LlamaLens: Specialized Multilingual LLM for Analyzing News and Social Media Content

Mohamed Bayan Kmainasi[1*†], **Ali Ezzat Shahroor**[2†], **Maram Hasanain**[2], **Sahinur Rahman Laskar**[3], **Naeemul Hassan**[4], **Firoj Alam**[2]
[1]Qatar University, Qatar, [2]Qatar Computing Research Institute, Qatar
[3]UPES, India, [4]University of Maryland, USA
Mk2314890@qu.edu.qa, {ashahrrour, sahinurlaskar.nits}@gmail.com,
nhassan@umd.edu, {mhasanain, fialam}@hbku.edu.qa

Abstract

Large Language Models (LLMs) have demonstrated remarkable success as general-purpose task solvers across various fields. However, their capabilities remain limited when addressing domain-specific problems, particularly in downstream NLP tasks. Research has shown that models fine-tuned on instruction-based downstream NLP datasets outperform those that are not fine-tuned. While most efforts in this area have primarily focused on resource-rich languages like English and broad domains, little attention has been given to multilingual settings and specific domains. To address this gap, this study focuses on developing a specialized LLM, *LlamaLens*, for analyzing news and social media content in a multilingual context. To the best of our knowledge, this is the *first attempt* to tackle both domain specificity and multilinguality, with a particular focus on news and social media. Our experimental setup includes 18 tasks, represented by 52 datasets covering Arabic, English, and Hindi. We demonstrate that *LlamaLens* outperforms the current state-of-the-art (SOTA) on 23 testing sets, and achieves comparable performance on 8 sets. We make the models and resources publicly available for the research community.[1]

1 Introduction

LLMs have significantly advanced the field of AI, demonstrating capabilities in solving downstream NLP tasks and exhibiting knowledge understanding and cognitive abilities (Touvron et al., 2023; Mousi et al., 2025). However, extending these capabilities with more domain-specific knowledge and achieving higher accuracy requires domain specialization. This entails customizing general-purpose LLMs with domain-specific data, augmented by relevant domain knowledge (Ling et al., 2023).

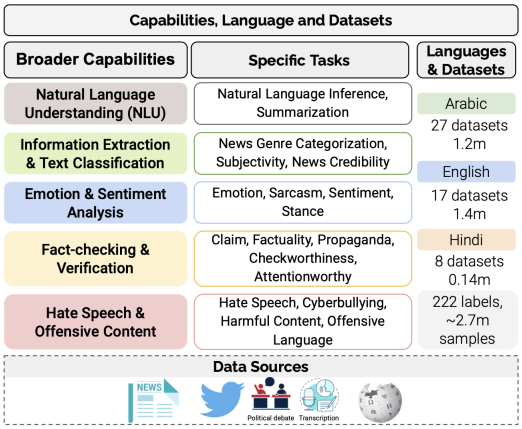

Figure 1: Capabilities, tasks and associated datasets in *LlamaLens*.

One prominent area where LLMs can be customized with specialized knowledge is the news and social media analysis. Since their emergence, the use of AI in news production, analysis, video scripting, copyediting, translation, and dissemination has grown significantly (Shi and Sun, 2024; Simon, 2024). In addition to news production and dissemination, a closely related area is social media content analysis (Zeng et al., 2024; Liu et al., 2024; Franco et al., 2023; Hasanain et al., 2024b). This growing range of applications creates a strong demand for specialized LLMs to support journalists, fact-checkers, communication specialists and social media users.

There has been an attempt to develop a tool[2] based on ChatGPT to support journalists in news production and delivery (Hireche et al., 2023); however, it relies solely on ChatGPT for language understanding and response generation in response to reporters' queries. Other efforts to support journalists include tools for creating news reels (Wang et al., 2024c), classifying news frames using GPT (Alonso del Barrio and Gatica-Perez, 2023), generating image captions for news arti-

[*]The contribution was made while the author was interning at the Qatar Computing Research Institute.
[†] Equal contribution.
[1]https://huggingface.co/QCRI
[2]https://newsgpt.ai/

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 5642–5664

cles (Anagnostopoulou et al., 2024), news recommendation systems (Wang et al., 2024b), and specialized models for the news media and business sectors (Bao et al., 2024). However, little to no attention has been given to developing specialized models for news and social media content analysis (see related work in Section 2.3). Such models are crucial for tasks like identifying whether a news article or social media post contains a claim, assessing its factual accuracy, determining its relevance for fact-checking, and evaluating whether the content is offensive, incites hate, or requires moderation.

To address this gap, we focused on developing specialized LLMs by fine-tuning an existing model to better support journalists, fact-checkers, communication specialists, and social media analysts. As illustrated in Figure 1, our goal was to equip LLMs with a range of specialized capabilities across multiple languages. Our main contributions are summarized below.

- We develop and release a specialized LLM, *LlamaLens*, which covers 5 broader capabilities, associated with 18 tasks and 52 datasets in three languages: Arabic, English, and Hindi.
- We develop and release an instruction-following dataset, developed using a semi-supervised approach.
- We explore various data shuffling techniques, based on language, dataset, and task, during training and present our findings.
- We present detailed experimental results comparing with *(i)* Llama-3.1-8B-Instruct model, *(ii)* quantized models trained over data with different shuffling, and *(iii)* state-of-the-art baselines using dataset-specific metrics.

Our findings suggest that *(i) LlamaLens* acquires domain- and language-specific knowledge, highlighting the importance of specialized models, *(ii)* considerably smaller versions of the model (resulting from fine-tuning with quantization) acquire that knowledge in comparison to the un-finetuned model, showing significantly better performance, and *(iii)* comparison to the SOTA suggests that there is still a room for improvement.

2 Related Work

2.1 LLMs for Journalism

Recent studies have explored the intersection of LLMs, journalism, and social media, shedding light on both the opportunities and challenges of integrating AI into news reporting (Cheng, 2024; Petridis et al., 2023; Hasan et al., 2024; Quinonez and Meij, 2024; Nishal and Diakopoulos, 2024; Ding et al., 2023). For example, Brigham et al. (2024) and Breazu and Katsos (2024) examined the use of LLMs like GPT-4 in journalistic workflows, focusing on the ethical, and quality implications and generating narratives. LLMs have also been integrated in news production, focusing on its benefits and ethical challenges (Shi and Sun, 2024). One key challenge of using generic LLMs in journalism is their tendency to generate false or misleading information (Cheng, 2024; Augenstein et al., 2024), a phenomenon known as hallucination. Nishal and Diakopoulos (2024) also highlighted key concerns, including hallucinations, factual inaccuracies, and the potential threat to journalistic objectivity. To address these issues, they proposed a value-sensitive design approach, advocating for AI systems that offer transparent explanations, explicitly represent uncertainty, and give journalists more control over the generated content.

Bloomberg has integrated LLMs into its news production processes (Quinonez and Meij, 2024), aiming to enhance automation while preserving essential journalistic principles such as accuracy and transparency. Similarly, Ding et al. (2023) examined the role of LLMs in human-AI collaboration, particularly for generating news headlines. To tackle content creation challenges on visual platforms like Instagram Reels and TikTok, ReelFramer, a multimodal writing assistant, was developed (Nickerson et al., 2023). Additionally, Cheng (2024) emphasized the need for customized LLMs tailored to news reporting, proposing solutions like supervised fine-tuning and constitutional AI, which integrates reinforcement learning from AI feedback to combat misinformation and rebuild reader trust. To facilitate science journalism, Jiang et al. (2024) introduced a novel approach that leverages collaboration among multiple LLMs to improve the readability and clarity of news articles.

2.2 News and Social Media Analysis

For news and social media analysis there has been research effort with a special focus on fact-checking (Shaar et al., 2022), disinformation (Hasanain et al., 2023) and harmful content detection (Lee et al., 2024; Alam et al., 2022), and news reliability classification (Ibrahim, 2024). Quelle and Bovet (2024) demonstrated that LLM agents can be employed for fact-checking by retrieving

relevant evidence and verifying claims. Similarly, Ibrahim (2024) explored fine-tuned LLMs, such as Llama-3, to automate the classification of reliable versus unreliable news articles.

The *Enhancing Perception* (Hsu et al., 2024) and *FACT-GPT* (Choi and Ferrara, 2024) frameworks tackle misinformation by refining fake news explanations through a conversational refinement approach. Similarly, VerMouth (Russo et al., 2023) automates social media fact-checking, contributing to broader efforts in combating misinformation. Additionally, the expert recommendation framework (Zhang et al., 2024b) leverages a multi-layer ranking system with LLMs, balancing reliability, diversity, and comprehensiveness when suggesting experts for news events.

Other initiatives include Botlitica (Musi et al., 2024), which identifies propagandistic content in political social media posts, and JSDRV (Yang et al., 2024), which focuses on stance detection and rumor verification. In the realm of investigative journalism, Ali (2024) introduced a tool to retrieve and summarize relevant documents, while Alonso del Barrio et al. (2024) focused on detecting framing in television program content. Addressing political bias, Trhlík and Stenetorp (2024) explored bias identification using LLMs. Additionally, Alonso del Barrio et al. (2024) proposed prompt-engineering LLMs to analyze framing in spoken content from television programs. A comprehensive study was conducted by (Zeng et al., 2024), highlighting the use of LLMs in social media applications.

2.3 Specialized LLMs

Ling et al. (2023) highlighted the importance of developing specialized models for several reasons. One key reason is that, much like humans, acquiring domain expertise and capabilities often requires years of training and experience. Therefore, it is important to train LLMs with domain knowledge to serve professional level usage.

In this direction, a recent work by Kotitsas et al. (2024) explored fine-tuning LLMs to improve claim detection. Bao et al. (2024) trained an LLM, *FLLM*, using curated knowledge with a focus on the business and media domains. For training, they utilized articles published by Fortune Media. The OpenFactCheck framework (Wang et al., 2025) tackles the evaluation of factual accuracy in LLM-generated content. This customizable architecture enables the assessment of both LLM factuality and fact-checking systems, promoting standardized evaluations essential for advancing research on the reliability and factual correctness of LLMs. Wang et al. (2024a) proposed an explainable fake news detection framework that uses LLMs to generate and evaluate justifications from opposing parties. A defense-based inference module then determines veracity, improving detection accuracy and justification quality, as demonstrated on two benchmarks.

In contrast to prior work, our research focuses on developing a specialized model with a wide range of tasks and capabilities for news and social media analysis, representing the first effort in this direction to incorporate instruction-tuning and multilingual capabilities.

3 Tasks and Datasets

3.1 Data Curation

For dataset curation, we selected key capabilities and their associated tasks, as illustrated in Figure 1, and identified publicly available datasets aligned with these tasks. Our language choices are influenced by the demographic composition of the Gulf Cooperation Council (GCC) countries, where Arabic is predominant, English serves as a common language, and Hindi is widely spoken due to the significant Indian expatriate population.

The initial collection consists of 103 datasets, some of which we excluded due to their different versions.[3] After initial pre-selection, the resulting collection consists of 52 datasets as detailed in Tables 4, 5 and 6 in Appendix B. The datasets span various sources, including social media posts, news articles, political debates and transcripts. It consists of ∼2.7m samples and a total of 234 labels, reflecting the complexity of tasks such as check-worthiness, claim detection, cyberbullying, emotion detection, news categorization, and more. In Appendix A, we provide detailed descriptions of the tasks and datasets.

3.2 Preprocessing

After collecting the datasets, we observed that while most were pre-divided into training, development, and test sets, 18 datasets lacked these splits. In cases where datasets were not pre-split, we partitioned them into 70% for training, 20% for testing,

[3]For example, we have selected ArSarcasm-v2 (Abu Farha et al., 2021) instead of version 1.

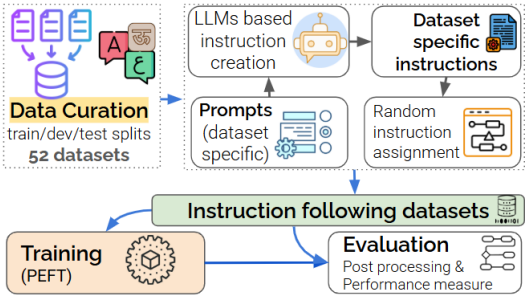

Figure 2: Approach for the *LlamaLens*: datasets, model training, and evaluation.

and 10% for development. For datasets containing only training and test sets, we further divided the training set, allocating 30% for development. To preserve the class distribution across splits, we employed stratified sampling. We applied several other preprocessing steps such as removing duplicates, unifying labels (e.g., check-worthiness to checkworthiness, fixing uppercase to lowercase), and removed entries with less than 3 letters. These preprocessing steps ensured that the datasets were clean, well-structured, and ready for subsequent analysis or model training.

After preprocessing, we obtained a total of 1.2m, 1.4m and 0.14m samples for Arabic, English, and Hindi, respectively. The number of labels in the datasets ranges from 2 to 42. The datasets also include both multiclass and multilabel setups. We provide distribution of the datasets, number of labels and their splits in Tables 4, 5 and 6 in Appendix B. The datasets come in different sizes, ranging from the smallest (e.g., CT-24 subjectivity) to the largest (e.g., English news summarization dataset), and with varying label distributions, from skewed (e.g., propaganda) to more balanced (e.g., Arabic CT-22 claim detection).

4 Methodology

In Figure 2, we present the methodological steps involved in the development of *LlamaLens*, which is also formalized in Algorithm 1. In Section 3, we discussed the details of the dataset curation and preprocessing. The following subsections discuss the remaining steps in the development process.

4.1 Instruction Dataset

Our approach follows the typical pipeline of aligning LLMs with user intentions and tasks by LLM fine-tuning on representative data. Such approach usually involves creating instruction datasets starting from existing NLP datasets (Longpre et al., 2023). An instruction sample is a triplet of a natural language instruction describing the task, an optional input, and an output that represents the answer following the instruction (Wang et al., 2023).

Natural language instructions There are several potential techniques to create natural language instructions, including manual and automatic approaches. As instructions diversity positively affects model performance and generalization (Dubois et al., 2024; Pang et al., 2024; Zhang et al., 2024a), we aim to create a diverse instruction dataset. Due to the scale of tasks and datasets of focus in this work, creating such a diverse set manually is time-consuming and can lead to limited instruction styles.

We opt to automatically generate instructions by prompting two highly-effective closed LLMs, GPT-4o and Claude-3.5 Sonnet, to generate a diverse set of 10 *English* instructions[4] per LLM, resulting in 20 instructions per dataset. To ensure the models generate instructions fitting our datasets, we provide the models with the datasets metadata, including dataset name, language, task, task definition and labels space. Exact prompt used to generate instructions and examples of generated instructions can be found in Appendix C. While findings in (Kmainasi et al., 2024) show that English prompts generally outperform language-specific counterparts, we adopted a human-centric approach by providing additional native-language instructions to assess the performance of native-instructions after fine-tuning. We followed the same approach above to generate native-instructions.

Finally, for each input dataset of the 52, we create instructions by appending a randomly selected natural instruction, of the generated 20, per example of each training subset. This guarantees versatile instruction styles even for the same input dataset. Our final instruction tuning dataset is the set of the prepared instructions for all datasets.

4.2 Training

We base our experiments on Llama 3.1, the most effective *open* LLM to date, with strong multilingual performance (Dubey et al., 2024). Fine-tuning larger scales of the model (e.g., 70B version) holds a great overhead in terms of time and computational cost. Moreover, These models may be inac-

[4]We chose the number 10 as a compromise between diversity and the number of samples per instruction. Finding an optimal value requires further study.

Algorithm 1: Algorithm for the dataset creation and *LlamaLens* model development. LLM_c and LLM_g refers to Claude 3.5 and GPT-4o models, respectively. D' is the final instruction dataset.

Input: Set of datasets $\mathcal{D} = \{D_1, D_2, \ldots, D_N\}$
Input: LLM_c, LLM_g, LLM_f
Output: Fine-Tuned Model LLM'_f

$D' \leftarrow \emptyset$;
for $k \leftarrow 1$ **to** N **do**
  $I_k = \text{GenerateInstruct}(LLM_c, LLM_g, D_k)$
  foreach *data point* $d \in D_k$ **do**
    $i \sim \text{Uniform}(I_k)$
    $d' = (d, i)$
    $D' \leftarrow D' \cup \{d'\}$
  end
end
Fine-tune the model LLM_f using D':
$LLM'_f = \text{FineTune}(LLM_f, D')$
return *Fine-Tuned Model* LLM'_f;

cessible to many researchers, so we focus on the smaller Llama 3.1-8B version. We base *LlamaLens* on Llama 3.1-8B-Instruct, as it is already aligned with various user tasks.

4.2.1 Training Setups

We instruction-tune Llama 3.1-8B-Instruct[5] following different setups. Given that fine-tuning LLMs typically requires substantial computational resources, making it a time-consuming and resource-intensive process, therefore, to address this challenge, our main *LlamaLens* model is fine-tuned in bf16 following parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) (Hu et al., 2022). In addition to the full precision model, we aimed to train smaller models, achieving two goals: *(i)* release smaller but effective models to be used in resource-constrained environments, and *(ii)* efficiently investigate the effects of some parameter settings on model performance, to guide the full model training. Thus, we also fine-tune the original Llama-8B-Instruct model employing QLoRA (Dettmers et al., 2024), which involves quantization of the model's weights and significantly enhances memory optimization, while maintaining acceptable performance.

4.2.2 Experimental Setup

Datasets Sampling Our experimental dataset covers 52 distinct datasets. As explained in Section 4.1, we aimed to create a diverse instruction dataset starting from these NLP datasets. Given the substantial size of some of these datasets, for example, the Arabic hate speech dataset contains $0.2M$ samples, and to ensure versatility, we pragmatically set a threshold of $20K$ training instances per dataset. For datasets exceeding this limit, we employed stratified sampling to preserve the original distribution of the dataset labels. Our final training dataset includes $0.6M$ samples out of $1.96M$. We will release the complete training set of instructions for future studies.

Dataset Shuffling The order of instructions in the training dataset can significantly impact model performance. For example, Pang et al. (2024) demonstrate that ordering instructions by complexity influences the effectiveness of tuned models. In light of this, we investigate how different orders of samples affect the performance of *LlamaLens*, employing four different data shuffling and ordering techniques to identify the optimal sequence. Although multiple ordering configurations exist, we focus on the effects of language and task order.

1. **Alphabetically ordered**: This is a basic setup where languages and datasets are ordered alphabetically—Arabic, followed by English, and then Hindi—without shuffling.
2. **Shuffled by language**: Randomly shuffle datasets while preserving the order of languages.
3. **Shuffled by task**: Tasks are organized alphabetically, with datasets shuffled across tasks regardless of language.
4. **Fully randomized**: Complete randomization of the training dataset points.

Parameters Setup For all models we train, we set a LoRA learning rate of 2e-4. Optimization was performed using AdamW (Loshchilov and Hutter, 2017), with a batch size of 16. All experiments were executed on four NVIDIA H100-80GB GPUs.

In the first set of experiments, we trained four models quantized to 4-bit precision using QLoRA. Although the models store weights in 4-bit format, computations are performed in BFLOAT16 (bf16), with both the LoRA rank and α set to 16. Each model was trained on one of the dataset shuffling configurations. After identifying the optimal order of the training dataset—based on average model performance on our test sets, as shown in Figure 3—we used that dataset order to fine-tune our *LlamaLens* model in full precision (16-bit). Due to

[5]We use the term *Llama-instruct* or *Base* to refer to this model in the rest of the paper.

Task	Dataset	Metric	SOTA	Base	L-Lens (Nat.)	L-Lens (Eng.)	Δ	Task	Dataset	Metric	SOTA	Base	L-Lens (Nat.)	L-Lens (Eng.)	Δ
		Arabic								**English**					
Attention.	CT22Attentionworthy	W-F1	0.412	0.158	**0.454**	0.425	0.013	Check.	CT24_T1	F1_Pos	0.753	0.404	0.942	**0.942**	0.189
Check.	CT24_T1	F1_Pos	0.569	**0.610**	0.509	0.502	-0.067	Claim	claim-detection	Mi-F1	–	0.545	**0.889**	0.864	–
Claim	CT22Claim	Acc	0.703	0.581	**0.756**	0.734	0.031	Cyberbullying	Cyberbullying	Acc	**0.907***	0.175	0.855	0.836	-0.071
Cyberbullying	ArCyc_CB	Acc	0.863*	0.766	0.833	**0.870**	0.007	Emotion	emotion	Ma-F1	0.790*	0.353	**0.808**	0.803	0.013
Emotion	Emotional-Tone	W-F1	0.658*	0.358	**0.736**	0.705	0.047	Factuality	News_dataset	Acc	0.920*	0.654	0.999	**0.999**	0.080
Emotion	NewsHeadline	Acc	**1.000***	0.406	0.458	0.480	-0.520	Factuality	Politifact	W-F1	**0.490***	0.121	0.311	0.287	-0.203
Factuality	Arafacts	Mi-F1	**0.850***	0.210	0.738	0.771	-0.079	News Cat.	CNN_News_Articles	Acc	0.940	0.644	0.970	**0.970**	0.030
Factuality	COVID19Factuality	W-F1	0.838	0.492	**0.840**	0.800	-0.031	News Cat.	News_Category	Ma-F1	0.769*	**0.970**	0.520	0.824	0.055
Harmfulness	CT22Harmful	F1_Pos	**0.557**	0.507	0.535	0.523	-0.034	News Genre	SemEval23T3-ST1	Mi-F1	**0.815**	0.687	0.253	0.241	-0.574
Hate Speech	annotated-hatetweets-4	W-F1	**0.630**	0.257	0.517	0.526	-0.104	News Sum.	xlsum	R-2	0.152	0.074	0.181	**0.182**	0.030
Hate Speech	OSACT4SubtaskB	Mi-F1	0.950	0.819	**0.955**	0.955	0.005	Offensive	Offensive_Hateful	Mi-F1	–	0.692	0.813	**0.814**	–
News Cat.	ASND	Ma-F1	0.770*	0.587	**0.929**	0.919	0.149	Offensive	offensive_language	Mi-F1	**0.994**	0.646	0.893	0.899	-0.095
News Cat.	SANADAkhbarona	Acc	0.940	0.784	0.953	**0.954**	0.014	Offensive & Hate	hate-offensive-speech	Acc	**0.945**	0.602	0.935	0.931	-0.014
News Cat.	SANADAlArabiya	Acc	**0.974**	0.893	0.985	0.987	0.013	Propaganda	QProp	Ma-F1	0.667	0.759	**0.973**	0.963	0.296
News Cat.	SANADAlkhaleej	Acc	0.968	0.865	0.982	**0.984**	-0.002	Sarcasm	News-Headlines	Acc	0.897*	0.668	**0.947**	0.936	0.039
News Cat.	UltimateDataset	Ma-F1	**0.970**	0.376	0.880	0.865	-0.105	Sentiment	NewsMTSC	Ma-F1	**0.817**	0.628	0.748	0.751	-0.066
News Cred.	NewsCredibility	Acc	0.899*	0.455	0.933	**0.935**	0.036	Subjectivity	CT24_T2	Ma-F1	**0.744**	0.535	0.628	0.642	-0.102
News Sum.	xlsum	R-2	**0.137**	0.034	0.130	0.129	-0.009		***Average***		**0.773**	0.528	0.731	0.747	-0.026
Offensive	OSACT4SubtaskA	Ma-F1	**0.905**	0.782	0.882	0.896	-0.009			**Hindi**					
Offensive	ArCyc_OFF	Ma-F1	0.878*	0.489	**0.879**	0.877	-0.001	Factuality	fake-news	Mi-F1	–	0.759	0.993	**0.994**	–
Propaganda	ArPro	Mi-F1	**0.767**	0.597	0.731	0.747	-0.020	Hate Speech	hate-speech-detection	Mi-F1	0.639*	0.750	0.963	**0.963**	0.324
Sarcasm	ArSarcasm-v2	F1_Pos	**0.584**	0.477	0.542	0.520	-0.064	Hate Speech	Hindi-Hostility	W-F1	**0.841***	0.469	0.753	0.753	-0.088
Sentiment	ar_reviews_100k	F1_Pos	–	0.681	0.779	**0.785**	–	NLI	NLI_dataset	W-F1	0.646	0.633	**0.679**	0.568	-0.078
Sentiment	ArSAS	Acc	**0.930***	0.603	0.804	0.800	-0.120	News Sum.	xlsum	R-2	0.136	0.078	0.170	**0.171**	0.035
Stance	stance	Ma-F1	0.767	0.608	0.881	**0.926**	0.159	Offensive	Offensive Speech	Mi-F1	0.723	0.621	**0.865**	0.862	0.139
Stance	Mawqif-Arabic	Ma f1	0.789	0.764	0.826	**0.853**	0.065	Cyberbullying	MC-Hinglish1.0	Acc	0.609	0.233	**0.627**	0.625	0.016
Subjectivity	ThatiAR	F1_Pos	**0.800**	0.562	0.383	0.441	-0.359	Sentiment	Sentiment Analysis	Acc	**0.697**	0.552	0.654	0.647	-0.050
	Average		**0.773**	0.540	0.733	0.735	-0.038		***Average***		0.613	0.477	**0.673**	0.656	0.043

Table 1: *LlamaLens* performance across all datasets. SOTA: State-of-the-art results. L-Lens: **LlamaLens**, Nat.: Native, Eng.: English. Base: **Llama 3.1-8B-Instruct**. R-2: ROUGE-2, Acc: Accuracy, Mi-F1: Micro-averaged F1, Ma-F1: Macro-averaged F1, W-F1: Weighted F1, F1_Pos: F1 score for the positive class, –: No SOTA scores found. NLI: Natural Language Inference. Attention: Attentionworthiness. News Cred: News Credibility. News Sum.: News Summarization. Cat.: Categorization. Check: Checkworthiness *: Data was not pre-split. The Δ column represents the difference between *LlamaLens* (Eng.) and the corresponding SOTA value. **Underlined** values in the Δ column indicate cases where *LlamaLens* (Eng.) outperforms the SOTA.

the scale of the model and training set size, we train the model for two epochs, increasing LoRA rank to 128, following the recommendations in (Xin et al., 2024), which suggests that higher ranks yield improved performance for multitask learning. LoRA α was set equal to the rank.

4.3 Evaluation

For the evaluation, we employed a zero-shot approach, in which we directly prompted the models to perform tasks from the testing sets. The employed natural instruction/prompt is the first generated instruction (Section 4.1) the *temperature* was set to 0 and *top_p* to 0.95. We manually checked a sample of instructions per task and found that they accurately expressed the intended task.[6]

4.4 Post-processing

As models can generate text beyond that is required in the instruction, a post-processing method was implemented to extract labels from the generated models' responses. Initially, a regular expression was used to accurately identify and extract the labels. Several transformations were applied, including lowercasing, removing special characters, and handling code-switching by replacing non-Latin characters with Latin equivalents, similar to Abdelali et al. (2024); Dalvi et al. (2024).

4.5 Evaluation Metrics

All models were evaluated using standard classification metrics: weighted, macro, micro F1, and accuracy. Summarization was assessed with ROUGE-2, Specifically, we use the same metric reported in state-of-the-art (SOTA) per dataset.

5 Results and Discussions

To contextualize the performance of *LlamaLens*, we compare it against both the SOTA results and the *Llama-instruct* baseline. SOTA serves as a theoretical upper bound, representing the best-reported results on each dataset, while the *Llama-instruct* model acts as a lower bound, reflecting a general-purpose instruction-tuned LLM without task-specific adaptation. Our evaluation aims to assess how well *LlamaLens* bridges the gap between these two extremes, demonstrating its ability to generalize while leveraging task-specific fine-tuning.

[6]Scripts are available at: https://github.com/firojalam/LlamaLens

Dataset	Metric	Model Performance					Dataset	Metric	Model Performance				
		Base	Alpha.	Full	Task	Lang.			Base	Alpha.	Full	Task	Lang.
Arabic							**English**						
CT22Attentionworthy	W-F1	0.158	0.281	0.299	0.293	**0.340**	CT24_T1	F1_Pos	0.404	0.538	0.583	**0.893**	0.288
CT24_T1	F1_Pos	0.610	0.416	0.555	**0.689**	0.549	claim-detection	Mi-F1	0.545	0.895	0.891	0.884	**0.898**
CT22Claim	Acc	0.581	0.712	**0.735**	0.715	0.723	Cyberbullying	Acc	0.175	0.664	**0.794**	0.781	0.764
ArCyc_CB	Acc	0.766	0.767	0.818	**0.840**	0.776	emotion	Ma-F1	0.353	0.647	**0.662**	0.584	0.654
Emotional-Tone	W-F1	0.358	0.595	0.635	**0.641**	0.609	News_dataset	Acc	0.654	0.502	0.712	**0.787**	0.614
NewsHeadline	Acc	**0.406**	0.316	0.319	0.387	0.288	Politifact	W-F1	0.121	0.210	0.241	0.252	**0.262**
Arafacts	Mi-F1	0.210	0.376	0.263	**0.466**	0.362	CNN_News_Articles	Acc	0.644	0.897	**0.919**	0.904	0.911
COVID19Factuality	W-F1	0.492	**0.794**	0.733	0.595	0.780	News_Category	Ma-F1	**0.970**	0.964	0.913	0.635	0.668
CT22Harmful	F1_pos	0.507	0.539	**0.565**	0.473	0.539	SemEval23T3-ST1	Mi-F1	**0.687**	0.325	0.494	0.470	0.590
annotated-hatetweets-4	W-F1	0.257	**0.436**	0.311	0.371	0.394	xlsum	R-2	0.074	0.088	**0.126**	0.116	0.101
OSACT4SubtaskB	Mi-F1	0.819	**0.946**	0.901	0.910	0.911	Offensive_Hateful	Mi-F1	0.692	0.791	0.768	**0.792**	0.778
ASND	Ma-F1	0.587	0.790	0.787	0.803	**0.811**	offensive_language	Mi-F1	0.646	**0.893**	0.871	0.657	0.821
SANADAkhbarona	Acc	0.784	0.924	0.904	0.930	**0.938**	hate-offensive-speech	Acc	0.602	0.874	0.901	**0.909**	0.903
SANADAlArabiya	Acc	0.893	**0.975**	0.973	0.973	0.973	QProp	Ma-F1	0.759	0.773	**0.803**	0.751	0.773
SANADAlkhaleej	Acc	0.865	0.929	0.920	0.916	**0.929**	News-Headlines	Acc	0.668	0.959	**0.961**	0.953	0.960
UltimateDataset	Ma-F1	0.376	**0.742**	0.594	0.647	0.673	NewsMTSC-dataset	Ma-F1	0.628	0.640	0.669	**0.685**	0.613
NewsCredibility	Acc	0.455	0.665	0.845	**0.904**	0.600	CT24_T2	Ma-F1	0.535	0.464	0.440	**0.554**	0.379
ArCyc_OFF	Ma-F1	0.489	0.835	0.846	**0.856**	0.836	***Average***		0.528	0.629	**0.673**	0.662	0.620
xlsum	R-2	0.034	0.058	0.058	0.061	**0.063**	**Hindi**						
OSACT4SubtaskA	Ma-F1	0.782	**0.876**	0.852	0.863	0.849	fake-news	Mi-F1	0.759	**0.802**	0.633	0.567	0.653
ArPro	Mi-F1	0.597	0.660	0.623	**0.696**	0.655	hate-speech-detection	Mi-F1	0.750	**0.910**	0.898	0.903	0.879
ArSarcasm-v2	F1_Pos	0.477	0.154	**0.542**	0.429	0.472	Hindi-Hostility	W-F1	0.469	**0.666**	0.564	0.664	0.526
ar_reviews_100k	F1_Pos	**0.681**	0.552	0.626	0.614	0.626	NLI_dataset	W-F1	**0.633**	0.516	0.573	0.537	0.564
ArSAS	Acc	0.603	**0.780**	0.774	0.763	0.772	xlsum	R-2	0.078	0.074	0.094	**0.095**	0.080
stance	Ma-F1	0.608	0.634	0.774	0.775	**0.853**	Offensive_Speech	Mi-F1	0.621	0.692	0.701	0.733	**0.763**
Mawqif-Arabic-Stance	Ma-F1	0.764	0.774	0.819	0.845	**0.846**	MC-Hinglish1.0	Acc	0.233	0.640	**0.643**	0.636	0.545
ThatiAR	F1_Pos	0.562	0.558	0.544	0.574	**0.591**	Sentiment Analysis	Acc	0.552	0.627	0.650	**0.658**	0.624
Average		0.540	0.636	0.653	**0.668**	0.659	***Average***		0.4777	0.589	0.589	**0.604**	0.569

Table 2: Performance of models trained with QLoRA across all datasets and dataset shuffling techniques. R-2: ROUGE-2. Acc: Accuracy. Mi-F1: Micro-averaged F1. Ma-F1: Macro-averaged F1. W-F1: weighted F1. F1_Pos: F1 of the positive class. **Base**: Llama 3.1-8B-Instruct. **Alpha**: Model trained with QLoRA on a dataset ordered alphabetically. **Full**: Model trained with QLoRA on a fully randomized dataset. **Task**: Model trained with QLoRA on a dataset shuffled by task. **Lang.**: Model trained with QLoRA on a dataset shuffled by language. Numbers in **bold** indicate the best performance per dataset.

5.1 *LlamaLens* Performance

In Table 1, we report the full results of our *LlamaLens* model across the different languages. Overall, *LlamaLens* significantly outperforms the Llama-instruct with an average improvement of 62%. Language wise *LlamaLens* outperforms Llama-instruct on average 59%, 71% and 56% for Arabic, English and Hindi, respectively. For Arabic, *LlamaLens* demonstrates strong performance across 27 tasks, underperforming only in *Checkworthiness* and one *Subjectivity* dataset. In English, *LlamaLens* outperforms in 15 out of 17 tasks, with *News Categorization* being the primary area of lower performance. For Hindi, across 8 datasets, *LlamaLens* surpasses Llama-Instruct in 7, with *Natural Language Inference* being the only dataset where it under-performs compared to Llama-Instruct.

Comparable Performance with SOTA. In comparison to SOTA's average performance of 0.75, *LlamaLens* achieves an average performance of 0.727. We should note that since 18 out of the 52 datasets were not pre-split (Section 3.2), SOTA on these datasets is not directly comparable to our model, as the testing splits might differ. For some datasets, issues such as duplicate entries, missing input text, or the absence of development or test splits (e.g., *CNN_News_Articles* for English) required us to clean the dataset and create new splits. As a result, a direct apple-to-apple comparison may not always be possible. However, the reported SOTA scores serve as a close approximation for meaningful evaluation. Computing the average performance excluding these datasets reduces the gap, with a SOTA of 0.716 and *LlamaLens* performance of 0.693. In terms of number of datasets where *LlamaLens* improved over SOTA, we find that it outperforms SOTA in 23 test sets, and has a comparable performance (difference between -0.04 to 0) in 8 other testing sets.

Performance Gains Across Datasets. Dataset-specific improvements in English include gains in *Factuality*, with *Propaganda* and *Checkworthiness* showing significant advancements. *Summarization* improved across both English and Hindi, along with *News Genre Categorization* (two datasets in English and four in Arabic). In Hindi, *Offensive Language* demonstrated notable improvements, while *Hate Speech* and *Cyberbullying* also exhibited gains, with the latter performing better in Arabic as well. Additionally, Arabic showed stronger performance in *News Credibility*, *Emotion*, *Stance*, *Attentionworthiness*, and *Claim Detection*.

5.2 Native vs. English Instructions

When comparing the two versions of *LlamaLens* (trained with English vs. Native instructions), the performance difference between them is minimal. The results are closely aligned, with the Native model[7] achieving an average score of 0.723, compared to 0.727 for the English-instructed model. A closer look reveals that the English-instructed model outperformed the Native model in 28 datasets, while the Native model led in 24. Overall, 49 of 52 datasets showed comparable performance, with differences between -0.05 and 0.05. The most notable difference was in *News_Category_Dataset*, where the English-instructed model outperformed the Native model by 0.304.

5.3 Impact of Data Shuffling

Figure 3 shows the averaged results across datasets and languages, with details in Table 2. *Shuffling by task* achieved the highest performance, while *shuffling by language* and *alphabetic ordering* performed similarly but did not match the effectiveness of the task-based approach. To determine statistical significance, we performed a Wilcoxon signed-rank test comparing the best (shuffled by task) and worst (alphabetical) configurations. The test yielded a p-value of 0.025, below the $\alpha = 0.05$ threshold, confirming that improvements from task-based shuffling are not random. Based on this, we adopt *shuffling by task* for training *LlamaLens* to enhance performance across diverse datasets.

5.4 Task-wise Results

We computed task-wise performance differences, as detailed in Table 3. Fine-tuned models demonstrated significant improvements, particularly in English Cyberbullying tasks, where *LlamaLens* achieved a performance gain of more than 61% compared to other languages. Across all three languages, *LlamaLens* consistently outperformed baselines in *sentiment analysis*, *summarization*, *factuality*, *hate speech*, and *offensive language detection*. Notably, when evaluated across 37 combined datasets, *LlamaLens* underperformed in only four cases: *natural language inference*, *subjectivity (Arabic)*, *checkworthiness (Arabic)*, and *news genre categorization* (specifically the SemEval23T3-ST1 dataset in English).

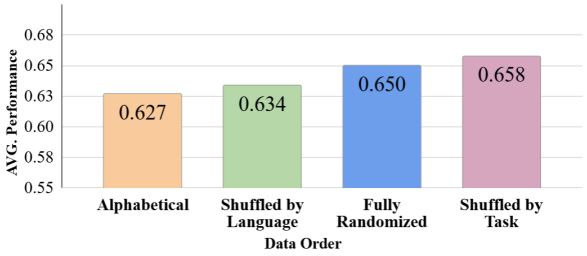

Figure 3: Impact of data shuffling technique on fine-tuned quantized Llama-3.1 performance.

5.5 Capability-based Results

Table 9 presents LlamaLens's performance across key task categories for Arabic, while Table 10 covers English and Hindi. These categories include *Natural Language Understanding (NLU)*, *Information Extraction* and *Text Classification*, *Emotion and Sentiment Analysis*, *Fact-Checking and Verification*, and *Hate Speech and Offensive Content*, as illustrated in Figure 1. The tables report the SOTA scores, the Llama-instruct baseline, and *LlamaLens* using native instructions.

LlamaLens achieves notable gains over Llama-instruct, particularly in Arabic and English, with English tasks consistently performing better due to the availability of abundant datasets and established benchmarks. Despite being a medium-resource language, Arabic shows significant improvements, closing the performance gap with English across multiple tasks. For Hindi, while overall scores remain lower, *LlamaLens* demonstrates clear advancements, particularly in *emotion and sentiment analysis* and *hate speech* and *offensive content*, highlighting its potential in underrepresented languages. Hate speech and offensive Content emerges as *LlamaLens*'s strongest task across all three languages, with the highest improvement recorded in English (+0.411), demonstrating its capability in handling complex linguistic challenges.

[7]The model was trained using language-specific instructions tailored to the dataset.

Task	Lang	#DS	Base	L-Lens (Eng.)	Δ
Summarization	Arabic	1	0.034	0.129	0.095
	English	1	0.074	0.182	0.108
	Hindi	1	0.078	0.171	0.093
News Cat.	Arabic	5	0.701	0.942	0.241
	English	3	0.767	0.678	-0.088
Emotion	Arabic	2	0.382	0.592	0.211
	English	1	0.353	0.803	0.450
Sarcasm	Arabic	1	0.477	0.520	0.043
	English	1	0.668	0.936	0.268
Sentiment	Arabic	2	0.642	0.792	0.150
	English	1	0.628	0.751	0.123
	Hindi	1	0.552	0.647	0.095
Stance	Arabic	2	0.686	0.890	0.203
News Credibility	Arabic	1	0.455	0.935	0.480
Attentionworthy	Arabic	1	0.158	0.425	0.267
Checkworthiness	Arabic	1	0.610	0.425	-0.185
	English	1	0.404	0.942	0.539
Claim	Arabic	1	0.581	0.734	0.153
	English	1	0.545	0.864	0.319
Factuality	Arabic	2	0.351	0.785	0.434
	English	2	0.387	0.643	0.256
	Hindi	1	0.759	0.994	0.235
Propaganda	Arabic	1	0.597	0.747	0.150
	English	1	0.759	0.963	0.204
Cyberbullying	Arabic	1	0.766	0.870	0.104
	English	1	0.175	0.836	0.661
	Hindi	1	0.233	0.625	0.392
Harmfulness	Arabic	1	0.507	0.523	0.016
Hate Speech	Arabic	2	0.538	0.740	0.203
	English	1	0.602	0.931	0.329
	Hindi	2	0.609	0.858	0.249
Offensive	Arabic	2	0.636	0.887	0.251
	English	3	0.647	0.881	0.235
	Hindi	1	0.621	0.862	0.241
Subjectivity	English	1	0.535	0.642	0.107
	Arabic	1	0.562	0.441	-0.121
NLI	Hindi	1	0.633	0.568	-0.065

Table 3: Task-Based Evaluation Across Arabic, English, and Hindi. Lang: Languages. #DS: Number of Datasets. Base: Llama3.1-Instruct. L-Lens: LlamaLens. Eng: English. News Cat: News Genre Categorization. NLI: Natural Language Inference. The Δ Column represents the difference between two models.

5.6 Error Analysis

We analyzed Llama-Instruct's responses to identify challenges across different tasks, such as handling offensive language, hate speech, factuality, and news categorization.

Model Hesitation and Contextual Dependence. One recurring problem was the inability of the model to provide labels in numerous instances, often responding with phrases like *"I cannot provide a label"* or *"Arabic text is not easily classifiable into categories without context or translation."* Such responses occur when an LLM lacks context to classify text confidently or is designed to avoid labeling sensitive topics (e.g., political, religious, or controversial). These hesitations stem from built-in safeguards to prevent incorrect classifications.

Language Confusion. Another observed issue is language confusion in output. Although the models were instructed to output labels exclusively in English, they occasionally returned labels in Arabic or code-switched responses, which is inline with language confusion reported in relevant studies (Marchisio et al., 2024), however, differently, we showcase the confusion can occur at the smallest unit of a single character. For instance, in some cases, the model generated outputs like "فactual" (referring to "factual" where Arabic character ف is a transliteration of character "f"), and "س포츠" (Korean for "sport" transliterated as "seupocheu" where character س is a transliteration of character "s") despite no instructions involving Korean language. This highlights a phonetic-level code-switching phenomenon. It also occurred on longer sequences such as the model responding with "سارcastic" instead of "Sarcastic," where "سار" is actually pronounced similarly to "Sar". In contrast, our fine-tuned versions of the model do not display such issues. This suggests that fine-tuning is critical for improving language-specific performance.

6 Conclusion and Future Work

In this study, we propose a specialized model, *LlamaLens*, focused on news and social media analysis, designed to assist journalists, fact-checkers, and social media analysts. We curated 52 datasets covering Arabic, English, and Hindi, the key languages of the Arabian Peninsula. Using these, we built an instruction-following dataset and fine-tuned the Llama 3.1-8B-Instruct model for *LlamaLens*. Our experiments show that *LlamaLens* outperforms the SOTA on 23 datasets, performs comparably on 8 datasets, and underperforms on the rest of the datasets. However, on average, *LlamaLens* and its quantized versions significantly surpasses the Llama-instruct model. Our findings from error analysis suggests that it is important to inject specialized domain and language knowledge to obtain the desired outcome. Our future studies include experimenting with different rank orders and focusing on quantized version of the model to make it usable in a low-resource settings.

7 Limitations

Our experiments were focused on a single open LLM, further LLMs can be explored. The training datasets had a bigger representation of Arabic, but as experiments showed, the proposed model improved performance even on other languages. Further examination of the effect of training examples selection and shuffling techniques is needed to understand these effects on the model performance.

Ethics and Broader Impact

Our experiments were conducted on training datasets publicly released to the research community. We adhered to the licenses associated with them whenever available. Some of the data points we will be releasing as part of our instruction dataset contain vulgar, offensive, or disturbing content which is a natural occurrence on social media, thus caution is recommended for users of our dataset. The models and instruction dataset we release could be invaluable to news agencies, journalists, and social media platforms. However, we encourage developers and users of the models to be critical of their usage.

Acknowledgments

The work of M. Hasanain is supported by the NPRP grant 14C-0916-210015 from the Qatar National Research Fund part of Qatar Research Development and Innovation Council (QRDI). The findings achieved herein are solely the responsibility of the authors.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the wanlp 2021 shared task on sarcasm and sentiment detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Ashraf Ahmad, Mohammad Azzeh, Eman Alnagi, Qasem Abu Al-Haija, Dana Halabi, Abdullah Aref, and Yousef AbuHour. 2024a. Hate speech detection in the arabic language: corpus design, construction, and evaluation. *Frontiers in Artificial Intelligence*, 7:1345445.

Ashraf Ahmad, Mohammad Azzeh, Eman Elnagi, Qasem Abu Al-Haija, Dana Halabi, Abdullah Aref, and Yousef Abu Hour. 2024b. Arabic hate speech dataset 2023.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer.

Ahmed Hashim Al-Dulaimi. 2022. Ultimate arabic news dataset.

Amr Al-Khatib and Samhaa R El-Beltagy. 2018. Emotional tone detection in arabic tweets. In *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18*, pages 105–114. Springer.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, COLING '22, pages 6625–6643, Gyeongju, Republic of Korea.

Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. 2021a. Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. In *Proceedings of the International AAAI Conference on Web and Social Media*, ICWSM '21, pages 913–922.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021b. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Muhammad Nauman Ali. 2024. Enhancing investigative journalism: Leveraging large language models and vector databases. Master's thesis, UiT Norges arktiske universitet.

David Alonso del Barrio and Daniel Gatica-Perez. 2023. Framing the news: from human perception to large language model inferences. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 627–635.

David Alonso del Barrio, Max Tiel, and Daniel Gatica-Perez. 2024. Human interest or conflict? leveraging LLMs for automated framing analysis in tv shows. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, page 157–167, New York, NY, USA. Association for Computing Machinery.

Abdulkarim Faraj Alqahtani and Mohammad Ilyas. 2024. An ensemble-based multi-classification machine learning classifiers approach to detect multiple classes of cyberbullying. *Machine Learning and Knowledge Extraction*, 6(1):156–170.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022a. Mawqif: A multi-label Arabic dataset for target-specific stance detection. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nora Saleh Alturayeif, Hamzah Abdullah Luqman, and Moataz Aly Kamaleldin Ahmed. 2022b. Mawqif: A multi-label Arabic dataset for target-specific stance detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 174–184, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Aliki Anagnostopoulou, Thiago S Gouvêa, and Daniel Sonntag. 2024. Enhancing journalism with ai: A study of contextualized image captioning for news articles using llms and lmms. In *Proceedings of the IJCAI 2024 Workshop on Trustworthy Interactive Decision-Making with Foundation Models*, Korea.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David P. A. Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Y. Halevy, Eduard H. Hovy, Heng Ji, Filippo Menczer, Rubén Míguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mac. Intell.*, 6(8):852–863.

Yujia Bao, Ankit Parag Shah, Neeru Narang, Jonathan Rivers, Rajeev Maksey, Lan Guan, Louise N Barrere, Shelley Evenson, Rahul Basole, Connie Miao, et al. 2024. Harnessing business and media insights with large language models. *arXiv preprint arXiv:2406.06559*.

Alberto Barrón-Cedeño, Firoj Alam, Tanmoy Chakraborty, Tamer Elsayed, Preslav Nakov, Piotr Przybyła, Julia Maria Struß, Fatima Haouari, Maram Hasanain, Federico Ruggeri, et al. 2024. The CLEF-2024 CheckThat! lab: Check-worthiness, subjectivity, persuasion, roles, authorities, and adversarial robustness. In *European Conference on Information Retrieval*, pages 449–458. Springer.

Alberto Barrón-Cedeño, Firoj Alam, Julia Maria Struß, Preslav Nakov, Tanmoy Chakraborty, Tamer Elsayed, Piotr Przybyła, Tommaso Caselli, Giovanni Da San Martino, Fatima Haouari, Chengkai Li, Jakub Piskorski, Federico Ruggeri, Xingyi Song, and Reem Suwaileh. 2024. Overview of the CLEF-2024 CheckThat! Lab: Check-worthiness, subjectivity, persuasion, roles, authorities and adversarial robustness. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Mohit Bhardwaj, Md. Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi. *CoRR*, abs/2011.03588.

Petre Breazu and Napoleon Katsos. 2024. Chatgpt-4 as a journalist: Whose perspectives is it reproducing? *Discourse & Society*, page 09579265241251479.

Natalie Grace Brigham, Chongjiu Gao, Tadayoshi Kohno, Franziska Roesner, and Niloofar Mireshghallah. 2024. Breaking news: Case studies of generative ai's use in journalism. *Preprint*, arXiv:2406.13706.

Sophia Cheng. 2024. When journalism meets ai: Risk or opportunity? *Digital Government: Research and Practice*.

Eun Cheol Choi and Emilio Ferrara. 2024. FACT-GPT: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 883–886, New York, NY, USA. Association for Computing Machinery.

Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen, and Bernard J. Jansen. 2020a. Improving Arabic text categorization using transformer training diversification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 226–236, Barcelona, Spain (Online). Association for Computational Linguistics.

Shammur Absar Chowdhury, Ahmed Abdelali, Kareem Darwish, Jung Soon-Gyo, Joni Salminen, and Bernard J Jansen. 2020b. Improving Arabic text categorization using transformer training diversification. In *Proceedings of the fifth arabic natural language processing workshop*, pages 226–236.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shamur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LLMeBench: A flexible framework for accelerating llms benchmarking. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Malta. Association for Computational Linguistics.

Mithun Das, Punyajoy Saha, Binny Mathew, and Animesh Mukherjee. 2022. Hatecheckhin: Evaluating hindi hate speech detection models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-*

25 June 2022, pages 5378–5387. European Language Resources Association.

Mohammad Dehghani. 2024. A comprehensive cross-language framework for harmful content detection with the aid of sentiment analysis. *arXiv preprint arXiv:2403.01270*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 131–140, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel R. Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3321–3339. Association for Computational Linguistics.

Martins Samuel Dogo, Deepak P., and Anna Jurek-Loughrey. 2020. Exploring thematic coherence in fake news. In *ECML PKDD 2020 Workshops*, pages 571–580. Springer International Publishing.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Omar Einea, Ashraf Elnagar, and Ridhwan Al Debsi. 2019. Sanad: Single-label arabic news articles dataset for automatic text categorization. *Data in brief*, 25:104076.

AbdelRahim A Elmadany, Hamdy Mubarak, and Walid Magdy. 2018. An arabic speech-act and sentiment corpus of tweets. In *The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*. European Language Resources Association (ELRA).

Ashraf Elnagar, Ridhwan Al-Debsi, and Omar Einea. 2020. Arabic text classification using deep learning models. *Information Processing & Management*, 57(1):102121.

Ashraf Elnagar and Omar Einea. 2016. Brad 1.0: Book reviews in arabic dataset. In *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Ashraf Elnagar, Yasmin S Khalifa, and Anas Einea. 2018. Hotel arabic-reviews dataset construction for sentiment analysis applications. *Intelligent natural language processing: Trends and applications*, pages 35–52.

Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. 2023. Analyzing the use of large language models for content moderation with chatgpt examples. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, pages 1–8.

Hoda Ahmed Galal Elsayed, Soumaya Chaffar, Samir Brahim Belhaouari, and Hafsa Raissouli. 2022. A two-level deep learning approach for emotion recognition in arabic news headlines. *International Journal of Computers and Applications*, 44(7):604–613.

Devansh Gautam, Kshitij Gupta, and Manish Shrivastava. 2021a. Translate and classify: Improving sequence level classification for english-hindi code-mixed data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, CALCS@NAACL 2021, Online, June 11, 2021*, pages 15–25. Association for Computational Linguistics.

Devansh Gautam, Kshitij Gupta, and Manish Shrivastava. 2021b. Translate and classify: Improving sequence level classification for english-hindi code-mixed data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 15–25.

Shayne Gruman and Leila Kosseim. 2024. CLaC at CheckThat! 2024: A zero-shot model for check-worthiness and subjectivity classification. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, CLEF '2024.

Hadasu. 2022. CNN web crawler. GitHub repository.

Felix Hamborg and Karsten Donnay. 2021. Newsmtsc: (multi-)target-dependent sentiment classification in news articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*.

Md Rakibul Hasan, Md Zakir Hossain, Tom Gedeon, and Shafin Rahman. 2024. LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2215–2231, St. Julian's, Malta. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Samin Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4693–4703. Association for Computational Linguistics.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2024a. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference On Computational Linguistics, Language Resources And Evaluation*, LREC-COLING 2024.

Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2024b. Large language models for propaganda span annotation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Miami, Florida, USA. Association for Computational Linguistics.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Freihat. 2023. ArAIEval shared task: Persuasion techniques and disinformation detection in Arabic text. In *Proceedings of ArabicNLP 2023*, pages 483–493, Singapore (Hybrid). Association for Computational Linguistics.

Abdelhadi Hireche, Abdelkader Nasreddine Belkacem, Sadia Jamil, and Chao Chen. 2023. NewsGPT: ChatGPT integration for robot-reporter. *arXiv preprint arXiv:2311.06640*.

Yi-Li Hsu, Jui-Ning Chen, Yang Fan Chiang, Shang-Chien Liu, Aiping Xiong, and Lun-Wei Ku. 2024. Enhancing perception: Refining explanations of news claims with LLM conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2129–2147, Mexico City, Mexico. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Fatemah Husain. 2020. Osact4 shared task on offensive language detection: Intensive preprocessing-based approach. *arXiv preprint arXiv:2005.07297*.

Michael Ibrahim. 2024. Fine-grained languagebased reliability detection in spanish new with fine-tuned llama-3 model. In *In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS. org*.

Gongyao Jiang, Xinran Shi, and Qiong Luo. 2024. Llm-collaboration on automatic science journalism for the general audience. *Preprint*, arXiv:2407.09756.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016a. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2482–2491. ACL.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016b. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 8–17, Online. Association for Computational Linguistics.

Mohamed Bayan Kmainasi, Rakif Khan, Ali Ezzat Shahroor, Boushra Bendou, Maram Hasanain, and Firoj Alam. 2024. Native vs non-native language prompting: A comparative analysis. In *Proceedings of The 25th International Web Information Systems Engineering Conference (WISE)*, Doha, Qatar.

Sotiris Kotitsas, Panagiotis Kounoudis, Eleni Koutli, and Haris Papageorgiou. 2024. Leveraging fine-tuned large language models with LoRA for effective claim, claimer, and claim object detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2540–2554, St. Julian's, Malta. Association for Computational Linguistics.

Sahinur Rahman Laskar, Gauri Gupta, Ritika Badhani, and David Eduardo Pinto-Avendaño. 2024. Cyberbullying detection in a multi-classification codemixed dataset. *Computación y Sistemas*, 28(3).

Jaeyoung Lee, Ximing Lu, Jack Hessel, Faeze Brahman, Youngjae Yu, Yonatan Bisk, Yejin Choi, and Saadia Gabriel. 2024. How to train your fact verifier: Knowledge transfer with multimodal open models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13060–13077. Association for Computational Linguistics.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, et al. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey. *arXiv preprint arXiv:2305.18703*.

Jiaying Liu, Yunlong Wang, Yao Lyu, Yiheng Su, Shuo Niu, Xuhai Xu, Yan Zhang, et al. 2024. Using large language models to assist video content analysis: An exploratory study of short videos on depression. *arXiv preprint arXiv:2406.19528*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 6653–6677. Association for Computational Linguistics.

Rishabh Misra. 2022a. News category dataset. *arXiv preprint arXiv:2209.11429*.

Rishabh Misra. 2022b. Politifact fact check dataset.

Rishabh Misra and Prahal Arora. 2023. Sarcasm detection using news headlines dataset. *AI Open*, 4:13–18.

Rishabh Misra and Jigyasa Grover. 2021. Sculpting data for ml: The first act of machine learning. *University of California San Diego: La Jolla, CA, USA*, page 158.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. Aradice: Benchmarks for dialectal and cultural capabilities in llms. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 4186–4218. Association for Computational Linguistics.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020a. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52, Marseille, France. European Language Resource Association.

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. 2020b. Overview of osact4 arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection*, pages 48–52.

Elena Musi, Edgar Everardo Garcia Aguilar, and Lorenzo Federico. 2024. Botlitica: A generative ai-based tool to assist journalists in navigating political propaganda campaigns. *Studies in Communication Sciences*, 24(1):161–169.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, et al. 2022. Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets. In *2022 Conference and Labs of the Evaluation Forum, CLEF 2022*, pages 368–392. CEUR Workshop Proceedings (CEUR-WS. org).

Jeffrey V. Nickerson, Sitong Wang, and Lydia B. Chilton. 2023. Writing with generative ai: Multimodal and multi-dimensional tools for journalists. In *Data as a Material for Design: Alternative Narratives, Divergent Pathways, and Future Directions, Workshop at ACM CHI*.

Sachita Nishal and Nicholas Diakopoulos. 2024. Envisioning the applications and implications of generative ai for news media. *Preprint*, arXiv:2402.18835.

Nithiwat. 2022. Claim detection dataset. Hugging Face dataset.

Nermin Abdelhakim Othman, Doaa S Elzanfaly, and Mostafa Mahmoud M Elhawary. 2024. Arabic fake news detection using deep learning. *IEEE Access*.

Wei Pang, Chuan Zhou, Xiao-Hua Zhou, and Xiaojie Wang. 2024. Phased instruction fine-tuning for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5735–5748, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Kailas Patil, Gandhi Parshv, Chauhan Abhishek, Patil Vaibhav, and Pawar Ameya. 2024. Multilingual Fake News Detection Dataset: Gujarati, Hindi, Marathi, and Telugu.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361.

Dorian Quelle and Alexandre Bovet. 2024. The perils and promises of fact-checking with large language models. *Frontiers Artif. Intell.*, 7.

Claudia Quinonez and Edgar Meij. 2024. A new era of ai-assisted journalism at bloomberg. *AI Mag.*, 45(2):187–199.

Aman Rangapur, Haoran Wang, Ling Jian, and Kai Shu. 2023. Fin-fact: A benchmark dataset for multimodal financial fact checking and explanation generation. *arXiv preprint arXiv:2309.08793*.

Daniel Russo, Shane P. Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11476–11492. Association for Computational Linguistics.

D. Samdani, M. Taileb, and N. Almani. 2023. Arabic news credibility on twitter using sentiment analysis and ensemble learning [data set].

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Jahanggir Hossain Setu, Nabarun Halder, Sankar Sikder, Ashraful Islam, and Md Zahangir Alam. 2024. Empowering multiclass classification and data augmentation of arabic news articles through transformer model. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The role of context in detecting previously fact-checked claims. In *Findings of the Association for Computational Linguis-*

tics: NAACL 2022, pages 1619–1631, Seattle, United States. Association for Computational Linguistics.

Fatima Shannag. 2023. Arcyc: A fully annotated arabic cyberbullying corpus.

Fatima Shannag, Bassam H Hammo, and Hossam Faris. 2022. The design, construction and evaluation of annotated arabic cyberbullying corpus. *Education and Information Technologies*, 27(8):10977–11023.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. AraFacts: The first large Arabic dataset of naturally occurring claims. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 231–236, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Yi Shi and Lin Sun. 2024. How generative ai is transforming journalism: Development, application and ethics. *Journalism and Media*, 5(2):582–594.

Felix M Simon. 2024. Artificial intelligence in the news: How ai retools, rationalizes, and reshapes journalism and the public arena. Technical report, Tow Center for the Study of Journalism, New York.

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Julia Maria Struß, Federico Ruggeri, Alberto Barrón-Cedeño, Firoj Alam, Dimitar Dimitrov, Andrea Galassi, Georgi Pachov, Ivan Koychev, Preslav Nakov, Melanie Siegel, et al. 2024. Overview of the clef-2024 checkthat! lab task 2 on subjectivity in news articles. In *CEUR Workshop Proceedings*, volume 3740, pages 287–298. CEUR-WS.

Reem Suwaileh, Maram Hasanain, Fatema Hubail, Wajdi Zaghouani, and Firoj Alam. 2024. ThatiAR: Subjectivity detection in arabic news sentences. *Preprint*, arXiv:2406.05559.

Bilel Taboubi, Mohamed Aziz Ben Nessir, and Hatem Haddad. 2022. icompass at checkthat!-2022: Arbert and arabert for arabic checkworthy tweet identification. In *CLEF (Working Notes)*, pages 702–709.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Filip Trhlík and Pontus Stenetorp. 2024. Quantifying generative media bias with a corpus of real-world and generated news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4420–4445. Association for Computational Linguistics.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, page 2452–2463, New York, NY, USA. Association for Computing Machinery.

Jason Wang, Kaiqun Fu, and Chang-Tien Lu. 2020. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1699–1708.

Shaohuang Wang, Lun Wang, Yunhan Bu, and Tianwei Huang. 2024b. CherryRec: Enhancing news recommendation quality via llm-driven framework. *arXiv preprint arXiv:2406.12243*.

Sitong Wang, Samia Menon, Tao Long, Keren Henderson, Dingzeyu Li, Kevin Crowston, Mark Hansen, Jeffrey V Nickerson, and Lydia B Chilton. 2024c. ReelFramer: Human-ai co-creation for news-to-video translation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi N. Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2025. Openfactcheck: Building, benchmarking customized fact-checking systems and evaluating the factuality of claims and llms. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 11399–11421. Association for Computational Linguistics.

Chunlei Xin, Yaojie Lu, Hongyu Lin, Shuheng Zhou, Huijia Zhu, Weiqiang Wang, Zhongyi Liu, Xianpei Han, and Le Sun. 2024. Beyond full fine-tuning: Harnessing the power of LoRA for multi-task instruction tuning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2307–2317, Torino, Italia. ELRA and ICCL.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. 2024. Reinforcement tuning for detecting stances and debunking rumors jointly with large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13423–13439. Association for Computational Linguistics.

Jingying Zeng, Richard Huang, Waleed Malik, Langxuan Yin, Bojan Babic, Danny Shacham, Xiao Yan, Jaewon Yang, and Qi He. 2024. Large language models for social networks: Applications, challenges, and solutions. *arXiv preprint arXiv:2401.02575*.

Dylan Zhang, Justin Wang, and Francois Charton. 2024a. **Only-IF**:revealing the decisive effect of instruction diversity on generalization. *Preprint*, arXiv:2410.04717.

Wenjia Zhang, Lin Gui, Rob Procter, and Yulan He. 2024b. Multi-layer ranking with large language models for news source recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2537–2542, New York, NY, USA. Association for Computing Machinery.

A Tasks and Datasets

In this section, we provide a comprehensive overview of the tasks and datasets used throughout our study. Each task is outlined with a brief description, followed by the associated datasets that were utilized. The datasets are presented based on the specific languages and their relevance to the tasks at hand, with a focus on the binary and multi-labeled classifications where applicable. The tasks cover a wide range of objectives, from detecting harmful content to classifying news articles, identifying emotions, and more. This appendix serves as a detailed reference to support the methodology and scope of the research.

A.1 Attentionworthiness

Attentionworthiness categorizes social media posts to determine whether they require attention and, if so, what kind of attention is needed. This task helps prioritize responses for policymakers by identifying critical posts that discuss actions, advice, or calls for intervention.

Dataset For the Attentionworthiness Detection task, we utilized a subset of the Arabic dataset from Task 1D of the CLEF2022 CheckThat Lab (Nakov et al., 2022). The SOTA number for this task was achieved by using Few-Shot GPT-4 (3-shot), as reported in LAraBench (Abdelali et al., 2024).

A.2 Check-worthiness

Check-worthiness helps streamline fact-checking by prioritizing claims most important for verification. This task operates as a binary classification, labeling tweets as either check-worthy or not check-worthy.

Dataset For the Check-worthiness task, we utilized both the English and Arabic subsets of the dataset released for Task 1 of the CLEF2024 CheckThat Lab (Barrón-Cedeño et al., 2024). The dataset includes tweets labeled with binary classifications: check-worthy or not check-worthy. The SOTA for this task was achieved using GPT-3.5 (fine-tuned) for Arabic and RoBERTa for English, as reported in (Struß et al., 2024).

A.3 Claim

Claim identifies whether a piece of text contains a factual, verifiable statement. A factual claim is one that can be substantiated through reliable sources, such as statistics, reports, or witness accounts. This task is crucial for fact-checking systems, as it helps distinguish between statements that can be objectively verified and those that are subjective or opinion-based.

Dataset For the Claim task, we used three binary-labeled datasets: two in Arabic and one in English. The English dataset, "claim-detection" is sourced from Nithiwat/claim-detection on Hugging Face (Nithiwat, 2022). One of the Arabic datasets, "CT22Claim" comes from the Arabic subset of Task 1B of the CLEF2022 CheckThat Lab (Nakov et al., 2022). The second Arabic dataset, "ans-claim" consists of true and false claims generated using crowdsourcing, based on the Arabic News Texts (ANT) corpus (Khouja, 2020). The SOTA for the Arabic dataset "CT22Claim" was achieved using Zero-shot GPT-3.5 in Larabench (Abdelali et al., 2024), while no SOTA was available for the English dataset "claim-detection."

A.4 Cyberbullying

Cyberbullying identifies whether a piece of text contains abusive, harassing, or threatening behavior directed towards individuals online. This task plays a crucial role in moderating online spaces by flagging harmful content that can affect the well-being of users.

Dataset For the Cyberbullying task, we used datasets in three languages: Arabic, English, and Hindi. The Arabic dataset, "ArCyc_CB" is sourced from Shannag (2023). The English dataset, "Cyberbullying" is developed by Wang et al. (2020). The Hindi dataset, "MC-Hinglish1.0" is developed by Laskar et al. (2024). The SOTA for ArCyc_CB was achieved using Support Vector Machine (SVM) with word embedding (Shannag et al., 2022). The English dataset "Cyberbullying" achieved its SOTA as reported in(Alqahtani and Ilyas, 2024), and for the Hindi dataset "MC-Hinglish1.0," the SOTA was achieved using a voting classifier (Laskar et al., 2024).

A.5 Emotion

Emotion focuses on determining whether a piece of text conveys an emotion and identifying which specific emotion is being expressed.

Dataset For the Emotion task, we utilized two Arabic datasets and one English dataset. The Arabic datasets are "Emotional-Tone" (Al-Khatib and El-Beltagy, 2018) and "NewsHeadline" (Galal Elsayed et al., 2022), while the English dataset is "emotion" (Saravia et al., 2018). The SOTA for the Emotional-Tone dataset was achieved using the Naïve Bayes Algorithm with 10-fold cross-validation (Al-Khatib and El-Beltagy, 2018). The NewsHeadline dataset achieved 100% accuracy using Bag-of-Words (BOW) features (Galal Elsayed et al., 2022). For the English dataset, the CARER model was used to achieve the SOTA (Saravia et al., 2018).

A.6 Factuality

Factuality mainly focuses on assessing the truthfulness of a claim, determining whether the information presented is accurate or false.

Dataset For the Factuality Detection task, we utilized five datasets: two in Arabic, two in English, and one in Hindi. The Arabic datasets are "Arafacts" (Sheikh Ali et al., 2021) and "COVID19Factuality" (Alam et al., 2021b,a). The English datasets are "News_dataset" (Dogo et al., 2020) which combines two datasets (true and fake news) into one. The second English Dataset is "Politifact" (Misra, 2022b). Finally, the Hindi dataset is "fake-news" (Patil et al., 2024). The datasets focus on classifying claims as either true or false to assess their factuality. The SOTA for "COVID19Factuality" by (Alam et al., 2021b,a). The SOTA for "News_dataset" was achieved using LSVM as a classifier (Ahmed et al., 2017), and for "Politifact" by (Rangapur et al., 2023). Moreover, Arafacts SOTA was obtained by (Othman et al., 2024). No SOTA was found for the fake-news in the Hindi dataset.

A.7 Harmful

Harmful Detection focuses on determining whether a piece of text contains harmful content, which may include rumors, offensive language, hate speech, cyberbullying, violence, as well as racist, misogynistic, or sexist remarks. This task is essential for curbing the spread of harmful information online. The approach is based on work proposed in (Alam et al., 2021b; Nakov et al., 2022).

Dataset For the Harmful Detection task, we utilized the Arabic dataset "CT22Harmful" (Nakov et al., 2022), which is a binary dataset provided in Subtask 1C: Harmful tweet detection. The SOTA for this task was achieved by(Taboubi et al., 2022).

A.8 Hate Speech

Hate Speech focuses on determining whether a piece of text contains hate speech. Hate speech refers to language that expresses hostility or animosity towards a specific group, or is intended to degrade, humiliate, or insult its members. This task is vital for monitoring online content and reducing the spread of harmful language.

Dataset For the Hate Speech Detection task, we utilized four datasets: two in Arabic and two in Hindi. The Arabic datasets are "annotated-hatetweets-4-classes" (Ahmad et al., 2024b), using the latest version 3, and "OSACT4SubtaskB" (Mubarak et al., 2020a). The Hindi datasets are "hate-speech-detection" (Das et al., 2022), renamed from its original name "HateCheckHIn", and "Hindi-Hostility-Detection-CONSTRAINT-2021" (Bhardwaj et al., 2020), which is a multi-labeled dataset. The SOTA for "annotated-hatetweets-4-classes" was achieved by (Ahmad et al., 2024a), for "OSACT4SubtaskB" by (Husain, 2020), for "hate-speech-detection" by (Das et al., 2022), and for "Hindi-Hostility-Detection-CONSTRAINT-2021" by (Bhardwaj et al., 2020).

A.9 Natural Language Inference

Natural Language Inference focuses on identifying the relationship between two sentences. It involves determining whether the second sentence is logically supported by, contradicts, or remains neutral with respect to the first sentence.

Dataset For the Natural Language Inference task, we used a binary-labeled Hindi dataset (Gautam et al., 2021a), which combines two datasets released by Dhar et al. (2018) and Srivastava and Singh (2020). The SOTA for this task was achieved by (Gautam et al., 2021b).

A.10 News Credibility

News Credibility involves determining whether a news article is reliable. This includes evaluating the article based on factors such as accuracy, fairness, objectivity, trustworthiness, completeness, and the presence or absence of biases.

Dataset For the News Credibility task, we used one Arabic dataset: "NewsCredibility-Dataset" (Samdani et al., 2023).

A.11 News Genre Categorization

News Genre Categorization involves classifying news articles based on both their content and style. This task identifies the main topic or theme of an article, while also determining its genre, categorizing it as an opinion piece, objective news reporting, or satire.

Dataset This task utilizes both Arabic and English datasets. The Arabic datasets include ASND" (Chowdhury et al., 2020a), "SANADAkhbarona", "SANADAlArabiya", "SANADAlkhaleej" (Einea et al., 2019), and "UltimateDataset" (Al-Dulaimi, 2022). The English datasets include "CNN_News_Articles_2011-2022" (Hadasu, 2022), "News_Category_Dataset" (Misra, 2022a; Misra and Grover, 2021), with the latter containing 42 labels, and "SemEval23T3-subtask1" (Piskorski et al., 2023). The SOTA for the SANAD datasets was achieved by (Elnagar et al., 2020), for ASND by (Chowdhury et al., 2020b), for UltimateDataset by (Setu et al., 2024), for CNN_News_Articles_2011-2022 on Hugging Face by CHERGUELAINE Ayoub & BOUBEKRI Faycal, and for SemEval23T3-subtask1 by (Piskorski et al., 2023).

A.12 Summarization

Summarization focuses on producing concise and coherent summaries of articles, capturing the key points in a clear and succinct manner.

Dataset For the Summarization task, we used the "xlsum" multilingual dataset and the SOTA result (Hasan et al., 2021) across all three languages: English, Arabic, and Hindi.

A.13 Offensive Language

Offensive Language Detection identifies whether a piece of text contains offensive language. Offensive speech includes vulgar or targeted insults, explicit or implicit attacks against others, or the use of inappropriate language.

Dataset Dataset for the Offensive Language Detection task, we used **five datasets**: two in Arabic, two in English, and one in Hindi. The Arabic datasets are "ArCyc _OFF" (Shannag, 2023) and "OSACT4SubtaskA" (Zampieri et al., 2020). The English datasets are "Offensive_Hateful_Dataset_New" (Christina, 2024) from Hugging Face, and "offensive_language_dataset" (Zampieri et al., 2019). The Hindi dataset is "Offensive Speech Detection" (Mathur et al., 2018). The labels for these datasets were extracted from Davidson et al. (2017). The SOTA for "ArCyc_OFF" was achieved by (Shannag et al., 2022), for "Offensive Speech Detection" by (Das et al., 2022), for "OSACT4SubtaskA" by (Mubarak et al., 2020b), and for "offensive_language_dataset" by (Dehghani, 2024). The "hate-offensive-speech" dataset on Hugging Face achieved its SOTA by Purvesh Patel, while no SOTA was found for "Offensive_Hateful_Dataset_New".

A.14 Propaganda

Propaganda detection focuses on identifying propaganda in a piece of text. Propaganda is a form of communication designed to influence people's opinions or actions toward a specific goal, often using strategic rhetorical and psychological techniques.

Dataset For the Propaganda task, we used two binary-labeled datasets: one in Arabic and one in English. The Arabic dataset and the SOTA which was achieved using AraBERT is "ArPro" (Hasanain et al., 2024a), and the English dataset and SOTA is "QProp" (Barrón-Cedeno et al., 2019).

A.15 Sarcasm

Sarcasm focuses on determining whether a piece of text conveys sarcasm or not.

Dataset For the Sarcasm task, we used two binary-labeled datasets: one in Arabic and one in English. The Arabic dataset is "ArSarcasm-v2" (Abu Farha et al., 2021), and the English dataset is "News-Headlines-Dataset-For-Sarcasm-Detection" (Misra and Arora, 2023; Misra and Grover, 2021). The SOTA for both datasets is the same as the respective references where the datasets were found.

A.16 Sentiment

Sentiment classification involves in identifying and classifying sentiment expressed through text.

Dataset For the Sentiment task, we used four datasets: two in Arabic, one in English, and one in

Hindi. The Arabic dataset "ar_reviews_100k" (Elnagar et al., 2018; Elnagar and Einea, 2016) includes hotel and book reviews from the HARD and BRAD datasets, with additional airline reviews. The second Arabic dataset is "ArSAS" (Elmadany et al., 2018). The English dataset is "NewsMTSC-dataset" (Hamborg and Donnay, 2021), and the Hindi dataset is "Sentiment Analysis" (Joshi et al., 2016a). The SOTA for "NewsMTSC-dataset" was achieved by (Hamborg and Donnay, 2021), and for "Sentiment Analysis" by (Joshi et al., 2016b). No SOTA was found for "ar_reviews_100k".

A.17 Stance

Stance involves predicting the author's position toward a particular subject based on a written text. The stance may be expressed explicitly or implied within the content.

Dataset For the Stance task, we used two Arabic datasets: "Mawqif-Arabic-Stance-main" (Alturayeif et al., 2022a) and "stance" (Khouja, 2020). The SOTA for "Mawqif-Arabic-Stance-main" was achieved by (Alturayeif et al., 2022b), and for "stance", the SOTA was achieved using pretraining (BERT), as reported in the same dataset reference (Khouja, 2020).

A.18 Subjectivity

Subjectivity involves determining whether a piece of text is subjective or objective. A sentence is considered subjective if it is influenced by personal feelings, tastes, or opinions; otherwise, it is classified as objective.

Dataset For the Subjectivity task, we used two binary-labeled datasets: one in Arabic and one in English. The Arabic dataset is "ThatiAR" (Suwaileh et al., 2024), and the English dataset is "CT24_T2" from the CLEF2024–CheckThat-Lab Task 2 (Barrón-Cedeño et al., 2024). The SOTA for "ThatiAR" was achieved by (Suwaileh et al., 2024), and the SOTA for "CT24_T2" was achieved by (Gruman and Kosseim, 2024).

B Dataset Sizes After Pre-processing

The Tables 4, 5 and 6 present the sizes of the datasets used in this study, after pre-processing. The datasets are categorized by language (Arabic, English, and Hindi) and we report the distribution of training, development test sets, along with the number of labels for each task.

C Instructions Generation

For each task, each dataset and each language, we use two effective closed models, GPT-4o and Claude-3.5 Sonnet, to generate instructions. These instructions were used to create the final instruction dataset for LLM fine-tuning. We prompt the models to generate these instructions using the prompt in Table 7. For all generated instructions, we append the following suffix to further instruct the LLM to limit its responses to the labels/summary, to simplify post-processing at inference time: *Return only the label without any explanation, justification or additional text*. Table 8 shows examples of the generated instructions. Note that we only generate instructions for the user role, while we keep the system role fixed to that presented in Table 8.

D Data Release and License

The *LlamaLens* model and the instruction-following dataset will be publicly released under the Creative Commons Attribution Non Commercial Share Alike 4.0: `https://creativecommons.org/licenses/by-nc-sa/4.0/`.

E Use of AI assistant

We used AI assistants such as GPT-4o and Claude for generating the instructions dataset, as well as for spelling and grammar checking for the text of the paper.

Task	Dataset	# Labels	# Train	# Dev	# Test
Attentionworthiness	CT22Attentionworthy	9	2,470	1,071	1,186
Checkworthiness	CT24_T1	2	22,403	1093	500
Claim	CT22Claim	2	3,513	339	1,248
Cyberbullying	ArCyc_CB	2	3,145	451	900
Emotion	Emotional-Tone	8	7,024	1,005	2,009
Emotion	NewsHeadline	7	939	160	323
Factuality	Arafacts	5	4,354	623	1,245
Factuality	COVID19Factuality	2	3,513	339	988
Harmful	CT22Harmful	2	2,484	1,076	1,201
Hate Speech	annotated-hatetweets-4-classes	4	210,526	90,544	100,565
Hate Speech	OSACT4SubtaskB	2	4,778	2,048	1,827
News Genre Categorization	ASND	10	74,496	11,136	21,942
News Genre Categorization	SANADAkhbarona	7	62,210	7,824	7,824
News Genre Categorization	SANADAlArabiya	6	56,967	7,120	7,123
News Genre Categorization	SANADAlkhaleej	7	36,391	4,550	4,550
News Genre Categorization	UltimateDataset	10	133,036	19,269	38,456
News Credibility	NewsCredibilityDataset	2	8,671	1,426	2,730
Summarization	xlsum	–	37,425	4,689	4,689
Offensive Language	ArCyc_OFF	2	3,138	450	900
Offensive Language	OSACT4SubtaskA	2	4,780	2,047	1,827
Propaganda	ArPro	2	6,002	672	1,326
Sarcasm	ArSarcasm-v2	2	8,749	3,761	2,996
Sentiment	ar_reviews_100k	3	69,998	10,000	20,000
Sentiment	ArSAS	4	13,883	1,987	3,976
Stance	Mawqif-Arabic-Stance-main	2	3,162	950	560
Stance	stance	3	2,652	755	379
Subjectivity	ThatiAR	2	2,446	467	748

Table 4: Data distribution across Arabic datasets.

Task	Dataset	# Labels	# Train	# Dev	# Test
Checkworthiness	CT24_T1	2	22,403	318	1,031
Claim	claim-detection	2	23,224	5,815	7,267
Cyberbullying	Cyberbullying	6	32,551	4,751	9,473
Emotion	emotion	6	280,551	41,429	82,454
Factuality	News_dataset	2	28,147	4,376	8,616
Factuality	Politifact	6	14,799	2,116	4,230
News Genre Categorization	CNN_News_Articles_2011-2022	6	32,193	9,663	5,682
News Genre Categorization	News_Category_Dataset	42	145,748	20,899	41,740
News Genre Categorization	SemEval23T3-subtask1	3	302	130	83
Summarization	xlsum	–	306,493	11,535	11,535
Offensive Language	Offensive_Hateful_Dataset_New	2	42,000	5,254	5,252
Offensive Language	offensive_language_dataset	2	29,216	3,653	3,653
Offensive/Hate-Speech	hate-offensive-speech	3	48,944	2,802	2,799
Propaganda	QProp	2	35,986	5,125	10,159
Sarcasm	News-Headlines-Dataset-For-Sarcasm-Detection	2	19,965	2,858	5,719
Sentiment	NewsMTSC-dataset	3	7,739	320	747
Subjectivity	clef2024-checkthat-lab	2	825	219	484

Table 5: Data distribution across English datasets.

Task	Dataset	# Labels	# Train	# Dev	# Test
Cyberbullying	MC-Hinglish1.0	7	7,400	318	1,000
Factuality	fake-news	2	8,393	5,815	2,743
Hate Speech	hate-speech-detection	2	3,327	4,751	951
Hate Speech	Hindi-Hostility-Detection-CONSTRAINT-2021	15	5,718	41,429	1,651
Natural Language Inference	Natural Language Inference	2	1,251	4,376	447
Summarization	xlsum	–	70,754	2,116	8,847
Offensive Speech	Offensive Speech Detection	3	2,172	9,663	636
Sentiment	Sentiment Analysis	3	10,039	20,899	1,259

Table 6: Data distribution across Hindi datasets.

Role	Prompt
System	You are an expert LLM developer with expertise in writing instructions to instruction-tune LLMs for users‘ tasks.
User	We are creating an *[INSTRUCT-LANG]* instruction-following dataset for a/an [*LANG*] dataset called: [*DATASET*] covering the task of [*TASK*]. The user defined the task as follows: [*TASK DEFINITION*]. For that task, the labels include: [*LABELS*]. Write 10 very diverse and concise English instructions. Return the instructions as strings in a list format as follows [].

Table 7: Prompts used to generate instructions through LLMs. *INSTRUCT-LANG* refers to the language, which can be Arabic, English, or Hindi. *LANG* also denotes the language, specifically Arabic, English, or Hindi. *TASK* refers to the task name. For each task, there is a *TASK DEFINITION*. *LABELS* refers to dataset-specific labels.

Model	Instruction	System Role
GPT-4o	Classify the given text as either ‘offensive’ or ‘not-offensive-hateful’. Return only the label without any explanation, justification or additional text.	You are a social media expert providing accurate analysis and insights.
Claude-3.5	Evaluate whether the given text is ‘offensive’ or ‘not-offensive-hateful’, considering vulgar or targeted attacks. Return only the label without any explanation, justification or additional text.	You are a social media expert providing accurate analysis and insights.

Table 8: Examples of instructions generated by two LLMs for the offensive language detection task on the English offensive_language_dataset, along with the pre-defined system role prompt.

Task	Dataset	Metric	SOTA	Base	L-Lens (Eng.)	L-Lens (Native)
Natural Language Understanding (NLU)						
Overall			0.137	0.034	0.129	0.130
News Sum	xlsum	R-2	0.137	0.034	0.129	0.130
Information Extraction & Text Classification						
Overall			0.923	0.660	0.869	0.864
News Cat	ASND	Ma-F1	0.770	0.587	0.919	0.929
News Cat	SANADAkhbarona	Acc	0.940	0.784	0.954	0.953
News Cat	SANADAlArabiya	Acc	0.974	0.893	0.987	0.985
News Cat	SANADAlkhaleej	Acc	0.986	0.865	0.984	0.982
News Cat	UltimateDataset	Ma-F1	0.970	0.376	0.865	0.880
News Credibility	NewsCredibility	Acc	0.899	0.455	0.935	0.933
Subjectivity	ThatiAR	F1_Pos	0.800	0.562	0.441	0.383
Emotion & Sentiment Analysis						
Overall			0.786	0.557	0.724	0.718
Emotion	Emotional-Tone	W-F1	0.658	0.358	0.705	0.736
Emotion	NewsHeadline	Acc	1.000	0.406	0.480	0.458
Sarcasm	ArSarcasm-v2	F1_Pos	0.584	0.477	0.520	0.542
Sentiment	ar_reviews_100k	F1_Pos	–	0.681	0.785	0.779
Sentiment	ArSAS	Acc	0.920	0.603	0.800	0.804
Stance	stance	Ma-F1	0.767	0.608	0.926	0.881
Stance	Mawqif-Arabic-Stance	Ma-F1	0.789	0.764	0.853	0.826
Fast-Checking & Verification						
Overall			0.689	0.407	0.663	0.671
Att.worthiness	CT22Attentionworthy	W-F1	0.412	0.158	0.425	0.454
Checkworthiness	CT24_T1	F1_Pos	0.569	0.610	0.502	0.509
Claim	CT22Claim	Acc	0.703	0.581	0.734	0.756
Factuality	Arafacts	Mi-F1	0.850	0.210	0.771	0.738
Factuality	COVID19Factuality	W-F1	0.831	0.492	0.800	0.840
Propaganda	ArPro	Mi-F1	0.767	0.597	0.747	0.731
Hate Speech & Offensive Content						
Overall			0.797	0.603	0.774	0.767
Cyberbullying	ArCyc_CB	Acc	0.863	0.766	0.870	0.833
Harmfulness	CT22Harmful	F1_Pos	0.557	0.507	0.523	0.535
Hate Speech	annotated-hatetweets-4	W-F1	0.630	0.257	0.526	0.517
Hate Speech	OSACT4SubtaskB	Mi-F1	0.950	0.819	0.955	0.955
Offensive	ArCyc_OFF	Ma-F1	0.878	0.489	0.877	0.879
Offensive	OSACT4SubtaskA	Ma-F1	0.905	0.782	0.896	0.882

Table 9: Arabic Capability-Based Evaluation Results. Base: Llama3.1-Instruct. L-Lens: LlamaLens. Eng.: English. Acc: Accuracy. W-F1: Weighted F1. Mi-F1: Micro-averaged F1. Ma-F1: Macro-averaged F1, R-2: ROUGE-2. Cat:Categorization. News Sum: News Summarization. Att.worthiness: Attentionworthiness.

Task	Dataset	Metric	SOTA	Base	L-Lens (Eng.)	L-Lens (Native)
English Capability-Based Evaluation						
Overall			0.152	0.074	0.182	0.181
News Sum	xlsum	R-2	0.152	0.074	0.182	0.181
Information Extraction & Text Classification						
Overall			0.817	0.709	0.669	0.593
News Cat	CNN_News_Articles	Acc	0.940	0.644	0.970	0.970
News Cat	News_Category	Ma-F1	0.769	0.970	0.824	0.520
News Genre	SemEval23T3-ST1	Mi-F1	0.815	0.687	0.241	0.253
Subjectivity	CT24_T2	Ma-F1	0.744	0.535	0.642	0.628
Emotion & Sentiment Analysis						
Overall			0.835	0.550	0.830	0.834
Emotion	emotion	Ma-F1	0.790	0.353	0.803	0.808
Sarcasm	News-Headlines	Acc	0.897	0.668	0.936	0.947
Sentiment	NewsMTSC	Ma-F1	0.817	0.628	0.751	0.748
Fact-Checking & Verification						
Overall			0.708	0.497	0.811	0.823
Checkworthiness	CT24_T1	F1_Pos	0.753	0.404	0.942	0.942
Claim	claim-detection	Mi-F1	–	0.545	0.864	0.889
Factuality	News_dataset	Acc	0.920	0.654	0.999	0.999
Factuality	Politifact	W-F1	0.490	0.121	0.287	0.311
Propaganda	QProp	Ma-F1	0.667	0.759	0.963	0.973
Hate Speech & Offensive Content						
Overall			0.949	0.529	0.870	0.874
Cyberbullying	Cyberbullying	Acc	0.907	0.175	0.836	0.855
Offensive	Offensive_Hateful	Mi-F1	–	0.692	0.814	0.813
Offensive	offensive_language	Mi-F1	0.994	0.646	0.899	0.893
Offensive & Hate	hate-offensive-speech	Acc	0.945	0.602	0.931	0.935
Hindi Capability-Based Evaluation						
Overall			0.391	0.356	0.369	0.425
NLI	NLI_dataset	W-F1	0.646	0.633	0.568	0.679
News Sum	xlsum	R-2	0.136	0.078	0.171	0.170
Emotion & Sentiment Analysis						
Overall			0.697	0.552	0.647	0.654
Sentiment	Sentiment Analysis	Acc	0.697	0.552	0.647	0.654
Fact-Checking & Verification						
Overall			–	0.759	0.994	0.993
Factuality	fake-news	Mi-F1	–	0.759	0.994	0.993
Hate Speech & Offensive Content						
Overall			0.703	0.518	0.801	0.802
Hate Speech	hate-speech-detection	Mi-F1	0.639	0.750	0.963	0.963
Hate Speech	Hindi-Hostility	W-F1	0.841	0.469	0.753	0.753
Offensive	Offensive Speech	Mi-F1	0.723	0.621	0.862	0.865
Cyberbullying	MC_Hinglish1	Acc	0.609	0.233	0.625	0.627

Table 10: Capability-Based Evaluation Results for English and Hindi. Base: Llama3.1-Instruct. L-Lens: LlamaLens. Eng.: English. Acc: Accuracy. W-F1: Weighted F1. Mi-F1: Micro-averaged F1. Ma-F1: Macro-averaged F1, R-2: ROUGE-2. News Sum: News Summarization. Cat:Categorization. NLI: Natural Language Inference.