# URDUBENCH: A Unified Benchmark for Evaluating Large Language Models on Native Urdu Tasks

Anonymous ACL submission

## Abstract

Large Language Models (LLMs) have driven rapid advances in natural language processing (NLP); however, low-resource languages such as Urdu, spoken by over 230 million people, remain severely underrepresented, limiting equitable deployment and widening multilingual performance gaps. Existing Urdu benchmarks are fragmented or translation dependent, lacking a unified framework for evaluating emerging efficient models on native, culturally grounded tasks. We present URDUBENCH, a comprehensive benchmark comprising 23 datasets across 19 tasks for Urdu LLM evaluation, covering natural language understanding, safety-critical moderation, and generation. The framework supports multi-GPU parallel execution, enabling scalable evaluation even in resource-constrained settings. Using this framework, we benchmark 14 open-weight instruction-tuned LLMs spanning nano (<1B), small (1–3B), and medium (up to 7B) parameter scales. Due to resource constraints and our focus on low-resource deployment settings, we do not include models larger than 7B parameters. Our evaluation emphasizes computationally efficient models suitable for real-world use in constrained environments. Results show pronounced performance disparities across model sizes and task categories, with persistent difficulties in Urdu sequence labeling and generation, and consistent gains from larger multilingual models.

## 1 Introduction

Large Language Models (LLMs) have significantly advanced natural language processing (NLP), showing strong performance across tasks such as question answering, summarization, reasoning, and generation. However, most of these advancements focus on high-resource languages like English, where abundant data and benchmarks enable effective evaluation and improvement. In contrast, low and mid resource languages, including Urdu spoken by over 230 million people worldwide (SIL International, 2022) remain underrepresented in LLM research, limiting equitable AI deployment and reinforcing performance gaps in multilingual systems.

At the same time, the ecosystem of open-source LLMs is undergoing a rapid transition toward smaller, efficient models. Sub-billion-parameter models (e.g., gemma-3-270m, granite-4.0-350m, Qwen2.5-0.5B-Instruct) and lightweight 1–3B models (e.g., gemma-3-1b-it, Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Qwen2.5-1.5B, granite 4.0 micro 3B, UrduLLaMA-1.1-Tiny) are increasingly capable of high-quality reasoning and instruction-following, while 4–7B models (e.g., Qwen2.5-7B-Instruct, phi-4 mini, Qwen 3 4B, gemma-3-4b-it, granite 4.0 h-tiny 7B) represent the emerging standard for deployable multilingual systems. Despite their practical relevance for resource-constrained settings, there exists *no systematic evaluation suite* designed to measure the Urdu capabilities of these model classes.

To address this gap, we introduce URDUBENCH, a comprehensive benchmark suite comprising 23 datasets across 19 tasks, designed to evaluate LLMs in Urdu. Our benchmark encompasses core NLP competencies, including core structural and classification tasks, ethical and safety-critical tasks, sentiment and emotion analysis, and generative tasks. Unlike translation-dependent multilingual resources, URDUBENCH integrates datasets that are culturally adapted, manually verified, and natively grounded in Urdu contexts. To facilitate community adoption, we release an open-source codebase for rapid evaluation, featuring a modular framework that allows seamless integration of new models or datasets.

We evaluate URDUBENCH on three model categories: **nano** (under 1B parameters) **small** (1–3B parameters) and **medium** (>3B–7B parameters). Our results reveal significant performance dispari-

ties across model sizes and tasks, underscoring the need for Urdu-specific fine-tuning and highlighting strengths in multilingual base models. Through this work, we aim to accelerate LLM development for Urdu and promote inclusive AI research.

The contributions of this paper are as follows:

- A novel benchmark, UrduBench, with 23 datasets spanning 19 tasks tailored for Urdu evaluation.
- An open-source, flexible evaluation framework for rapid testing of LLMs on Urdu tasks.
- Comprehensive analysis of 14 LLMs across three parameter scales, focusing on resource efficient models.

The remainder of the paper is organized as follows: Section 2 reviews related work, Section 3 details the construction of UrduBench, Section 4 elaborates the framework of UrduBench, Section 5 and Section 6 presents our experiments and results, and Section 7 concludes with future directions.

## 2 Related Work

Several benchmarks have been proposed to evaluate LLMs across diverse linguistic and cultural contexts. IndicGenBench (Singh et al., 2024) and the Indic LLM Leaderboard (Kumara and CognitiveLab Team, 2024) focus on Indic languages using translated versions of English benchmarks, covering tasks such as summarization, translation, QA, and reasoning. Similarly, SeaExam and SeaBench (Liu et al., 2025) provide native-crafted benchmarks for Southeast Asian languages, targeting both educational assessment and conversational evaluation.

For South Asian contexts, PakBBQ (Hashmat et al., 2025) and PakPolBias (Nadeem et al., 2025) address bias-related evaluation in Urdu and other Pakistani languages, focusing on social and political dimensions. While valuable, these benchmarks are task-specific and do not provide broad coverage of core NLP, safety, and generation tasks for Urdu. (Tahir et al., 2025) evaluated pre-trained LLMs on a range of Urdu NLP tasks; however, the study relied on a static set of now-outdated models and lacked a reusable evaluation framework, making results difficult to reproduce and requiring substantial re-engineering to assess newly released models, thereby underscoring the need for a systematic and extensible benchmarking suite.

Language-agnostic benchmarking frameworks such as LLMeBench (Dalvi et al., 2024) enable rapid evaluation across tasks and languages but do not provide native, culturally grounded datasets for low-resource languages. Broader multilingual efforts include AfroBench (Ojo et al., 2025) and IberBench (Ángel González et al., 2025), which aggregate large collections of datasets across African and Iberian languages, respectively, spanning both general-purpose and region-specific tasks.

Classic English benchmarks such as GLUE (Wang et al., 2019b), WinoGrande (Sakaguchi et al., 2019), and recent dynamic evaluations like LiveBench (White et al., 2024) have shaped LLM evaluation methodology but remain English-centric. Specialized benchmarks such as AD-LLM (Yang et al., 2025) target specific phenomena (e.g., anomaly detection) rather than comprehensive language evaluation.

In contrast to prior work, UrduBench provides the first unified, native benchmark for Urdu that jointly evaluates structural NLP, safety-critical moderation, sentiment and emotion analysis, and generation tasks, accompanied by a reproducible evaluation framework tailored to efficient open-weight LLMs. While existing evaluation frameworks such as LM Evaluation Harness and Lighteval have greatly facilitated standardized LLM assessment but they were insufficient for our purposes because they offer limited coverage of native Urdu-specific tasks.

## 3 UrduBench

UrduBench is a unified benchmarking framework for large-scale, reproducible evaluation of Urdu NLP tasks and models. Rather than introducing new datasets, it contributes by carefully aggregating, standardizing, and curating 23 existing high-quality Urdu datasets across 19 diverse tasks. This addresses the fragmentation and inconsistency in prior Urdu resources, enabling fair comparisons via uniform prompting and metrics. The framework reduces duplicated engineering effort by ensuring consistent handling of inputs, intermediate representations, and outputs through a pipeline-oriented execution model with structured key-value mappings.

UrduBench consists of three core components, dataset, task, and Model, coordinated by a central-

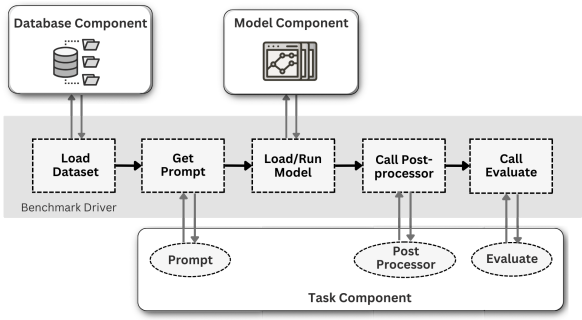


Figure 1: Overall architecture of URDUBENCH. A centralized Benchmark Driver orchestrates the Dataset, Task, and Model components through a unified execution pipeline.

ized Benchmark Driver, as shown in Figure 1.

### 3.1 Dataset Component

The Dataset Component is responsible for loading and preprocessing data. It supports multiple data formats and provides a standardized interface for accessing training and test splits. All datasets are registered through abstract interfaces, enabling easy extension with new datasets while maintaining consistency across tasks.

### 3.2 Model Component

The Model Component encapsulates model-specific inference logic behind a minimal and uniform interface. Each model implements a `prompt()` method that accepts role-based messages and returns generated text. To ensure robustness during large-scale evaluation, the framework includes an optional safety wrapper that catches runtime errors without interrupting the overall benchmarking process.

### 3.3 Task Component

The Task Component defines how inputs are converted into prompts, how model outputs are post-processed, and how predictions are evaluated. Each task specifies three core operations: prompt construction, output post-processing, and metric computation. By decoupling task logic from both datasets and models, URDUBENCH enables high reusability and consistent evaluation across different models and tasks.

### 3.4 Execution and Interaction

Benchmarking experiments are executed via a unified command-line interface, with representative examples shown in Listing 1. The framework supports three main evaluation modes, each configurable with the number of few-shot examples (`n_shots`):

1. **Single Task-Model Evaluation**: Evaluates a specific task on a selected model, with optional sample limits for debugging.
2. **Model-Centric Evaluation**: Evaluates one model across all registered tasks under a consistent few-shot setting.
3. **Task-Centric Evaluation**: Evaluates one task across all available models under identical few-shot conditions.

All intermediate results and artifacts are cached to ensure reproducibility and efficient re-execution.

```
# Single task-model evaluation
python run_bench.py --task <task_name>
    --model <model_name> --n_shots <
    n_shots>

# Model-centric evaluation (model across
     all tasks)
python run_bench.py --model <model_name>
     --n_shots <n_shots>

# Task-centric evaluation (task across
    all models)
python run_bench.py --task <task_name>
    --n_shots <n_shots>
```

Listing 1: Illustrative command-line interaction modes supported by UrduBench with configurable few-shot settings.

## 4 Framework Design and Properties

URDUBENCH is designed around core engineering principles to enable fair, reproducible, and scalable benchmarking for low-resource languages. The framework draws on best practices from modular LLM evaluation frameworks Dalvi et al. (2024) and unified benchmarks Gao et al. (2023); Wang et al. (2019a).

URDUBENCH employs a modular design where datasets, tasks, and model providers are registered through lightweight abstract interfaces. This decouples configuration from execution, facilitating easy extension without altering the core engine and ensuring performance differences reflect model capabilities rather than pipeline variations. The framework strictly separates task logic from model inference: tasks manage prompt construction, post-processing, and evaluation, while models implement only a minimal `prompt()` method. This ensures consistent prompting and decoding across

models, which is essential as small prompt variations can significantly influence results (Wang et al., 2019a).

### 4.1 Reproducibility and Execution Features

The framework incorporates fault-tolerant execution with configurable caching to resume after transient failures. All experiments are fully specified through configuration files and command-line arguments, with persistent storage of artifacts for transparent reporting. It natively supports systematic cross-product evaluation in both model-centric and task-centric modes, as well as partial dataset execution for rapid debugging.

### 4.2 Language-Independent Core

Although motivated by Urdu, URDUBENCH makes no language-specific assumptions in its core. Tokenization, script handling, and evaluation criteria are fully delegated to task definitions, allowing straightforward extension to other languages.

### 4.3 MMR-Guided Few-Shot Example Selection

To enhance few-shot performance, URDUBENCH uses Maximum Marginal Relevance (MMR) for example selection. MMR balances relevance to the test instance and diversity among selected examples, reducing redundancy that is common in random or similarity-based sampling — especially in low-resource languages like Urdu. A balanced candidate pool is first created using stratified sampling. Multilingual embeddings are then used with an MMR selector ($\lambda = 0.5$) to iteratively retrieve high-quality examples. Selected shots are cached for efficiency. This approach provides more representative demonstrations and improves performance on fine-grained tasks compared to naive baselines.

## 5 Experimental Setup

We conduct a large scale benchmarking study using URDUBENCH to evaluate open weight instruction tuned LLMs across diverse Urdu NLP tasks under zero-shot and five shot settings. All experiments were performed on NVIDIA L40 and A100 GPUs using deterministic inference.

### 5.1 Dataset Collection and Task Coverage

URDUBENCH comprises 23 datasets spanning a wide range of Urdu NLP tasks, with strong emphasis on ethical and safety-critical applications relevant to real-world Urdu online platforms. We selected these datasets based on high annotation quality, preference for native Urdu content, broad task coverage, and prior research adoption. Alternatives with small sample sizes, noisy labels, or poor documentation were excluded. Figure 2 provides an overview of the task categories, datasets, and evaluated models grouped by parameter scale. The Datasets sizes are reported in Table 1.

Table 1: Datasets sizes used in URDUBENCH.

| Dataset Name | Size |
|---|---|
| PeerJ Fake News | 4097 |
| EmoThreat (Task A) | 9750 |
| Urdu Nastalique Emotions Dataset | 378 |
| Hate Speech Categorization | 8702 |
| CLE-Hate Speech | 5432 |
| ISE-Hate Corpus | 21759 |
| UWQ-22 | 8596 |
| NER MK-PUCIT | 99999 |
| COUNTER | 1198 |
| CLE Urdu POS Tagset | 22522 |
| Aspect-Based Sentiment | 2902 |
| Sentiment Tweets | 10008 |
| CLE Meeting Corpus | 10 |
| CORPURES | 1614 |
| HASOC (Subtask A) | 2400 |
| HASOC (Subtask B) | 9950 |
| SIB200 | 903 |
| Bible Translation | 7657 |
| Quran Translation | 6200 |
| FLORES200 | 1096 |
| Cyberbullying | 12393 |
| Belebele | 899 |
| ProSOUL | 11,574 |

#### 5.1.1 Core Structural and Classification Tasks

Foundational tasks include named entity recognition using the MK-PUCIT dataset (Kanwal et al., 2019), part-of-speech tagging with the CLE Urdu POS Tagset, news categorization via the COUNTER dataset (Sharjeel et al., 2017), intent detection using UWQ-22 (Shams and Aslam, 2022), machine reading comprehension using Belebele (Bandarkar et al., 2024) and topic categorization via SIB200 (Adelani et al., 2024)

#### 5.1.2 Ethical and Safety-Critical Tasks

A key focus of URDUBENCH is ethical and safety-critical evaluation. This category covers hate speech detection using the ISE-Hate Corpus (Akram et al., 2023) and CLE-HateSpeech (Ali et al., 2021), propaganda detection with ProSOUL (Kausar et al., 2020), abusive language and threat detection from HASOC subtasks A and B (Das et al., 2021), cyberbullying iden-

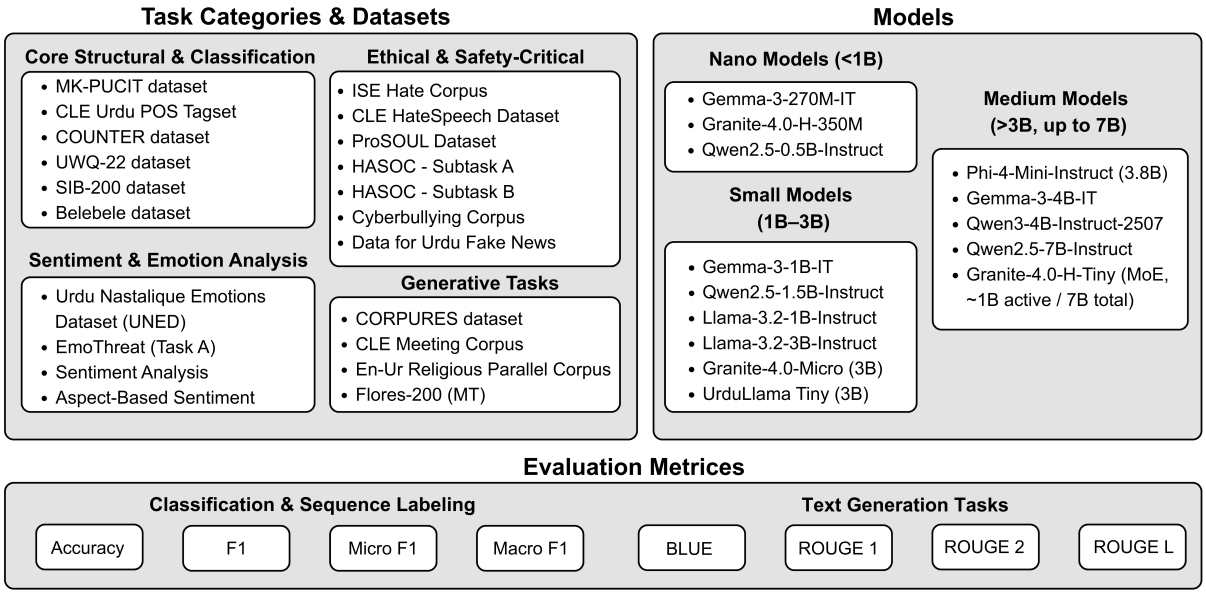


Figure 2: Summary of Task Categories, Datasets, and Model Scales in UrduBench

tification (Adeeba et al., 2024), and fake news detection (Farooq et al., 2023).

#### 5.1.3 Sentiment and Emotion Analysis

Subjective language understanding is evaluated through sentiment analysis datasets supporting both general and aspect-based sentiment (Muhammad and Burney, 2023; ul Haq et al., 2020), alongside emotion classification using the Urdu Nastalique Emotions Dataset (Bashir et al., 2023) and EmoThreat (Task A) (Ashraf et al., 2022).

#### 5.1.4 Generative Tasks

To assess generation in morphologically rich Urdu, UrduBench includes extractive summarization using CORPURES (Humayoun and Akhtar, 2022), abstractive summarization via the CLE Meeting Corpus (Sadia et al., 2024), English–Urdu machine translation based on the English-Urdu Religious Parallel Corpus (Jawaid and Zeman, 2011) and Machine Translation with FLORES200 (NLLB Team, 2022).

### 5.2 Model Categories

UrduBench evaluates a representative set of open-weight, instruction-tuned large language models spanning multiple parameter scales and architectural families. The benchmark includes models from the Qwen series (Qwen2.5 and Qwen3) (Team, 2024, 2025b), Gemma 3 (Team, 2025a), Granite 4.0 (IBM Research, 2025), Phi-4 (kinfey, 2025), Llama 3.2 (Meta AI, 2024), and the domain-adapted UrduLLaMA-1.1-Tiny. Models are grouped into three categories based on parameter count: **nano**(<1B), **small**(1B–3B), and **medium** (>3B, up to 7B).

This model selection was done to assess LLM performance under *low-resource and compute-constrained settings*, which are common in real-world deployments. Nano and small models emphasize efficiency, reduced memory footprint, and deployability on limited hardware, making them suitable for edge devices. Medium models are included to study scaling effects while remaining within a practical upper bound for single or few GPU environments. Additionally, the inclusion of UrduLLaMA-1.1-Tiny, a continued pre-trained model on approximately 800M Urdu tokens built upon Llama 3.2 3B-Instruct, allows direct evaluation of the impact of language-specific adaptation compared to general-purpose multilingual models.

## 6 Results and Discussion

We evaluated 14 instruction-tuned LLMs on 23 tasks using UrduBench under both zero-shot and five-shot settings using Maximum Marginal Relevance (MMR) few-shot selection to ensure diverse and representative in-context examples, following the hardware and experimental setup described in Section 5. The number of shots can be further increased as needed. All evaluations were exe-

| Task | G-270M | Gr-350M | Q-0.5B | G-1B | Q-1.5B | L-1B | L-3B | Gr-3B | L-Tiny | P-4B | G-4B | Q-7B | Q3-4B | Gr-T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Core Classification Tasks** | | | | | | | | | | | | | | |
| CLE Urdu Hate Speech | 0.438 | 0.425 | 0.494 | 0.456 | 0.544 | 0.496 | 0.533 | 0.277 | 0.403 | 0.556 | **0.616** | 0.590 | 0.585 | 0.539 |
| UWQ-22 Intent | 0.293 | 0.299 | 0.308 | 0.320 | 0.196 | 0.292 | 0.329 | 0.209 | 0.276 | 0.377 | **0.432** | 0.404 | 0.328 | 0.401 |
| NER MK-PUCIT | N/A | 0.254 | 0.262 | 0.337 | 0.274 | 0.079 | 0.198 | 0.006 | 0.178 | 0.414 | **0.591** | 0.552 | 0.586 | 0.503 |
| POS Tagging | 0.015 | 0.037 | 0.023 | 0.025 | 0.101 | 0.013 | 0.107 | 0.005 | N/A | 0.089 | 0.131 | **0.222** | 0.187 | 0.196 |
| News Categorization | 0.106 | 0.126 | 0.107 | 0.552 | 0.535 | 0.426 | 0.771 | 0.125 | 0.507 | 0.719 | 0.789 | **0.816** | 0.801 | 0.748 |
| Extractive Summarization | 0.393 | 0.507 | 0.480 | **0.522** | 0.489 | 0.312 | 0.397 | 0.425 | 0.368 | 0.417 | 0.374 | 0.313 | 0.357 | 0.287 |
| Topic Categorization | 0.030 | 0.250 | 0.306 | 0.574 | 0.480 | 0.229 | 0.483 | **0.775** | 0.388 | 0.697 | 0.765 | 0.768 | 0.774 | 0.694 |
| Machine Reading Comprehension | 0.101 | 0.188 | 0.202 | 0.332 | 0.311 | 0.169 | 0.314 | 0.530 | 0.318 | 0.399 | 0.557 | 0.580 | **0.695** | N/A |
| **Ethical and Safety-Critical Tasks** | | | | | | | | | | | | | | |
| CLE Hate Speech Categorization | 0.227 | 0.192 | 0.214 | 0.156 | 0.125 | 0.092 | 0.126 | 0.132 | N/A | 0.197 | 0.000 | **0.338** | 0.246 | 0.120 |
| ISE Hate | 0.378 | 0.530 | 0.491 | 0.382 | 0.332 | 0.461 | 0.534 | 0.448 | 0.494 | 0.670 | **0.716** | 0.639 | **0.716** | 0.653 |
| Abusive Language | 0.336 | 0.507 | 0.422 | 0.540 | 0.337 | 0.490 | 0.377 | 0.374 | 0.448 | 0.546 | 0.624 | **0.660** | 0.638 | 0.531 |
| Cyberbullying | 0.067 | 0.120 | 0.131 | 0.055 | 0.079 | 0.062 | 0.148 | 0.000 | 0.030 | 0.196 | **0.242** | 0.191 | 0.188 | 0.214 |
| Propaganda Detection | 0.542 | 0.491 | 0.513 | 0.472 | 0.317 | 0.479 | 0.628 | 0.486 | 0.507 | 0.551 | 0.579 | **0.717** | 0.618 | 0.288 |
| Threat Detection | 0.486 | 0.506 | 0.471 | 0.149 | 0.258 | 0.520 | 0.402 | 0.246 | 0.172 | 0.558 | 0.455 | **0.583** | 0.506 | 0.545 |
| Fake News | 0.367 | 0.370 | 0.422 | 0.531 | 0.517 | 0.550 | 0.499 | 0.542 | 0.429 | 0.303 | 0.671 | 0.565 | **0.730** | 0.415 |
| **Sentiment and Emotion Analysis** | | | | | | | | | | | | | | |
| Sentiment Analysis | 0.370 | 0.245 | 0.250 | 0.545 | 0.367 | 0.180 | 0.297 | 0.258 | 0.326 | 0.455 | 0.635 | 0.624 | **0.651** | 0.533 |
| Emotion Analysis Task A | 0.293 | 0.171 | 0.293 | 0.277 | 0.267 | 0.293 | 0.242 | 0.174 | 0.167 | **0.342** | 0.341 | 0.289 | 0.339 | 0.334 |
| Paragraph Emotion | 0.122 | 0.065 | 0.082 | 0.251 | 0.299 | 0.141 | 0.341 | 0.167 | N/A | 0.332 | 0.158 | **0.393** | 0.376 | 0.361 |
| Aspect-Based Sentiment | 0.104 | 0.128 | 0.170 | 0.183 | 0.146 | 0.064 | 0.192 | 0.078 | 0.170 | 0.147 | 0.267 | **0.286** | 0.246 | 0.219 |
| **Generative Tasks** | | | | | | | | | | | | | | |
| Abstractive Summarization | 0.032 | 0.000 | 0.033 | 0.000 | 0.064 | 0.140 | 0.034 | 0.067 | 0.000 | 0.000 | 0.000 | 0.000 | **0.150** | 0.000 |
| Bible MT | 1.732 | 0.031 | 0.065 | 1.532 | 0.872 | 1.532 | 4.202 | 5.949 | **6.501** | 0.872 | 0.575 | 1.910 | 6.097 | **6.501** |
| Quran MT | 0.579 | 0.000 | 0.000 | 0.044 | 0.258 | 0.044 | 0.323 | **2.710** | 1.547 | 0.338 | 0.000 | 0.156 | 2.278 | 1.547 |
| Machine Translation | 0.243 | 0.246 | 0.624 | 3.218 | 0.907 | 1.657 | 7.190 | 6.013 | **10.028** | 4.785 | 9.925 | 2.697 | 4.872 | 8.670 |

Table 2: URDUBENCH evaluation results (zero-shot) across 23 tasks and 14 models. Model abbreviations: G-270M = gemma 3 270M, G-1B = gemma 3 1B, G-4B = gemma 3 4B, Q-0.5B = Qwen 2.5 0.5B, Q-1.5B = Qwen 2.5 1.5B, Q-7B = Qwen 2.5 7B, Q3-4B = Qwen 3 4B, L-1B = Llama 3.2 1B, L-3B = Llama 3.2 3B, L-Tiny = UrduLLaMA-1.1-Tiny, P-4B = Phi-4 mini, Gr-350M = granite 4.0 350M, Gr-3B = granite 4.0 micro 3B, Gr-T = granite 4.0 h-tiny 7B.

cuted through the unified framework, leveraging fault-tolerant inference, partial-run capability, and reproducible execution semantics. Table 2 and Table 3 reports the macro F1-score for core structural and classification tasks, ethical and safety-critical tasks, sentiment and emotion analysis tasks, and the BLEU score for generative tasks under the zero-shot and five-shot settings respectively.

## 6.1 Zero Shot Evaluation

### 6.1.1 Core Structural and Classification Tasks

In core structural and classification tasks, medium models substantially outperform nano and small models. This pronounced scaling effect is particularly evident in news categorization, where medium models reach up to 0.816 (Qwen-7B), while nano models on average hover around 0.11, suggesting that larger models better capture the lexical and contextual diversity required for accurate Urdu text classification. Named entity recognition and intent classification also follow this trend, with medium models consistently exceeding 0.5 macro-f1, whereas nano and most small models struggle below 0.35 due to insufficient capacity for handling Urdu's rich morphology and script-specific tokenization challenges.

### 6.1.2 Ethical and Safety-Critical Tasks

Small models perform relatively well on ethical and safety-critical tasks compared to their weaker results on core structural and classification tasks, with medium models only slightly better. Results show that binary signals like hate, abuse, or propaganda transfer effectively from multilingual pretraining even to nano models. Medium models still excel on nuanced subtasks (Qwen2.5-7B-Instruct 0.660 Abusive Language; gemma-3-4b-it 0.716 ISE Hate), but nano models (granite-4.0-350m, Qwen2.5-0.5B-Instruct) often exceed 0.53–0.54 on binary detection, despite near-random performance on core structural and classification tasks. Fine-grained challenges remain across scales (Cyberbullying <0.242; CLE Hate Speech Categorization <0.338), showing socio-cultural subtleties in Urdu toxicity.

### 6.1.3 Sentiment and Emotion Analysis

In Sentiment and Emotion Analysis, medium models lead with 0.651, small with 0.545 and nano with 0.370 macro-F1, showing a wider perfor-

| Task | G-270M | Gr-350M | Q-0.5B | G-1B | Q-1.5B | L-1B | L-3B | Gr-3B | L-Tiny | P-4B | G-4B | Q-7B | Q3-4B | Gr-T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Core Classification Tasks** | | | | | | | | | | | | | | |
| CLE Urdu Hate Speech | 0.477 | 0.556 | 0.411 | 0.578 | 0.581 | 0.470 | 0.605 | 0.699 | 0.470 | 0.683 | 0.728 | 0.611 | **0.734** | 0.647 |
| UWQ-22 Intent | 0.289 | 0.338 | 0.285 | 0.298 | 0.353 | 0.258 | 0.198 | 0.439 | 0.272 | 0.406 | 0.478 | 0.466 | **0.490** | 0.441 |
| NER MK-PUCIT | 0.027 | 0.186 | 0.063 | 0.092 | 0.195 | 0.077 | 0.123 | **0.232** | 0.096 | 0.141 | 0.145 | 0.169 | 0.103 | 0.133 |
| POS Tagging | 0.008 | 0.060 | 0.037 | 0.035 | 0.082 | 0.055 | 0.008 | 0.103 | 0.054 | 0.104 | 0.162 | 0.155 | **0.176** | 0.145 |
| News Categorization | 0.233 | 0.233 | 0.094 | 0.372 | 0.092 | 0.378 | 0.314 | 0.740 | 0.285 | 0.638 | 0.601 | **0.748** | 0.673 | N/A |
| Extractive Summarization | 0.398 | 0.391 | 0.439 | 0.512 | 0.545 | 0.370 | 0.361 | 0.503 | 0.440 | 0.407 | **0.558** | 0.451 | 0.448 | 0.473 |
| Topic Categorization | 0.496 | 0.397 | 0.147 | 0.591 | 0.645 | 0.478 | 0.700 | 0.785 | 0.545 | 0.755 | 0.743 | 0.781 | **0.841** | 0.747 |
| Machine Reading Comprehension | 0.221 | 0.210 | 0.203 | 0.236 | 0.319 | 0.185 | 0.333 | 0.535 | 0.273 | 0.447 | 0.556 | 0.585 | **0.666** | N/A |
| **Ethical and Safety-Critical Tasks** | | | | | | | | | | | | | | |
| CLE Hate Speech Categorization | 0.253 | 0.231 | 0.238 | 0.289 | 0.213 | 0.286 | 0.072 | 0.387 | 0.197 | 0.340 | 0.444 | **0.495** | 0.373 | 0.423 |
| ISE Hate | 0.377 | 0.531 | 0.526 | 0.561 | 0.516 | 0.377 | 0.296 | 0.622 | 0.459 | 0.481 | **0.712** | 0.703 | 0.668 | N/A |
| Abusive Language | 0.511 | 0.459 | 0.608 | 0.442 | 0.448 | 0.624 | 0.591 | 0.611 | 0.433 | 0.502 | **0.653** | 0.616 | 0.612 | 0.638 |
| Cyberbullying | 0.133 | 0.137 | 0.123 | 0.172 | 0.193 | 0.076 | 0.180 | 0.177 | 0.119 | 0.188 | 0.234 | **0.255** | 0.244 | 0.249 |
| Propaganda Detection | 0.355 | 0.437 | 0.565 | **0.623** | 0.373 | 0.481 | 0.398 | 0.595 | 0.534 | 0.437 | 0.408 | 0.596 | 0.475 | 0.429 |
| Threat Detection | 0.453 | 0.479 | 0.485 | 0.316 | 0.223 | 0.503 | 0.224 | 0.462 | 0.396 | 0.349 | 0.379 | **0.540** | 0.401 | 0.430 |
| Fake News | 0.436 | 0.495 | 0.491 | 0.483 | 0.394 | 0.409 | 0.482 | **0.512** | 0.501 | 0.435 | 0.459 | 0.442 | 0.481 | 0.488 |
| **Sentiment and Emotion Analysis** | | | | | | | | | | | | | | |
| Sentiment Analysis | 0.264 | 0.305 | 0.339 | 0.417 | 0.461 | 0.364 | 0.203 | 0.589 | 0.352 | 0.506 | 0.601 | 0.579 | **0.626** | 0.582 |
| Emotion Analysis Task A | 0.117 | 0.252 | 0.223 | 0.325 | 0.260 | 0.243 | 0.255 | 0.313 | 0.334 | 0.337 | 0.388 | 0.322 | 0.307 | **0.432** |
| Paragraph Emotion | 0.032 | 0.127 | 0.149 | 0.203 | 0.181 | 0.268 | 0.095 | 0.232 | 0.199 | 0.252 | 0.253 | **0.291** | 0.282 | 0.286 |
| Aspect-Based Sentiment | 0.258 | 0.081 | 0.214 | 0.224 | 0.255 | 0.164 | 0.103 | 0.278 | 0.097 | 0.267 | 0.238 | **0.328** | 0.278 | N/A |
| **Generative Tasks** | | | | | | | | | | | | | | |
| Abstractive Summarization | 0.000 | 0.000 | 0.005 | 0.000 | 0.010 | 0.008 | 0.008 | 0.000 | 0.012 | **0.050** | 0.030 | 0.005 | 0.014 | 0.001 |
| Bible MT | 0.977 | 0.585 | 0.802 | 2.324 | 1.035 | 0.938 | 1.860 | 4.678 | 6.895 | 4.102 | **9.674** | 2.955 | 7.801 | 6.507 |
| Quran MT | 0.667 | 0.862 | 0.762 | 2.305 | 2.196 | 0.833 | 2.566 | 5.371 | 5.687 | 3.417 | **8.071** | 3.349 | 6.110 | 6.356 |
| Machine Translation | 0.708 | 0.402 | 0.646 | 2.509 | 1.138 | 1.671 | 5.417 | 5.556 | 6.795 | 4.278 | **9.443** | 3.067 | 5.767 | 8.169 |

Table 3: UrduBench evaluation results (five-shot) across 23 tasks and 14 models. Model abbreviations: G-270M = gemma 3 270M, G-1B = gemma 3 1B, G-4B = gemma 3 4B, Q-0.5B = Qwen 2.5 0.5B, Q-1.5B = Qwen 2.5 1.5B, Q-7B = Qwen 2.5 7B, Q3-4B = Qwen 3 4B, L-1B = Llama 3.2 1B, L-3B = Llama 3.2 3B, L-Tiny = UrduLLaMA-1.1-Tiny, P-4B = Phi-4 mini, Gr-350M = granite 4.0 350M, Gr-3B = granite 4.0 micro 3B, Gr-T = granite 4.0 h-tiny 7B.

mance gap than in ethical and safety-critical tasks. Small models struggle as these tasks require deeper contextual understanding of subjective and culturally nuanced language. Medium models excel on sentiment (Qwen3-4B-Instruct-2507 0.651) and paragraph-level emotion (Qwen2.5-7B-Instruct 0.393), while aspect-based sentiment remains challenging across all sizes (<0.29). Emotion classification shows moderate scale gains (Phi-4-mini instruct and medium Qwen/gemma 0.34–0.39), but nano/small models hover near 0.17–0.27, indicating limited transfer from multilingual pretraining.

#### 6.1.4 Generative Tasks

In Generative Tasks, all models perform poorly, with medium-sized models achieving only marginally higher BLEU scores than smaller counterparts, highlighting a stark contrast with the comprehension oriented tasks elsewhere in the benchmark. While multilingual pretraining enables reasonable understanding and classification of Urdu text even in sub billion parameter models, the ability to produce coherent, contextually appropriate Urdu output remains severely limited across the board. Translation and summarization demand not only linguistic knowledge but also faithful stylistic preservation and cultural nuance, capabilities that current instruction-tuned models lack due to insufficient exposure to high-quality Urdu generative data during training.

### 6.2 Five-Shot Evaluation

#### 6.2.1 Core Structural and Classification Tasks

Core structural and classification tasks exhibit a clear scaling trend, with medium-scale models significantly outperforming nano and small models across most tasks. The highest performance is seen in Topic Categorization (0.841 with Qwen3-4B). In contrast, sequence labeling remains challenging, with POS Tagging recording the lowest scores (best 0.176) and NER staying below 0.232 even for stronger models. Machine Reading Comprehension and UWQ-22 Intent show moderate gains (highest 0.666 and 0.490), and Extractive Summarization achieves a modest best score of 0.558.

#### 6.2.2 Ethical and Safety-Critical Tasks

Ethical and safety critical tasks demonstrate more robust performance across different model sizes, especially on binary detection subtasks. Medium

model (Gemma 3 4B) achieved strong results on ISE Hate and Abusive Language (0.712 and 0.653). However, CLE Hate Speech Categorization remains challenging, with nano models scoring below 0.29, while some medium models exceed 0.49. Similarly, in Cyberbullying Detection, all models perform poorly, with all scores remaining below 0.26. In contrast, in Propaganda Detection, Gemma 3 1B achieved the highest macro F1 score of 0.623 across all model categories.

#### 6.2.3 Sentiment and Emotion Analysis

These tasks reveal the largest performance gap across model scales. Gemma 3 1B model performed better than most of the small models in the Sentment analysis task but Qwen3 4B performed the best among all the models (0.626). Paragraph-level Emotion and Aspect-Based Sentiment prove especially difficult, with even medium models scoring below 0.33. Emotion classification shows moderate scaling effects. These findings indicate that fine-grained Urdu sentiment and emotion analysis remains limited by model capacity and cultural nuance, even with MMR-enhanced few-shot prompting.

#### 6.2.4 Generative Tasks

Generative tasks remain highly challenging even in the five-shot setting. Abstractive Summarization continues to be the most difficult task, with scores close to zero across nearly all models and only marginal improvement (up to 0.05). In contrast, the machine translation tasks perform substantially better: Bible MT reaches up to 9.674, Quran MT up to 8.071, and general Machine Translation up to 9.443 with medium-scale models, clearly outperforming smaller ones.

## 7 Conclusion

In this paper, we introduce URDUBENCH, the first comprehensive benchmark for evaluating large language models on native Urdu tasks. It includes 23 datasets across 19 tasks covering natural language understanding, safety-critical moderation, sentiment and emotion analysis, and generative capabilities, addressing challenges in a low-resource setting. We accompany the benchmark with an open source, modular evaluation framework that ensures reproducible, few shot assessments through uniform prompting, post-processing, and metric computation, while facilitating seamless extension to new models and datasets. We conduct a large-scale evaluation of 14 open weight instruction-tuned LLMs across nano (<1B), small (1–3B), and medium (up to 7B) scales under zero-shot and five-shot settings. Results (macro F1 and BLEU) show clear performance gaps: larger multilingual models such as Qwen and Gemma outperform smaller ones, while challenges persist in sequence labeling, fine-grained toxicity detection, and long-context generation. Sub-billion models perform reasonably on binary safety tasks, indicating potential for lightweight moderation, though stronger linguistic and generative abilities require larger models. Five-shot results improve over zero-shot on several tasks, though gains vary. By releasing URDUBENCH, we aim to support systematic progress in Urdu LLMs, highlight gaps in low-resource settings, and provide a scalable benchmark extensible to other underrepresented languages. Future work could explore few-shot and proprietary models to further advance equitable AI for Urdu speakers.

## Limitations

URDUBENCH is evaluated exclusively in zero-shot and five-shot settings to ensure consistent and reproducible comparisons across heterogeneous models. While the framework supports higher-shot prompting and fine-tuning, these regimes are not explored in this work and may provide further insights into model performance on Urdu tasks. Although the evaluation framework is language-agnostic, the current benchmark and experiments are instantiated exclusively for Urdu. The evaluation is further restricted to open-sourced, lightweight, and mid-sized models (up to 7B parameters), reflecting practical deployment considerations in resource-constrained environments. Due to budget constraints and the limited accessibility of closed-source, SOTA LLMs, we do not include such models in our evaluation. Nevertheless, the proposed framework is designed to be extensible, and it supports evaluation through external APIs, enabling future incorporation of larger-scale and closed-source models.

## References

Farah Adeeba, Muhammad Irfan Yousuf, Izza Anwer, Sardar Umair Tariq, Abdullah Ashfaq, and Malik Naqeeb. 2024. Addressing cyberbullying in urdu tweets: a comprehensive dataset and detection system. *PeerJ Comput. Sci.*, 10:e1963.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen,

Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *Preprint*, arXiv:2309.07445.

Muhammad Akram, Khurram Shahzad, and Maryam Bashir. 2023. Ise-hate: A benchmark corpus for interfaith, sectarian, and ethnic hatred detection on social media in urdu. *Information Processing & Management*, 60.

Muhammad Ali, Ehsan Ul Haq, Sahar Rauf, Kashif Javed, and Sarmad Hussain. 2021. Improving hate speech detection of urdu tweets using sentiment analysis. *IEEE Access*.

Noman Ashraf, lal khan, Sabur Butt, Hsien-Tsung Chang, Grigori Sidorov, and Alexander Gelbukh. 2022. Multi-label emotion classification of urdu tweets. *PeerJ Computer Science*, 8:e896.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 749–775. Association for Computational Linguistics.

Muhammad Farrukh Bashir, Abdul Rehman Javed, Muhammad Umair Arshad, Thippa Reddy Gadekallu, Waseem Shahzad, and Mirza Omer Beg. 2023. Context-aware emotion detection from low-resource urdu language using deep neural network. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(5).

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. Llmebench: A flexible framework for accelerating llms benchmarking.

Mithun Das, Somnath Banerjee, and Punyajoy Saha. 2021. Abusive and threatening language detection in urdu using boosting based and BERT based models: A comparative approach. *CoRR*, abs/2111.14830.

Muhammad Shoaib Farooq, Ansar Naseem, Furqan Rustam, and Imran Ashraf. 2023. Fake news detection in urdu language using machine learning. *PeerJ Computer Science*, 9:e1353. Https://doi.org/10.7717/peerj-cs.1353.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. A framework for few-shot language model evaluation.

Abdullah Hashmat, Muhammad Arham Mirza, and Agha Ali Raza. 2025. Pakbbq: A culturally adapted bias benchmark for qa.

Muhammad Humayoun and Naheed Akhtar. 2022. Corpures: Benchmark corpus for urdu extractive summaries and experiments using supervised learning. *Intelligent Systems with Applications*, 16:200129.

IBM Research. 2025. Granite 4.0 language models. `https://github.com/ibm-granite/granite-4.0-language-models`. Accessed: 2025-10-01.

Bushra Jawaid and Daniel Zeman. 2011. Word-order issues in english-to-urdu statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95.

Safia Kanwal, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. 2019. Urdu named entity recognition: Corpus generation and deep learning applications.

Soufia Kausar, Bilal Tahir, and Amir Mehmood. 2020. Prosoul: A framework to identify propaganda from online urdu content.

kinfey. 2025. Welcome to the new phi-4 models - microsoft phi-4-mini phi-4-multimodal. Microsoft Tech Community Blog, Accessed: December 31, 2025.

Adithya Sreedharan Kumara and CognitiveLab Team. 2024. Indic LLM leaderboard and IndicEval suite. `https://huggingface.co/spaces/Cognitive-Lab/indic_llm_leaderboard`. Accessed: 2025-12-10.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. SeaExam and SeaBench: Benchmarking LLMs with local multilingual questions in Southeast Asia. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6119–6136, Albuquerque, New Mexico. Association for Computational Linguistics.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Accessed: December 31, 2025.

Khalid Bin Muhammad and S. M. Aqil Burney. 2023. Innovations in urdu sentiment analysis using machine and deep learning techniques for two-class classification of symmetric datasets. *Symmetry*, 15(5).

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Framing political bias in multilingual llms across pakistani languages.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik

Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

Bareera Sadia, Farah Adeeba, Sana Shams, and Kashif Javed. 2024. Meeting the challenge: A benchmark corpus for automated urdu meeting summarization. *Information Processing & Management*, 61(4):103734.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Sana Shams and Muhammad Aslam. 2022. Improving user intent detection in urdu web queries with capsule net architectures. *Applied Sciences*, 12:11861.

Muhammad Sharjeel, Rao Nawab, and Paul Rayson. 2017. Counter: corpus of urdu news text reuse. *Language Resources and Evaluation*, 51.

SIL International. 2022. Urdu. https://www.ethnologue.com/language/urd. Ethnologue: Languages of the World (25th edition, accessed 2025).

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages.

Munief Hassan Tahir, Sana Shams, Layba Fiaz, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking the performance of pre-trained LLMs across Urdu NLP tasks. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 17–34, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ehsan ul Haq, Sahar Rauf, Sarmad Hussain, and Kashif Javed. 2020. Corpus of aspect-based sentiment for urdu political data.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, and 1 others. 2024. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 4.

Tiankai Yang, Yi Nian, Li Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan A. Rossi, Kaize Ding, Xia Hu, and Yue Zhao. 2025. AD-LLM: Benchmarking large language models for anomaly detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1524–1547, Vienna, Austria. Association for Computational Linguistics.

José Ángel González, Ian Borrego Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Ríos, Angelo Basile, and Marc Franco-Salvador. 2025. Iberbench: Llm evaluation on iberian languages. *Preprint*, arXiv:2504.16921.