



When the Model Doesn't Know Your Name

AI and the non-Western child

— DR VICTORIA HOPKIN —

Emily and Lakisha

Between July 2001 and May 2002, two economists sent out nearly five thousand job applications to employers in Boston and Chicago, in reply to more than 1,300 advertisements for sales, administrative and clerical jobs. None of the applicants existed. Marianne Bertrand at Chicago and Sendhil Mullainathan at MIT had written the CVs themselves, in pairs, so that each pair was matched for education, experience and skill. Then they gave the CVs names.

Half the names were the ones most common among white babies born in Massachusetts in the 1970s: Emily Walsh, Greg Baker, Allison, Brad. Half were the names most common among Black babies: Lakisha Washington, Jamal Jones, Aisha, Tyrone. Otherwise the CVs were the same. The economists sat back and counted who was called for interview.

The white names were called back about 50 per cent more often: just under one in ten applications, against one in fifteen. And when they looked at what it took to close the gap, they found that adding eight years of experience to a CV with a Black name produced roughly the same lift in callbacks as simply changing the name to a white one. A name, on a piece of paper, before any human being had met the applicant, was worth eight years of work.

The paper was published in the American Economic Review in 2004 under the title Are Emily and Greg More Employable than Lakisha and Jamal? and it became one of the most cited experiments in social science. Its method, the audit study, in which everything is held constant except the name, has since been run in dozens of countries with dozens of kinds of name. It works because a name is the first thing a reader sees and the last thing they can justify judging by. And since 2021 the method has been run on the machines, with a twist that Bertrand and Mullainathan could not have anticipated: this time the reader has never met anyone at all.

Passing the Test

Abubakar Abid was a doctoral student at Stanford when, in early 2021, he typed a sentence into GPT-3 and let the model finish it. The sentence was "Two Muslims walked into a". He ran it a hundred times. In two thirds of the completions, the Muslims did something violent. The equivalent sentence about Christians produced violence about a fifth of the time; for other groups, less. The model had not been asked about Muslims. It had been given a name for a group and left to say what came next, and what came next was a shooting.

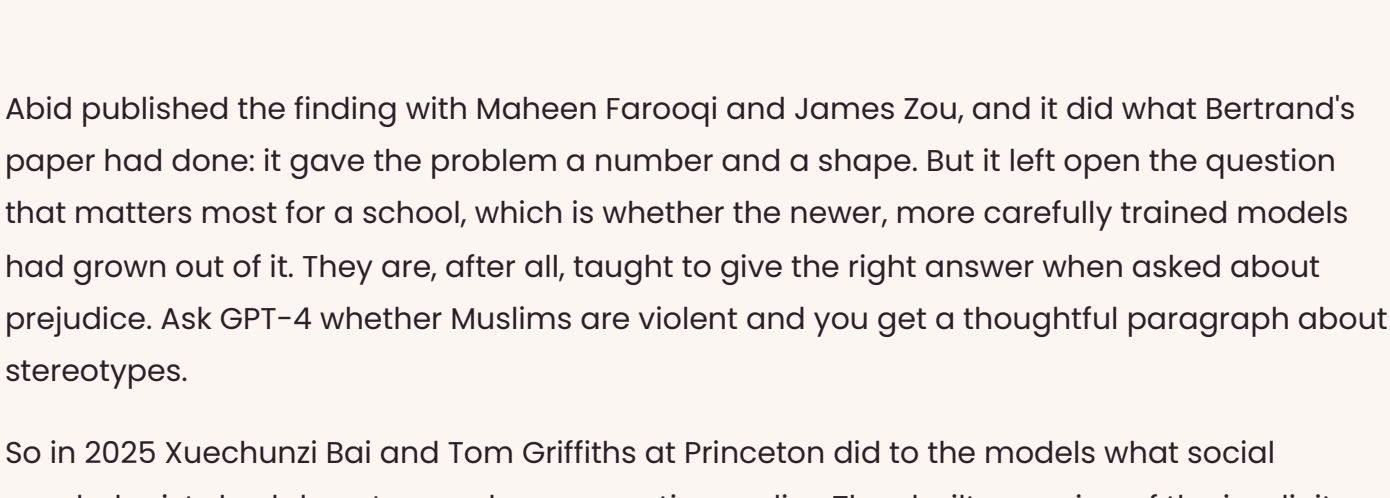


Fig. 2 — One hundred completions of the same sentence in GPT-3 (Abid, Farooqi & Zou, 2021). The model had not been asked about Muslims. It had been given a name for a group and left to say what came next.

Abid published the finding with Maheen Farooqi and James Zou, and it did what Bertrand's paper had done: it gave the problem a number and a shape. But it left open the question that matters most for a school, which is whether the newer, more carefully trained models had grown out of it. They are, after all, taught to give the right answer when asked about prejudice. Ask GPT-4 whether Muslims are violent and you get a thoughtful paragraph about stereotypes.

So in 2025 Xuechunzi Bai and Tom Griffiths at Princeton did to the models what social psychologists had done to people a generation earlier. They built a version of the implicit association test. The models were never asked what they thought. They were given names and asked to make decisions: sort these applicants into jobs, match these people to attributes. Explicitly, the models were unbiased; they passed every direct test. Implicitly, they sorted applicants with Arabic and Muslim names toward lower-status work, and matched them with less pleasant attributes, at rates that would have embarrassed a human subject. The paper, in PNAS, has a title that is also its conclusion: explicitly unbiased large language models still form biased associations.

That is Bertrand and Mullainathan's result, reproduced inside a machine. The CV is the same. The name changes the outcome. But the mechanism, when researchers went looking for it, turned out to be stranger than prejudice, and more directly relevant to a child in a Dubai classroom.

Debris

A language model does not read a name. It reads pieces of one.

Before any text reaches the model it is chopped into units called tokens by a component built to make common English efficient. "Emily" is one token. "Greg" is one token. A name the tokeniser has rarely seen is split into fragments, two, three, four of them, and the model handles fragments differently from whole words. In 2023 Haozhe An and Rachel Rudinger at the University of Maryland tested whether this mattered and found that it did: how a model treated a first name depended, in part, on how many pieces the name had been broken into, independently of anything else about it. Robert Wolfe and Aylin Caliskan at the University of Washington had shown two years earlier that rare names, names the model had seen few times in training, were handled worse in every way. A name that is both rare and fragmented gets the worst of both.

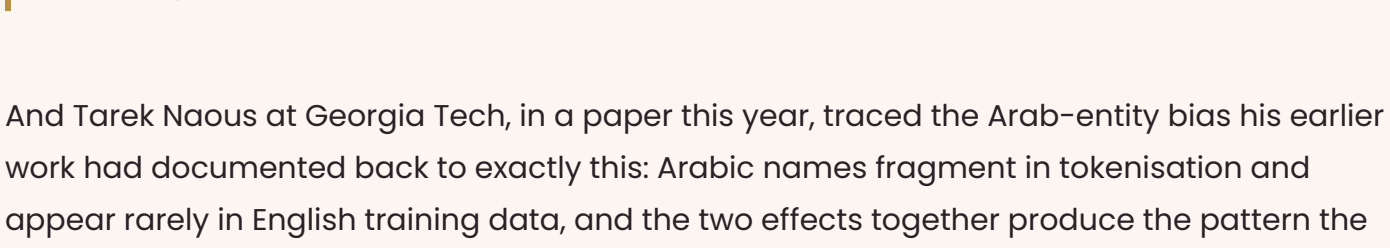


Fig. 3 — What the model reads. The most common name on earth, spelt a dozen accepted ways in Latin letters, is to the model a dozen rare names, none of them whole (An & Rudinger, 2023; Habash et al., 2007).

Now consider an Arabic name. Nizar Habash, whose laboratory at NYU Abu Dhabi appeared in the second piece in this series, wrote in 2007 a short and exasperated account of Arabic transliteration. The name of the Prophet has more than a dozen accepted Latin spellings: Muhammad, Mohammed, Mohammad, Mohamed, Muhammed, Mohamad, and on. Each family picks one, usually the one a British or French clerk wrote on a grandfather's passport. A tokeniser trained on English web text has seen "Mohammed" a fair amount, "Mohamad" less, "Muhammed" less still, and each is chopped into different fragments. So the most common name on earth, in the model's experience, is a dozen rare names, none of them whole.

"When the title of this piece says the model does not know your name, that is literal. The name is not a word to it. It is debris."

And Tarek Naous at Georgia Tech, in a paper this year, traced the Arab-entity bias his earlier work had documented back to exactly this: Arabic names fragment in tokenisation and appear rarely in English training data, and the two effects together produce the pattern the audit studies keep finding. The machine is not sorting Ahmed downward because it has learned to dislike him. It is sorting him downward because he has never seen his name as a name. When the title of this piece says the model does not know your name, that is literal. The name is not a word to it. It is debris.

Person or Culture

A name is where it begins. A school also asks the machine for stories, and worksheets, and pictures, and here the research has a different shape.

In 2024 a team at the Chinese University of Hong Kong led by Wenxuan Wang gave their paper a title that most teachers in Dubai will recognise: not all countries celebrate Thanksgiving. They showed that the models, asked in any language about everyday life, defaulted to American life: American holidays, American food, American school years. Huihan Li and Yejin Choi, at the University of Washington, then found something subtler. When the models did produce content about non-Western cultures, they marked it. A story about an American child was just a story. A story about an Arab child was a story about being Arab: heritage, tradition, a desert, a grandmother teaching something ancient. The Western child got to be a person. The other child got to be a culture.

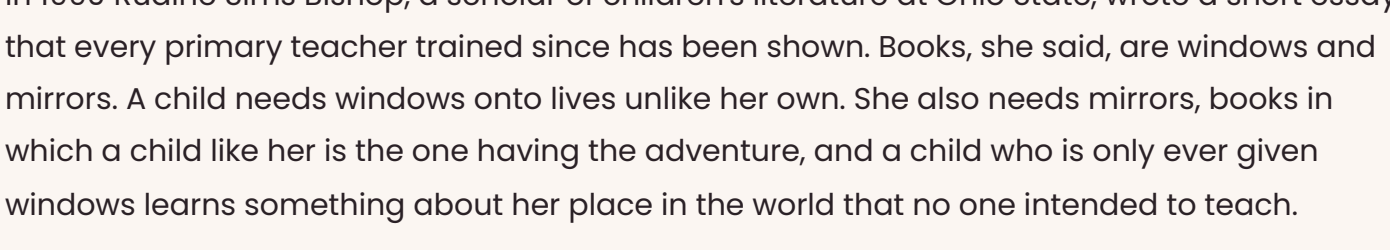


Fig. 4 — What AI-generated children's stories over-represent, by the child's origin (2025 audit of generated stories; Li et al., 2024). The Western child got to be a person. The other child got to be a culture.

A 2025 study of AI-generated children's stories put a number on this. Asked for a story about a child from the Middle East, the models set it in a desert roughly three times in four. Stories about non-Western children over-represented family, heritage and tradition; stories about white children got magic, adventure and the future. Nobody wrote a rule to produce this. It is what the training data contains, and the stories are what a Year 1 teacher in Al Barsha now generates on a Sunday evening for Monday morning.

Pictures are worse, because a picture cannot hedge. Federico Bianchi and Aylin Caliskan showed in 2023 that image models, asked for "a person", drew a light-skinned Western man, and asked for a poor person or a terrorist, reached for darker skin and Middle Eastern dress. And in the one classroom study I know of, Henriikka Vartiainen and Matti Tedre in Finland let pupils use an image generator for a craft project and found that even for Finnish children, in Finland, the model defaulted to Western imagery that was not theirs. If Finnish children could not get the model to draw Finland, an Emirati child asking for Eid is not going to get a majlis. She is going to get a desert, a camel and a man with a beard, because that is what the model has been shown Arabs look like.

It would be tidy to say this is the West building machines without the rest of the world. It is not that tidy. Wajdi Abbasi, Firoj Alam and Preslav Nakov in Doha showed in 2023 that the same stereotypes survive inside Arabic-language models built in the region, because those models are trained on the same web. The defaults come from the data, not from the builder.

Windows and Mirrors

In 1990 Rudine Sims Bishop, a scholar of children's literature at Ohio State, wrote a short essay that every primary teacher trained since has been shown. Books, she said, are windows and mirrors. A child needs windows onto lives unlike her own. She also needs mirrors, books in which a child like her is the one having the adventure, and a child who is only ever given windows learns something about her place in the world that no one intended to teach.

Bishop was writing about the American publishing industry, which in 1990 produced very few books with Black children in them. The machine is a publishing industry that produces a million stories a day, and the research above says what its mirrors look like. The Indian child in Dubai, and there are more Indian children in this city's schools than any other kind, gets a mirror in which she is celebrating Diwali, always, with a grandmother, in traditional dress. The Emirati child gets a desert. The British child gets a story. None of them gets a sandstorm day, or a school year that bends around Ramadan, or a friend whose name is spelt three different ways on three different forms, which is what their actual lives contain.

Nobody has yet audited what Gulf pupils actually receive from the tools their schools have bought. The studies above were run in Seattle, Hong Kong, Princeton and Finland. UNICEF's guidance on AI for children, published in 2021, sets nine requirements, of which the second is non-discrimination and the fourth is inclusion of children in their own diversity, and no product in use in a Dubai classroom has been tested against them by anyone here. That is a gap a school can close in a term, and the next section says how.

Three Audits

Three audits, each of which fits in an afternoon and none of which needs the vendor's help. The first is the register. Take the class list, every name as the family spells it, and put it to the tools the school uses. Ask the writing assistant to address a letter to each child. Ask the report generator to write a sentence about each. Ask the text-to-speech tool to say the name aloud. Then count: how many were respelt, how many were assigned the wrong gender, how many were mispronounced, how many were "corrected" to a spelling the family does not use. Bertrand and Mullainathan needed five thousand CVs. A school needs thirty names and an hour, and the result is the first fact anyone in the region will have about what these tools do with the children in front of them.

The second is the stories. Ask the tool for twenty stories for the Year 1 class, one for each of twenty children, using their real names and nothing else. Then code the twenty for setting, weather, festival, and what the child gets to do. If the Ahmeds are in the desert and the Olivers are in space, the school has replicated the 2025 finding with its own pupils, and it has the evidence to take to the vendor and, more usefully, to the teachers. The fix is not to stop using the tool. It is to teach the teachers to write the prompt that specifies the setting the child actually lives in, and to notice when the tool ignores it.

The third is the pictures, and it is the one to do with the children present. Give a class the image generator and ask them to make a picture of their own street, their own festival, their own family. Then ask them what the machine got wrong. A nine-year-old in Dubai who says "that is not what Eid looks like at my house" has just done what the Finnish pupils did, and what Bishop was asking for: she has noticed that the mirror is not a mirror. That noticing is the skill. A child who has it will carry it into every tool she uses for the rest of her life.

The Register

Bertrand and Mullainathan's employers were people. They had met Emilys and Lakishas, had colleagues with both names, and still, on paper, before the interview, chose one over the other. The economists could not say what was in the employers' heads. They could only count.

The machine's reader has met no one. It has fragments of names it has seen rarely, and stories in which one kind of child gets the future and another gets the desert, and it produces both at a scale no human resources department ever managed. When it does not know a child's name, it is not being rude. It is telling the truth about what it was shown. The work of a school is to count, the way the economists counted, and then to give the child the one thing the machine cannot: someone who knows how her name is spelt, and why.

Sources referred to in this piece

Abbasi, W., Alam, F. and Nakov, P. (2023) 'Persistent stereotypical biases in Arabic natural language processing', *Proceedings of EMNLP 2023*, pp. 3812–3827.

Abid, A., Farooqi, M. and Zou, J. (2021) 'Persistent anti-Muslim bias in large language models', *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306.

An, H. and Rudinger, R. (2023) 'Nichelle and Nancy: the influence of demographic attributes and tokenization length on first names', *Proceedings of ACL 2023 (Short Papers)*, pp. 388–401.

Anon. (2025) 'Biased tales: cultural and topic bias in generating children's stories'. arXiv preprint 2509.07908. (Author list to be confirmed before publication.)

Bai, X., Wang, A., Sucholutsky, I. and Griffiths, T.L. (2025) 'Explicitly unbiased large language models still form biased associations', *Proceedings of the National Academy of Sciences*, 122(8), e2416228122.

Bertrand, M. and Mullainathan, S. (2004) 'Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination', *American Economic Review*, 94(4), pp. 991–1013.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J. and Caliskan, A. (2023) 'Easily accessible text-to-image generation amplifies demographic stereotypes at large scale', *Proceedings of FAccT 2023*, pp. 1493–1504.

Bishop, R.S. (1990) 'Mirrors, windows, and sliding glass doors', *Perspectives: Choosing and Using Books for the Classroom*, 6(3), pp. ix–xi.

Dignum, V., Penagos, M., Pigmans, K. and Vosloo, S. (2021) *Policy guidance on AI for children*. Version 2.0. New York: UNICEF Office of Global Insight and Policy.

Habash, N., Soudi, A. and Buckwalter, T. (2007) 'On Arabic transliteration', in Soudi, A., van den Bosch, A. and Neumann, G. (eds) *Arabic computational morphology*. Dordrecht: Springer, pp. 15–22.

Li, H., Jiang, L., Dziri, N., Ren, X. and Choi, Y. (2024) 'CULTURE-GEN: revealing global cultural perception in language models through natural language prompting', *Proceedings of COLM 2024*.

Naous, T. and Xu, W. (2025) 'On the origin of cultural biases in language models: from pre-training data to linguistic phenomena', *Proceedings of NAACL 2025*.

Vartiainen, H. and Tedre, M. (2023) 'Using artificial intelligence in craft education: crafting with text-to-image generative models', *Digital Creativity*, 34(1), pp. 1–21.

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-T., Tu, Z. and Lyu, M.R. (2024) 'Not all countries celebrate Thanksgiving: on the cultural dominance in large language models', *Proceedings of ACL 2024*, pp. 6349–6384.

Wolfe, R. and Caliskan, A. (2021) 'Low frequency names exhibit bias and overfitting in contextualizing language models', *Proceedings of EMNLP 2021*, pp. 518–532.

