LANGUAGE TECH Artificial Intelligence / NLP / Research Commentary

Generative AI in Urdu: Who Is Building for Pakistan's Native Language?

Urdu is spoken by roughly 230 million people, yet it makes up a sliver of the text the world's AI models are trained on. A student's side project, a telecom-backed government MOU, and a two-decade-old university lab are each trying to close that gap — on their own.

By **Amna Sheikh** BUITEMS | Published: July 2026
amna.sheikh@gmail.com | Position: Computational Linguistics

Introduction

In January 2026, a Pakistani student studying in the United States, Taimoor Hassan, released Qalb — a large language model trained on 1.97 billion Urdu tokens and pitched as the world's largest LLM built exclusively for the language. Coverage called it Pakistan's own "Urdu ChatGPT," and for a country where most generative AI tools still think in English before translating into Urdu, the milestone was real.

But Qalb was not built in a vacuum, and it is not alone. A government-backed consortium had already signed on to build a national Urdu model over a year earlier, and a university language lab in Lahore has been quietly producing the speech and text tools that any Urdu AI system depends on since long before "generative AI" was a household phrase. The question worth asking is not whether Pakistan is building for Urdu — it clearly is — but whether these efforts add up to something bigger than the sum of their parts.

> ***"Urdu has more than 230 million speakers worldwide — and less than 0.1 percent of the internet's text to learn from. That mismatch, not ambition, is the real bottleneck."***

Three Different Bets on the Same Language

Qalb sits at one end of the spectrum: a small, founder-led team building Urdu as a first-class language at the architecture and dataset level, rather than fine-tuning an English model and hoping it translates well. It has been evaluated across more than seven international benchmarking frameworks and is aimed squarely at commercial use — customer service, education platforms, and voice-based AI agents for local businesses.

At the other end sits the state. NUST, the National Information Technology Board, and telecom operator Jazz signed an MOU to build a government-backed indigenous LLM centred on Urdu, with datasets extending to Pashto and Punjabi, aimed at healthcare, education, and agriculture use cases. And underneath both efforts is infrastructure few outside academia know exists: UET Lahore's Center for Language Engineering has spent over two decades building Urdu speech recognition, optical character recognition, and part-of-speech tagging tools, releasing much of it as free web services that any developer — including the newer entrants — can build on.

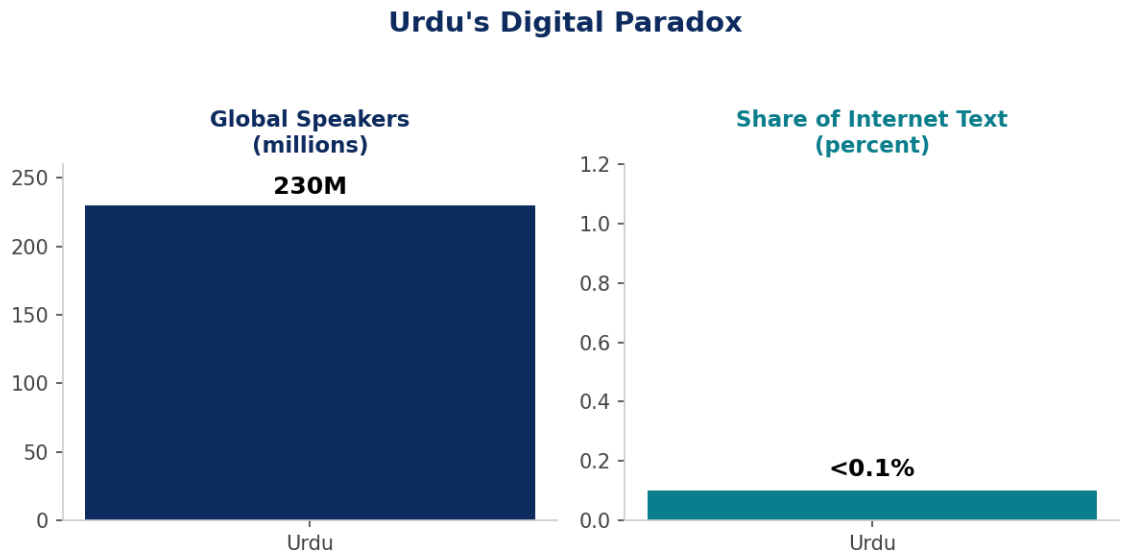

Figure 1: Urdu's digital paradox — global speakers versus share of internet text available to train AI models. Source: Digital Watch Observatory; NITB/Jazz MOU reporting.

Key Findings & Impact

What is working is proof of concept. Qalb's benchmarking claims, however they hold up under independent scrutiny, show that a small team without a national lab's budget can now train a credible Urdu model — a threshold that did not exist even two years ago. Meanwhile, global platforms are following the demand signal: Meta AI has rolled out Urdu-language support for its assistant, suggesting the market opportunity is no longer a niche argument academics have to make to funders.

What isn't working yet is coordination. A student startup, a telecom-backed government consortium, and a university lab are all drawing — separately — on the same scarce pool of digitised Urdu text, and none of the public reporting suggests they are pooling data or benchmarks. Most also default to formal, standard-script Urdu, while the way Pakistanis actually type day to day is heavily code-mixed Roman Urdu — "yaar kal ka match dekha?" — a register that remains poorly represented in any of these training sets.

What's Next?

The single highest-leverage move available to Pakistan's Urdu AI ecosystem is not another model launch — it is a shared data commons. Linking the Center for Language Engineering's two decades of linguistic assets to the newer web-scale corpora being assembled for Qalb and the NUST-led consortium, under a common licence, would do more for model quality than any one team's compute budget. Regional language datasets for Pashto and Punjabi, already scoped under the NITB-Jazz MOU, should be built on the same shared foundation rather than as separate side projects.

Pakistan does not lack people willing to build Urdu AI — a student, a telecom operator, and a university lab all made that clear within the space of two years. What it lacks is a mechanism for those builders to hand each other the one resource none of them can generate fast enough alone: a living, growing, shared corpus of the language itself.

About the Author

Amna Sheikh is a Computational Linguistics. She researches low-resource language modelling for South Asian languages. She has collaborated with regional language-technology labs on Urdu speech and text corpora.

Keywords: Generative AI; Urdu NLP; Low-Resource Languages; Large Language Models; Pakistan AI Policy

References

[1] The Express Tribune (2026) *Pakistani Student Develops World's Largest Urdu AI Model 'Qalb'.* Karachi: The Express Tribune, 18 January.

[2] Haq, R. (2024) *Pakistan to Develop Urdu LLM for Generative AI.* Haq's Musings, 7 November.

[3] Center for Language Engineering, KICS, UET Lahore (2020) *CLE NLP Webservices Documentation.* Lahore: UET.