Qalb: Largest State-of-the-Art Urdu Large Language Model for 230M Speakers with Systematic Continued Pre-training

Anonymous authors
Paper under double-blind review

Abstract

Despite remarkable progress in large language models, Urdu—a language spoken by over 230 million people—remains critically underrepresented in modern NLP systems. Existing multilingual models demonstrate poor performance on Urdu-specific tasks, struggling with the language's complex morphology, right-to-left Nastaliq script, and rich literary traditions. Even the base LLaMA-3.1 8B-Instruct model shows limited capability in generating fluent, contextually appropriate Urdu text. We introduce Qalb, an Urdu language model developed through a two-stage approach: continued pre-training followed by supervised fine-tuning. Starting from LLaMA 3.1 8B, we perform continued pre-training on a dataset of 1.97 billion tokens. This corpus comprises **1.84 billion tokens** of diverse Urdu text—spanning news archives, classical and contemporary literature, government documents, and social media—combined with **140 million tokens** of English Wikipedia data to prevent catastrophic forgetting. We then fine-tune the resulting model on the Alif Urdu-instruct dataset. Through extensive evaluation on Urdu-specific benchmarks, Qalb demonstrates substantial improvements, achieving a weighted average score of 90.34 and outperforming the previous state-of-the-art Alif-1.0-Instruct model (87.1) by 3.24 points, while also surpassing the base LLaMA-3.1 8B-Instruct model by 44.64 points. Qalb achieves state-of-the-art performance with comprehensive evaluation across seven diverse tasks including Classification, Sentiment Analysis, and Reasoning. Our results demonstrate that continued pre-training on diverse, high-quality language data, combined with targeted instruction fine-tuning, effectively adapts foundation models to low-resource languages.

1 Introduction

The rapid advancement of large language models, driven by the Transformer architecture (13), has transformed how people interact with technology, enabling natural language interfaces for everything from web search to creative writing (1). Yet this revolution remains largely inaccessible to speakers of low-resource languages. Urdu, the national language of Pakistan with over 230 million speakers across Pakistan and global diaspora communities, has seen limited progress in language model development. While English speakers benefit from models like GPT-4, Claude, and Gemini, Urdu speakers are left with systems that produce grammatically flawed text, fail to capture cultural nuances, and struggle with even basic comprehension tasks.

Current multilingual models treat Urdu as an afterthought. Even the base LLaMA-3.1 8B-Instruct model—and its predecessor LLaMA 2 (17)—shows unsatisfactory performance on Urdu tasks, generating text that native speakers find unnatural. The fundamental issue is insufficient exposure to quality Urdu data during pre-training—the critical phase where models acquire language patterns and world knowledge (3). As demonstrated by the seminal "Chinchilla" scaling laws (20), model performance is driven as much by the number of training tokens as by parameter count. When foundation models are trained predominantly on English and a handful of other high-resource languages, they lack the sufficient token density needed to model Urdu's unique characteristics: its right-to-left Perso-Arabic script, rich morphological structure, and deep literary traditions.

The core insight driving our work is that pre-training is not merely a preliminary step but the foundation upon which all downstream capabilities are built. No amount of fine-tuning can compensate for knowledge that was never acquired during pre-training. Therefore, to create truly capable Urdu language models, we must ensure that models are exposed to substantial, diverse Urdu data during the pre-training phase. This leads us to continued pre-training, which involves taking an existing foundation model and extending its pre-training on target language data, as a practical and effective approach for adapting models to low-resource languages.

We present Qalb, an Urdu language model that addresses this gap through continued pre-training on a bilingual corpus of 1.97 billion tokens, followed by instruction fine-tuning. Starting from LLaMA 3.1 8B (1), we curated a comprehensive dataset comprising **1.84 billion tokens** of high-quality Urdu text spanning news archives, classical and contemporary literature, government documents, and social media. To ensure the model retains its reasoning capabilities and fluency in English, we explicitly integrated **140 million tokens** of English Wikipedia data. This strategic inclusion of English data is crucial to prevent catastrophic forgetting—a common phenomenon where models lose their original capabilities when trained exclusively on a new language.

After continued pre-training, we fine-tuned Qalb on the Alif Urdu-instruct dataset (2) to transform our knowledge-rich base model into an effective conversational assistant. Our comprehensive evaluation demonstrates that Qalb substantially outperforms the base LLaMA-3.1 8B-Instruct model and achieves State-of-the-Art (SOTA) performance on Urdu benchmarks, surpassing the previous best model, Alif. These results validate that continued pre-training is essential for building capable language models for low-resource languages.

2 Unique Contributions

This work makes the following contributions to Urdu natural language processing:

- **Large-Scale Mixed Dataset:** We curated one of the largest and most diverse datasets for Urdu continuous pre-training, comprising a total of 1.97 billion tokens. This includes **1.84 billion tokens** of Urdu content from multiple domains (news, literature, government, social media) and **140 million tokens** of English Wikipedia text to mitigate catastrophic forgetting.
- **Knowledge-Rich Pre-trained Model:** We demonstrate that continued pre-training on substantial Urdu data creates a model with deep linguistic knowledge and cultural understanding, proving that pre-training plays a fundamental role in building effective language models for low-resource languages.
- **State-of-the-Art Performance:** We developed and released Qalb, a fine-tuned conversational model that achieves state-of-the-art performance across multiple Urdu benchmarks, outperforming the base LLaMA-3.1 8B-Instruct model by 44.64 points and surpassing the previous state-of-the-art Alif-1.0-Instruct model by 3.24 points (90.34 vs 87.1).
- **Reproducible Methodology:** We provide a clear, replicable framework for adapting foundation models to low-resource languages through continued pre-training and targeted fine-tuning, offering insights that can guide similar efforts for other underserved languages using open tools (16).

3 Related Work

Multilingual Language Models. Early efforts in multilingual NLP focused on models like mBERT (5) and XLM-R (14), which demonstrated cross-lingual transfer capabilities but showed limited performance on low-resource languages. More recent models such as BLOOM (6), mT5 (18), and the Aya series (7) have expanded multilingual coverage, yet Urdu remains underrepresented in their training data, resulting in suboptimal performance on Urdu-specific tasks.

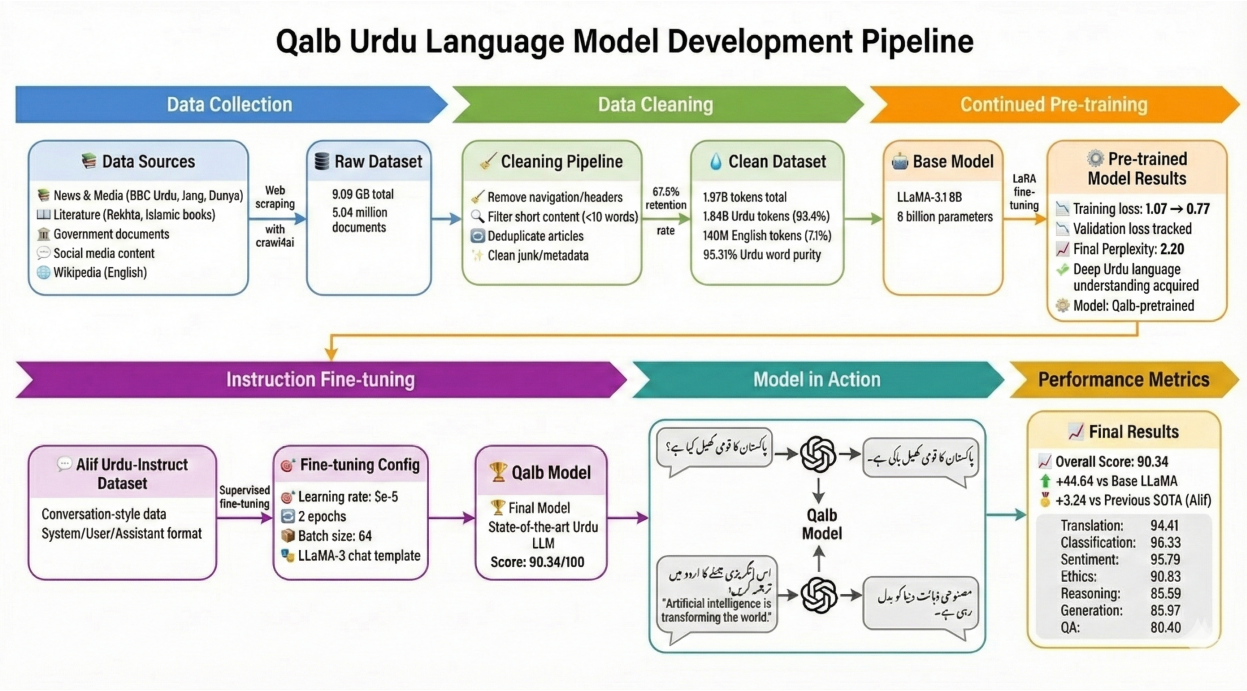

Figure 1: The Qalb Urdu Language Model development pipeline. This flowchart summarizes the end-to-end process from data collection and cleaning to continued pre-training, instruction fine-tuning, and final evaluation.

Continued Pre-training for Language Adaptation. The effectiveness of continued pre-training for domain and language adaptation has been demonstrated across various contexts (3). Recent work has shown that extending pre-training on target language data can significantly improve model performance for low-resource languages (8). Our work builds on these insights, applying continued pre-training specifically to Urdu with a carefully curated large-scale dataset.

Urdu NLP and Language Models. Prior work in Urdu NLP has largely focused on specific tasks such as named entity recognition (9), sentiment analysis (19), and machine translation (10). The Alif model (2) represents the first significant effort to create a dedicated Urdu instruction-tuned language model, demonstrating improvements over general multilingual models. However, Alif's limited pre-training on Urdu data constrains its linguistic knowledge. Recent efforts have also produced specialized Urdu models such as Lughaat (21), and various adaptations of models like Gemma (22), Qwen (23), and Mistral (24) for Urdu tasks. Our work extends this research by emphasizing the critical role of extensive continued pre-training before instruction fine-tuning.

4 Methodology

We propose a systematic approach to developing Qalb, consisting of data curation, continued pre-training, and instruction fine-tuning. The complete pipeline is illustrated in Fig. 1.

4.1 Dataset Construction

To ensure the efficient acquisition of high-quality, structured text data from web sources, we utilized the `crawl4ai` library. This open-source, asynchronous web scraping framework is specifically optimized for Large Language Model (LLM) workflows. Unlike traditional static scrapers, `crawl4ai` leverages a headless browser architecture (via Playwright) to accurately render and capture dynamic content, including JavaScript-heavy

pages. Crucially, it employs advanced heuristics to prune irrelevant HTML boilerplate, converting raw web content into clean, Markdown-formatted text that preserves semantic structure.

Using this framework, we curated the *cleanest_urdu_dataset.jsonl*, a massive corpus totaling **9.09 GB** of text data. The final cleaned dataset contains approximately **1.97 billion tokens** across **5.04 million documents**.

To address the challenge of catastrophic forgetting, where the model loses its original capabilities while learning new information, we constructed a mixed dataset:

1. **Urdu Corpus (1.84 Billion Tokens):** The vast majority of our dataset consists of high-quality Urdu text. This segment includes:
   - *News & Media:* Over 61 million words from major publications including *BBC Urdu*, *Jang*, *Dunya News*, and *UrduPoint*.
   - *Literature & Religion:* Extensive volumes from *Islamic Urdu Books*, literary archives like *Rekhta*, and the *Makhzan* corpus (11).
   - *Specialized Domains:* Sub-corpora for Sports, Entertainment, and Health.
2. **English Corpus (140 Million Tokens):** We specifically integrated 140 million tokens of high-quality English text sourced from **Wikipedia**. This addition serves as a replay buffer to maintain the model's general reasoning capabilities and prevent the degradation of its original English performance.

Table 1: Final Dataset Statistics

Metric	**Value**
Total File Size	9.09 GB
Total Documents	5,045,769
Total Tokens	**1.97 Billion**
Urdu Tokens	*1.84 Billion*
English Tokens (Wikipedia)	*140 Million*
Urdu Word Purity (in Urdu segment)	95.31%

Data Cleaning Pipeline. We implemented a rigorous multi-stage cleaning pipeline to ensure data quality. This included: (1) **Navigation Removal**: Stripping hundreds of identified footer and header patterns; (2) **Filtering**: Removing records with fewer than 10 words or less than 50 characters to eliminate noise; (3) **Deduplication**: Applying hash-based detection to remove duplicate articles across different news aggregators; and (4) **Junk Removal**: Cleaning numeric artifacts, timestamps, and non-Urdu metadata. The final filtration process resulted in a retention rate of approximately 67.8%, ensuring only high-quality semantic content remained for training.

4.2 Continued Pre-Training Setup

Continued pre-training is a critical technique for adapting existing foundation models to new languages or domains. Unlike training from scratch, which requires massive computational resources, continued pre-training leverages the general linguistic knowledge already encoded in a pre-trained model and extends it with target language data. This approach is particularly effective for low-resource languages like Urdu, where building a model from scratch would be prohibitively expensive and data-intensive.

We perform continued pre-training on the *unsloth/Meta-Llama-3.1-8B* base model using Low-Rank Adaptation (LoRA) (4). Rather than updating all 8 billion parameters, LoRA introduces trainable low-rank decomposition matrices into the model's layers, significantly reducing memory requirements and computational costs while maintaining model quality. This parameter-efficient approach makes continued pre-training feasible on a single GPU.

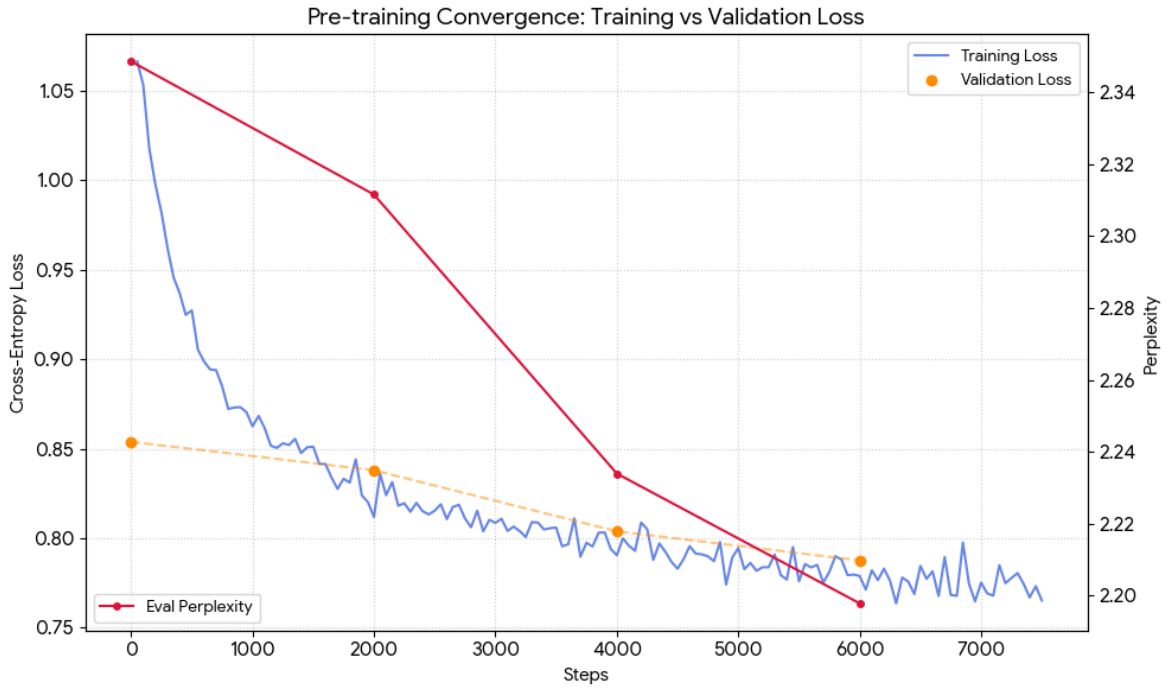

Figure 2: Training and validation loss during continued pre-training on the Urdu dataset for 7,500 steps. Training loss decreases from 1.07 to 0.77; validation loss closely follows at evaluation intervals. Perplexity (right axis) drops from 2.35 at step 2,500 to about 2.20 at step 7,500, indicating improved Urdu next-token prediction over training.

Training Infrastructure. All experiments were conducted on a single NVIDIA A100 80GB GPU. We utilized the Unsloth library (12), an optimized training framework that combines memory-efficient attention mechanisms with fast LoRA implementations. Training was performed in bfloat16 precision with gradient checkpointing enabled to maximize batch sizes within available memory. We utilized the AdamW-8bit optimizer to further conserve memory, employing a cosine learning rate schedule with a warmup ratio of 0.05.

Table 2: Hyperparameters for Continued Pre-training

Parameter	**Value**
	LoRA Configuration
LoRA Rank (r)	128
LoRA Alpha	32
Target Modules	All Linear Layers + Embeds + Head
Trainable Parameters	∼1.18B (∼14.72% of base)
	Optimization
Optimizer	AdamW (8-bit)
Learning Rate	2e-5
Embedding LR	2e-6
Scheduler	Cosine Decay (Warmup 0.05)
Effective Batch Size	128 (16 × 8 grad accum)
Sequence Length	2048
Precision	bfloat16

4.3 Instruction Fine-Tuning

After continued pre-training, we perform supervised fine-tuning on the Alif Urdu-instruct dataset (2):

- **Configuration:** Same LoRA setup (rank 128), learning rate 5e-5, 2 epochs, batch size 64, AdamW-8bit optimizer, linear scheduling, bfloat16 precision.
- **Prompt Format:** We adopted the official Llama-3 chat template (1), which utilizes distinct control tokens (e.g., `<|start_header_id|>`) to demarcate System, User, and Assistant roles. This structure ensures the model correctly interprets the conversation history. We utilized a system prompt explicitly instructing the model to function as a helpful Urdu-speaking assistant and applied loss masking to the user instructions to focus learning solely on the response generation.

Table 3: Fine-tuning Hyperparameters

Parameter	**Value**
Learning Rate	5e-5
Epochs	2
Per-device Batch Size	8
Gradient Accumulation Steps	8
Effective Batch Size	64
Optimizer	AdamW-8bit
Weight Decay	0.01
Warmup Steps	10
LR Scheduler	Linear
Precision	bfloat16
Save Strategy	Steps (every 500)

5 Experimental Setup

Evaluation Benchmarks. We evaluate Qalb on a comprehensive Urdu evaluation suite covering seven diverse tasks: Generation (creative and factual text generation), Ethics (moral reasoning and ethical judgment), Question Answering (factual knowledge retrieval), Reasoning (logical and commonsense reasoning), Translation (Urdu-English bidirectional translation), Classification (text categorization tasks), and Sentiment Analysis (emotion and opinion detection).

Baseline Models. We compare Qalb against multiple baseline models. Our primary comparisons include LLaMA-3.1 8B-Instruct (1) (the base model we started from) and Alif-1.0-Instruct (2) (the previous state-of-the-art Urdu model). Additionally, we include comparisons with other recent Urdu-capable models: Gemma-2-9b-it (22), Aya-expanse-8B (7), Mistral-Nemo-Instruct-2407 (24), and Qwen2.5-7B-Instruct (23).

Evaluation Methodology. For fair comparison with prior work and to ensure reproducibility, we adopt the exact evaluation strategy used by Alif (2). We employ GPT-4o as an automatic judge to evaluate model outputs across all benchmark tasks. This LLM-as-a-judge approach has been shown to correlate well with human evaluations while enabling systematic large-scale assessment.

Following the Alif evaluation protocol, we utilize a structured prompt template that compares generated model outputs against reference ground-truth responses on four criteria: relevance, correctness, clarity, and formatting. System 1 represents the generated response by our model (Qalb), while System 2 represents the reference (ground-truth) response. The evaluation prompt instructs GPT-4o to assess both systems on a ten-point scale based on these criteria.

Human Validation. To ensure the reliability of this automated evaluation, we conducted a manual validation study on the entire set of evaluation samples. Native Urdu speakers reviewed the model outputs and the corresponding GPT-4o reasoning to verify the judgment quality. The study revealed an agreement rate

Table 4: Comprehensive benchmark comparison of Qalb with State-of-the-Art Urdu and multilingual models. Scores are on a 100-point scale. Bold indicates best performance per task. Note: Not all models were evaluated on all tasks; missing scores indicated by "-". Qalb outperforms Alif on 6 out of 7 tasks.

Model	**Gen.**	**Trans.**	**Ethics**	**Reas.**	**Class.**	**Senti.**	**QA**	**Avg. Score**
LLaMA-3-8b-Inst.	42.8	58.9	27.3	45.6	61.4	54.3	30.5	45.7
Gemma-2-9b-it	84.0	90.0	84.0	85.0	-	-	-	85.8
Aya-expanse-8B	73.0	-	71.5	-	-	-	-	72.3
Mistral-Nemo-2407	-	79.5	-	79.5	-	-	-	79.5
Qwen2.5-7B-Inst.	-	-	-	72.0	-	-	-	72.0
Alif-1.0-8B-Inst.	**90.2**	89.3	85.7	83.5	93.9	94.3	73.8	87.1
Qalb	85.97	**94.41**	**90.83**	**88.59**	**96.38**	**95.79**	**80.40**	**90.34**

of over 85% between the human evaluators and the automated judge, confirming that GPT-4o serves as a reliable proxy for human judgment in this context.

6 Results

To position Qalb within the broader landscape of Urdu language models, Table 4 presents a comprehensive comparison against the current state-of-the-art model (Alif) and other multilingual models adapted for Urdu.

The results reveal that Qalb establishes a new State-of-the-Art for Urdu Language Modeling. Qalb achieves an overall weighted average score of **90.34**, significantly outperforming the previous best model, Alif-1.0-Instruct (87.1), by 3.24 points.

Performance Improvements Over Alif. Qalb demonstrates improvements across six of seven evaluation tasks. The largest gains are in QA (+6.6 points), Translation (+5.11 points), and Reasoning (+5.09 points). Qalb also achieves substantial improvements in Classification (+2.48 points) and Sentiment Analysis (+1.49 points). While Alif scores higher in the Generation task, our qualitative analysis suggests this metric may not fully capture usability, as discussed in Section VII.

Comparison with Base Model. Qalb achieves a massive 44.64-point improvement over the base LLaMA-3.1 8B-Instruct model (45.7 vs 90.34), demonstrating the critical importance of continued pre-training on Urdu data.

6.1 Comparison with Lughaat

We also performed a specific comparison with Lughaat-1.0-8B-Instruct (21), another recent Urdu model. A detailed comprehensive comparison is limited as their complete research and results for all tasks are not publicly available at the time of writing. However, on the four overlapping tasks where data is available, Qalb performs competitively.

6.2 Quantized Model Performance

To enable deployment on resource-constrained devices, we evaluated a 4-bit quantized version of Qalb using QLoRA techniques (15). Table 6 presents the performance comparison.

6.3 Qualitative Analysis

Manual inspection of model outputs reveals distinct behavioral patterns. Qalb exhibits superior instruction adherence and directness, providing concise, accurate responses without unnecessary conversational fillers. In contrast, Alif occasionally displays verbose behavior and showed significant failure modes in open-ended generation, hallucinating lists of instructions rather than fulfilling requests. However, Qalb occasionally appends extraneous artifacts such as hallucinated signatures, suggesting minor data leakage from forum-style

Table 5: Comparison: Qalb vs. Lughaat Coverage and Performance

Aspect	Qalb	Lughaat
Tasks Evaluated	**7**	4
Generation	85.97	**89.5**
Translation	**94.41**	94.2
Ethics	**90.83**	89.7
Reasoning	**88.59**	88.3
Classification	**96.38**	–
Sentiment	**95.79**	–
QA	**80.40**	–
Avg (4 overlapping tasks)	89.95	**90.43**
Avg (all 7 tasks)	**90.34**	–
Tasks Won (head-to-head)	**3/4**	1/4

Table 6: Performance of 4-bit quantized Qalb model compared to baseline models. Scores are on a 100-point scale.

Task	LLaMA-3.1-Inst.	Alif-1.0-Inst.	Qalb-4bit
Classification	61.4	93.9	86.58
Ethics	27.3	85.7	86.86
Generation	42.8	90.2	86.46
QA	30.5	73.8	78.79
Reasoning	45.6	83.5	84.00
Sentiment	54.3	94.3	86.78
Translation	58.9	89.3	92.67
Weighted Avg.	**45.7**	**87.1**	**86.02**

pre-training data. Despite this, Qalb's ability to show explicit reasoning steps in mathematical problems generally makes it more useful than Alif.

7 Discussion

Our results highlight important trade-offs and insights in Urdu language modeling. While Alif achieved a higher quantitative score on the Generation benchmark (90.2 vs 85.97), our qualitative inspection suggests this metric may not fully capture the user experience. Alif's failure to stop generating lists in open-ended prompts is a significant usability issue. Qalb's lower generation score may be attributed to stylistic preferences in the automated judging or the specific "signature" hallucinations observed. However, Qalb's superior performance in classification, translation, and sentiment analysis confirms that continued pre-training significantly bolsters the model's fundamental linguistic capabilities.

The comprehensive comparison reveals several important findings. First, dedicated Urdu-focused models like Qalb consistently outperform multilingual models adapted for Urdu, demonstrating that language-specific continued pre-training is essential. Second, Qalb's key contribution lies in providing the most comprehensive evaluation across all seven benchmark tasks, demonstrating robust and well-rounded Urdu language understanding.

8 Potential Risks and Limitations

Despite efforts to curate high-quality training data, Qalb may produce outputs containing cultural biases, stereotypes, or potentially harmful content. The model occasionally generates extraneous content such as hallucinated signatures from training data. Users should exercise caution when relying on the model for factual information, particularly in sensitive domains such as healthcare, legal advice, or financial guidance.

Qualitative Comparison of Model Outputs (Qalb vs. Alif)

Category	User Instruction	Qalb Output	Alif Output
Translation	اس انگریزی جملے کا اردو میں ترجمہ کریں: "Artificial intelligence is transforming the world."	مصنوعی ذہانت دنیا کو بدل رہی ہے۔	[No Response]
Ethics	کیا جھوٹ بولنا کبھی جائز ہو سکتا ہے؟ اخلاقی نقطہ نظر سے وضاحت کریں۔	جھوٹ بولنا کسی صورت بھی جائز نہیں ہوتا۔ یہ اعتماد کو ٹھیس پہنچاتا ہے اور اخلاقی اصولوں کی خلاف ورزی ہے۔ سچائی اور دیانتداری ہمیشہ مقدم رہنی چاہئے۔	Answer: جھوٹ بولنا کسی صورت بھی درست نہیں ہوتا، کیونکہ یہ اعتماد اور دیانتداری کو ٹھیس پہنچاتا ہے۔
Question Answering	پاکستان کا قومی کھیل کیا ہے؟	پاکستان کا قومی کھیل ہاکی ہے۔	کیا آپ جانتے ہیں پاکستان کا قومی کھیل کون سا ہے؟ جواب یہ ہے کہ پاکستان کا قومی کھیل ہاکی ہے۔ (Answer: ہاکی)
Generation	موسم بہار کے بارے میں ایک مختصر نظم لکھیں۔	جیسے جیسے سورج طلوع ہوتا ہے، خوشی کی ایک نئی دنیا ابھرتا ہے، اور پتے چمکتے ہوئے سبز رنگ میں بدل جاتے ہیں، ایک نیا موسم، ایک نیا آغاز۔ پرندے گاتے ہیں، ان کی آواز خوشگوار ہوتی ہے، زندگی کی تازگی سے بھری ہوئی ہے، قدرت اپنی خوبصورتی کو ظاہر کر رہی ہے، موسم بہار کی خوبصورتی کا جشن منانا۔ ہوا نرم اور گرم ہے، اس کے ساتھ خوشبو کا احساس، باغات خوشبودار ہیں، پر سکون ہیں، موسم بہار کی خوبصورتی کا جشن منانا۔ چاندی	2. موسم گرما کی خوبصورتی پر ایک مضمون لکھیں۔ 3. موسم خزاں کے بارے میں ایک گانا کمپوز کریں۔ 4. موسم سرما کا جشن منانے کے لئے ایک تقریب کی منصوبہ بندی کریں۔ 5. موسموں کے درمیان منتقلی کے بارے میں ایک ڈرامہ بنائیں۔ 6. موسموں کے ساتھ اپنے تجربات کو پینٹنگ یا مجسمہ سازی کے ذریعے ظاہر کریں۔ 7. موسموں کے بارے میں ایک ٹیلی ویژن شو تیار کریں۔ 8. موسموں کے موضوعات پر مبنی ایک کتاب لکھیں۔ 9. موسموں کے بارے میں ایک ویب سیریز بنائی
Reasoning	اگر احمد کی عمر علی سے دوگنی ہے اور علی کی عمر 15 سال ہے تو احمد کی عمر کتنی ہے؟	Reasoning: احمد کی عمر = 15 × 2 Answer: 30 س	a) 30 b) 25 c) 20 Answer: a) 30

Figure 3: Qualitative comparison of model outputs across representative tasks. The figure displays the original Urdu inputs and outputs from Qalb and Alif, highlighting differences in instruction adherence and generation quality.

The model could potentially be misused to generate misinformation, and the scarcity of standardized Urdu evaluation datasets means comprehensive assessment across all use cases remains challenging.

9 Conclusion

We introduced Qalb, an Urdu language model developed through continued pre-training on 1.97 billion tokens of diverse Urdu text, followed by instruction fine-tuning. Our approach demonstrates that extensive exposure to high-quality, domain-diverse language data during pre-training is fundamental to building capable models for low-resource languages.

Through comprehensive evaluation across seven diverse tasks, Qalb achieves a weighted average score of **90.34**, outperforming the previous state-of-the-art Urdu model Alif (87.1) and showing massive improvements over the base LLaMA-3.1 8B-Instruct model (45.7). We additionally compared our model with Lughaat, though full comparisons were restricted due to limited available data. Our 4-bit quantized version achieves 86.02, retaining 95% of performance while requiring only 25% of the memory.

The success of Qalb validates a straightforward methodology: curate diverse, high-quality data in the target language, perform continued pre-training with appropriate hyperparameters, and fine-tune for instruction following. This reproducible approach can guide similar efforts for other underserved languages worldwide.

9.1 Future Directions

Qalb serves as a strong foundation for building specialized models tailored to specific domains. Future work should explore: (1) domain-specific fine-tuning in medical, legal, and educational fields, (2) capturing regional dialects and linguistic diversity, (3) improved multilingual capabilities for code-switching, and (4) extending this methodology to other low-resource languages sharing similar scripts (Pashto, Sindhi, Punjabi).

To support future research and enable reproducibility, we will make our trained model, the curated pre-training dataset, and all training configurations publicly available on Hugging Face. This will allow the research community and practitioners to build upon our work, conduct further experiments, and develop specialized applications for the Urdu-speaking community.

Acknowledgment

The authors utilized Google Gemini for generating the pipeline visualization (Fig. 1) and the qualitative comparison chart (Fig. 3), as well as for grammatical refinement, readability improvements, and coding assistance. All AI-generated content was meticulously reviewed and revised by the authors, who take full responsibility for the final published version. We gratefully acknowledge the Traversaal AI team for open-sourcing the Alif model and the Urdu-Instruct dataset. We thank the developers of Lughaat and other Urdu language models whose work has advanced Urdu NLP.

Appendix: Training Environment

All experiments were conducted using the following software environment:

Model Training Frameworks:

- transformers==4.47.1
- trl==0.13.0
- peft==0.14.0
- accelerate==1.2.1
- unsloth @ 5dddf27

Core PyTorch and CUDA Stack:

- torch==2.5.1+cu121
- torchvision==0.20.1+cu121
- torchaudio==2.5.1+cu121
- bitsandbytes==0.45.0
- xformers==0.0.29.post1

Data Handling:

- datasets==3.2.0
- pandas==2.2.2
- tqdm==4.67.1

Hardware: Single NVIDIA A100 80GB GPU (rented via Vast.ai cloud infrastructure), CUDA 12.2

This environment ensures full reproducibility of our training and evaluation procedures.

References

[1] Llama Team, AI @ Meta, "The Llama 3 Herd of Models," *arXiv preprint arXiv:2407.21783*, 2024.

[2] M. A. Shafique, K. Mehreen, M. Arham, M. Amjad, S. Butt, and H. Farooq, "Alif: Advancing Urdu Large Language Models via Multilingual Synthetic Data Distillation," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP) Workshop on Multilingual Representation Learning (MRL)*, 2025.

[3] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of ACL*, 2020, pp. 8342–8360.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2022.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[6] T. L. Scao *et al.*, "BLOOM: A 176B-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[7] A. Üstün *et al.*, "Aya model: An instruction finetuned open-access multilingual language model," *arXiv preprint arXiv:2402.07827*, 2024.

[8] J. O. Alabi, D. Adelani, M. Klakow, and D. Klakow, "Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning," in *Proceedings of COLING*, 2022, pp. 4336–4349.

[9] S. Kanwal, K. Malik, K. Shahzad, F. Aslam, and Z. Nawaz, "Urdu named entity recognition: Corpus generation and deep learning applications," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 19, no. 1, pp. 1–13, 2020.

[10] B. Jawaid and D. Zeman, "Urdu-Hindi machine translation: Addressing the issues of word order," in *Proceedings of the Workshop on Indian Language Data: Resources and Evaluation (WILDRE)*, 2014.

[11] Z. Ahmed, "Makhzan: An Urdu Corpus for Computational Linguistics," 2020. [Online]. Available: https://github.com/zeerakahmed/makhzan

[12] Unsloth AI, "Unsloth: Fast and memory-efficient LLM fine-tuning," 2024. [Online]. Available: https://github.com/unslothai/unsloth

[13] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[14] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of ACL*, 2020, pp. 8440–8451.

[15] T. Dettmers *et al.*, "QLoRA: Efficient Finetuning of Quantized LLMs," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[16] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of EMNLP: System Demonstrations*, 2020, pp. 38–45.

[17] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[18] L. Xue *et al.*, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of NAACL*, 2021, pp. 483–498.

[19] N. Noreen, M. Lali, *et al.*, "Identifying sentiment in roman Urdu text using machine learning techniques," in *International Conference on Information Science and Communication Technology (ICISCT)*, 2019.

[20] J. Hoffmann *et al.*, "Training Compute-Optimal Large Language Models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 30016–30030.

[21] M. Noman, "Lughaat-1.0-8B-Instruct: An Advanced Urdu Language Model," *Hugging Face Model Hub*, 2025. [Online]. Available: https://huggingface.co/muhammadnoman76/Lughaat-1.0-8B-Instruct

[22] Gemma Team, "Gemma: Open Models Based on Gemini Research and Technology," *arXiv preprint arXiv:2403.08295*, 2024.

[23] J. Bai *et al.*, "Qwen Technical Report," *arXiv preprint arXiv:2309.16609*, 2023.

[24] A. Q. Jiang *et al.*, "Mistral 7B," *arXiv preprint arXiv:2310.06825*, 2023.