



EPIC-PP-2026-012 · RESEARCH ARTICLE · AI GOVERNANCE

Whose Arabic? Diglossia, Dialect Gaps and Representational Justice in Arabic Large Language Models

EPIC – Emirati Perspectives and Insight Centre

Whose Arabic? Diglossia, Dialect Gaps and Representational Justice in Arabic Large Language Models

EPIC – EMIRATI PERSPECTIVES AND INSIGHT CENTRE

EPIC-PP-2026-012 · AI Governance · Published 20 October 2026

© 2026 EPIC – Emirati Perspectives and Insight Centre. Released under Creative Commons Attribution 4.0 International (CC BY 4.0): <https://creativecommons.org/licenses/by/4.0/>

EPIC is an independent research initiative in formation. It is not yet a registered or licensed entity. EPIC does not accept donations, payments or contracts.

Abstract

Arabic is spoken by hundreds of millions of people, but it is not one variety. Arabic speakers write and speak formally in Modern Standard Arabic and live much of their lives in regional dialects that differ widely from it and from each other. A new generation of Arabic-centric large language models, many built with public support in the Gulf states, promises to serve Arabic speakers in their own language. This article asks a question those projects have not fully faced: whose Arabic will these systems speak? Drawing on recent benchmark studies, it shows that Arabic models remain markedly weaker in dialects than in the standard language, that leading evaluation resources are built mainly around the standard language, and that some widely used dialect resources contain no data from the United Arab Emirates, Kuwait or Bahrain at all. The article then asks what kind of wrong such gaps involve. It distinguishes a distributive wrong (unequal quality of service), a recognitional wrong (failure to understand speakers on their own terms) and a standardising wrong (technology that quietly turns one register into the norm). Against both an "MSA-only" view and a demand for full dialect parity, it defends a principle of repertoire fidelity: systems should serve the full range of registers that their users actually move between, and should let users, not models, choose the register of a reply. It ends with proposals for evaluation, data partnerships, reporting and procurement, aimed at Gulf governments that fund national models.

Keywords: Arabic large language models; Diglossia; Linguistic justice; Representational harm; Language technology policy; Gulf states

I Introduction

In May 2025 the Technology Innovation Institute in Abu Dhabi announced Falcon Arabic, a model trained on native Arabic data that, according to its developers, covers Modern Standard Arabic and regional dialects (Technology Innovation Institute, 2025). It joined a

fast-growing group of Arabic-centric large language models. Jais, developed in the United Arab Emirates, was introduced in 2023 as an open bilingual Arabic-English model (Sengupta et al., 2023). ALLaM, developed by Saudi Arabia's national AI centre, followed in 2024 (Bari et al., 2024). These projects are often presented, by their developers and by the governments that fund them, as a matter of linguistic and cultural self-determination: a way to ensure that Arabic speakers are not left as second-class users of technologies built mainly in and for English.

This article takes that aim seriously and asks what it requires. Its starting point is a fact familiar to every Arabic speaker and to every linguist of Arabic, but easily lost in technical announcements: Arabic is diglossic. Most Arabic speakers use Modern Standard Arabic (MSA) for formal writing, news, scholarship and official life, and a regional dialect for nearly everything else, including family conversation, much social media, humour and intimacy (Ferguson, 1959; Holes, 2004). The dialects differ from MSA in vocabulary, morphology and pronunciation, and from each other to the point that speakers from distant regions may struggle to understand one another (Bouamor et al., 2018). An "Arabic" model therefore faces a question that an English model mostly does not: which Arabic?

The evidence reviewed in Section 3 suggests that, so far, the answer has been mainly MSA. Recent studies find that Arabic-centric models are better than general multilingual models at handling dialects, but that their performance on dialect tasks still lags well behind their performance on MSA, especially when they are asked to produce dialect rather than understand it (Mousi et al., 2025). Leading evaluation resources are built mainly in MSA (Koto et al., 2024). Even a landmark dialect resource, the MADAR corpus, covers 25 cities across the Arab world but includes none in the UAE, Kuwait or Bahrain (Bouamor et al., 2018).

The aim of this article is not to criticise particular projects, several of which are doing important work, but to clarify what kind of moral problem these gaps represent and what would count as solving it. Section 2 states the method. Section 3 sets out the linguistic background and the empirical evidence. Section 4 distinguishes three wrongs that dialect gaps may involve: distributive, recognitional and standardising. Section 5 considers and rejects two simple responses, an "MSA-only" view and a demand for full dialect parity, and defends instead a principle of repertoire fidelity. Section 6 draws out concrete implications for evaluation, data, reporting and procurement, with particular attention to the Gulf. Section 7 answers objections.

2 Method

The article combines a critical review of published technical literature on Arabic language technology with normative analysis drawn from political philosophy and the ethics of language. No new experiments were run and no human participants were in-

volved. Technical claims are taken from peer-reviewed papers, conference proceedings, preprints from the developers concerned, and official announcements, and are reported in the terms their authors use. Where the article draws inferences from these sources, it says so.

3 Diglossia, Data and the Dialect Gap

3.1 One language, many varieties

Ferguson (1959) introduced the term diglossia for situations in which two varieties of a language coexist in a community with a clear division of functions: a "high" variety, learned through schooling and used in formal settings, and a "low" variety, acquired at home and used in ordinary life. Arabic was one of his four defining cases. Later work has shown that the reality is more continuous than a simple two-level model suggests. Educated speakers move along a spectrum, mixing standard and dialect features according to topic, audience and purpose (Holes, 2004). But the core asymmetry remains. MSA is the variety of prestige, of print and of the state. The dialects are the varieties in which most people actually speak.

This asymmetry has a political history. Suleiman (2003) shows how the standard language became a central symbol of Arab national identity in the twentieth century, and how dialects were often treated as corruptions to be overcome. At the same time, the standard language carries real value. It connects speakers across the Arab world, it is the language of a shared literary and religious heritage, and it gives access to a vast body of writing. Any account of justice in Arabic language technology must keep both facts in view: the dialects are the living languages of their speakers, and MSA is a genuine common good.

3.2 The data problem

Large language models learn from text, and the text available in Arabic is unevenly distributed across varieties. Most published, edited and formally archived Arabic is in MSA. Dialects appear mostly in informal and spoken sources, such as social media, messaging and broadcast speech, which are harder to collect, harder to clean, and often written without standard spelling (Habash, 2010). Arabic as a whole is already underrepresented compared with English. Jais, one of the most Arabic-focused models, was pretrained on roughly 72 billion Arabic tokens alongside some 232 billion English tokens and code, because the Arabic data available for such training was so much smaller (Sengupta et al., 2023). Within Arabic, the dialect share is smaller again.

The same pattern appears in evaluation. ArabicMMLU, a major benchmark of knowledge and reasoning in Arabic built from school examinations across several Arab countries, is written entirely in MSA (Koto et al., 2024). This is a reasonable design choice for a benchmark of school knowledge. But when a central yardstick of progress is MSA-only, models can improve on it without improving in the varieties people mostly speak. What gets measured shapes what gets built.

Dialect resources do exist, and some are excellent. The MADAR corpus provides parallel translations of the same sentences into the dialects of 25 cities, a remarkable achievement (Bouamor et al., 2018). The NADI shared tasks have made fine-grained dialect identification a recognised research problem, using data from 21 Arab countries (Abdul-Mageed et al., 2020). The Gumar corpus collects about 110 million words of Gulf Arabic from online novels, annotated by sub-dialect (Khalifa et al., 2016). Yet coverage remains uneven. MADAR's Gulf region is represented by Mosul, Baghdad, Basra, Doha, Muscat, Riyadh and Jeddah, with Yemen treated separately; no city in the UAE, Kuwait or Bahrain is included (Bouamor et al., 2018). Any model evaluated on MADAR's Gulf test sets is therefore evaluated on varieties that do not include the dialects spoken by the citizens of three Gulf states, including the state that funds two of the region's best-known model families.

3.3 Evidence of a gap

The most systematic recent evidence comes from AraDiCE, a set of benchmarks for dialectal and cultural capabilities in Arabic language models (Mousi et al., 2025). Its authors created dialect versions of established benchmarks by machine translation and human post-editing, and built a new test of regional cultural knowledge for the Gulf, Egypt and the Levant. Their findings are nuanced. Arabic-centric models such as Jais and AceGPT outperform general multilingual models on dialect tasks, which shows that targeted investment pays off. But the authors also report that models "generally struggle" with dialect identification, generation and translation, that they "often rely on MSA knowledge for distinguishing between dialects", and that they are better at understanding dialects than at producing them. In their translation experiments, scores for translating from dialects into English were many times higher than scores for translating into dialects, for every model tested.

Two further findings matter for the argument here. First, performance varied by dialect in ways that seem to track closeness to MSA and online presence. The authors suggest that Gulf generation was stronger because of the dialect's closer resemblance to MSA, and that stronger Egyptian cultural knowledge may reflect Egypt's larger population and online footprint (Mousi et al., 2025). Second, these are exactly the patterns one would expect if dialect performance were a by-product of MSA and of sheer data volume rather than a design goal.

Cross-linguistic work points the same way. Kantharuban et al. (2023), studying dialect gaps across several languages, found that the strongest predictor of a model's performance on an under-resourced dialect was its linguistic closeness to a well-resourced variety. Dialects that are distant from the standard, or spoken by smaller or less digitally present communities, are left furthest behind. Related research has shown that even when language models operate in Arabic, they may default to Western cultural associations, for instance in names, foods and practices (Naous et al., 2024). Representation of a language is not the same as representation of the people who speak it.

The empirical picture, then, is not that Arabic models fail at dialects. It is that they serve dialect speakers less well than MSA users, that the gap is largest for generation, and that its distribution favours varieties that are close to the standard or well represented online. The question is what, if anything, is wrong with that.

3.4 Speech, mixing and script

Three features of real Arabic use make the gap wider than text benchmarks suggest. The first is speech. Voice interfaces are becoming a main route into AI systems, and spoken Arabic is overwhelmingly dialectal. A person asking a voice assistant about a medical appointment, a traffic fine or a school form will usually do so in dialect, often in the dialect of their home region rather than the country they live in. Benchmarks built from written MSA tell us little about how systems handle this.

The second is mixing. Arabic speakers frequently combine registers within a single sentence, moving between dialect and MSA, and many also mix Arabic with English or other languages, particularly in the Gulf, where English is widely used in business and education. Holes (2004) describes this movement along a continuum as a normal feature of educated speech rather than a lapse. Evaluation sets that present clean MSA or clean dialect sentences, as most do, do not test what systems do with the mixed input that users actually produce.

The third is script. Dialect writing has no single agreed spelling, and online users often write Arabic in Latin characters with numerals standing for some sounds. Bouamor et al. (2018) note that orthographic variation in dialectal Arabic makes it difficult for computational models to identify and reason about dialect words. A model that treats non-standard spelling as error will systematically underserve the people who write most informally, who are often also the people with the least formal schooling.

None of these features is exotic. Together they describe the ordinary linguistic life of a large share of Arabic speakers. They also explain why progress on standard benchmarks can coexist with persistent failure in everyday use.

Table 1 summarises the resources discussed in this section, what they cover and what they show.

Table 1 Selected Arabic resources and evaluations discussed in this article

Resource	Varieties covered	Relevance to the argument
ArabicMMLU (Koto et al., 2024)	MSA only; school-exam questions from several Arab countries	A central knowledge benchmark that cannot register dialect performance
MADAR corpus (Bouamor et al., 2018)	Dialects of 25 cities plus MSA; Gulf region represented by Iraqi, Qatari, Omani and Saudi cities	No city in the UAE, Kuwait or Bahrain; widely used for dialect evaluation
NADI 2020 (Abdul-Mageed et al., 2020)	Country- and province-level dialect identification across 21 countries	Shows dialect identification is a recognised but unsolved task
Gumar corpus (Khalifa et al., 2016)	About 110 million words of Gulf Arabic from online novels, annotated by sub-dialect	Shows substantial Gulf resources can be built from chosen sources
AraDiCE (Mousi et al., 2025)	MSA, Levantine and Egyptian benchmarks; cultural test for Gulf, Egypt and Levant	Arabic-centric models lead, but all models are weaker in dialects, especially in generation

4 Three Wrongs

Blodgett et al. (2020), in an influential survey of "bias" in language technology, argued that researchers should state clearly what harm they have in mind, and distinguished between allocational harms, in which resources or opportunities are unfairly distributed, and representational harms, in which systems represent some groups in ways that demean, erase or misdescribe them. Building on this distinction, we separate three wrongs that dialect gaps may involve. They are independent: a system could commit one without the others.

4.1 The distributive wrong

The simplest wrong is unequal quality of service. If a public service chatbot, an educational tool or a health information system works well for users who write in MSA and poorly for those who write in dialect, then the benefits of the technology are unequally distributed. Because command of MSA is acquired mainly through formal education, the gap is likely to track educational and social advantage. Older people, people with less schooling, migrants from other Arab countries whose dialects are less represented, and people in moments of stress, when they are least likely to switch into formal register, may all be served worse.

This wrong is serious, but it is also the easiest to see and to address, because it can be measured. The UAE Charter for the Development and Use of Artificial Intelligence commits to addressing algorithmic bias so as to make AI "inclusive and accessible to everyone", and to ensure "equal technological benefits" without exclusion (UAE Government, 2024). A dialect gap in public-facing systems is a direct instance of the problem that commitment names.

4.2 The recognitional wrong

A second wrong is harder to measure. Taylor (1994) argued that a person's identity is shaped in part by recognition or its absence, and that misrecognition can inflict real harm by reflecting back to people a demeaning or diminished picture of themselves. Patten (2014) has developed a liberal theory of minority rights around a principle of equal recognition: where the state supports cultural and linguistic practices, it should extend comparable support to the practices of different groups, rather than treating one as the default.

The dialects are not minority languages in the usual sense; they are the everyday varieties of majorities. But the structure of the wrong is similar. When a system treats a user's dialect as noise to be interpreted through MSA, misreads it, or answers it only in the standard register, it communicates that the user's ordinary way of speaking is not quite proper language. Fricker's (2007) notion of hermeneutical injustice is also relevant. Hermeneutical injustice occurs when a person's experience cannot be properly understood because the shared interpretive resources needed to make sense of it are lacking. Much of what people most need to say, about illness, family conflict, grief or local life, is said in dialect, with dialect vocabulary and idiom. A system that understands such speech poorly fails people precisely where understanding matters most.

4.3 The standardising wrong

The third wrong is the most structural. Bourdieu (1991) described language as a market in which some ways of speaking have more value than others, and in which institutions such as schools and states impose a legitimate language against which all other varieties are judged. Language technologies are now becoming one of those institutions. When millions of people write to, and read from, the same small number of systems, the variety those systems produce exerts pressure on usage. If the systems answer in MSA, correct dialect spellings towards standard forms, or reward formal prompts with better answers, they encourage users to adapt to the machine.

This pressure is not necessarily malign. Some standardisation has real benefits, and the spread of MSA through schooling and broadcasting has helped Arabic speakers communicate across borders. But there is a difference between a public decision to promote a standard language, made through education policy and open to debate, and a technical

default that achieves similar effects without anyone deciding it. Bender et al. (2021) warned that large language models trained on large, poorly documented datasets tend to encode the dominant voices in their data. In the Arabic case, the dominant voice is not only English; within Arabic, it is MSA and the best-resourced dialects. The standardising wrong is the risk that technology narrows the living diversity of Arabic as a side effect of what was easiest to build.

4.4 How the wrongs interact

The three wrongs reinforce one another. Unequal service gives users a reason to switch into MSA when they address a machine. Each such switch produces more MSA interaction data, which developers then use to improve their systems, which widens the gap further. Over time, the recognitional message that dialect is not proper language for serious matters is taught not by any explicit statement but by the experience of being understood better in one register than another. And as more everyday writing passes through systems that prefer the standard, the standardising pressure grows.

This feedback loop is what distinguishes the present situation from earlier forms of diglossia. Schools and broadcasters have always favoured MSA, but they did not sit inside people's private conversations, answering their questions about health, money and family. Generative systems increasingly do. The scale and intimacy of the interaction give the technology's default register an influence that no earlier institution had. For that reason, the default is a matter of public concern, not merely a product feature.

It is also worth noting who bears the cost. The users most affected are not, in the main, the professional and educated users who are comfortable moving into MSA, but those for whom the dialect is the language in which they think and feel most naturally. A technology funded in the name of the Arabic language, but tuned to the register of its most educated speakers, risks redistributing advantage towards those who already have it.

5 Repertoire Fidelity

5.1 Two simple answers

Faced with these wrongs, two simple responses suggest themselves. The first, which we will call the MSA-only view, holds that Arabic models should focus on the standard language. MSA is shared across the Arab world, it is the language of education and government, it has the most data, and a single standard target allows models to serve the widest audience. On this view, dialect support is a luxury, and perhaps even undesirable, since it might fragment a shared language.

The second response, dialect parity, holds that every major dialect deserves performance equal to MSA, and that anything less is unjust. On this view, developers should invest until the gap is closed for each variety.

Both views fail. The MSA-only view mistakes the language people are taught for the language they live in. It accepts the distributive and recognitional wrongs as the price of efficiency, and it endorses the standardising wrong as a goal. It also misdescribes how Arabic speakers actually use language: as Holes (2004) shows, educated speakers routinely mix registers, and much written communication, especially online, is in dialect or in mixed forms. A model that serves only MSA does not serve a coherent population of users; it serves one register of every user.

Dialect parity fails in a different way. There is no fixed list of Arabic dialects. Varieties shade into each other along continua of geography, community and generation, and any list of "major dialects" will itself be a political choice that favours some communities over others. Full parity for every variety is not achievable, and treating it as the standard of justice would make every real system unjust by definition. Parity also ignores the genuine value of MSA. Users sometimes want the standard register, for instance when drafting a formal letter or reading about a technical subject, and a model that defaulted to dialect would serve them badly.

5.2 The principle

We propose instead a principle of repertoire fidelity. Most Arabic speakers do not have a single language. They have a repertoire: a set of registers, including MSA, their own dialect and often features of other dialects, among which they move according to situation. A language technology serves such users justly to the extent that it is faithful to their repertoire. This has three parts.

First, comprehension across the repertoire. A system should understand users in whichever register they choose to write or speak, including their own dialect and mixtures of dialect and standard, with performance that does not fall sharply when they move away from MSA. Comprehension is where the distributive and recognitional wrongs bite hardest, because a system that misunderstands a user cannot help them at all. Comprehension is also, according to the evidence reviewed above, where models are already closest to adequate, so the goal is within reach (Mousi et al., 2025).

Second, user control over production. A system should produce text in the register the user wants, not the register the model finds easiest. By default, it might mirror the register of the user's input. It should allow users to request MSA, their dialect or a mixed register explicitly. And it should never treat dialect input as an error to be corrected unless correction is what the user asked for. This addresses the standardising wrong by moving the choice of register from the model to the person.

Third, honesty about limits. Where a system's performance in a given variety is weak, it should say so, both in public documentation and, where appropriate, in the interface. A model that answers fluently in MSA while misunderstanding a dialect query does more harm than one that signals its uncertainty. Honesty is also what makes the principle achievable: no system will serve every variety equally well, but every system can be transparent about where it falls short.

An example makes the principle concrete. Suppose an elderly user in Sharjah writes to a public health chatbot, in Emirati Arabic and with informal spelling, to ask whether a medicine she has been prescribed can be taken while fasting. A system faithful to her repertoire would understand the question as written, without requiring her to restate it in MSA. It would answer in a register she can follow, by default close to her own, and it would offer the formal register if she or a family member preferred it. If it were uncertain about a dialect term, it would ask rather than guess. None of this requires the system to be equally fluent in every Arabic variety. It requires only that it serve the person in front of it on her own linguistic terms.

Repertoire fidelity differs from parity in that it does not require equal performance in every variety. It requires that the gaps be small enough that users are not excluded, that the choice of register belong to users, and that remaining gaps be visible. It differs from the MSA-only view in that it treats dialects as part of the language people are entitled to be served in, not as noise around it.

5.3 Fit with linguistic justice

The principle fits a wider tradition of thought about linguistic justice. Van Parijs (2011), writing about the global spread of English, defends the practical value of a common lingua franca while insisting on what he calls parity of esteem among languages: the spread of a shared language must not be allowed to signal that other languages, and their speakers, are of lesser worth. Repertoire fidelity applies the same idea inside Arabic. MSA can and should serve as a shared medium, but its role as a medium should not turn into a judgment that the dialects are inferior. Patten's (2014) principle of equal recognition supports the same conclusion from a different angle: where public resources are spent on language technology, they should not systematically favour the registers of the already advantaged.

The principle also connects to work on decolonial and relational approaches to AI. Mohamed et al. (2020) argue that AI development should attend to whose knowledge and whose ways of life are encoded as defaults, and Birhane (2021) argues for an ethics that begins from the experience of those most affected by a system rather than from abstract fairness metrics. In the Arabic case, those most affected by dialect gaps are not foreign users but ordinary citizens of the very states building the models. The irony is worth stat-

ing: national AI projects justified as a defence of Arabic against English could, if they serve only the standard register, reproduce within Arabic the same hierarchy they set out to resist.

6 Implications for Gulf AI Policy

The Gulf states are unusually well placed to act on these arguments. They fund national models, they buy AI systems for public services at scale, and they have explicit cultural policies on the Arabic language. In 2012 the UAE announced an Arabic Language Charter, whose stated objective was "to protect the essence of the language and enhance its usage", as part of a strategy to make the country a global centre of excellence for Arabic; the same announcement included plans to enhance Arabic content online and to promote Arabic as a language of science and technology (Gulf News, 2012). The UAE's AI Charter commits to mitigating algorithmic bias, to inclusion and to transparency (UAE Government, 2024). Repertoire fidelity offers a way to bring these commitments together. Five measures follow.

6.1 Evaluate by variety

Public funders and buyers should require that Arabic models be evaluated, and the results published, separately for MSA and for a defined set of dialects, including the dialects of the funding country's own population. The evidence in Section 3 shows that a single "Arabic" score can hide large gaps. Evaluation should cover both comprehension and generation, since the gap is largest in generation, and it should include mixed-register input of the kind real users produce. Existing benchmarks such as AraDiCE are a starting point, but they do not yet cover every Gulf variety. Building evaluation sets for Emirati, Kuwaiti and Bahraini Arabic, among others, is a modest and well-defined task that Gulf research institutions are well placed to lead.

6.2 Build dialect data with communities, not just from them

Closing comprehension gaps requires dialect data, and much dialect data is personal: conversations, voice messages, social media posts. Harvesting it without consent would trade one wrong for another. Gulf institutions should instead build dialect resources through consented partnerships, for instance with schools, universities, broadcasters, cultural heritage bodies and volunteer contributors, with clear terms about use. Datasheets documenting how each dataset was collected, from whom and with what consent should be standard practice (Geburu et al., 2021). The Gumar corpus, built from publicly available online novels and annotated by sub-dialect, shows that substantial Gulf resources can be built from carefully chosen sources (Khalifa et al., 2016).

6.3 Report honestly

Model documentation should state, in a standard format, how a model performs across Arabic varieties and where it is known to be weak. Model cards, proposed as a way to report performance across groups and conditions, provide a ready template (Mitchell et al., 2019). For public-facing deployments, interfaces should allow users to see which registers a system supports well, and systems should signal uncertainty when input falls outside them.

6.4 Give users the choice of register

Public services that deploy Arabic chatbots should allow users to choose the register of replies, and should by default mirror the user's register rather than switching to MSA. Systems should not silently "correct" dialect input into standard forms. These are small design choices with large effects on whether users feel served or corrected.

6.5 Procure for fidelity

Finally, governments can use their purchasing power. Procurement rules for AI in public services, such as health, education, justice and social support, can require minimum performance in the dialects of the population served, evaluated independently. Because Gulf governments are among the largest buyers of AI services in the region, such requirements would create a market incentive for every provider, local or foreign, to close dialect gaps.

6.6 Treat Arabic language technology as cultural infrastructure

The five measures above share a premise: that Arabic language models funded by public money are not only commercial products but part of the cultural infrastructure of Arabic-speaking societies, comparable to dictionaries, broadcasting and school curricula. Infrastructure of this kind has traditionally been governed with explicit attention to whose language it carries. The Arabic Language Charter shows that Gulf governments already think of Arabic in these terms (Gulf News, 2012). Extending that thinking to language models means asking, for each publicly funded system, which varieties it was built to serve, how well it serves them, and who decided. It also means involving linguists, educators and community representatives, not only engineers, in setting those goals.

7 Objections

7.1 Too many dialects

It may be objected that Arabic has too many varieties for any of this to be practical. The objection is right that full coverage is impossible, which is why the article rejects parity. But repertoire fidelity does not require covering every variety equally. It requires that funders choose, transparently, which varieties to evaluate and improve, that they include the varieties of the populations their systems serve, and that remaining gaps be disclosed. A public chatbot in Abu Dhabi does not need to generate Moroccan Darija fluently. It does need to understand the Emirati Arabic of its users and the dialects of the large Arab expatriate communities it also serves.

7.2 Promoting MSA is legitimate cultural policy

A second objection holds that states are entitled to promote MSA, and that technology which encourages its use serves a legitimate public aim. The article agrees that promoting MSA can be legitimate. But legitimacy depends on how it is done. A public decision to strengthen MSA in schools and official life is open to debate and accountable. A technical default that produces the same effect by serving dialect speakers worse is neither. Moreover, repertoire fidelity is compatible with promoting MSA: a system that understands dialect input and offers MSA output when users want it may help people move between registers, rather than penalising them for where they start.

7.3 The market will solve it

A third objection holds that commercial competition will close dialect gaps without policy, since providers that serve dialect speakers better will win more users. Competition will help, and recent models already show progress (Technology Innovation Institute, 2025). But the cross-linguistic evidence suggests that market forces favour varieties that are close to the standard and abundant online (Kantharuban et al., 2023; Mousi et al., 2025). Smaller and less digitised varieties will be served last, if at all, without deliberate investment. That is the classic case for public action.

7.4 Dialect support will divide Arabic speakers

A fourth objection is that investing in dialects will weaken the shared standard and, with it, the cultural unity of Arabic speakers that the standard language has long symbolised (Suleiman, 2003). The worry deserves respect, since the shared standard is a real achievement. But it rests on an assumption that the evidence does not support: that serving people in their dialect draws them away from the standard. Repertoire fidelity does the opposite of fragmenting. It treats the standard and the dialects as parts of one reper-

toire, and it asks technology to move with users across that repertoire rather than forcing them to one end of it. A system that understands a Kuwaiti user in Kuwaiti Arabic, an Egyptian user in Egyptian Arabic and both in MSA is, if anything, a stronger bridge between them than one that understands neither unless they switch registers. Unity that depends on the inability of technology to understand ordinary speech is a fragile kind of unity.

8 Conclusion

The Gulf's investment in Arabic language models is among the most important developments in global AI outside the United States and China. Its stated purpose, to ensure that Arabic speakers are served by generative AI in their own language, is admirable. This article has argued that the purpose is not yet fully realised, because "their own language" is not a single variety. Current models, and the benchmarks used to measure them, are oriented towards Modern Standard Arabic. Dialect performance lags, especially in generation, and its distribution favours varieties that are close to the standard or well represented online. Some widely used dialect resources omit the dialects of the UAE, Kuwait and Bahrain entirely.

These gaps involve three distinct wrongs: unequal service, failure of recognition, and a quiet standardisation of a living language. The appropriate response is neither to accept MSA as the sole target nor to demand an impossible parity, but to seek repertoire fidelity: comprehension across the registers people use, user control over the register of replies, and honesty about limits. Gulf governments, as funders, regulators and buyers, can make that principle real through evaluation by variety, consented data partnerships, honest reporting, user choice of register and procurement for fidelity. If they do, the Arabic models built in the region could become not only a defence of Arabic against English but a model of how language technology can serve a language as its speakers actually live it.

References

- Abdul-Mageed, M., Zhang, C., Bouamor, H., & Habash, N. (2020). NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (pp. 97–110). Association for Computational Linguistics. <https://aclanthology.org/2020.wanlp-1.9/>
- Bari, M. S., Alnumay, Y., Alzahrani, N. A., Alotaibi, N. M., Alyahya, H. A., AlRashed, S., Mirza, F. A., Alsubaie, S. Z., Alahmed, H. A., Alabduljabbar, G., Alkhathran, R., Almushayqih, Y., Alnajim, R., Alsubaihi, S., Al Mansour, M., Alrubaian, M., Alammari, A., Alawami, Z., Al-Thubaity, A., ... Khan, H. (2024). ALLaM: Large language models for Arabic and English (arXiv:2407.15390). arXiv. <https://doi.org/10.48550/arXiv.2407.15390>

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2), Article 100205. <https://doi.org/10.1016/j.patter.2021.100205>
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.485>
- Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., & Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association. <https://aclanthology.org/L18-1535/>
- Bourdieu, P. (1991). *Language and symbolic power* (J. B. Thompson, Ed.; G. Raymond & M. Adamson, Trans.). Harvard University Press.
- Ferguson, C. A. (1959). Diglossia. *Word*, 15(2), 325–340. <https://doi.org/10.1080/00437956.1959.11659702>
- Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Gulf News. (2012, April 23). UAE to be a 'centre of excellence' for Arabic language. <https://gulfnews.com/uae/government/uae-to-be-a-centre-of-excellence-for-arabic-language-1.1012508>
- Habash, N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool. <https://doi.org/10.1007/978-3-031-02139-8>
- Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties* (Rev. ed.). Georgetown University Press.
- Kantharuban, A., Vulić, I., & Korhonen, A. (2023). Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7226–7245). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.481>

- Khalifa, S., Habash, N., Abdulrahim, D., & Hassan, S. (2016). A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 4282–4289). European Language Resources Association. <https://aclanthology.org/L16-1679/>
- Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., & Baldwin, T. (2024). ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 5622–5640). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-acl.334>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). Association for Computing Machinery. <https://doi.org/10.1145/3287560.3287596>
- Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4), 659–684. <https://doi.org/10.1007/s13347-020-00405-8>
- Mousi, B., Durrani, N., Ahmad, F., Hasan, M. A., Hasanain, M., Kabbani, T., Dalvi, F., Chowdhury, S. A., & Alam, F. (2025). AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 4186–4218). Association for Computational Linguistics. <https://aclanthology.org/2025.coling-main.283/>
- Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024). Having beer after prayer? Measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16366–16393). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.862>
- Patten, A. (2014). *Equal recognition: The moral foundations of minority rights*. Princeton University Press.
- Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., Afzal, O. M., Kamboj, S., Pandit, O., Pal, R., et al. (2023). Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models (arXiv:2308.16149). arXiv. <https://doi.org/10.48550/arXiv.2308.16149>
- Suleiman, Y. (2003). *The Arabic language and national identity: A study in ideology*. Edinburgh University Press.
- Taylor, C. (1994). The politics of recognition. In A. Gutmann (Ed.), *Multiculturalism: Examining the politics of recognition* (pp. 25–73). Princeton University Press.

Technology Innovation Institute. (2025, May 21). Middle East's leading AI powerhouse TII launches two new AI models: Falcon Arabic, the first Arabic model in the Falcon series, and Falcon-H1. <https://www.tii.ae/news/middle-east-s-leading-ai-powerhouse-tii-launches-two-new-ai-models-falcon-arabic-first-arabic>

UAE Government. (2024). The UAE Charter for the Development and Use of Artificial Intelligence. The Official Portal of the UAE Government. <https://u.ae/en/about-the-uae/strategies-initiatives-and-awards/policies/Ai/The-UAE-Charter-for-the-Development-and-Use-of-Artificial-Intelligence>

Van Parijs, P. (2011). Linguistic justice for Europe and for the world. Oxford University Press.