Futures

What idioms teach us about AI and its evolution

Research into how AI interprets idioms into images and text reveals a lot about the technology – and us, finds Nicole Kobie

Counting the instances of the letter "r" in "strawberry". Generating an image to show kicking the bucket, monkey business or another idiom – or *no* elephants in a room. These are all examples of viral pranks played on large language models (LLMs), and they reveal plenty about how these systems work, when they don't and how they're slowly getting better.

When it comes to idioms, LLMs perform much better when sticking with text than they do with images. Ask ChatGPT et al to explain a phrase in text, and you're likely to receive a detailed and even accurate answer.

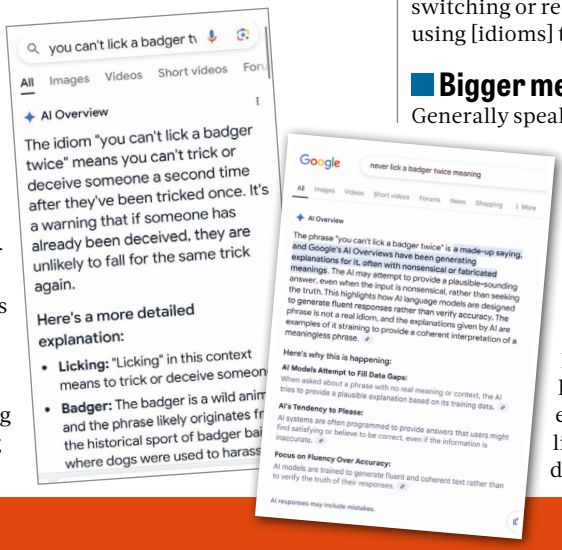
But it's still possible to trick AI by making up idioms. On social media, one suggestion was: "You can't lick a badger twice." It felt real enough to fool Google's AI Overview (see "Make your own idiom", p28), but ChatGPT 4o returns a more sensible result, saying it's not a common phrase, suggesting a few potential meanings from similar idioms and asking for more information, which is as close as these LLMs get to saying they don't know.

Understanding LLMs

Tom Pickard, a researcher at the University of Sheffield, studies how AI manages idioms, in part to help develop systems that are better at deciphering the weird figures of speech created by humans. It turns out that idioms are a great way to help people understand how LLMs work, and why they sometimes don't.

That really comes down to the fact that these systems don't "understand" our language. In fact, they rip it apart, reducing words and letters into tokens devoid of any meaning. "The tokens, the words are broken up in training, it doesn't process text, it processes multi-dimensional vectors and numbers," Pickard said at the Turing Institute's AI conference.

Now, such examples are trickery on our behalf. We humans are trying to fool the AI into returning a silly result. Beyond creating fodder for social media posts or amusing articles on technology sites, this does have a real purpose: we can learn how LLMs work, or don't. Doing so in an entertaining way is a bonus.



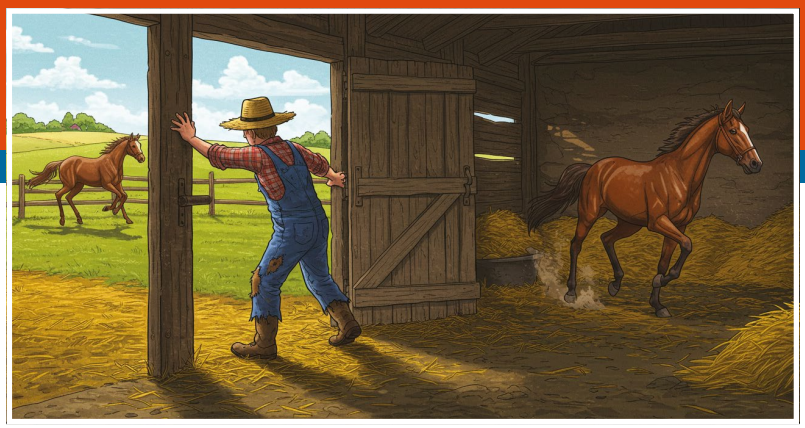
BELOW That well-known idiom "you can't lick a badger twice"

As Pickard notes, the success of "pranks" such as asking Google's AI Overviews about licking badgers comes down to the fact that computer scientists and engineers didn't consider that people would try to trick their tools. "If you talk to Siri or the Google voice-activated tech and put on a silly voice or an exaggerated accent it may not recognise what you're saying very well," he told me. "To an extent this doesn't really matter and there are more important edge-cases to focus on – if the tool doesn't recognise your actual accent or speech patterns, pick up on you code-switching or recognise that you're using [idioms] that's a bigger deal."

Bigger means better?

Generally speaking, the belief is that LLMs will get better as they get bigger. That does apply to idioms, but it's not as straightforward as it may seem, Pickard notes.

"My colleagues and I have a paper from last year where we explored using LLMs like ChatGPT for idiom detection and they do



ABOVE Shutting the stable door after the horse bolted, as imagined by AI

get better with size, but they don't keep up with humans or with smaller models which have been tuned to this particular task," he said.

More recent models, including DeepSeek, have included a technique called reasoning that has improved outputs without having to necessarily build ever bigger LLMs, although they're still pretty huge. But that technique doesn't always improve idioms, either.

"The model needs to be able to produce an accurate definition of an idiom before it can 'reason' about whether that definition applies in a particular context," Pickard explained.

He points to how humans unpick a strange phrase, perhaps saying in a meeting "shutting the stable door after the horse bolted". A colleague who had never heard that before would understand you weren't talking about farm animals, and probably be able to guess what you meant.

"That kind of meta-reasoning just isn't present in the AI outputs, and that doesn't surprise me at all, because what they're doing doesn't actually resemble reasoning in a meaningful sense – they're producing vaguely reasoning-shaped text, not actually thinking through a problem," he said. "It's fascinating that this process seems to increase the chance of them getting the right answer on lots of benchmarks, but I think it's a bit disingenuous to call it 'reasoning', and it doesn't appear to be helping them very much with idiom processing."

Quick changes

Beyond idioms, there have been a few cases of LLM mix-ups going viral. In one early example, ChatGPT was asked how many instances of the letter "r" were in "strawberry" – and it simply couldn't accept there were three. In another, OpenAI's image generator DALL-E was asked to create an image of a room without an elephant, but every room prominently featured a pachyderm.

When such examples hit the headlines, the systems are often immediately tweaked to avoid such issues in the future.

"When these kinds of viral weaknesses are fixed rapidly, I'm pretty sure the developers are patching something in the interface layer rather than updating the models themselves," Pickard said. "You'll get a better answer now about the number of 'r's in 'strawberry' than you once did, not because something has fundamentally been changed about the input processing but because that interface either includes a kind of hard-coding, or maybe because some sort of module has been added to allow for input text to be analysed at the character level."

And that makes it hard to study this aspect of LLMs, though the nature of LLMs means variation is inherent in outputs. "All of this... makes these large, opaque systems pretty poor subjects for scientific study in a lot of ways," said Pickard. "If I publish something which says that a shiny new model is great at interpreting sarcasm, we don't know what it is that makes that model good at this task, and you're unlikely to be able to replicate my study because there'll be a new version in the meantime."

He added: "From a scientific point of view, I think we can do more by studying smaller, more specialised and openly available models, and then potentially have what we learn be applied to the larger systems."

Either way, that's some of the joy of getting to use these systems while they're being developed: we get to see the cracks and watch them get patched over. It's an education in how these systems work; someone new to LLMs even now may find it harder to spot flaws now they aren't as obvious.

Generally speaking, the belief is that LLMs will get better as they get bigger

Why does it matter?

We need to know how LLMs work. Call it AI literacy, but in many ways it's an extension of digital or algorithmic understanding that many of us have failed to keep pace with over the past two decades.

The elephant in the room

Picture a room with no elephants. It's pretty easy for most of us – it's just any room, after all – but ask a LLM and the results tend to prominently include an elephant. Perhaps the pachyderm is peering in through a window, represented in a painting on the wall, or there's not one but two elephants.

What's happening? This reflects the fact that AI is trained by massive data sets, and when it comes to images the LLM relies on captions or labelling. A room with no elephant would simply be labelled as a room and wouldn't mention elephants in the caption. "It doesn't learn negative associations," said Pickard. "It learns to spot things that are mentioned in the caption."

This is a classic example, if there can be such a thing as a classic in toddler-aged technology. But here's where research in AI can be a challenge. Prompt ChatGPT for a room without an elephant, and it will spit out a classy but boring living room, with a few sofas and a coffee table, with no animals in sight. But ask Bing, which uses OpenAI's DALL-E 3, the most recent version of its image generation tool, and the room is once again full of elephants.

Ten years ago, the *topic du jour* for tech ethics wasn't AI bias but the risk of algorithmic black boxes. Simply put, if we don't know how an algorithm comes to a "decision", we don't know when it makes mistakes or how to make them more accountable or accurate. It's simply numbers in, answers out.

AI, and in particular deep learning, exacerbates that because we don't really know what the systems are "thinking", for lack of a better word. We set these models loose on huge sets of data in order to "learn", but it's often not clear why an AI thinks that an image is of a dog rather than a wolf, a classic example in which researchers eventually understood that their model couldn't tell the animals apart, but organised them via their environmental surroundings, such as on a sofa inside or in the wilderness surrounded by snow.

Image: ?????????????

Image: ?????????????

That lack of understanding what the system is doing matters when we start to apply AI to everything from sifting through CVs for job applications to deciding who stays in prison or is released or powering the entire civil service, as the UK government hopes to achieve. "The UK government, for whatever reason, is 'betting the farm' on using generative AI, and I'm not sure they should be," Pickard said. "It doesn't understand things – and I think this is where it becomes quite important."

Now, AI could be helpful, but anyone using such systems – or having such systems used on them – should understand the limitations. For example, using AI to transcribe GP appointments or benefits interviews could be beneficial, but it may also trip up on figures of speech or confusingly phrased language. And that matters when English isn't your first language.

"My go-to example is to imagine a chatbot of some sort in a healthcare setting – maybe it's trying to do something similar to the NHS 111 phone line and triage less severe problems, point people to the right services," Pickard said. "If I phone up or open the chat window and tell it I'm feeling 'blue', I should probably be signposted to mental health resources rather than treated for hypoxia."

ABOVE/LEFT The UK government is betting the farm on generative AI

We get to see the cracks and then watch them get patched over

And that's particularly important with other languages. "Translation is another good example – metaphors, idioms and other figurative language are some of the most challenging things for human translators to handle, and I think improving models' 'understanding' of them will lead to appreciably better automatic translations," he said.

Figuring out idioms and LLMs won't just help us understand AI better, it could help us understand each other better, too. ●

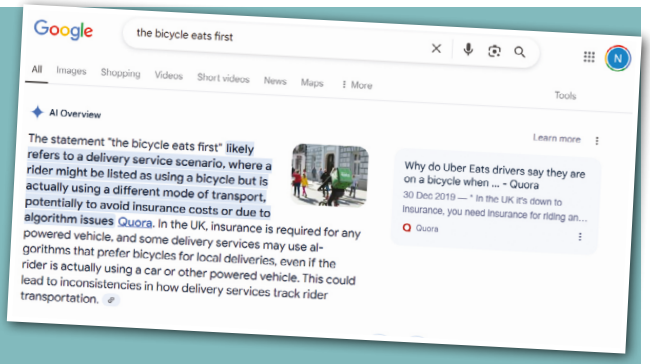
Make your own idiom

Google's AI Overview – the one that pops up unasked for at the top of you search results – until recently had a funny quirk that sadly seems to have disappeared. (Perhaps AI developers should fix the disinformation rather than the fun stuff.)

Type in an idiom followed by "meaning" or "explained", and AI Overview will generate a miniature explainer for you – very handy for helping anyone from a different cultural or language background.

For example, try searching for "raining cats and dogs idiom meaning and sentence". At the time of writing, Google's AI Overview comes up with this: "The idiom 'raining cats and dogs' means it is raining very heavily. It's≈a figurative way to describe a downpour that is unusually intense. Sentence example: 'We had to cancel the picnic because it was raining cats and dogs.'"

But what happens if you make up an idiom? In a Threads post, author Meaghan Wilson Anastasios shared her attempt of "peanut butter platform heels". Google's AI Overview decided that was a "reference to a scientific experiment where peanut butter was used to demonstrate the creation of diamonds under high pressure". (It is not.)



RIGHT It's fun testing Google's AI with made-up idioms

Other examples started doing the rounds, including "the bicycle eats first" (which apparently means you should prioritise nutrition when cycling or that food delivery services should prioritise using bicycles over cars or motorbikes) and "a loose dog won't surf", which AI Overview decided was a playful way of saying something isn't going to happen or work out. And then there's "never throw your poodle at a pig", which it interpreted as a humorous way of saying not to waste good things on someone who won't appreciate them. Apparently, it's derived from the biblical phrase about not casting pearls before swine.

Another that's popped up on social media is "you can't lick a badger twice", which AI Overview decided meant you can't trick someone a second time. But after the example phrase was included in a *Futurism* article, Google is now aware of it. The new AI-generated response says it's "a made-up saying, and Google's AI Overviews have been generating explanations for it, often with nonsensical or fabricated meanings". The response adds that this highlights how AI language models are "designed to generate fluent responses rather than verify accuracy". Indeed.