

Arabic LLMs: Initial Pillars Evaluation and the Future of Localization

[Assem Hijazi – Experience, Knowledge & Ideas](#)

As part of my initial thoughts of localization, I have decided to start checking the Arabic support in LLMs.

We now see new models coming out of the UAE, Saudi Arabia, Qatar, Egypt, and even independent research labs. They carry local names — like **Jais**, **Allam**, **Fanar**, **SILMA**, **Arcee-Meraj** — and big ambitions. "which is good news"

That got me thinking: **How do we evaluate Arabic LLMs fairly, beyond marketing headlines?**

The Strategy: A Pillar-Based Evaluation Framework

To organize my thinking, I started building a **set of 19 evaluation pillars** to benchmark Arabic LLMs — even if just from public data for now.

Those pillars are mainly technical I did not yet engage the main localization pillars to ease the starting.

Here are a few examples:

- Are they built **from scratch**, or fine-tuned on existing models like LLaMA or Qwen?
- Do they support **Modern Standard Arabic** only, or also dialects like Egyptian or Gulf Arabic?
- How many parameters are they built on — 7B? 13B? 70B?
- Do they support **RAG** (Retrieval-Augmented Generation)?
- Can users build their **own private knowledgebases** by uploading files or links?
- Are they **aligned with national strategies**, and **used in actual government projects**?
- What's the **general market perception** so far?

This framework doesn't claim to be final — it's a **first version**, based on open-source insights, not hands-on testing. But it's a starting point. And I believe it's the kind of structured thinking we need as we invest more into building sovereign, Arabic-first AI.

Why We Need a Structured Framework to Evaluate Arabic LLMs

As I started looking into Arabic large language models, one thing became clear: there's a growing list of them — but **very little structure** in how we evaluate them.

Each model seems to have its own story, its own data source, architecture, or mission — and most of the comparisons out there are either:

- Marketing-driven summaries ("fastest," "most advanced")
- Or limited to leaderboard rankings that don't reflect Arabic-specific priorities (like dialect handling or government integration).

So instead of doing a subjective review or relying only on global LLM benchmarks, I decided to take a **step back** and ask:

What are the right questions we should be asking about Arabic LLMs?

That's how I landed on the idea of a **pillar-based evaluation framework**.

This framework helps bring clarity by defining a **set of common factors** — technical, linguistic, strategic, and practical — that can be used to benchmark Arabic LLMs fairly, regardless of who built them or how popular they are.

I'm not claiming this is final. It's a **living framework**, and it will evolve as I continue this exploration, work with experts, and (hopefully) test the models more deeply myself.

But it gives us a starting point — and a way to begin comparing models not just by their size or source code, but by how **useful, accessible, and regionally aligned** they really are.

Again this does not include the full localization aspect of the main evaluation pillars

The Pillars: How We're Going to Evaluate Arabic LLMs

So instead of just saying "this model is good" or "that one is fast," I figured we need a better way to look at things. Something clear. Something repeatable. That's how I came up with this idea: **Let's build a list of evaluation pillars** — kind of like a checklist or scorecard for Arabic LLMs.

These pillars help answer questions like:

- Was this model really built from scratch, or just fine-tuned from something else?
- Does it support real Arabic use — not just formal MSA, but dialects too?
- Can it be used by governments? Enterprises? The average developer?
- Is it open-source? Can I plug in my own files and build on top of it?

In the end, I came up with **19 core pillars** — technical ones, strategic ones, and some that look at real-world adoption.

Here's the full list of the 19 pillars I'm using:

#	Factor	Description
1	Origin & Independence	Is it built from scratch or adapted from another model?
2	Model Size	How many parameters? (7B, 13B, 70B, etc.)
3	Training Data	What kind of data is it trained on? Is it transparent?
4	Domain Focus	Is it general-purpose or built for a specific domain (e.g., legal, medical)?
5	Language Coverage	MSA only? Or does it support dialects too?
6	Regional Reach	Is it used across the Arab world, or just one country?
7	Model Type	Open-source, closed, or hybrid?
8	Hosting & Infrastructure	Trained locally or on external clouds?
9	Tokenizer Design	Is the tokenizer Arabic-friendly or just reused from English models?
10	Fine-Tuning Strategy	SFT, RLHF, LoRA, DPO — what's behind it?
11	Multilingual Strength	Is it Arabic-dominant, or equally good at English too?
12	Real Use Cases	Where is it actually being used?
13	Benchmark Results	Has it ranked on Arabic leaderboards like OALL?
14	Ecosystem & Community	Any open dev tools, docs, GitHub repos, or Hugging Face models?
15	National Alignment	Is it part of a country's AI plan or vision?
16	RAG Support	Can it retrieve info from external documents?
17	User Knowledgebase Support	Can I upload files or links and chat with my own data?
18	Government Integration	Is it being used in public services or government tools?
19	Market Reputation	What's the general vibe — trusted? hyped? ignored?

MSA : Modern Standard Arabic (MSA)

It's the formal version of Arabic used in:

- News broadcasts
- Official documents
- Books, articles, and school textbooks
- Government communication
- Most digital platforms (if they support Arabic)

Pillar 5 is asking:

- Does the LLM only support **Modern Standard Arabic**?
- Or does it also understand and generate content in **regional dialects** like:

Supporting **MSA** is a minimum requirement. Supporting **dialects** = real-life usability, especially for chatbots, customer service, voice assistants, and social content.

I will add this to the localization pillar, i might regroup the pillars into new group and grouping pillars , still working on them

Quick First Scan – Where Arabic LLMs Stand Today

Once I had the pillars in place, I went hunting — checking articles, blog posts, leaderboards, Hugging Face models, and LinkedIn threads. I wasn't testing the models directly (not yet), but I wanted to get a feel for the scene.

And honestly? **There's a lot going on.** More than I expected.

Here are some early impressions from this first scan:

♦ **Some Models Are Built Seriously From the Ground Up**

- **Jais** (from G42 & MBZUAI in the UAE) seems to be the closest thing we have to an Arabic-first LLM that's trained almost entirely from scratch — using real Arabic datasets, modern Arabic, and dialects.
- Others, like **Falcon**, are trained in the region (by TII) but were more focused on English at first.

♦ **Fine-Tuned Models Are Leading the Pack**

- Models like **SILMA**, **Arcee-Meraj**, **Allam** (Saudi), and **Fanar** (Qatar) are fine-tuned versions of base models like Qwen or Gemma — but still show impressive performance in Arabic.
- Some of them are topping Arabic leaderboards like the **Open Arabic LLM Leaderboard (OALL)**.

◆ RAG and Custom Knowledgebase Support Is Rare (But Coming)

- A few models (like **Cohere's Arabic RAG model**) are built for this — but most others are still focused on core text generation.
- End-user upload support (like ChatGPT's "Custom GPTs" where users can add files and links) isn't common yet in regional models — but it should be.

◆ Government Backing Is a Big Deal

- Several of these models are either funded by or aligned with national strategies — Saudi's Allam, Qatar's Fanar, UAE's Jais.
- That's exciting, but it also means adoption may be **limited to official projects** (at least for now).

◆ Community Ecosystems Are Still Growing

- Open-source support is strong for some models (especially Jais, Falcon, and SILMA).
- But developer docs, APIs, fine-tuning instructions — those are still limited or scattered.
- The "plug-and-play" experience still isn't quite there yet for many of them.

This was just the **first pass**. I haven't touched the models yet, haven't stress-tested them, and haven't explored all the dialect support. That's the plan for later.

But at a high level, one thing is clear: **We've started — but we still have a long way to go.**

Confirmed Customized Arabic LLMs from the Arab World (as of 2025)

Here's a consolidated list of the key Arabic-first or Arabic-customized models developed in the Arab world: "as per my research"

#	Model Name	Country/Org	Notes
1	Jais	UAE (G42 + MBZUAI)	Most “from-scratch” Arabic LLM, trained on 398B tokens, dialect-aware
2	Falcon	UAE (TII)	Trained in UAE but English-first; not Arabic-focused initially
3	Allam	Saudi Arabia (SDAIA)	Saudi sovereign model, Arabic-focused, based on Falcon
4	Fanar	Qatar (MOTC)	Qatar’s national Arabic LLM, also based on Falcon or similar
5	SILMA v1.0	Egypt (SILMA AI)	Fine-tuned from Gemma, ranked #1 on OALL at launch
6	SILMA Kashif	Egypt	RAG-optimized 2B version of SILMA for doc retrieval
7	Arcee-Meraj	UAE (Arcee AI)	Built on Qwen2, heavily customized for Arabic, top ranked
8	AceGPT	Egypt (with multi-national contributions)	Arabic instruction-tuned model, part of AceGPT project
9	Yehia-7B-preview	Egypt/Saudi (linked to Allam project)	Preview fine-tuned model on top of Allam
10	Noor (TBA)	UAE (TII)	In development — a dedicated Arabic model from Falcon creators
11	Qalam (Rumored)	Jordan or UAE	Mentioned in regional research circles, no public release yet

Total Count (as of now):

- **Actively available models:** ~9
- **In development or preview stage:** ~2

Full Evaluation Applied to 11 Arabic LLMs

After building the framework, I applied all **19 evaluation pillars** to **11 Arabic LLMs** — including both established models (like Jais, Arcee-Meraj, SILMA, Allam) and newer or in-progress ones (like Fanar, Noor, and Qalam).

Since this creates a **massive comparison table**, I decided to keep things tidy and provide the **full evaluation as an Excel sheet** you can download, explore, or build on.

The link to the file, if you can not download ask me i will send you the file [Arabic LLM](#)

Sample Preview – 4 Models × 7 Pillars

Just to give you a quick taste of how it works, here’s a small snippet from the evaluation: We’ll take **4 well-known models** and apply **7 core pillars**:

Model	Built from Scratch	Parameters	Domain	Language Type	Open-source	Government Alignment	RAG Support
Jais	✔ Yes	13B / 30B	General-purpose	MSA + Dialects	✔ Yes	✔ UAE	✘
Arcee-Meraj	✘ (Qwen-based)	72B	General-purpose	MSA	Partial	✘ Private	✔
SILMA v1.0	✘ (Gemma-based)	9B	General-purpose	MSA	■	✘ Private	✔ (Kashif)
Allam	✘ (Falcon-based)	7B	General-purpose	MSA	✘ No	✔ Saudi Arabia	✘

This preview shows how the pillars help break things down clearly — not just in terms of size or name, but in real strategy and regional alignment.

Localization: Beyond Language

I've seen firsthand — through nearly 20 years working with Microsoft — how difficult and nuanced true Arabic localization is. It's not a "feature." It's an **ongoing commitment**. And honestly, it's a commitment that needs to come more from **within the region**, not just from global AI companies.

When we talk about evaluating Arabic LLMs, it's easy to fall into technical comparisons — parameters, benchmarks, datasets. But there's something bigger that can't be measured in billions of tokens or leaderboard scores:

Localization.

Arabic isn't just a language. It's a world of dialects, identities, cultures, religions, traditions, and regional sensitivities.

So while technical benchmarks are important, I believe any serious evaluation of Arabic LLMs must also look at:

- **Dialect coverage:** Are we just seeing Modern Standard Arabic, or are dialects like Egyptian, Levantine, Maghrebi, and Khaleeji truly supported?
- **Cultural alignment:** Do these models reflect Arab values, humor, etiquette, social norms?
- **Religious sensitivity:** Can the model engage respectfully with Islamic content, religious expressions, and theological nuance?
- **Support for heritage:** Does it handle diacritics, poetry structures, and classical texts — not just modern prose?
- **Local context understanding:** Is it aware of regional events, public figures, traditions, and institutions?

The more we want satisfaction, trust, and adoption across the Arab world, the more we need LLMs to be built **with the region in mind** — not just translated for it.

And that might mean **new pillars** in future versions of my evaluation — pillars focused on:

- Localization depth
- Cultural compatibility
- Trustworthiness in sensitive contexts
- Regional representation

Because in the end, true support for Arabic isn't just about language. It's about **identity**