Launching TelecomGPT-Arabic Initiative

The Region's First Arabic LLM Framework For Telecom

Dr. Lina Bariah

AI FOR GOOD SUMMIT, GENEVA

9 JUL 2025

Arabic in the Digital World

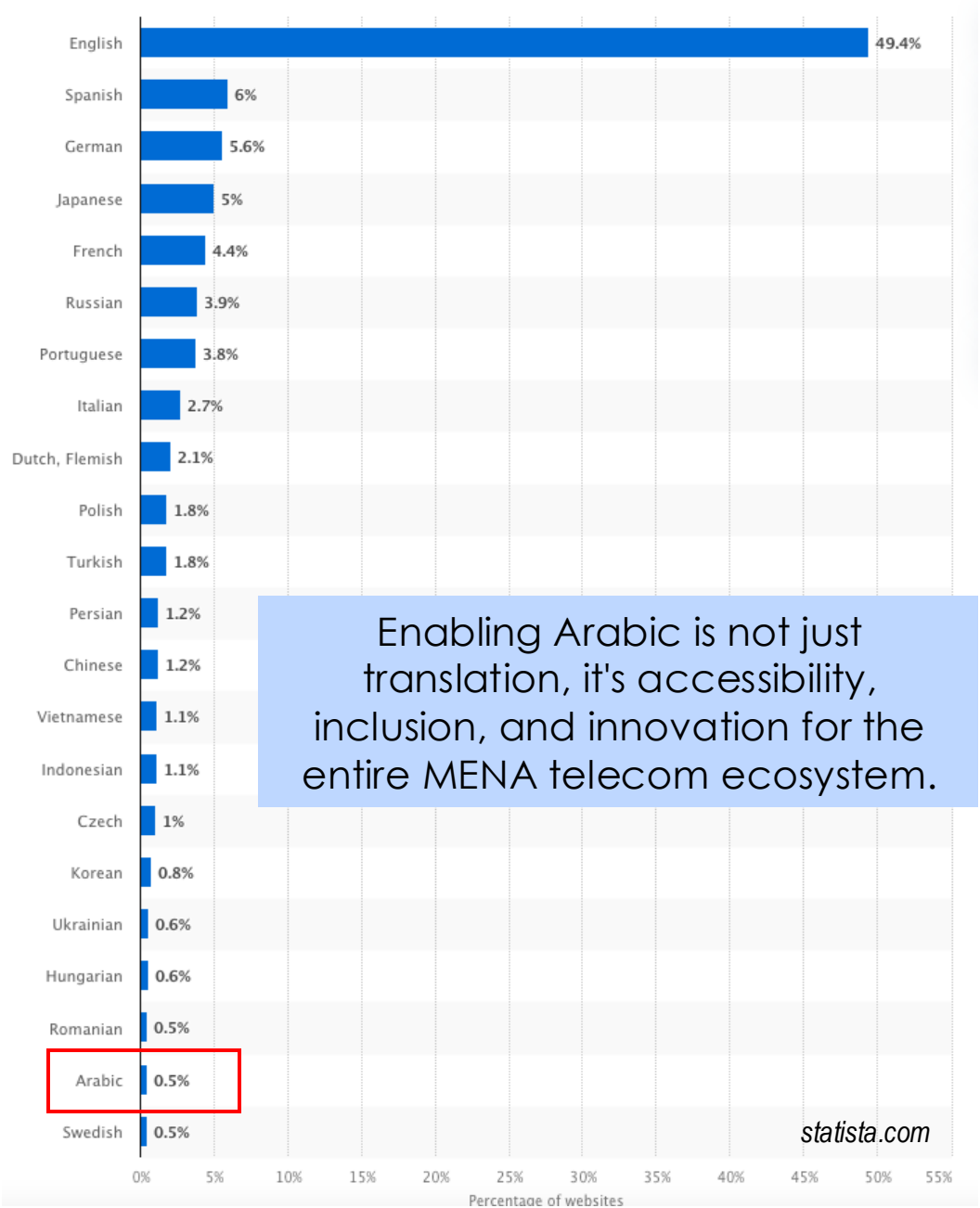
As of February 2025, 49.4% of web content is in **English**, with Arabic text accounting for only 0.5%!

But Arabic Telecom? ~ Almost negligible! (~0%)

"Local regulations (UAE, KSA, Egypt) encouraging Arabic-first interfaces, yet, Arabic Telecom data is nearly absent from AI systems, limiting accessibility, innovation, and regional relevance."

So .. Unlocking Arabic content is key to enabling:

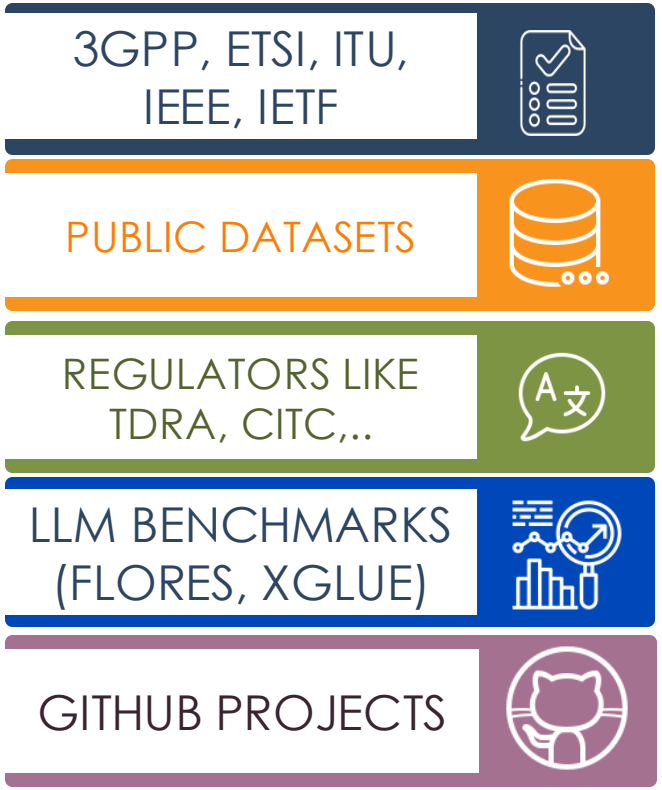
- Localized LLMs (e.g., TelecomGPT-Arabic)
- Multilingual RAG systems
- Arabic-speaking AI agents
- Arabic speech recognition & content moderation
- Regional innovation & regulatory alignment



Enabling Arabic is not just translation, it's accessibility, inclusion, and innovation for the entire MENA telecom ecosystem.

Arabic is Absent from Telecom AI

There is no mainstream telecom dataset in Arabic!



- Publish only in English
- 0% Arabic
- Bilingual summaries, but technical content = English
- Arabic included in general NLP, not in telecom
- No Arabic documentation

Arabic Telecom AI faces unique challenges due to the complexity of Arabic grammar, the wide variety of regional dialects, and the scarcity of high-quality datasets.

Arabic LLMs Activities

Huawei launched an Arabic LLM in Egypt amid the region's rising demand for generative AI.

Mistral releases regional model focused on Arabic language and culture.

Google also announced new Arabic-specific features for its generative AI platform Gemini.

Noor is recognized as one of the largest Arabic NLP models, trained on 10 billion parameters.

Jais is an Arabic-English bilingual LLM, by G42 & MBZUAI, trained on high-quality Arabic and English datasets.

Fanar is an Arabic AI Large Language Model developed by the Qatar Computing Research Institute.

> **Google** expects more **Arabic** large language models to emerge in the coming years, as AI developers' interest in one of the world's most widely-spoken languages grows
> - *The National*

And many more multilingual general-purpose LLMs ..

BUT .. How these models perform with Telecom Terminologies?

Why Arabic Telecom?

Telecom is a strategic sector in the MENA region (e.g., 5G/6G deployments, digital transformation, AI regulation)

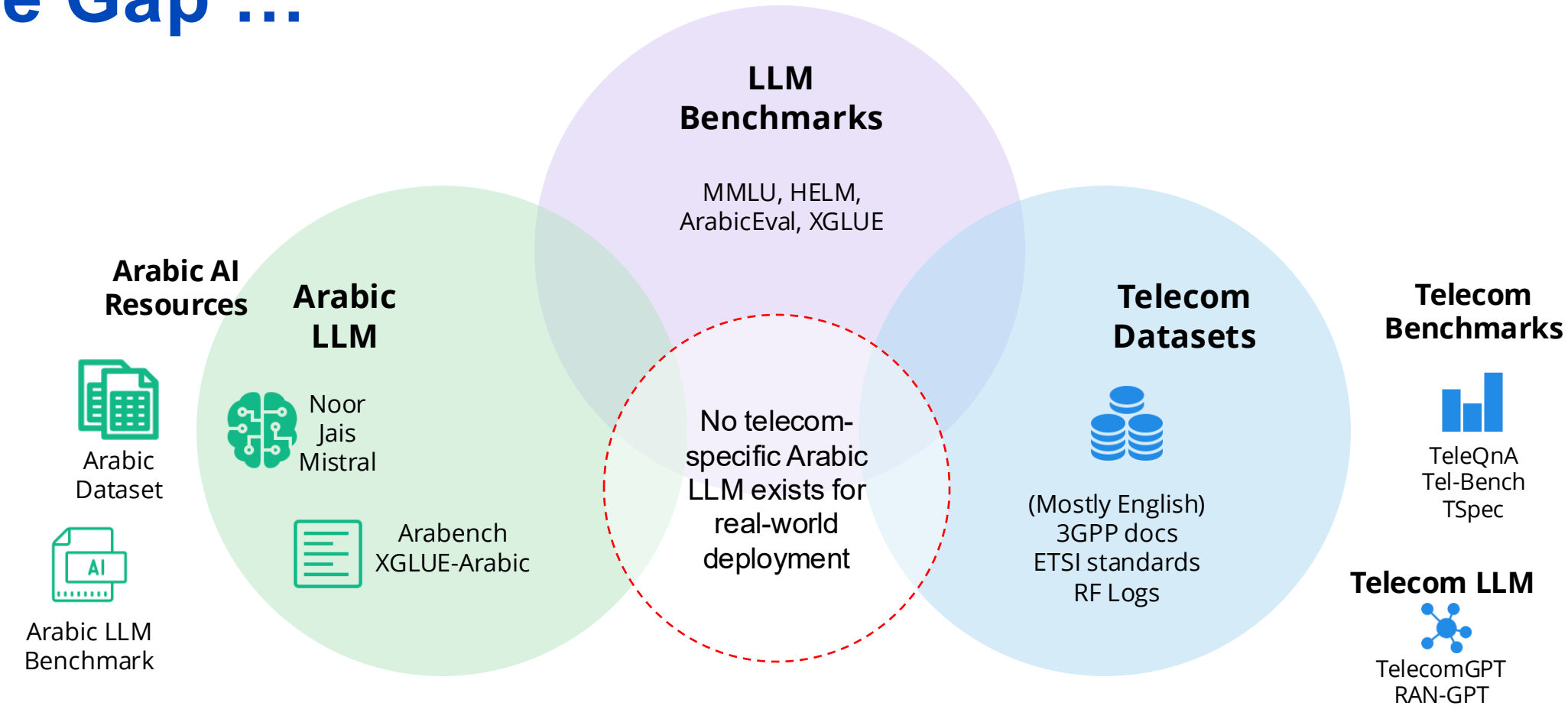
Arabic-first AI systems are needed for:

- Call centers
- RAG agents for local telecom docs
- Arabic interfaces for self-service apps

But! .. we **cannot**:

- Evaluate LLMs on telecom understanding
- Compare general-purpose Arabic LLMs vs. fine-tuned ones
- Drive progress toward Arabic TelecomGPT

The Gap …



TelecomGPT-Arabic bridges this gap — the first sovereign Arabic LLM for telecom use cases

If we don't speak telecom in Arabic, we exclude millions.

Market Need and Key Challenges for Arabic-Native AI in Telecom

$180B MENA Telecom Market Needs Arabic-Native AI

TelecomGPT-Arabic

LLMs alone aren't enough. Domain grounding is essential.

TelecomGPT-Arabic is the region's first sovereign telecom AI assistant trained on Arabic-language telco tasks.
While general multilingual LLMs exist, they fail to handle domain-specific terminology, telecom acronyms, and regulatory context in Arabic.
That's why we built a telecom-native LLM — fluent in both Arabic and telecom.

Why can't we use multilingual LLMs to translate Telecom datasets directly?

Domain-Specific Terminology Is Often Misunderstood

"Bearer," "handover," "paging" → often misinterpreted literally

Acronyms and Abbreviations Are Misused

Terms like "RRC," "NAS," "eNB," "QoS"

Contextual Precision Is Critical

Same word can mean different things in call flow vs. policy document vs. log

Using ChatGPT ..

Translate to Arabic

"Approaches to addressing the issue of modeling VHF/UHF for time percentages less than 1%"

الأساليب المتبعة لمعالجة مشكلة نمذجة ترددات VHF/UHF لنسب زمنية أقل من 1%

But, According to ITU Radio Regulations Resolutions and Recommendations Volume 3:

أساليب بحث الأثر التراكمي لتحديد مناطق التنسيق للمحطات الأرضية ذات الكثافة العالية (في الخدمتين الثابتة والمتنقلة)؛

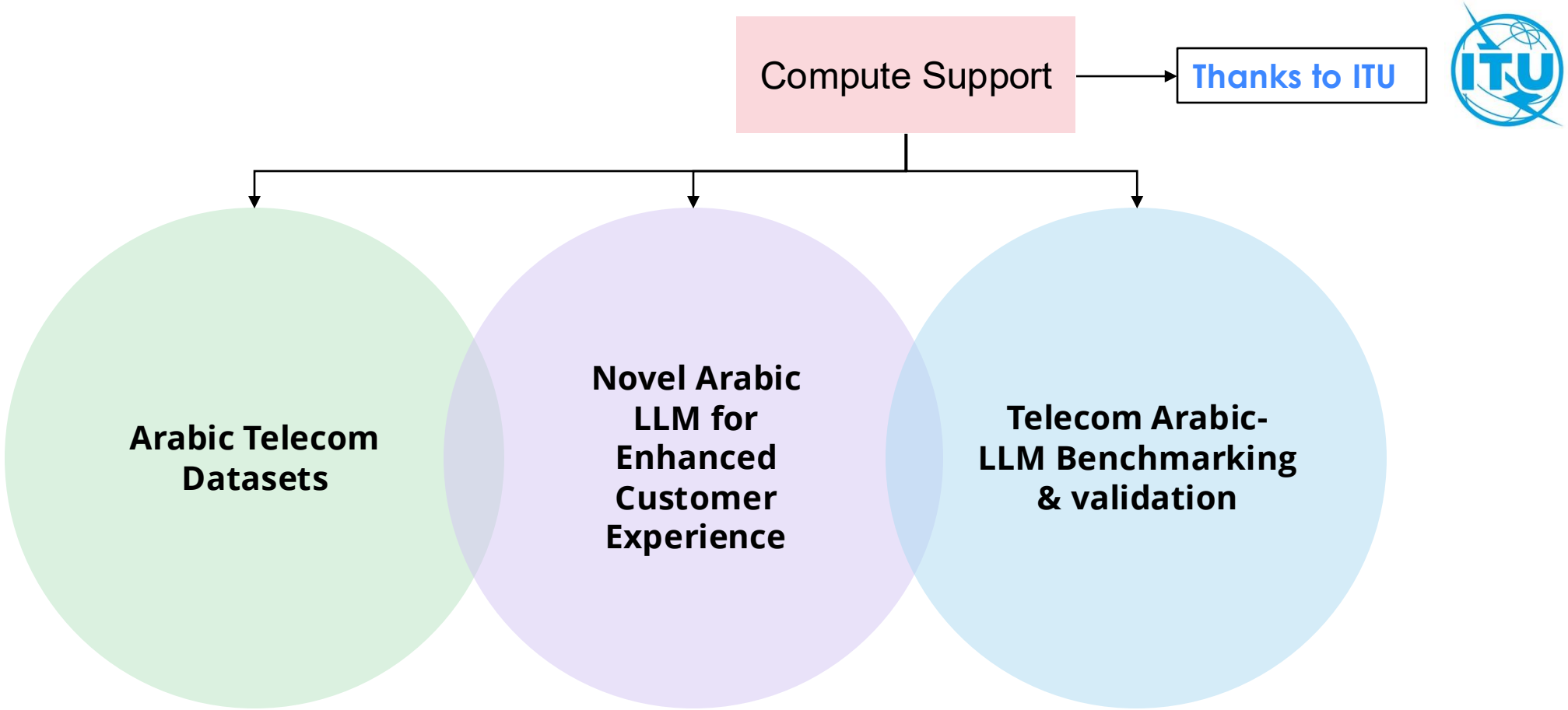
أساليب تناول مسألة وضع نماذج لترددات الموجات المترية/الديسيمترية (VHF/UHF) لنسب زمنية تقل عن 1 في المائة؛

أساليب دراسة كثافة بخار الماء بالنسبة لأسلوب الانتشار (1) في منطقتي المناخ المطري B وC؛

إدخال تحسينات في أسلوب الانتشار (2) لمعالجة مسألة التبعية لزاوية الارتفاع، وإزاحة مركز كفاف أسلوب

Who are the partners?

TelecomGPT-Arabic



TelecomGPT-Arabic to be released in Q3 - 2025