Introduction to AI Narratives (R255 Topic 7: Narratives in Artificial Intelligence and Machine Learning)

Alessandro Trevisan
19th January 2024

Summary

In my research, I am interested in the task of benchmarking the outputs generated by LLMs in response to open-ended questions. This project derives from the observation that all of the mainstream benchmarks for LLMs (including ARC, HellaSwag, MMLU, WinoGrande, TruthfulQA, and GSM8K) rely on multiple-choice questions to evaluate the quality of the responses provided by these models. (Multiple choice question questions are typically designed to assess the reasoning, general knowledge, and calculation skills of LLMs). However, as different critics have argued, including Lianmin Zheng et al., these types of benchmarks are not always appropriate to evaluate current state-of-the-art LLMs, which have in many cases come to master such close-ended questions (except for the calculation ones).[1] Indeed, these benchmarks may be misleading, because higher results in the tests contained in them might not necessarily equate to improvements in LLMs that can actually be *felt* by the vast majority of users. This is what Melanie Mitchell is getting at, with reference to the Gemini family of multimodal models released by Google DeepMind in December last year when she states that 'it's not obvious […] that Gemini is actually substantially more capable than GPT-4', despite performing better in 30 out of 32 benchmarks.[2] So, because of the inappropriateness of current benchmarks, and because, as I will argue, they are highly specific, meaning that they test only specific types of intelligence, I propose that what is created, as a result, is a *narrative* about the capabilities of LLMs which might not really reflect the reality. (In the case of Gemini, this narrative may be used even for commercial purposes).

And yet, despite all these problems with benchmarks that rely on close-ended questions, there still has not been a general effort to evaluate (and possibly measure), the outputs generated by LLMs in response to open-ended questions. (An open-ended question is a type of question that cannot be answered with a simple "yes" or "no" response. Instead, it encourages the person, or, in this case, the chatbot, being questioned to provide more detailed and expansive answers. Open-ended questions often begin with words like "how," "why," "what," or "tell me about"). Indeed, evaluating the outputs generated by a language model in response to an open-ended question is notoriously challenging due to the subjective nature of this task, in contrast to the more objective assessment of yes/no answers. One approach to benchmarking LLM-generated responses to open-ended questions that has been suggested involves using a state-of-the-art LLM to evaluate the responses generated by another LLM. This benchmarking strategy is detailed by the afore-mentioned Zheng et al. in a paper titled 'Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena' and published just this past October (so, as you can tell, this is still a fresh research field). However, as we shall see, the current LLMs, even the

[1] Lianmin Zheng, et al., 'Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena' (arXiv, 2023) <https://doi.org/10.48550/arXiv.2306.05685>.
[2] Melanie Mitchell, quoted by Melissa Heikkilä and Will Douglas Heaven, 'Google Deepmind's New Gemini Model Looks Amazing—but Could Signal Peak AI Hype, *MIT Technology Review*' <https://www.technologyreview.com/2023/12/06/1084471/google-deepminds-new-gemini-model-looks-amazing-but-could-signal-peak-ai-hype/> [accessed 31 December 2023].

very big ones, often struggle to evaluate the expressiveness and aesthetic qualities of a text in a way that aligns with human preferences. Therefore, what I am trying to say is that they do not always make reliable evaluators of text. We shall see this in greater detail in a moment.

One of the key areas in which LLMs struggle pertains to evaluating the quality of *narratives*. Therefore, right now, in my research, I am focusing on automating the task of benchmarking the narratives generated by LLMs. I would like to specify that out of all the possible forms of LLM-generated text which I could have set out to benchmark, I chose to focus first on *narrative* communication because I believe that proficiency in storytelling, as evidenced in the work of prominent narratologists like Marie-Laure Ryan, requires the application of different types of intelligence. As Ryan put it, 'stories tell us about problem solving, about the interplay in life of planned action and random events, about the feelings of other people, about the time-bound nature of human experience, about success and failure, and they could very well help us interpret life according to these categories'.[3] I am thus interested in this idea of narrative and narrative communication as the seat of an inherent humanity, as it were: in the idea that, in an LLM's narrative outputs, we may observe not merely its capacity to tell a compelling story but also its logical—problem solving—intelligence, its spatio-temporal intelligence, and also its emotional intelligence. In this sense, I would argue that narrative provides a good testing ground for machine intelligence.

(To briefly recap: my objective here is to use an LLM to evaluate the narrative outputs, the stories, created by another LLM).

The problem with this, however, is that current, off-the-shelf LLMs are not really capable of distinguishing good narratives from bad narratives, as it turns out. LLMs, indeed, struggle to evaluate narratives, and narrative communication is something they still do not have the full grasp of, as you might have noticed in your own interactions with language models. Here is an example: I asked ChatGPT 3.5 to pretend it was a university-level marker for a unit in Creative Writing and to compare two different stories, one of which had been written by it while the other was 'Hills Like White Elephants' by Ernest Hemingway; it was quite interesting that ChatGPT went on to give the story it had written a mark of 95/100, while the Hemingway story only got 85/100.

What I think this suggests is the frequent inappropriateness of using LLMs for evaluating text (in the way proposed by Zheng et al.).

The way in which *I* am attempting to get around this problem and automate this process of benchmarking is by finetuning an open-source LLM on a labelled dataset comprising 'good' and 'bad' narratives (these could be short stories by Hemingway and ChatGPT-authored stories, respectively), so that the finetuned model would then be able to distinguish the two, and hopefully discern that the story by Hemingway possesses greater artistic merit than the one written by ChatGPT.

Research question (part 1): How can we define a good narrative?

Now, of course, some of the questions which you may be asking yourselves at this point are: how can we describe a narrative? And, furthermore, what is a 'good' narrative and what is

[3] Marie-Laure Ryan, 'Narratology and Cognitive Science: A Problematic Relation', *Style*, 44.4 (2010), 469–95 (p. 484).

a ‘bad’ narrative? Aside from the simple example quoted above (with the Hemingway stories as the ‘good’ stories and the ChatGPT stories as the ‘bad’ ones), this is an issue that is at the center of my research and that will become increasingly more important as language models get better at crafting narratives and at evaluating them. Indeed, these are complex and fascinating questions which have been at the centre of literary criticism for centuries.

To understand the history and possible future directions of this debate, I would like to draw your attention to a parallel, which I propose, between the task of benchmarking LLMs and canon formation. The process of selecting a canon, in literature, involves discerning which narratives are worthy of being propagated (whether that be orally or in written form). The word ‘canon’, as specified by classicist George A. Kennedy, derives from ‘the Greek *kanôn* (perhaps derived from a Semitic word for “reed”), meaning a straight rod or bar used by a weaver or carpenter, then a rule or model in law or in art’.[4] I would like to underline here that the canon was conceived, by the Greeks, as a measuring standard, analogous, in many ways, to a benchmark. The status of the canon as a sort of litmus test for assessing the artistic merit of a work of art is evidenced by Kennedy when he specifies: ‘[i]n the fourth century B.C. Polycrates carved a statue called “The Canon,” which established artistic proportions for representation of the human figure. The earliest application of *kanôn* to describe written texts is a statement in the third chapter of the *Letter to Pompeius* by Dionysius of Halicarnassus that Herodotus is the best canon (that is, “model”) of Ionic historiography and Thucydides of Attic’.

The point I would like to make here is that the canon itself, much like different AI benchmarks, constitutes a narrative (in the case of the canon a narrative about what artistic characteristics we value as a culture). As you might have guessed, in fact, the canon is subjective, malleable, by all means not etched in stone, but under constant mutation. The canon undergoes constant rejection and revision. Kennedy brings the example of Sappho of Lesbos, who was ‘the only woman writer included in any of the ancient canons; none of her works were copied into codex manuscripts [in the Middle Ages] and as a result most were lost. What we have today are two poems quoted by male writers, some other brief quotations, and fragments that have been recovered in modern times on pieces of papyrus in Egypt. What amounted to censorship of Greek lyric poetry by early medieval scribes perhaps reflects distaste for many of its themes, especially homosexual love, but the Aeolic dialect of the poetry […] was also a negative factor’. You can see here how what was considered canonical by one generation of literary critics is completely ignored (censored) by another.

I think it is useful to focus a bit more on this struggle between different narratives vying for transmission to future generations of readers. Jonathan Swift even wrote a short satire on the revision of the classics which happened in 17th/18th century France and England, titled ‘The Battle of the Books’: the satire depicts a literal battle, symbolising the struggle for canonisation, between books in the King’s Library.[5]

The literary historian and theorist Franco Moretti, in his provocatively titled essay ‘The Slaughterhouse of Literature’ (2000), describes this battle, this conflict between different books

[4] George A. Kennedy, ‘The Origin of the Concept of a Canon and Its Application to the Greek and Latin Classic’, in *Canon Vs. Culture: Reflections on the Current Debate*, ed. by Jan Groak, <https://learning.oreilly.com/library/view/canon-vs-culture/9780815308898/19_Chapter07.html> [accessed 8 January 2024].

[5] Jonathan Swift, *The Battle of the Books, and Other Short Pieces*, ed. by Henry Morley, 1996 <https://www.gutenberg.org/ebooks/623> [accessed 19 January 2024].

as an evolutionary struggle between narratives.[6] He brings the example of 19th and early 20th-century detective stories, arguing that the ones which ended up surviving (which ended up being *canonised*) are the ones which contain specific plot devices, or, to borrow a term from evolutionary biology, specific traits, which guaranteed their commercial success: Moretti postulates that it was Doyle's specific use of clues, the fact that, within Sherlock Holmes stories, they are often *necessary* to the solution of the mystery and *decodable* by the reader, which made these texts particularly successful.

Moretti suggests that, ultimately, it is the readers, those who buy the books, and not, crucially, the university professors, who make the canon. To support this idea one need only turn to George Orwell's essay 'Good Bad Books', wherein Orwell argues that even books that do not necessarily display great artistic merit ('bad books') may still end up being canonised and, thus, bought by many successive generations of readers.[7] This might be because these 'bad', unoriginal books are still be able to convey true emotions and to move readers, for instance. So, to summarise: the canon is created by the *readers, the people who buy the books, the market*.

Research question (part 2): Recognising dominant AI narratives

Thinking about the multiplicity of literary narratives in which AI goes rogue and poses an existential threat to humanity (e.g. Mary Shelley's *Frankenstein*, H. G. Wells' novels, Karel Čapek's play 'R.U.R.', etc.), I would suggest that so many of these texts were propagated primarily because they were highly memorable, and thus sold more than the ones which perhaps included a more nuanced, less extreme interaction between humans and machines.

The importance of the commercial appeal of an AI narrative for its propagation is something that is suggested by a report published in 2018 by the UK House of Lords Select Committee on Artificial Intelligence, wherein multiple AI researchers stated that 'the public have an unduly negative view of AI and its implications, which in their view had largely been created by Hollywood depictions and sensationalist, inaccurate media reporting'.[8] So, you can see here how there is a suggestion that the market, in a way, whether it be the box-office market or the news media one, has substantially influenced the public's perception of AI.

(I thought I would mention that the book *AI Narratives: A History of Imaginative Thinking About Intelligent Machines*—edited by Stephen Cave, Kanta Dihal, and Sarah Dillon—includes many chapters on 'more sophisticated stories about AI […], in contrast to the narratives that currently dominate' (as the editors specify, pp. 9-10): I would really recommend that you check out this book if you are interested in the topic!)

Up until now we have been talking chiefly about AI narratives as narratives expressed in literary form (i.e. novels, plays, film or TV scripts). But, of course, not all narratives come in literary form. I would therefore like to go back to the idea of benchmarks as narratives. What I would like to suggest is that narratives about AI reside not just in popular entertainment for

---

[6] Franco Moretti, *Distant Reading* (London: Verso, 2013).
[7] George Orwell, 'Good Bad Books', in *Shooting an Elephant* (London; New York: Penguin Classics, 2009).
[8] UK House of Lords Select Committee on Artificial Intelligence (2018 report), quoted in Cave, Stephen, et al., eds., *AI Narratives: A History of Imaginative Thinking about Intelligent Machines* (Oxford: OUP, 2020).

the laypeople, for non-specialists, but that they can also infiltrate themselves into technical conversations among machine learning engineers and researchers. In *AI Narratives*, the editors specify that '[n]arratives of intelligent machines *matter* because they form the backdrop against which AI systems are being developed, and against which these developments are interpreted and assessed' (p. 7). Furthermore, Dillon and Schaffer-Goddard highlight that 'the narratives with which AI researchers themselves engage can influence their 'career choice, research focus, community formation, social and ethical thinking, and science communication' (p. 8). **Even the work of AI researchers may be influenced by different narratives surrounding their field.**

In my research, I am now endeavouring to argue that benchmarks are engendered by deep cultural narratives, by cultural understandings of computing and intelligence, and that, in turn, they can substantially influence the way LLMs are received and developed. One of the questions I am asking is: what kind of intelligence are we actually testing with current benchmarks? It seems to me that these benchmarks reflect an *ethnocentric* conception of intelligence: not only are the tests that form these datasets in English, but they can also comprise questions closely modelled on prominent academic ability tests (like the SAT) created by and for Western institutions. Furthermore, some of these questions (like the ones on US history included within MMLU) may even imply a Western perspective through the overtness of their cultural specificity. Thus, one issue which I am interested in discussing is: are these benchmarks propagating a narrative in which intelligence is equivalent to academic ability in Western higher education institutions?

How have narratives impacted the development of AI systems?

In the final section of my talk, I would like to illustrate how the Turing Test itself, what I think of as a 'proto-benchmark', helped originate narratives which then directly informed the way certain early chatbots were programmed. To explore this argument, I find it useful to emphasize the distinction between a specific AI program (i.e. the actual lines of code it comprises) and the way people perceive or think about that AI program. To do this I would like to turn to the concept of 'computational assemblage', discussed by John Johnston: he notes that, '[i]n this framework, every computational machine is conceived of as a material assemblage (a physical device) conjoined with a unique *discourse* that explains and justifies the machine's operation and purpose' [my emphasis].[9]

This notion of 'computational assemblage' is a very important one to have in mind when looking, for instance, at early chatbots, as suggested by Simone Natale.[10] This is because even the engineers who created these chatbots often made them to fit within specific narratives about intelligence which arose as a response to the Turing Test. What I mean is that these chatbots were created so that they could 'imitate intelligence': the Turing Test is, after all, the 'Imitation Game', and, as such, is about how well the outputs generated by a computer fit

[9] John Johnston, *The Allure of Machinic Life: Cybernetics, Artificial Life, and the New AI* (Cambridge, UNITED STATES: MIT Press, 2010), p. x <http://ebookcentral.proquest.com/lib/cam/detail.action?docID=3338928> [accessed 28 December 2023].

[10] Simone Natale, 'How to Create a Bot: Programming Deception at the Loebner Prize Competition', in *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test*, ed. by Simone Natale (Oxford University Press, 2021), <https://doi.org/10.1093/oso/9780190080365.003.0006>.

within our narratives about intelligence, and not about whether the computer is 'actually' intelligent.

Multiple significant occasions in which the development of a chatbot was *guided* by widespread AI narratives unfolded during the Loebner Prize competition (which ran annually between 1991 and 2019 and which was conceived by American entrepreneur Hugh Loebner). As part of this competition, computer programmers could submit a chatbot to participate in a series of Turing Tests. At the end of the competition, the chatbot that had performed most convincingly in these tests would be crowned champion. What is interesting is that many of the chatbots that participated were given a distinctive personality in an attempt to fake an inherent *human-ness*: very soon, the programmers who decided to participate in the Loebner Prize competition figured out that a chatbot 'was more credible if it imitated not only humans' skills and abilities but also their shortcomings', as highlighted by Natale. Specifically, these purported shortcomings which the chatbots were programmed to feature could in turn serve to mask the shortcomings of these machines as dialogue partners. This strategy originated chatbots like PC Therapist, programmed by Joseph Weintraub and winner of the 1991 Loebner Prize, which simulates a jester: therefore, the frequent irrelevant and nonsensical response provided by this chatbot were often justified by the judges as the irreverent expressive style of a jester. Then came TIPS, a chatbot programmed by Thomas Whalen, which won the 1994 Loebner Prize by pretending to be a janitor at the University of Eastern Ontario who worked night shifts and did not read or watch television, and thus had a highly limited knowledge of world events. Of course, this backstory that was given to TIPS was intended to somehow justify the bot's shortcomings in its general knowledge or in its expressive abilities. Furthermore, this strategy, the 'strategy of the programmed shortcomings', as I call it, was then adopted by programmers who created chatbots which pretended to be non-native English speakers, for instance, so as to mask their occasional use of simple a simple diction and syntax. There are many different examples of similar chatbots, but I hope you get the gist.

I would here argue that the evolutionary argument introduced earlier in reference to Franco Moretti and the birth of the canon is once more germane in describing the narratives spun around actual AI programs, like the chatbots PC Therapist and TIPS. I say this because the overarching narrative of the 'programmed shortcomings' was adopted by different developers precisely *because* it guaranteed the survival, the relevance, the victory at the Loebner Prize of their chatbots. The success of these chatbots, indeed, depended almost entirely on the credibility of the narratives which they propose in their conversations: these AI programs were indeed quite basic, quite disappointing actually, when you look at their workings, but it was the *discourse* that surrounded them, to go back to Johnston's definition of 'computational assemblage', which created a sense of *illusion*, the illusion that indeed they were intelligent.

I acknowledge that the development of these chatbots as part of the Loebner Prize competition cannot be genuinely regarded as serious AI research, as it did not really help push the field forward in any significant way. Nonetheless, I believe it highlights how the narrative at the basis of the Turing Test, the notion that a given AI program under examination might be intelligent, influenced how some of the early chatbots were programmed. And this, I suggest, provides us with an interesting lens through which to look at the modern benchmarks and tests for AI, to consider whether they too, much like the Turing Test, favour the propagation of certain narratives about computing and intelligence.

I would like to use the example of the Loebner Prize competition to reiterate this point: narratives about AI are crucial, as noted by Dillon and Schaffer-Goddard, because of their potential to shape the development of AI.