Aqal: First Urdu Reasoning Large Language Model through Continued Pretraining and GRPO-Based Reinforcement Learning

Azher Ali
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST), Islamabad, Pakistan
aali.msds2024seecs@seecs.edu.pk

1 Motivation and Problem Context

Recent advances in Large Language Models (LLMs) show that increasing data, parameters, and alignment techniques improves complex reasoning and multi-step problem solving, especially in English. However, these capabilities vary among languages.

Urdu, spoken by hundreds of millions, remains underrepresented in reasoning-focused LLM research. While multilingual models produce fluent Urdu text, they often struggle with structured reasoning tasks such as math word problems and logical inference. This highlights a gap between surface fluency and deeper reasoning skills.

The lack of reasoning-optimized Urdu LLMs limits educational, professional, and research applications. This project argues that these weaknesses come from insufficient Urdu-specific training and alignment, not from linguistic issues. By continuing pretraining, supervised fine-tuning (Chen et al., 2025), and reinforcement learning designed for Urdu reasoning, we aim to create stronger reasoning-focused Urdu LLMs.

2 Precise Problem Statement

Although multilingual LLMs can generate Urdu text, their reasoning skills in Urdu are significantly weaker than in English. This project addresses the problem of existing multilingual LLMs having limited ability to perform structured, multi-step reasoning in Urdu.

More formally, let M_0 represent a pretrained multilingual LLM. When tested on Urdu reasoning tasks T_{ur}, the model shows lower accuracy, higher hallucination rates, and weaker logical consistency compared to its performance on similar English tasks T_{en}. This gap suggests that the model's internal representations are not well-suited for reasoning in Urdu.

The goal of this project is to create a systematic training pipeline that converts M_0 into a reasoning-optimized Urdu model M_{ur} through three steps: continued pretraining (CPT) on large Urdu datasets, supervised fine-tuning (SFT) on selected Urdu reasoning data, and GRPO-based reinforcement learning for better reasoning alignment.

The central problem can therefore be stated as follows:

How can reasoning accuracy, logical coherence, and step-by-step analytical capability of a multilingual LLM in Urdu be significantly improved through structured multi-stage training?

3 Research Question and Hypotheses

The primary research question guiding this study is:

Can the reasoning performance of a multilingual base LLM in Urdu be significantly improved through a combination of continued pretraining, supervised fine-tuning, and GRPO-based reinforcement learning?

To address this question, three hypotheses are proposed.

First, continued pretraining on large-scale Urdu corpora will improve linguistic grounding and token distribution alignment. By reducing perplexity on Urdu text, the model is expected to produce more syntactically and semantically coherent outputs, thereby forming a stronger foundation for reasoning.

Second, supervised fine-tuning on curated Urdu reasoning datasets will directly enhance structured reasoning capabilities. Because reasoning tasks require explicit exposure to step-by-step examples, SFT is expected to significantly increase final-answer accuracy and step-level correctness.

Third, GRPO-based reinforcement learning will further improve logical coherence and reduce hallucination by optimizing a reward function that explicitly favors structured reasoning and correct

conclusions. Reinforcement learning is hypothesized to improve alignment beyond what supervised learning alone can achieve.

4 Task Formalization

The project focuses on generating structured reasoning in Urdu. The model takes an Urdu prompt that presents a reasoning problem. These prompts can involve math calculations, logical reasoning, analytical deduction, or everyday reasoning scenarios.

Formally, given an input sequence x in Urdu representing a reasoning problem, the model must produce an output sequence y that meets four criteria: linguistic correctness, a step-by-step reasoning structure, logical consistency, and the correctness of the final answer.

For example, consider a mathematical word problem in Urdu that asks for time calculation based on speed and distance. The expected output should not just be a number; it must include a structured explanation with the reasoning steps. This requirement sets the task apart from simply predicting an answer and connects it to chain-of-thought reasoning.

The challenges of the task include a lack of high-quality reasoning datasets in Urdu, possible errors in web-scraped data, different forms of Urdu script, and mixing in English. Moreover, computational limits require using parameter-efficient methods for training instead of retraining very large models completely.

5 Related Work and Research Gap

Multilingual LLMs such as Google Research's work on multilingual transformers and Meta AI's large-scale language models demonstrate broad cross-lingual capability, yet most reasoning-optimized research centers on English benchmarks like OpenAI's GPT series and related alignment studies. Instruction tuning and reinforcement learning approaches, particularly Reinforcement Learning from Human Feedback (RLHF) (Kirk et al., 2023), have proven effective for reasoning enhancement in English-dominant systems (Ouyang et al., 2022; Wei et al., 2022). However, comparable systematic investigations for Urdu remain largely absent.

Existing Urdu NLP research primarily focuses on machine translation, sentiment analysis, and text classification. Structured reasoning tasks—such as mathematical word problems, logical inference, and multi-step analytical reasoning—in Urdu are significantly underexplored. Furthermore, reinforcement learning–based alignment methods have not been specifically adapted or rigorously evaluated for Urdu reasoning performance.

The research gap lies at the intersection of three domains: low-resource language modeling, reasoning-oriented alignment, and reinforcement learning optimization. While prior studies have explored these areas independently, no comprehensive work has systematically examined how continued pretraining, supervised fine-tuning, and reinforcement learning interact to enhance structured reasoning in Urdu. This project addresses that gap through controlled empirical experimentation and ablation analysis.

6 Methodology Overview

The proposed methodology consists of a sequential three-stage pipeline.

In the first stage, continued pretraining is performed on a large-scale Urdu corpus. The objective is to reduce perplexity and improve linguistic representation in Urdu. This stage adapts the model's internal distribution to better reflect Urdu syntax, vocabulary, and discourse patterns.

In the second stage, supervised fine-tuning is conducted using a curated dataset of Urdu reasoning problems formatted in chain-of-thought style. The model is trained using standard cross-entropy loss to generate structured reasoning steps followed by a final answer. This stage explicitly teaches the model how to reason step by step in Urdu.

In the third stage, GRPO-based reinforcement learning is applied. A reward function is constructed to measure answer correctness, logical coherence, and structural consistency. The model is optimized to maximize expected reward, thereby aligning its outputs with reasoning objectives beyond supervised signal.

Performance comparisons are conducted across four model variants: the base model, CPT-only model, CPT+SFT model, and CPT+SFT+GRPO model. This design enables clear attribution of performance gains to each training stage.

7 Success Criteria and Evaluation Plan

Evaluation will combine quantitative and qualitative metrics. Quantitatively, final-answer accuracy and exact match scores will measure correctness.

Step-level logical consistency metrics will assess coherence across reasoning steps. Perplexity reduction will evaluate improvements from continued pretraining.

Human evaluation will complement automated metrics. Expert annotators will assess reasoning clarity, logical soundness, and hallucination frequency. Statistical significance testing will determine whether improvements over baseline are meaningful.

Success is defined as a statistically significant improvement in reasoning accuracy and logical coherence compared to the base multilingual model.

8 Risks and Challenges

The development of a reasoning-optimized Urdu Large Language Model involves several significant challenges spanning data availability, optimization stability, linguistic variability, and computational constraints. One of the primary risks is the limited availability of high-quality Urdu reasoning datasets. Unlike English, which benefits from extensive structured benchmarks for mathematical and logical reasoning, Urdu lacks curated chain-of-thought datasets. This scarcity may hinder the model's ability to learn generalized reasoning patterns and increase the risk of domain-specific overfitting, particularly if the dataset is narrow or translation-based.

Reinforcement learning introduces additional complexity. GRPO-based optimization (Guo et al., 2025) is sensitive to reward design, and poorly calibrated rewards may encourage superficial formatting rather than genuine logical coherence. Furthermore, reinforcement learning can destabilize training, potentially degrading linguistic fluency or causing policy collapse if updates are not carefully controlled. Ensuring stable alignment while preserving language quality is therefore a non-trivial challenge.

Supervised fine-tuning on limited reasoning data also carries the risk of memorization rather than true reasoning generalization. Without proper validation protocols, improvements may reflect pattern recall rather than transferable inference capability. Additionally, Urdu's orthographic variation, inconsistent diacritic usage, and frequent code-switching with English complicate preprocessing and tokenization, potentially affecting model efficiency and consistency.

These risks will be mitigated through careful dataset augmentation, structured reward design, conservative optimization strategies, validation-based early stopping, and standardized Urdu normalization pipelines. Addressing these challenges systematically is essential to ensure reliable and generalizable reasoning improvements.

9 Reflection

Reasoning requires structured multi-step inference rather than mere text continuation. Achieving logical coherence demands alignment between intermediate reasoning steps and final conclusions. In low-resource languages like Urdu, this challenge is amplified by limited structured supervision and fewer alignment-focused studies.

Furthermore, reinforcement learning introduces additional complexity in reward design and optimization stability. Balancing linguistic fluency with logical correctness is inherently difficult, particularly when training data is limited.

Therefore, building a reasoning-optimized Urdu LLM is not a straightforward extension of multilingual modeling but a complex research problem that combines language adaptation, supervised learning, and reinforcement alignment. Its successful implementation would contribute both technically and scientifically to the broader field of multilingual reasoning systems.

References

Zihong Chen, Wanli Jiang, Jinzhe Li, Zhonghang Yuan, Huanjun Kong, Wanli Ouyang, and Nanqing Dong. 2025. Graphgen: enhancing supervised fine-tuning for llms with knowledge-driven synthetic data generation. *arXiv preprint arXiv:2505.20416*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.