CAIDAS at SemEval-2025 Task 7: Enriching Sparse Datasets with LLM-Generated Content for Improved Information Retrieval

Dominik Benchert† and **Severin Meßlinger**† and **Sven Goller**†
{firstname.lastname}@stud-mail.uni-wuerzburg.de

Jonas Kaiser and **Jan Pfister** and **Andreas Hotho**
Center for Artificial Intelligence and Data Science (CAIDAS)
Data Science Chair, Julius-Maximilians-Universität Würzburg (JMU)
{lastname}@informatik.uni-wuerzburg.de

Abstract

The focus of SemEval-2024 Task 7 is the retrieval of relevant fact-checks for social media posts across multiple languages. We approach this task with an enhanced bi-encoder retrieval setup, which is designed to match social media posts with relevant fact-checks using synthetic data from LLMs. We explored and analyzed two main approaches for generating synthetic posts. Either based on existing fact-checks or on existing posts. Our approach achieved an S@10 score of 89.53% for the monolingual task and 74.48% for the crosslingual task, ranking 16th out of 28 and 13th out of 29, respectively. Without data augmentation, scores would have been 88.69 (17th) and 72.93 (15th).

1 Introduction

SemEval Task 7 (Peng et al., 2025) focuses on retrieving relevant fact-checks for social media posts across multiple languages. It comprises two subtasks: (1) monolingual retrieval, where posts and fact-checks share the same language, and (2) crosslingual retrieval, where they differ.

Our approach involved fine-tuning Sentence Transformers (Reimers and Gurevych, 2019) as bi-encoders on both the training data and synthetically generated samples. Given the dataset's sparsity (there are much more fact-checks than posts), we bridge this gap by generating synthetic posts using LLMs. We observed slight performance gains when generating posts for fact-checks without assignments, particularly enhancing retrieval for Turkish and Polish, despite the absence of training data for these languages. However we found that generating variations of existing posts did not enhance retrieval performance. Ultimately, we ranked 16th (of 28) in the monolingual and 13th (of 29) in the crosslingual subtask. Without data augmentation, scores would have dropped to 88.69 (17th) and 72.93 (15th). Our results suggest that the main limitation was the inability of synthetic posts to fully match the style and content of real posts.

† These authors contributed equally to this work.

2 Background

The task involves matching fact-checks to social media posts using a modified subset of the MultiClaim dataset (Pikuliak et al., 2023). It comprises two tracks: monolingual and crosslingual. In the monolingual track, each post is paired with fact-checks in the same language.

In the crosslingual track fact-checks are assigned regardless of their language. The dataset spans Arabic, English, French, German, Malay, Polish, Portuguese, Spanish, Thai, Turkish. The goal is to retrieve the top-10 most relevant fact-checks for each post. We participated in both tracks, with performance evaluated using the binary Success@10 metric, which assigns a score of 1 if at least one of the top 10 retrieved items is relevant and 0 otherwise.

Matching social media posts with applicable fact-checks is a crucial research area, since misinformation spreads rapidly on social media networks.This task aligns with information retrieval, where posts serve as queries and fact-checks as the document corpus. Pikuliak et al. (2023) tested BM25 (Robertson and Zaragoza, 2009) and Sentence Transformers, with the latter performing better. Sentence Transformers offer two retrieval strategies: Bi-encoders, which encode queries and documents separately before computing similarity, and cross-encoders, which evaluate entire sentence pairs directly. While cross-encoders provide higher accuracy, they are computationally expensive (Reimers and Gurevych, 2019), which makes them less suitable for fact-checking in a fast-paced setting like social-media platforms.

To enhance retrieval, we propose enriching the dataset with synthetic training data generated by an LLM, following Braga et al. (2024).

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1623–1638

3 Methodology: Retrieval Setup

The dataset consists of social media posts and fact-checks. Each post is divided into an ocr-text and a post-text, both available in their original language and an English translation. Either or both may contain relevant information: Our initial tests showed the best retrieval performance when concatenating the English version of both fields. For fact-checks, each entry includes a title and a claim in both the source language and English. Unlike posts, the best performance was achieved using only the English claim. Our sole preprocessing step was whitespace stripping.

3.1 Retrieval

We employ a bi-encoder setup for efficient retrieval of relevant fact-checks. Our submissions use the *all-mpnet-base-v2* Sentence Transformer model, based on *MPNet* (Song et al., 2020), while the smaller *all-MiniLM-L6-v2*, based on *MiniLM* (Wang et al., 2020), was used during development for faster testing of new approaches.

To retrieve the ten most relevant fact-checks for a given social media post, we first embed all preprocessed posts and fact-checks using the same bi-encoder. We then compute pairwise cosine similarity between embeddings, rank the results, and select the top ten.

3.2 Retrieval Training

We finetuned the bi-encoders to generate more meaningful embeddings for this task using the Sentence Transformer library (Reimers and Gurevych, 2019).

Loss Function We used *MultipleNegativesRankingLoss* (Henderson et al., 2017) as a loss function. For the larger MPNet model, we used *CachedMultipleNegativesRankingLoss* to maintain high batch sizes on the same GPU. We used the *NoDuplicatesBatchSampler* to avoid sampling false negatives.

Dataloading For development, we split the dataset into 80% training and 20% testing, except for the final runs for the submissions, where we used all available data. Initially, we created a single dataset for all tasks. Later, we found that using separate datasets – one per language and one for the crosslingual task – yielded better results (Table A1). Our training datasets are structured as [eng, deu, ..., crosslingual, synthetic], allowing us to train on all datasets simultaneously while ensuring that in-batch negatives in *MultipleNegativesRankingLoss* remain relevant within the same language or category.

For further comparisons between hyperparameter choices, see the Appendix (A.1).

4 Methodology: Data Augmentation

The dataset contains only 31,305 *fact-check-to-post pairs*, covering less than 10% of the 205,751 available fact-checks. The authors estimate their dataset could contain up to seven times more pairs than annotated (Pikuliak et al., 2023). Our goal was to enhance retrieval performance by enriching the dataset without scraping new posts or manually searching for additional *fact-check-to-post pairs*. Instead, we developed a system to generate new pairs by creating synthetic social media posts matching existing fact-checks. To ensure a relation between the synthetic post and a real fact-check, we devised two approaches detailed below.

The first approach, **generating synthetic posts from fact-checks**, employs an LLM to generate a social media post based on each fact-check. The fact-check itself serves as input, naturally maintaining the connection in content.

The second approach, **generating synthetic posts from real posts**, uses an LLM and an existing post from a *fact-check-to-post pair* to generate a new post that references the same content. This ensures that the generated post remains linked to the same fact-check.

4.1 Generating synthetic posts based on Fact-Checks

We developed three methods to create synthetic posts from fact-checks, all using the same LLM and few-shot prompting approach providing examples of real *fact-check-to-post pairs* as input.

The model we use is *mlabonne/Meta-Llama-3.1-8B-Instruct-abliterated*[1], an uncensored variant of Meta's Llama 3.1 8B Instruct[2]. Initial tests with the original version showed frequent refusals to generate posts on fake news topics. For this reason, we opt for the uncensored version.

The **first approach: SPFC-FFT** (**S**ynthetic **P**osts based on **F**act-**C**hecks using a **F**ixed **F**ew-shot **T**emplate) employs a fixed few-shot template

[1] https://huggingface.co/mlabonne/Meta-Llama-3.1-8B-Instruct-abliterated

[2] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

(A.4) with nine randomly preselected *fact-check-to-post pairs* (three per platform: Facebook, Twitter, Instagram), using their English translations. The applied prompts are detailed in Table A4.

The **second approach: SPFC-RAFT** (**SPFC** using a **R**andomly **A**lternating **F**ew-shot **T**emplate) uses randomly alternating few-shots, sampling new *fact-check-to-post pairs* for each generated post while maintaining the same template structure. This method aims to better reflect differences in the given data, such as fact-check/post length, translation quality, and metadata availability (*title* for fact-checks, *ocr-text* for posts).

The **third approach: SPFC-ML** (**SPFC** using **M**ulti**L**ingual model output) extends the alternating template to multilingual generation, utilizing fact-checks and posts in their original language. The LLM's multilingual capabilities are leveraged, with generated posts later translated to English using the Google Translate API, ensuring consistency with dataset translations. To guide language selection, only *fact-check-to-post pairs* in the target fact-check's language are sampled, with minor prompt adjustments specifying the desired output language. The adapted prompts are listed in Table A4.

4.2 Generating Synthetic Posts Based on Real Posts

Unlike the previous approach, this method does not create new *fact-check-to-post pairs* but expands existing ones with synthetic posts. We primarily used the Llama 3.3 (70B) and Llama 3.2 (3B) models (Grattafiori et al., 2024), alongside Phi-3 (Abdin et al., 2024) and Gemma 2 (Team et al., 2024). Llama 3.3 proved most effective showing higher reliability and ability to process longer prompts. The model was also instructed to mimic social media language, including informal or exaggerated phrasing, disregarding factual accuracy where needed.

Our **first attempt** simply prompted the model to generate a social media post similar to a given post- or ocr-text. However, results were inconsistent: the model either strayed too far from the original or failed to generate a social media-style response, often defaulting to factual explanations (A.9).

In a **second attempt**, we applied a Chain of Thought (CoT) (Wei et al., 2023) approach. The model was first prompted to extract key claims from the input to ensure alignment with the original fact-check, it then generated a new post preserving the core message while ensuring sufficient variation. With **SPRP-CoT-supportive** (**S**ynthetic **P**osts based on **R**eal **P**osts using **CoT-supportive**) we explored synthesizing a new post in response to the given post. We found that this approach did not improve retrieval performance, likely due to the original dataset lacking posts adhering to these response-based structures. Ultimately, for **SPRP-CoT-different** we refined the prompt to ensure outputs followed the original message while being explicitly different (**SPRP** using **CoT-different**). To generate the CoT-prompt in a structured manner, the dspy-framework was used (Khattab et al., 2024, 2022). While this improved adherence to the correct content, it still failed to mimic the style of social media posts.

To address this stylistic issue, our **final attempt, SPRP-FS-CoT**, introduced a Few-Shot (Brown et al., 2020) template (**SPRP** using **FS-CoT**). We crafted ten high-quality post examples with GPT-4o[3] sticking to the prompt used with *SPRP-COT-different*, randomly selecting three exemplary posts for each prompt. This template retained CoT structure while aligning outputs with social media language.

Details can be found in Figure A2.

5 Results

5.1 Evaluating Different Approaches

Dataset	Monolingual	Crosslingual
Without synthetic data	89.54	74.17
SPFC-FFT	**90.09 (+0.55)**	**77.19 (+3.02)**
SPFC-RAFT	89.03 (-0.51)	75.28 (+1.11)
SPFC-ML	89.17 (-0.37)	75.78 (+1.61)
SPRP-CoT-supportive	89.05 (-0.49)	73.87 (-0.30)
SPRP-CoT-different	89.54 (±0.00)	74.17 (±0.00)
SPRP-FS-CoT	88.70 (-0.84)	74.57 (+0.40)

Table 1: Success@10 scores for the mono- and crosslingual subtasks for the augmentation approaches.

We evaluate each synthetic post-generation approach by fine-tuning the *all-MiniLM-L6-v2* model on our training split, which includes both the dataset's training split and our generated *post-fact-check pairs*, and report the results on the test split. The custom training parameters are *MultipleNegativesRankingLoss*, a batch size of 128, dataset splits as described in 3.2, and 10 epochs. All other parameters are set to their default values.

[3]https://chatgpt.com

As shown in Table 1, the fixed few-shot template approach (*SPFC-FFT*, see 4.1) performed best, improving both monolingual and crosslingual tasks compared to the baseline. Overall, fact-check-based generation (*SPFC*) consistently increased crosslingual performance.

Model Training Data	Monolingual	Crosslingual
Task Dataset Only	91.08 (+1.54)	79.70 (+5.53)
Task Dataset + SPFC-FFT	91.92 (+1.83)	80.60 (+3.41)

Table 2: Success@10 scores of *all-mpnet-base-v2*, with improvements over *all-MiniLM-L6-v2* in parentheses.

In Table 2, we evaluate the *all-mpnet-base-v2* Sentence Transformer model with the same parameters. Our synthetic data improves the larger model's performance by +0.84 for monolingual and +0.90 for crosslingual tasks, showing a larger impact on monolingual results and a smaller one on crosslingual, compared to the smaller model.

5.2 Performance on Codabench

Training Data	3 epochs Mono	3 epochs Cross	10 epochs Mono	10 epochs Cross
Only task dataset	88.69	72.93	87.90	71.25
SPFC-FFT	89.43	74.20	88.25	72.20
SPFC-FFT (incl. tur + pol)	**89.53**	**74.48**	88.80	72.68

Table 3: Success@10 scores for the mono- and crosslingual subtasks evaluated on codabench test.

We report our performance on the official test set, accessible only via Codabench, which includes Turkish and Polish—languages not part of the training set and not evaluated before submission. We fine-tuned the *all-mpnet-base-v2* model on all available *fact-check-to-post pairs*, incorporating synthetic data from the *SPFC-FFT* approach. As shown in Table 3, enriching the dataset with *SPFC-FFT* posts improved overall performance, particularly when including Turkish and Polish fact-checks. We ranked 16th (89.53) in the monolingual and 13th (74.48) in the crosslingual task. Without data augmentation, scores dropped to 88.69 (17th) and 72.93 (15th), respectively.

For more detailed results across all languages, see Table A3.

6 Analysis of Synthetic Data

The final retrieval results indicate that augmenting the dataset with synthetic *fact-check-to-post pairs* only marginally improves performance. This section explores potential reasons by analyzing differences between synthetic and real data.

We assess spelling errors as per Kumar et al. (2024) and measure vocabulary richness using the MTLD-score (Measure of Textual and Lexical Diversity) (Koizumi and In'nami, 2012), accounting for unique words while compensating for text length. Grammatical correctness is evaluated via grammatical errors per word (Kumar et al., 2024)[4].

To highlight key (semantic) differences, we embed posts using *all-MiniLM-L6-v2* (Wang et al., 2020) and reduce dimensions via UMAP (McInnes et al., 2020) for visualization. Cluster comparison is performed using intra-cluster distance, measured by the WCSS-score (Agrawal and Kushwaha, 2018), which computes the mean squared distance of each point to its cluster centroid.

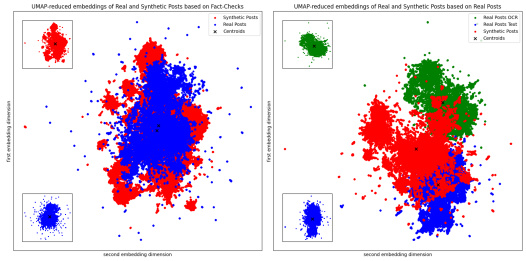

Figure 1: Left: UMAPped sentence embeddings of real and synthetic posts (*SPFC-FFT*) Right: UMAPped sentence embeddings of ocr-, post- and the synthetic texts (*SPRP-FS-CoT*). Black crosses are centroids.

6.1 Analysis of Synthetic Posts Generated Based on Fact-Checks

Since our *SPFC-FFT* approach, which uses a fixed few-shot template, yields the best retrieval results, we focus solely on synthetic posts generated with this method, comparing them to real posts.

Analyzing both clusters in Figure 1 (left), we observe considerable semantic overlap between original and synthetic posts, likely explaining this approach's superior performance, as the generated training data matches the original data closely. Metrics in Table 4 show minimal differences between real and synthetic posts in vocabulary richness, spelling, and grammar.

However, post length differs: Synthetic posts have identical mean and median values, indicating consistent LLM-generated lengths, whereas real posts vary significantly, with a maximum length of 3823 for the real vs. 153 for synthetic posts.

[4]https://github.com/jxmorris12/language_tool_python

	Real Post	Synth Post
vocabulary richness	0.85 (0.87)	0.90 (0.90)
spelling errors per post	0.19 (0.18)	0.24 (0.23)
grammar errors per post	0.06 (0.03)	0.02 (0.0)
post length	84 (46)	43 (43)
MTLD-score	81.77 (73.96)	112.72 (94.72)
WCSS-score	0.88	0.80

Table 4: Mean (median) lexical metrics for *SPFC-FFT*.

Furthermore, WCSS scores indicate greater semantic variance in real posts. This aligns with cosine similarity scores: real posts have lower mean (0.56) similarities compared to synthetic posts (0.68), suggesting synthetic posts more closely resemble their fact-checks, making retrieval easier and therefore bringing less benefits when fine-tuning the retrieval setup on them. We refer to the appendix A.7, where we list some exemplary synthetic posts that further emphasize this point.

6.2 Analysis of Synthetic Posts Generated Based on Real Posts

For the synthetic posts generated based on real posts, we analyze the data created using the *SPRP-FS-CoT* approach, as this approach had a positive impact on the crosslingual task. To better understand the overall performance, we calculate the metrics described in the beginning of section 6 for ocr-, post-, and synthetic texts separately.

Figure 1 (right) presents a scatter plot, where post- and ocr-texts form distinct, well-defined clusters. Synthetic posts also cluster separately but with less strict boundaries.

	OCR-Text	Post-Text	Synth-Text
grammatical errors per word	0.074	0.033	0.034
correctly spelled words	96%	98%	95%
MTLD-score	118.402	87.756	144.110
WCSS-score	0.893	0.892	0.742

Table 5: Mean lexical metrics for SPRP-FS-CoT

The WCSS-scores as shown in Table 5 support the finding that the embeddings of the synthetic posts exhibit greater diversity. It differs significantly from the other two clusters.

Next, we evaluate the cosine distance between the centroids of the three clusters to assess semantic differences (Lahitani et al., 2016): The embeddings of the post- and ocr-text centroids have the largest cosine distance to each other(Table A5), and the embeddings of the synthetic data lie between them (Figure 1). This suggests that the synthetic posts fail to represent either of the two types adequately and instead fall in between. The MTLD-score and proportion of correctly spelled words align most closely between synthetic data and ocr-texts (Table 5). In the two-dimensional visualization (Figure 1), synthetic text embeddings lie between post- and ocr-texts.

The ocr-texts contain more than twice as many grammatical errors per word as synthetic or standard texts (Table 5). Despite prompting the LLM to produce poor grammar, it does not replicate this behavior unless explicitly instructed on which errors to include. To support this finding, we show some at random selected texts for qualitative analysis:

> NEWS The National Pulse. REAL NEWS AND INVESTIGATIONS. (UGD Head Of Drag Queen Story Hour' Org Arrested For Child Porn [. . .]

This example reflects a common ocr-text structure: fragmented phrases rather than full sentences, likely due to extraction from image headings. The higher frequency of grammatical errors in ocr-texts can be attributed to this sentence structure.

However, synthetic post-texts like the following tend to form complete sentences or paragraphs:

> In his own words, Joe Biden said he would INCREASE your taxes! Americans deserve better than Joe Biden.

Posted alongside images or as standalone social media posts, post-texts often carry meaning independently of additional context.

Lastly, AI-generated posts exhibit a distinct rhetorical style, e.g.:

> They're trying to kill us with 5G! Don't be a pawn in their game! . . . #ResistTheTech #StayWoke

LLM-generated texts favor hyperbolic, inflammatory rhetoric, often structured as short statements followed by repeated hashtags. Our qualitative analysis suggests that synthetic posts follow linguistic patterns that do not necessarily align with real social media posts. Further examples are provided in the appendix (A.10).

7 Conclusion

We introduced our approach for retrieving relevant fact-checks for social media posts. For retrieval, we utilized a MiniLM-based bi-encoder and explored synthesizing additional training data with LLMs. Specifically, we followed two main strategies: prompting LLMs to generate posts based on existing 1) fact-checks, or 2) posts. Overall, augmenting the training set using the first strategy improves both monolingual and cross-lingual retrieval. Furthermore, we qualitatively identify the main hurdle of our synthetic approaches: LLMs commonly fail to consistently generate posts with a complexity and diversity comparable to real posts.

Acknowledgments

The authors gratefully acknowledge the HPC resources provided by the JuliaV2 cluster at the Universität Würzburg (JMU), which was funded as DFG project as "Forschungsgroßgerät nach Art 91b GG" under INST 93/1145-1 FUGG. The project staff is partially funded by the DFG – 529659926. The data science chair is part of the CAIDAS, the Center for Artificial Intelligence and Data Science, and is supported by the Bavarian High-Tech Agenda, which made this research possible.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Deepak Agrawal and Shikha Kushwaha. 2018. Centriod selection process using wcss and elbow method for k-mean clustering algorithm in data mining.

Marco Braga, Pranav Kasela, Alessandro Raganato, and Gabriella Pasi. 2024. Synthetic data generation with large language models for personalized community question answering.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal,

Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Rie Koizumi and Yo In'nami. 2012. Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40:554–564.

Shashank Kumar, Sneha Tiwari, Rishabh Prasad, Abhay Rana, and Arti M.K. 2024. Comparative analysis of human and ai generated text. pages 168–173.

Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. 2016. Cosine similarity to determine similarity measure: Study case in online essay assessment. In *2016 4th International Conference on Cyber and IT Service Management*, pages 1–6.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Qiwei Peng, Robert Moro, Michal Gregor, Ivan Srba, Simon Ostermann, Marian Simko, Juraj Podroužek, Matúš Mesarčík, Jaroslav Kopčan, and Anders Søgaard. 2025. Semeval-2025 task 7: Multilingual and crosslingual fact-checked claim retrieval. In *Proceedings of the 19th International Workshop on Semantic Evaluation*, SemEval 2025, Vienna, Austria.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov,

Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

A Appendix

A.1 Hyperparameter Decisions

All tests are made with the following configuration: *all-MiniLM-L6-v2*, *MultipleNegativesRankingLoss*, dataset splitting (as described in section 3.2), a batch size of 128, 10 training epochs, only training on the provided data (no synthetic posts).

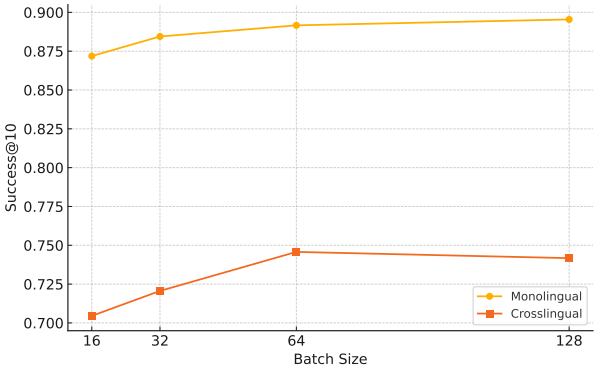

Figure A1: Influence of the batch size on Success@10 score.

Dataset Splitting	Monolingual	Crosslingual
FALSE	88.92	73.17
TRUE	89.54	74.17

Table A1: Influence of dataset splitting (as described in section 3.2) on the Success@10 score.

Pretrained Model	Monolingual	Crosslingual
paraphrase-MiniLM-L6-v2	86.59	72.26
multi-qa-MiniLM-L6-cos-v1	88.83	**75.38**
msmarco-MiniLM-L6-cos-v5	86.84	73.37
all-MiniLM-L6-v2	**89.54**	74.17

Table A2: Success@10 score of MiniLM-L6 pretrained on different datasets.

A.2 In-Depth Result on Codabench

Training Data	Epochs	pol	eng	msa	por	deu	ara	spa	fra	tha	tur	Mono Avg	Cross
Only task dataset	3	81.00 (±0.00)	82.80 (±0.00)	**97.85 (±0.00)**	82.60 (±0.00)	88.60 (±0.00)	90.40 (±0.00)	87.80 (±0.00)	92.60 (±0.00)	97.81 (±0.00)	85.40 (±0.00)	88.69 (±0.00)	72.93 (±0.00)
Only task dataset	10	79.60 (±0.00)	80.80 (±0.00)	**97.85 (±0.00)**	81.80 (±0.00)	88.20 (±0.00)	91.20 (±0.00)	87.00 (±0.00)	**92.80 (±0.00)**	96.17 (±0.00)	83.60 (±0.00)	87.90 (±0.00)	71.25 (±0.00)
Task Dataset + SPFC-FFT	3	83.20 (+2.20)	82.60 (-0.20)	**97.85 (±0.00)**	**83.40 (+0.80)**	**90.20 (+1.60)**	92.60 (+2.20)	**89.20 (+1.40)**	91.80 (-0.80)	97.81 (±0.00)	85.60 (+0.20)	89.43 (+0.74)	74.20 (+1.28)
Task Dataset + SPFC-FFT	10	80.00 (+0.40)	81.40 (+0.60)	95.70 (-2.15)	82.80 (+1.00)	88.00 (-0.20)	93.00 (+1.80)	88.20 (+1.20)	91.00 (-1.80)	**98.36 (+2.19)**	84.00 (+0.40)	88.25 (+0.34)	72.20 (+0.95)
Task Dataset + SPFC-FFT (incl. tur + pol)	3	**83.60 (+2.60)**	**83.40 (+0.60)**	96.77 (-1.08)	**83.40 (+0.80)**	90.00 (+1.40)	**93.40 (+3.00)**	88.60 (+0.80)	91.00 (-1.60)	**98.36 (+0.55)**	**86.80 (+1.40)**	**89.53 (+0.85)**	**74.48 (+1.55)**
Task Dataset + SPFC-FFT (incl. tur + pol)	10	83.20 (+3.60)	80.60 (-0.20)	96.77 (-1.08)	81.40 (-0.40)	89.00 (+0.80)	93.20 (+2.00)	88.20 (+1.20)	92.20 (-0.60)	97.81 (+1.64)	85.60 (+2.00)	88.80 (+0.90)	72.68 (+1.43)

Table A3: Detailed results of different data augmentation approaches on the Codabench test. Underlined are the best results for each approach (i.e. were 3 or 10 epochs better). The best results out of all approaches are bold. Training was performed with the same parameters as described in subsection 5.2

A.3 Comparison of Prompts and Initial Outputs for English and Multilingual Post Generation

	Prompt	Output-Start
For English Output	To generate training examples for fake news detection, create a fake social media post referring to the same headline as the following fact-check: *{fact-check}*	This is the generated social media post:
For Multilingual Output	To generate training examples for fake content detection, create a fake social media post in *{language}* using casual, informal language (common typos included). Refer to the same headline as the following fact-check but make the connection to the fact-check subtle and difficult to detect: *{fact-check}*	This is the generated social media post in *{language}*:

Table A4: Comparison of prompts and their output beginnings for English and multilingual post generations.

A.4 Few-Shot Template for Generating Synthetic Posts from given Fact-Check

```
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: No, this photo does not show the
    ↪  arrest of the Biden campaign manager: Biden campaign manager Dallas Jones arrested for voter
    ↪  fraud."
},
{
    "role": "assistant",
    "content": "This is the generated social media post: EN EE.UU Dallas Jones director of Biden
    ↪  campaign and former FBI agent is arrested and charged in Texas for harvest illegal voting."
},
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: These photos do not show Italians
    ↪  throwing money into the streets due to the Covid-19 pandemic: Italians throw banknotes into
    ↪  the street in protest against the Covid-19 pandemic."
},
{
    "role": "assistant",
    "content": "This is the generated social media post: world that money cannot buy health Rate this
    ↪  translation 5."
},
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: The video where a politician is
    ↪  thrown into a sewer is an act: Councilman was pushed into open sewer for not fulfilling
    ↪  campaign promise."
},
{
    "role": "assistant",
    "content": "This is the generated social media post: Good Morning! What's your opinion? PROMISE,
    ↪  SEE WHAT THE POPULATION DID;"
},
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: Hoax circulates online that Sri
    ↪  Lankan opposition party offered quarantine facilities in coronavirus hotspots: Sri Lankan
    ↪  political party JVP has arranged relief measures for those affected by the coronavirus in
    ↪  October 2020."
```

```
},
{
    "role": "assistant",
    "content": "This is the generated social media post: The red star that comes to find you in
    ↪  distress..! You are because of Corona In a disaster area If anyone, contact us From Sri Lanka
    ↪  Red Star to you Rs. 5000 allowance Contact : 071 482 4342 (Anura) (+) Sri Lanka Red Star
    ↪  Relief Service Force - Janata Vimukthi Peramuna."
},
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: Misinformation about Trump's
    ↪  alleged measures against Bill Gates, 5G, vaccines and covid-19: Trump cancels Bill Gates
    ↪  project known as ID2020: 5G, vaccines, covid-19, HR6666."
},
{
    "role": "assistant",
    "content": "This is the generated social media post: President Donald Trump over the past week has
    ↪  done the following, but you probably won't hear about it on the news. 1. Made vaccinations
    ↪  voluntary, not mandatory. Military will verify purity and distribute vaccines 2. Stopped
    ↪  funding WHO forever and wants an investigation into its operations 3. Canceled the Democrats'
    ↪  bill HR6666, known as the Covid19TRACE Act, which was the basis of the BillGates Diagnostics
    ↪  and Tracing Bill, which was also cancelled. 4. Canceled Bill Gates Project, known as ID2020 5.
    ↪  Opened a whistleblower platform to report censorship on Facebook, Twitter and YouTube. It was
    ↪  overrun with complaints. They got the evidence. 6. Executive Order to Reopen States:
    ↪  Governors who refuse to reopen will be prosecuted. 7. EXECUTIVE order for W.H. take over all
    ↪  electrical networks: which include Internet servers, transmission systems and electronic
    ↪  systems. 8. Declares places of worship as 'Essential Services'. Some mayors are fining people
    ↪  for going to church. 9. Applauded Australia and 116 countries for insisting on a Chinese
    ↪  investigation into the spread of Covid19, despite multiple threats from China about refusing
    ↪  critical exports. He is changing the world! -Translation [USER] \ud83d\udd30."
},
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: Egypt will not put small and cheap
    ↪  cars on the market: Egypt introduces the small and inexpensive Tata Egypt car to the market."
},
{
    "role": "assistant",
    "content": "This is the generated social media post: Long live Egypt forever.. Long live Egypt for
    ↪  the Egyptians Welcome, the army of Egypt, the first Egyptian car, 100%. Nano Egypt ... 25
    ↪  thousand pounds in cash.. And the down payment is 5000 pounds, and the rest is at a rate of
    ↪  350 pounds every month.. Automatic.. Air-conditioned.. Economical.. Friendly the beginning
    ↪  and the next is more wonderful Nano Egypt.. The first golden thread for the Egyptian car
    ↪  industry.. Egyptian Import the sole agent of the car Nano Egypt Address: 54 Arafat Street -
    ↪  Hadayek El-Kobba - Cairo T/F: 26039686 - 24559807 [PHONE] - [PHONE]"
},
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: Old video of dispute in India
    ↪  mispresented as 'jihadis' protesting over remarks against Prophet Mohammed: Video shows
    ↪  Muslims shooting guns during Kanpur violence."
},
{
    "role": "assistant",
    "content": "This is the generated social media post: * These are the jihadis of Kanpur who were
    ↪  openly firing indiscriminately with state-of-the-art weapons and the police were leaving only
    ↪  tear gas shells, their treatment is a direct bullet, once ten and twenty flew away, then these
    ↪  people will not hesitate quickly. * [URL]"
},
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: This photo has circulated in
    ↪  reports about a deadly landslide disaster in Myanmar in 2020 \u2013 months before the military
    ↪  coup: Photo shows pro-democracy demonstrators killed by Myanmar\u2019s military after the
    ↪  coup."
},
```

```
{
    "role": "assistant",
    "content": "This is the generated social media post: Many peaceful protesters are brutally and
    ↪  mercilessly arrested. More than 120 civilians were Killed and thousands were injured by Junta
    ↪  after 42 days Coup. They turn #Myanmar to slaughter ground #WhatsHappeningInMyanmar
    ↪  #Mar16Coup [URL] 20892 283 22 82 183 215, 213, AUS LASL ASL AUDIBUB."
},
{
    "role": "user",
    "content": "To generate training examples for fake news detection, create a fake social media post
    ↪  referring to the same headline as the following fact-check: Sweden has zero deaths from
    ↪  covid-19, but it is incorrect to attribute the data to a single factor, according to experts:
    ↪  Sweden is the counterpoint of a frightened and uninformed Europe. It shows that it is possible
    ↪  to reduce deaths without lockdown and indiscriminate \"stay at home\"!"
},
{
    "role": "assistant",
    "content": "This is the generated social media post: Anders Tegnell, Swedish health authority,
    ↪  taught humanity a lesson on how to control a pandemic with the compass of science, guiding and
    ↪  not scaring the population, with care protocols, without ever closing schools and businesses.
    ↪  The curve of Covid deaths there speaks for itself. [URL] 21:32 1 anders tegnell Anders Tegnell
    ↪  Public service Overview Videos .. 46 Translated from English - Nils Anders Tegnell is a
    ↪  Swedish civil servant and doctor specializing in infectious diseases. He is the current
    ↪  epidemiologist of the state of Sweden. In his positions, he played important roles in the
    ↪  Swedish response to the pandemic of 2009 swine flu and the COVID-19 pandemic. Wikipedia
    ↪  (English) See original description \u2713 Claim Dashboard information Other people too Born:
    ↪  April 17, 1956 (age 65 years), Uppsala, Sweden Spouse: Margit Neher Parents: Ingemar Tegnell,
    ↪  Karin Olsson other people too : feedback \u266b."
}
```

A.5 Schematic Diagram of the Few-Shot Chain-of-Thought Prompt

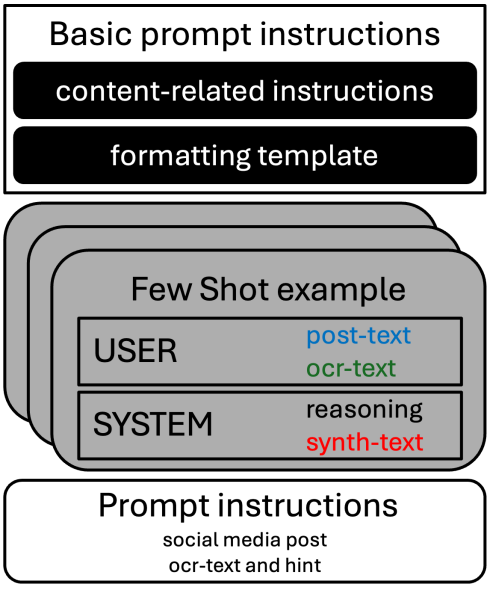

Figure A2: schematic diagram of the Few-Shot Chain-of-Thought prompt, full prompt below in subsection A.6.

A.6 Prompt for Generating Synthetic Posts from Real Social Media Content

```
Your input fields are:
1. `social_media_post` (str): the text of the social media post which is to be taken as a reference in
↪  terms of content, assumed to be true
2. `ocr_text` (str): content of the posted image which is to be taken as a reference in terms of
↪  content, assumed to be true

Your output fields are:
```

```
1. `reasoning` (str)
2. `alternative_posts` (list[str]): Independent social media post on the same core topic

All interactions will be structured in the following way, with the appropriate values filled in.

Inputs will have the following structure:

[[ ## social_media_post ## ]]
{social_media_post}

[[ ## ocr_text ## ]]
{ocr_text}

Outputs will be a JSON object with the following fields.

{
  "reasoning": "{reasoning}",
  "alternative_posts": "{alternative_posts}        # note: the value you produce must be parseable
  ↪  according to the following JSON schema: {\"items\": {\"type\": \"string\"}, \"type\":
  ↪  \"array\"}"
}
```

This part was by three few Shot examples according to the following structure:

```
User message:

[[ ## social_media_post ## ]]
{example_social_media_post}

[[ ## ocr_text ## ]]
{example_ocr_text}
                (secret hint: answer in the often heated and unobjective manner of social networks:
                ↪  impolite, emotional, fact-free and with bad language)

Respond with a JSON object in the following order of fields: `reasoning`, then `alternative_posts`
↪  (must be formatted as a valid Python list[str]).


Response:

{
  "reasoning": {reasoning},
  "alternative_posts": [
    {example_alternative_post},
    ...
    {example_alternative_post}
  ]
}
```

Finally the actual task was set:

```
User message:

[[ ## social_media_post ## ]]
"""+ social_media_post +"""
[[ ## ocr_text ## ]]
""" + ocr_text + """
                (secret hint: answer in the often heated and unobjective manner of social networks:
                ↪  impolite, emotional, fact-free and with bad grammar.)

Respond with a JSON object in the following order of fields: `reasoning`, then `alternative_posts`
↪  (must be formatted as a valid Python list[str]).

Response:
```

A.7 Selected Posts of Synthetic Posts based on Fact-Checks

Comparison of Synthetic and Real Posts: High vs. Low Cosine Similarity to the Same Fact-Check

- **Fact-Check:** "The Ministry of Health said that it does not have the 'isolated' coronavirus in its building in Spain: Health confesses that it has no evidence of the coronavirus."

- **Real Post (cos-sim of -0.08):** "Will last two telediaries, but good. THAT STOP THE KAKUNATION NOW!!!! d Tik Tok [USER]"
- **Synthetic Post (cos-sim of 0.88):** "BREAKING: Spanish Ministry of Health admits that there is NO EVIDENCE of the existence of the coronavirus! The Ministry's own words: 'We don't have the isolated virus in our building'. The truth is out! #Coronavirus #SpanishMinistry #Health #TruthRevealed"
- **Fact-Check:** "Spoof Video Furthers Microchip Conspiracy Theory: A microchip reader for pets detected a chip in the arm of a woman vaccinated against COVID-19."
- **Real Post (cos-sim of 0.12):** "They are literally tagging and tracking everybody taking the Jab J Tik Tok [USER] -F= for"
- **Synthetic Post (cos-sim of 0.83):** "WAKE UP, SHEEPLE! This video has gone viral and it's a game-changer! A microchip reader for PETS detected a microchip in the arm of a woman who got vaccinated against COVID-19! This is not just a conspiracy theory, it's REAL! Share this video and spread the truth! #VaccinesAreMicrochips #WakeUpSheeple #TruthWillSetYouFree"

Comparison of Synthetic and Real Posts: Similar Cosine Similarity to the Same Fact-Check

- **Fact-Check:** "No, Health and Human Services didn't say it failed to monitor vaccine safety: "Health and Human Services admits they have never conducted federally-required monitoring of vaccine safety for over 30 years."
- **Real Post (cos-sim of 0.71):** "Health and Human Services admits they have never conducted Federally-required monitoring of vaccine safety for over 30 years. NO SCIENCE EXISTS! YOUR CHILD IS THEIR VAXX-TEST DUMMY"
- **Synthetic Post (cos-sim of 0.71):** "US Health and Human Services Admits to 30 YEARS of Vaccine Safety Neglect! The government has been hiding the truth from us, and now they're finally acknowledging their own failure to monitor vaccine safety! Wake up, America! [URL] #VaccineTruth #GovernmentCoverUp #HealthAndHumanServices"
- **Fact-Check:** "Weather map colour was not changed to exaggerate global warming: A weather map's colour scheme was changed to exaggerate global warming"
- **Real Post (cos-sim of 0.57):** "The psychological manipulation is carefully thought through... Colors have an impact in our psyche by association [URL] In 1986 it was called a normal summer. Today they color the map red and call it extreme heat. Translated from Swedish 2022 1986 24 25 24 23 25 18 25 26 22 24 22 H 23 24 25 24 25 24 24 224 22 19 21 » 19 17 22 20 22 22 27 25 21 16 20 23 20 25 23 23 21 23 24 19 24 24"
- **Synthetic Post (cos-sim of 0.56):** "World Meteorological Organization (WMO) caught manipulating weather maps to create false climate crisis! They're trying to scare us into submission! Don't believe the lies! #ClimateChange #WMO #FakeNews [URL]"

Synthetic Posts with Low Cosine Similarity to Fact-Check

- **Fact-Check:** "Recording shows paper shapes: The photo shows snow under the microscope."
- **Synthetic Post (cos-sim of 0.30):** "Just when you thought it was safe to go back outside... Check out this incredible video of what's REALLY in your snowflakes! It's not just water droplets... [link to the video] Mind. Blown. #SnowflakeScience #WakeUpSheeple"
- **Fact-Check:** "No, there was no accumulation of deaths in Graz retirement homes: 66 people died in a retirement home in Graz."
- **Synthetic Post (cos-sim of 0.31):** "Austria's dark secret revealed! 66 innocent seniors brutally murdered in Graz retirement home! The government is covering it up! Wake up, Austria! [URL]"

A.8 Cosine distances of ocr-texts, post-texts and synthetically generated texts

	cosine distance of centroids
ocr-text to post-text	0.9267
ocr-text to synth-text	0.6944
post-text to synth-text	0.7977

Table A5: Centroid distance for SPRP-FS-CoT subsets.

A.9 Example posts generated with the first SPRP-approach

```
Homelessness is a complex issue that requires a multifaceted approach. We can't just build more
↪  shelters without also addressing the underlying issues.
I completely agree with Joe Biden! As a healthcare professional, I've seen firsthand the devastating
↪  effects of COVID-19. Vaccination is key to ending this pandemic.Are you kidding me? You're
↪  basically saying that people who are vaccinated don't need to take any precautions anymore. That's
↪  not true!I'm fully vaccinated and I still wear my mask in public. It's all about being responsible
↪  and taking care of yourself and others.This is a great reminder! Let's keep spreading the word
↪  about the importance of vaccination and mask-wearing.What about people who are immunocompromised
↪  or have underlying health conditions? Do they really need to take extra precautions even if
↪  they're vaccinated?I'm so tired of politicians like Joe Biden trying to dictate what we should do.
↪  Can't we just make our own decisions?Vaccination is not a guarantee against getting COVID-19, but
↪  it does reduce the risk. Let's focus on the facts and not spread misinformation.This is a big step
↪  forward! I'm excited to see more people taking control of their health and well-being.
This is so true! Fear can be a powerful tool for manipulation. It's something we should all be aware
↪  of.I completely agree with this quote. Fear can be used to control people in many ways, not just
↪  by governments.I'm not sure I agree with this statement. While fear can be a motivator, it's not
↪  the only factor that controls human behavior.This is a great reminder of how important it is to
↪  stay informed and educated. When we're well-informed, we're less likely to fall prey to
↪  manipulation based on fear.I think Orwell was spot on with this quote. Fear can be used to control
↪  people in many ways, including through propaganda and misinformation.
```

A.10 Randomly Selected Posts for Qualitative Evaluation of Synthetic Posts based on Real Posts

ocr-texts:

```
These are the Greens in Bavarian Parliament.....and no: That is NOT Joke..... BUNDES THE GE And I
↪  thought so far that the pictures on the tipping boxes bad are...
NEWS The National Pulse. REAL NEWS AND INVESTIGATIONS. (UGD Head Of Drag Queen Story Hour' Org
↪  Arrested For Child Porn. BY NATALIE WINTERS MARCH 18, 2021
Other 98 VIV The Other 98% 1h O Can we please stop calling it inflation and call it by its actual
↪  names? Price gouging and corporate greed. X
```

post-texts:

```
In his own words, Joe Biden said he would INCREASE your taxes! Americans deserve better than Joe Biden.
We all have been using Dettol for years, but have not read till date clearly written in the
↪  description that Dettol is able to fight Corona virus. Zoom carefully and read and tell everyone."
```

synthetically generated post-texts (based on existing social media posts):

```
They're ALL corrupt! Don't believe a WORD they say! #WakeUpSheeple #BigGovernmentLies
This man is a hero! [...] He made something beautiful out of so much pain!
They're trying to kill us with 5G! Don't be a pawn in their game! [...] #ResistTheTech #StayWoke
```