

EMERGING TRENDS

# Emerging trends: Smooth-talking machines

Kenneth Ward Church  and Richard Yue 

Institute for Experiential AI, Northeastern University, San Jose, CA, USA

**Corresponding author:** Kenneth Ward Church; Email: [k.church@northeastern.edu](mailto:k.church@northeastern.edu)

(Received 15 August 2023; accepted 17 August 2023)

## Abstract

Large language models (LLMs) have achieved amazing successes. They have done well on standardized tests in medicine and the law. That said, the bar has been raised so high that it could take decades to make good on expectations. To buy time for this long-term research program, the field needs to identify some good short-term applications for smooth-talking machines that are more fluent than trustworthy.

**Keywords:** Large language models; Hallucinations; ChatGPT; Responsible AI

## 1. Amazing successes

There is considerable excitement recently in large language models (LLMs). ChatGPT is amazingly fluent. There are plenty of caveats, of course, but the strengths are more obvious than the weaknesses, especially on first impression:

*ChatGPT is, quite simply, the best artificial intelligence chatbot ever released to the general public.<sup>a</sup>*

ChatGPT has been successful on standardized tests in medicine (Kung *et al.*, 2023):

*ChatGPT performs at or near the passing threshold for the US medical licensing exam. This is pretty remarkable. . . these are hard tests that doctors typically spend several years studying to be able to do.<sup>b</sup>*

ChatGPT has also been successful on bar exams.<sup>c</sup>

## 2. Back-peddling

### 2.1 Irrational exuberance and unrealistic expectations

The bar has been raised so high that it may be impossible to meet expectations. A number of articles on ChatGPT (such as the article in footnote c) lead with successes and end with cautionary caveats:

*You shouldn't expect a computer to hang a shingle. . . anytime soon, but. . . It's best to think of ChatGPT as autocorrect on steroids. . . anyone who uses the internet knows that the*

<sup>a</sup><https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html>

<sup>b</sup><https://youtu.be/J3nV3w5Y1Nk?t=139>

<sup>c</sup><https://www.enjuris.com/blog/news/chatgpt-passes-the-bar-exam/>



*internet is, well, not always accurate. What's more, we don't know precisely what information ChatGPT is being fed.*

Unfortunately, it may be impossible to back-peddle and set realistic expectations after leading with successes (that sound too good to be true).<sup>d</sup>

## 2.2 Fluency != intelligence

Many people find ChatGPT's super-human fluency super-impressive, especially on first impression. ChatGPT is so fluent that many people assume it must be "intelligent."

Of course, there is no logical reason why there should be a relationship between fluency and intelligence. One could imagine a system that does well in terms of fluency metrics such as perplexity and BLEU (Papineni *et al.*, 2002), but fails on downstream metrics such as WER (word error rate) and TER (translation error rate)<sup>e</sup> (Dorr *et al.*, 2011). Metrics often point in the same direction; in speech recognition, for example, it is common to assume a correlation between perplexity and WER (Klakow and Peters, 2002), though these days, machine learning tends to have more faith in extrinsic evaluations than intrinsic evaluations. From this perspective, perhaps we should be more impressed by extrinsic accomplishments than intrinsic fluency.

## 2.3 IQ testing and vocabulary size

Vocabulary size is even easier to measure than fluency. In educational testing, there is a well-established tradition of testing IQ by measuring vocabulary size. Here is a discussion on this topic by ChatGPT itself:

**Prompt:** *If one's vocabulary size is an indication of their intellectual capacity, can one increase their vocabulary size? As increasing your overall intelligence (IQ) is theoretically impossible.*

**ChatGPT:** *Yes, one can increase their vocabulary size. This can be done through reading, writing, and listening to a variety of texts and media. Additionally, using flashcards, playing word games, and studying vocabulary lists can be helpful. It is important to note that vocabulary size is just one aspect of intelligence, and many factors contribute to a person's overall intelligence. Additionally, intelligence is a complex concept that is not fully understood and is not always accurately measured by IQ tests.*

It is not hard to find outputs from ChatGPT that are impressively good, as well as impressively bad. It is tempting to cherry-pick in both directions, but that would be more entertaining than productive.

## 2.4 What is the difference between a hallucination and a con?

Many people are impressed by people that communicate effectively, perhaps because such people are often well-read and successful. It is natural to generalize this rule to machines, and many people will do so, for better and for worse.

Obviously, machines have different strengths and weaknesses. Machines are better than people on some tasks and worse on other tasks. No one would be impressed by a machine that is better than we are at spelling correction.

<sup>d</sup><https://aiguide.substack.com/p/did-chatgpt-really-pass-graduate>

<sup>e</sup><https://machinetranslate.org/ter>

So why are we impressed by a machine with super-human fluency? We fear the public will be disappointed when they realize that machines can be fluent, but not intelligent (or honest, truthful, trustworthy).

After the euphoria of first impressions, the public might view a smooth-talking machine as a “con”:<sup>f</sup>

*Smooth talking, soft spoken, con man*  
*Smooth talking, soft spoken, con man*  
*You stole all of my love*  
*Then you washed me off your hands*

Later verses of this song include:

*Your words were well reversed lies*  
*You make me change my mind*

and

*Now I'm smooth talking swinging too*

The song ends with:

*Smooth talkin', smooth walkin'*  
*soft spoken, slick workin'*  
*hip swinging, fast talkin' too*

Bots are optimized for fluency, not for “scout’s honor”:

*A Scout is trustworthy. A Scout tells the truth. He is honest, and he keeps his promises.*  
*People can depend on him.*<sup>g</sup>

## 2.5 Good applications for smooth-talking machines

There have been many booms and busts in Artificial Intelligence (AI). AI Winters often follow periods of “irrational exuberance” (like the current excitement with LLMs). There could be another AI Winter if we do not find a way to deal with unrealistic expectations in the near future.

About 30 years ago, we wrote a paper on good applications for crummy machine translation (Church and Hovy, 1993). Even though machine translation did not work very well at the time, we argued that it would help advance the field in the long-term to look for promising short-term use cases. We needed a few quick successes to support the field to buy time for longer-term investments in more fundamental improvements. It was clear at the time that it would take decades to make good on expectations.

Looking back on the history of machine translation, the field is in better shape now than it was then. We now have more successes and more realistic expectations. Machine translation is now good enough to be used by many people for many purposes. Of course, though there will always be plenty of opportunities for improvement, the technology is now producing demonstrable value, and users are no longer expecting magic. That said, it took the field many decades to get to where we are.

Similar comments may also apply to LLMs. The bar has been raised so high that it could take decades to make good on expectations. In the meantime, we should be on the lookout for short-term quick hits to buy time for more fundamental improvements.

<sup>f</sup><https://didapelled.bandcamp.com/track/smooth-talkin-con-man>

<sup>g</sup>[https://www.boyscouttrail.com/content/content/scout\\_law-1760.asp](https://www.boyscouttrail.com/content/content/scout_law-1760.asp)

## 2.6 You have no idea how much we're using ChatGPT

A candidate “good application for a smooth-talking machine” is to collaborate with students on essays. Owen Terry, a rising sophomore at Columbia wrote: *I'm a Student. You have no idea how much we're using ChatGPT.*<sup>h</sup> In an interview on NPR,<sup>i</sup> he identified some of ChatGPT's strengths and weaknesses. Machines are more qualified than people for some sub-tasks, and people are more qualified for other sub-tasks. To make the collaboration successful, we need to assign sub-tasks appropriately to the more qualified member of the collaboration. According to Owen Terry, 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.

Owen refers to his process as cheating, but a professor, Inara Scott, does not see it that way. She appears later in the NPR story above and praises Owen's process as a creative way to learn creative writing. She continues to suggest teachers should encourage the use of these tools in a collaborative way that takes advantage of the strengths and weaknesses of humans and machines. Machines are more fluent, but less trustworthy.

## 3. Limitations

### 3.1 Hallucinations: embarrassing computer errors

Hallucinations are one of the more serious challenges for LLMs. Hallucinations have become a nice way, in machine learning, to refer to embarrassing computer errors:

*To err is human, it takes a computer to really foul things up* (Stewart, 1985).

Some computer errors are more embarrassing than others. LLMs are designed for the average case, but a single error can cause a product to be canceled, if the error is bad enough.<sup>jk</sup>

The hallucination in Figure 1<sup>l</sup> is not that bad, but it is not good. Lesley Stahl, a reporter on the popular television show, CBS 60 Minutes, asked ChatGPT: “Who is Lesley Stahl?” CBS highlighted in yellow the incorrect assertion that she worked for NBC. CBS and NBC might seem similar to a large language model (and to many of us), but the difference is important to them.

It is natural to downplay the magnitude of these types of errors and suggest they will be fixed “in the next release.” It is possible that there will be a quick fix in the next few years, but we fear hallucinations may be more challenging than that.

Section 4.2 will review some of the history of our field, which includes periods when empiricism was more popular, as well as periods when rationalism was more popular. Both approaches have their strengths and weaknesses. LLMs are more closely aligned with empiricism, which may account for the strength with fluency, as well as the weakness with truth (hallucinations).

### 3.2 Responsible AI: risks 1.0, 2.0 and 3.0

There has been considerable criticism of deep nets on at least three grounds:

1. Hallucinations,
2. Stochastic Parrots (Bender *et al.*, 2021): *Is that all there is?*,<sup>m</sup> and
3. Responsible AI

<sup>h</sup><https://www.chronicle.com/article/im-a-student-you-have-no-idea-how-much-were-using-chatgpt>

<sup>i</sup><https://www.wbur.org/hereandnow/2023/05/22/chatgpt-academia>

<sup>j</sup><https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

<sup>k</sup><https://www.zdnet.com/article/microsoft-sued-for-racist-application/>

<sup>l</sup><https://www.youtube.com/watch?v=1wzPr4cUoMQ&t=463s>

<sup>m</sup><https://www.youtube.com/watch?v=LCRZZC-DH7M>

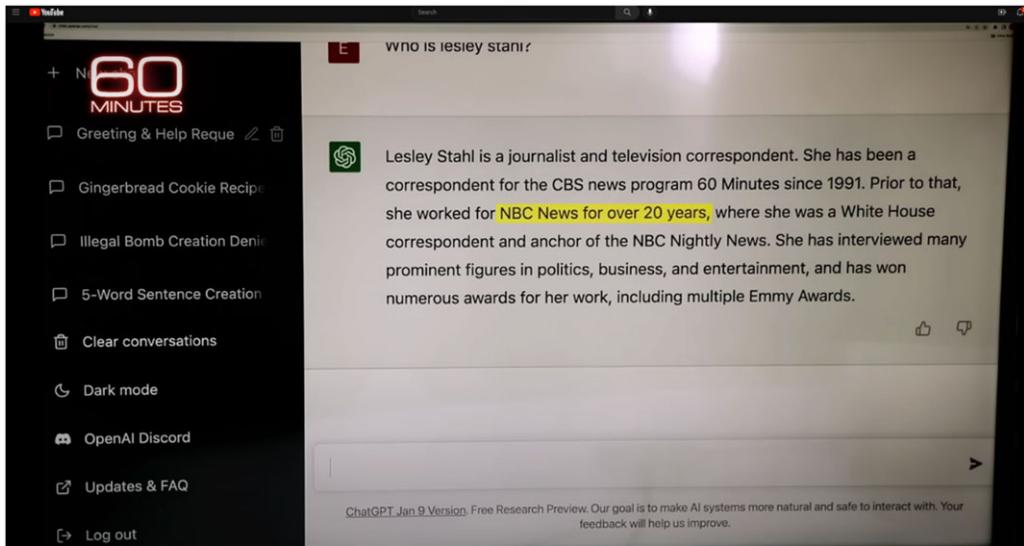


Figure 1. ChatGPT hallucinates on the CBS television show: “60 Minutes.”

Hallucinations were discussed above. This section will summarize the discussion of Responsible AI in our last two emerging trends articles (Church *et al.*, 2022; Church and Chandrasekar, 2023), where we introduced three risks:

1. Risks 1.0: unfair, biased
2. Risks 2.0: addictive, dangerous, deadly, and insanely profitable
3. Risks 3.0: proliferation of spyware

It is bad to treat people badly (Risks 1.0), but worse to kill them (Risks 2.0), and even worse to do so with malicious intent (Risks 3.0). Many of these problems involve incentives.

A major challenge for regulation is to address gaps between business cases and public interest. 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 social media companies to stop trafficking in misinformation given that it is so effective and so insanely profitable. The CBS television show, 60 Minutes, ran similar stories on whistle-blowers in tobacco companies<sup>n</sup> and social media.<sup>o</sup> In both cases, the companies appeared to know more than they were willing to share about risks to public health and public safety.

Given these incentives, attempts to build toxicity classifiers may not be effective. Social media companies have discovered that toxicity is profitable, and therefore, if we gave them a toxicity classifier, they may well use the classifier in the reverse direction to maximize toxicity in order to maximize profits.

Conflict is similar to toxicity. Both are profitable. Thus far, companies that benefit from Risks 2.0 tend to be in the social media business, and companies that benefit from conflict have been in the defense industry. But as spyware continues to proliferate, the spyware business will soon expand into other sectors. Spyware will be cheaper and more effective than divorce lawyers, and most legitimate and illegitimate ways of getting “even.”

<sup>n</sup>[https://www.youtube.com/watch?v=1\\_-Vu8LrUDk](https://www.youtube.com/watch?v=1_-Vu8LrUDk)

<sup>o</sup>[https://www.youtube.com/watch?v=\\_Lx5VmAdZSI](https://www.youtube.com/watch?v=_Lx5VmAdZSI)

### 3.3 ChatBots, ELIZA, and Responsible AI

Modern chatbots are remarkably similar to Weizenbaum's ELIZA program. The first author was a TA for Weizenbaum when he first started graduate school in 1978. Weizenbaum was horrified by how seriously people took ELIZA.

His paper on ELIZA (Weizenbaum, 1966) was written more than a decade earlier. Most of the paper describes the technical details behind ELIZA, but the paper ends with some cautionary notes, pushing back on the temptation to sell ELIZA as more than it is:

*The intent of the above remarks is to further rob ELIZA of the aura of magic. . . Seen in the coldest possible light, ELIZA is a translating processor. . .*

A recent article in the Guardian<sup>P</sup> calls out some of the similarities between Weizenbaum's ELIZA and ChatGPT:

*By the time Weizenbaum died, AI had a bad reputation. The term had become synonymous with failure. . . Getting computers to perform tasks associated with intelligence, like converting speech to text, or translating from one language to another, turned out to be much harder than anticipated.*

*Today, the situation looks rather different. We have software that can do speech recognition and language translation quite well. We also have software that can identify faces and describe the objects that appear in a photograph. This is the basis of the new AI boom that has taken place since Weizenbaum's death. Its most recent iteration is centred on "generative AI" applications like ChatGPT, which can synthesise text, audio and images with increasing sophistication.*

The article turns from these more promising accomplishments to more pessimistic remarks about Responsible AI:

*Certain of Weizenbaum's nightmares have come true. . . Weizenbaum would probably be heartened to learn that AI's potential for destructiveness is now a matter of immense concern. . . the EU is finalising the world's first comprehensive AI regulation, while the Biden administration has rolled out a number of initiatives around "responsible" AI*

## 4. Constructive suggestions for addressing hallucinations

What can be done about hallucinations? We can imagine three paths forward:

1. Give up; hallucinations demonstrate that LLMs are hopelessly flawed.
2. Use Search to Verify Assertions (Section 4.1)
3. Revive Rationalism (Section 4.2)

### 4.1 Fact-checking with search

One suggestion for dealing with hallucinations is fact-checking. Consider the hallucination in Figure 1. If one does a Google search for "Lesley Stahl works for which company," Google returns a paragraph from Wikipedia with the correct answer, CBS. As one would expect, there is no mention of the hallucination, NBC.

A challenge for this approach is to identify the assertions in ChatGPT's output that would benefit from fact-checking. A simple special case is acronyms.

<sup>P</sup><https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai>

**Table 1.** Opportunity for Fact-Checking: Translation of Acronyms

Input French		Output English		
LF	SF	LF	SF (gold)	SF (Google)
Acides gras trans	AGT	Trans fatty acids	TFA	TGA
Anémie de Fanconi	AF	Fanconi Anemia	FA	AF

**Table 2.** Search will find more documents matching the good combinations than the bad combinations

Good Combinations	Bad Combinations
Trans fatty acids (TFA)	Trans fatty acids (TGA)
Fanconi Anemia (FA)	Fanconi Anemia (AF)

Suppose we want to translate technical abstracts from French to English. It turns out that Google Translate is better at translating long forms (LFs) than short forms (SFs). The examples in Table 1 were taken from these two contexts from two abstracts:

1. *De nombreux facteurs de risque participent au développement de cette pathologie, parmi lesquels les acides gras trans (AGT).*
2. *... une diminution d'expression de 12 gènes mutés dans l'anémie de Fanconi (AF)*

Because Google is better at translating long forms than short forms, we suggest fact-checking the SFs. That is, use the English LF (the third column from Table 1) to generate some candidates for the English SF. Fact-checking takes an English LF and a candidate English SF as input and searches a large document collection (or the web) for evidence of the combination such as: “LF (SF).”

In this way, search can be used to reduce hallucinations from translation (such as the last column in Table 1). That is, search will find more documents matching the good combinations in Table 2 than the bad combinations.

More generally, search can be used to filter out many hallucinations by looking for evidence to support assertions. Acronyms, of course, are a relatively easy case. The CBS/NBC hallucination in Figure 1 is more challenging. Given an arbitrary essay from ChatGPT, it will be challenging to list the assertions that need to be fact-checked, and even more challenging to construct search queries to verify those assertions.

#### 4.2 Revival of rationalism from 1970s

The previous discussion on fact-checking is one way to deal with hallucinations. A more ambitious alternative is to revisit rationalism. When we created EMNLP in the 1990s, we were advocating a pivot away from hard problems (rationalism) toward easier problems (empiricism).

As discussed in Church (2011), there was a 10-year transition starting in the late 1980s. At the beginning of that transition, there were almost no statistical papers in our field. A decade later, there were almost no non-statistical papers. Students who started after this period may believe the field has always been empirical and will always be that way, but that is not the way it was.

In fact, Minsky and Chomsky rebelled against the previous generation (Firth, Harris), and our generation returned the favor.

- 1950s: Empiricism (Firth, Harris, Skinner)
- 1970s: Rationalism (Minsky, Chomsky)
- 1990s: Empiricism (EMNLP)

Zellig Harris was Chomsky's thesis advisor. Firth is remembered for his famous quote, which has become popular again with the revival of empirical methods.

*You shall know a word by the company it keeps* (Firth, 1957).

What was the motivation for our rebellion? Why did we revive empiricism in the 1970s? And would we take a different position to address hallucinations in LLMs?

In the late 1980s, Minsky and Chomsky had been banging their heads on fundamental issues such as AI-complete problems and long-distance dependencies. Our revival of empirical methods was motivated by pragmatic considerations, as well as frustration with the lack of progress during a severe AI Winter. The field had been attempting to do too much, and was achieving too little:

*What motivated the revival of empiricism in the 1990s? What were we rebelling against? The revival was driven by pragmatic considerations. The field had been banging its head on big hard challenges like AI-complete problems and long-distance dependencies. We advocated a pragmatic pivot toward simpler more solvable tasks like part of speech tagging. Data was becoming available like never before. What can we do with all this data? We argued that it is better to do something simple (than nothing at all). Let's go pick some low hanging fruit. Let's do what we can with short-distance dependencies. That won't solve the whole problem, but let's focus on what we can do as opposed to what we can't do.*

LLMs grew out of this pragmatic approach. This research program emphasized methods that achieved amazing fluency, but also dodged hard problems, such as the problems that Minsky and Chomsky were interested in. We avoided those problems because they are hard. But if we want to address challenges mentioned in footnote [g](#) such as truthfulness and trustworthiness, then it may be necessary to revive rationalism from the 1970s.

## 5. Conclusions

What can be done about hallucinations? We discussed three paths forward:

1. Low road: Give up; hallucinations demonstrate that LLMs are hopelessly flawed.
2. Middle road: Use Search to Verify Assertions (Section [4.1](#))
3. High road: Revive Rationalism (Section [4.2](#))

The middle road is the most promising, especially in the short term. The high road may be necessary in the long term, but it is very ambitious.

The low road is unrealistic. There have been attempts to pause AI,<sup>[9](#)</sup> but it is not clear how a pause could be effective. The good guys will play by the rules, but others will continue to behave irresponsibly. We may not like what is happening, but it is not clear how to put the genie back into the bottle.

*Is there a proper prayer for the czar?  
May God bless and keep the Czar far away from us<sup>r</sup>*

While the middle road is not as ambitious as the high road, the middle road will not be easy, and it will take considerable time. In the meantime, we need to lower expectations and find some good short-term applications for smooth-talking machines that are more fluent than trustworthy. We should position the technology as a thesaurus (rather than a magic trick).

<sup>9</sup><https://pauseai.info/>

<sup>r</sup><https://www.youtube.com/watch?v=8jZFnKZcids>

## References

- Bender E. M., Gebru T., McMillan-Major A. and Shmitchell S.** (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623.
- Church K.** (2011). A pendulum swung too far. *Linguistic Issues in Language Technology* 6(5), 1–27.
- Church K., Schoene A., Ortega J. E., Chandrasekar R. and Kordoni V.** (2022). Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable. *Natural Language Engineering* 29(2), 483–508.
- Church K. W. and Chandrasekar R.** (2023). Emerging trends: Risks 3.0 and proliferation of spyware to 50,000 cell phones. *Natural Language Engineering* 29(3), 824–841.
- Church K. W. and Hovy E. H.** (1993). Good applications for crummy machine translation. *Machine Translation* 8(4), 239–258.
- Dorr B., Olive J., McCary J. and Christianson C.** (2011). *Machine Translation Evaluation and Optimization*. New York: Springer, pp. 745–843.
- Firth J. R.** (1957). A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*. Oxford: Blackwell.
- Klakow D. and Peters J.** (2002). Testing the correlation of word error rate and perplexity. *Speech Communication* 38(1–2), 19–28.
- Kung T. H., Cheatham M., Medenilla A., Sillos C., De Leon L., Elepaño C., Madriaga M., Aggabao R., Diaz-Candido G., Maningo J., Tseng V. and Dagan A.** (2023). Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS Digital Health* 2(2), e0000198.
- Papineni K., Roukos S., Ward T. and Zhu W.-J.** (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA: Association for Computational Linguistics, pp. 311–318.
- Stewart K. K.** (1985). Editorial: 'To err is human, it takes a computer to really foul things up!'. *Journal of Automatic Chemistry* 7(4), 169.
- Weizenbaum J.** (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1), 36–45.