Why Large Language Models Appear to be Intelligent and Creative: Because They Generate Bullshit!

Daniel M. Berry,
University of Waterloo

The Hype

These days, many are falling for the AI hype

and are proclaiming that

large-language models (LLMs),

such as ChatGPT,

are truly intelligent and creative.

Examples:

“October 13, 2023. ChatGPT changes everything! This and other smooth-talking artificial intelligences will soon be sentient! If they’re not already!” [reported by John Horgan]

and

“… we’ve reached a momentous point. Large language models, or LLMs, can often seem to wield something close to human intelligence, at least to us non-experts.” [Yejin Choi]

Hype Despite Debunkers

This hype is despite the patient, careful explanations by tech-savvy debunkers.

Even I Almost Believed

Even I, programming since 1965,

upon seeing ChatGPT in operation,

thought

“Finally, here’s an AI that might actually be intelligent!”

I had to slap myself across the face and think carefully

to push that thought aside.

Reality

After all, ChatGPT is just a learned machine (LM)

trained on a humongous database of

unvalidated crap that is found

out there in the Internet,

Reality, Cont'd

It's programmed to construct sentences

that have a very high probability of looking like

the typical native-English speaker's almost-grammatically-correct writing

Oy!!!!

The question remains …

Why do so many, even tech-savvy, people

perceive

LLMs and their chatbots

to be intelligent and creative?

Two Recent Publications

Based on two recent publications,

Hicks et al
Turpin and Kara-Yakoubian et al,

I think I have put my finger on it.

These two publications report research that assumes Harry Frankfurt's 2005 definitions of "bullshit" and of "soft bullshit".

ON BULLSHIT

Harry G. Frankfurt

PRINCETON UNIVERSITY PRESS

PRINCETON AND OXFORD

Frankfurt on BS

“It is just this lack of connection

to a concern with truth —

this indifference to how things really are —

that I regard as of the essence of bullshit.”

Hard vs. Soft BS

Hard BS: with purposeful intent

Soft BS: with no particular intent, e.g., shooting the bull

Ethics and Information Technology (2024) 26:38
https://doi.org/10.1007/s10676-024-09775-5

ORIGINAL PAPER

ChatGPT is bullshit

Michael Townsen Hicks[1] · **James Humphries**[1] · **Joe Slater**[1]

Published online: 8 June 2024

Abstract
Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called "AI hallucinations". We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

Keywords Artificial intelligence · Large language models · LLMs · ChatGPT · Bullshit · Frankfurt · Assertion · Content

Introduction

Large language models (LLMs), programs which use reams of available text and probability calculations in order to create seemingly-human-produced writing, have become increasingly sophisticated and convincing over the last several years, to the point where some commentators suggest that we may now be approaching the creation of artificial general intelligence (see e.g. Knight, 2023 and Sarkar, 2023). Alongside worries about the rise of Skynet and the use of LLMs such as ChatGPT to replace work that could and should be done by humans, one line of inquiry concerns what exactly these programs are up to: in particular, there is a question about the nature and meaning of the text produced, and of its connection to truth. In this paper, we argue against the view that when ChatGPT and the like produce false claims they are lying or even hallucinating, and in favour of the position that the activity they are engaged in is bullshitting, in the Frankfurtian sense (Frankfurt, 2002, 2005). Because these programs cannot themselves be concerned with truth, and because they are designed to produce text that *looks* truth-apt without any actual concern for truth, it seems appropriate to call their outputs bullshit.

We think that this is worth paying attention to. Descriptions of new technology, including metaphorical ones, guide policymakers' and the public's understanding of new technology; they also inform applications of the new technology. They tell us what the technology is for and what it can be expected to do. Currently, false statements by ChatGPT and other large language models are described as "hallucinations", which give policymakers and the public the idea that these systems are misrepresenting the world, and describing what they "see". We argue that this is an inapt metaphor which will misinform the public, policymakers, and other interested parties.

The structure of the paper is as follows: in the first section, we outline how ChatGPT and similar LLMs operate. Next, we consider the view that when they make factual errors, they are lying or hallucinating: that is, deliberately uttering falsehoods, or blamelessly uttering them on the basis of misleading input information. We argue that neither of these ways of thinking are accurate, insofar as both lying and hallucinating require some concern with the truth of their statements, whereas LLMs are simply not designed to accurately represent the way the world is, but rather to

✉ Michael Townsen Hicks
Michael.hicks@glasgow.ac.uk

James Humphries
James.Humphries@glasgow.ac.uk

Joe Slater
Joe.Slater@glasgow.ac.uk

[1] University of Glasgow, Glasgow, Scotland

Hicks et al

Hicks et al consider ChatGPT to be a machine that generates soft BS in the Frankfurtian sense.

ChatGPT is designed to generate cogent text

that reads as though it was written by a native-English-speaking human being.

Hicks et al, Cont'd

There is no requirement that the generated text

bears any relation to the truth.

The data from which ChatGPT's LLM learns

are not vetted for truth.

Hicks et al, Cont'd

Thus, the LLM is indifferent —

even careless —

as to the truth.

It BSs.

An LLM has *no* intent.

So it's a soft BSer

OpenAI and Haigh Agree

OpenAI, the creators of the LLM, say

“ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers.”

and thus admit that ChatGPT is indifferent to the truth.

This observation was independently confirmed by Haigh.

DOI:10.1145/3708554

Thomas Haigh

Historical Reflections

Artificial Intelligence Then and Now

From engines of logic to engines of ~~bullshit~~?

Measuring Bullshit in the Language Games played by ChatGPT

Alessandro Trevisan, Harry Giddens, Sarah Dillon, Alan F. Blackwell
University of Cambridge

Manuscript prepared for submission to Critical AI
https://www.dukeupress.edu/critical-ai

Abstract

Generative large language models (LLMs), which create text without direct correspondence to truth value, are widely understood to resemble the uses of language described in Frankfurt's popular monograph *On Bullshit*. In this paper, we offer a rigorous investigation of this topic, identifying how the phenomenon has arisen, and how it might be analysed. In this paper, we elaborate on this argument to propose that LLM-based chatbots play the 'language game of bullshit'. We use statistical text analysis to investigate the features of this Wittgensteinian language game, based on a dataset constructed to contrast the language of 1,000 scientific publications with typical pseudo-scientific text generated by ChatGPT. We then explore whether the same language features can be detected in two well-known contexts of social dysfunction: George Orwell's critique of politics and language, and David Graeber's characterisation of bullshit jobs. Using simple hypothesis-testing methods, we demonstrate that a statistical model of the language of bullshit can reliably relate the Frankfurtian artificial bullshit of ChatGPT to the political and workplace functions of bullshit as observed in natural human language.

Independent Corroboration

Trevisan et al empirically show that

an LLM-based chatbot, such as ChatGPT,

usually produces Frankfurtian BS,

because the LLM is knowingly trained on data that

have not been vetted for truth.

Evolutionary Psychology
April-June 2021: 1–10

DOI: 10.1177/14747049211000317
journals.sagepub.com/home/evp
SAGE

Bullshit Ability as an Honest Signal of Intelligence

Martin Harry Turpin[1,*], Mane Kara-Yakoubian[2,*], Alexander C. Walker[1], Heather E. K. Walker[3], Jonathan A. Fugelsang[1], and Jennifer A. Stolz[1]

Abstract

Navigating social systems efficiently is critical to our species. Humans appear endowed with a cognitive system that has formed to meet the unique challenges that emerge for highly social species. Bullshitting, communication characterised by an intent to be convincing or impressive without concern for truth, is ubiquitous within human societies. Across two studies ($N = 1{,}017$), we assess participants' ability to produce satisfying and seemingly accurate bullshit as an honest signal of their intelligence. We find that bullshit ability is associated with an individual's intelligence and individuals capable of producing more satisfying bullshit are judged by second-hand observers to be more intelligent. We interpret these results as adding evidence for intelligence being geared towards the navigation of social systems. The ability to produce satisfying bullshit may serve to assist individuals in negotiating their social world, both as an energetically efficient strategy for impressing others and as an honest signal of intelligence.

Keywords

intelligence, social navigation, bullshit, social signaling, individual differences

Turpin and Kara-Yakoubian et al

Turpin and Kara-Yakoubian et al

explain how,
hypothesize that, and
prove empirically that,

a human's ability to BS convincingly

is taken instinctively by other humans

as an honest signal of

the [first] human's intelligence.

BSing Well Requires Intelligence

Humans evidently understand instinctively that

to be able to tell lies or nontruths

that appear to be true

requires intelligence.

ChatGPT BSs

Because LLMs generate soft BS,

an LLM's output appears to a human

the same

as human-generated BS appears to a human.

Thereofore,

an LLM’s output

is taken instinctively by humans

as an honest signal of

the LLM’s intelligence,

even though the LLM has no such intelligence.

Particularly, …

those that don’t understand the reality of an LLM

feel in their guts that

the LLM is truly intelligent.

Compounding the‘ Effect

This effect might be compounded by my observation that

many non-techies understand “AI” as

“an artificial being that is truly intelligent”

rather than its intended meaning as

“faked intelligence”.

Why LLMs Appear Creative

From all this, it becomes clear also why LLMs are perceived as creative.

Many Definitions of Creativity

Many definitions of creativity and creative ideas.

Well studied in the literature.

My Favorite Definition

The one I find most operative is that

a creative, innovative idea is

an exception, from the norm,

that is perceived, in retrospect, to be

a good idea after all.

An Exception

An exception can range from

an inadvertent failure to follow a procedure or rules (mistake, Eureka)

to

an intentional deviation from the current conventions or styles (thinking out of the box, brainstorming).

Example

A composer decides to deviate from prevailing style of music,

tries sequences of notes until E finds one that sounds good to er ears.

Example, Cont'd

If enough concert goers agree with em,

the composition is considered a creative innovation.

Eventually, this innovative composition becomes part of the norm.

An LLM is a BSer

A fraction of what it produces are exceptions.

Some exceptions are seen by humans,

upon examination,

to be good ideas after all.

Ergo, the LLM is perceived as creative.

Corroboration

William J. Broad in NYT, reports that scientists

have come to a similar conclusion,

and are parlaying LLMs

to speed up breakthroughs

and maybe “even win the Nobel Prize”

Conclusion

Why do LLMs *appear* to be

intelligent and creative?

In a nutshell, it’s

it’s because they generate BS!

because they are BSers!

Future Work Needed

To validate this hypothesis,

it will be necessary to

conduct an experiment like that of Turpin and Kara-Yakoubian et al

in which humans will estimate the intelligence of the authors of

a random mixture of

output from ChatGPT and
output from humans.