UrduFactCheck: An Agentic Fact-Checking Framework for Urdu with Evidence Boosting and Benchmarking

Sarfraz Ahmad* **Hasan Iqbal*** **Momina Ahsan** **Numaan Naeem**
Muhammad Ahsan Riaz Khan **Arham Riaz** **Muhammad Arslan Manzoor**
Yuxia Wang **Preslav Nakov**
MBZUAI
{sarfraz.ahmad, hasan.iqbal, preslav.nakov}@mbzuai.ac.ae

Abstract

The rapid adoption of Large Language Models (LLMs) has raised important concerns about the factual reliability of their outputs, particularly in low-resource languages such as Urdu. Existing automated fact-checking systems are predominantly developed for English, leaving a significant gap for the more than 200 million Urdu speakers worldwide. In this work, we present UrduFactBench and UrduFactQA, two novel hand-annotated benchmarks designed to enable fact-checking and factual consistency evaluation in Urdu. While UrduFactBench focuses on claim verification, UrduFactQA targets the factuality of LLMs in question answering. These resources, the first of their kind for Urdu, were developed through a multi-stage annotation process involving native Urdu speakers. To complement these benchmarks, we introduce UrduFactCheck, a modular fact-checking framework that incorporates both monolingual and translation-based evidence retrieval strategies to mitigate the scarcity of high-quality Urdu evidence. Leveraging these resources, we conduct an extensive evaluation of twelve LLMs and demonstrate that translation-augmented pipelines consistently enhance performance compared to monolingual ones. Our findings reveal persistent challenges for open-source LLMs in Urdu and underscore the importance of developing targeted resources. All code and data are publicly available at https://github.com/mbzuai-nlp/UrduFactCheck.

1 Introduction

In recent years, the way we find and share information has changed dramatically. Large language models (LLMs) like GPT-4o (OpenAI et al., 2023) are now capable of answering questions, generating articles, and even holding conversations that sound convincingly human.

Despite all mentioned strengths, these models sometimes make mistakes and do so with surprising confidence, even when they're wrong. This problem, known as "hallucination" (Bang et al., 2023; Borji, 2023; Tie et al., 2024), is especially troubling when technology is used in important areas such as healthcare, finance, or law (Chuang et al., 2024; Geng et al., 2024; Wang et al., 2024b).

At the same time, social media platforms have become a main source of news and information for millions of people worldwide. These platforms are also a source of fake news and misinformation. During major events such as the 2016 U.S. Presidential Election and the Brexit referendum, false narratives were used to manipulate public opinion at scale (Allcott and Gentzkow, 2017; Pogue, 2017; Vosoughi et al., 2018). The rapid algorithm driven spread of such content, especially on TikTok, Facebook, and Twitter, has reduced public trust in institutions and increased political polarization (Zimmer et al., 2019; Trilling et al., 2017). This trend worsened during the COVID-19 pandemic, which not only increased public awareness of misinformation but also revealed its risks in real time. The World Health Organization (WHO) warned that we were facing not only a pandemic but also an 'infodemic', a surge of false or misleading information about the virus on social media (Humprecht, 2020; Arechar et al., 2023; Organization et al., 2023).

Despite the growing momentum of fact-checking efforts, most initiatives focus on English-language content (Guo et al., 2022), leaving a gap for other widely spoken languages. Urdu is the national language of Pakistan, holds official status in several Indian states, and is spoken by about 232 million people worldwide, yet it accounts for less than 0.5% of all online content.[1]

* Equal contribution.

[1] https://www.icls.edu/blog/most-spoken-languages-in-the-world

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22788–22802
November 4-9, 2025

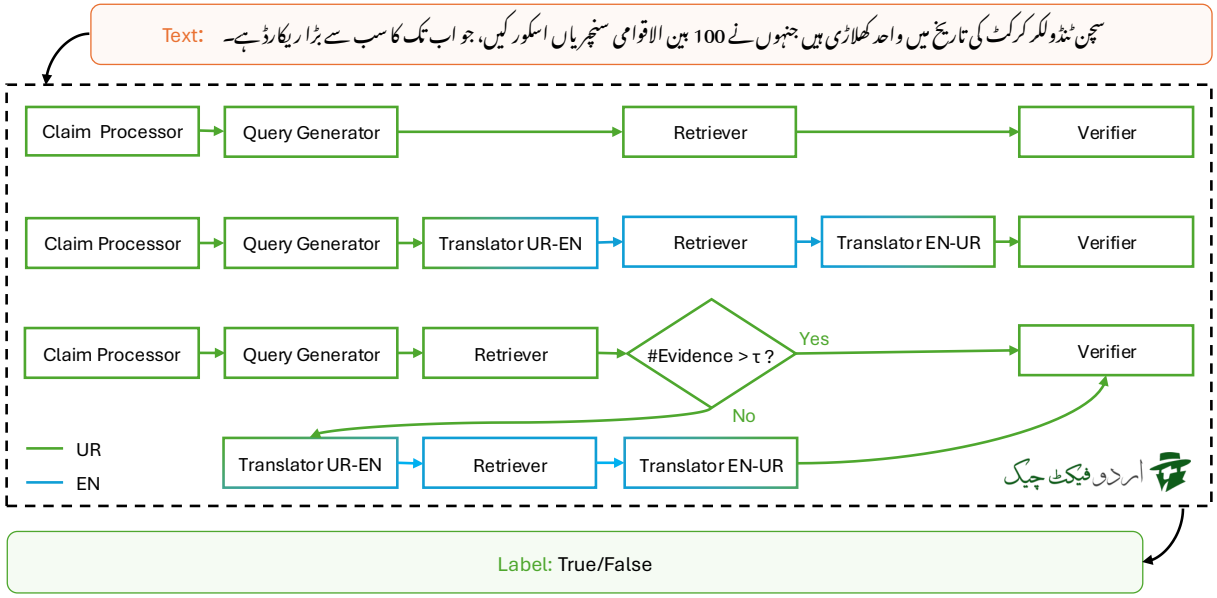

Figure 1: Three core fact-checking pipelines of URDUFACTCHECK: **(top)** end-to-end Urdu framework, **(middle)** translation-augmented retrieval, and **(bottom)** threshold-based rerouting using τ when No. of Evidence $< \tau$. Green indicates Urdu context, while blue indicates English context.

The development of automated fact-checking tools for Urdu remains limited, reflecting a broader trend in Natural Language Processing (NLP), where low-resource languages are often routinely left behind due to a lack of annotated data and low representation in LLM pretraining corpora (Du et al., 2021; Raja et al., 2023). These gaps are concerning given the frequent spread of misinformation in Urdu on social media. False content circulates in various forms, including jokes, memes, and deliberate disinformation (Amjad et al., 2022). While there has been progress in applying cross-lingual methods to tasks such as hate speech and rumor detection in (Glavaš et al., 2020; Haider et al., 2023), these advances have not extended to automated fact-checking in Urdu. To address this, we introduce three resources for factuality assessment in Urdu:

URDUFACTBENCH: A manually curated benchmark for claim verification, enabling the evaluation of fact-checking systems and comparisons to automated fact-checkers. It combines multiple claim datasets into a unified resource and standardizes their labels into a consistent format for evaluation.

URDUFACTQA: A manually annotated dataset for evaluating the factual accuracy of LLMs on Urdu QA tasks. We use it to assess 12 state-of-the-art LLMs, and it serves as the first benchmark to measure the factuality of LLM answers in Urdu.

URDUFACTCHECK: A fact-checking pipeline for claim verification and LLM evaluation in Urdu. It features a modular design with monolingual and translation-based retrieval, and uses a thresholded evidence boosting technique as illustrated in Figure 1.

URDUFACTCHECK builds on recent modular frameworks such as LOKI (Li et al., 2025), OPENFACTCHECK (Wang et al., 2025; Iqbal et al., 2024), and FIRE (Xie et al., 2025), and addresses key challenges in Urdu fact-checking:

- Detecting factual errors in free-form text;
- Enhancing quality and coverage of evidence;
- Evaluating factuality of LLMs on Urdu;
- Analyzing system component contributions;

Our main contributions are two new Urdu datasets for factuality evaluation, supported by an open-source framework for systematic experimentation. These resources provide a foundation for future research on factuality in low-resource languages. In addition, they enable direct comparisons between proprietary and open-source LLMs on Urdu inputs. By making both datasets and framework publicly available, we aim to encourage reproducible research and facilitate extensions to other underrepresented languages.

2 Related Work

Prior efforts in Urdu fact-checking have focused mainly on fake news classification. The URDU-FAKE@FIRE2021 (Amjad et al., 2022) shared task framed this as a binary problem, highlighting generalization issues under domain shifts. AX-TO-GRIND (Harris et al., 2023) introduced a larger annotated dataset and applied multilingual models like MBERT and XLNET. More recently, HOOK AND BAIT URDU (Harris et al., 2025) released the largest Urdu fake news corpus, leveraging LoRA fine-tuning of LLAMA-2 for mono- and multilingual detection. While these systems improved classification and dataset scale, they do not provide end-to-end factuality pipelines.

In parallel, several Urdu QA datasets have emerged. UQA (Arif et al., 2024), a span-preserving translation of SQUAD2.0 (Rajpurkar et al., 2018), has been used to benchmark multilingual models like MBERT and XLM-ROBERTA. Other corpora such as UQUAD1.0 (Kazi and Khoja, 2021) support extractive QA but do not assess factual correctness. These resources target reading comprehension and general QA, rather than fact-checking or systematic evaluation of generated responses.

Multilingual benchmarks such as X-FACT (Gupta and Srikumar, 2021) have tested LLM factuality across several low-resource languages, but Urdu remains underrepresented in these evaluations. At the same time, tools like FACTSCORE (Min et al., 2023), FACTOOL (Chern et al., 2023), and FACTCHECKGPT (Wang et al., 2024a) have advanced metrics, retrieval strategies, and modularity in fact-checking, but they are developed mainly for English and lack Urdu-specific datasets. This absence limits systematic evaluation of factuality in Urdu and restricts the transfer of existing methods to this language.

In contrast, this work introduces URDUFACT-BENCH and URDUFACTQA, the first datasets for fact-checking and factuality evaluation in Urdu. URDUFACTBENCH supports claim verification, while URDUFACTQA evaluates the factual accuracy of LLM-generated responses. To support their use, we also provide URDUFACTCHECK, a modular framework with claim processing, evidence retrieval, and verification. Together, these resources provide the first foundation for factuality evaluation in Urdu, addressing gaps from prior work on classification or general QA.

3 Datasets

To provide a foundation for evaluating automated fact-checkers and measuring the factuality of LLMs, we draw on three claim verification datasets and two QA datasets, curating them into URDU-FACTBENCH and URDUFACTQA, respectively.

3.1 Dataset Collection

Given the limited availability of factual datasets in Urdu, we adapted established English datasets through a multi-stage process under expert supervision. For claim verification, we selected three datasets: BINGCHECK (Li et al., 2024), FACTCHECK-BENCH (Wang et al., 2024a), and FACTOOL (Chern et al., 2023). From FACTOOL, which spans multiple domains, we extracted a subset requiring world knowledge, referred to as FACTOOL-QA.

For factual question answering, we used two QA datasets: SIMPLEQA (Wei et al., 2024) and FRESHQA (Vu et al., 2024). Together, these five datasets cover varied claim structures, domains, and formats, allowing evaluation of both claim-level verification and the factual behavior of LLMs. We emphasized claims that require external knowledge, as this is central to factuality tasks. For FACTCHECK-BENCH and BINGCHECK, we simplified the original four-label scheme (*supported*, *partially supported*, *not supported*, *refuted*) into binary: *True* for *supported/partially supported*, *False* for *refuted*, and removal of *not supported*. This standardization ensured consistency across datasets.

To reduce class imbalance in BINGCHECK, which contained 3,581 *True* and only 42 *False* claims, we sampled 100 *True* instances for the test set. This produced a more balanced dataset and enabled evaluation metrics to better capture performance across both classes.

3.2 Translation and Annotation

To begin the translation process, we used three LLMs: GPT-4o (OpenAI et al., 2023), GEMINI-1.5-PRO (Team et al., 2024), and CLAUDE-SONNET-3.5 (Anthropic, 2024b). Each model was tested by translating 50 claims and 50 QA pairs in a few-shot setup, and the translations were reviewed by expert annotators for fluency, adequacy, and correctness. Based on these qualitative assessments by annotators, GPT-4o was selected as the most suitable model for translation.

This machine-generated translation was not intended as a final product, but rather as a way to accelerate the annotation workflow and reduce the manual workload for human annotators. By leveraging GPT-4o, annotators were able to dedicate their efforts to quality assurance, validation, and refinement rather than translating from scratch.

To improve translation quality, annotators created 100 demonstration examples (20 per dataset) for the LLM's few-shot setup. A custom prompt template, guided by linguistic rules, instructed the model to generate grammatically correct Urdu while preserving technical terms and transliterations. The guidelines also addressed left-to-right numerals in right-to-left flow and correct placement of acronyms (see Appendix A). For optimal few-shot performance, Max Marginal Relevance (MMR) was used to select the most relevant examples. The pipeline was implemented with LangChain[2] and an output parser, using default OpenAI Library parameters for GPT-4o.

After translation, each dataset underwent dual annotation: one expert reviewed the output, and a second independently validated it for linguistic consistency, factual correctness, and cultural appropriateness. A custom annotation portal further streamlined review and verification (see Appendix B). This workflow ensured all datasets met high standards of quality and reliability for factuality evaluation in Urdu. Native Urdu-speaking annotators were employed to ensure the highest linguistic and cultural quality in the datasets. All annotators were required to be senior high-school graduates at minimum, with higher educational qualifications preferred, and both parents being from and residing in Urdu-speaking regions. This careful selection process helped guarantee not only fluency but also deep cultural familiarity with the language. The final translated datasets resulted in the following two resources:

UrduFactBench: Comprising the claim datasets BingCheck, FacTool-QA, and Factcheck-Bench, this benchmark serves as the ground truth for evaluating the performance of automated fact-checkers in Urdu (see Table 1).

UrduFactQA: Consisting of the QA datasets SimpleQA and FreshQA, this benchmark is designed for evaluating the factuality capabilities of LLMs in Urdu (see Table 2).

Dataset	#True	#False	Total
Factcheck-Bench	472	159	631
FacTool-QA	177	56	233
BingCheck	100	42	142
UrduFactBench	749	257	1006

Table 1: Statistics of UrduFactBench.

Dataset	Size
SimpleQA	4,326
FreshQA	600
UrduFactQA	4926

Table 2: Statistics of UrduFactQA.

Together, these benchmarks fill a critical gap in Urdu NLP by providing the first resources for reproducible, benchmarked research on factuality in low-resource settings.

4 Framework

To address the challenge of evaluating factuality in Urdu free-form text, we present UrduFactCheck, a set of three end-to-end pipelines specifically tailored for the Urdu language. The base framework consists of four core agent modules: ClaimProcessor, QueryGenerator, Retriever, and Verifier, drawing on well-established automated fact-checking frameworks, as illustrated in Figure 1 (Li et al., 2025; Iqbal et al., 2024; Xie et al., 2025; Chern et al., 2023).

4.1 Prompt Engineering for Core Modules

The UrduFactCheck framework adopts an agentic architecture, where each module functions as a specialized agent with a distinct role. Prompts are designed to handle Urdu's linguistic and contextual challenges, ensuring coherent outputs across the pipeline.

The ClaimProcessor (CP) decomposes text into atomic, check-worthy claims. The QueryGenerator (QG) produces two query types per claim: (i) question-based queries that conceal the fact and (ii) direct claim-based queries to improve retrieval. The Retriever (RTV) uses the Google SERP API[3] without prompt engineering, while the Verifier (VFR) assigns factuality labels, provides reasoning, and suggests corrections. Each prompt includes two–three examples, with full templates in Appendix C.

[2] https://www.langchain.com

[3] https://serper.dev

4.2 Evidence Boosting

A major challenge in automated fact-checking for Urdu is the limited availability of high-quality evidence in the language. To address this, URDUFACTCHECK implements a multi-strategy evidence retrieval approach, consisting of three distinct strategies, that dynamically adapts to the difficulty and resource needs of each claim.

Monolingual Retrieval: This is a straightforward approach in which, for every Urdu query q_{ur}, the system retrieves evidence E_{ur} in Urdu. This method ensures language consistency and computational efficiency, but struggles to provide relevant results for niche or globally underrepresented topics due to the scarcity of reliable Urdu web content. As a result, the evidence retrieved can sometimes be insufficient or only loosely related to the original claim.

Translated Retrieval: This strategy seeks to overcome the limitations of monolingual retrieval by translating the Urdu query q_{ur} into English q_{en} and conducting the web search in English to obtain evidence E_{en}. The retrieved evidence is then translated back into Urdu, resulting in $E_{\text{en-ur}}$, to maintain consistency with downstream modules. While this translation-based approach significantly improves evidence recall and quality by leveraging abundant English online sources, it incurs higher computational overhead and introduces potential risks of semantic drift during back-translation.

Thresholded Translated Retrieval: This approach combines the efficiency of monolingual retrieval with the robustness of translation-based search using a dynamic fallback mechanism. We introduce a thresholded evidence retrieval function, $\mathcal{R}(q_{\text{ur}}, \tau)$, which first attempts direct Urdu retrieval. The sufficiency of evidence E_{ur} is assessed by comparing its cardinality $|E_{\text{ur}}|$ to a predefined threshold τ which represents the minimum evidence count.

If $|E_{\text{ur}}| \geq \tau$, the system proceeds with E_{ur} for factual verification. Otherwise, q_{ur} is translated into English (q_{en}) and additional evidence E_{en} is retrieved using English search. This evidence is then translated back into Urdu $E_{\text{en-ur}}$. In such cases, both E_{ur} and $E_{\text{en-ur}}$ are combined for downstream verification.

$$\mathcal{R}(q, \tau) = \begin{cases} E_{\text{ur}}, & \text{if } |E_{\text{ur}}| \geq \tau \\ E_{\text{ur}} \cup E_{\text{en-ur}}, & \text{otherwise} \end{cases} \quad (1)$$

As shown in Equation 1, this adaptive approach allows the system to default to efficient monolingual retrieval while guaranteeing broader verification coverage when Urdu evidence is insufficient. In such cases, the system dynamically invokes translation-based retrieval and combines evidence from both languages to strengthen verification. This design ensures that retrieval quality does not degrade even in scenarios where Urdu web content is sparse.

To support these transitions, we engineered dedicated prompts for Urdu-to-English and English-to-Urdu translation. All translation is performed by an LLM agent, which preserves meaning and maintains consistency across the pipeline. This setup not only improves recall but also ensures that retrieved evidence is usable in downstream modules. Overall, the tiered retrieval framework enables URDUFACTCHECK to balance accuracy and cost, while directly addressing the central challenge of evidence scarcity in Urdu fact-checking.

5 Experiments

To evaluate URDUFACTCHECK and our benchmarks, we conducted three experiments: (i) analyzing the effect of evidence thresholding on retrieval and verification, (ii) benchmarking automated fact-checkers with URDUFACTBENCH, and (iii) assessing LLM factuality with URDUFACTQA. We also report API costs for proprietary LLMs, GPU rental for open-source models, search engine query expenses, and total fact-checking time. Open-source experiments ran on an NVIDIA RTX 6000 GPU ($0.79/hour), with each SerpAPI query costing about $0.00105.

5.1 Threshold Tuning

A key hyperparameter in the retrieval pipeline is the evidence threshold τ, which specifies the minimum number of Urdu snippets required before falling back to translation-based retrieval. To study trade-offs between recall, accuracy, and efficiency, we vary $\tau \in \{1, 3, 5, 7, 9\}$ and evaluate performance on the FACTCHECK-BENCH subset of URDUFACTBENCH, recording verification accuracy and retrieval cost. This identifies the optimal threshold balancing recall and cost. All experiments use GPT-4O-MINI as the backbone model, with temperature 0 and a 2500-token limit; other parameters remain default.

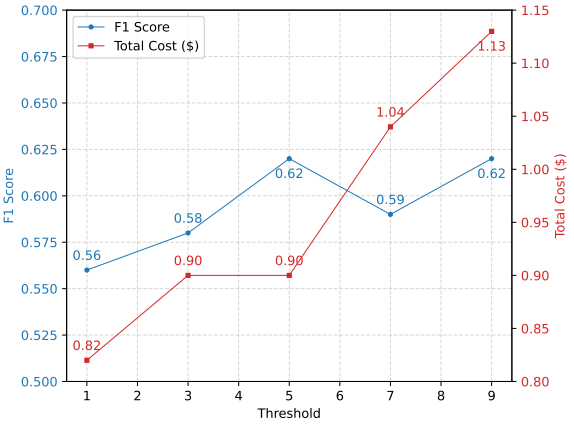

Figure 2: Effect of evidence threshold τ on fact-checking performance and cost for the FACTCHECK-BENCH subset of URDUFACTBENCH. The blue line (left axis) shows the F1 score, while the red line (right axis) shows the total cost ($). Higher thresholds increase cost but can improve F1 up to an optimal range before plateauing.

Results Analysis Figure 2 shows the effect of varying the evidence threshold τ on both F1 score and total retrieval cost. As τ increases from 1 to 5, F1 score improves, but a dip is observed at $\tau = 7$ before rising again at $\tau = 9$. This pattern is not noise, but rather reflects that higher thresholds may introduce more loosely related or noisy snippets, which can hinder verification accuracy despite increasing the volume of evidence. The increase in cost is expected, as larger τ values trigger more frequent fallback to translation-based retrieval.

Overall, setting τ in the range of 3–5 appears to provide a favorable balance between improved accuracy and manageable computational cost, making it a practical choice for deployment in resource-constrained or cost-sensitive scenarios. Based on these results, we use $\tau = 5$ as the default threshold for subsequent experiments, unless otherwise specified.

5.2 LLM Selection for Fact-Checking

Before conducting a full evaluation of URDUFACTCHECK, we first examined the performance of a set of LLMs on the FACTCHECK-BENCH subset of URDUFACTBENCH. The aim of this step was to identify which models provide the best balance between accuracy and cost, so that they could be used as backbone verifiers in later benchmarking experiments. We chose FACTCHECK-BENCH for this initial study because it provides a diverse set of claims and has significant number of samples, making it a suitable testbed for comparing models under controlled conditions.

The models included in this evaluation fall into two groups. The proprietary group consisted of the GPT series (OpenAI et al., 2023, 2024) and the Claude series (Anthropic, 2024a), while the open-source group included MISTRAL-INST 7B (Jiang et al., 2023) and LLAMA3.1-INST 8B (Grattafiori et al., 2024). This selection allowed us to compare high-resource proprietary models with widely used open-source alternatives that are more accessible but often trained with smaller scale resources.

Results Analysis: The results in Table 3 show that proprietary models generally outperform open-source ones on precision and recall for true and false labels. GPT-4.1 achieved the strongest overall performance, confirming its capacity for accurate factual verification in Urdu. However, smaller proprietary variants such as GPT-4.1-MINI and GPT-4O-MINI also performed competitively, reaching similar F1 scores at a fraction of the computational cost. In contrast, the open-source models MISTRAL-INST 7B and LLAMA3.1-INST 8B showed weaker performance, particularly on false labels, which suggests limitations in their ability to reliably detect incorrect claims in Urdu.

A key observation from this experiment is the cost-performance trade-off. While large proprietary models provide the highest accuracy, smaller variants deliver comparable results at lower cost, making them practical for large-scale evaluation. Based on these findings, we selected GPT-4O and GPT-4O-MINI as backbone models for subsequent experiments. These represent a complementary pair: one offering accuracy and the other efficiency. This ensured that benchmarking of URDUFACTCHECK was grounded in a balance of performance and efficiency for Urdu factual verification, while also providing a basis for future comparisons with other models.

5.3 Fact-Checker Benchmarking

To evaluate URDUFACTCHECK, we conducted experiments on two additional URDUFACTBENCH subsets: FACTOOL-QA and BINGCHECK. As no end-to-end fact-checking systems exist for Urdu, direct comparisons are limited. We include FACTOOL, but it produced unreliable outputs on Urdu text. For this reason, we did not extend comparisons to English-based fact-checkers, as their results would not provide a fair basis for Urdu verification, underscoring the need for Urdu-specific evaluation.

LLM	LLM + Search Cost ($)	Label = True			Label = False		
		Prec	Recall	F1	Prec	Recall	F1
GPT-4.1	6.06+2.35	0.92	0.56	0.70	0.39	0.85	0.54
GPT-4.1-MINI	1.10+2.06	0.88	0.61	0.72	0.40	0.75	0.52
GPT-4O	7.42+2.32	0.90	0.56	0.69	0.38	0.80	0.52
GPT-4O-MINI	0.35+1.87	0.92	0.48	0.63	0.36	0.87	0.51
CLAUDE-SONNET	21.6+2.66	0.90	0.44	0.59	0.34	0.85	0.49
CLAUDE-HAIKU	5.71+2.73	0.85	0.40	0.54	0.30	0.79	0.44
MISTRAL-INST 7B	1.84+1.22	0.80	0.39	0.52	0.30	0.62	0.40
LLAMA3.1-INST 8B	4.02+2.15	0.84	0.43	0.57	0.32	0.65	0.42

Table 3: Fact-checking performance and cost of different language models on the FACTCHECK-BENCH subset of URDUFACTBENCH, using Thresholded Retrieval ($\tau = 5$). Results are reported separately for *Label = True* and *Label = False* in terms of precision, recall, and F1, alongside the combined LLM and search cost. Each cell is color-coded from red (lowest) to green (highest) within its column to highlight relative performance.

Framework	LLM	LLM + Search Cost ($)	URDUFACTBENCH- FACTOOL-QA						URDUFACTBENCH- BINGCHECK						Urdu Language		
			Label = True			Label = False			Label = True			Label = False					
			Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1	CP	RTV	VFR
Random	-	-	0.58	0.77	0.66	0.46	0.26	0.33	0.58	0.77	0.66	0.46	0.26	0.33	-	-	-
Always True	-	-	1.00	0.76	0.86	0.00	0.00	0.00	1.00	0.76	0.86	0.00	0.00	0.00	-	-	-
Always False	-	-	0.00	0.00	0.00	1.00	0.24	0.39	0.00	0.00	0.00	1.00	0.24	0.39	-	-	-
FACTOOL	GPT-4O	4.67+1.57	0.75	0.50	0.60	0.43	0.59	0.50	0.82	0.47	0.60	0.38	0.76	0.50	×	×	×
	GPT-4O-MINI	0.21+1.22	0.72	0.48	0.56	0.41	0.61	0.49	0.84	0.46	0.59	0.39	0.79	0.52	×	×	×
URDUFACTCHECK	GPT-4O	4.87+1.61	0.84	0.63	0.72	0.35	0.63	0.45	0.87	0.41	0.56	0.39	0.86	0.54	✓	✓	✓
	GPT-4O-MINI	0.22+1.24	0.87	0.53	0.65	0.33	0.75	0.46	0.87	0.45	0.59	0.34	0.84	0.48	✓	✓	✓
URDUFACTCHECK TH-TR-3	GPT-4O	5.02+1.72	0.84	0.62	0.71	0.35	0.64	0.45	0.88	0.41	0.56	0.40	0.85	0.54	✓	✓	✓
	GPT-4O-MINI	0.24+1.37	0.83	0.48	0.61	0.29	0.68	0.41	0.87	0.47	0.61	0.40	0.84	0.55	✓	✓	✓
URDUFACTCHECK TH-TR-5	GPT-4O	5.45+2.19	0.83	0.65	0.73	0.34	0.57	0.43	0.83	0.41	0.55	0.38	0.81	0.52	✓	✓	✓
	GPT-4O-MINI	0.24+1.37	0.87	0.50	0.64	0.33	0.77	0.46	0.93	0.50	0.65	0.44	0.91	0.59	✓	✓	✓
URDUFACTCHECK TH-TR-7	GPT-4O	5.20+2.38	0.84	0.67	0.75	0.35	0.59	0.44	0.80	0.40	0.53	0.35	0.77	0.49	✓	✓	✓
	GPT-4O-MINI	0.28+1.59	0.87	0.53	0.66	0.34	0.79	0.48	0.89	0.48	0.62	0.42	0.86	0.56	✓	✓	✓
URDUFACTCHECK TH-TR-9	GPT-4O	6.12+2.67	0.87	0.53	0.66	0.34	0.79	0.48	0.80	0.41	0.54	0.36	0.77	0.49	✓	✓	✓
	GPT-4O-MINI	0.30+1.66	0.85	0.53	0.66	0.33	0.71	0.45	0.90	0.53	0.67	0.44	0.86	0.58	✓	✓	✓
URDUFACTCHECK TR	GPT-4O	8.87+2.23	0.90	0.70	0.79	0.44	0.75	0.56	0.79	0.55	0.65	0.39	0.67	0.50	✓	✓	✓
	GPT-4O-MINI	0.46+1.38	0.88	0.58	0.70	0.37	0.78	0.50	0.92	0.55	0.69	0.45	0.88	0.60	✓	✓	✓

Table 4: Comparison of fact-checking performance across frameworks on URDUFACTBENCH subsets (FACTOOL-QA and BINGCHECK). Each cell is color-coded from red (lowest) to green (highest) within its column to highlight relative performance. Metrics are reported separately for *Label = True* and *Label = False*, along with precision (Prec), recall, and F1. Cost values denote the combined expense of the LLM and search.

Our evaluation included three variants of URDUFACTCHECK: (i) a monolingual retrieval pipeline, (ii) a fully translated retrieval pipeline (TR), and (iii) a thresholded translated retrieval pipeline (TH-TR). For the TH-TR variant, we varied the evidence threshold parameter τ across values 3, 5, 7, and 9, resulting in six distinct configurations of the system. All experiments were performed using two backbone language models: GPT-4o and GPT-4o-MINI. This setup allowed us to study the trade-off between retrieval quality, accuracy, and cost under different system configurations.

Results Analysis: The results are reported in Table 4. Across both subsets, all URDUFACTCHECK variants outperform FACTOOL and trivial baselines. Among the three approaches, translation-based pipelines (TR and TH-TR) give the strongest results. For example, the TR variant with GPT-4o yields F1 scores up to 0.79 on true labels, showing the benefit of using English evidence to supplement limited Urdu web content. The TH-TR variants also achieve high accuracy while controlling cost, with $\tau = 5$ providing a favorable balance between retrieval quality and efficiency.

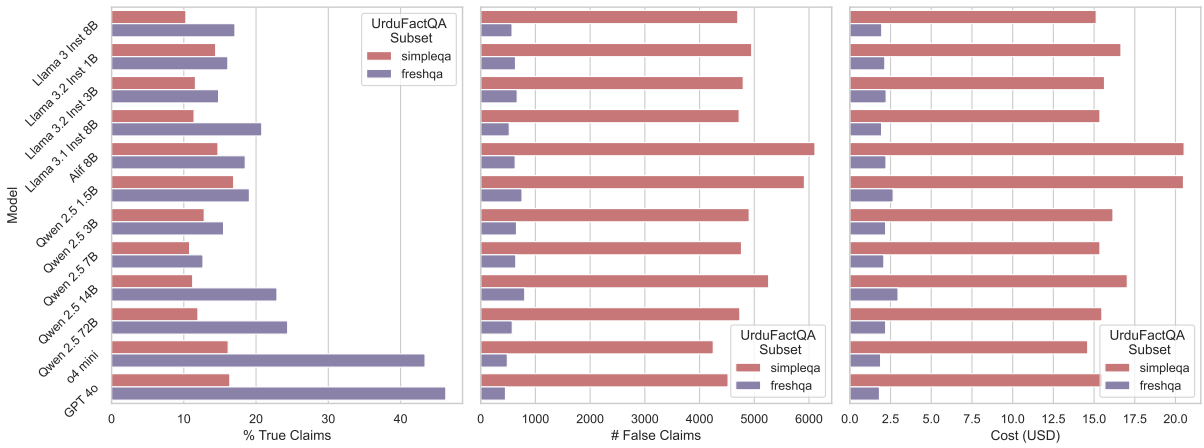

Figure 3: Automatic factuality evaluation results for 12 SOTA LLMS on URDUFACTQA using URDUFACTCHECK-TR. *left:* the percentage of true claims, *center:* the number of false claims, and *right:* the cost of using URDUFACTCHECK-TR in USD.

In addition to accuracy, cost is an important factor in benchmarking fact-checking pipelines. Our analysis shows that GPT-4O-MINI achieves performance close to GPT-4O but at a much lower cost, making it a practical choice for large-scale or resource-constrained applications. These findings demonstrate that while URDUFACTCHECK can be configured for higher accuracy through translation-based retrieval, lightweight variants offer a cost-effective alternative without a major loss in performance. This flexibility allows users to select configurations that best match their accuracy requirements and available resources.

5.4 Evaluating LLM Factuality

We evaluated the factual accuracy of twelve LLMs on the URDUFACTQA benchmark. The evaluation included both proprietary and open-source models. The proprietary group consisted of GPT-4O and O4-MINI. The open-source group included ALIF (8B), a Urdu specific model, as well as the LLAMA-3-INST (8B), LLAMA-3.1-INST (8B), LLAMA-3.2-INST (1B, 3B), and QWEN 2.5 (1.5B, 3B, 7B, 14B, 72B) families.

For each question in URDUFACTQA, responses were generated from all twelve models. Proprietary models used default OpenAI API parameters, while open-source models followed Hugging Face defaults. All outputs were automatically evaluated for factuality using the translation-based URDUFACTCHECK pipeline (TR variant) with GPT-4O-MINI as verifier and Google Serper for retrieval. Prompts and templates from URDUFACTQA were applied consistently.

Results Analysis: The results are shown in Figure 3. Proprietary models such as GPT-4O and O4-MINI produced the highest percentage of factually correct responses across both subsets of URDUFACTQA. In these models, the percentage of true claims reached up to 46%. Open-source models, including ALIF-8B, the LLAMA-3 series, and the QWEN series, showed lower factual accuracy, often below 25% true claims. Within the open-source group, QWEN 2.5 (72B) was the strongest model, though still behind proprietary systems.

The two subsets of URDUFACTQA also showed different levels of difficulty. The SIMPLEQA subset contained a larger number of questions (4,326) and produced lower percentages of true claims, reflecting its greater challenge for LLMs. The FRESHQA subset, with fewer questions (600), resulted in a higher proportion of correct answers and fewer false claims overall. This difference highlights the importance of dataset size and question design in assessing model factuality. Computational costs were comparable across models, with differences mainly due to model size rather than evaluation procedure.

In summary, the evaluation shows that proprietary models currently provide better factual accuracy for Urdu question answering, while open-source models have more limited performance. These findings confirm the value of large instruction-tuned models for factuality tasks in low-resource languages, and they emphasize the need for further development of competitive open-source models for Urdu.

Major Issue	Error Type Description	FACTCHECK-BENCH			FACTOOL			BINGCHECK			#Total
		I	II	III	I	II	III	I	II	III	
A. Dataset Issue	**1. Invalid claim:** When the input consists only of a single entity such as a person's name or a place name or the claim is simply not check-worthy. For Example: نیو یارک or وہ چائے پی رہا تھا۔	0	0	0	0	0	0	0	0	0	0
	2. Unclear, ambiguous, or subjective claim: When the statement is vague, open to multiple interpretations, or not objectively verifiable. For Example: انہوں نے میدان مار لیا۔	0	1	0	0	2	2	1	2	1	9
	3. False gold labels: Cases where the original annotated label is incorrect. For Example the following statement is erroneously gold labeled as true: سبز چائے پینے سے ذیابیطس مکمل طور پر ختم ہو جاتی ہے۔	0	0	0	0	0	0	0	0	0	0
B. Knowledge Issue	**4. Requires complex scientific or domain-expert knowledge:** Cases where specialized expertise is necessary to verify the claim. For Example: دعویٰ کہ بلیک ہول کے قریب وقت کی رفتار کم ہو جاتی ہے۔	2	4	6	1	2	3	0	0	2	20
	5. Inaccurate parametric knowledge: Cases where LLM-based verifiers make incorrect judgments because of erroneous or outdated knowledge stored in their parameters. For Example: دعویٰ کہ پاکستان کا دارالحکومت کراچی ہے۔	4	8	5	4	2	5	0	0	3	31
	6. Insufficient or inaccurate externally collected knowledge (evidence): Cases where verification fails due to (i) no external evidence retrieved and the model relies only on its own reasoning; (ii) evidence collected is incomplete and does not cover all aspects of a complex claim; (iii) evidence itself is inaccurate or contains errors. For Example: دعویٰ کہ قراقرم میں دنیا کی پانچ بلند ترین چوٹیاں ہیں۔	11	5	1	6	6	2	12	9	7	59
C. LLM Reasoning	**7. Incorrect reasoning:** Cases where the model draws invalid conclusions, either by dismissing relevant information from the claim or by introducing unsupported details.	3	2	6	9	8	6	4	5	6	49
	8. Strict reasoning: Cases where the model relies too rigidly on a single source of knowledge: (i) strictly depending on collected evidence without using commonsense leads to wrong verification, whereas combining both could yield the correct result; (ii) overly strict reasoning based on parametric knowledge. For Example: دعویٰ کہ فلو اور انفلوئنزا دو مختلف بیماریاں ہیں۔).	0	0	0	0	0	2	2	3	1	8
D. Debatable Opinion	**9. Debatable opinions:** Cases involving controversial or disputed topics where multiple perspectives exist and no single view is universally accepted. For Example: (دعویٰ کہ موہنجوداڑو کی تہذیب آریاؤں نے قائم کی تھی	0	0	2	0	0	0	1	1	0	4
Total		20	20	20	20	20	20	20	20	20	180

Table 5: **Datasets error distribution, grouped into nine fine-grained types under four major issues.**. "I" represents the base URDUFACTCHECK pipeline, "II" refers to the version with translation-augmented retrieval, and "III" denotes the version with threshold-based rerouting.

5.5 Error Analysis

Table 5 shows the distribution of nine error types, based on 20 randomly sampled errors from each evaluation. Most errors stem from **Knowledge Issues**, particularly insufficient or inaccurate external evidence (59 cases) and inaccurate parametric knowledge (31 cases). **LLM Reasoning** errors (49 incorrect reasoning, 8 strict reasoning) also reveal that models often dismiss relevant information, hallucinate unsupported details, or apply overly rigid reasoning strategies.

Importantly, URDUFACTCHECK's translation-augmented and thresholded translation-augmented retrieval significantly reduce errors caused by low-quality evidence, confirming the effectiveness of multilingual strategies for Urdu fact-checking. While **Dataset Issues** and **Debatable Opinions** appear less often, their inclusion ensures coverage of ambiguity and contested claims.

These findings demonstrate the complementary value of our two benchmarks: URDUFACTBENCH enables fine-grained study of verification errors, while URDUFACTQA highlights how factuality challenges surface in generative QA. Together, they provide a comprehensive basis for advancing factuality research in Urdu and related low-resource languages.

6 Conclusion

We introduced two new benchmarks for factuality evaluation in Urdu: URDUFACTBENCH for claim verification and URDUFACTQA for factual question answering. Built using translation and dual annotation, these are the first publicly available datasets enabling systematic study of factuality in Urdu, and filling a major resource gap for low-resource languages.

In order to demonstrate the utility of these new datasets, we further developed URDUFACTCHECK, a modular framework that integrates claim processing, evidence retrieval, and verification. It serves as a testbed for benchmarking fact-checking strategies and LLMs in Urdu. Using these resources, we conducted extensive experiments, including evaluations of twelve LLMs, revealing clear differences between proprietary and open-source models and underscoring both challenges and opportunities in Urdu factuality.

Our main contribution is the release of two high-quality benchmarks, supported by an open framework for experimentation. We hope that these resources will provide a foundation for advancing factuality assessment in low-resource languages, will foster cross-lingual transfer of methods, and will offer practical tools to address misinformation in Urdu and beyond.

Limitations and Future Work

While URDUFACTCHECK represents a significant step forward in factuality evaluation for Urdu, several limitations remain:

Evaluation Datasets The effectiveness of URDUFACTCHECK relies heavily on the quality and diversity of the evaluation datasets. Although we have incorporated multiple benchmarks to ensure broad domain coverage, inherent biases and coverage gaps persist. Certain specialized domains may be underrepresented, potentially limiting the system's robustness and generalizability for all types of factual claims.

Latency and Cost Automatic fact-checking with URDUFACTCHECK can incur substantial computational costs and latency, particularly when leveraging high-accuracy models and multi-stage retrieval strategies. These resource requirements may pose challenges for real-time applications or users with budgetary constraints.

Quality of Machine Translation The framework relies on machine translation when retrieving and processing evidence across Urdu and English. Despite careful prompt engineering and post-editing, translation errors can introduce semantic drift, loss of nuance, or context misinterpretation, potentially affecting both evidence quality and factuality judgments.

Temporal Limitations Currently, URDUFACTCHECK does not explicitly model the temporal dynamics of factuality. As facts may change over time, especially in rapidly evolving domains, this can lead to mismatches between system judgments and the present state of knowledge. We are actively working on methods to integrate temporal awareness into future versions of the framework.

Dependence on External Knowledge Sources The framework's reliance on external knowledge bases and web search engines introduces variability in the availability, reliability, and timeliness of evidence. Since web content is dynamic and not always up to date, the factual accuracy of retrieved information cannot be guaranteed in all scenarios.

Limited Human Evaluation While we perform automated evaluation of LLM outputs using URDUFACTCHECK, comprehensive human annotation and double-checking across all benchmarks is limited by human annotator availability and budget constraints. Automated metrics may not always fully capture nuanced or context-dependent factual errors that human experts could identify.

Handling Ambiguity and Subjectivity Some claims and questions may be inherently ambiguous, subjective, or context-dependent. The current framework is not equipped to distinguish between subjective assertions, nuanced opinions, or multi-faceted claims, which may impact the accuracy of factuality judgments in such cases.

7 Ethical Statement

The development and deployment of URDUFACTCHECK are guided by ethical principles to ensure responsible use and positive societal impact:

Transparency and Accountability We prioritize transparency by making our code, data, and evaluation protocols publicly available. This enables independent scrutiny and fosters community trust. We invite users and researchers to report issues and biases, promoting continual improvement of the framework.

Bias Mitigation We acknowledge the existence of potential biases in both language models and evaluation datasets. By integrating diverse benchmarks and supporting research into fair fact-checking, we aim to minimize the influence of bias on factuality assessments.

Social Impact Improving the factual accuracy of LLM outputs is central to combating misinformation and supporting informed public discourse. We believe URDUFACTCHECK can contribute meaningfully to these goals, especially in low-resource linguistic communities.

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Maaz Amjad, Sabur Butt, Hamza Imam Amjad, Alisa Zhila, Grigori Sidorov, and Alexander Gelbukh. 2022. Overview of the shared task on fake news detection in Urdu at FIRE 2021. *ArXiv preprint*, abs/2207.05133.

Anthropic. 2024a. The Claude 3 model family: Opus, Sonnet, Haiku. *Anthropic*.

Anthropic. 2024b. Claude 3.5 sonnet model card addendum. *Anthropic*.

Antonio A Arechar, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro, and 1 others. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, 7(9):1502–1513.

Samee Arif, Sualeha Farid, Awais Athar, and Agha Ali Raza. 2024. UQA: Corpus for Urdu question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, LREC-COLING '24, pages 17237–17244, Torino, Italia. ELRA and ICCL.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, IJCNLP-AACL '23, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Ali Borji. 2023. A categorical archive of ChatGPT failures. *ArXiv preprint*, abs/2302.03494.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. FacTool: Factuality detection in generative AI–A tool augmented framework for multi-task and multi-domain scenarios. *ArXiv preprint*, abs/2307.13528.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. DoLa: Decoding by contrasting layers improves factuality in large language models. In *Proceedings of the Twelfth International Conference on Learning Representations*, ICLR '24, Vienna, Austria. OpenReview.net.

Jiangshu Du, Yingtong Dou, Congying Xia, Limeng Cui, Jing Ma, and Philip S Yu. 2021. Cross-lingual COVID-19 fake news detection. In *Proceedings of the 2021 International Conference on Data Mining Workshops*, ICDMW '21, pages 859–862. IEEE.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL '24, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, COLING '20, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-Fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, IJCNLP '21, pages 675–682, Online. Association for Computational Linguistics.

Samar Haider, Luca Luceri, Ashok Deb, Adam Badawy, Nanyun Peng, and Emilio Ferrara. 2023. Detecting social media manipulation in low-resource languages. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23, pages 1358–1364, Austin, TX, USA.

Sheetal Harris, Jinshuo Liu, Hassan Jalil Hadi, Naveed Ahmad, and Mohammed Ali Alshara. 2025. Benchmarking Hook and Bait Urdu news dataset for domain-agnostic and multilingual fake news detection using large language models. *Scientific Reports*, 15(1):15553.

Sheetal Harris, Jinshuo Liu, Hassan Jalil Hadi, and Yue Cao. 2023. Ax-to-Grind Urdu: Benchmark dataset for Urdu fake news detection. In *Proceedings of the 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications*, TrustCom '23, pages 2440–2447. IEEE.

Edda Humprecht. 2020. How do they debunk "fake news"? A cross-national comparison of transparency in fact checks. *Digital journalism*, 8(3):310–327.

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Nenkov Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. OpenFactCheck: A unified framework for factuality evaluation of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, EMNLP '24, pages 219–229, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel,

Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Samreen Kazi and Shakeel Khoja. 2021. Uquad1.0: Development of an Urdu question answering training data for machine reading comprehension. *ArXiv preprint*, abs/2111.01543.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2025. Loki: An open-source tool for fact verification. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, COLING '25, pages 28–36, Abu Dhabi, UAE. Association for Computational Linguistics.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024. Self-Checker: Plug-and-Play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics*, NAACL '24, pages 163–181, Mexico City, Mexico. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP '23, pages 12076–12100, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. GPT-4 technical report. *ArXiv preprint*, abs/2303.08774.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024. OpenAI o1 system card. *ArXiv preprint*, abs/2412.16720.

World Health Organization and 1 others. 2023. An overview of infodemic management during COVID-19 pandemic, january 2020–july 2022. *An overview of infodemic management during COVID-19 pandemic, January 2020–July 2022*.

David Pogue. 2017. How to stamp out fake news. *Scientific American*, 316(2):24–24.

Eduri Raja, Badal Soni, and Samir Kumar Borgohain. 2023. Fake news detection in Dravidian languages using transfer learning with adaptive finetuning. *Engineering Applications of Artificial Intelligence*, 126:106877.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '18, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, abs/2403.05530.

Jiessie Tie, Bingsheng Yao, Tianshi Li, Syed Ishtiaque Ahmed, Dakuo Wang, and Shurui Zhou. 2024. LLMs are imperfect, then what? An empirical study on LLM failures in software engineering. *ArXiv preprint*, abs/2411.09916.

Damian Trilling, Petro Tolochko, and Björn Burscher. 2017. From newsworthiness to shareworthiness: How to predict news sharing based on article characteristics. *Journalism & mass communication quarterly*, 94(1):38–60.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024a. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics*, EMNLP '24, pages 14199–14230, Miami, Florida, USA. 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 '25, pages 11399–11421, Abu Dhabi, UAE. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024b. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, EMNLP '24, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *ArXiv preprint*, abs/2411.04368.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. FIRE: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics*, NAACL '25, pages 2901–2914, Albuquerque, New Mexico. Association for Computational Linguistics.

Franziska Zimmer, Katrin Scheibe, Mechtild Stock, and Wolfgang G Stock. 2019. Fake news in social media: Bad algorithms or biased users? *Journal of Information Science Theory and Practice*, 7(2):40–53.

A Pre-Translation Prompt for Dataset Generation

To speed up annotation for URDUFACTQA and URDUFACTBENCH, we designed a structured pre-translation prompt. It instructed the LLM to translate claim–label pairs into formal Urdu. The full prompt appears in Listing 1.

```
You are an expert Urdu translator. Your task
is to translate the following claim-label
pairs from English to Urdu.

### Instructions
- Translate both the claim and label into
formal, fluent Urdu.
- Use correct masculine/feminine grammatical
forms in Urdu.
- Translate proper nouns only if a widely
accepted Urdu version exists (e.g., "India" →
بھارت, "Syria" → شام).
- Avoid translating proper nouns when they
appear in the name of an organization.
- Retain technical or factual terms (e.g.,
award names, organization names) in
transliterated form, where appropriate.
- Translate dates into proper Urdu format
(e.g., "Jan 1, 2020" → "2020 یکم جنوری").

### Important Formatting Guidelines
1. English acronyms and abbreviations (e.g.,
IEEE, NASA, UNESCO):
- Do not translate or transliterate.
- Place them at a natural position in the Urdu
sentence (ideally after the date or subject).
- Avoid starting Urdu sentences with acronyms
or left-to-right (LTR) text.

2. Western numerals and LTR elements (e.g.,
2022, 7.8.8, Notepad++):
- Do not convert numerals to Urdu words.
- Always place an Urdu phrase before such
elements to maintain proper right-to-left
(RTL) sentence flow.
- This applies to acronyms, version numbers,
software/product names, etc.

Incorrect (structurally broken):
a. فرینک روزن بلیٹ ایوارڈ کس کو دیا گیا؟ IEEE سال 2010
میں
b. 2 of January of 2019
c. رگبی یورپ چیمپئن شپ کا حصہ بننے والے اسپین اور رومانیہ
کے درمیان رگبی میچ 27 فروری 2022 کو اسپین کے لیے تمام
کنورژنز کس کھلاڑی نے اسکور کیے؟ 2022
Correct (natural Urdu structure):
a. IEEE فرینک روزن بلیٹ ایوارڈ سال2010 میں کس کو دیا
گیا؟
b. سال 2019 میں 2 جنوری کو
c. رگبی یورپ چیمپئن شپ 2022 کا حصہ بننے والے اسپین اور
رومانیہ کے درمیان رگبی میچ میں 27 فروری 2022 کو اسپین
اسپین کے لیے تمام کنورژنز کس کھلاڑی نے اسکور کیے؟

3. Ensure the final Urdu sentence is:
- Grammatically correct
- Visually aligned for RTL display
- Fluent and natural to read

### Examples
Here are a few examples of claims and expected
translations:
{examples}

### Translation
claim: {claim}
label: {label}

### Format Instructions:
{format_instructions}
```

Listing 1: Pre-translation prompt used to guide LLM in generating Urdu translations of claims and labels. The prompt defines translation rules for proper nouns, acronyms, numerals, and dates, ensuring consistent and fluent output across the dataset.

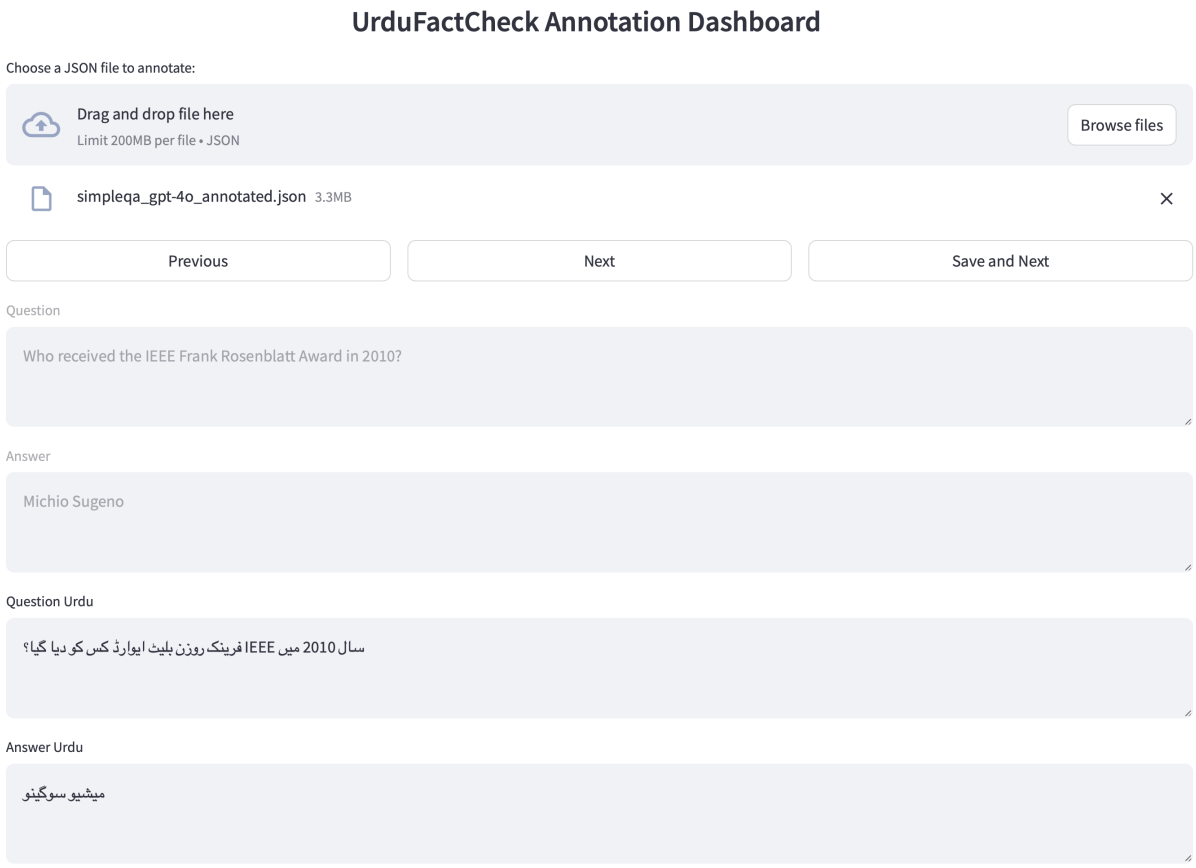

Figure 4: URDUFACTCHECK annotator dashboard built in Streamlit.

B URDUFACTCHECK Annotator Dashboard

As shown in Figure 4, we developed a dedicated annotator dashboard to streamline dataset creation and quality assurance. Implemented in Streamlit, the dashboard offered expert annotators a simple interface to translate and review claim–label pairs, making it easier to ensure high-quality and consistent Urdu translations.

C URDUFACTCHECK Prompts

This section presents the prompts that power the core modules of URDUFACTCHECK.

CLAIMPROCESSOR Prompt

آپ کو ایک ایسا متن دیا گیا ہے جس میں علم کے دعوے شامل ہیں۔
دعویٰ ایک بیان ہے جو کچھ سچ یا جھوٹ ہونے کا دعویٰ کرتا ہے،
جس کی تصدیق انسانوں سے کی جا سکتی ہے۔ آپ کا کام یہ ہے
کہ آپ دیے گئے متن میں سے ہر دعوے کو درست طریقے سے
شناخت اور نکالیں۔ پھر، کسی بھی کورفرنس یعنی ضمیر یا
دوسرے حوالہ دینے والے اظہار کو دعوے کی وضاحت کے لیے
الفاظ سے کم اور 15 حل کریں۔ ہر دعویٰ مختصر یعنی
خود مختار ہونا چاہیے۔
متن اردو میں دیا گیا ہے اور دعوے اردو میں نکالے جانے چاہئیں۔
آپ کا جواب صرف نیچے دیے گئے فارمیٹ میں ہونا چاہیے۔
اس کے علاوہ کوئی اور اضافی نوٹس یا وضاحت شامل نہ کریں۔

جواب کا فارمیٹ:
]
}
"claim" : الفاظ سے کم ہو 15 یقین دہانی کرائیں کہ دعویٰ
اور مکمل خیال فراہم کرے۔ کورفرنس کو دعوے کی
,"وضاحت کے لیے حل کریں
,{
...
[
یہاں دو مثالیں دی گئی ہیں:
{examples}

اب یہ مکمل کریں، صرف جواب کی شکل میں، کوئی اور الفاظ نہیں:
{input} :[text]
:[response]

Listing 2: Our claim extraction prompt, which guides the identification of factual claims in Urdu text, with rules for co-reference resolution and concise, self-contained phrasing.

C.1 QUERYGENERATOR Prompt

```
آپ ایک سوالات بنانے والے ہیں جو صارفین کو دیے گئے
دعوے کو تلاش کے انجن کے ذریعے تصدیق کرنے میں مدد کرتے ہیں۔
آپ کا بنیادی کام دو موثر اور شک انگیز تلاش کے انجن کے
سوالات تیار کرنا ہے۔ یہ سوالات صارفین کو دیے گئے دعوے
کی حقیقت کو تنقیدی طور پر جانچنے میں مدد فراہم کریں گے۔
سوالات اردو میں ہونے چاہئیں اور سوالات اردو میں بنائے جائیں۔
آپ کو صرف نیچے دیے گئے فارمیٹ میں جواب دینا ہوگا
پائیتھون کی فہرست میں سوالات۔ براہ کرم اس فارمیٹ کی
سختی سے پیروی کریں۔ کچھ اور واپس نہ کریں۔ اپنا جواب
سے شروع کریں۔ '['

جواب کا فارمیٹ: ['2سوال', '1سوال']
یہاں دو مثالیں دی گئی ہیں:
{examples}

اب یہ مکمل کریں، صرف جواب کی شکل میں، کوئی اور الفاظ نہیں:
[claim]: {input}
[response]:
```

Listing 3: Query generation prompt used by the QUERYGENERATOR module to create search engine queries for fact-checking claims. The prompt ensures queries are generated in Urdu and follow a specific format for effective claim verification.

C.2 VERIFIER Prompt

```
آپ کو ایک ٹکڑا دیا گیا ہے۔ آپ کا کام یہ ہے کہ آپ یہ
شناخت کریں کہ آیا دیے گئے متن میں کوئی حقیقت کی
غلطیاں ہیں۔ جب آپ دیے گئے متن کی حقیقت کو پرکھ
رہے ہوں، تو آپ ضرورت کے مطابق فراہم کردہ شواہد
کا حوالہ دے سکتے ہیں۔ فراہم کردہ شواہد مددگار ہو سکتے ہیں۔
بعض شواہد ایک دوسرے سے متضاد ہو سکتے ہیں۔ آپ کو
شواہد کو احتیاط سے استعمال کرنا چاہیے جب آپ دیے گئے
متن کی حقیقت کا اندازہ لگائیں۔

جواب ایک ڈکشنری ہونی چاہیے جس میں چار کلیدیں ہوں
"reasoning" (وجہ), "factuality" (حقیقت)
"error" اور (غلطی) "correction" (تصحیح),
جو بالترتیب آپ کی وجہ، یہ کہ آیا دیے گئے متن میں کوئی حقیقتی
غلطی ہے یا نہیں "Boolean True or False"
اور غلطی کی وضاحت، اور تصحیح فراہم کریں۔
وجہ، غلطی اور تصحیح اردو میں ہونی چاہیے۔

جواب کا فارمیٹ:
{
"reasoning": کیوں دی گئی عبارت حقیقت پر مبنی ہے یا نہیں؟
جب آپ یہ کہتے ہیں کہ کوئی چیز حقیقت پر مبنی نہیں ہے،
تو آپ کو اپنے فیصلے کی حمایت کرنے کے لیے متعدد
شواہد فراہم کرنے ہوں گے۔
"error": اگر عبارت حقیقت پر مبنی ہے تو 'None'، ورنہ
غلطی کی وضاحت کریں۔
"correction": اگر کوئی غلطی ہو تو تصحیح شدہ عبارت
فراہم کریں۔
"factuality": اگر دی گئی عبارت حقیقت پر مبنی True
ہے ورنہ False
}

اب یہ مکمل کریں، صرف جواب کی شکل میں، کوئی اور الفاظ نہیں:
[claim]: {claim}
[evidence]: {evidence}
[response]:
```

Listing 4: Verification prompt used by the VERIFIER module to assess claim factuality based on retrieved evidence. The prompt ensures systematic evaluation and structured output with reasoning, error identification, and corrections in Urdu.

Urdu to English Translator Prompt

```
You are given a piece of text in Urdu. Your
task is to translate it into English. The
translation should be accurate and maintain
the original meaning of the text. Please
ensure that the translation is grammatically
correct and coherent in English.
DO NOT RESPOND WITH ANYTHING ELSE. ADDING ANY
OTHER EXTRA NOTES THAT VIOLATE THE RESPONSE
FORMAT IS BANNED.

{input}
```

Listing 5: Prompt for translating Urdu text into English.

C.3 English to Urdu Translator Prompt

```
You are given a piece of text in English.
Your task is to translate it into Urdu. The
translation should be accurate and maintain
the original meaning of the text. Please
ensure that the translation is grammatically
correct and coherent in Urdu. DO NOT RESPOND
WITH ANYTHING ELSE. ADDING ANY OTHER EXTRA
NOTES THAT VIOLATE THE RESPONSE FORMAT IS
BANNED.
{input}
```

Listing 6: Prompt for translating English text into Urdu.