Natural Language Engineering (2024), **30**, pp. 417–427
doi:10.1017/S1351324923000578

EMERGING TRENDS

Emerging trends: When can users trust GPT, and when should they intervene?

Kenneth Church

Northeastern University, Boston, MA, USA
Email: k.church@northeastern.edu

(Received 19 December 2023; accepted 19 December 2023; first published online 16 January 2024)

Abstract

Usage of large language models and chat bots will almost surely continue to grow, since they are so easy to use, and so (incredibly) credible. I would be more comfortable with this reality if we encouraged more evaluations with humans-in-the-loop to come up with a better characterization of when the machine can be trusted and when humans should intervene. This article will describe a homework assignment, where I asked my students to use tools such as chat bots and web search to write a number of essays. Even after considerable discussion in class on hallucinations, many of the essays were full of misinformation that should have been fact-checked. Apparently, it is easier to believe ChatGPT than to be skeptical. Fact-checking and web search are too much trouble.

Keywords: ChatGPT; fluency; trustworthiness; human-in-the-loop; evaluation in situ; fact-checking

1. Introduction

Much has been written about GPT including Dale (2021), the most read paper in this journal.[a] ChatGPT is extremely popular.

> *ChatGPT sets record for fastest-growing user base.*[b]

Why is GPT so popular? The fact that so much has been written on this question suggests that the definitive answer to that question has not been written (yet). And it is unlikely that it will be written soon. That said, GPT is remarkably easy to use, and the outputs are written in an authoritative style that appears to be credible (even when it is making stuff up).

Is ChatGPT a Google Killer? Much has also been written on this question. As the cliche goes, it is difficult to make predictions, especially about the future, but the recent excitement over retrieval-augmented generation (Lewis *et al.* 2020) suggests there is plenty of room for synergy between chat bots and web search.

Our last emerging trends article Church and Yue (2023) on "Smooth-Talking Machines" suggested Chat bots have amazing strengths (fluency) and amazing weaknesses (trustworthiness). Search can help with trustworthiness. Chat bots have a serious problem with hallucinations. Search can be used in a fact-checking mode to mitigate some of this risk. But, as a practical matter, fact-checking is hard work, and probably won't happen as much as it should.

[a]https://www.cambridge.org/core/journals/natural-language-engineering/most-read
[b]https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

2. Evaluation of prompt engineering

How do we evaluate prompt engineering? InstructEval[c] (Chia *et al.* 2023) combines a number of other benchmarks[d] (Dua *et al.* 2019; Hendrycks *et al.* 2020; Chen *et al.* 2021; Frohberg and Binder 2022; Srivastava *et al.* 2023). These methods represent the current consensus on how to evaluate deep nets, but it is important to evaluate more than just the machine. We want to measure both what the machine does and what ordinary users (and my students) will do with the machine. In addition to benchmarks such as InstructEval, we should also run evaluations with humans-in-the-loop.

OpenAI, the organization behind ChatGPT, believes it is important to roll out new GPTs incrementally because it takes time for the community to learn how to use the next GPT, and because it takes time for OpenAI to appreciate how the community will take advantage of the opportunities. The process is very much a collaboration between people and machines. We would like to see evaluations that are more collaborative than traditional evaluations of deep nets.

3. Homework: use ChatGPT to write essays

This article will describe a homework assignment that I gave to my class on NLP (Natural Language Processing).[e] The homework asked the students to use tools such as chat bots and web search to write a number of essays. I was surprised how well ChatGPT did on some of these essays, but I was also surprised how much the students tended to believe ChatGPT, even when it was making stuff up. Even after considerable discussion in class on hallucinations, many of the student essays contained misinformation that the students should have fact-checked, as we will see.

There is considerable interest in how misinformation and toxicity spread online (Qazvinian *et al.* 2011; Vicario *et al.* 2016; Shu *et al.* 2017; Zubiaga *et al.* 2017; Cinelli *et al.* 2021). Work on machine learning classifiers (Poletto *et al.* 2020; Fortuna *et al.* 2021)[f] is unlikely to help much, given the incentives. Just as we cannot expect tobacco companies to sell fewer cigarettes and prioritize public health ahead of profits, so too, it may be asking too much of companies (and countries) to stop trafficking in misinformation given that it is so effective and so insanely profitable (Church *et al.* 2023). If we gave these companies a toxicity classifier that just worked, given current incentives to maximize profits, they should run the classifier in the reverse direction to maximize toxicity, since toxicity is profitable.

The homework mentioned above raises an additional concern. Not only should we be concerned about the supply of misinformation, but we should also be concerned about demand for misinformation. As mentioned above, my students could have easily fact-checked their homework, but they chose not to do so. They were prepared to believe much of what ChatGPT says, because of how it says what it says, and ease-of-use. It is easier to believe ChatGPT than to be skeptical. Fact-checking and web search are too much trouble.

4. Lay of the land

Table 1 summarizes my interpretation of what the students did with large language models (LLMs). There are also columns for what the students could have done with traditional NLP and web search. LLMs such as ChatGPT are amazingly good on metaphors, probably better than non-native speakers of English. Most of the students are international students. They did not grow up in America and they are not familiar with Americana (such as metaphors involving American baseball). Although web search also works well on metaphors and Americana, my students tended

[c] https://declare-lab.net/instruct-eval/
[d] https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/hhh_alignment
[e] https://kwchurch.github.io/teaching/2023-fall/CS6120/assignments/assignment.04/index.html
[f] https://paperswithcode.com/task/fake-news-detection

Table 1. LLMs have amazing strengths and amazing weaknesses

Task	Traditional NLP	Web search	LLMs
Metaphor	"AI Complete"	Very good	Amazingly good
Documentation	NA	Useful	Amazingly good
Outlines	NA	NA	Useful
Directions	NA	Useful	Poor
Quotes	NA	Useful	Amazingly bad
References	NA	Useful	Amazingly bad
Perspective	NA	Useful	Amazingly bad

to prefer ChatGPT, perhaps because it is the new hot thing, or perhaps because bots are easier to use (and require fewer clicks than web search).

I was surprised how well ChatGPT did on metaphors, but I was also surprised how poorly it did on references. If you ask ChatGPT for references on some topic, it will make up papers that do not stand up to fact-checking. We will discuss some of the amazingly good cases in Section 5 and some of the amazingly bad cases in Section 6.

4.1 You have no idea how much we're using ChatGPT

The homework assignment was inspired by Owen Terry's *I'm a Student. You have no idea how much we're using ChatGPT.*[g] We discussed Terry's essay in Section 2.6 of Church and Yue (2023). In an interview on NPR,[h] Terry, a rising sophomore at Columbia, identified some of ChatGPT's strengths and weaknesses. ChatGPT is good at producing thesis statements and outlines, but it does not capture the student's style, and it is worse on quotes. If you ask for quotes, it makes stuff up.

4.2 Zero-tolerance for misinformation

Quotes and the other amazingly bad cases in Table 1 involve *hallucinations*. The term, hallucination, has become a nice way of referring to what we used to call bugs and computer errors:

> *To Err is Human; To Really Foul Things Up Requires a Computer*[i]

There should be no excuse for making stuff up. And it should be even worse for someone to traffic in misinformation. If someone cheated on their CV, they could be fired, even years after the offense. So too, there should be little tolerance for handing in homework with misinformation, especially when it is so easy to fact-check with search.

It should not be the teacher's responsibility to fact-check the homework. To discourage the spread of misinformation, there need to be prohibitively high penalties in the court of public opinion for trafficking in misinformation. To stop the spread of misinformation, we need to go after all parties including suppliers, users, and market makers. If we do not address the problem, misinformation could lead to a loss of confidence in all things digital.

[g] https://www.chronicle.com/article/im-a-student-you-have-no-idea-how-much-were-using-chatgpt
[h] https://www.wbur.org/hereandnow/2023/05/22/chatgpt-academia
[i] https://quoteinvestigator.com/2010/12/07/foul-computer/

https://doi.org/10.1017/S1351324923000578 Published online by Cambridge University Press

4.3 Adversarial attacks

There is a considerable literature on adversarial attacks on LLMs (Jia and Liang 2017; Morris *et al.* 2020; Camburu *et al.* 2020; Wang *et al.* 2021; Li *et al.* 2021; Ziegler *et al.* 2022; Wang *et al.* 2023). We do not see as many papers on adversarial attacks on web search, perhaps because it is too easy to find misinformation on the web. When we look at logs of speech APIs and agents like SIRI, we find lots of kids having fun at the expense of the API, but we see less "prank calling" attacks against web search:

> *Do you have Sir Walter Raleigh in a can? Better let him out!*[j]

One might view search engine optimization as an adversarial attack against web search, though Google documentation prefers to view the subject as an opportunity to teach content owners how to work with Google, as opposed to how to work against Google.[k] I prefer to view arbitrage as a way to teach the market makers such as Google and ChatGPT how to do better in the future.

I am not sure why there is so much literature on adversarial attacks against ChatGPT given some of the observations in Table 1. It is really not hard to generate prompts that will produce hallucinations:

1. Ask ChatGPT for quotes
2. Ask ChatGPT for references
3. Ask ChatGPT to crawl links, quotes, references
4. Ask ChatGPT to evaluate an expression with more than two big numbers

It is not hard to find opportunities for improvement. Consider Chain-of-Thought Prompting (Wei *et al.* 2022). This paper is widely cited because it offers a constructive workaround to an obvious weakness in ChatGPT. ChatGPT is not particularly good at decomposing complex tasks into two or more simpler tasks. ChatGPT lacks principles such as *superposition*,[l] a key principle in linear algebra and linear systems.[m] Intermediate representations and compositionality used to be hot topics in linguistics and computer science. End-to-end systems are easier to optimize and may perform well on benchmarks, but modularity has other advantages:

1. explanation,
2. capturing relevant generalizations (Chomsky 1957, 1965), and
3. model size (number of parameters) (Minsky and Papert 1969).

4.4 Human-in-the-loop

Much of the adversarial literature above is focused on use cases with no human in the loop. In Church and Yue (2023), we envisioned a human-in-the-loop collaboration of humans with machines. To make this work, we need to characterize which subtasks are appropriate for humans and which are appropriate for machines. Evaluations in our field tend to focus on how well the machines are doing by themselves, as opposed to which tasks are appropriate for humans and which are appropriate for machines.

The adversarial literature above is asking the right questions:

1. *What ChatGPT does well?*
2. *What ChatGPT does not do well?* (Wang *et al.* 2023)

[j] https://pin.it/4bqqmcP
[k] https://developers.google.com/search/docs/fundamentals/seo-starter-guide
[l] https://en.wikipedia.org/wiki/Superposition_principle
[m] https://en.wikipedia.org/wiki/Linear_system

https://doi.org/10.1017/S1351324923000578 Published online by Cambridge University Press

but we were hoping for answers that would be more helpful in human-in-the-loop settings. For example, the first answer to the second question above in Wang *et al.* (2023) is

> *The absolute performance of ChatGPT on adversarial and OOD classification tasks is still far from perfection even if it outperforms most of the counterparts.*

This answer is not that useful in human-in-the-loop settings. We already know that the machine is more fluent than trustworthy (on average) (Church and Yue 2023). What we would like to know is: when can we trust the machine, and when should we ask the human to step in.

5. Amazingly good

5.1 Metaphor

I expected the following sports metaphors to be hard, especially since most of the students in the class are not native speakers of English. But I was surprised how well ChatGPT did on these.

I asked the students to use chat bots and search engines to explain what the following terms mean. Which sports are these terms from? What do they mean in that sport? How are they used metaphorically outside of that sport? If the term is more common in one English speaking country than another, can you name a country that is likely to use the term?

1. *cover all the bases*
2. *drop the ball*
3. *dunk*
4. *fumble*
5. *get on base*
6. *hit a home run*
7. *out in left field*
8. *punt* (as a noun)
9. *punt* (as a verb)
10. *ragging the puck*
11. *run out the clock*
12. *sticky wicket*
13. *strike out*

In most cases, ChatGPT is at least as good as Google, and easier to use. Users like instant answers, where they do not have to click on anything. That said, content providers and developers of search engines depend on behavioral signals like clicks to make improvements. In the long run, without these behavioral signals, results are likely to degrade over time. Chat bots may be doing well in the short term, but it may be in their best interest to – and good for the internet ecosystem – if chat bots work with content suppliers, rather than dis-intermediate them.

For more obscure metaphors like *ragging the puck*, Google and Bing return a link[n] with an excellent answer, whereas ChatGPT does not explain the metaphorical sense. The version of Bing chat (which uses ChatGPT) provides two definitions with links. I would not have known this idiom if I had not spent so much time reading Canadian Hansards (Gale and Church 1991). In the Hansards, ragging the puck means to waste time with great skill. It is something that you accuse the other guy of doing, or deny that you are doing, because the filibuster is not a thing in the Canadian Parliament.

[n]https://idioms.thefreedictionary.com/ragging+the+puck

https://doi.org/10.1017/S1351324923000578 Published online by Cambridge University Press

Metaphor is a classic hard problem in computational linguistics. There is a considerable literature on this topic (Carbonell 1980; Fass and Wilks 1983; Martin 1990; Hobbs 1992; Gedigian *et al.* 2006; Krishnakumaran and Zhu 2007; Lakoff 2008; Lakoff and Johnson 2008; Shutova 2010; Mohammad *et al.* 2016), but most of us that have been working in computational linguistics for a few decades have rather low expectations. We grew up with certain stereotypes. The 1966 TV show, "Get Smart," introduced a robot character, Hymie, in season 1, episode 19.[o] The running gag was that Hymie was too literal. When Max, the leading character, tells Hymie, "you knock me out," the robot does exactly that.[p]

Given this stereotype, it is surprising how well ChatGPT does on: "you knock me out."

Prompt : *What does "you knock me out" mean?*

ChatGPT : *The phrase "you knock me out" is an idiomatic expression in English that is often used to convey a strong positive reaction to someone or something. When someone says, "You knock me out," they are typically expressing admiration, surprise, or amazement. It's a way of saying that the person or thing being referred to has a powerful or impressive effect on the speaker. For example, if someone is really impressed by a performance, a joke, or someone's charisma, they might say, "Wow, you really knock me out!" It's a colloquial and informal way of expressing appreciation or being captivated by someone or something.*

5.2 Documentation

When I constructed the homework, I expected ChatGPT to work well on documentation. Last summer, some undergraduates taught me that they prefer ChatGPT to Stack Overflow. To test this observation on my class of masters students, I asked them to do the following with vectors from MUSE (which they had used previously):

> *Some people are finding ChatGPT to be useful for documentation (e.g., as a replacement for Stack Overflow). Suppose I have a set of vectors from MUSE and I want to index them with annoy and/or faiss.*
>
> 1. *Find some documentation for how to do this.*
> 2. *If possible, find some examples.*
> 3. *Write a program that creates the index.*
> 4. *Write a program that queries the index.*
> 5. *Provide a short description of what approximate nearest neighbors does, and what is it useful for.*
> 6. *Do a literature survey to find some primary papers on approximate nearest neighbors, as well as some survey papers on the topic*
>
> *As with the previous two questions, in addition to answering the specific questions, I am more interested in which tools you found useful, and how you used them. Did you already know the answers? Did you use any tools, and were they useful? Which tools were useful, and which were not? What worked and what did not? Do you have any useful comments about what is good for what, and what is not good for what?*

[o] https://www.imdb.com/title/tt0587500/
[p] https://www.youtube.com/watch?v=MPhg_Hhsvhg

https://doi.org/10.1017/S1351324923000578 Published online by Cambridge University Press

Most students used ChatGPT, as is, for this question. ChatGPT was amazingly good on questions 1–4.

6. Amazingly bad

6.1 Made-up references

But ChatGPT was amazingly bad on question 6. A number of the students returned the same awful answer:

> Title: "A Survey of Nearest Neighbor Search Algorithms"
> Authors: Yufei Tao, Dongxiang Zhang
> Link: Survey Paper (link to https://arxiv.org/abs/1904.06188)

Note that this paper does not exist, though there are a number of papers with similar titles such as Abbasifard *et al.* (2014). The authors can be found in Google Scholar: Zhang Dongxiang and Yufei Tao, though I was unable to find a paper that they wrote together. Worse, the link points to Amanbek *et al.* (2020), a completely different paper on a different topic with a different title and different authors. I believe the students used ChatGPT to find this non-existent paper. I had hoped that the students would do more fact-checking than they did, especially after having discussed hallucinations in class, but users do not do as much fact-checking as they should. Perhaps the court of public opinion needs to increase the penalties for trafficking in misinformation.

This case is similar to a recent case where a lawyer relied on A.I. He "did not comprehend" that the chat bot could lead him astray. The bot crafted a motion full of made-up case law.

> *As Mr. Schwartz answered the judge's questions, the reaction in the courtroom, crammed with close to 70 people who included lawyers, law students, law clerks and professors, rippled across the benches. There were gasps, giggles and sighs... "I continued to be duped by ChatGPT. It's embarrassing," Mr. Schwartz said.*[q]

> *The episode, which arose in an otherwise obscure lawsuit, has riveted the tech world, where there has been a growing debate about the dangers – even an existential threat to humanity – posed by artificial intelligence. It has also transfixed lawyers and judges.*

In addition to computer errors (hallucinations), there were also some human errors. I'm not sure the students (and ChatGPT) understand the difference between primary literature and secondary literature. One student confused *approximate nearest neighbors* with *shortest paths*. A survey paper is not the same as a paper whose title contains the string: "survey." That said, I am much more concerned with trafficking in computer-generated misinformation (bogus references) than human errors by well-meaning students that are making excellent progress on the material in class.

6.2 Essays that are not only wrong but lack depth and perspective in ways that could be dangerous

Several questions on the homework asked the students to write essays:

> *Please use ChatGPT, Google and whatever other tools might be useful to do this assignment. The point is not so much to solve the problems, but to learn how to use these tools effectively, and to discover their strengths and weaknesses.*

[q]https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

https://doi.org/10.1017/S1351324923000578 Published online by Cambridge University Press

In retrospect, I wish I had been more explicit about asking for fact-checking. One of the student essays contained the following paragraph:

> *During the First Opium War (1839–1842), the British government was led by the Conservative Party under Prime Minister Sir Robert Peel. The opposition, primarily the Whigs, had varying views on the war. Some opposed it on moral grounds, criticizing the ethics of trading in opium, while others were concerned about the potential impact on international relations and trade.*

This paragraph includes a number of factual errors. While the dates are correct,[r] the Conservatives were in the opposition at the time. Peel was the Prime Minister of England under a Conservative Government, but not at that time.[s]

In fact, the Opium War had little to do with opium. Neither the government (Whigs) nor the opposition (Conservatives) wanted to have anything to do with the drug trade. The Whigs had just abolished slavery and considered the drug trade to be a form of slavery. The conservatives also objected to the drug trade, though for different reasons. They viewed the drug trade as bad for business (in textiles and tea). The name of the conflict, *Opium Wars*, comes from an editorial on March 23, 1840, in the conservative newspaper: *The Times*, which argued that

> *The British would be saddled with the massive expense of an unnecessary foreign campaign that would cost far more than the entire value of the lost opium.* Platt (2019), p. 393.

The government was put in an awkward corner because, Charles Elliot, their representative in China mishandled the situation. He convinced the drug smugglers to give him their drugs in return for British IOUs, and then he handed over the drugs to the Chinese authorities for destruction. When Parliament discovered that they could not afford to make good on the IOUs, they thought they could use force to get the Chinese to pay the 2 million pounds for the lost opium.

Here is the question that I gave to the students:

1. *What were the Opium Wars?*
2. *Where did the name come from?*
3. *Summarize the conflict from multiple perspectives, including:*
   (a) *England*
   (b) *China*
   (c) *India*
   (d) *United States*
   (e) *France*
4. *In the English parliament, which party was in power at the time? Did the party in power agree with the opposition at the time? If not, what were the two positions? Listen to this.*[t] *Can you use ChatGPT and/or Google to find evidence to support Platt's description of the opposition to the Opium War in England? You may want to listen to much more of this YouTube video because it has answers to many of these questions. The bigger question is what will be the future of academic historians? Will technology make it easier for historians to do research? Or will technology replace historians?*

I was hoping the students would listen to the YouTube video (footnote [20]). That video explains Platt's description of the conservative position. More seriously, I was hoping the students would

[r]https://en.wikipedia.org/wiki/First_Opium_War
[s]https://en.wikipedia.org/wiki/Robert_Peel
[t]https://www.youtube.com/watch?v=17WF0v48vGw&t=1663s

https://doi.org/10.1017/S1351324923000578 Published online by Cambridge University Press

appreciate the difference between ChatGPT and an academic discussion by a historian that had just published a book on this topic (Platt 2019).

Most of the essays from the students repeated output from ChatGPT more or less as is. These essays tended to include a number of factual errors, but more seriously, the essays lack depth and perspective. In Platt (2019), p. 444, Platt argued that Napoleon understood that it would be foolish for Britain to use its short-term advantage in technology to humiliate the Chinese. Eventually, the Chinese would do what they have done (become stronger). Since the 1920s, these events are referred to as the "century of humiliation" by the authorities in China.[u] Platt makes it clear that the current Chinese government is using this terminology to motivate its efforts to compete with the West in technologies such as artificial intelligence. When we discussed these essays in class, I tried to argue that over-simplifying the truth, and taking the Western side of the conflict, could be dangerous and could lead to a trade war, if not a shooting war.

7. Conclusions

But despite such risks, usage of ChatGPT will almost surely continue to grow, since it is so easy to use, and so (incredibly) credible. I would be more comfortable with this reality if we encouraged more usage with humans-in-the-loop, with a better characterization of when the machine can be trusted and when humans should intervene.

We have seen that LLMs (and ChatGPT) have much to offer and can be a useful tool for students. However, there are risks. Users of LLMs, including students and everyone else, should be made aware of strengths (fluency) and weaknesses (trustworthiness). Users should be expected to do their own fact-checking before trafficking in misinformation. Laziness is inexcusable. Users are responsible for what they say, whether or not it came from a chat bot.

That said, realistically, laziness is also inevitable. Chat bots are not going away. If others are conservative with the truth and with fact-checking, then we will become conservative with belief. Credibility may disappear before chat bots go away.

References

Abbasifard M.R., Ghahremani B. and Naderi H. (2014). A survey on nearest neighbor search methods. *International Journal of Computer Applications* **95**(25), 39–52.

Amanbek Y., Singh G., Pencheva G. and Wheeler M.F. (2020). Error indicators for incompressible darcy flow problems using enhanced velocity mixed finite element method. *Computer Methods in Applied Mechanics and Engineering* **363**, 112884.

Camburu O.-M., Shillingford B., Minervini P., Lukasiewicz T. and Blunsom P. (2020). *Make up your mind! adversarial generation of inconsistent natural language explanations*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 4157–4165, Online.

Carbonell J.G. (1980). *Metaphor - a key to extensible semantic analysis*. In *18th Annual Meeting of the Association for Computational Linguistics*, *Philadelphia, Pennsylvania, USA*. Association for Computational Linguistics, pp. 17–21.

Chen M., Tworek J., Jun H., Yuan Q., de Oliveira Pinto H.P., Kaplan J., Edwards H., Burda Y., Joseph N., Brockman G., Ray A., Puri R., Krueger G., Petrov M., Khlaaf H., Sastry G., Mishkin P., Chan B., Gray S., Ryder N., Pavlov M., Power A., Kaiser L., Bavarian M., Winter C., Tillet P., Such F. P., Cummings D., Plappert M., Chantzis F., Barnes E., Herbert-Voss A., Guss W.H., Nichol A., Paino A., Tezak N., Tang J., Babuschkin I., Balaji S., Jain S., Saunders W., Hesse C., Carr A.N., Leike J., Achiam J., Misra V., Morikawa E., Radford A., Knight M., Brundage M., Murati M., Mayer K., Welinder P., McGrew B., Amodei D., McCandlish S., Sutskever I. and Zaremba W. (2021). Evaluating large language models trained on code.

Chia Y.K., Hong P., Bing L. and Poria S. (2023). Instructeval: towards holistic evaluation of instruction-tuned large language models. arXiv preprint arXiv:2306.04757.

Chomsky N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

[u]https://en.wikipedia.org/wiki/Century_of_humiliation

https://doi.org/10.1017/S1351324923000578 Published online by Cambridge University Press

Church K., Schoene A., Ortega J.E., Chandrasekar R. and Kordoni V. (2023). Emerging trends: unfair, biased, addictive, dangerous, deadly, and insanely profitable. *Natural Language Engineering*, **29**(2), 483–508.

Church K.W. and Yue R. (2023). Emerging trends: smooth-talking machines. *Natural Language Engineering* **29**(5), 1402–1410.

Cinelli M., Morales G.D.F., Galeazzi A., Quattrociocchi W. and Starnini M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* **118**. https://doi.org/10.1073/pnas.2023301118.

Dale R. (2021). Gpt-3: what's it good for? *Natural Language Engineering* **27**(1), 113–118.

Dua D., Wang Y., Dasigi P., Stanovsky G., Singh S. and Gardner M. (2019). *DROP: a reading comprehension benchmark requiring discrete reasoning over paragraphs*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, *Minneapolis, Minnesota*. Association for Computational Linguistics, pp. 2368–2378.

Fass D. and Wilks Y. (1983). Preference semantics, ill-formedness, and metaphor. *American Journal of Computational Linguistics* **9**(3-4), 178–187.

Fortuna P., Soler J. and Wanner L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing and Management* **58**(3), 102524.

Frohberg J. and Binder F. (2022). *CRASS: a novel data set and benchmark to test counterfactual reasoning of large language models*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, *Marseille, France*. European Language Resources Association, pp. 2126–2140.

Gale W.A. and Church K.W. (1991). *Identifying word correspondences in parallel texts*. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

Gedigian M., Bryant J., Narayanan S. and Ciric B. (2006). *Catching metaphors*. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, *New York City, New York*. Association for Computational Linguistics, pp. 41–48.

Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D. and Steinhardt J. (2020). Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300.

Hobbs J.R. (1992). Metaphor and abduction. In *Communication From an Artificial Intelligence Perspective: Theoretical and Applied Issues*. Berlin, Heidelberg: Springer, pp. 35–58.

Jia R. and Liang P. (2017). *Adversarial examples for evaluating reading comprehension systems*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, *Copenhagen, Denmark*. Association for Computational Linguistics, pp. 2021–2031.

Krishnakumaran S. and Zhu X. (2007). *Hunting elusive metaphors using lexical resources*. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, *Rochester, New York*. Association for Computational Linguistics, pp. 13–20.

Lakoff G. (2008). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago: University of Chicago Press.

Lakoff G. and Johnson M. (2008). *Metaphors We Live by*. Chicago: University of Chicago Press.

Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474.

Li D., Zhang Y., Peng H., Chen L., Brockett C., Sun M.-T. and Dolan B. (2021). *Contextualized perturbation for textual adversarial attack*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp. 5053–5069. Online.

Martin J.H. (1990). *A Computational Model of Metaphor Interpretation*. Cambridge, MA: Academic Press Professional, Inc.

Minsky M. and Papert S. (1969). *Perceptron: An Introduction to Computational Geometry*, vol. **19**(18), expanded edn. Cambridge: The MIT Press, p. 2.

Mohammad S., Shutova E. and Turney P. (2016). *Metaphor as a medium for emotion: An empirical study*. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, *Berlin, Germany*. Association for Computational Linguistics, pp. 23–33.

Morris J., Lifland E., Yoo J.Y., Grigsby J., Jin D. and Qi Y. (2020). *TextAttack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 119–126. Online.

Platt S.R. (2019). *Imperial Twilight: The Opium War and the End of China's Last Golden Age*. Vintage.

Poletto F., Basile V., Sanguinetti M., Bosco C. and Patti V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* **55**(2), 477–523.

Qazvinian V., Rosengren E., Radev D.R. and Mei Q. (2011). *Rumor has it: identifying misinformation in microblogs*. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, *Edinburgh, Scotland, UK*. Association for Computational Linguistics, pp. 1589–1599.

Shu K., Sliva A., Wang S., Tang J. and Liu H. (2017). Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newsletter* **19**(1), 22–36.

Shutova E. (2010). *Models of metaphor in NLP*. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden*. Association for Computational Linguistics, pp. 688–697.

Srivastava A., Rastogi A., Rao A., Shoeb A.A.M., Abid A., Fisch A., Brown A. R., Santoro A., Gupta A., Garriga-Alonso A., et al. (2023). Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Vicario M.D., Bessi A., Zollo F., Petroni F., Scala A., Caldarelli G., Stanley H.E. and Quattrociocchi W. (2016). The spreading of misinformation online. *Proceedings of The National Academy of Sciences of The United States of America* **113**(3), 554–559.

Wang B., Xu C., Wang S., Gan Z., Cheng Y., Gao J., Awadallah A.H. and Li B. (2021). Adversarial glue: a multi-task benchmark for robustness evaluation of language models. ArXiv, abs/2111.02840.

Wang J., Hu X., Hou W., Chen H., Zheng R., Wang Y., Yang L., Huang H., Ye W., Geng X., Jiao B., Zhang Y. and Xie X. (2023). On the robustness of chatgpt: an adversarial and out-of-distribution perspective. ArXiv, abs/2302.12095.

Wei J., Wang X., Schuurmans D., Bosma M., Xia F., Chi E., Le Q.V., Zhou D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* **35**, 24824–24837.

Ziegler D.M., Nix S., Chan L., Bauman T., Schmidt-Nielsen P., Lin T., Scherlis A., Nabeshima N., Weinstein-Raun B., Haas D., Shlegeris B. and Thomas N. (2022). Adversarial training for high-stakes reliability. ArXiv, abs/2205.01663.

Zubiaga A., Aker A., Bontcheva K., Liakata M. and Procter R. (2017). Detection and resolution of rumours in social media. *ACM Computing Surveys (CSUR)* **51**, 1–36.

Cite this article: Church K (2024). Emerging trends: When can users trust GPT, and when should they intervene? *Natural Language Engineering* **30**, 417–427. https://doi.org/10.1017/S1351324923000578