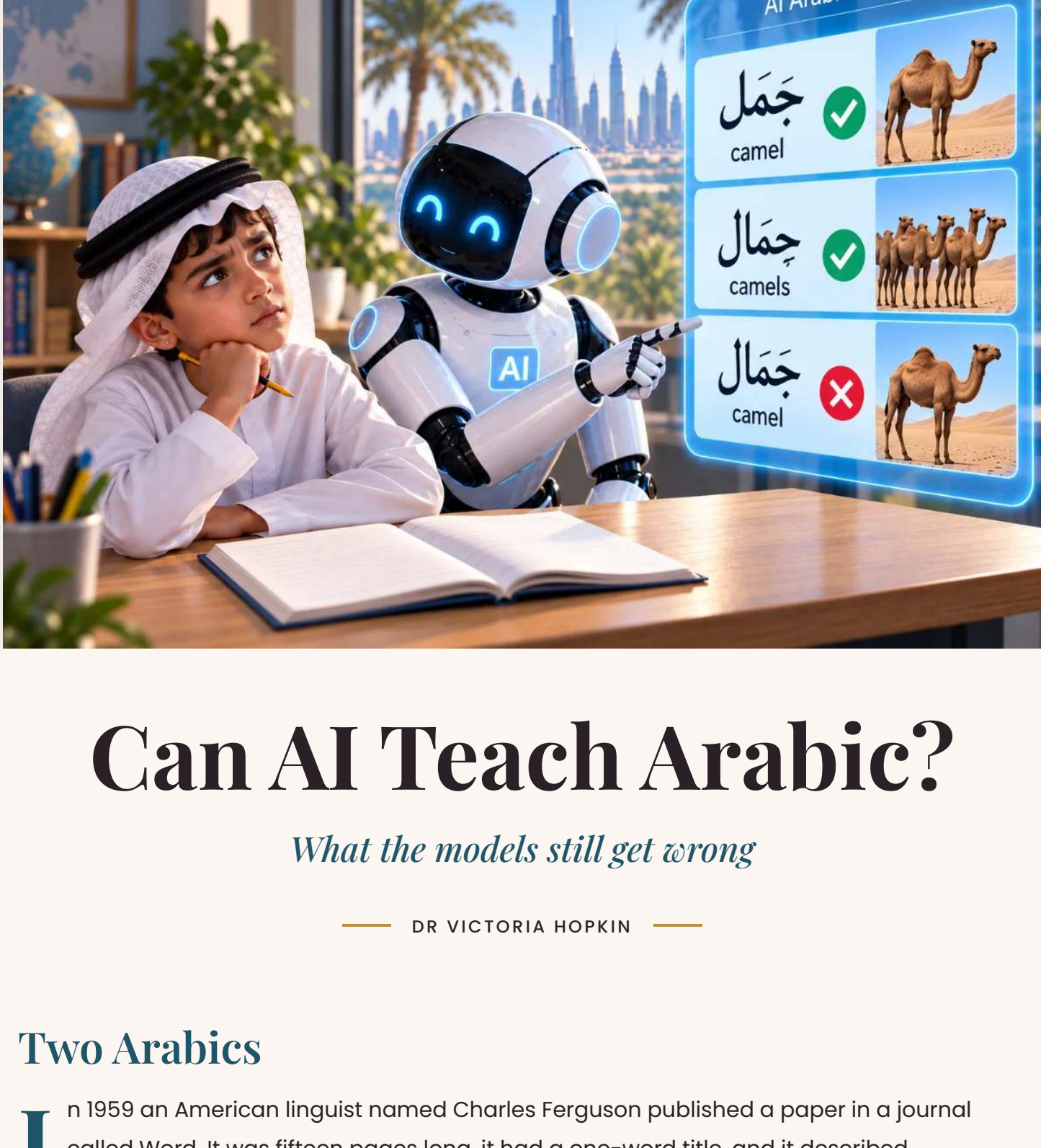




Can AI Teach Arabic?

What the models still get wrong

— DR VICTORIA HOPKIN —

Two Arabics

In 1959 an American linguist named Charles Ferguson published a paper in a journal called Word. It was fifteen pages long, it had a one-word title, and it described something that everyone who spoke Arabic already knew and nobody had written down.

Ferguson had spent years with Arabic, and he had noticed a pattern that did not fit the way linguists then thought about languages. In the Arab world there was not one Arabic but two, living side by side in the same people, and the two were kept apart by an unwritten rule about where each was allowed to go. He made a list of occasions. A sermon in a mosque. A speech in parliament. A news broadcast. A university lecture. Poetry. A newspaper editorial. For all of these there was one Arabic, formal and classical, which he called H, for high. Then the other list. Conversation with family and friends. Instructions to a waiter. A radio soap opera. Folk stories. The caption on a political cartoon. For all of these there was a second Arabic, which he called L, for low, and which was never written and never taught.

"The language they used every hour of every day was the one they would not admit to speaking."

And then Ferguson recorded the detail that makes the paper worth reading sixty-seven years later. When he asked educated speakers about L, they often denied it. Some told him they did not use it. Some told him it was not real Arabic, or had no grammar, or was a corruption that ought to die out. They said this in L, to a foreigner, in the ordinary course of a conversation. The language they used every hour of every day was the one they would not admit to speaking.

He called the arrangement diglossia. The word stuck, the paper became one of the most cited in the history of sociolinguistics, and the pattern it described is still the first thing any teacher of Arabic has to understand. It is also, it turns out, the first thing the machines got wrong.

Ferguson's Machines

Nathan Robinson works at Cohere, an AI company in Toronto, and in 2024 he and three colleagues set out to answer a question that sounds simple. How good are the large language models at Arabic? The answer depended entirely on which Arabic you meant, and so they built a test that asked the question Ferguson would have asked.

They took nine models, including the ones behind the most widely used chatbots and two built in the Gulf, and eight varieties of spoken Arabic, from Morocco to the Gulf. Then they measured four things separately. Could the model understand a dialect when it was written down? Could it produce the dialect when asked to? When it produced it, was the Arabic good? And did the model know which Arabic belonged in which situation, the sermon or the soap opera, which is Ferguson's question exactly.

The models understood the dialects well. Given a sentence in Egyptian or Gulf or Levantine Arabic, they knew what it meant. Asked to answer in that same dialect, they mostly would not. They slid back toward H, toward the formal written Arabic of the news broadcast, and when they did attempt the dialect the result was often thin, or wrong, or a few local words dropped into an otherwise formal sentence. Understanding ran far ahead of production, across every model and every dialect.

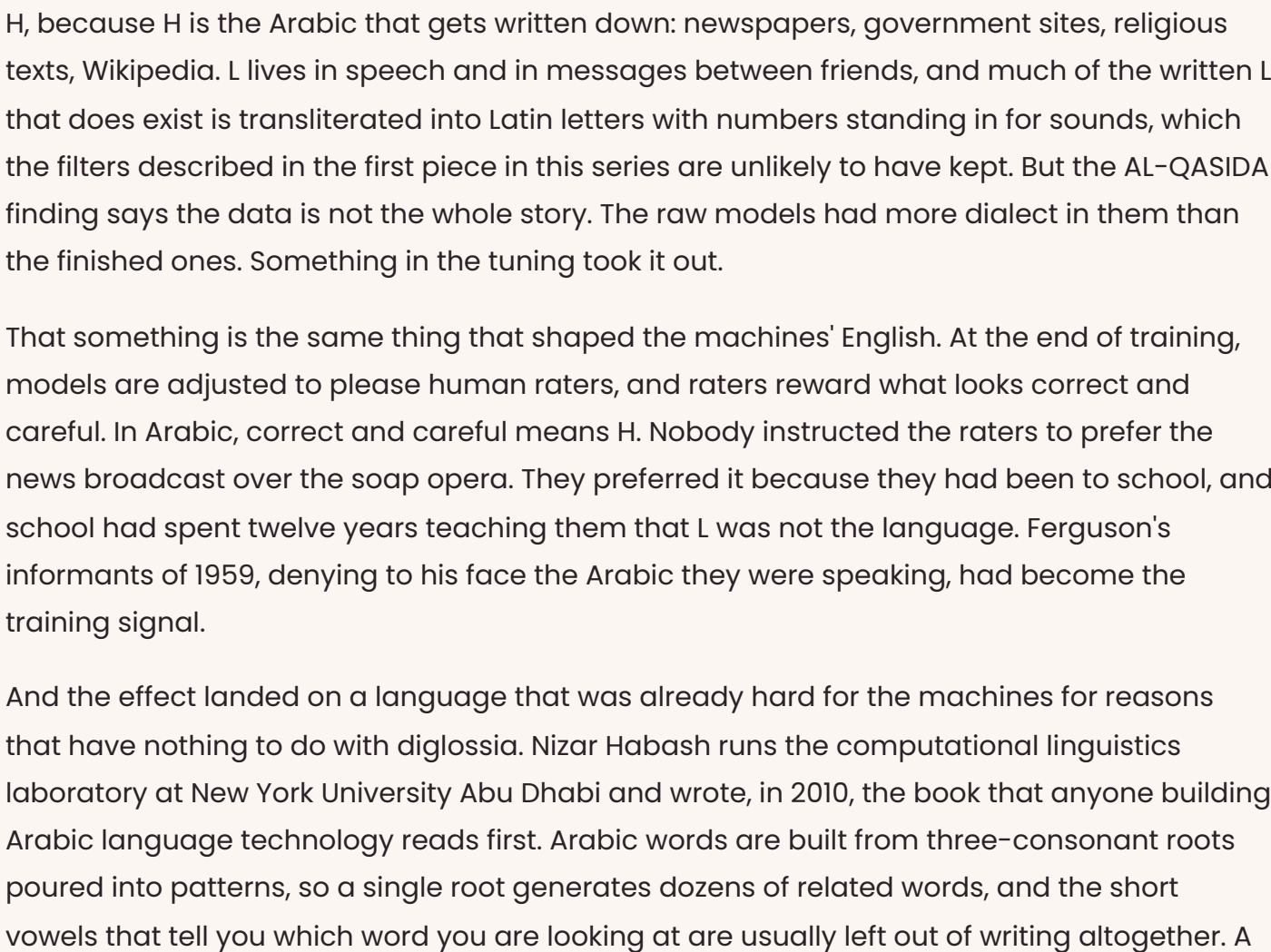


Fig. 2 — The AL-QASIDA pattern (Robinson et al., 2025), shown schematically. Comprehension runs far ahead of production — and the finished, polished versions of the models produced less dialect than the raw ones.

Then came the result that gives the study its point. Robinson and his colleagues compared the raw versions of the models, the ones straight out of training, with the finished versions that had been through the final stage of tuning, the process that makes a model helpful and polite and safe. The finished versions were worse at dialect than the raw ones. The polishing had pushed them toward H. Whatever the models had picked up of spoken Arabic from the web, the last step of making them presentable had taught them to leave it out.

The team called their benchmark AL-QASIDA, and published it in 2025. A second group at Qatar Computing Research Institute, led by Firoj Alam, reached a similar place by a different road with a benchmark called AraDiCE. The Arabic-centric models built in the region did better on dialect than the multilingual ones from California, but every model was weak at producing it. The pattern was Ferguson's pattern. The machines had learned to understand L and to insist on H, and, like his educated informants, they would not speak the language they plainly knew.

Roots and Vowels

Why would that be? Part of the answer is in the data. Arabic on the internet is overwhelmingly H, because H is the Arabic that gets written down: newspapers, government sites, religious texts, Wikipedia. L lives in speech and in messages between friends, and much of the written L that does exist is transliterated into Latin letters with numbers standing in for sounds, which the filters described in the first piece in this series are unlikely to have kept. But the AL-QASIDA finding says the data is not the whole story. The raw models had more dialect in them than the finished ones. Something in the tuning took it out.

That something is the same thing that shaped the machines' English. At the end of training, models are adjusted to please human raters, and raters reward what looks correct and careful. In Arabic, correct and careful means H. Nobody instructed the raters to prefer the news broadcast over the soap opera. They preferred it because they had been to school, and school had spent twelve years teaching them that L was not the language. Ferguson's informants of 1959, denying to his face the Arabic they were speaking, had become the training signal.

And the effect landed on a language that was already hard for the machines for reasons that have nothing to do with diglossia. Nizar Habash runs the computational linguistics laboratory at New York University Abu Dhabi and wrote, in 2010, the book that anyone building Arabic language technology reads first. Arabic words are built from three-consonant roots poured into patterns, so a single root generates dozens of related words, and the short vowels that tell you which word you are looking at are usually left out of writing altogether. A reader supplies them from context. A machine has to guess.

For children, the vowels are written in. Every early reader, every Quran, every primary textbook carries the full set of marks above and below the letters, because a six-year-old cannot supply them from context yet. So the natural thing, if you wanted a machine to help a six-year-old read, would be to give it the vowelised text. In March of this year Habash, Go Inoue, Bashar Alhafni and Timothy Baldwin reported what happens when you do. The models split the vowelised words into far more pieces than the unwoulded ones, and their performance on standard tests fell. The text written for children to read is the text the models read worst.

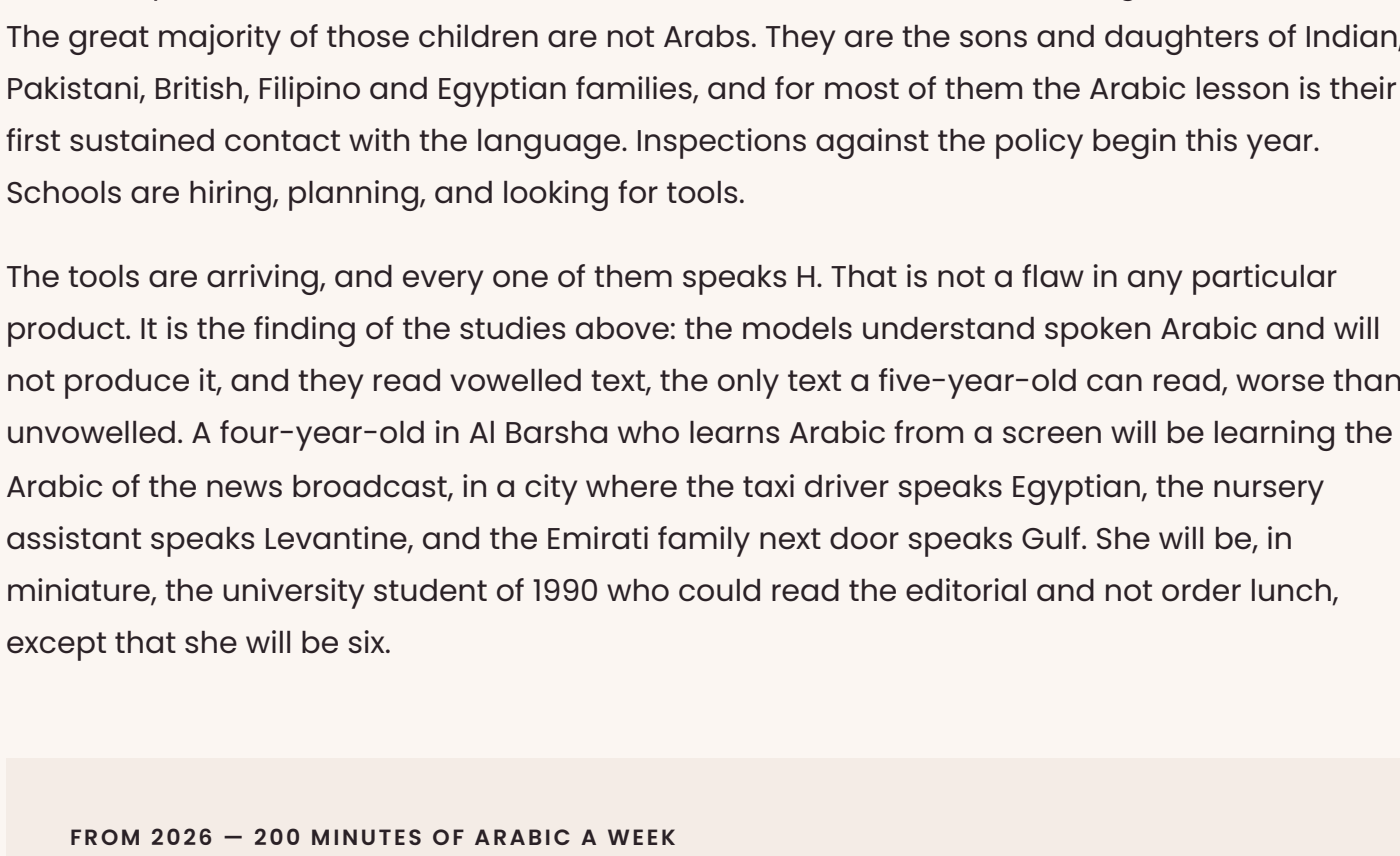


Fig. 3 — The same sentence, tokenized. Models chop vowelised words into far more pieces than unwoulded ones, and performance falls (Inoue, Alhafni, Habash & Baldwin, 2026). The text a five-year-old can read is the text the models read worst.

Catherine Arnett and Benjamin Bergen at the University of California, San Diego, asked the broader question in 2025: why do language models do badly on languages with rich morphology at all? Their answer was that morphology was largely a bystander. The models were struggling because their tokenisers, the components that chop text into units before anything is learned, split such languages into many more fragments, and because there was less text to learn from per language. Arabic was not too complicated for the machines. It had been chopped up badly and fed to them in the wrong variety.

One Language

While the machines were settling on H, the people who teach Arabic had spent twenty years moving in the opposite direction.

For most of the twentieth century, Arabic was taught to foreigners the way Ferguson's informants would have wanted: H only, the language of the newspaper and the lecture hall, with L treated as a regrettable local habit the student might pick up later, by accident, in a taxi. The results were well known and rarely said aloud. Students spent four years learning on Arabic that no one would speak to them, arrived in Cairo or Amman or Dubai, and inouere they could read the editorial and not order lunch.

Munther Younes at Cornell was one of the first to break with this, building a course in the 1990s that taught Levantine dialect and formal Arabic together from the first week. Mahmoud Al-Batal, then at the University of Texas, gathered the argument into a book in 2018 whose title states its case: Arabic as One Language. The dialect is not a corruption of the standard, the argument runs, and the standard is not a purer form of the dialect. They are one repertoire, used by one speaker, and a learner should be taught to move between them the way a native speaker does. The approach is now mainstream in the strongest university programmes and is spreading into schools.

"The pedagogy had spent sixty years catching up with how Arabic is actually used. The technology, in three years, put the old rule back."

So the sequence, put side by side, is this. In 1959 Ferguson described two Arabics and the denial that kept them apart. Over the following decades, teachers slowly stopped denying, and began teaching both. Then the machines arrived, learned both from the web, and were tuned by a process that taught them to deny one. The pedagogy had spent sixty years catching up with how Arabic is actually used. The technology, in three years, put the old rule back.

Two Hundred Minutes

Here is where this stops being a story about linguistics and becomes a story about a policy.

In 2025 the Knowledge and Human Development Authority in Dubai issued its policy on Arabic in early childhood education. From this school year, children aged four to six in the emirate's private schools are to receive 200 minutes of Arabic a week, rising to 300 from 2027. The great majority of those children are not Arabs. They are the sons and daughters of Indian, Pakistani, British, Filipino and Egyptian families, and for most of them the Arabic lesson is their first sustained contact with the language. Inspections against the policy begin this year. Schools are hiring, planning, and looking for tools.

The tools are arriving, and every one of them speaks H. That is not a flaw in any particular product. It is the finding of the studies above: the models understand spoken Arabic and will not produce it, and they read vowelised text, the only text a five-year-old can read, worse than unwoulded. A four-year-old in Al Barsha who learns Arabic from a Egyptian, the nursery assistant speaks Levantine, and the Emirati family next door speaks Gulf. She will be, in miniature, the university student of 1990 who could read the editorial and not order lunch, except that she will be student.

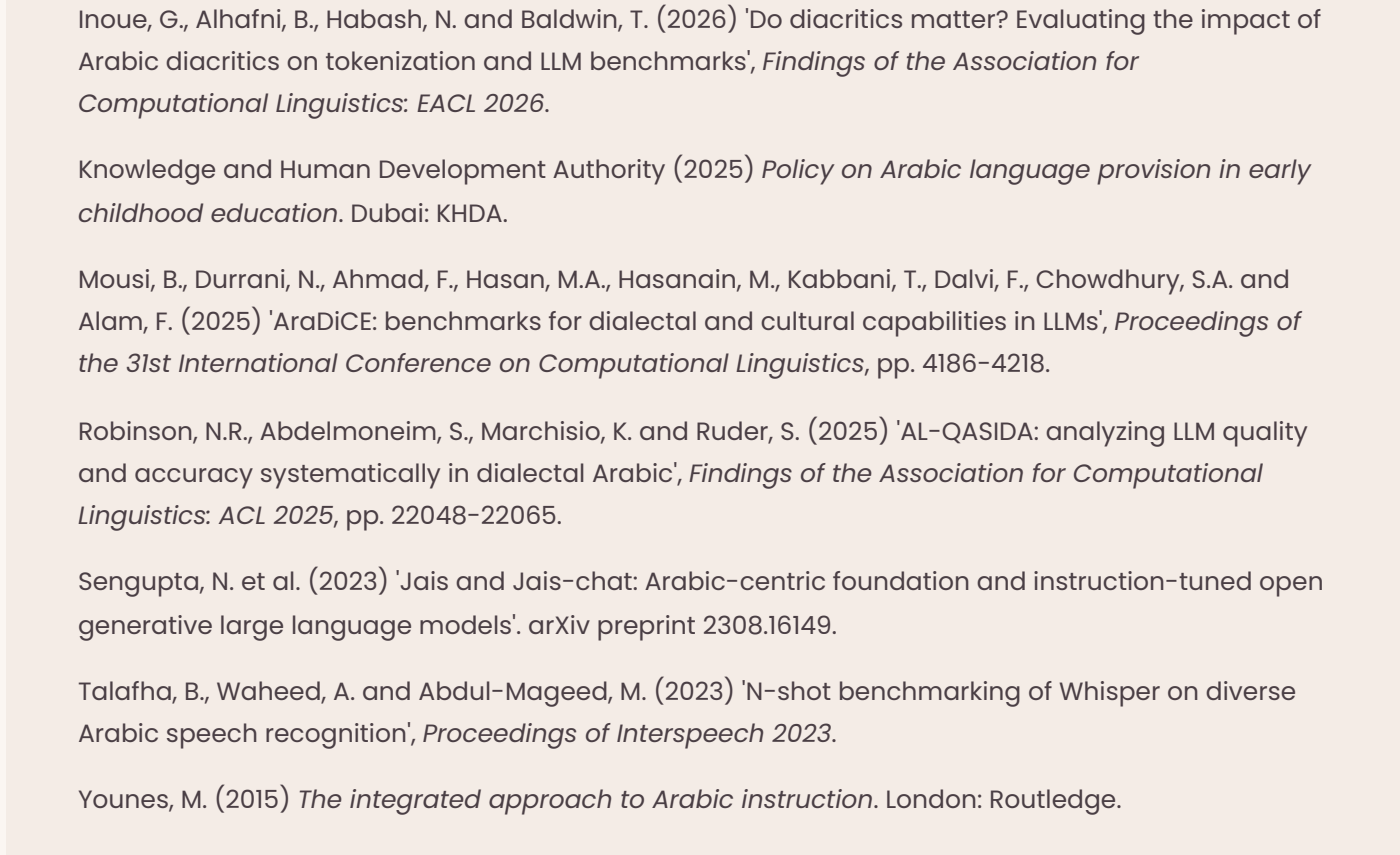


Fig. 4 — The KHDA policy on Arabic in early childhood education (2025): ages four to six, in Dubai's private schools, most of whom are meeting the language for the first time. No published study anywhere has tested an AI Arabic tool with this age group.

The speaking side is worse still. Bashar Talafha and colleagues at the University of British Columbia tested Whisper, the speech recognition system underneath most AI speaking tools, on Arabic in 2023, and found that its accuracy fell sharply the moment the speech left formal Arabic for dialect. So a child who speaks to an AI tutor in the Arabic she hears around her will be poorly heard, and then answered in an Arabic she does not hear anywhere.

I should say what has and has not been studied. Ahmed Allaihy and Mona Zaki, in a paper this year in Computer Assisted Language Learning, had experienced teachers rate AI-generated Arabic reading materials and found them usable with correction, which is real evidence and the closest thing to a classroom trial that exists. It was conducted with university students. There is no published study anywhere of an AI Arabic tool used with children aged four to six, which is precisely the group the Dubai policy covers. At LEAP we are building that study now, with the two schools in our panel. The policy has created the largest experiment in early Arabic acquisition the region has seen, and nobody is yet measuring what the machines do inside it.

Three Tests

None of this is an argument against the tools. It is an argument for knowing which Arabic they speak before a school puts one in front of a child. Three tests, each of which a Head of Arabic can run in an afternoon.

The first is Ferguson's test. Take any tool that claims to teach Arabic and speak to it in dialect, the way a parent would speak to a child: kayf halak, shu ismak, wein rayeh. Then look at what comes back. If the answer is in the Arabic of the news broadcast, the tool has passed Robinson's comprehension test and failed his production test, and it will teach a child to understand her neighbours and never to speak to them. A tool that answers in dialect, or that can be told which dialect to use and holds to it, is rare, and worth paying for.

The second is Habash's test. Give the tool the text a child actually reads: fully vowelised, every mark in place. Ask it to read it aloud, to explain a word, to write a sentence in the same style. Then take the marks off and ask again. If the tool is visibly better without the vowels, it was built for adults and pointed at children, and it will be least reliable on exactly the text the youngest learners need.

The third is Al-Batal's test, and it is about the school rather than the software. Does the Arabic programme treat dialect as part of the language, or as a distraction from it? A school that has adopted the integrated approach will find the machines pulling against its curriculum every day, and will need to teach children, as the first piece in this series argued for English, to notice when the tool is correcting a feature rather than an error. A school that still teaches H only will find the tools agreeable, and will be reproducing, with better graphics, the classroom Ferguson described in 1959.

One more thing worth saying, because the story is too easily told as the West building machines without the Arab world. The best-performing models on dialect in both benchmarks were built in the Gulf: Jais in Abu Dhabi, AllLaM in Riyadh, Fanar in Doha. Fanar's team went so far as to build a tokeniser that respects Arabic roots. The region is not a bystander here. The honest version of the criticism is that its own models still cannot do what a five-year-old needs: hear dialect, speak it back, and read vowelised text well. Those are engineering problems with a known shape, and the region is better placed than anyone to solve them, if it knows they are there.

The Informants

Ferguson died in 1998, and he did not live to see a machine take up the position of his informants. But he would have recognised it at once. Ask the model what it thinks of dialect and it will tell you, in careful H, that dialects are a rich part of the Arab heritage. Speak to it in dialect and it will answer in the Arabic of the sermon, as though you had not. It knows L. It has been taught, in the last step of its making, that L is not the language.

A child in Dubai, four years old, does not know any of this. She knows the Arabic she hears at the school gate, and from this year she will spend two hundred minutes a week learning another one. What she gets from the machine in those minutes is a choice someone else has already made, in a tuning process in another country, without a linguist in the room. The work of a school is to know that the choice was made, to test for it, and, where the tool speaks only the Arabic nobody speaks, to make sure she also hears the one everybody does.

Sources referred to in this piece

Al-Batal, M. (ed.) (2018) *Arabic as one language: integrating dialect in the Arabic language curriculum*. Washington, DC: Georgetown University Press.

Allaihy, A. and Zaki, M. (2025) 'Evaluation of AI-generated reading comprehension materials for Arabic language teaching', *Computer Assisted Language Learning*.

Arnett, C. and Bergen, B. (2025) 'Why do language models perform worse for morphologically complex languages?', *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6607–6623.

Bari, M.S. et al. (2025) 'ALLaM: large language models for Arabic and English', *Proceedings of ICLR 2025*.

Fanar Team (2025) 'Fanar: an Arabic-centric multimodal generative AI platform'. arXiv preprint 2501.13944.

Ferguson, C.A. (1959) 'Diglossia', *Word*, 15(2), pp. 325–340.

Habash, N.Y. (2010) *Introduction to Arabic natural language processing*. San Rafael, CA: Morgan & Claypool.

Inoue, G., Alhafni, B., Habash, N. and Baldwin, T. (2026) 'Do diacritics matter? Evaluating the impact of Arabic diacritics on tokenization and LLM benchmarks', *Findings of the Association for Computational Linguistics: EACL 2026*.

Knowledge and Human Development Authority (2025) *Policy on Arabic language provision in early childhood education*. Dubai: KHDA.

Moussi, B., Durrani, N., Ahmad, F., Hasan, M.A., Hasanain, M., Kabbani, T., Dalvi, F., Chowdhury, S.A. and Alam, F. (2025) 'AraDiCE: benchmarks for dialectal and cultural capabilities in LLMs', *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4186–4218.

Robinson, N.R., Abdelmoneim, S., Marchisio, K. and Ruder, S. (2025) 'AL-QASIDA: analyzing LLM quality and accuracy systematically in dialectal Arabic', *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22048–22065.

Sengupta, N. et al. (2023) 'Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models'. arXiv preprint 2308.16149.

Talafha, B., Waheed, A. and Abdul-Mageed, M. (2023) 'N-shot benchmarking of Whisper on diverse Arabic speech recognition', *Proceedings of Interspeech 2023*.

Younes, M. (2016) *The integrated approach to Arabic instruction*. London: Routledge.

