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

Anonymous ACL submission

Abstract

The rapid use of large language models (LLMs) has raised critical concerns regarding the factual reliability of their outputs, especially in low-resource languages such as Urdu. Existing automated fact-checking solutions overwhelmingly focus on English, leaving a significant gap for the 200+ million Urdu speakers worldwide. In this work, we introduce URDUFACTCHECK, the first comprehensive, modular fact-checking framework specifically tailored for Urdu. Our system features a dynamic, multi-strategy evidence retrieval pipeline that combines monolingual and translation-based approaches to address the scarcity of high-quality Urdu evidence. We curate and release two new hand-annotated benchmarks: URDUFACTBENCH for claim verification and URDUFACTQA for evaluating LLM factuality. Extensive experiments demonstrate that URDUFACTCHECK, particularly its translation-augmented variants, consistently outperforms baselines and open-source alternatives on multiple metrics. We further benchmark twelve state-of-the-art (SOTA) LLMs on factual question answering in Urdu, highlighting persistent gaps between proprietary and open-source models. URDUFACTCHECK's code and datasets are open-sourced and publicly available at `[URLredacted]`.

1 Introduction

In recent years, the way we find and share information has changed dramatically. Large language models (LLMs) like GPT-4o (Achiam 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., 2023; Geng et al., 2023).

At the same time, social media platforms have become the main source of news and information for millions of people worldwide. Unfortunately, these platforms are also a hotbed for rumors, fake news, and viral misinformation. As noted during major world events such as the 2016 U.S. Presidential Election and the Brexit referendum, false narratives have been used to manipulate public opinion at scale (Allcott and Gentzkow, 2017; Pogue, 2017; Vosoughi et al., 2018). The rapid and algorithm-driven dissemination of such content, especially on platforms like TikTok, Facebook, and Twitter, has led to the erosion of public trust in institutions and an increase in political polarization (Zimmer et al., 2019; Trilling et al., 2017). This phenomenon was further exacerbated during the COVID-19 pandemic, which not only heightened public awareness of misinformation but also revealed its dangers in real-time. The World Health Organization (WHO) famously warned that we were facing not just a pandemic but also an 'infodemic', a surge of false or misleading information about the virus circulating on social media platforms (Humprecht, 2020; Arechar et al., 2023; Organization, 2023).

Despite the growing momentum of fact-checking efforts in recent years, most initiatives remain focused exclusively on English-language content (Guo et al., 2022), which leaves a substantial gap for other major languages. Urdu is the national language and lingua franca of Pakistan, holds official status in several Indian states, and is spoken by an estimated 232 million people worldwide, including both native and second-language speakers. Yet, it has a minimal digital footprint, accounting for less than 0.5% of all online content (ICLS, 2024). This gap is particularly concerning given the high prevalence of fake news and misinformation circulating in Urdu on social media platforms. Such content often spreads rapidly, sometimes in the

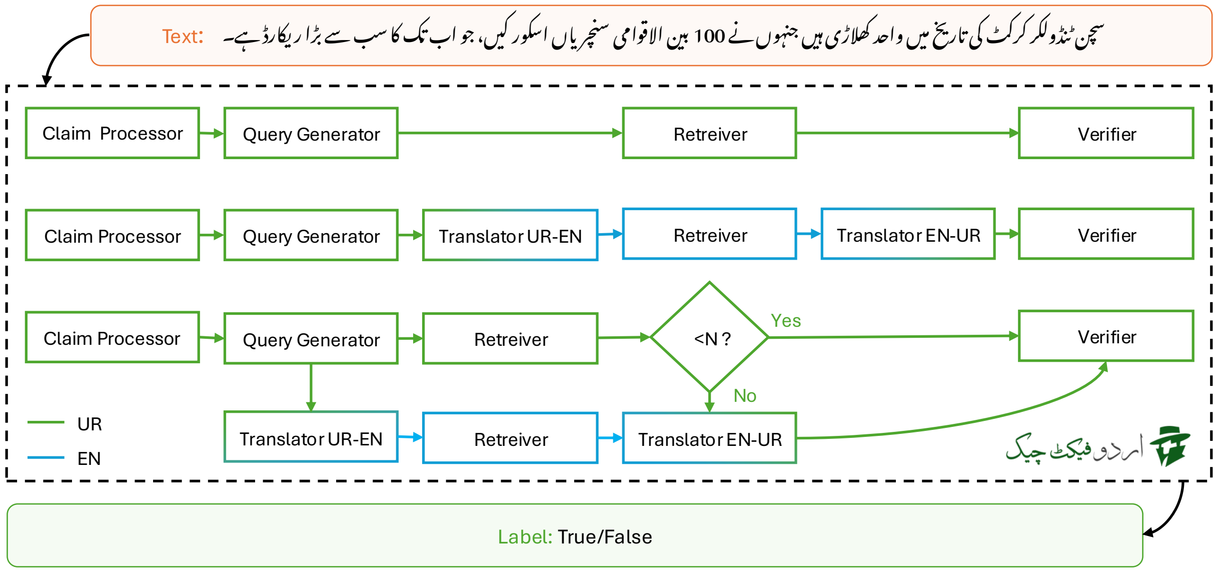

Figure 1: Three core fact-checking pipelines of URDUFACTCHECK. **First pipeline (top)** provides an end-to-end urdu based fact-checking framework. **Second pipeline (middle)** introduces translators to improve the issue of low-quality evidence. **Third pipeline (bottom)** introduces thresholded translation to reduce cost overhead.

form of jokes or memes, but also as genuine disinformation (Amjad et al., 2022). While there has been some encouraging progress in applying cross-lingual transfer learning to low-resource languages, particularly for hate speech and rumor detection tasks (Glavaš et al., 2020; Haider et al., 2023), these advances have rarely extended to automated fact-checking in Urdu. As a result, the area remains largely unexplored and underserved, highlighting a pressing need for research and specialized tools aimed at supporting factuality assessment in the Urdu language. To address this, we introduce **URDUFACTCHECK**.

URDUFACTCHECK is inspired by recent advances in modular fact-checking frameworks, such as LOKI (Li et al., 2024), OPENFACTCHECK (Wang et al., 2025, 2024; Iqbal et al., 2024), and FIRE (Xie et al., 2024), and is designed to address several key challenges in Urdu:

- Identifying factual errors in free-form text;
- Boosting the quality and availability of evidence in contexts;
- Systematically evaluating the factuality capabilities of LLMs for Urdu;
- Determining which automated fact-checker performs best and identifying the pipeline components that contribute most to overall verification accuracy;

To address these challenges, this work introduces three core resources.

URDUFACTCHECK: An end-to-end fact-checking pipeline tailored for the Urdu language, capable of detecting factual errors in free-text input. To mitigate the issue of sparse and low-quality Urdu evidence, the pipeline also incorporates thresholded evidence boosting technique for improved retrieval and verification as shown in Figure 1.

URDUFACTQA: A hand-annotated dataset specifically constructed to evaluate the factual accuracy of LLMs on Urdu QA tasks. Using URDUFACTQA, the factual performance of 12 state-of-the-art LLMs is systematically assessed.

URDUFACTBENCH: A manually curated benchmark for claim verification in Urdu, supporting thorough evaluation of automated fact-checking systems. This benchmark enables direct comparison between URDUFACTCHECK and other existing automated fact-checkers.

In summary, URDUFACTCHECK marks a significant step toward democratizing fact-checking technologies for low-resource languages. It provides a practical, extensible, and open-source solution to help researchers, journalists, and developers evaluate factuality in Urdu texts, whether generated by humans or machines, and sets the stage for future cross-lingual and culturally inclusive fact-checking systems.

2 Related Work

Prior efforts in Urdu fact-checking have focused mainly on classification. The URDU-FAKE@FIRE2021 (Amjad et al., 2022) shared task addressed fake news detection as a binary classification problem, revealing generalization challenges under domain shifts. AX-TO-GRIND (Harris et al., 2023) expanded this space by introducing a large-scale annotated Urdu dataset and applying multilingual models like MBERT and XLNET to achieve strong performance. More recently, HOOK AND BAIT URDU (Harris et al., 2025) introduced the largest fake news corpus for Urdu to date, leveraging LoRA-based fine-tuning of LLAMA-2 for both monolingual and multilingual fake news detection, and achieving high F1-scores and accuracy. While these systems represent important advances in classification performance and dataset scale, they do not support end-to-end factuality pipelines or evaluation of generated text.

In parallel, several Urdu Question-Answering (QA) datasets have been proposed. UQA (Arif et al., 2024) uses span-preserving translation of SQUAD2.0 (Rajpurkar et al., 2018) and serves as a benchmark for multilingual models including MBERT and XLM-ROBERTA. Other corpora such as UQUAD1.0 (Kazi and Khoja, 2021) support extractive QA but do not assess the factuality of model outputs. Notably, these available QA datasets are primarily focused on general QA and reading comprehension tasks, rather than on fact-checking or evaluating the factual correctness of generated responses.

While multilingual benchmarks such as X-FACT (Gupta and Srikumar, 2021) test LLM factuality across several low-resource languages, Urdu remains under-represented in these evaluations. Additionally, recent tools like FACTSCORE (Min et al., 2023), FACTOOL (Chern et al.), and FACTCHECK-GPT (Wang et al., 2023) have advanced metrics, retrieval, and modularity in fact-checking, but are typically built for English and lack support for Urdu-specific tasks or datasets.

3 Datasets

To enable rigorous evaluation of automated fact-checkers and the factual capabilities of LLMs, we curated a diverse collection of five datasets spanning the tasks of claim verification and factual QA, introducing both URDUFACTQA and URDUFACT-BENCH.

3.1 Dataset Collection

Given the near-complete absence of high-quality factual datasets in Urdu, we created a multi-stage process to bring proven English-language resources into Urdu with careful expert supervision at each step. For claim verification, we selected three datasets: BINGCHECK (Li et al., 2023), FACTCHECK-BENCH (Wang et al., 2023), and FACTOOL (Chern et al., 2023). FACTOOL offers factuality claims from a range of domains, so we specifically selected examples that require world knowledge to verify. This subset is referred to as FACTOOL-QA. To evaluate the factual question answering abilities of LLMs, we included two widely-used QA datasets: SIMPLEQA (Wei et al., 2024) and FRESHQA (Vu et al., 2023). These datasets were selected for their diversity in claim structure, domain coverage, and format, ensuring a comprehensive evaluation of both claim-level factuality and broader LLM factual behavior. Particular emphasis was placed on claims that require models to leverage external knowledge, as this is a critical aspect of robust factual verification.

For both FACTCHECK-BENCH and BINGCHECK, we simplified the original four-label classification scheme (*supported*, *partially supported*, *not supported*, *refuted*) into a binary format. Specifically, *supported* and *partially supported* were merged into *True*, *refuted* was mapped to *False*, and instances labeled as *not supported* were excluded. This binarization ensures consistency across all datasets and streamlines the evaluation process by focusing on clear factuality decisions.

To address class imbalance in BINGCHECK (which contained 3,581 *True* claims and only 42 *False* claims), we sampled 100 *True* claims for inclusion in our test set. This approach provided a more balanced and manageable dataset, allowing evaluation metrics to better reflect system performance on both classes.

3.2 Translation and Annotation

To kick-start the translation process, GPT-4O, a state-of-the-art LLM, was used to generate initial Urdu translations for all datasets in a few-shot setup. 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 further improve machine translation quality, expert annotators hand-crafted 100 demonstration examples (20 per dataset), which were incorporated as examples in the LLM's few-shot setup. A custom prompt template, enriched with formal linguistic guidelines, was developed to support the model in producing grammatically correct and fluent Urdu. The LLM was instructed to retain proper technical terms and noun transliterations, handle left-to-right (LTR) numerals and tokens with right-to-left (RTL) sentence flow, and avoiding improper placement of acronyms or LTR content in Urdu syntax (see Appendix A for details). For optimal few-shot performance, Max Marginal Relevance (MMR) was used to select the most relevant examples. The translation pipeline was implemented using LANGCHAIN[1] with an output parser to ensure structured and accurate outputs. For GPT-4O default parameters of OpenAI Library were used.

Following machine translation, every dataset underwent a rigorous dual-annotation process. Each translated dataset was first reviewed by one expert annotator and then independently validated by a second annotator, providing checks for linguistic consistency, factual correctness, and cultural appropriateness. The annotation process was further supported by a custom-built annotation portal that enabled annotators to efficiently view and verify translations (see Appendix B for details). This workflow ensured that all datasets ultimately 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 for full statistics)

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

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

URDUFACTQA: Consisting of the QA datasets SIMPLEQA and FRESHQA, this resource is designed for evaluating the factuality capabilities of LLMs in Urdu. (see Table 2 for full statistics)

Together, these resources help bridge the severe resource gap in Urdu NLP and enable 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., 2024; Iqbal et al., 2024; Xie et al., 2024; Chern et al., 2023).

4.1 Prompt Engineering for Core Modules

URDUFACTCHECK framework adopts an agentic architecture, where each module operates as a specialized agent fulfilling a distinct role in the fact-checking process. The initial step in building the URDUFACTCHECK framework involves carefully designing prompts for each agent. Each prompt is crafted to address the unique linguistic and contextual challenges of Urdu, ensuring that every agent produces coherent outputs throughout the pipeline.

The prompt for CLAIMPROCESSOR not only decomposes input text into atomic claims (concise, self-contained statements that can be independently fact-checked) but also ensures that each claim is decontextualized and check-worthy. This careful prompt engineering guides the agent to consistently identify claims suitable for factual verification.

For QUERYGENERATOR, the prompt is tailored to generate context-aware web search queries for each atomic claim. Specifically, two types of queries are generated for each claim: a question-based query that conceals the explicit fact, and a direct claim-based query that includes the exact factual statement. This dual-query approach ensures both broad coverage in evidence retrieval, maximizing the likelihood of finding authoritative information to support or refute each claim.

RETRIEVER, in contrast, does not require prompt engineering, as it directly interacts with SERP APIs to conduct web searches and return summarized snippets for each query. In our framework, we utilize the Google SERP API[2] to retrieve relevant web content.

Finally, the prompt for VERIFIER enables it to thoroughly evaluate the collected evidence and assign a factuality label (true or false) to each atomic claim. Beyond basic labeling, the prompt also guides the agent to also provides clear *reasoning* for its decision, explains its understanding based on the available evidence, and, where appropriate, suggests *corrections* for the atomic claim.

Each prompt also includes two-to-three examples to guide the agents' outputs. Detailed prompt templates are provided in Appendix C.

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.

[2]https://serper.dev

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 robust verification coverage for claims with insufficient Urdu evidence, dynamically invoking and combining the translation-based strategy only when necessary.

To facilitate seamless transitions between Urdu and English, we engineered dedicated prompts for both Urdu-to-English and English-to-Urdu translation tasks. All translation is performed by an LLM agent, ensuring accurate and context-aware conversions that preserve the original meaning throughout the evidence retrieval pipeline.

This tiered retrieval framework allows URDUFACTCHECK to adaptively maximize recall and reliability while minimizing unnecessary cost, effectively addressing the core challenge of evidence scarcity in low-resource languages like Urdu.

5 Experiments

To assess the effectiveness of URDUFACTCHECK and the utility of our annotated benchmarks, we conduct a comprehensive suite of experiments. Specifically, we (i) investigate the impact of varying the evidence threshold parameter on retrieval and verification performance, (ii) benchmark the accuracy of automated fact-checkers using URDUFACTBENCH dataset, and (iii) evaluate the factuality of state-of-the-art LLMs with URDUFACTQA.

We report both the API costs associated with proprietary LLMs and the GPU rental expenses for open-source models, as well as the costs incurred from search engine queries and the overall time required for fact-checking. Experiments with open-source models were run on an NVIDIA RTX 6000 GPU, costing approximately $0.79 per hour. Each search query issued through SerpAPI resulted in an estimated cost of $0.00105 per query.

5.1 Threshold Tuning

A key hyperparameter in the evidence retrieval pipeline is the evidence threshold τ, which determines the minimum number of Urdu evidence snippets required before triggering fallback to translation-based retrieval. To understand the trade-offs between recall, accuracy, and computational efficiency, we systematically vary τ across the set 1, 3, 5, 7, 9. For each threshold value, we evaluate the system's performance on the FACTCHECK-BENCH subset of URDUFACTBENCH, recording factual verification accuracy and retrieval cost. This analysis enables us to identify the optimal threshold setting that balances high recall and low cost.

All threshold experiments are conducted using GPT-4O-MINI as the backbone language model, with a temperature setting of 0 and a maximum token limit of 2500. All other parameters are kept at their default values.

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 and peaks at thresholds 5 and 9. This suggests that requiring a moderate amount of Urdu evidence before falling back to translation-based retrieval maximizes verification performance. However, this gain comes with a trade-off in cost, which remains stable between thresholds 3 and 5 but rises sharply at 7 and 9.

Overall, setting τ in the range of 3–5 appears to provide a favorable balance between improved

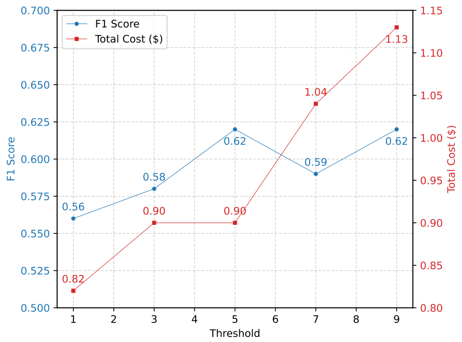

Figure 2: Impact of evidence threshold τ on F1 score (left axis, blue) and total cost (right axis, red) on the FACTCHECK-BENCH subset.

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 Language Models

To enable a fair comparison for automated fact-checking in Urdu, we evaluate a range of state-of-the-art (SOTA) language models. Our experiments include proprietary models from two leading families—GPT models (OpenAI, 2024a,b) and Claude models (Anthropic, 2024) as well as two high-performing open-source models: MISTAL-INST 7B (Jiang et al., 2023) and LLAMA3.1-INST 8B (Dubey et al., 2024). This evaluation provides insight into the current capabilities of LLMs for factual verification tasks in Urdu.

Results Analysis Table 3 shows performance comparison of different LLMs. Proprietary models consistently achieve better results than open-source alternatives, likely due to their greater size and more advanced training in reasoning and tool use. Within the proprietary category, the most recent offerings—such as GPT-4.1 from OpenAI—stand out with the highest performance. Interestingly, the budget friendly GPT-4.1-MINI and GPT-4O-MINI also give a comparable performance and also offers an impressive 7̃00x reduction in cost. This suggests that for fact-checking tasks, the absolute best-performing models may not be necessary.

5.3 Fact-Checker Benchmarking

To comprehensively evaluate URDUFACTCHECK, we benchmarked its performance on the two remaining subsets of URDUFACTBENCH: FACTOOL-QA and BINGCHECK. As there are

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 comparisons between different language models for URDUFACTCHECK (Thresholded Retrieval $\tau = 5$) on FACTCHECK-BENCH.

currently no end-to-end fact-checking systems designed for Urdu, direct comparisons with other tools are limited. We attempted to include FACTOOL in our evaluation; however, it produced unsatisfactory results on Urdu inputs. For this reason, we did not consider other English-based fact-checkers, since their outputs would not provide a fair basis for Urdu-language verification.

Our benchmarking involved three core variants of URDUFACTCHECK: monolingual, translated (TR), and thresholded translated (TH-TR). For the thresholded variant, we experimented with evidence thresholds τ set at 3, 5, 7, and 9, yielding a total of six different URDUFACTCHECK configurations. All experiments were conducted using two backbone LLMs: GPT-4O and GPT-4O-MINI.

Results Analysis Table 4 demonstrates that all URDUFACTCHECK variants outperform FACTOOL and trivial baselines by a substantial margin. Among the URDUFACTCHECK approaches, translation-based methods (TR and TH-TR) achieve the best overall performance, especially with GPT-4O, which obtains F1 scores as high as 0.79 for true labels. This highlights the importance of accessing English evidence in supplementing limited Urdu web content. The thresholded translated (TH-TR) variants offer strong results, particularly with moderate threshold settings such as $\tau = 5$, balancing accuracy and computational cost effectively. Cost analysis reveals that GPT-4O-MINI delivers competitive accuracy relative to GPT-4O, but at a dramatically lower operational cost. This makes it an attractive backbone for scenarios with budget constraints.

5.4 Evaluating LLM Factuality

To assess the factual accuracy of large language models in Urdu, we evaluated twelve state-of-the-art LLMs using the URDUFACTQA benchmark. Our evaluation covers both proprietary and open-source models. The proprietary group includes GPT-4O and O4-MINI, while the open-source group consists of ALIF (8B), which is specifically tailored for Urdu, as well as 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).

For each question in URDUFACTQA, responses were collected from all twelve models. GPT-based models were run with default decoding parameters from the OpenAI API, while open-source models used the default generation settings provided in the Hugging Face Transformers library. After generating free-form answers, we automatically evaluated their factuality using the translation-based URDUFACTCHECK pipeline (TR variant), with GPT-4O-MINI as the verifier and Google Serper as the evidence retriever, applying the factuality prompts from URDUFACTQA.

Results Analysis As illustrated in Figure 3, proprietary models such as GPT-4O and O4-MINI achieve the highest percentages of factually correct responses on both the SIMPLEQA and FRESHQA subsets, with true claim rates ranging upto 46%. In contrast, open-source models i.e. ALIF-8B, LLAMA-3 series, and modelQwen series yield lower factual accuracy, typically achieving less than 25% true claims, although QWEN 2.5 (72B) stands out as the most competitive among them.

The questions in the SIMPLEQA subset are relatively more challenging for LLMs, resulting in a lower percentage of true claims compared to FRESHQA. Additionally, the lower number of false claims observed in FRESHQA is largely due to its smaller dataset size (600) compared to SIMPLEQA (4,326). Computational costs are relatively similar across the board.

Overall, these results indicate that while proprietary models deliver the best factual accuracy in Urdu, progress in open-source models remains limited. This highlights both the value of high-resource, instruction-tuned LLMs for factual question answering in low-resource languages and the need for continued development of competitive open-source solutions for Urdu.

6 Conclusion

In this work, we introduced URDUFACTCHECK, the first comprehensive automated fact-checking system tailored for the Urdu language. Our modular, agentic framework addresses key challenges

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: Performance comparisons between different frameworks across multiple datasets. Red indicates the lowest performance, while green indicates the highest.

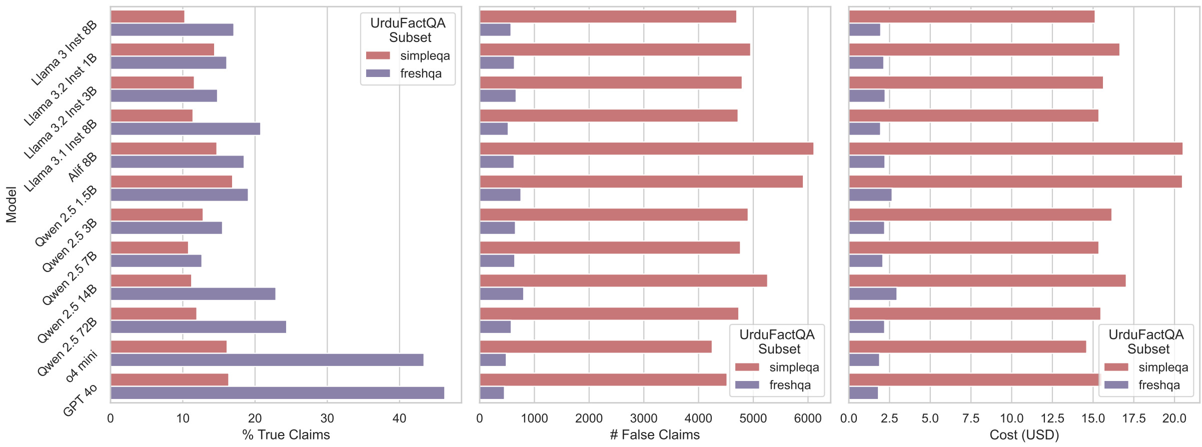

Figure 3: Automatic factuality evaluation results for 12 SOTA LLMS on URDUFACTQA using URDUFACTCHECK-TR. *left:* the percentage of true claims, *center:* the number of false claims, and *right:* the cost of using URDUFACTCHECK-TR in USD.

in low-resource factual verification, including the scarcity of high-quality Urdu evidence and the absence of existing end-to-end Urdu fact-checking tools. Through a multi-strategy evidence retrieval pipeline and curated translation prompts, URDUFACTCHECK dynamically balances retrieval accuracy, recall, and computational cost.

We also presented two new annotated benchmarks: URDUFACTBENCH for claim verification and URDUFACTQA for evaluating the factual capabilities of large language models in Urdu. Extensive experiments demonstrated that URDUFACTCHECK, particularly its translation-augmented variants, consistently outperforms existing baselines and open-source frameworks, establishing robust new standards for Urdu factual verification.

Our large-scale evaluation of twelve state-of-the-art LLMs on URDUFACTQA further highlighted a persistent gap between proprietary and open-source models in Urdu factuality, while also underscoring the growing potential of models specifically tailored for Urdu. By making our system, datasets, and evaluation framework publicly available, we hope to spur further research in low-resource fact-checking, bridge the digital resource gap for Urdu, and provide scalable solutions to combat misinformation in linguistically diverse settings.

7 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 resource 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.

8 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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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 arXiv:2207.05133*.

Anthropic. 2024. Introducing the next generation of claude.

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. *arXiv preprint arXiv:2405.01458*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, and 1 others. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv: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 arXiv:2307.13528*.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios, 2023. *URL https://arxiv. org/abs/2307.13528*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 8 others. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2023. A survey of language model confidence estimation and calibration. *arXiv preprint arXiv:2311.08298*.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. Xhate-999: Analyzing and detecting abusive language across domains and languages. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2106.09248*.

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*, pages 1358–1364.

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 *2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 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.

ICLS. 2024. 10 most spoken languages in the world in 2025 — icls.edu. `https://www.icls.edu/blog/most-spoken-languages-in-the-world`.

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. Openfactcheck: A unified framework for factuality evaluation of llms. *arXiv preprint arXiv:2408.11832*.

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 arXiv:2111.01543*.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2024. Loki: An open-source tool for fact verification. *arXiv preprint arXiv:2410.01794*.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2023. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

OpenAI. 2024a. Hello gpt-4o.

OpenAI. 2024b. Introducing openai o1-preview.

World Health Organization. 2023. An overview of infodemic management during the covid-19 pandemic, january 2020–july 2022.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

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 arXiv: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 1 others. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, and 1 others. 2023. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. *arXiv preprint arXiv:2311.09000*.

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi Georgiev, Jiahui Geng, and Preslav Nakov. 2024. Openfactcheck: A unified framework for factuality evaluation of llms. *arXiv preprint arXiv:2405.05583*.

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*, pages 11399–11421.

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 arXiv:2411.04368*.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2024. Fire: Fact-checking with iterative retrieval and verification. *arXiv preprint arXiv:2411.00784*.

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

This prompt was used to perform pre-translation of all data instances, significantly accelerating the annotation process for both URDUFACTQA and URDUFACTBENCH.

```
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., "January 1, 2020" → "یکم جنوری 2020").

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

Here are a few examples of claims and expected translations:
<EXAMPLES>

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

### Formated Instructions:
{format_instructions}
```

Figure 4: Prompt for pre-translation before expert annotation of URDUFACTBENCH and URDUFACTQA

B URDUFACTCHECK Annotator Dashboard

To streamline dataset creation and quality assurance, we developed a dedicated annotator dashboard for URDUFACTCHECK using Streamlit. This dashboard was provided to expert annotators to simplify the annotation process, making it easier to review translations and ensure high-quality, consistent data across both URDUFACTQA and URDUFACTBENCH.

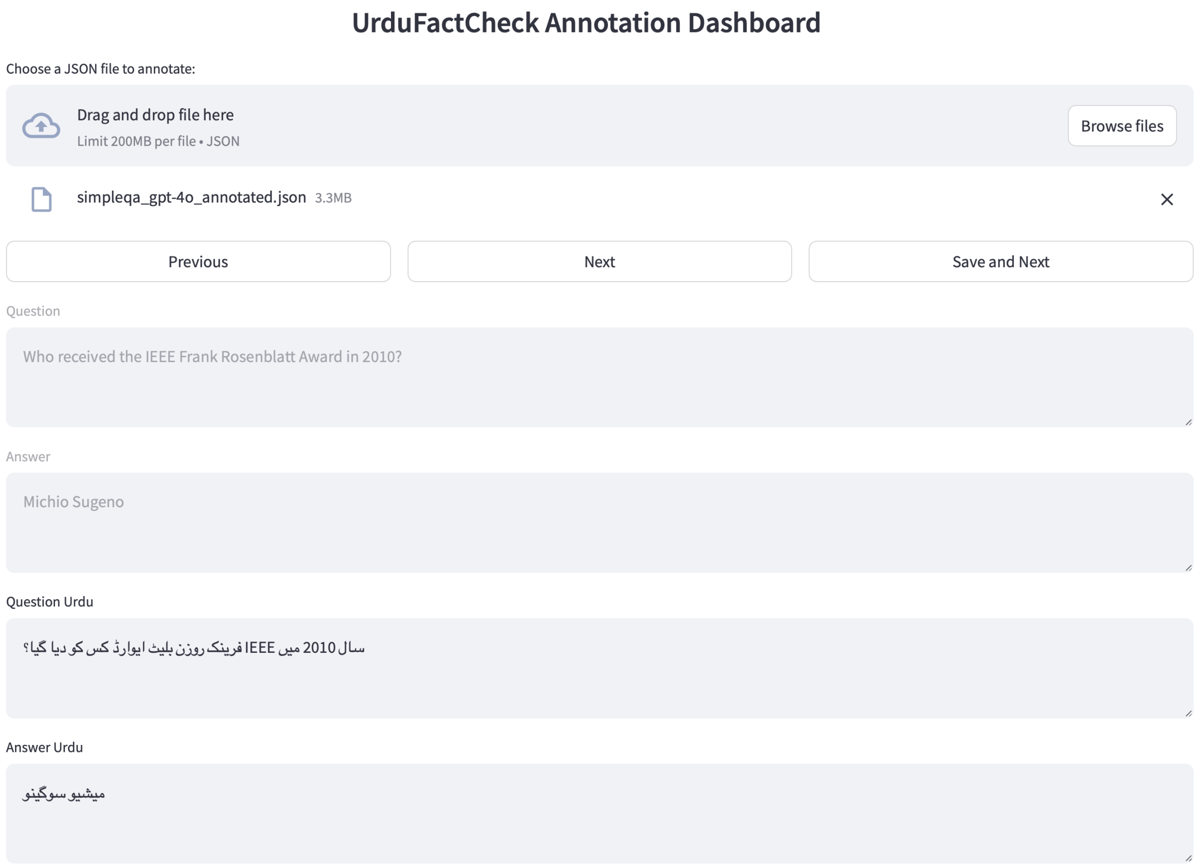

Figure 5: URDUFACTCHECK Annotator Dashboard.

C URDUFACTCHECK prompts

This section provides the custom-designed prompts used for the core modules of URDUFACTCHECK, including the CLAIMPROCESSOR, QUERYGENERATOR, and VERIFIER, each tailored to handle the unique linguistic and contextual challenges of Urdu factuality evaluation.

C.1 Claim Processor Prompt

```
CLAIM_PROCESSOR_PROMPT = {
    "system": "براہ کرم وہ دعویٰ فراہم کریں جس کی آپ حقیقت جانچنا چاہتے ہیں۔",
    "user": """
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 (ضمیر یا دوسرے حوالہ دینے والے اظہار) کو دعوے کی وضاحت کے لیے حل کریں۔ ہر دعویٰ مختصر (15 الفاظ سے کم) اور 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": "یقین دہانی کرائیں کہ دعویٰ 15 الفاظ سے کم ہو اور مکمل خیال فراہم کرے۔ کورفرنس کو دعوے کی وضاحت کے لیے حل کریں",
}},
…
]


یہاں دو مثالیں دی گئی ہیں:
[text]: کرکٹ میچ میں شعیب ملک نے 50 رنز بنائے۔ وہ پاکستان کے بہترین کھلاڑی ہیں۔ شعیب ملک کو اگلے میچ میں شامل کیا جائے گا۔
[response]: [{{"claim": "شعیب ملک نے 50 رنز بنائے"}}, {{"claim": "شعیب ملک پاکستان کے بہترین کھلاڑی ہیں"}}, {{"claim": " شعیب ملک کو اگلے میچ میں شامل کیا جائے گا"}}]

[text]: لاہور میں موسم خوشگوار رہا۔ لوگ پارکوں میں چلنے پھرنے گئے۔ حکام نے کہا کہ کل بارش ہو سکتی ہے۔
[response]: [{{"claim": "لاہور میں موسم خوشگوار رہا"}}, {{"claim": "لوگ پارکوں میں چلنے پھرنے گئے"}}, {{"claim": " حکام نے کہا کہ کل بارش ہو سکتی ہے"}}]

[text]: {input}
[response]:
""",
}
```

Figure 6: CLAIMPROCESSOR prompt.

C.2 Urdu to English Translator Prompt

```
URDU_TO_ENGLISH_TRANSLATION_PROMPT = {
    "system": "You are a helpful assistant.",
    "user": """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}
""",
}
```

Figure 7: Urdu to English translator prompt.

C.3 Query Generator Prompt

```
QUERY_GENERATION_PROMPT = {
    "system": "آپ ایک سوالات بنانے والا ہیں جو دیے گئے دعوے کو تصدیق کرنے کے لیے موثر اور جامع تلاش کے انجن کے سوالات تیار کرتا ہے۔ آپ صرف پائیتھون کی فہرست کی شکل میں جواب دیں گے (کسی اور الفاظ میں نہیں!)",
    "user": """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]:
[
    'سوال1',
    'سوال2'
]

یہاں تین مثالیں ہیں:
دعویٰ: ٹویٹر کے سی ای او بل گیٹس ہیں۔
جواب: ["ٹویٹر کے سی ای او کون ہیں؟", "سی ای او ٹویٹر"]

دعویٰ: مائیکل فیلپس تمام اوقات کے سب سے زیادہ سجاوٹی اولمپین ہیں۔
جواب: ["تمام اوقات کے سب سے زیادہ سجاوٹی اولمپین کون ہیں؟", "مائیکل فیلپس"]

دعویٰ: چیٹ جی پی ٹی کو گوگل نے بنایا ہے۔
جواب: ["چیٹ جی پی ٹی کو کس نے بنایا؟", "چیٹ جی پی ٹی"]

اب یہ مکمل کریں، صرف جواب کی شکل میں، کوئی اور الفاظ نہیں:
دعویٰ: {input}
جواب:
""",
}
```

Figure 8: QueryGenerator prompt.

C.4 English to Urdu Translator Prompt

```
ENGLISH_TO_URDU_TRANSLATION_PROMPT = {
    "system": "You are a helpful assistant.",
    "user": """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}
""",
}
```

Figure 9: English to Urdu translator prompt.

C.5 Verifier Prompt

```
VERIFICATION_PROMPT = {
"system": "آپ ایک شاندار معاون ہیں۔",
"user": """
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 تین کلیدیں ہوں - "reasoning" (وجہ)، "factuality" (حقیقت) اور "error" (غلطی) اور "correction" (تصحیح)، جو بالترتیب آپ کی وجہ، یہ کہ آیا
دیے گئے متن میں کوئی حقیقتی غلطی ہے یا نہیں (Boolean - True یا False)، اور غلطی کی وضاحت، اور تصحیح فراہم کریں۔
وجہ، غلطی اور تصحیح اردو میں ہونی چاہیے۔

یہ ہے دی گئی عبارت
{text]: {claim]
یہ ہیں فراہم کردہ شواہد
{evidences]: {evidence]
آپ کو صرف نیچے دیے گئے فارمیٹ میں جواب دینا چاہیے۔ کچھ اور واپس نہ کریں۔ اپنے جواب کا آغاز '{{' سے کریں۔
[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": True اگر دی گئی عبارت حقیقت پر مبنی ہے، ورنہ False۔
}}
""",
}
```

Figure 10: VERIFIER prompt.