TIFIN at CheckThat! 2025: Reasoning-Guided Claim Normalization for Noisy Multilingual Social Media Posts

Notebook for the CheckThat! Lab at CLEF 2025

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

[1]TIFIN

Abstract

We address claim normalization for multilingual misinformation detection - transforming noisy social media posts into clear, verifiable statements across 20 languages. The key contribution demonstrates how systematic decomposition of posts using Who, What, Where, When, Why and How questions enables robust cross-lingual transfer despite training exclusively on English data. Our methodology incorporates finetuning Qwen3-14B using LoRA with the provided dataset after intra-post deduplication, token-level recall filtering for semantic alignment and retrieval-augmented few-shot learning with contextual examples during inference. Our system achieves METEOR scores ranging from 41.16 (English) to 15.21 (Marathi), securing third rank on the English leaderboard and fourth rank for Dutch and Punjabi. The approach shows 41.3% relative improvement in METEOR over baseline configurations and substantial gains over existing methods. Results demonstrate effective cross-lingual generalization for Romance and Germanic languages while maintaining semantic coherence across diverse linguistic structures.

Keywords

claim normalization, misinformation detection, multilingual NLP, social media analysis

1. Introduction

Misinformation represents the foremost global threat for 2025, according to the World Economic Forum's Global Risks Report [1], while false news spreads up to 10 times faster than accurate reporting on social media platforms [2]. Social media giants have recently abandoned traditional fact-checking programs in favor of community-driven approaches [3], creating new gaps in verification systems precisely when misinformation campaigns target everything from elections to disaster response. The noisy nature of social media posts makes it challenging to identify important claims that require manual fact-checking, forcing researchers to develop automated solutions for processing the overwhelming volume of misleading content. Our work addresses this critical challenge through CheckThat! Lab CLEF 2025 Task 2: Claim Normalization, which focuses on transforming chaotic social media posts into clear, verifiable statements across 20 languages. This text generation task requires systems to extract core assertions from noisy posts and present them in normalized forms suitable for fact-checking pipelines, representing a fundamental step toward scaling verification efforts to match the speed and volume of misinformation spread.

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

*Corresponding author.

†These authors contributed equally.

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

1.1. Task Overview

CheckThat! Lab CLEF 2025 [4] [5] Task 2 [6] introduces the problem of simplifying noisy social media posts into normalized claims that fact-checkers process efficiently. The task operates across 20 languages including English, Arabic, German, French, Spanish, Hindi and 14 others, requiring systems to handle diverse linguistic structures and cultural contexts. Participants face two distinct settings: monolingual, where training, development and test data exist for the same language and zero-shot, where only test data exists for the target language. The monolingual setting covers 13 languages with full datasets, while the zero-shot setting evaluates generalization across 7 languages including Dutch, Romanian, Bengali, Telugu, Korean, Greek and Czech. Posts originate from various social media platforms including Twitter, Reddit and Facebook, sourced from Google Fact-check Explorer to ensure real-world relevance. Systems generate normalized claims evaluated using METEOR score, measuring the quality of simplified text against human-annotated ground truth. This research addresses the practical challenge faced by fact-checkers who must process thousands of posts daily, extracting verifiable claims from content laden with hashtags, mentions, emojis and informal language that obscures the core assertions requiring verification.

2. Related Work

At its core, claim normalization is an abstractive generation task closely related to summarization, but with key differences. Sequence-to-sequence models like BART [7] or T5 [8] have advanced general-purpose summarization. Controlled summarization techniques allow setting summary length or focus [9] [10] [11]. However, generic summaries may omit critical facts or introduce hallucinations, making them unreliable for fact-checking. For example, Kryściński et al. (2020) [12] showed that abstractive models often add contradictory information. Utama et al. (2022) [13] and Durmus et al. (2020) [14] developed QA-based checks for factual consistency. Claim normalization instead prioritizes factual precision and context-independence: the generated claim must be fully verifiable on its own. Sundriyal et al. [15] note that unlike typical summaries, normalized claims "must be self-contained and verifiable". This means, for example, resolving entities or adding minimal context so that the claim cannot be misunderstood when isolated (e.g. clarifying that "Bird" refers to the scooter company, rather than the animal).

In practice, many systems treat normalization as a specialized summarization. For instance, Reddy et al. (2024) [16] recast document-level claim extraction as extractive summarization followed by decontextualization: they extract central sentences and then use a QA-based model to expand them into stand-alone claims. This approach yielded higher relevance (precision@1) and fact consistency in their test cases. Similarly, models trained for text summarization (T5/BART/PEGASUS) have been applied directly as baselines for normalization. However, the unique goal of preserving a single factual assertion often calls for tailored strategies.

3. Methodology

3.1. Model Architecture

We employ Qwen3-14B [17] as our base model due to its strong multilingual capabilities and efficient architecture for fine-tuning across diverse languages. Qwen3-14B demonstrates robust performance on multilingual tasks, achieving 79.69 on the MMMLU benchmark, while maintaining computational efficiency, making it well-suited for our cross-lingual claim normalization objectives. The model's strong multilingual foundation provides an ideal starting point for fine-tuning across diverse languages without sacrificing performance on cross-lingual understanding tasks. We fine-tune the model using Low-Rank Adaptation (LoRA) [18] with 4-bit quantization for memory efficiency. Our training configuration includes:

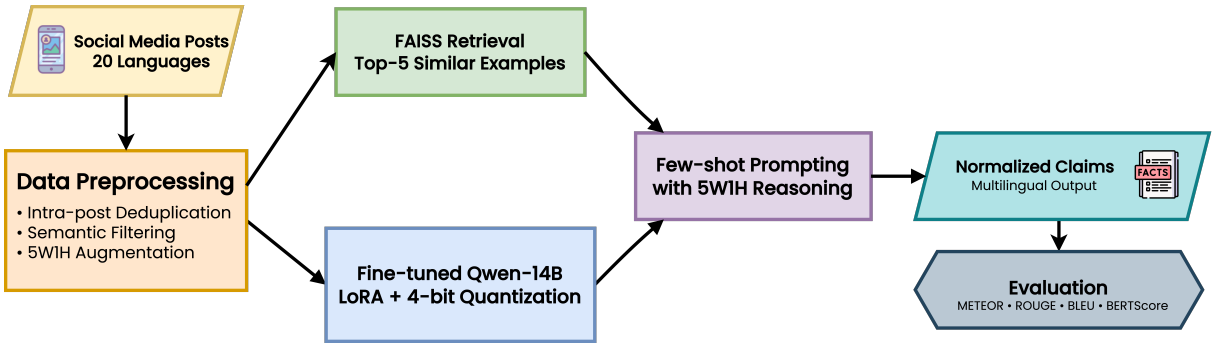

Figure 1: System workflow for multilingual claim normalization using fine-tuned Qwen3-14B with 5W1H reasoning framework and retrieval-augmented few-shot prompting.

- LoRA rank $r = 16$, scaling factor $\alpha = 32$, dropout rate 0.05
- Target modules: attention and projection layers
- Training epochs: 3
- Per-device batch size: 6, gradient accumulation steps: 4 (effective batch size: 24)
- Optimizer: paged AdamW 8-bit with learning rate 3×10^{-4}
- Precision: bfloat16 with gradient checkpointing enabled
- Hardware: Single NVIDIA A100 GPU (40GB VRAM)

3.2. Data Preprocessing

The CheckThat! Lab CLEF 2025 Task 2 dataset (Table 1) encompasses 26,399 instances across 20 languages, representing one of the most comprehensive multilingual collections for claim normalization research. The dataset exhibits significant linguistic diversity, with English comprising the largest subset (13,830 instances), followed by Spanish (4,336), Portuguese (2,183) and French (1,469). Thirteen languages provide complete training, development and test splits for monolingual evaluation, ranging from high-resource languages like English to lower-resource languages such as Tamil (252 instances) and Polish (304 instances). Seven additional languages—Bengali, Czech, Greek, Korean, Dutch, Romanian and Telugu—are included exclusively for zero-shot evaluation with 1,068 test instances total. Post lengths vary dramatically across languages and cultural contexts, from concise Tamil posts averaging 26 words to verbose Czech posts averaging 332 words, while normalized claims maintain relative consistency (8.87-19.85 words) across all languages.

3.2.1. Data Cleaning and Quality Control

We first address the inherent noise in social media posts through intra-post deduplication, identifying and removing repeated sentences within individual posts using MD5 fingerprinting of normalized text segments. This eliminates redundant content while preserving unique information. More critically, we filter post-claim pairs based on semantic alignment to ensure meaningful correlations. Using token-level recall between posts and their corresponding normalized claims, we retain only pairs with recall scores above 0.4, effectively removing instances where claims bear insufficient relation to their source posts.

Table 2 illustrates representative cases where posts and claims exhibit poor semantic alignment, justifying our recall-based filtering approach. The highlighted example demonstrates a particularly egregious case where the post discusses health workers during COVID-19, while the assigned claim addresses an unrelated conspiracy theory about empty body bags.

3.2.2. Data Augmentation through 5W1H Framework

To enhance the model's reasoning capabilities and expand training signal, we augment each original post-claim pair with structured 5W1H reasoning components [19]. For every training instance, we

Table 1
Dataset overview for CheckThat! Lab CLEF 2025 Task 2: Claim Normalization across 20 languages.

Language	Train	Dev	Test	Total	Avg. Post Length	Avg. Claim Length
Arabic	470	118	100	688	116.66	10.98
German	386	101	100	587	199.57	11.55
English	11374	1171	1285	13830	96.53	14.43
French	1174	147	148	1469	281.08	13.38
Hindi	1081	50	100	1231	29.25	19.03
Marathi	137	50	100	287	24.13	13.34
Indonesian	540	137	100	777	198.88	9.33
Punjabi	445	50	100	595	28.09	17.82
Polish	163	41	100	304	241.50	8.87
Portuguese	1735	223	225	2183	188.04	14.94
Spanish	3458	439	439	4336	226.70	13.60
Tamil	102	50	100	252	26.02	19.85
Thai	244	61	100	405	119.58	2.51
Zero-shot Languages						
Bengali	-	-	81	81	24.73	-
Czech	-	-	123	123	332.09	-
Greek	-	-	156	156	300.68	-
Korean	-	-	274	274	159.32	-
Dutch	-	-	177	177	192.04	-
Romanian	-	-	141	141	240.28	-
Telugu	-	-	116	116	24.62	-

Table 2
Examples demonstrating the necessity of recall-based filtering for post-claim alignment. Posts with low token-level recall (< 0.4) show poor semantic correlation with their assigned claims, requiring removal from the English training dataset.

Post	Claim	Recall
Photo Before Landing Of PK-320	Image shows Pakistani plane moments before crash in Karachi in May 2020	0.09
Strong people these health workers for Covid 19 ... they carry the dead bodies with one hand	Authorities planted empty body bags in 'fake' pandemic plot	0.00
AC MASJID MELEDAK, 2 JEMAAH MENINGGAL DUNIA AC MASJID MELEDAK, 2 JEMAAH MENINGGAL DUNIA AC MASJID MELEDAK, 2 JEMAAH MENINGGAL DUNIA None	Photo shows a fatal mosque blast in Bangladesh	0.00
Vladmir Putin has dropped 800 Tigers and lions across the country to push people to stay home..sana all Russia: Containment:	This photo shows a lion patrolling Russian streets during coronavirus lockdown	0.00
"Say it...you stand with.....?? ZELENSKYY 2018 5 @chrisskyarmy1 45"	Photo shows Volodymyr Zelensky holding a jersey featuring a swastika	0.00

systematically generate intermediate reasoning steps that decompose the post according to What (subject/topic), Who (individuals/organizations), Where (location), When (timing), How (process) and Why (causation). This augmentation transforms each simple post-claim pair into a rich training example that includes both the reasoning process and the final normalized claim. The expanded format provides the model with explicit guidance on how to systematically analyze social media posts before generating claims, effectively multiplying the learning signal from each original training instance. The prompt utilized has been described in Appendix A.

3.2.3. Dataset Composition

The final preprocessed dataset consists exclusively of English-language posts and their corresponding normalized claims, now enriched with structured reasoning annotations. We focus on English-language content to ensure consistency in linguistic patterns and reduce complexity during the initial training phase, while leveraging the base model's strong multilingual and reasoning capabilities for potential

cross-lingual transfer during inference. The combination of quality filtering and 5W1H augmentation results in a more robust training set that teaches the model both what to extract and how to reason through the extraction process.

3.3. Context Augmentation and Retrieval

Inspired by the GPT-RE framework for in-context learning in relation extraction [20], we implement a retrieval-augmented approach to address context-deficient posts using dense embeddings. We index the training set using FAISS [21] with embeddings from OpenAI's text-embedding-3-small model. For each post, we retrieve the top-5 most similar posts based on cosine similarity. Posts identified as semantic subsets of longer, more informative posts are replaced with their supersets during training. Following the GPT-RE methodology, during inference, the top-5 similar posts serve as few-shot examples in the prompt, providing contextual guidance for claim generation. This retrieval-based few-shot learning approach enables the model to leverage relevant examples from the training data to better understand the structure and style of effective claim normalization.

3.4. Final Dataset

The final preprocessed dataset consists exclusively of English-language posts and their corresponding normalized claims, now enriched with structured reasoning annotations. The combination of quality filtering and 5W1H augmentation results in a more robust training set that teaches the model both what to extract and how to reason through the extraction process. We evaluated our approach across 13 languages: English, German, French, Spanish, Hindi, Marathi, Punjabi, Arabic, Polish, Dutch, Bengali, Tamil and Telugu. Our primary focus centered on improving the English training set, with other languages serving as cross-lingual evaluation benchmarks to assess model generalization capabilities.

3.5. Evaluation

We evaluate model performance using standard text generation metrics: BLEU [22], ROUGE-1, ROUGE-2, ROUGE-L [23], METEOR [24] and BERTScore [25]. METEOR serves as our primary optimization metric due to its emphasis on semantic similarity over exact lexical matching, which aligns better with the goals of claim normalization.

4. Results

Our fine-tuned Qwen3-14B model demonstrates robust multilingual claim normalization capabilities across 14 languages, achieving consistent performance despite training exclusively on English data. The model exhibits strong generalization with ROUGE-1 F1 scores ranging from 2.26 (Bengali) to 46.98 (English) and METEOR scores spanning 15.21 (Marathi) to 41.16 (English). Notably, BERTScore maintains relatively high consistency across languages (83.25-95.28), indicating that the model preserves semantic coherence even when lexical overlap varies significantly. This suggests that our 5W1H reasoning framework effectively transfers cross-lingually, enabling the model to extract factual claims despite linguistic differences.

Romance Languages demonstrate exceptional performance, with Spanish (ROUGE-1 F1: 45.7, METEOR: 39.06), French (40.57, 34.41), Italian (27.9, 36.76) and Portuguese (30.92, 23.31) achieving the highest scores after English. This pattern indicates strong cross-lingual transfer within the Romance family, likely due to shared linguistic structures and cognate relationships with Latin-derived vocabulary.

Germanic Languages show moderate performance, with German achieving ROUGE-1 F1 of 30.58 and METEOR of 26.42, while Dutch records 24.89 and 17.2 respectively. The performance gap between Germanic and Romance languages suggests that morphological and syntactic similarities to English training data play a crucial role in transfer effectiveness.

Table 3
Multilingual claim normalization results across 13 languages using our fine-tuned Qwen3-14B model with 5W1H reasoning and retrieval-augmented few-shot prompting.

Language	ROUGE-1			ROUGE-2			ROUGE-L			BLEU-4	METEOR	BERTScore
	P	R	F1	P	R	F1	P	R	F1			
English (eng)	**47.88**	**49.14**	**46.98**	**30.11**	**30.55**	**29.46**	**43.92**	**44.96**	**43.11**	**22.20**	**41.16**	**90.41**
Spanish (spa)	46.05	48.62	45.70	27.38	28.86	27.08	40.69	43.01	40.41	20.27	39.06	87.70
Arabic (ara)	8.00	7.33	7.50	2.75	2.75	2.75	7.60	6.93	7.10	18.42	37.05	93.46
Tamil (ta)	29.50	27.57	27.90	9.33	8.83	8.67	26.50	27.03	27.22	18.29	36.76	95.50
French (fra)	39.69	46.13	40.57	23.21	27.19	23.72	34.69	40.52	35.55	13.23	34.41	86.77
Punjabi (pa)	9.00	8.33	8.57	0.00	0.00	0.00	9.00	8.33	8.57	10.57	26.85	92.51
German (deu)	30.64	33.36	30.58	15.74	17.29	15.81	27.40	30.42	27.68	10.24	26.42	85.47
Hindi (hi)	10.33	10.08	9.87	2.50	2.83	2.57	10.00	9.75	9.54	10.00	26.04	92.83
Telugu (te)	18.39	17.27	17.24	4.31	4.31	4.31	18.39	17.27	17.24	9.40	25.02	95.29
Polish (pol)	31.08	33.21	30.92	15.54	17.59	15.82	29.08	31.00	28.91	10.03	23.31	84.82
Bengali (bn)	3.09	2.58	2.26	0.00	0.00	0.00	3.09	2.58	2.26	6.29	20.30	90.62
Dutch (nld)	26.01	25.85	24.89	8.40	8.51	8.07	22.65	22.81	21.80	4.49	17.20	83.25
Marathi (mr)	8.83	5.94	6.18	1.60	1.50	1.55	8.83	5.94	6.18	6.29	15.21	89.31

South Asian Languages exhibit variable performance patterns. Hindi achieves reasonable scores (ROUGE-1 F1: 9.87, METEOR: 26.04), while Bengali and Marathi show limited lexical overlap but maintain semantic coherence as evidenced by their BERTScore values (90.37 and 88.51 respectively).

Arabic presents an interesting case with low lexical overlap scores (ROUGE-1 F1: 7.5) but high semantic preservation (BERTScore: 93.46), indicating that while surface-level matching is limited, the model successfully captures underlying claim semantics.

Notably, our English results (ROUGE-1 F1: 46.98, METEOR: 41.16) substantially outperform the CACN baseline [15] on their CLAN dataset (ROUGE-1 F1: 38.64, METEOR: 35.10), demonstrating the effectiveness of our fine-tuning approach with structured reasoning.

As shown in Table 4, our ablation study on English data reveals the substantial impact of each methodological component. The baseline configuration without Chain-of-Thought reasoning or few-shot retrieval achieves moderate performance (ROUGE-1 F1: 36.03, METEOR: 29.13). Introducing the 5W1H reasoning framework yields significant improvements across all metrics (ROUGE-1 F1: +4.23, METEOR: +4.98), demonstrating that structured decomposition enhances claim extraction quality. The addition of retrieval-augmented few-shot examples further amplifies performance substantially (ROUGE-1 F1: +6.72, METEOR: +7.05), with the combined approach achieving a 30.4% relative improvement in ROUGE-1 F1 and 41.3% in METEOR compared to the baseline. This progression validates our hypothesis that systematic reasoning combined with contextual examples enables more accurate and semantically coherent claim normalization.

Table 4
Ablation study showing the progressive impact of Chain-of-Thought (CoT) reasoning and few-shot retrieval on English claim normalization performance.

Configuration	ROUGE-1			ROUGE-2			ROUGE-L			BLEU-4	METEOR	BERTScore
	P	R	F1	P	R	F1	P	R	F1			
w/o CoT + w/o Few-Shot	36.93	37.99	36.03	16.43	17.10	16.02	32.40	33.36	31.63	9.27	29.13	88.71
w/ CoT + w/o Few-Shot	40.16	43.15	40.26	20.68	21.89	20.41	35.92	38.13	35.62	13.86	34.11	89.29
w/ CoT + w/ Few-Shot	**47.88**	**49.14**	**46.98**	**30.11**	**30.55**	**29.46**	**43.92**	**44.96**	**43.11**	**22.20**	**41.16**	**90.41**

Figure 2 illustrates the qualitative improvements achieved through our progressive enhancement approach. The base model generates claims that closely mirror the original post structure, while the addition of 5W1H reasoning produces more focused and coherent claims. The combination of structured reasoning with retrieval-augmented examples yields the most concise and professionally formatted normalized claims, demonstrating how each component contributes to improved claim quality.

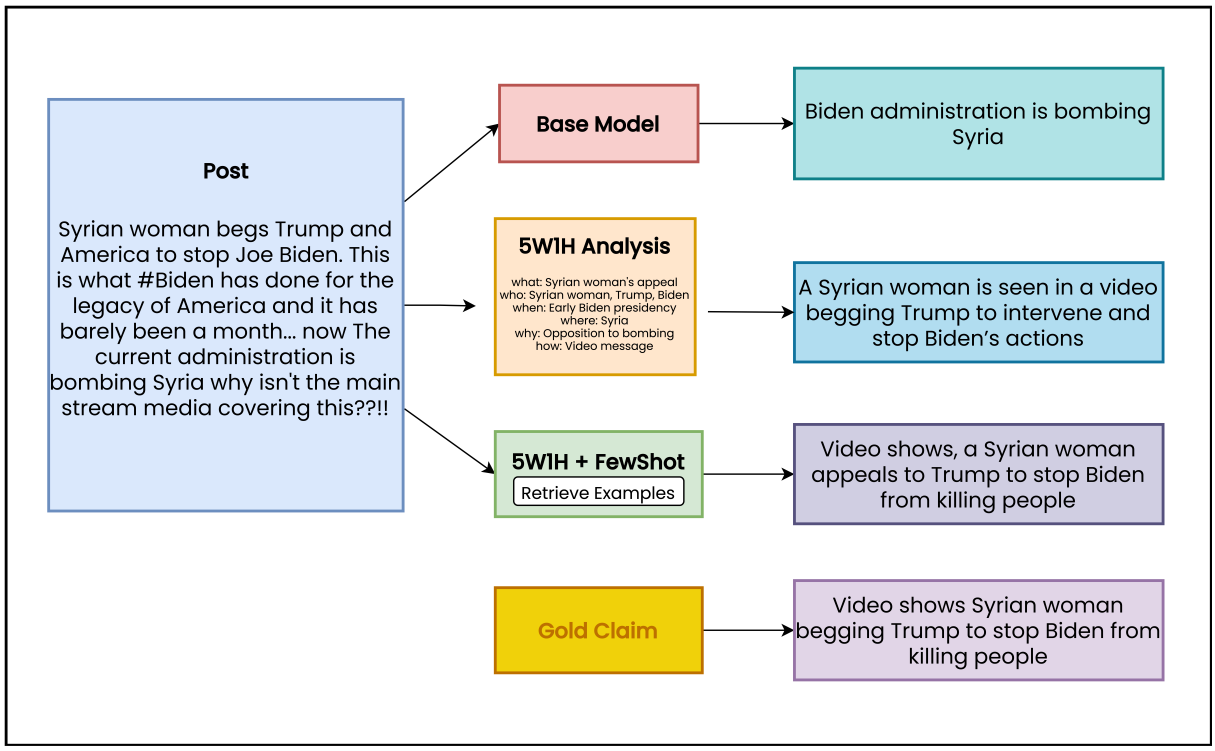

Figure 2: Progressive improvement in claim normalization quality across three configurations, showing the impact of 5W1H reasoning and few-shot retrieval on output coherence and conciseness. More examples in Appendix B

5. Conclusion and Future Work

We developed a comprehensive approach for multilingual claim normalization using fine-tuned Qwen-14B enhanced with structured 5W1H reasoning, retrieval-augmented few-shot prompting and semantic filtering techniques. Our results across 14 languages demonstrate that systematic decomposition of social media posts enables effective cross-lingual transfer despite training exclusively on English data, achieving competitive performance with third rank on the English leaderboard and fourth rank on Dutch and Punjabi leaderboards of the CheckThat! 2025 Task 2. We observed that combining structured reasoning frameworks with retrieval-based contextual examples captures the majority of performance gains while maintaining computational efficiency. Future work includes language-specific fine-tuning to accomodate additional low-resource languages, testing generalizability across different social media platforms and investigating integration with complete fact-checking pipelines for end-to-end misinformation detection systems.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) and ChatGPT in order to: perform grammar and spelling checks, improve writing style and paraphrase and reword sections for clarity and conciseness. After using these tool(s)/service(s), the author(s) thoroughly reviewed, critically evaluated and edited all content to ensure accuracy and alignment with research objectives. The author(s) take(s) full responsibility for the publication's content.

References

[1] M. Elsner, G. Atkinson, S. Zahidi, 2025. URL: https://reports.weforum.org/docs/WEF_Global_Risks_Report_2025.pdf.

[2] P. Dizikes, Study: On twitter, false news travels faster than true stories, 2018. URL: https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308.

[3] J. Calma, Meta is leaving its users to wade through hate and disinformation, 2025. URL: https://www.theverge.com/2025/1/7/24338127/meta-end-fact-checking-misinformation-zuckerberg.

[4] 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.

[5] 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.

[6] 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.

[7] 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, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[8] 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, CoRR abs/1910.10683 (2019). URL: http://arxiv.org/abs/1910.10683. arXiv:1910.10683.

[9] A. M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, in: L. Màrquez, C. Callison-Burch, J. Su (Eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 379–389. URL: https://aclanthology.org/D15-1044/. doi:10.18653/v1/D15-1044.

[10] Y. Kikuchi, G. Neubig, R. Sasano, H. Takamura, M. Okumura, Controlling output length in neural encoder-decoders, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 1328–1338. URL: https://aclanthology.org/D16-1140/. doi:10.18653/v1/D16-1140.

[11] A. Fan, D. Grangier, M. Auli, Controllable abstractive summarization, in: A. Birch, A. Finch, T. Luong, G. Neubig, Y. Oda (Eds.), Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 45–54. URL: https://aclanthology.org/W18-2706/. doi:10.18653/v1/W18-2706.

[12] W. Kryscinski, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9332–9346. URL: https://aclanthology.org/2020.emnlp-main.750/. doi:10.18653/v1/2020.emnlp-main.750.

[13] P. Utama, J. Bambrick, N. Moosavi, I. Gurevych, Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 2763–2776. URL: https://aclanthology.org/2022.naacl-main.199/. doi:10.18653/v1/2022.naacl-main.199.

[14] E. Durmus, H. He, M. Diab, FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Lin-

guistics, Association for Computational Linguistics, Online, 2020, pp. 5055–5070. URL: https://aclanthology.org/2020.acl-main.454/. doi:10.18653/v1/2020.acl-main.454.

[15] 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. URL: https://aclanthology.org/2023.findings-emnlp.439/. doi:10.18653/v1/2023.findings-emnlp.439.

[16] R. Gangi Reddy, S. C. Chinthakindi, Y. R. Fung, K. Small, H. Ji, A zero-shot claim detection framework using question answering, in: N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, S.-H. Na (Eds.), Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6927–6933. URL: https://aclanthology.org/2022.coling-1.603/.

[17] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

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

[19] Y. Cao, Y. Lan, F. Zhai, P. Li, 5w1h extraction with large language models, 2024. URL: https://arxiv.org/abs/2405.16150. arXiv:2405.16150.

[20] Z. Wan, F. Cheng, Z. Mao, Q. Liu, H. Song, J. Li, S. Kurohashi, Gpt-re: In-context learning for relation extraction using large language models, 2023. URL: https://arxiv.org/abs/2305.02105. arXiv:2305.02105.

[21] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, CoRR abs/1702.08734 (2017). URL: http://arxiv.org/abs/1702.08734. arXiv:1702.08734.

[22] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[23] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[24] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. URL: https://arxiv.org/abs/1904.09675. arXiv:1904.09675.

A. 5W1H Prompt

Below are the prompt templates used for our 5W1H reasoning framework during model training and inference.

A.1. System Prompt

Listing 1: System prompt for 5W1H claim normalization

```
You are an AI assistant that analyzes social media posts to extract factual claims. For each post,
     you will analyze it using the WH questions framework and extract the main factual claim. Make
     sure to reflect same language the post is mentioned in. If the post is in Hindi, respond in
    Hindi. Your output must be valid JSON with the following structure:

{
  "what": "Subject or topic of the post",
  "who": "Key individuals, organizations, or groups mentioned",
  "where": "Location information (if mentioned)",
  "when": "Time information (if mentioned)",
  "how": "Process information (if described)",
  "why": "Reason or motivation information (if explained)",
  "claim": "The single main factual crisp claim made in the post within 10-15 words"
}
If information for a particular field is not available, use an empty string. Also if information
    is not clearly written, don't assume anything from your end. Always stick to the post, don't
    add anything from your end. Keep things concise.
```

A.2. User Prompt Template

Listing 2: User prompt template for structured claim analysis

```
Carefully analyze the following social media post and answer each question thoughtfully to
    identify the main factual claim:

Post: {post}

Please answer each of these questions, based only on what is stated in the post:
1. What is the subject/topic of the post?
2. Who is the post talking about (key individuals, organizations, or groups)?
3. Where is this situation taking place (if mentioned)?
4. When did this situation take place (if mentioned)?
5. How did the situation take place (if described)?
6. Why did the situation take place (if explained)?

After answering these questions, extract the main factual claim being made in the post in a single
    , clear, concise sentence.

Provide your response in the specified JSON format:
{
  "what": "...",
  "who": "...",
  "where": "...",
  "when": "...",
  "how": "...",
  "why": "...",
  "claim": "..."
}
```

B. Configuration Examples

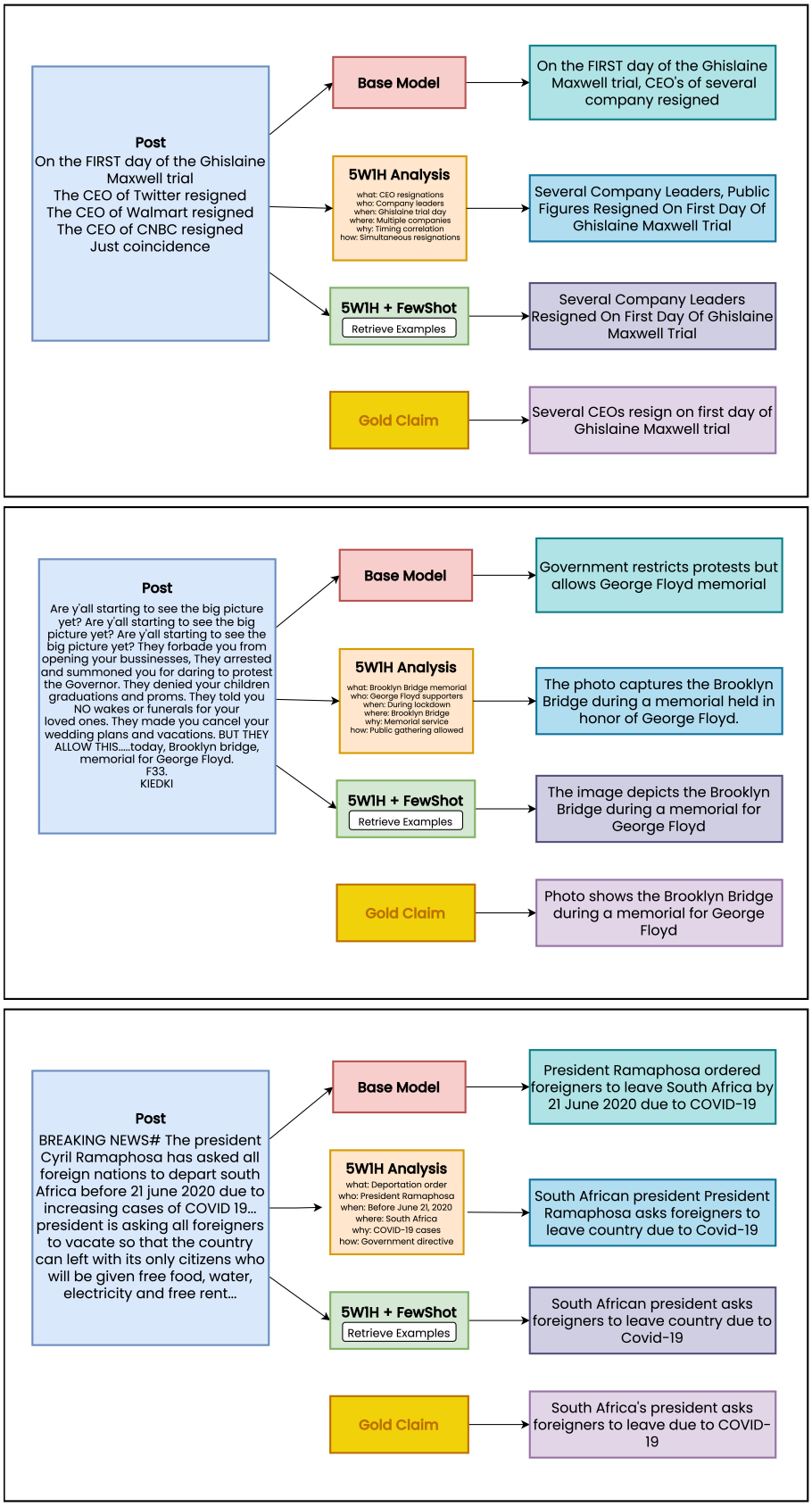

Figure 3: More examples for illustrating progressive enhancement through 5W1H reasoning and retrieval