Output first: Rethinking bullshit in large language models

Jeremy Wyatt

Plan for the talk

- There is an emerging and fast-moving debate about whether large language models (LLMs) are 'bullshitters,' 'bullshit machines,' or 'bullshit generators'
- For the most part: participants in these debates have employed notions of bullshit that are grounded in Harry Frankfurt's seminal work
- My proposal: instead of using a broadly Frankfurtian account of bullshit to evaluate LLMs, we should use an account that is squarely focused on the quality of LLMs' *outputs*

Plan for the talk

- My main aims during the talk:
  - To review the influential Frankfurtian account of LLM bullshit due to Hicks et al. (2024) and one of the main criticisms of this account
    - The criticism: Hicks et al.'s account should be replaced by an *output-based* account of LLM bullshit (Gunkel and Coghlan 2025)
  - To build upon this criticism by developing a detailed output-based account of LLM bullshit that is based on Florian Cova's recent work

Hicks et al. on LLMs and bullshit

- Following the release of ChatGPT in November 2022, many authors argued (or simply claimed) that ChatGPT and other LLMs are bullshitters
  - For instance: Agüera y Arcas (2022), Bergstrom and Ogbunu (2023), McQuillan (2023), Marcus (2022), Mollick (2022), Narayanan and Kapoor (2022), and Sundar and Liao (2023)

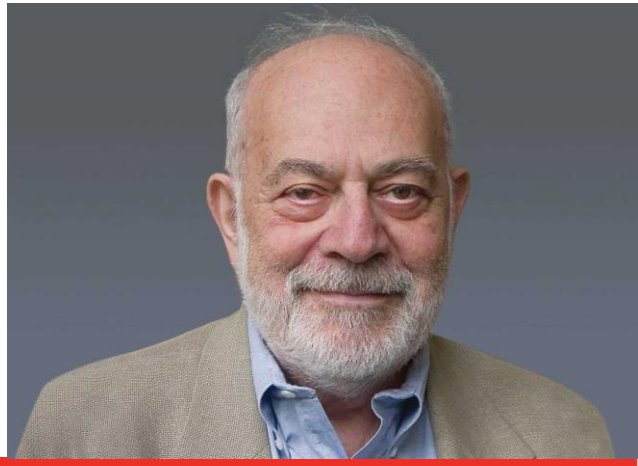
Hicks et al. on LLMs and bullshit

- However: the spark for the recent debates about LLMs and bullshit was Hicks et al.'s 'ChatGPT is bullshit' (2024)
- Hicks et al. target the common claim that when LLMs confidently produce inaccurate or nonsensical outputs, they are *hallucinating*

> At OpenAI, we're working hard to make AI systems more useful and reliable. Even as language models become more capable, one challenge remains stubbornly hard to fully solve: hallucinations. By this we mean instances where a model confidently generates an answer that isn't true. Our new research paper argues that language models hallucinate because standard training and evaluation procedures reward guessing over acknowledging uncertainty. (Kalai et al., 'Why language models hallucinate,' 2025)

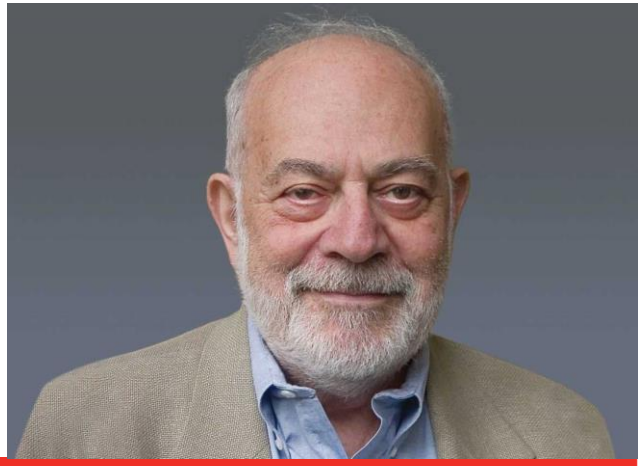
Hicks et al. on LLMs and bullshit

- Hicks et al. argue that instead of thinking of LLMs as hallucinators (or confabulators or liars), we should think of them as *bullshitters*
- In doing so: they draw on Harry Frankfurt's highly influential account of bullshit



Hicks et al. on LLMs and bullshit

- In essence, Frankfurt argues that a person *X* bullshits iff:
  - *i.* *X* asserts a proposition *p* while being indifferent to whether *p* is true or false and
  - *ii.* *X* asserts *p* with the intention to deceive *X*'s audience about *X*'s indifference towards *p*'s truth
- Frankfurt's view is standardly called a *process-based* (or 'activity-centered') account of bullshit
  - Process-based accounts aim primarily to identify the features in virtue of having which processes (e.g. the performance of a speech act) count as bullshit (or better: bullshitting)



Hicks et al. on LLMs and bullshit

- Hicks et al. suggest that we can use the resources of Frankfurt's account to distinguish between the *genus* of bullshit and two *species* thereof:
  - *Bullshit (genus):* Any utterance produced where a speaker has indifference towards the truth of the utterance
  - *Hard bullshit:* Bullshit produced with the intention to mislead the audience about the utterer's agenda
  - *Soft bullshit:* Bullshit produced without the intention to mislead the hearer regarding the utterer's agenda

Hicks et al. on LLMs and bullshit

- Notably: when Hicks et al. use 'utterance,' they presumably mean to be talking about *assertions* (or perhaps what Searle (1979, pp. 12-13) calls 'assertives')
  - For instance: when I ask a question, I am indifferent towards its truth-value, since I don't take it to be truth-apt
  - However: Hicks et al. presumably wouldn't hold that it is for this reason a bullshit question
- Hicks et al.'s position: LLMs are soft bullshitters and may be hard bullshitters, depending on whether we should regard them as intending to mislead their audience about their agenda

Gunkel and Coghlan's objection

- Hicks et al.'s views have attracted a great deal of criticism, and I'll focus on an objection due to Gunkel and Coghlan (2025)
- The objection:
  - In place of Hicks et al.'s process-based account of LLM bullshit, we should develop an *output-based* account
    - Output-based accounts aim primarily to identify the features in virtue of having which outputs of some causal process or another count as bullshit
    - Put differently: in an output-based account, "the shit wears the trousers" (Cohen 2002, p. 324)
- Gunkel and Coghlan offer three reasons for favouring an output-based account

Gunkel and Coghlan's objection

- *Reason #1:* Unlike Hicks et al.'s account, an output-based account can secure the result that some LLM outputs are bullshit
  - Gunkel and Coghlan agree with Hicks et al. that LLMs produce bullshit at least some of the time
  - However: they point out that LLMs which are fine-tuned for truth/correctness appear to track truth
  - For this reason: Hicks et al.'s contention that such LLMs are wholly indifferent towards the truth of their assertions is misleading
  - This undermines the ability of Hicks et al.'s account to explain why we should categorise (some) of these LLMs' outputs as bullshit
  - By contrast: an output-based account can straightforwardly explain this
  - The basic explanation: even LLMs that are fine-tuned for truth/correctness sometimes produce outputs with flaws like "falsity, exaggeration, illogicality, damning shallowness, and pseudo-profundity" (Gunkel and Coghlan 2025, p. 4)

Gunkel and Coghlan's objection

- *Reason #2:* Unlike Hicks et al.'s account, an output-based account doesn't problematically *anthropomorphise* LLMs
  - Hicks et al. anthropomorphise LLMs by analogising them to human (soft) bullshitters, but this analogy is questionable
  - While the human (soft) bullshitter is unconcerned with whether certain of their assertions are true/false, they are still *capable* of having such a concern
  - By contrast: it is highly contentious whether LLMs are capable of having such a concern
  - An output-based account needn't anthropomorphise LLMs in this way, as it is primarily focused on certain flaws in their outputs, not their mental states

Gunkel and Coghlan's objection

- *Reason #3:* Unlike Hicks et al.'s account, an output-based account is attractively *selective*
  - An account of LLM bullshit should be selective in the sense that it entails that while LLMs sometimes produce bullshit, they don't always do so

Gunkel and Coghlan's objection

- To see the point here, contrast two cases:
  - *Case #1:* I ask ChatGPT for a list of five articles on deflationary theories of truth that have been published in the last ten years. ChatGPT promptly produces what looks like a list of five articles of the requested sort. However, when I use Google to determine whether these articles exist, it turns out that none of them do.
    - It would be natural for me to respond 'Enough with the bullshit ChatGPT. Stop making things up and consult a reliable source like PhilPapers'
  - *Case #2:* I upload a draft of one of my articles to Claude and ask it to evaluate the paper as if it were a hard-headed referee at *Mind*. The report that Claude produces identifies some significant flaws in my arguments that I hadn't noticed. As a result, I'm able to improve my article.
    - In this case, we wouldn't typically describe Claude's report as bullshit
    - Instead: we would say that it offers a helpful evaluation of my article

Gunkel and Coghlan's objection

- Hicks et al.'s account is not selective, as it entails that LLMs always produce soft (and potentially hard) bullshit
- As a result: Hicks et al. are committed to saying that both ChatGPT and Claude produce bullshit in the above cases
- By contrast: an output-based account is selective, given that some—but not all—LLM outputs are e.g. false, exaggerated, illogical, damningly shallow, or pseudo-profound
- In particular: adopting an output-based account enables us to explain why ChatGPT's output above is bullshit while Claude's isn't
  - An initial explanation: ChatGPT's output is both inaccurate and completely fabricated, whereas Claude's output is accurate and insightful

Gunkel and Coghlan's objection

- Admittedly: Gunkel and Coghlan don't offer *decisive* reasons to favour an output-based account of LLM bullshit over a process-based account
- For instance: it may be possible to rebut Reasons #1 and #3 by developing a process-based account that is more sophisticated than that of Hicks et al.
- Moreover: as both LLMs and our understanding of them become more advanced, it may ultimately be desirable to anthropomorphise LLMs to a significant degree
- However: at the very least, I take Gunkel and Coghlan to provide strong reasons to seriously consider an output-based account of LLM bullshit

Three criticisms of Hicks et al.

- That said, there is an important lacuna Gunkel and Coghlan's criticism: they don't develop a detailed output-based account of LLM bullshit
  - In what follows, I will aim to develop such an account

Cova's output-based account

- Florian Cova (2024) has recently developed a nuanced, output-based account of bullshit
- In motivating the account: he draws on cases of *truth-tracking bullshit*
    - *Pompous French restaurant:* Shane is a New Zealander visiting Nice and he doesn't speak much French. He scans the menu at a fancy French restaurant and points randomly to the name of a dish, asking the waiter what the dish is. The waiter (kindly speaking in English) says, 'That dish is the Prince of the Seas in his Sauce of the Fruits of Provence.' Shane thinks that the dish sounds delicious, so he orders it. Later, Shane is very disappointed when the waiter brings him sardines in oil.

Cova's output-based account

- It is natural to regard the waiter's assertion in this case as bullshit
- Nevertheless: Cova points out that the assertion is truth-tracking insofar as "there is some connection between the name of the dish and its content, to the extent that trained consumers can guess what the dish contains. Rather than inventing names at random, the person who wrote the menu intended for the names to track the content of the dishes[.]" (2024, p. 578)

Cova's output-based account

- Cova takes cases like this to reveal an important feature of bullshit: "bullshit is not necessarily meaningless, false, or lacking in evidence or respect for truth, but…it can always be *deflated*…[B]ullshit is something that seems to have much value at first sight, but scratch the surface and you will see that it is *empty*. To put it simply: bullshit is just *hot air*." (2024, p. 586)
- Cova's proposed account of bullshit:
  - (BS_C) An assertion *N* is bullshit iff (i) *N* is presented as or appears at first sight as making an interesting contribution to a certain inquiry *I*, but (ii) *N* would turn out, on closer inspection by a minimally competent inquirer, to make a much less interesting contribution

Cova's output-based account

- To unpack (BS_C), we should briefly consider Cova's accounts of:
  a. What an *interesting contribution* to an inquiry amounts to;
  b. What makes someone a *minimally competent inquirer*; and
  c. What it is to *closely inspect* an assertion

Cova's output-based account

- (Interesting) An assertion *N* is an interesting contribution to inquiry *I* iff (i) *N* answers a question that matters to the participants in *I* and (ii) *N* answers this question in a way that is either surprising or illuminating (i.e. *N* must either bring new, unexpected, or game-changing information, or help the participants in *I* better integrate the information they already had in a coherent explanatory framework)
- (MCI) Person *P* is a minimally competent inquirer relative to an assertion *N* iff (i) *P* has a basic comprehension of the terms that are used in *N* and (where applicable) the theories in which those terms play a role; (ii) *P* is able to identify obviously fallacious arguments; and (iii) *P* has a sufficient amount of *basic knowledge*, i.e. *P* knows what most people in their epistemic community know
- (CI) Person *P* closely inspects assertion *N* iff (i) *P* aims to determine what *N* means and (ii) *P* seeks easily accessible information to determine whether *N* is both non-trivial and true, or at least based on reasonable evidence

Cova's output-based account

- These definitions make it easy enough to see why (BS_C) secures the result that the French waiter's assertion is bullshit
  - It initially seems to Shane that the assertion makes an interesting contribution to his inquiry
  - Specifically: the assertion appears to answer a question that matters to Shane—'What is in the dish whose name I just pointed to?'
  - Moreover: given the pompous name of the dish, it seems to Shane that the assertion answers this question in a surprising and/or illuminating way

  - However: a minimally competent inquirer would be able to determine on closer inspection that the assertion *doesn't* answer Shane's question in a surprising or illuminating way
  - In particular: such an inquirer would come to recognise that the dish's unnecessarily pretentious name *misleads* Shane into thinking that the dish to whose name he pointed is complex, unusual, and appetising when in fact it is just sardines in oil

Cova's output-based account

- Cova recognises that in addition to an account of bullshit, we need an account of *bullshitting*
- Interestingly: his proposed account of bullshitting is process-based, rather than output-based:
  - (BT_C) *X* is bullshitting when (i) *X* engages in a communicative act *C* but is more concerned with the general (affective) impression their act will have on a given target *T* than about the particular propositions they will get *T* to endorse and (ii) *X* is aware that *X* has no good reason to think that the impression *X* is trying to convey is warranted by reasons presented in or hinted at by *C*

Cova's output-based account

- Cova takes this account to be attractive for three main reasons:
  a. It allows for the possibility that *X* bullshits without producing bullshit;
  b. It allows for the possibility that *X* produces bullshit without bullshitting; and
  c. It is nevertheless based upon the core idea animating (BS_C), that of "something that is more impressive at first sight than it really is or purported to be" (2024, p. 594)

Streamlining Cova's account

- I take Cova's account of bullshit to be sophisticated and plausible
- However, I want to offer a *streamlined* output-based account that preserves the spirit of Cova's account using much less machinery:
  - (BS) An assertion *A* that is offered as a contribution to inquiry *I* is bullshit iff *A* doesn't actually further any of the aims of *I* (cf. Sundell 2026)

Streamlining Cova's account

- Like (BS_C), (BS) is a way of spelling out the basic idea that while bullshit outputs initially appear to be valuable, they are in fact just 'hot air'
  - *Note:* this idea is in line with sense 1 for the noun 'bullshit' in the *Oxford English Dictionary*, according to which it means rubbish or nonsense
- According to (BS): a bullshit assertion initially appears to be valuable insofar as it is offered as a contribution to an inquiry *I*
  - In fact, though, the assertion doesn't further any of *I*'s aims, so in connection with *I*, the assertion it is worthless/empty/just hot air

Streamlining Cova's account

- A notable strength of (BS): it unifies the *bullshit-making features*
  - Feature *F* is a bullshit-making feature iff an assertion *A*'s having *F* makes it more probable that *A* is bullshit

Streamlining Cova's account

- An initial list of the bullshit-making features, where *A* is offered as a contribution to some inquiry *I*:
  - Being meaningless
  - Being meaningful but trivial
  - Being meaningful and non-trivial but insufficiently grounded in available evidence
  - Being meaningful, non-trivial, and sufficiently grounded in available evidence but irrelevant to *I*'s aims
  - Being vague
  - Being ambiguous
  - Being misleading
  - Being inaccurate
  - Being completely fabricated
- If *A* exhibits any of these features, then that makes it more probable (and perhaps certain) that while *A* is offered as a contribution to *I*, it actually fails to further any of *I*'s aims

Streamlining Cova's account

- Additionally: like (BS_C), (BS) secures the result that in the pompous French restaurant case, the waiter's assertion is bullshit
  - This assertion is bullshit because it doesn't further any of the aims of Shane's inquiry
  - The main aim of this inquiry is to determine what dishes (if any) Shane should order from the menu
  - The subsidiary aims of the inquiry are e.g. to determine whether the menu lists any dishes that are appetising to Shane and to determine whether Shane can afford any of the dishes on the menu
  - The waiter's assertion doesn't help Shane to pursue any of these aims
  - Instead: it simply misleads Shane into thinking that the dish to whose name he pointed is complex, unusual, and appetising when it is just sardines in oil

Bullshit LLM outputs

- We can now return to LLMs, focusing on three details:
  i. (BS) can be straightforwardly applied to LLMs
  ii. When thinking about LLM *bullshitting*, we have strong reasons to adopt an output-based account, rather than a process-based account like Cova's
  iii. Since LLMs are often branded as 'bullshitters,' 'bullshit machines,' or 'bullshit generators,' it is worthwhile to reconsider whether familiar LLMs are *bullshitters* in any significant sense

Bullshit LLM outputs

- When applied to LLMs, (BS) is attractively selective—it secures the result that while some LLM outputs are bullshit, not all are
- For instance: return to the ChatGPT and Claude cases
- ChatGPT produces what looks like a list of five articles on deflationary theories of truth published in the last ten years
  - The aim of my inquiry is to identify these articles, and it initially seems that ChatGPT's response furthers this aim
  - However: it becomes clear upon inspection that ChatGPT's response doesn't further this aim insofar as none of the putative articles in its list actually exist
  - As a result: ChatGPT's response is bullshit

Bullshit LLM outputs

- By contrast: Claude's report on my draft article identifies significant flaws in my arguments that I had missed
  - The aim of my inquiry is to identify ways in which the draft can be improved, and Claude's report furthers this aim
  - As a result: Claude's response to my prompt is not bullshit

Bullshit LLM outputs

- *Note:* In applying (BS) to LLMs, I don't mean to deny that there are interesting questions about the processes that may cause an LLM to produce bullshit
  - For instance: if an LLM makes a bullshit assertion, we can try to determine whether this is due to gaps in its training data, features of its architecture, flaws in the user's prompt, or some other factor(s)
- The advantage of applying (BS) to LLMs is that we are able to draw a sharp distinction between process-related questions such as these and the output-related questions upon which we have been focusing

LLM bullshitting

- Turn now to LLM *bullshitting*
- Cova's process-based account of bullshitting runs as follows:
  - (BT_C) X is bullshitting when (i) X engages in a communicative act C but is more concerned with the general (affective) impression their act will have on a given target T than about the particular propositions they will get T to endorse and (ii) X is aware that X has no good reason to think that the impression X is trying to convey is warranted by reasons presented in or hinted at by C

LLM bullshitting

- (BT_C) may be a promising account of human bullshitting
- However, if applied to LLMs, it would have several highly contentious implications:
  - That LLMs can engage in communicative acts;
  - That LLMs can be concerned about things; and
  - That LLMs can be aware of things
- As a result: it is worthwhile to consider whether there is an alternative account of LLM bullshitting that pairs well with (BS) and doesn't have these implications

LLM bullshitting

- Let a *bullshit-conducive process* be a process that more often than not produces bullshit
  - *Examples:* speculating about issues of which one has little or no knowledge, drafting a mandatory yet insincere diversity statement for a job application, writing a script for a car commercial
- An alternative, output-based account of bullshitting:
  - (BT) Entity *y* is bullshitting at time *t* iff *y* is executing a bullshit-conducive process at *t*

LLM bullshitting

- (BT) pairs well with (BS) as both are output-based and (BT) incorporates the notion of bullshit that is delivered by (BS)
- Additionally: (BT) is compatible with the other two motivations that Cova offers for (BT_C): that it is possible for *y* to bullshit without producing bullshit and that it is possible for *y* to produce bullshit without bullshitting

LLM bullshitting

- (BT) allows for bullshitting without bullshit because a bullshit-conducive process is a process that produces bullshit *more often than not*
  - Given this fact: if *y* executes a bullshit-conducive process, then it is likely but not guaranteed that *y* will produce bullshit
- Moreover: (BT) allows for bullshit without bullshitting because (BT) doesn't entail that the *only* way to produce bullshit is to execute a bullshit-conducive process
  - It may be unlikely that a process that isn't bullshit-conducive will produce bullshit, but given (BT) and (BS), this is possible

Are LLMs bullshitters?

- Lastly: given (BS) and (BT), we should consider whether there is any significant sense in which LLMs are *bullshitters*
- Michael Wreen observes that there are weaker and stronger senses in which someone can be a bullshitter:

> In the weakest sense, a humbug or a bullshitter is someone who produces humbug or bullshit on a given occasion, just as a smoker in the weakest sense is someone who smokes on a given occasion. More strongly, a humbug or a bullshitter is someone who has a disposition to produce humbug or bullshit, just as a smoker in a stronger sense is someone who has a disposition to smoke. Paralleling this distinction is a stronger one: a humbug or a bullshitter is someone who (qualifications aside) deliberately produces humbug or bullshit on an occasion; a humbug or a bullshitter is someone who has a disposition to do so. (Wreen 2013, pp. 113-114)

Are LLMs bullshitters?

- For present purposes: we can set aside the issue of whether LLMs *deliberately* produce their outputs, which leaves Wreen's first and second senses of being a bullshitter
- It is worthwhile to introduce a bit of nuance here
  - This is because: as Wreen indicates, *context* plays a key role when we are determining whether someone or something is a bullshitter in either of these senses
- A revised articulation of these two senses of being a bullshitter:
  - (BR_I) An entity *y* is an *instance bullshitter* relative to context *c* iff *y* produces some bullshit in *c*
  - (BR_D) An entity *y* is a *dispositional bullshitter* relative to context *c* iff *y* is disposed to produce some bullshit in *c*

Are LLMs bullshitters?

- LLMs are clearly instance bullshitters relative to some contexts, given that they sometimes produce bullshit (consider the ChatGPT case)
- However: they aren't instance bullshitters relative to all contexts, since they sometimes produce outputs that aren't bullshit (consider the Claude case)

Are LLMs bullshitters?

- LLMs are also dispositional bullshitters relative to certain contexts, as they are disposed to produce bullshit in those contexts
  - For instance: suppose that an LLM has been trained to respond directly to all user queries rather than abstaining when it lacks access to information about the topic raised in a query
  - Suppose further that Samuel asks the LLM in what year Sally Hemings was born, and the LLM lacks access to any information about Sally Hemings, as it has only been trained on a corpus of articles about medicine published between 2000 and 2025
  - Relative to this context: the LLM is disposed to produce bullshit

Are LLMs bullshitters?

- However: LLMs aren't dispositional bullshitters relative to all contexts, as there are some contexts in which they aren't disposed to produce bullshit
  - Suppose that an LLM has been trained to only respond directly to a user's query when it has access to information about the topic raised in the query
  - Suppose further that Samuel asks the LLM in what year Sally Hemings was born, and the LLM has access to a great deal of information about Sally Hemings, as it has been trained on an extensive corpus about early American history
  - Relative to this context: the LLM is disposed to respond accurately to Samuel's query and is thus not disposed to produce bullshit

Summary of the talk

- Picking up on Gunkel and Coghlan's criticism of Hicks et al. and Cova's recent work, I have offered what I take to be a promising output-based account of LLM bullshit
- I have also supplemented this account with an output-based account of LLM bullshitting and an explanation of why LLMs are bullshitters relative to some contexts but not others
- The key idea underpinning my proposal: even if process-based accounts of LLM bullshit are untenable, we can still think of LLMs as producing bullshit, bullshitting, and being bullshitters—as long as we put their outputs first

Kia ora/thanks!