MMA at CheckThat! 2025: Multilingual Claim Normalization of Social-Media Posts*

Notebook for the CheckThat Lab at CLEF 2025

Mariam Saeed[1,*], Mazen Yasser[1,*], Marwan Torki[1] and Nagwa ElMakky[1]

[1]Department of Computer and Systems Engineering, Alexandria University, Egypt

Abstract

This paper presents the work of the MMA team for Task 2 – Claim Normalization of the CheckThat! 2025 shared task. The task aims to convert noisy and informal social media posts into concise, check-worthy claims suitable for fact-checking. Our experiments focused on the monolingual setup of the task. While we submitted runs for all languages provided in the task, we provided additional experiments tailored for Arabic. We explored a range of approaches, including T5-based sequence-to-sequence models, zero-shot prompting, fine-tuning LLMs, and data augmentation techniques. The best-performing configurations were selected for each language. In particular, our Arabic model, using `umt5` with data augmentation, achieved a strong score of 0.4584, placing third among all submissions. Spanish achieved the highest score of 0.5094 with the base `umt5` model, while languages such as Polish and German showed lower performance. These results demonstrate the effectiveness of our multilingual strategies and the impact of data augmentation in improving performance, particularly for low-resource languages.

Keywords

Claim Normalization, Misinformation, Social Media, Fact-Checking

1. Introduction

The digital age has transformed social media platforms into primary channels for information. However, this rapid and unregulated flow of content has facilitated the widespread circulation of false or misleading information shared without malicious intent [1]. Unlike traditional news articles, social media posts often lack formal structure and editorial oversight, frequently characterized by personal opinions, emotional language, and an absence of standardized formatting. This unstructured nature complicates the assessment of content veracity. In addition, the design of social media platforms, which often rewards engagement over accuracy, can help with the spread of such content [2]. Traditional methods are slow and labor-intensive, making them insufficient for keeping up with the fast pace of online content [3].

To mitigate previous problems, researchers are developing automated fact-checking systems and collaborative verification methods to address the challenges posed by the diverse formats, varying source credibility, and complex user interactions in online environments [4]. A key step in enhancing the efficiency and effectiveness of fact-checking is extracting the main claims from complex social media posts. This process involves distilling lengthy or ambiguous content into concise, clear, and verifiable statements, enabling fact-checkers to focus on specific points rather than extraneous information [5]. This task is particularly crucial during critical events such as elections, public health crises, or conflicts, where misinformation can significantly influence public opinion and behavior [6].

In this study, we present our approach for Task 2 of the CLEF2025 CheckThat! Lab [7], which focuses on claim extraction and normalization from unstructured social media content. We developed multiple transformer-based models to extract the main claims from posts, aiming to transform informal content into concise, standardized statements. To enhance the robustness of our models, we employed

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

*Corresponding author.

es-mariamzaho4@alexu.edu.eg (M. Saeed); es-mazen2215@alexu.edu.eg (M. Yasser); mtorki@alexu.edu.eg (M. Torki); nagwamakky@alexu.edu.eg (N. ElMakky)

data augmentation techniques, including the collection of relevant posts from fact-checking resources and the utilization of large language models (LLMs) to refine the collected data into suitable training examples. Our experiments demonstrate that our contributions can enhance this process across 13 languages provided in the shared task's monolingual setup, thereby supporting the broader goal of effective misinformation detection and fact-checking.

2. Related Work

To better understand how to extract and normalize social media claims, we look at related work in four main areas. First, we explain how claims are defined and show the importance of normalization. Second, we explore how sequence-to-sequence models help generate claims from longer posts. Then, we discuss how data augmentation can improve model performance when training data is limited. Finally, we investigate the use of large language models (LLMs) to perform claim extraction and normalization more effectively even with little or no fine-tuning.

2.1. Claim Definition and Normalization

A claim in the context of fact-checking is typically defined as a declarative statement asserting a piece of information as true, which can be subsequently verified for its accuracy [8]. On social media, claims are often embedded within informal language, opinions, and multimedia content, making their precise identification challenging. The unstructured and often ambiguous nature of these platforms necessitates a clear definition of what constitutes a verifiable claim to enable effective automated processing [9].

Claim normalization refers to the process of transforming an extracted claim into a standardized, concise, and context-independent statement that is suitable for verification [10]. This often involves removing redundancies and converting informal or colloquial language into a more formal and unambiguous representation. Normalization is crucial because the same underlying claim can be expressed in various ways across different posts or platforms. Standardizing these expressions allows fact-checkers to consolidate efforts, identify duplicate claims, and efficiently retrieve relevant evidence from knowledge bases or previously checked claims [11]. Research in this area has explored rule-based systems, semantic similarity measures, and more recently, generative models to achieve effective normalization [12].

2.2. Sequence-to-Sequence Transformer Models for Claim Generation

Sequence-to-sequence (Seq2Seq) models, particularly those based on the Transformer architecture [13], have become the standard for many natural language generation tasks, including abstractive summarization, machine translation, and dialogue generation. Their ability to capture long-range dependencies and contextual information through attention mechanisms makes them well-suited for transforming longer, unstructured social media posts into concise claims. In the context of claim extraction, these models can be trained to transform a social media post into a shorter, claim-like sentence that encapsulates its core verifiable assertion.

Several studies have demonstrated the effectiveness of Transformer-based models like BART [14] and T5 [15] for abstractive summarization, a task closely related to claim generation. These models are pre-trained on large text corpora and can be fine-tuned on specific datasets to generate coherent and relevant summaries. Adapting these architectures for claim extraction involves fine-tuning them on datasets where social media posts are paired with their corresponding main claims [16]. The goal is to generate a statement that is not an extraction of a sentence from the original post but a potentially novel sentence that accurately represents the core claim.

2.3. Data Augmentation

The performance of deep learning models, including Seq2Seq Transformers, is heavily dependent on the availability of large, high-quality training datasets. For specialized tasks like claim extraction from

social media, particularly in multiple languages, such datasets are often scarce or expensive to create. Data augmentation techniques are employed to artificially expand the training set, thereby improving model generalization and robustness [17].

Common NLP data augmentation methods include back-translation, translating a sentence to a target language and then back to the original, synonym replacement, and paraphrasing [18]. For Seq2Seq tasks, techniques like noising, sentence shuffling, and synthetic data generation using pre-trained language models have also been explored. Studies have shown that data augmentation can significantly boost the performance of Transformer models in low-resource scenarios or for tasks requiring nuanced understanding, such as claim generation, by exposing the model to a wider variety of linguistic expressions of similar underlying claims.

2.4. Leveraging LLMs for Claim Extraction and Normalization

Recent advancements in Large Language Models (LLMs), such as GPT-3.5, GPT-4 [19], and instruction-tuned models such as Qwen [20, 21], have significantly enhanced their ability to perform complex natural language processing tasks without the need for task-specific fine-tuning. This zero-shot capability is particularly beneficial for tasks like claim extraction and normalization, which often suffer from a scarcity of annotated data.

Moreover, the inherent reasoning abilities of LLMs have been harnessed to improve the accuracy of claim verification. Kojima et al. [22] showed that prompting LLMs with phrases such as "Let's think step by step" can significantly enhance their zero-shot reasoning performance across various tasks, including fact verification .

In the realm of claim extraction, LLMs have demonstrated proficiency in identifying and structuring claims from unstructured text. For instance, Liu et al. [23] introduced a self-prompting framework that enables LLMs to perform zero-shot relation extraction effectively by generating synthetic samples that encapsulate specific relations, thereby guiding the model without explicit training data . Similarly, Sundriyal et al. [24] propose the CACN framework, leveraging chain-of-thought prompting and in-context learning with large language models to enhance the claim normalization process.

3. Methods

We investigate multilingual fine-tuning with T5-based architectures, zero-shot and fine-tuned large language models, and data augmentation techniques. Each approach is described in detail in the following subsections. The quantitative analysis and comparison between different approaches are given in section 4.

3.1. Multilingual T5-Based Training

We adopted the `google/umt5` model[25], a multilingual version of T5 designed for cross-lingual generation tasks. We fine-tuned it for post-to-claim generation and explored two training configurations:

- **Unified Multilingual Training**: We trained a single `umt5` model on the entire multilingual dataset. This approach leverages shared cross-lingual representations, enabling low-resource languages to benefit from richer ones through transfer learning.
- **Language-Specific Training**: We trained separate `umt5` models per language. While this setup goes beyond cross-lingual transfer, it allows each model to better specialize in the linguistic nuances of its respective language, potentially improving claim generation fidelity.

3.2. Zero-Shot Prompting

We evaluated the generative capability of Qwen 2.5 [20] (Qwen2.5-7B and Qwen2.5-32B) in a zero-shot setting. This setting assesses the out-of-the-box generalization performance of the pre-trained model for claim normalization. The models were prompted directly with the following instruction:

System Prompt

Given a social media post. Your task is to extract the post claim.
You must filter the noise from the post.
You must respond with the same language of the post.
Ignore repetition, rhetorical flourishes, and background anecdotes.
Keep only the essential elements:
• Who is involved (names, age, nationality).
• What is being claimed will happen or has happened.
• Where/When only if they are integral to the claim.
The claim must only include words from the post and do not use any external words.
The claim must be in the style of title for an article.
The claim must be Declarative Short Sentence with preserving names and places.
DO NOT add, omit, or translate information. Preserve names and wording where possible.

3.3. Instruction Fine-Tuning with Key Point Intermediate Step

To improve the performance of the Qwen2.5-7B model, we applied parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA)[26]. The training pipeline involved a two-stage strategy:

1. **Key Point Extraction**: Using Qwen2.5-32B in a zero-shot setting, we extracted key points from each training example (post and normalized claim) with the following system instruction:

   Training System Prompt

   Given a social media post, your task is to:
   - Identify and extract the important key points presented in this post.
   - Then, you have to use all these key points to extract the main claim of that post.
   You must answer in the same language as the given post. The output must be in the following format:
   <keypoints>
   ['key1','key2',....]
   </keypoints>
   <claim>
   claim of the post
   </claim>

   This produced a curated intermediate dataset of post-key point pairs.
2. **Claim Generation from Key Points**: We fine-tuned Qwen2.5-7B model on this augmented dataset to map input posts to final claims via the extracted key points. The objective was to focus the model's attention on important information and reduce the noise from irrelevant content.

3.4. Arabic-Specific Data Augmentation and Modeling

While previous sections focused on multilingual and monolingual experiments across 13 languages, we dedicated additional effort to improving performance on Arabic, given its linguistic complexity. This involved both model selection and targeted data augmentation tailored to Arabic content.

We experimented with the following variants of T5 models, specifically ara-t5 models [27], specialized for Arabic:

- `ara-t5-v2`, a general-purpose Arabic T5 model,
- `ara-t5`, trained for title generation tasks,
- `ara-t5-tweet`, fine-tuned on Arabic tweet data.

In addition to modeling variations, we enriched the Arabic training data using the Google Fact Check Tools API [28]. This API provides access to a curated set of verified claims and related content. We scraped Arabic post-claim pairs from this resource and incorporated them into our training pipeline. The goal was to increase the amount of high-quality, claim-specific data for Arabic, which is typically underrepresented.

These augmented samples were used to fine-tune the multilingual `umt5`, which showed superior performance and was ultimately selected for final submission. An example of the augmented Arabic data is shown below:

Scrapped sample

Post:
تداولت صفحات وحسابات على مواقع التواصل الاجتماعي صوراً ومقاطع فيديو ادعت أنها تُظهر شوارع باريس صباح اليوم بعد انتهاء حفلة رأس السنة، وتظهر فيها كميات كبيرة من النفايات. إلا أن هذا الادعاء غير صحيح، فالصور ومقاطع الفيديو المتداولة تعود إلى أماكن أخرى وفي تواريخ مختلفة، وليست لباريس بعد احتفالات رأس السنة. بعضها يعود إلى نابولي في إيطاليا، وبعضها الآخر من احتفالات سابقة في باريس نفسها.
Claim:
الصور ومقاطع الفيديو المتداولة التي تدعي أنها تُظهر شوارع باريس بعد حفلة رأس السنة هي صور ومقاطع فيديو قديمة أو من أماكن أخرى.

4. Experiments

4.1. Datasets and Metrics

For our experiments, we utilized the dataset provided by the CheckThat! 2025 Task 2: Claim Normalization [7]. The task is offered in two distinct setups:

- **Monolingual Setup**: This setup provides training, development, and test datasets for 13 languages: Arabic, German, English, French, Hindi, Marathi, Indonesian, Punjabi, Portuguese, Spanish, Tamil, Thai, and Polish.
- **Zero-Shot Setup**: For the remaining 7 languages: Bengali, Czech, Greek, Korean, Romanian, Telugu, and Dutch, only the test datasets are provided, without accompanying training or development data.

Our participation was exclusively in the monolingual setup, focusing on languages with complete datasets. Table 1 presents the aggregated statistics of the dataset across the training, development, and test splits for these 13 languages.

Table 1
Aggregated Dataset Statistics.

Split	Number of Posts
Train	21,309
Development	2,638
Test	4,065

For evaluation, we employed the METEOR score [29], as specified by the task organizers. METEOR is designed to evaluate the quality of machine-generated text by comparing it to reference texts. It calculates a harmonic mean of unigram precision and recall, with recall weighted higher than precision. Additionally, METEOR incorporates features such as stemming, synonymy matching, and a fragmentation penalty to account for word order, making it more aligned with human judgment.

4.2. Training Setup

We conducted our experiments using a single NVIDIA A6000 GPU. Our training encompassed two primary approaches: fine-tuning T5-based models and fine-tuning large language models (LLMs) using Low-Rank Adaptation (LoRA).

4.2.1. T5-Based Models

For fine-tuning T5-based models, we employed the following hyperparameters:

- **Number of epochs**: 20
- **Learning rate**: 5e-4
- **Learning rate scheduler**: Cosine with 90 warmup steps
- **Optimizer**: AdamW
- **Weight decay**: 0.01
- **Effective batch size**: 32

During evaluation, we selected the best checkpoint and then utilized the following decoding parameters:

- **Maximum sequence length**: 512
- **Number of beams**: 5
- **Top-p**: 0.85
- **Top-k**: 40

4.2.2. Large Language Models with LoRA

For fine-tuning large language models using LoRA, we adopted a parameter-efficient approach, adjusting only a subset of the model's parameters. The hyperparameters for this setup were:

- **Number of epochs**: 5
- **Learning rate**: 1e-5
- **Effective batch size**: 8
- **LoRA rank**: 8

4.3. Results and Analysis

We conducted a comprehensive evaluation across multiple training paradigms to assess the impact of multilingual modeling, zero-shot generalization, and parameter-efficient fine-tuning. Our experiments use the METEOR score as the evaluation metric across 13 languages in the monolingual setting.

Multilingual vs. Language-Specific Models. We first trained a single `umt5` model jointly on all 13 languages. While this setup benefits from shared multilingual patterns, it showed limitations in capturing language-specific nuances. To address this, we trained a separate `umt5` model for each language. As shown in Table 2, language-specific models outperformed the multilingual one in several languages. However, in a few cases like Marathi and Punjabi, the multilingual model demonstrated better performance, likely due to limited monolingual data for those languages. Since the monolingual setup outperformed the multilingual one in most cases and is better suited to capturing the linguistic characteristics of each language, we adopted it for our final submission across all languages—except for Arabic, where the multilingual `umt5` model with data augmentation achieved superior results.

Zero-Shot Inference with LLMs. We evaluated the zero-shot capabilities of two large language models `Qwen2.5-7B` and `Qwen2.5-32B`. As expected, the 32B variant achieved better performance across almost all languages, highlighting the positive correlation between model size and generalization in zero-shot settings. However, variants lagged behind fine-tuned models, especially in low-resource languages like Hindi and Punjabi.

Parameter-Efficient Fine-Tuning. The results were competitive, achieving an average improvement over the zero-shot 7B baseline. For example, in Tamil, LoRA fine-tuning reached 0.437 compared to 0.156 in zero-shot 7B and 0.295 in zero-shot 32B. This demonstrates that parameter-efficient tuning can significantly close the gap to full fine-tuning with much lower resource requirements.

Arabic-Specific Modeling. In addition to our multilingual experiments, we performed a focused evaluation on Arabic using various variants of the `AraT5` model to evaluate their suitability for claim normalization.

Their METEOR scores on the development and the test sets are shown in Table 3. Among them, `ara-t5-v2` yielded the best performance. However, despite promising results, these models were outperformed by the multilingual `umt5`, especially when combined with data augmentation strategies.

To further boost performance, we extended the training data using scraping post-claim examples using the Google Fact Check Tools API. As shown in Table 4, the best result (0.4584 on the test set) was obtained using `umt5` trained with the full augmented dataset, confirming the advantage of cross-lingual modeling with targeted data enhancement.

Final Results on the Test Set. Table 4 presents our final submitted results along with the final rank on the leaderboard. Due to evaluation issues with the Thai language, we submitted predictions for 12 out of the 13 target languages. For each language, the best-performing configuration was selected. The Arabic model achieved a strong score of 0.4584 using `umt5` with targeted data augmentation, while Spanish attained the highest overall score of 0.5094 with the base `umt5` model. On the other hand, languages such as Polish and German yielded comparatively lower scores, suggesting the need for additional data augmentation or model adaptation to improve performance.

Table 2
Development set METEOR scores for different models across 13 languages.

Language	UMT5 (All)	UMT5 (Mono)	Qwen2.5-7B Zero-shot	Qwen2.5-32B Zero-shot	Qwen2.5-7B-LoRA
Arabic	0.353	**0.401**	0.332	0.335	0.322
German	**0.268**	0.2186	0.121	0.165	0.192
English	0.38	**0.415**	0.247	0.262	0.311
French	**0.362**	0.2457	0.2	0.288	0.307
Hindi	**0.262**	0.258	0.055	0.214	0.22
Marathi	**0.39**	0.219	0.142	0.322	0.32
Indonesian	0.392	**0.3959**	0.2	0.273	0.31
Punjabi	0.278	0.21	0.08	**0.319**	0.25
Polish	0.217	0.148	0.147	0.204	**0.22**
Portuguese	0.44	**0.4787**	0.22	0.302	0.346
Spanish	0.42	**0.4966**	0.21	0.29	0.31
Tamil	0.395	0.271	0.156	0.295	**0.437**
Thai	0.231	-	0.054	0.0533	0.116

5. Conclusion

In this paper, we presented our contribution to Task 2 of the CLEF2025 CheckThat! Lab, addressing the multilingual challenge of claim extraction and normalization from social media content. We developed transformer-based models, primarily focusing on the monolingual setup across 13 languages. Our

Table 3
Arabic-specific METEOR scores for different model configurations on the development and test sets.

Model	Develpment Score	Test Score
AraT5-v2 (general-purpose)	**0.427**	0.4243
AraT5 (title generation)	0.3914	0.418
AraT5 (tweet-based)	0.1397	-
UMT5 (Mono)+ scraped data	0.4195	**0.4584**

Table 4
Final test set METEOR scores per language. The best configurations alongside our overall leaderboard rank are reported.

Language	METEOR	Best Configuration	Final Rank
Arabic	0.4584	UMT5(Mono) + Augmentation	**3**
German	0.1556	UMT5 (Mono)	8
English	0.3841	UMT5 (Mono)	8
French	0.2469	UMT5 (Mono)	9
Hindi	0.2641	UMT5 (Mono)	6
Marathi	0.2793	UMT5 (Mono)	**3**
Indonesian	0.3089	UMT5 (Mono)	5
Punjabi	0.1834	UMT5 (Mono)	7
Polish	0.1243	UMT5 (Mono)	8
Portuguese	0.4719	UMT5 (Mono)	4
Spanish	0.5094	UMT5 (Mono)	4
Tamil	0.3468	UMT5 (Mono)	7

approach combined multilingual T5 variants, zero-shot prompting of large language models, and parameter-efficient fine-tuning with LoRA, alongside a novel keypoint-guided generation strategy. Data augmentation, especially via the Google Fact Check Tools API, played a vital role in improving model performance in low-resource settings, e.g., for Arabic language. Our experiments demonstrated the effectiveness of these methods, with umt5 models and keypoint-assisted strategies yielding strong results. While our work is limited to the monolingual track, future efforts will target improved keypoint generation and broader support for zero-shot settings. Our findings highlight the potential of tailored multilingual systems in enhancing claim verification pipelines and combating misinformation online.

Declaration on Generative AI

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

References

[1] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online, science 359 (2018) 1146–1151.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective, ACM SIGKDD explorations newsletter 19 (2017) 22–36.

[3] D. Graves, Understanding the promise and limits of automated fact-checking, Reuters Institute for the Study of Journalism (2018).

[4] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. Da San Martino, Automated fact-checking for assisting human fact-checkers, 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. 4551–4558. URL: https://doi.org/10.24963/ijcai.2021/619. doi:10.24963/ijcai.2021/619, survey Track.
[5] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.
[6] V. La Gatta, C. Wei, L. Luceri, F. Pierri, E. Ferrara, Retrieving false claims on twitter during the russia-ukraine conflict, in: Companion proceedings of the ACM web conference 2023, 2023, pp. 1317–1323.
[7] M. Sundriyal, T. Chakraborty, P. Nakov, Overview of the CLEF-2025 CheckThat! lab task 2 on claim normalization, in: 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.
[8] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, arXiv preprint arXiv:1803.05355 (2018).
[9] A. Vlachos, S. Riedel, Fact checking: Task definition and dataset construction, in: Proceedings of the ACL 2014 workshop on language technologies and computational social science, 2014, pp. 18–22.
[10] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, et al., Overview of checkthat! 2020: Automatic identification and verification of claims in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11, Springer, 2020, pp. 215–236.
[11] D. S. Nielsen, R. McConville, Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset, in: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, 2022, pp. 3141–3153.
[12] A. Hanselowski, H. Zhang, Z. Li, D. Sorokin, B. Schiller, C. Schulz, I. Gurevych, Ukp-athene: Multi-sentence textual entailment for claim verification, arXiv preprint arXiv:1809.01479 (2018).
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).
[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.
[16] B. Chen, B. Chen, D. Gao, Q. Chen, C. Huo, X. Meng, W. Ren, Y. Zhou, Transformer-Based Language Model Fine-Tuning Methods for COVID-19 Fake News Detection, 2021, pp. 83–92. doi:10.1007/978-3-030-73696-5_9.
[17] C. Khosla, B. S. Saini, Enhancing performance of deep learning models with different data augmentation techniques: A survey, in: 2020 international conference on intelligent engineering and management (ICIEM), IEEE, 2020, pp. 79–85.
[18] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, arXiv preprint arXiv:1901.11196 (2019).
[19] OpenAI, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.
[20] Q. Team, qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.
[21] Q. Team, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.
[22] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2023. URL: https://arxiv.org/abs/2205.11916. arXiv:2205.11916.
[23] S. Liu, Y. Li, J. Li, S. Yang, Y. Lan, Unleashing the power of large language models in zero-shot relation extraction via self-prompting, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 13147–13161. URL: https://aclanthology.org/2024.

findings-emnlp.769/. doi:10.18653/v1/2024.findings-emnlp.769.
[24] M. Sundriyal, T. Chakraborty, P. Nakov, From chaos to clarity: Claim normalization to empower fact-checking, 2024. URL: https://arxiv.org/abs/2310.14338. arXiv:2310.14338.
[25] H. W. Chung, X. Garcia, A. Roberts, Y. Tay, O. Firat, S. Narang, N. Constant, Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/forum?id=kXwdL1cWOAi.
[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.
[27] E. M. B. Nagoudi, A. Elmadany, M. Abdul-Mageed, AraT5: Text-to-text transformers for Arabic language generation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 628–647. URL: https://aclanthology.org/2022.acl-long.47.
[28] Google fact check tools, 2024. https://toolbox.google.com/factcheck/explorer.
[29] A. Lavie, A. Agarwal, Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, StatMT '07, Association for Computational Linguistics, USA, 2007, p. 228–231.