
|---|---|---|
| 1 | DS@GT | 0.5650 |
| 2 | dfkinit2b | 0.5021 |
| 3 | AKCIT-FN | 0.3866 |
| 4 | Factiverse and IAI | 0.3099 |
| 5 | MMA | 0.3089 |
| – | *Baseline* | *0.2825* |
| 6 | OpenFact | 0.2445 |
| 7 | UmuTeam | 0.1305 |

| | Arabic | |
|---|---|---|
| 1 | dfkinit2b | 0.5037 |
| 2 | DS@GT | 0.5035 |
| 3 | MMA | 0.4584 |
| 4 | OpenFact | 0.4175 |
| 5 | TIFIN | 0.3705 |
| 6 | AKCIT-FN | 0.3277 |
| 7 | Factiverse and IAI | 0.2457 |
| – | *Baseline* | *0.2186* |
| 8 | UmuTeam | 0.0003 |

| | Spanish | |
|---|---|---|
| 1 | DS@GT | 0.6077 |
| 2 | dfkinit2b | 0.5539 |
| 3 | AKCIT-FN | 0.5213 |
| 4 | MMA | 0.5094 |
| 5 | aryasuneesh | 0.3906 |
| 5 | TIFIN | 0.3906 |
| 6 | Factiverse and IAI | 0.3821 |
| 7 | OpenFact | 0.3710 |
| 8 | Investigators | 0.3447 |
| – | *Baseline* | *0.3294* |
| 9 | UmuTeam | 0.2048 |

| | Polish | |
|---|---|---|
| 1 | DS@GT | 0.4065 |
| 2 | dfkinit2b | 0.3961 |
| 3 | AKCIT-FN | 0.2798 |
| 4 | OpenFact | 0.2666 |
| 5 | TIFIN | 0.2331 |
| 5 | manan-tifin | 0.2331 |
| 6 | Factiverse and IAI | 0.1964 |
| – | *Baseline* | *0.1594* |
| 7 | MMA | 0.1243 |
| 8 | UmuTeam | 0.0742 |

| | German | |
|---|---|---|
| 1 | DS@GT | 0.3859 |
| 2 | dfkinit2b | 0.3469 |
| 3 | AKCIT-FN | 0.2652 |
| 4 | Factiverse and IAI | 0.2644 |
| 5 | aryasuneesh | 0.2642 |
| 5 | TIFIN | 0.2642 |
| 6 | OpenFact | 0.2319 |
| 7 | MMA | 0.1556 |
| – | *Baseline* | *0.1100* |
| 8 | UmuTeam | 0.1039 |

| | Tamil | |
|---|---|---|
| 1 | dfkinit2b | 0.6316 |
| 2 | AKCIT-FN | 0.5197 |
| 3 | DS@GT | 0.4702 |
| 4 | OpenFact | 0.4681 |
| 5 | aryasuneesh | 0.3676 |
| 5 | TIFIN | 0.3676 |
| 6 | MMA | 0.3468 |
| – | *Baseline* | *0.1855* |
| 7 | UmuTeam | 0.0196 |
| 8 | Factiverse and IAI | 0.0043 |