Natural Language Engineering (2021), **27**, pp. 113–118
doi:10.1017/S1351324920000601

CAMBRIDGE
UNIVERSITY PRESS

INDUSTRY WATCH

GPT-3: What's it good for?

Robert Dale

Language Technology Group
E-mail: rdale@language-technology.com

Abstract

GPT-3 made the mainstream media headlines this year, generating far more interest than we'd normally expect of a technical advance in NLP. People are fascinated by its ability to produce apparently novel text that reads as if it was written by a human. But what kind of practical applications can we expect to see, and can they be trusted?

1. Introduction

The mid-year release of OpenAI's GPT-3 language model, with its ability to generate natural language texts that can be remarkably hard to distinguish from human-authored content, was this year's big AI news item. It received coverage in both the technical and mainstream media far in excess of what you'd normally expect for a technical advance in NLP.

Here's a sample of headlines from the tech industry press:

- *ZDNet*, 1st June: 'OpenAI's gigantic GPT-3 hints at the limits of language models for AI'[a]
- *MIT Technology Review*, 20th July: 'OpenAI's new language generator GPT-3 is shockingly good – and completely mindless'[b]
- *Wired*, 22nd July: 'Did a person write this headline, or a machine? GPT-3, a new text-generating program from OpenAI, shows how far the field has come – and how far it has to go'[c]
- *The Verge*, 30th July: 'OpenAI's latest breakthrough is astonishingly powerful but still fighting its flaws'[d]

Reading just the titles gives an accurate flavour of the tone of this coverage: an acknowledgement of just how impressive the technology is, but tempered with a recognition of its limitations.

At least on the basis of the headlines, coverage in the mainstream media was a little more alarmist, expressing both awe and anxiety in response to GPT-3's capabilities:

- *BBC News*, 24th July: 'Have we seen our future?'[e]
- The New York Times, 29th July: 'How do you know a human wrote this? Machines are gaining the ability to write, and they are getting terrifyingly good at it'[f]

[a] https://www.zdnet.com/article/openais-gigantic-gpt-3-hints-at-the-limits-of-language-models-for-ai/
[b] https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/
[c] https://www.wired.com/story/ai-text-generator-gpt-3-learning-language-fitfully/
[d] https://www.theverge.com/21346343/gpt-3-explainer-openai-examples-errors-agi-potential
[e] https://www.bbc.com/news/technology-53530454
[f] https://www.nytimes.com/2020/07/29/opinion/gpt-3-ai-automation.html

https://doi.org/10.1017/S1351324920000601 Published online by Cambridge University Press

- *The Economist*, 6th August: 'A new AI language model generates poetry and prose: GPT-3 can be eerily human-like – for better and for worse'[g]
- The UK's *Telegraph*, 26th August: 'Forget deepfakes – we should be very worried about AI-generated text'[h]

Headlines that hint at the overtaking of the human race by smart machines make for good click-bait, but if you read any of these articles, you'll see that they are generally less sensationalist than their titles might suggest, and in fact they usually reflect quite well the limitations of the technology that have been commented upon in the more technical press.

Regardless, in terms of the column inches devoted to it, the release of GPT-3 has clearly been the most significant AI news event of the year. But once you get past the 'wow' factor, what's this technology actually good for? What kinds of commercial applications might we expect to see? And are there applications we should discourage?

2. Some history

First, let's review how we got here.

OpenAI was founded as a non-profit research organisation in late 2015 via a collective pledge of US$1B from a group of industry heavyweights, including Sam Altman (Y Combinator), Greg Brockman (Stripe), Reid Hoffman (LinkedIn), Elon Musk (Tesla) and Peter Thiel (Palantir). Its mission is to ensure that artificial general intelligence (AGI) benefits all of humanity; and in line with that mission, the goal of being the first to develop AGI.

Over its first few years, the organisation publicly released a number of software artefacts, but nothing that made headlines outside of the relevant communities of interest.

Then, in February 2019, OpenAI announced GPT-2 (for Generative Pre-trained Transformer 2), a large unsupervised transformer language model with 1.5B parameters trained on 40GB of text, or roughly 10B tokens. When used to repeatedly predict the next word in a text based on the preceding context, the model was capable of generating very coherent and plausible-sounding output, although it was also capable of outputting gibberish.

In its announcement, OpenAI stated 'Due to our concerns about malicious applications of the technology, we are not releasing the trained model'[i]. This immediately drew criticism from many who saw the claim that the technology was so dangerous that it had to be locked up as simply a means of generating hype and media interest. I have no idea whether that figured into OpenAI's strategy, although subsequent interviews given by Sam Altman, OpenAI's CEO, suggest that the company was and is pretty serious about the responsible release of the technology it develops.

In any case, the full 1.5B parameter model was eventually released in November 2019, following intermediate releases embodying increasingly larger language models: a 'small' 124M parameter model in February, a medium 355M model in May, and a 774M model in August.

Presaging the attention that GPT-3 would later generate, and no doubt alerted by the suggestion that the technology was dangerous, a number of mainstream media outlets picked up on the story, demonstrating the technology by allowing it to be a co-contributor. *The New Yorker*'s John Seabrook discussed predictive-text technology more generally in an interactive piece that, at various points in the article, lets you view GPT-2's contributions based on the preceding human-authored content;[j] and *The Economist* got GPT-2 to answer a youth essay question on climate change and had a team of judges assess the results.[k]

[g] https://www.economist.com/science-and-technology/2020/08/06/a-new-ai-language-model-generates-poetry-and-prose

[h] https://www.telegraph.co.uk/technology/2020/08/26/forget-deepfakes-ai-generated-text-should-worried/

[i] https://openai.com/blog/better-language-models/

[j] https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker

[k] https://www.economist.com/open-future/2019/10/01/how-to-respond-to-climate-change-if-you-are-an-algorithm

https://doi.org/10.1017/S1351324920000601 Published online by Cambridge University Press

3. GPT-3: A generator to trump all others

In June 2020, OpenAI announced GPT-3, a new language model more than 100 times larger than GPT-2, with 175B parameters and 96 layers trained on a corpus of 499B tokens of web content, making it by far the largest language model constructed to date. At the time of writing, the closest contenders are considerably smaller, with Microsoft's T-NLG and Google's T5-11B both being less than a tenth of GPT-3's size. And size, it seems, does matter: as it turned out, the texts created by GPT-3 were much more likely to sound coherent than those of its predecessor.[l]

Again, the model itself was not made available; instead, access was to be provided via an API, thus giving the model's creators more control over its use. At the time of writing, a beta version of the API is up and running, but you'll have to get on the presumably rather long wait-list if you want access. In the interim, some information on future pricing has leaked:[m] via a typical SaaS tiered pricing model, there's a free tier that gets you 100,000 generated tokens, a US$100-per-month tier that gets you 2M tokens, and a US$400-per-month tier that gets you 10M.

These prices have been criticised as being on the high side; they're certainly more than the sub-US$50 per month price-point that's typical of many other SaaS offerings and steep enough to lock out all but the keenest lone researchers. On the other hand, there are no obvious comparators on the basis of which we might establish what counts as a reasonable price, and arguably it's actually pretty cheap for access to a language model that is estimated to have a compute cost of US$4.6M per training run – and that's only a fraction of the overall total development and running costs.[n]

Meanwhile, back in March 2019, the non-profit OpenAI had restructured as a 'capped-profit' company, the stated reason being that this was necessary to be able to raise the kind of capital required to fund the company's high cost of research and maintain a pace of development competitive with major industry players like Google. Following this change, in July 2019, Microsoft agreed to invest US$1B in OpenAI over the next decade; and just over a year later, in September 2020, Microsoft obtained an exclusive licence to GPT-3. The consequences of this deal are unclear, but it's likely that the API access will be unaffected, whereas Microsoft's customers might eventually see the benefits of GPT-3 in a range of applications effectively for free.

As we noted in the introduction, the predominant sentiment in the media coverage of GPT-3 that has appeared subsequent to its release has been one of awe, sometimes followed by an expression of concern for the future of humanity now that AGI appeared to be in sight, eventually resolving to a recognition that we're still a long way from Skynet.

There's been a lot of praise for the technology's capabilities, especially in text generation. The typical use of the API involves providing a prompt and some initial text to get the model going, along with some optional parameter fiddling. Some of the outputs produced are truly breathtaking in their plausibility and believability as candidates for being human-authored text. The key word in that previous sentence, though, is 'some'; we'll get back to that below. There's insufficient space to include examples here, but you've probably seen some already, and if not, you can easily find them all over the web via your favourite search engine.[o]

Apart from the obvious application of text generation, the technology has also been lauded for its results in a wide range of other areas, some quite surprising: so you'll easily find examples and discussion of the model's capability in generating poetry, playing chess, doing arithmetic and writing web interface code on the basis of requirements expressed in natural language. It really is hard not to be impressed.

But, as noted above, there is a 'but'.

[l] Gwern Branwen suggests that, for fiction generation, the number of samples he'd need to consider before finding one worth showing off has fallen from 100 in the case of GPT-2 to five in the case of GPT-3: on that metric, at least, GPT-3 is 20 times better. This does, however, appear to require careful prompt refinement. See https://www.gwern.net/GPT-3.

[m] https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/

[n] https://lambdalabs.com/blog/demystifying-gpt-3/

[o] A comprehensive set of examples is provided on Gwern Branwen's website, along with extensive and detailed technical discussion: see https://www.gwern.net/GPT-3.

https://doi.org/10.1017/S1351324920000601 Published online by Cambridge University Press

4. Can you trust a transformer?

As has been widely noted, the technology has its limitations. The following have been identified by many observers:

- Its outputs may lack semantic coherence, resulting in text that is gibberish and increasingly nonsensical as the output grows longer.
- Its outputs embody all the biases that might be found in its training data: if you want white supremacist manifestos, GPT-3 can be coaxed to produce them endlessly.
- Its outputs may correspond to assertions that are not consonant with the truth.

As a consequence of these weaknesess, many of the impressive outputs that have been demonstrated are the results of cherry-picking: you run the API with the same prompt a few times, then pick the best result, rejecting those which sound less convincing or are just plain rubbish.

The Guardian attracted a lot of flak for being, in the eyes of many commentators, misleading in printing a news story entitled ‘A robot wrote this entire article. Are you scared yet, human?’.[p] The newspaper had set GPT-3 an assignment: convince us that robots come in peace. An editorial footnote admits

> GPT-3 produced eight different outputs, or essays. Each was unique, interesting and advanced a different argument. *The Guardian* could have just run one of the essays in its entirety. However, we chose instead to pick the best parts of each, in order to capture the different styles and registers of the AI.

To anyone who has ever looked at a pile of rejected GPT-3 outputs, this sounds more than just a little disingenuous; and the newspaper's subsequent claim that ‘editing GPT-3's op-ed was no different to editing a human op-ed’ might be considered insulting by the newspaper's human contributors. At the time of writing, *The Guardian* hadn't published the full set of outputs, so it's possible that a few of them are ‘unique’ and ‘interesting’ because they read like acid trip fiction.

Incoherent output is certainly a problem. But in terms of the weaknesses identified above, it's the last of the three that I want to single out as a major concern. Ultimately, there is nothing other than the fortuitous alignment of its textual statistics to make GPT-3 lean towards uttering statements which accord with reality. This is an important characteristic in determining what kinds of applications of the technology are appropriate.

In the case of *The Guardian*'s playful experiment, there was never any suggestion that GPT-3's output be taken seriously, or that it should be measured for its truth or falsity. But in those situations where truth is important, then we have a problem. This is most evident when GPT-3 is used as a question-answering system. Yes, it often does provide the correct answer to the question posed; but it often does not, and unless you already know the answer to the question ahead of time, you can't tell which of the two scenarios you're faced with. In line with much other reporting, Kelsey Piper in *Vox* acknowledges the scope for error, saying ‘GPT-3 can even correctly answer medical questions and explain its answers . . . though you should not trust all its answers’.[q] But that sounds like rather flawed logic: if you know you can't trust *some* of its answers, then you can't trust *any* of them.

Of course, these observations are not new. As Gary Marcus and Ernest Davis conclude in their piece in the *MIT Technology Review*:[r] ‘It's a fluent spouter of bullshit, but even with 175B parameters and 450 gigabytes of input data, it's not a reliable interpreter of the world’. OpenAI CEO Sam Altman himself underlined the limitations when he tweeted ‘The GPT-3 hype is way too much . . .

[p] https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

[q] https://www.vox.com/future-perfect/21355768/gpt-3-ai-openai-turing-test-language

[r] https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

https://doi.org/10.1017/S1351324920000601 Published online by Cambridge University Press

it still has serious weaknesses and sometimes makes very silly mistakes.'[s] I'm glad he came out and said this – but to suggest that it 'makes mistakes' already assumes some notion of intent, which GPT-3 lacks.

5. Unreliable doesn't mean useless

The bottom line is that GPT-3 is, to borrow a term from literary criticism, an 'unreliable narrator': its credibility, at least in regard to some key application use cases, is compromised by the fact that it is untethered to the truth.

This is not to say that GPT-3 is devoid of practical application; far from it. But it means that some use cases are appropriate and some are not.

The production of creative fiction, provided it is clearly identified as such, is of course perfectly fine; likewise poetry. And I'm sure we'll see many fantasy games where GPT-3 authors machine contributions to dialogues.

But, as already hinted, it seems to me that question-answering or advice-giving systems, where it's important that the resulting answer be true, are a risk too far. I expect that OpenAI's vetting process for requests for API access will rule out applications that attempt to offer advice in critical areas like health, but the demarcation lines here are fuzzy. For example, FitnessAI[t], which uses GPT-3 to answer questions about fitness, is already up and running. Now, I'm sure this application uses all kinds of pre- and post-filtering to avoid dealing with questions whose answers carry health risk, but it's hard to see how we can ensure that it won't at some point provide misinformation that leads to injury.

On the other hand, applications where a human stays in the loop are much safer, and we're already seeing a slew of these. Almost all of these are effectively augmented writing tools that take a user's textual input and provide an alternative version of that input, either longer or shorter depending on the application in question.

OtherSideAI's Quick Response[u] generates full-length emails in your style of writing given an outline of the key points you want covered; at the time of writing, the company had just raised US$2.6M in a seed funding round. Compose.ai[v] and Magic Email[w] also offer email-writing applications; and in a similar vein, Kriya.ai[x] generates personalised introduction requests: just what you need if you're trying to build up your LinkedIn network—you get 200 intros for US$9.

Dover.ai[y] rewrites your short job description into a longer variant.

Copy.ai[z] writes ad copy given a product description; you can sign up for a free trial. Taglines[aa] is another copy-writing tool, generating taglines based on product or service descriptions.

It's crucial to the success of these applications that in each case you can choose to accept, reject or edit the output generated, so you're not required to blindly place your trust in what is, ultimately, the wisdom of the web.

But things get trickier when you as a user may not be in a position to properly assess the outputs. Machine translation to or from a language you don't know is an already extant instance of this, and the reason why most of us would be willing to rely on an MT system to give us the gist of a news article but wouldn't be comfortable relying on it to translate a legally binding contract without some human review. And so I'm more wary of apps like that demonstrated by Michael

[s] https://twitter.com/sama/status/1284922296348454913

[t] https://app.fitnessai.com/knowledge/

[u] https://www.othersideai.com/; video at https://twitter.com/i/status/1285776335638614017.

[v] https://compose.ai/

[w] https://magicemail.io/

[x] https://www.kriya.ai/

[y] https://www.dover.io/tools/job-description-rewriter; This didn't appear to be working when I tried it.

[z] https://www.copy.ai/

[aa] https://www.taglines.ai/

https://doi.org/10.1017/S1351324920000601 Published online by Cambridge University Press

Tefula, which turns legalese into plain English:[ab] this sounds great in principle, but relying on the output would seem to carry a level of risk. Even if used as a writing assistance tool by a lawyer who's in a position to knowledgably post-edit the results, there's the risk of learned over-reliance – a danger that, ultimately, none of these applications is immune to.

6. Should we be worried after all?

We started out by observing that, despite the teasing of mainstream press headlines to the contrary, GPT-3 doesn't signal the beginning of the end for humanity. But that doesn't mean we shouldn't be concerned about the potential misuses of the technology. Helpfully, something of an 'ethics in AI' industry has grown up in the last few years, so it's unlikely that dubious uses of GPT-3 will avoid scrutiny. And, as we observed earlier, OpenAI itself is concerned about responsible use; here's a particularly relevant paragraph from their blog:[ac]

> One key factor we consider in approving uses of the API is the extent to which an application exhibits open-ended versus constrained behaviour with regard to the underlying generative capabilities of the system. Open-ended applications of the API (i.e., ones that enable frictionless generation of large amounts of customisable text via arbitrary prompts) are especially susceptible to misuse. Constraints that can make generative use cases safer include systems design that keeps a human in the loop, end user access restrictions, post-processing of outputs, content filtration, input/output length limitations, active monitoring and topicality limitations.

This, of course, is to be applauded, although it remains to be seen how workable these processes will be as the number of applications of the technology ramps up, and the ethical issues around specific cases get muddier; there are likely to be echoes here of the kinds of content moderation dilemmas faced by Facebook and Twitter. I understand why OpenAI's characterisation of what is safe, and what is not, has to be couched in pretty general terms, but it seems to me that they could go further. From where I sit, one maxim is incontrovertible:

> To the extent that a use case places importance on the truth of the outputs provided, it is not a good fit for GPT-3.

But that might not go down well with Marketing.

[ab] See the video at https://twitter.com/i/status/1287425989878915074.
[ac] https://openai.com/blog/openai-api/

Cite this article: Dale R (2021). GPT-3: What's it good for? *Natural Language Engineering* **27**, 113–118. https://doi.org/10.1017/S1351324920000601

https://doi.org/10.1017/S1351324920000601 Published online by Cambridge University Press