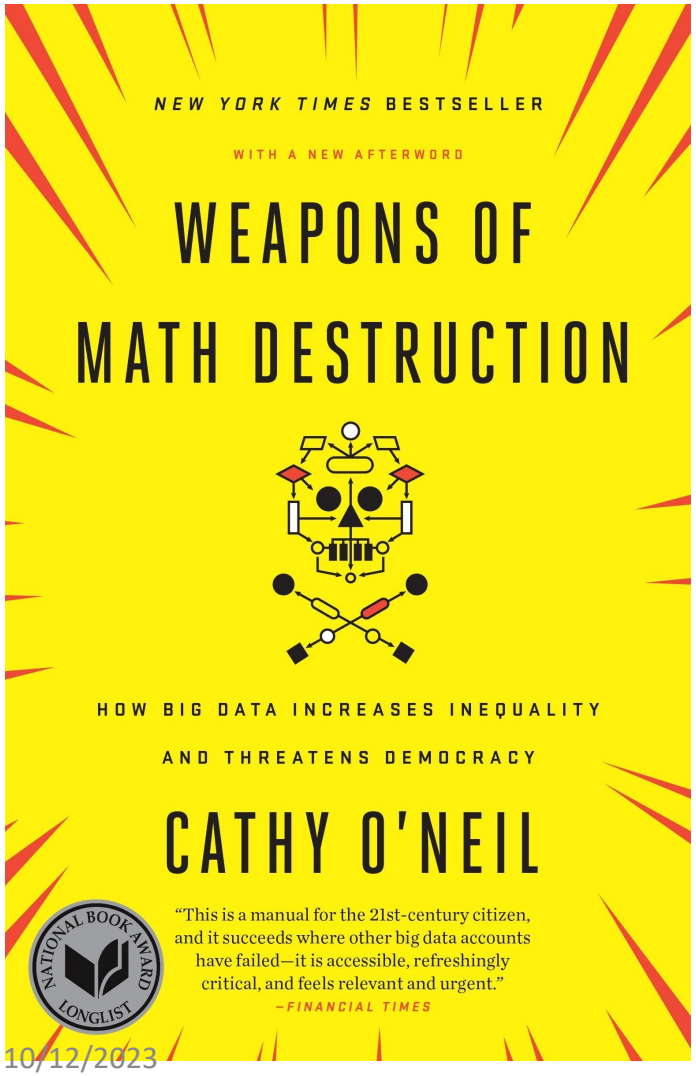
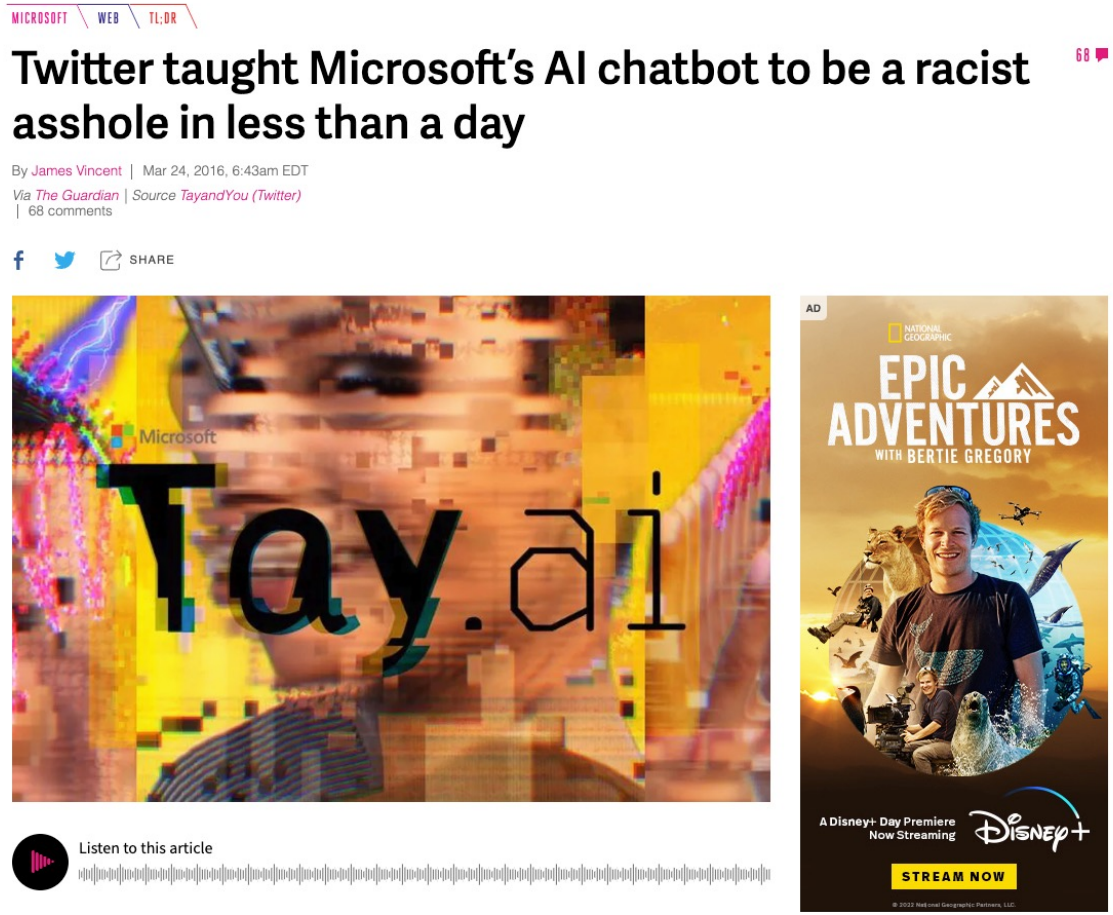
The Easy, the Hard and the Ugly

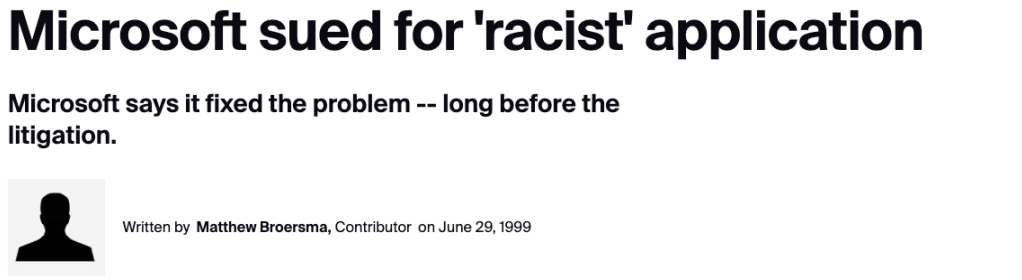
- ✓Easy
  - ✓Inference (*fit*)
  - ✓Fine-Tuning (*predict*)
- ✓Hard
  - ✓Pre-training
- ➢Ugly (Responsible AI)
  - Bias
  - Toxicity
  - Misinformation
  - Hallucinations
  - Plagiarism

History of Irresponsible AI

Risk (5 years ago) Product gets canceled



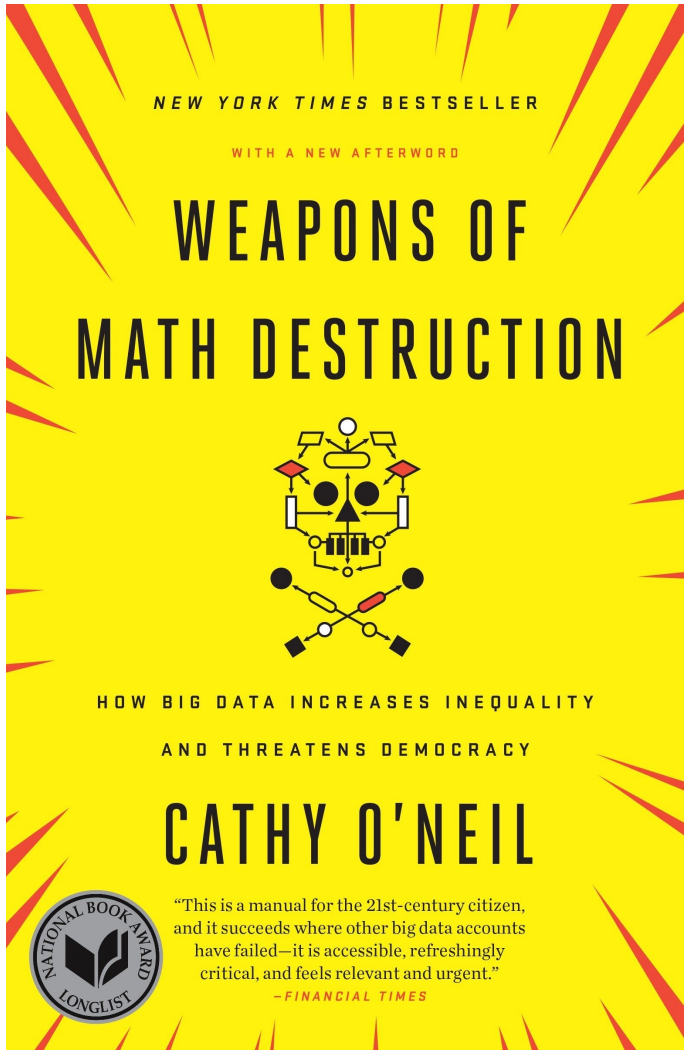




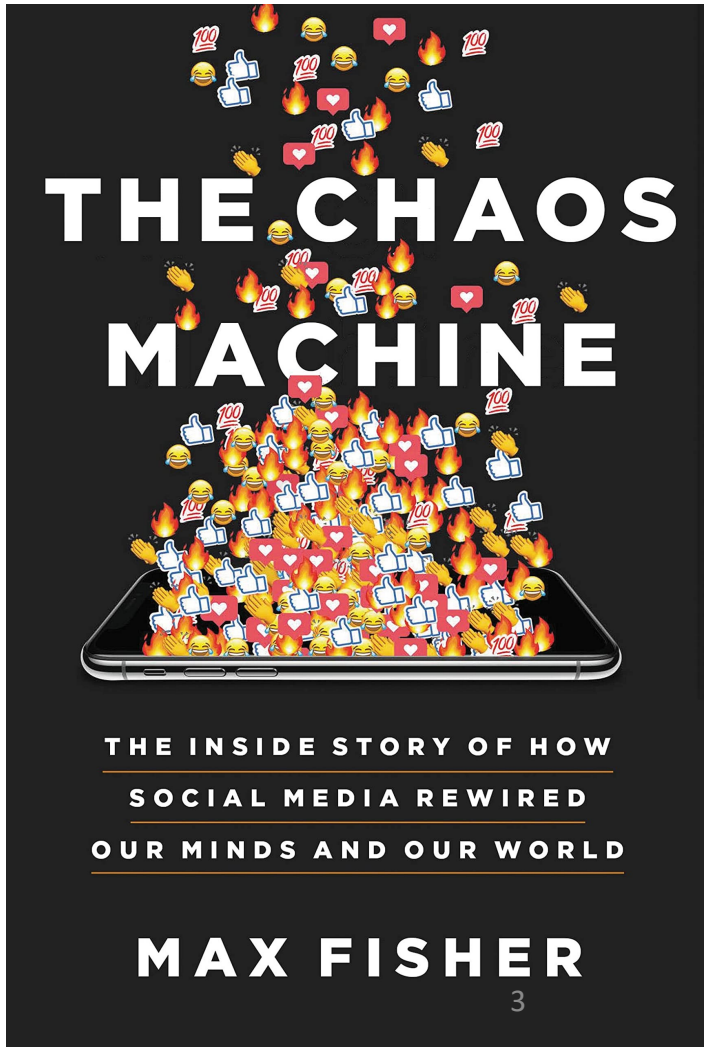
Are we losing ground??? Are we at fault???

- Old risks: (work in progress)
  - Bias, Fairness
- New risks: (bigger than us)
  - Genocide, Insurrection
  - Root causes:
    - ML + Social Media → Addiction
    - Max Engagement → Dangerous
    - Insanely profitable:
      - Companies & Countries
  - Long book, but no mention of our efforts to address old risks

2016

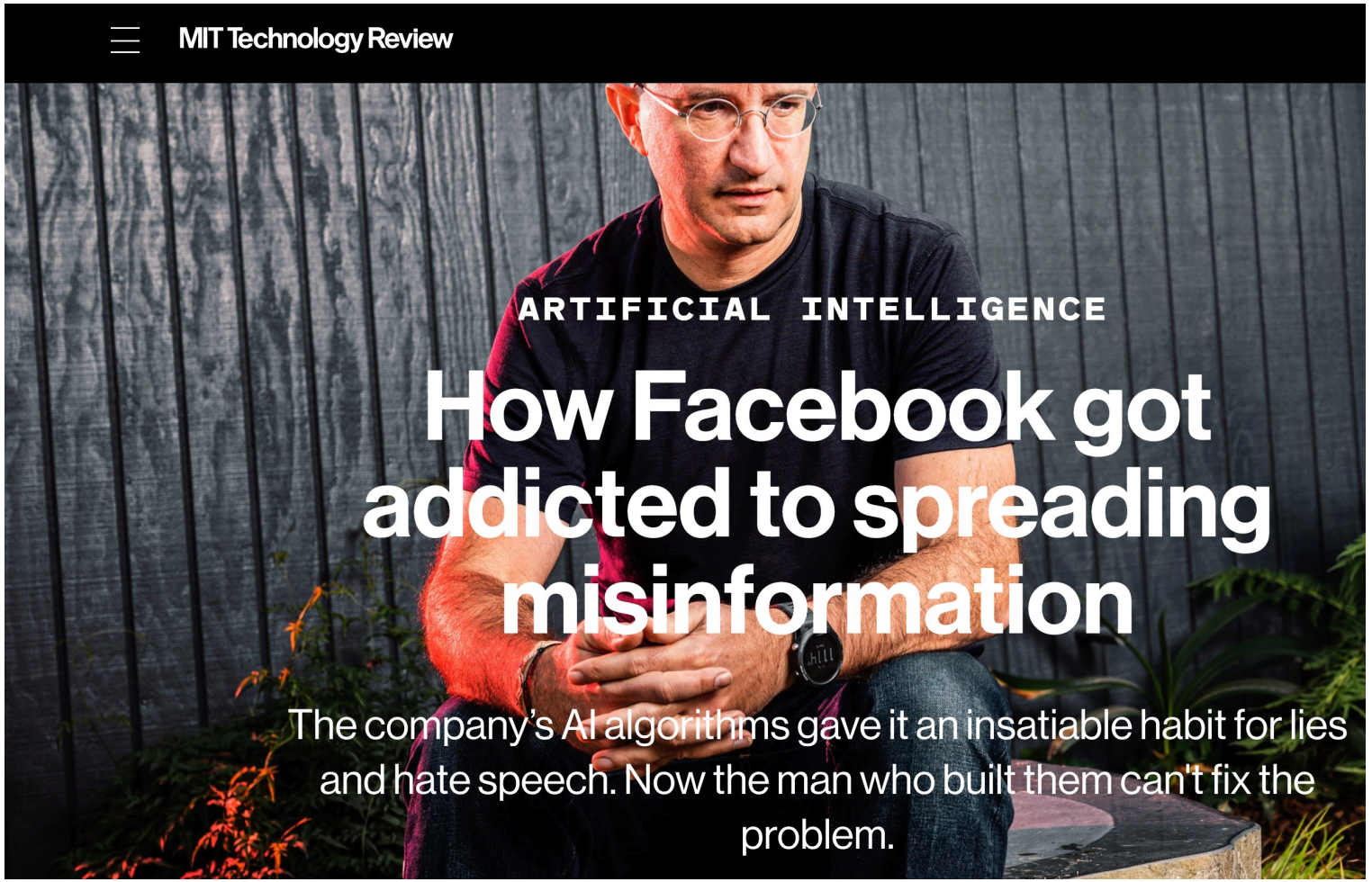


2022



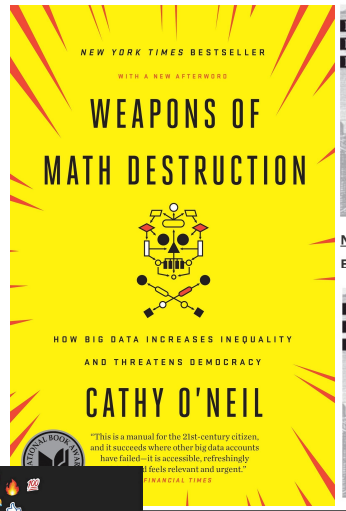
Reporter wanted to talk about new risks; Accused Facebook of pivoting to old risks

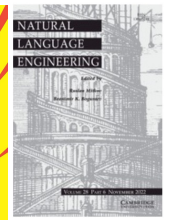
https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/



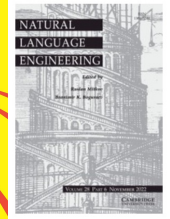
Ugly: Responsible AI

- Incentives matter
  - Risks 1.0 (2016)
    - Unfair, Biased
  - Risks 2.0 (2022)
    - Addictive, dangerous, deadly
    - and insanely profitable
  - Risks 3.0 (2023)
    - Malware
    - Spyware
- Challenge for Regulation
    - Business case ≠ Public Interest (Health, National Security)
  - Tobacco companies maximize sales; ditto for fast food & junk food
  - Risks 2.0 (Toxicity): Good for social media companies; we ❤️ click bait
  - Risks 3.0 (Conflict): Good for defense industry





Natural Language Engineering



Emerging trends: Unfair, biased, addictive, dangerous, deadly, and insanely profitable

Published online by Cambridge University Press: **19 December 2022**

Kenneth Church, Annika Schoene, John E. Ortega, Raman Chandrasekar and Valia Kordoni

Show author details

Article | Figures | Metrics

Save PDF | Share | Cite | Rights & Permissions

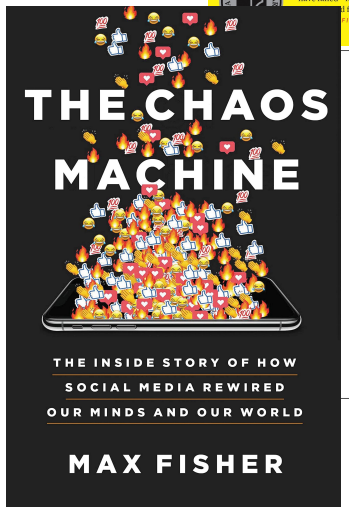
Emerging trends: Risks 3.0 and proliferation of spyware to 50,000 cell phones

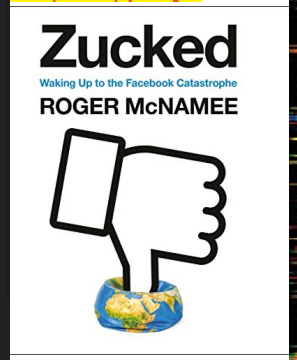
Published online by Cambridge University Press: **19 May 2023**

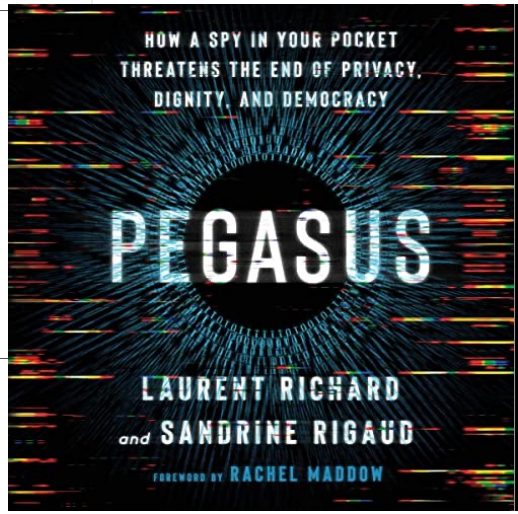
Kenneth Ward Church and Raman Chandrasekar

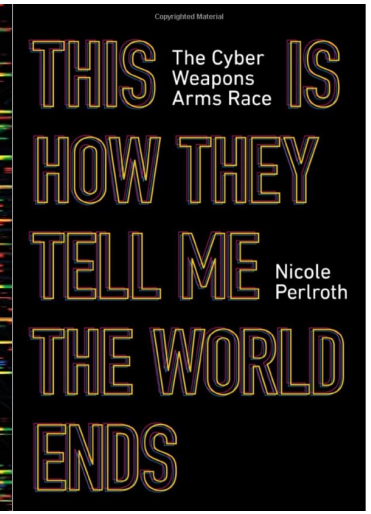
Show author details

Article | Metrics









Winter is Coming

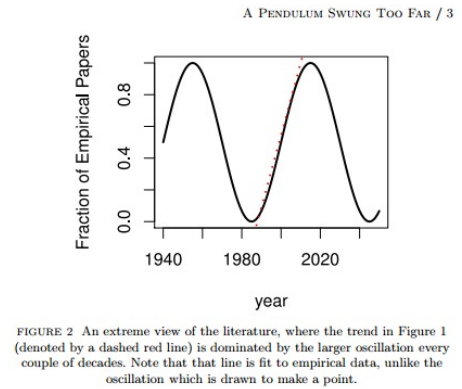


FIGURE 2 An extreme view of the literature, where the trend in Figure 1 (denoted by a dashed red line) is dominated by the larger oscillation every couple of decades. Note that that line is fit to empirical data, unlike the oscillation which is drawn to make a point.

- Pendulum Swung Too Far
  - There have been many AI Winters
  - Often, after ``irrational exuberance''
  - (like current excitement with nets)

- We tend to be impressed by people that speak/write well
  - Fluency → well-read → success → smart
- Machines are better than people on many tasks (spelling),
  - Now that machines are more fluent than people, ***are they smarter***?
- Fear: AI Winter
  - there will be disappointment
  - when public realizes
    - ***fluency ≠intelligence***

Ugly: Outline

- **Benchmarking** (Omar)
- Smooth-Talking Machines/Trust (Ken)

Evaluating Evaluations: A Perspective on Benchmarks

https://github.com/kwchurch/WSDM_2024_tutorial

The importance of evaluation

- More and more benchmarks, data sets and evaluations available
- Industry ships a new product/feature …
  … users test it, high expectations, harsh reaction …
  … back to the lab to fix things, iterate; very expensive process
- As a community:
  - Are we really evaluating the right stuff?
  - Are we using solid principles to construct and maintain benchmarks?
  - What are we learning?

Principles (Spoiler Alert)

- Characterize use case & audience
- Validity:
  - Relevance of task to use case
- Reliability:
  - Inter-rater agreement
- Realistic workloads
- Labeling and annotation
  - Documentation:
    - How was it done?
    - Availability of instructions
- Maintenance
  - Include a feedback loop mechanism to maximize adoption
  - Workload Evolution
  - Lessons learned and addendum(s)
  - Lifecyle and deprecation
- High standards
  - (for high-stakes use cases)
    - Use of established software engineering and data management techniques
      - (e.g., code review, versioning, configurations, dependencies, and testing).
    - How was the data sourced?
      - Provenance?
      - Can we data set be generated again easily?
    - Clean and well-documented data model

Validity and Reliability

Krippendorff (2018) *Content analysis: An introduction to its methodology*

- Reliability is about data, and validity is about truth:
  - Reliability:
    - The attribute of Data on which researchers can rely in answering their Research questions – Krippendorff, p. 411
  - Validity:
    - The quality of a claim to be as stated, true, or correct. – Krippendorff, p. 413
- Validity assumes a hypothesis/claim.
  - Hard to test validity without a clearly stated hypothesis
- There are more papers on reliability than validity in our field(s).
  - 18k matches for “inter-annotator agreement” in ACL Anthology
  - https://aclanthology.org/search/?q=inter-annotator+agreement

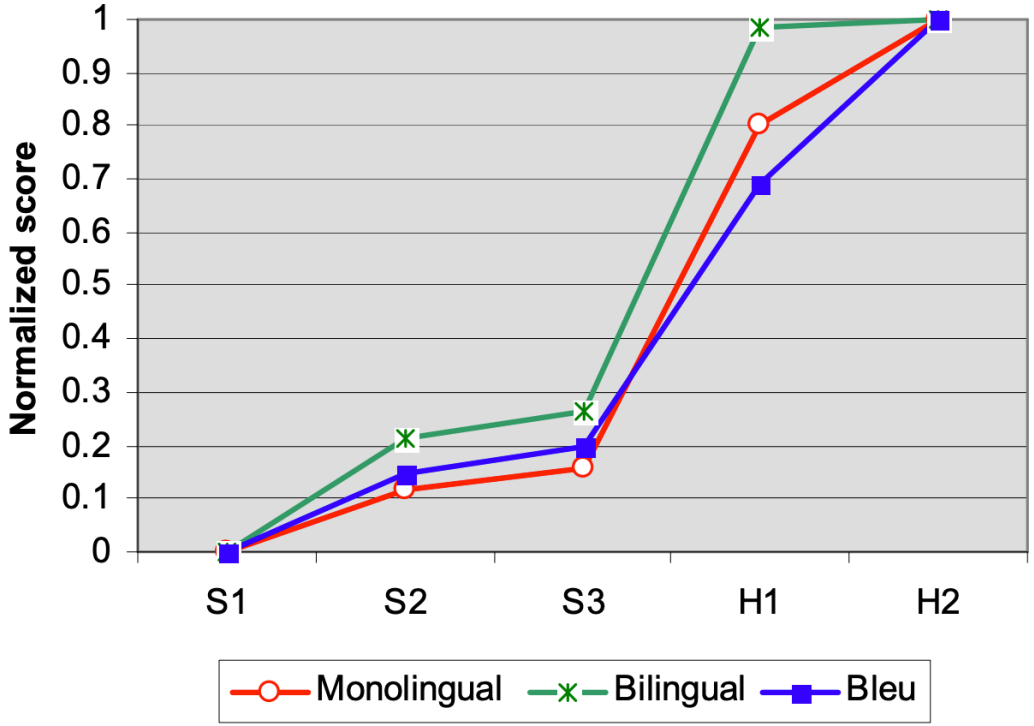
Reliability of BLEU https://aclanthology.org/P02-1040/

- The original paper asks 3 excellent questions about reliability
  - How reliable is the difference in BLEU metric?
  - What is the variance of BLEU score?
  - If we were to pick another random set of 500 sentences,
    - would we still judge S3 to be better than S2?

Reliability of BLEU

- In their Figure 7,
  - they use BLEU to compare three machine translation systems (S1, S2 and S3) and
  - two human (non-professional) translators (H1 and H2).
- They report that humans score better than machines.
- They conclude:
  - "BLEU's strength is that it correlates highly with human judgments
    - by averaging out individual sentence judgment errors over a test corpus
    - rather than attempting to divine the exact human judgment for every sentence:
    - *quantity leads to quality*."

Figure 7: BLEU vs Bilingual and Monolingual Judgments



https://aclanthology.org/P02-1040/

Validity of BLEU

- Less discussion of validity in original paper
- Some papers raise serious questions about validity
  - of BLEU and other metrics
  - at least, for some use cases that go beyond the original BLEU paper

- It is hard to talk about validity without a clearly stated use case
  - https://aclanthology.org/D17-1238/
    - *This paper shows that state-of-the-art automatic evaluation metrics for NLG systems*
      - *do not sufficiently reflect human ratings,*
      - *which stresses the need for human evaluations.*
  - https://aclanthology.org/J18-3002
    - *Overall, the evidence supports using BLEU for diagnostic evaluation of MT systems*
      - *(which is what it was originally proposed for),*
    - *but does not support using BLEU outside of MT, for evaluation of individual texts, or for scientific hypothesis testing.*

Use Cases

- BLEU was originally proposed
  - to compare a small number of systems for DARPA competitions.
- But soon after BLEU was introduced,
  - Och suggested using BLEU for a very different use case.
- It had been standard practice to to use
  - different metrics for testing and training.
- Och found that if one is going to use BLEU to evaluate systems,
  - then his system would do better in the evaluation if he also used BLEU for training.
- Och's suggestion did well in competitions,
  - but raises questions about reliability and validity

Consequence of Evaluation: Proposed Scale

- Minor: e.g., SOTA-chasing, leaderboards
- Moderate: e.g., Multitask learning
  - Generalizing results over workloads and tasks.
- Major: e.g., Och training
  - Significant consequences for system performance
- Mission Critical: e.g., SPEC
  - What should I buy? And what performance should I expect on my workloads?
  - Go/No-go decisions

Averaging: Arithmetic vs. Geometric

- There are many benchmarks in our field:
  - e.g., GLUE, SciRepEval, MS MARCO and Big Bench
- Many designed for SOTA-chasing,
  - but hopefully, results will generalize to more important use cases.
- More likely to generalize to more important use cases
  - if they were designed to do so in the first place

- SPEC was designed to report performance relative to baseline
  - How much better is the candidate CPU relative to VAX 11/780?
  - On the user's (unspecified) workload?
- Mashey argues that
  - geometric means generalize better
  - over workloads than arithmetic means.
- Mashey suggests that results on our benchmarks would
  - generalize beyond less important SOTA-chasing use cases
  - if we replaced arithmetic means with geometric means in GLUE
  - (and many of our other benchmarks).

SPEC: A benchmark for evaluating CPUs

Code	App Area	Lines	Remarks
gcc	Compiler	87,800	CNU C Compiler V1.35, compiles 76 soruces, 10% I/O
Espresso	Logic Design	14,800	PALs generation tool, heuristic minimization, little paging
Li	Interpreter	7700	Lisp interpreter (XLIST 16), solves 8-queens problem using recursive backtracking, many jumps/loops
Eqntott	Logic design	3500	Creates truth tables; >95% of time in qsort
Compress	Data compression	1500	Compress/decompress 1MB file 20 times using adaptive Lempel-Ziv coding
Sc	Spreadsheet	8500	Spreadsheet app based on the Unix "curses"

Table 1: SPEC CINT92 suite (from Table 2 in [8])

Code	App Area	Lines	Remarks
Spice2g6	Circuit Design	18,900	Analog circuit simulation tool, unvectorizable, unparallelizable, uses cache
Doduc	Physics, simulation	5300	Monte Carlo simulation of thermohydraulic neclear reactor, unvectorizable, many jumps/loops
Fpppp	Quantum chemistry	2700	Electron integral, unvectorizable, no jumps (good for pipelining)
Tomcatv	Geometry	200	Mesh generation, 90-98% vectorizable, exercises data cache
Nasa7	Kernels	1300	Some kernels are vectorizable
Mdljdp2	Chemistry	4500	Motion equations for 500-atom model
Wave5	Physics	7600	Particle and Maxwell's equations
Ora	Optics	500	Ray Tracing
Alvinn	Robotics	300	Neural network propagation training
Ear	Medicine	5200	Ear simulation using FFT
Mdljsp2	Chemistry	3900	Single precision of Mdljdp2
Swm256	Simulation	500	Shallow water equations system
Su2cor	Quantum physics	2500	Masses of elementary particles, 98.5% vectorizable
Hydro2d	Astrophysics	4500	Calactic jets, 99.5% vectorizable

Table 2: SPEC CFP92 suite (from Table 3 in [8])

Challenges for Validity and Reliability

- Unrepresentative Samples
- Test/Train Splits: Interpolation vs. Extrapolation
- Leakage
- Labeling and Inter-annotator Agreement
- Too many (irrelevant) tasks

Unrepresentative Samples

- Ideally, a benchmark should be
  - a representative sample
  - of a larger population of interest.
- Balanced Corpora:
  - 1960s: Brown Corpus
  - 1990s: British National Corpus
- Current view: catch as catch can
  - *there is no data like more data* – Mercer
- Representative samples → More credible generalizations

	Semantic Scholar		Citation Task	
	%	N	%	N
$\|A\|$	47%	99M	82%	21,885
$\|L\|$	53%	111M	97%	25,850
$\|A \cup L\|$	70%	145M	99%	26,378
$\|A \cap L\|$	31%	65M	80%	21,357
totals	100%	208M	100%	26,657

Table 3: Comparison of Semantic Scholar with a benchmark (SciRepEval Cite task). There are large mismatches in $|A|$ (papers with abstracts) and $|L|$ (papers with links in G).

Test/Train Splits: Interpolation ≠ Extrapolation

It's Difficult to Make Predictions, Especially About the Future

- It is common for graphs to evolve over time.
  - For example, the academic literature is growing very quickly,
  - doubling every nine years
- Benchmarks such as OGB focus on static snapshots from a few years ago,
  - missing opportunities to encourage the community to study growth and timeliness.
- Random splits are common in graph learning benchmarks, e.g., WN18RR,
  - a popular Knowledge Graph Completion (KGC) task based on WordNet

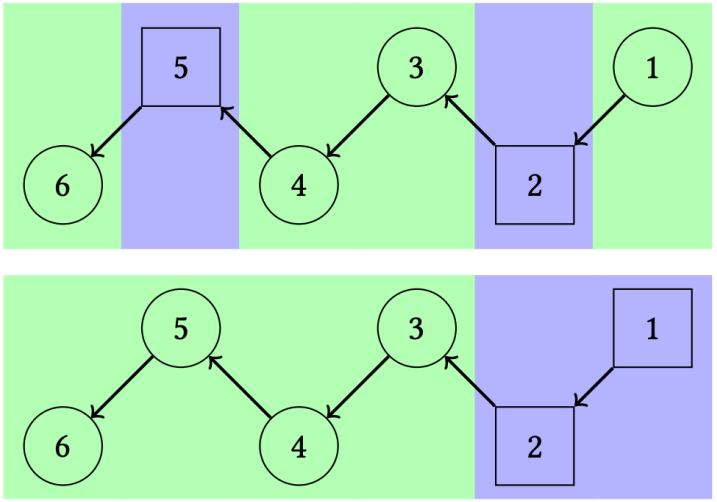


Figure 1: Random Splits (top) vs Causal Split (bottom).

Paper	Year	Title
1	2018	[...] Photograment imaging
2	2016	Convenient probe of S(1D2)[...]
3	2005	Megapixel ion imaging [...]
4	2003	Direct current slide imaging [...]
5	1995	profiles of CI(2Pj) photoframents [...]
6	1988	Adiabatic dissociation of [...]

Table 4: 1 cites 2, 2 cites 3,..., 5 cites 6

Leakage

- Leakage is common in many benchmarks
  - There is considerable discussion of leakage in SciDocs
    - (see 4.2 of https://aclanthology.org/2022.emnlp-main.802/)
  - WN18 → WN18RR (WordNet benchmarks)
    - If x is-a y (a car is a vehicle),
      - then there will be two links between *x* and *y*: hypernym and hyponym
    - Since WN18 randomly splits links into test and train,
      - one of these links is likely to be in test and the other in train
    - Unfortunately, WN18RR corrects some (but not all) of the leakage
      - See table 4 of https://aclanthology.org/2021.emnlp-main.501/
    - Despite this leakage, there are many papers on WN18RR
      - https://paperswithcode.com/dataset/wn18rr

Labeling and Inter-annotator Agreement

- The documentation on SciRepEval makes it clear that some labels are “silver” (less reliable) [underlining added]:
  - *... a new large-scale field of study (FoS) multi-label training set of more than 500K papers with **silver** FoS labels based on publication venue*
- We compared FoS labels in SciRepEval with FoS labels in MAG and found large differences.
  - More agreement in some fields (Computer Science)
    - Less agreement in History, Sociology and Art.
  - It is possible that the annotators
    - are more familiar with Computer Science
    - than History, Sociology and Art.

Too many (irrelevant) tasks

- SciRepEval example
- There are so many tasks that
  - some will be more relevant to our use cases,
  - and others will be less relevant
- The FoS task, for example, classifies documents into 23 fields of study.
  - The FoS task is probably not relevant to recommendation use cases

Principles (Conclusion)

- Characterize use case & audience
- Validity:
  - Relevance of task to use case
- Reliability:
  - Inter-rater agreement
- Realistic workloads
- Labeling and annotation
  - Documentation:
    - How was it done?
    - Availability of instructions
- Maintenance
  - Include a feedback loop mechanism to maximize adoption
  - Workload Evolution
  - Lessons learned and addendum(s)
  - Lifecyle and deprecation
- High standards
  - (for high-stakes use cases)
    - Use of established software engineering and data management techniques
      - (e.g., code review, versioning, configurations, dependencies, and testing).
    - How was the data sourced?
      - Provenance?
      - Can we data set be generated again easily?
    - Clean and well-documented data model

Ugly: Outline

- ✓ Benchmarking (Omar)
- **Smooth-Talking Machines/Trust** (Ken)

Natural Language Engineering (2023), **29**, pp. 1402–1410
doi:10.1017/S1351324923000463

CAMBRIDGE UNIVERSITY PRESS

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

(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

Too Good to be True

- Standardized tests in medicine, law, etc.
- Press
  - *ChatGPT is, quite simply, the best artificial intelligence chatbot ever released to the general public.* -- New York Times
- Back-peddling
  - *You shouldn't expect a computer to hang a shingle... anytime soon, but...*
  - *It's best to think of ChatGPT as autocomplete on steroids...*
  - *anyone who uses the internet knows that*
    - *the internet is, well, not always accurate.*
  - *What's more, we don't know precisely what information ChatGPT is being fed.*

Fluency ≠ Intelligence

- People want to believe in chatbots

- What is the difference between a hallucination and a con?

https://didapelled.bandcamp.com/track/smooth-talkin-con-man

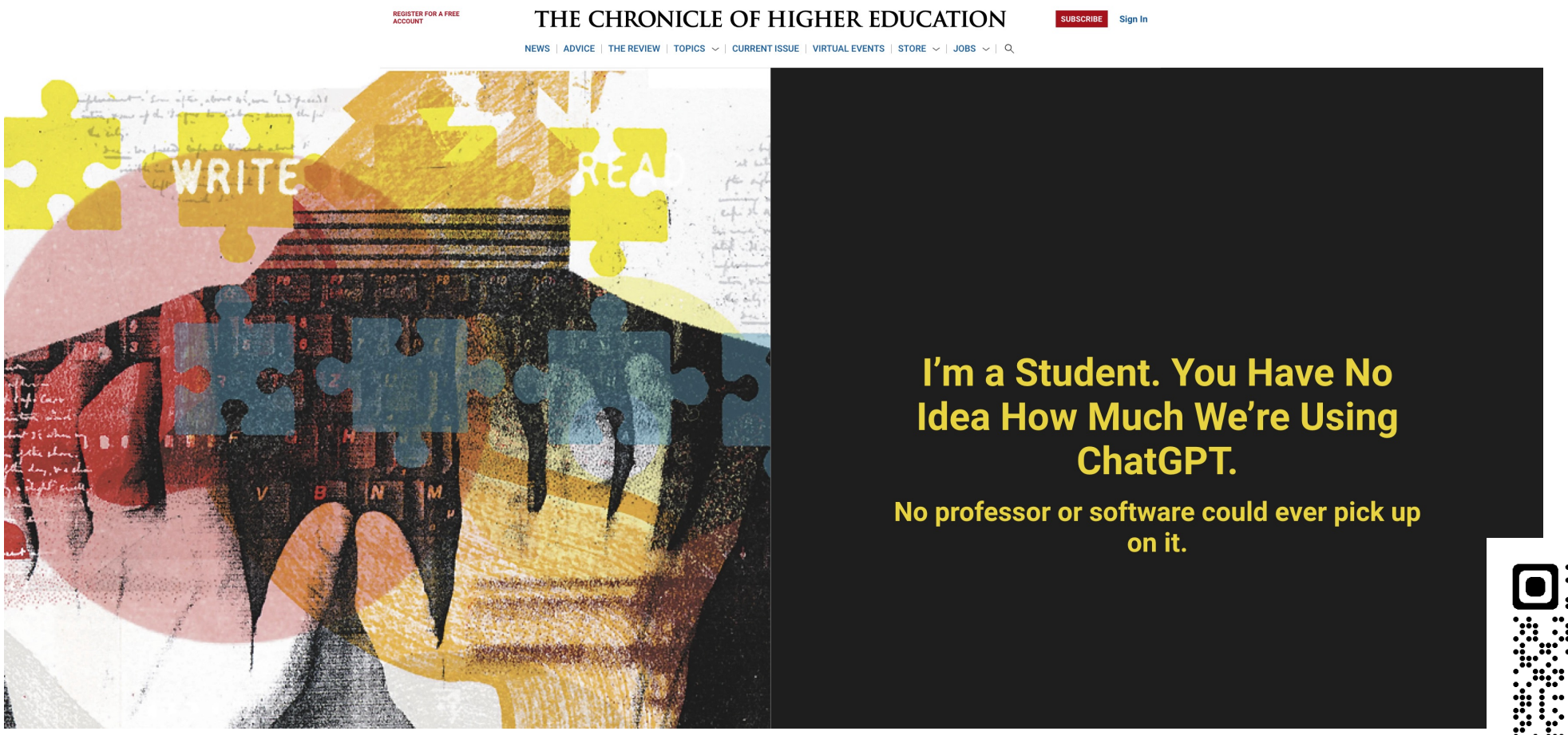
Smooth talking, soft spoken, con man
Smooth talking, soft spoken, con man
You stole all of my love
Then you washed me off your hands

Your words were well reversed lies
You make me change my mind
…
Now I'm smooth talking swinging too
…
Smooth talkin', smooth walkin'
soft spoken, slick workin'
hip swinging, fast talkin' too

Good Applications for Smooth-Talking Machines

- There have been booms and busts in AI
  - Busts → AI Winters
- *Good Applications for Crummy Machine Translation* – Church & 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.
- Similar comments 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.

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



https://github.com/kwchurch/WSDM_2024_tutorial

Essay Writing Subtasks:
Human-machine Collaboration

Chatbots are good at

- Thesis statements
- Outlines

Chatbots are not good at

- Capturing student’s voice
- Quotes (makes them up)

Is it cheating to use a chatbot?

Yes

- According to Owen Terry, author of
  - *You have no idea how much we're using ChatGPT*

No

- A professor, Inara Scott, on NPR
  - https://www.wbur.org/hereandnow/2023/05/22/chatgpt-academia
  - Creative way to learn creative writing

Good applications:
Take advantage of strengths and avoid weaknesses

Strengths

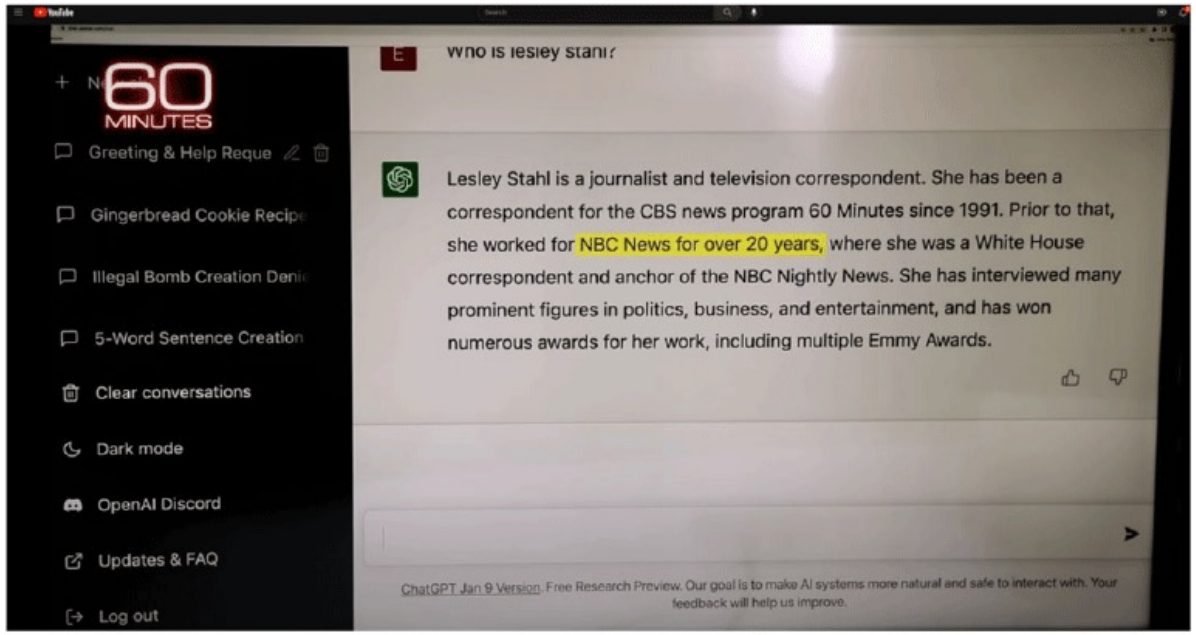
- Fluency

Thesaurus

Complement with fact-checking

Weaknesses

- Hallucinations



Smooth-Talking Conclusions

- Low Road: Give up; hallucinations are too hard
- ➢**Middle Road**: Use search to verify assertions (fact-checking)
- High Road: Revive Rationalism
  - 1950s: Empiricism (Firth, Harris, Skinner)
  - 1970s: Rationalism (Minsky, Chomsky)
  - 1990s: Empiricism (EMNLP)

Natural Language Engineering (2024), 1–11
doi:10.1017/S1351324923000578

CAMBRIDGE UNIVERSITY PRESS

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)

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

https://github.com/kwchurch/WSDM_2024_tutorial

Evaluation of Human-in-the-Loop
Homework: use tools to help them write essays

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

https://github.com/kwchurch/WSDM_2024_tutorial

Use Cases:
Some work well and some do not

Amazingly Good

- *cover all the bases*
- *drop the ball*
- *dunk*
- *fumble*
- *get on base*
- *hit a home run*
- *out in left field*
- *ragging the puck*
- *sticky wicket*

Amazingly Bad

- Made-up Reference
  - Title: “A Survey of Nearest Neighbor Search Algorithms”
  - Authors: Yufei Tao, Dongxiang Zhang
  - Link: Survey Paper (link to https://arxiv.org/abs/1904.06188)

- *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.*

- Factual Errors
- Lack of Depth/Perspective

- Question for students
  - *What were the Opium Wars?*
  - *Where did the name come from?*
  - *Summarize the conflict from multiple perspectives, including:*
    - *England*
    - *China*
    - *India*
    - *United States*
    - *France*

In fact, the Opium War had little to do with opium

- Neither the government (Whigs) nor the opposition (Conservatives)
  - supported drugs.
- The Whigs had just abolished slavery
  - and viewed drugs to be a form of slavery.
  - The conservatives viewed drugs 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 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 did not want to make good on the IOUs,
  - they thought they could force
  - the Chinese to pay for the lost opium.

Lack of Perspective → Danger
Chatbots ≠ Historian (Platt)

- Most of the essays from the students repeated output from ChatGPT more or less as is.
- These essays contained 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.
- 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 AI
- 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