Scrutinizing the foundations: could large Language Models be solipsistic?

Author: Andreea Esanu
Affiliation: New Europe College – Institute for Advanced Study, Bucharest, Romania
E-mail: aesanu@nec.ro

In artificial intelligence (AI) literature, "delusions" are generally characterized as the generation of false, nonsensical or unfaithful output from reliable source content (Ji *et al*., 2023, p.3). The occurrence of delusions questions the ability of generative AI models to work in real-world contexts and have real-world applications, raising a serious challenge to the development of the field in the future. For example, in medical applications, reports generated from patient data using generative AI models may contain unfaithful information that could put the patients' lives at risk (*ibidem*). This means that the delusion problem needs somehow to be addressed. There is an extensive literature on computer-generated delusions, which ranges from tackling hallucinations, like the production of nonsensical images, in Computer Vision (Baker & Kanade, 2006) to dismantling nonsensical or factually false text generated by (natural) Language Models, but this literature is predominantly taxonomic, focusing especially on ways to classify various delusions: visual- *vs* text-based delusions, intrinsic (the generated output contradicts the source content) *vs* extrinsic (the generated output is not verifiable from the source content) delusions (Ji *et al*., 2023, p.4) etc.

In a recent research paper titled "Shaking the Foundations: delusions in sequence models for interaction and control" (Ortega *et al.*, 2021), a group of scientists from DeepMind successfully presented a well-defined formal treatment of the delusion problem for an entire class of generative AI models (Ortega & Braun, 2009; Rezende *et al.*, 2020) focused on modeling purposeful adaptive behavior. While Ortega's *et al.* (2021) result does not comprehensively explain all types of computer-generated delusions or their sources, nor does it aim to do so, it bears significance across a wide range of generative AI models gathered under the umbrella of sequence models or transformers (*e.g.*, foundation models, language models, reinforcement-learning models etc.), in which purposeful adaptive behavior is expected to occur. As the authors show, in sequence models or transformers, delusions are in fact auto-suggestive, *i.e.*, they are self-induced and self-propagated by the models themselves due to confounding in the underlying stochastic model. Confounding, in this context, means that the models fail to distinguish their own actions (like generating an output) from observations by which they represent the source content.

In the case of large Language Models and natural language processing based on sequence models, the literature documenting and classifying computer-generated delusions is growing at an extremely fast pace (Ji *et al.*, 2023; Li *et al.*, 2023; Zheng *et al.*, 2023) but little has been said about auto-suggestive delusions (Ortega *et al.*, 2021). Usually, auto-suggestive delusions are equated

with exposure bias (He *et al.*, 2021), but exposure bias does not say much about the nature of the delusions it produces. To this purpose we may employ a different notion: *i.e.*, the delusion is triggered by the fact that the language model treats both the observations (the ground-truth data) and its own actions (the model's own samples generated from the ground-truth data) as language tokens. The consequence is a form of confounding (Pearl, 2009), in the sense that the model ends up taking its own actions *as* observations – as evidence about the world, thus generating self- or auto-suggestive delusions.

Many problem formulations in the evolving field of generative AI still lack a straightforward formalization, so this neat formal result, *i.e.*, the presence of stochastic confounding in sequence models, is notably valuable for a conceptual analysis of delusions in the field of machine learning. Sequence models or transformers trained using exclusively self-supervised learning (Amatriain *et al.*, 2023) are most prone to auto-suggestive delusions, because the manner in which these models sample data from the training set is entirely up to them. In the family of large Language Models, relevant examples in this category are older transformer models based on BERT or GPT-3. Nevertheless, even more recent hybrid transformer models, like those based on GPT-3.5 (such as ChatGPT), are exposed to this kind of delusion, although they are trained using self-supervised learning followed by a human-in-the-loop fine tuning or reinforcement learning with human feedback – RLHF fine tuning (*ibidem*), which constrains the models into sampling relevant data from the training set and minimizes exposure bias. The primary reason for the occurrence of auto-suggestive delusions is that neither transformer models with human-in-the-loop fine tuning (or reinforcement learning with human feedback – RLHF) explicitly address the underlying problem of *confounding*, although the exposure bias is acknowledged and reduced through approaches like RLFH.

In short, whenever a confounder of the type discussed here (Ortega *et al.*, 2021) is identified in a sequence model or transformer, such as a large Language Model of the GPT-3.5 sort, it presents an opportunity to ask whether the resulting self- or auto-suggestive delusions of the model could be likened, not so much to "stochastic parroting" (Bender *et al.*, 2021), but rather to what in the philosophy of language and mind may be called *a private language*, yet in a weak sense – that is, a language that exhibits a *systematic* deviation from what we generally call the human language. The main rationale for asking this question is that, even though the model could learn a stochastic representation of the world (the human language included), it could still take its own actions, *e.g.*, the generated text, as evidence (as ground-truth data) about the world (the human language included), creating *systematically* wrong representations, and so "altering" both the world and the meaning of words.

This further raises the question as to whether the presence of self- or auto-suggestive delusions could indicate that, at least in theory, large Language Models are likely to be also *solipsistic*, that

is, likely to become further and further disconnected from the world, instead of just and simply biased.

The plan of the paper is as follows. I will begin by providing a brief overview of exposure bias in sequence models, including language models. Following that, I will introduce the formal framework that elucidates the probabilistic delusions capable of explaining exposure bias in a broad manner. This will provide the basis for discussing self- or auto-suggestive delusions in sequence models. Moving on to the third section, I will analyze self- or auto-suggestive delusions by proposing an analogy with the rule-following problematic originating in the philosophy of mind and language. Lastly, in the fourth section, I will argue that this comprehensive approach leads to the suggestion that sequence models, large Language Models in particular, may develop in a manner that touches upon solipsism, understood here as a gradual tendency to go further and further astray in representing the world, and the emergence of a private language in a weak sense, that is a made-up language, progressively detached from human language.

1. Stochastic parrots and exposure bias

There is a rapidly expanding body of literature focused on the diverse issues arising from large Language Models. Among them, an apparently inherent tendency towards randomness, which was observed in the behavior of these models, has earned them the moniker of "stochastic parrots" (Bender *et al.*, 2021). It is posited that, despite their proficiency in multiple languages, these models lack a genuine comprehension of linguistic meaning, relying instead on a statistical matching of linguistic tokens (*i.e.*, words) to generate various forms of text. Yet these forms of text are often hallucinatory, suggesting that Language Models are nothing but random machines.

This characterization, however, offers a broad and somewhat unsatisfactory representation of computer-generated delusions within large Language Models. Alternative, more systematic approaches seek to furnish precise definitions of delusion and to establish classifications grounded in their distinct sources and manifestations (Ji *et al.*, 2023; Li *et al.*, 2023). Thus, some delusions result from the heuristics of data collection and mismatches in the data (Dhingra *et al.*, 2019). Others, derive from the model itself, *i.e.*, from aspects related to training and inference in the models. Some models learn imperfect representations of context, in the sense that they "learn wrong correlations between different parts of the training data" (Ji *et al.*, 2023, p. 8; Aralikatte *et al.*, 2021).

Other models draw incorrect inferences due to exposure bias. Exposure bias refers to the discrepancy between how a model is trained and how it is used at inference time. During the training of a language model, the model is typically exposed to ground-truth data, meaning it is provided with the correct or target sequence of language tokens (Brown *et al.*, 2020), at each step

of the sequence generation. This allows the model to learn and adjust its parameters based on ground-truth information. These models are trained using a maximum likelihood estimation (MLE) objective. This objective encourages the model to generate text that maximizes the likelihood of the correct next token given the previous context: $P\ (T_{n+1} \mid T_1, \ldots, T_n)$. In short, the model learns to predict the most likely token T_i at each step in a sequence. In this phase, the model has access to perfect information.

However, during the inference or generation phase, the model is not provided with ground-truth information. Instead, it starts by *generating* a sequence of tokens based on its own predictions from training, one token at a time (see Fig.1), and uses its own generated tokens as input for generating subsequent tokens. The exposure bias, then, arises from this difference between training and inference. Since the model is not exposed to ground-truth information during inference, errors can accumulate, as it generates a sequence, and these errors can compound over time (Ji *et al.*, 2023) producing delusions. "The web is full of actions (text) produced by many other agents, mostly people, but recently by machines too, such as GPT-3. Language models (...) are often pre-trained with self-supervising learning techniques. These pre-trained models are agents that can generate actions by conditioning on previous actions." (Ortega *et al.*, 2021, p. 9). So, once a model's own actions enter the mix, chances are that the model will delude itself.

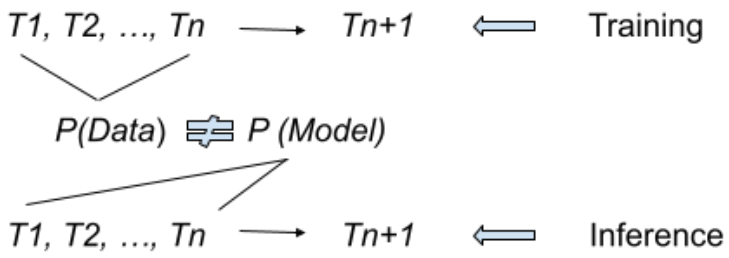

Figure 1. A graphical representation of exposure bias. During generation, the model is fed ground-truth data tokens, from *P(Data*). During inference, the model instead uses tokens from the model's own samples *P(Model).* Source: (He *et al.*, 2021: 1)

This kind of delusion, however, is far more interesting than a delusion generally understood as a "generated content that is nonsensical or unfaithful to the provided source content" (Ji *et al.*, 2023, p. 4; Maynez *et al.*, 2020), because of its peculiar nature. Even we, as humans, experience it and we often find it compelling; we experience it, for instance, when we take what we *do* (*e.g.*, what we hold to be the case) as what *is* (*e.g.*, what is factually the case).

In some sense, this is a solipsist delusion: it occurs whenever we equate our actions or beliefs with objective reality. Now, what is even more fascinating is that we can encounter the same kind of delusion in machines that are devoid of any subjectivity as we know it. Their delusions can be called *self-* or *auto-suggestive* delusions because such machines can behave quite like us, *i.e.*, they

are prone to taking what they do as something that they see. And as Pearl & Mackenzie (2018) show, this is essentially a problem of confounding[1]: $P(T_{n+1} \mid T_{1:n}) \neq P(T_{n+1} \mid do\ (T_{1:n}))$.

2. Probabilistic delusions, a formal approach

Probabilistic models are, in fact, prone to producing delusions. This is something that anyone who has ever delved into causal inference has heard about. In this section, we will tackle two standard types of probabilistic delusions: colliders (which are interesting for introducing the topic of how our mind plays tricks on us) and confounders (which are specifically relevant to our discussion).

Even though *pure* probabilistic associations seem incomprehensible to us[2], they are, in fact, quite common around us. Think, for example, a simple experiment like the next one (Pearl & Mackenzie, 2018, p. 185). We conduct a series of one hundred simultaneous coin tosses using two coins, Coin 1 and Coin 2. Using a table, we only record the outcome of a toss when at least one of them displays Heads. Eventually, this will end up counting roughly 75 entries in the table. Upon examining the table, we then notice that the tosses of the two coins are not entirely *unrelated*; but in every instance in which Coin 1 landed on Tails, Coin 2 landed on Heads. In other words, whenever Coin 1 displays Tails, it is certain that Coin 2 displays Heads. Now, this is a purely probabilistic association without an underlying common cause; and even if we have the illusion that the coin tosses are not causally independent, in fact they are. "The correlation that we observe is, in the purest and most literal sense, an illusion. Or perhaps, even delusion: that is, *an illusion we brought upon ourselves by choosing which events to include in our data-set and which to ignore*." (Pearl & Mackenzie, 2018, p. 185).

For such a delusion, a perspicuous formalism exists, called causal diagrams, which exposes the nature of the problem quite neatly. The causal diagram formalizing the delusion in our experiment is called a *collider* (Pearl, 2009). A collider is a variable that is influenced by two other variables, thus leaving the misleading impression that the two other variables are somehow connected (see Fig.2).

[1] "Confounding (...) should be defined as anything that leads to a discrepancy between $P(Y \mid X)$ and$P(Y \mid do(X))$." (Pearl & Mackenzie, 2018, p.143)

[2] As Pearl & Mackenzie (2018) note, "...we find it utterly incomprehensible that there is a probabilistic association. Our brains are not prepared to accept causeless correlations and we need training." (p.183).

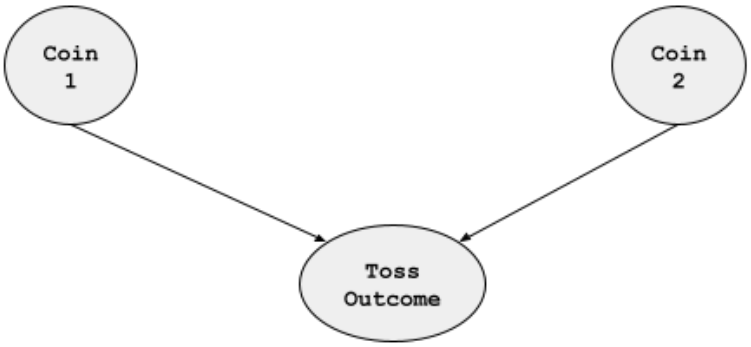

Figure 2. Causal diagram representing a collider. The two arrows starting in Coin 1 and Coin 2, which are coin tosses, pointing into Toss Outcome, which is the joint outcome of the two tosses, represent the fact that whenever Coin 1 displays Tails, Coin 2 displays Heads.

Another type of causal diagrams formalizing delusions in probabilistic models are called *confounders* (Pearl, 2009). Confounders are, in the formal sense, the opposite of colliders (see Fig.3).

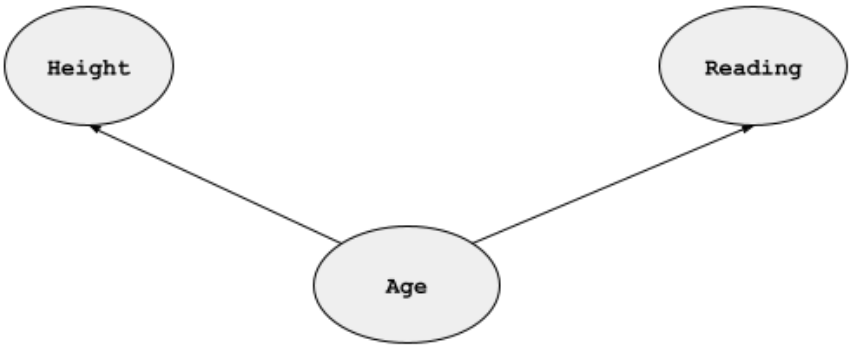

Figure 3. Causal diagram representing a confounder, usually called a "fork" (Pearl & Mackenzie, 2018, p. 112). Age, from which the arrows start, is called a common cause or a confounder of both children's Height and Reading proficiency, such that Height and Reading appear to be statistically correlated.

A confounder is, usually, an underlying variable that influences two other variables, such that it induces a spurious correlation between them. Reading proficiency in children, for example, is strongly positively correlated with children's height, although reading proficiency does not influence height or vice versa. Nevertheless, one can predict reading proficiency from height or vice versa, because of the underlying factor, age.

Confounders are, in fact, a source of important probabilistic delusions in sequence models. In order to see this more clearly, let us take a look at a familiar toy problem called the "prize or frog" (Ortega *et al.*, 2021), represented in Fig.4 below. We have two boxes; in one box there is a frog, in the other one there is a prize, say a candybar. The aim is to open the box containing a candybar,

thus avoiding the frog. Let us first define our three variables: Θ (theta) = the configuration of the box (candybar or frog); A = the action ("open box 1" or "open box 2") and O = the outcome (get a candybar or get a frog).

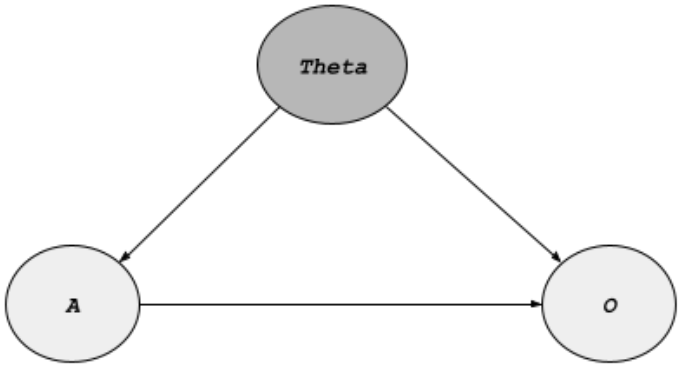

Figure 4. A causal diagram representing the "prize or frog" problem. A (the action) and O (the outcome of the action) are confounded by Θ (the configuration of the box), which is known by the expert. The arrow from A to O also represents the fact that the outcome is dependent on the action. Source: (Ortega *et al.*, 2021, p. 3).

Two problem scenarios can be imagined, one in which the model is only asked to predict the next successful action, and another one in which the model is asked to choose the next successful action, both based on recording the previous actions of an expert (who knows in which box there is a candybar and in which there is a frog). The prediction scenario corresponds to the training phase in a sequence model, while the choice scenario corresponds to the inference phase (see section 1 above).

In the training phase, in which the aim is to predict the action which will result in finding a candybar by watching an expert do it, a probabilistic model is devised to characterize the problem: $P(A_{n+1} = a_{n+1} \mid a_{1:n}, o_{1:n})$ where a is an action to be taken and o is an observation recording the outcome of the action. Basically, given that Θ is not known to the agent, so it's a latent variable in the model, the task will be to predict the action A and the observation O, such that the probability of A leading to outcome O is maximized.

A sequence model will do that by using the probability distributions $P(A)$ and $P(O \mid A)$, given the conditional probability formula:

(Cond) $P(a, o) = P(a)\, P(o \mid a)$.
Thus:
i) The model will start by predicting $P(a) =$ ½ (because it does not know in which box there is a candybar).

ii) Then it will observe the expert issue an action $A = a$.
iii) Since the expert who issues action *a* knows the configuration of the boxes, the subsequent observation O = *candybar* will follow with certainty. Therefore, given the dependency of *A* on Θ, the posterior probability $P(o \mid a)$ will be 1, since:
$P(o \mid a) = 1$ when $o = candybar$ and $P(o \mid a) = 0$ when $o = frog$.
iv) So, $P(a)$ will be updated to 1, such that:
v) $P(a, o) = 1$.

Let us now move to the second scenario, in which the model is required to *choose* by itself the action which will result in finding a candybar. Given that it does not know the configuration Θ of the boxes, the task is difficult. It will have to choose an action A such that the probability of A leading to the outcome O is maximized. As before, it will do it by using the probability distributions *P(A)* and *P* (*O* | *A*), given the conditional probability formula:

(Cond) $P(a, o) = P(a)\, P(o \mid a)$.
Thus:
i) it will first suggest $P(a) = ½$ (because it does not know the configuration of the boxes).
ii) then it will *choose* an action $A = a$ by sampling it from $P(A)$, which is the probability distribution of the expert's past actions.
iii) since all expert's past actions are successful (because A is dependent on Θ), the subsequent observation $O = candybar$ will follow with certainty, hence $P(O \mid A) = 1$.
iv) So, $P(a)$ will be updated to 1, such that:
v) $P(a, o) = 1$.
And yet, the model will not be successful, because the probability of *a* should still be $P(a) = ½$.

The explanation for this error is that here we have a *confounder*: by training the model only on expert-generated data, the model falsely infers from observations of the expert's actions that *all* actions sampled from the probability distribution $P(A)$ are successful actions. Thus, the knowledge the expert possesses about the configuration of the boxes creates a spurious correlation in the model between actions and their outcomes. The model could have sampled any other action from the distribution $P(A)$ and the outcome would have been the same, that is:
$P(O = candybar \mid a) = 1$ (Ortega *et al.*, 2021).

This is precisely a delusion in the general sense of Pearl & Mackenzie (2018), that is "an illusion we brought upon ourselves by choosing which events to include in our data-set and which to ignore." (p.185). In particular, it is a delusion resulting from confounding: *i.e.*, from taking something that we do as something that we see.

3. Auto-suggestive delusions, an analogy with rule following

If we are to take an even more abstract stance to probabilistic delusions in sequence models, the following considerations will provide a stimulating perspective.

"This was our paradox: no course of action could be determined by a rule, because any course of action can be made out to accord with the rule. If every course of action can be brought into accord with the rule, then it can also be brought into conflict with it. And so there would be neither accord nor conflict here." (Wittgenstein, 1958, §201)

The way in which the rule-following paradox is usually articulated within the philosophy of language and mind is not only intriguing but also remarkably versatile. It possesses the generality required for accommodating various formulations, reaching far beyond its original scope (see Peacocke & Kripke, 1985). The trick that makes the case of sequence models relatable to the rule-following paradox is the perplexing result: $P(O = 1 \mid a) = 1$, for any a that is sampled from $P(A)$, obtained in the previous sections. In other words, although actions are not easy to predict from previous actions and observations of their outcomes (*e.g.*, by determining a rule, such as a marginal distribution), it is significant that actions are never *groundless*.

One straightforward interpretation of the rule-following paradox is, for example, that a rule is unable to uniquely dictate a specific course of action because, by merely observing actions, one cannot determine the general rule or the next action. The paradox, however, is not that perplexing and not really a paradox once we analyze it as a problem of "inverse" probabilities (Bayes) as we did previously, in scenario 1 of the "prize or frog" problem. The idea can be illustrated by looking at the two cases below:

(I)
```
Whenever I'm hurt, I cry.
I am hurt.
---hence--->
I cry.
```

(II)
```
(Whenever I'm hurt, I cry.)
I cry.
---hence--->
Am I hurt?
```

In Bayesian language, case (I) is a forward-probability problem: I know that I am hurt, and I want to know the probability of me crying. Case (II), on the other hand, is an inverse-probability problem: I know that I cry, and I want to know the probability of me being hurt. It is interesting that the Bayes formula allows one to switch between cases (I) and (II), depending on the amount of information one has in each of the cases, thus deriving the probability of an event in one direction from the probability of an event in the other direction.

(forward probability) $P(X, Y) = P(Y \mid X)\, P(X)$.

(inverse probability) $P(X, Y) = P(X \mid Y)\, P(Y)$.

(Bayes formula): $P(Y \mid X)\, P(X) = P(X \mid Y)\, P(Y)$.

In our example, case (I) provides more information than case (II), because in the first case one knows the rule. So, by applying Bayes' formula, which allows to derive case (II) from case (I), one can easily determine the inverse probability $P(Hurt \mid Cry)$ as:

$P(Hurt \mid Cry) = P(Hurt, Cry) \,/\, P(Cry)$.

This means that the rule-following paradox has no true bearing in a purely *factual* setting. Once the probability of the rule $(Hurt, Cry)$ is determined from the forward probability, one can use it to update the inverse probability, and thus determine to which extent the fact that I cry indicates that I am hurt.

In a more sophisticated interpretation, however, the rule-following paradox could suggest something more interesting, which resembles scenario 2 in the previous "prize or frog problem". Suppose, this time, that our setting is not factual and our purpose is not to merely predict an action from a rule (or a marginal distribution), but instead *act* by applying the rule: that is, the setting is now *normative*. The paradox, in this interpretation, implies that one cannot apply or follow a rule by merely imitating someone who is known to apply the rule, because there is no way to determine which actions do indeed count as applying the rule, and which fail to do so. For example, one can pretend to be making a fine-dining dish by merely watching, hearing, and reproducing what Gordon Ramsey does in his fancy kitchen. Nevertheless, such an approach often ends up in culinary disaster, even when no step of the recipe is missed. The reason for this is deeply embedded in the normativity of rule following: no matter what action the apprentice chooses to perform, she deems it as *correct* because it is what the expert was seen doing. In this sense, predicting an action from observing actions conforming to the rule and acting by applying a rule are entirely two different things.

In order to see this more clearly, let us consider another two cases (simplified from the "prize or frog" problem).

(Ib) *Expert*

```
1)Whenever fire,
call the firefighters,
kill fire.
2)(Fire, hidden) Call the firefighters



---hence--->
3) Kill fire
```

(IIb) *Agent*

```
1) (Whenever fire,
call the firefighters,
kill fire.)
2)Call the firefighters (sampled
from the expert's actions, who does
not call the firefighters for no
reason)
---hence--->
3) Kill Fire, with probability P=1
```

```
                                        (given the expert-induced
                                        association between calling
                                        firefighters and fire)
---hence--->                            ---hence--->
4) Call the firefighters,               4) Call the firefighters,
with probability P=1.                   with probability P=1.
```

First of all, the transition from step (2) to step (3) in the Agent scenario can be labeled as a probabilistic delusion because a confounder is present: *i.e.*, the *normative* or expert-induced association between calling firefighters and extinguishing a fire (see Fig.5 below[3]). Second, the transitions in steps (3) and (4) mark the *auto-suggestive delusion* because, given the spurious normative association between actions and their outcomes, which cannot be made explicit in the model, the agent ends up deviating completely from reality: even if in reality there is no fire, the agent will keep on summoning the firefighters to extinguish a fire, in an endless loop, without realizing that it is wrong. The explanation for the auto-suggestive delusion is hence that, due to the confounder present in the model, the agent *is unable to distinguish its actions from those of the expert*, treating its own actions (step 2) as expertise (step 3), and then acting accordingly (step 4)[4].

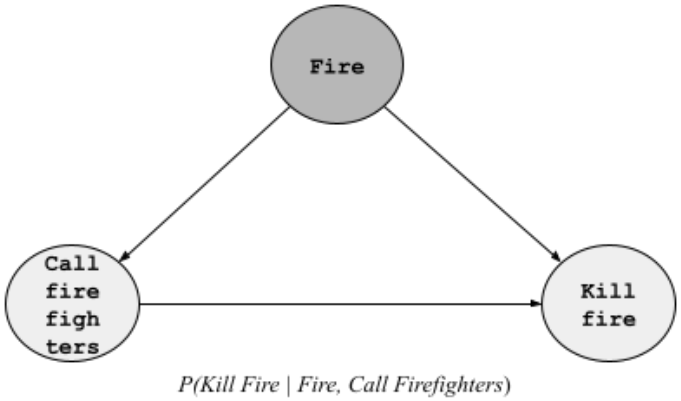

Figure 5. Diagram representing the self-delusion problem (Ortega *et al.*, 2021:5). Conditioning on the model's self-generated action (Call firefighters) leads to wrong inferences about the outcome (Kill Fire), because, due to confounding, an action and its outcome are both determined by the state of the world (*i.e.*, by whether there is a fire or not).

[3] In order to draw the diagram at least three (random) variables are required, so I modified the example a little, in order to accommodate three variables.

[4] As Ortega *et al*. (2021) point out: "The reason is subtle: the model update triggered by the collected data differs depending upon whether the data was generated by the model itself (*i.e*. actions) or outside (*i.e.* observations), and mixing them up leads to incorrect inferences. These take the form of *self-delusions* where the model takes its own actions as evidence about the world (...) due to the presence of confounding variables." (p. 2).

In the diagram above, the variable labeled as "Fire" (which records the state of affairs existent in the world, which the expert knows in order to ground their actions) is a confounding factor, introducing in the probabilistic model a spurious correlation between the action "Call firefighters" and the singular outcome "Kill fire". Because of this spurious correlation, the posterior probability denoted in the model as *P (Kill Fire | Call Firefighters*) becomes equal to 1, regardless of how things stand in the world. So, the agent never understands that something is wrong with its representation of the world. *Mutatis mutandis*, someone who wishes to learn how to make a dish by imitating Gordon Ramsey's moves in the kitchen will probably fail to understand that they failed to prepare Gordon Ramsey's dish, because they have no idea how it should taste. So probably they will keep on going about it the wrong way. This understanding of auto-suggestive delusions aligns, in fact, with another core aspect of the rule-following paradox, which emphasizes the idea that to follow a rule is something that is somehow entrenched in what we do, turning what we do into a correct or an incorrect action[5]. As Ortega *et al.* (2021) spell it out, "imitation requires knowing the reasons behind actions" (p. 8).

What makes an action correct or incorrect is captured in the reasons for doing it (in the "why"-s for taking that specific action). Following a rule resides in knowing what a correct or incorrect action is with respect to that rule. The expert, in our example, calls the firefighters because there is an actual fire that is getting out of hand. Their action is the correct one – it is done for the right reason and it leads to the extinguishing of an actual fire. Imitating the expert and therefore producing a correct action requires the model to understand the expert's reason for taking that specific action. But the model does not understand the reason behind the expert's action. Only due to confounding, it sometimes happens that the model takes the action that conforms to the rule (*i.e.,* call the firefighters when there is an actual fire). In this sense, the model exhibits only *apparent* expertise.

That the model's expertise is only apparent becomes evident once we get rid of the confounder and notice that the behavior of the model changes. When we block the confounder (which contains, in fact, the expert's reasons for their actions), the model starts to behave in a random manner (see Fig.6). Note now that the confounder „Fire" is blocked by forcing a value on the action „Call firefighters", in accordance with the general logic of *do*-calculus (Pearl, 2009; Pearl & Mackenzie, 2018). To force a value on an action means simply *to do* something.

In our case, the agent can choose at random between calling or not calling the firefighters, that is: $do(a) = 1 \; or \; do(a) = 0$, which under the back-door adjustment in *do*-calculus and given the uncertainty of „Fire", yields:

(F) $P(o \mid do\ (a)) = \sum_F P(o \mid F, a)\ P(F) = 1/2$, or in other words:

[5] It is interesting how such idea comes up in the *Philosophical Investigations*: "For what we thereby show [by means of the rule-following paradox - n.a.] is that there is a way of grasping a rule (...) which, from case to case of application, is exhibited in what we call 'following the rule' and 'going against it'." (Wittgenstein, 1958, §201)

$P(Kill\ Fire \mid do(Call\ Firefighters)) =$
$\sum_{Fire} P(Kill\ Fire \mid Fire, Call\ Firefighters)\ P(Fire) = 1/2.$

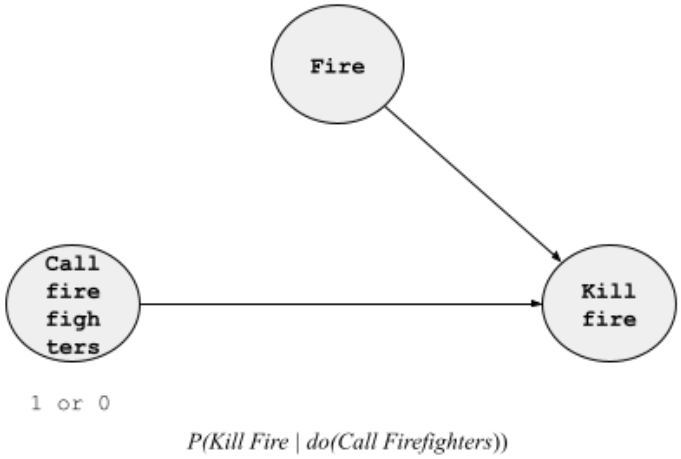

Figure 6. A causal diagram representing the blocking of the confounder Fire, by forcing a value on the action Call firefighters, according to the rules of *do*-calculus.

Formula (F) simply states that the probability to extinguish a fire, given that the firefighters were called, remains uncertain. Yet this result – unlike the paradoxical one in (IIb) – makes sense: selecting an action at random (to call or not to call the firefighters) does not provide any evidence about the real world, whose state (whether there is a fire to be extinguished or not) remains uncertain based on that evidence.

Also, that this is essentially a problem of confounding is evident from:

$P(Kill\ Fire \mid Call\ Firefighters) \neq P(Kill\ Fire \mid do(Call\ Firefighters))$.

And if the example we used so far seems artificial, we can turn to a more familiar example of auto-suggestive delusion in large LMs. Let us say we ask ChatGPT, the chat-box based on GPT-3.5, to provide us with a list of references on a given topic. At a certain point, however, we notice that the reference list it has compiled for us contains a mix of relevant authors and titles, but some listings are completely wrong, such as Ludwig Wittgenstein being listed as the author of *Wittgenstein on Rules and Private Language*, which would imply that Ludwig Wittgenstein actually wrote a book about himself. This kind of delusion or "hallucination" has been reported by a large number of sources (see Sobieszek & Price, 2022). In fact, it has a straightforward formal explanation once we return to our notion of confounding.

Let us look at Fig.7 below. The diagram emphasizes the underlying fact that some thematic grouping (or Theme) operates as a confounder in the model, inducing a spurious association between authors and titles.

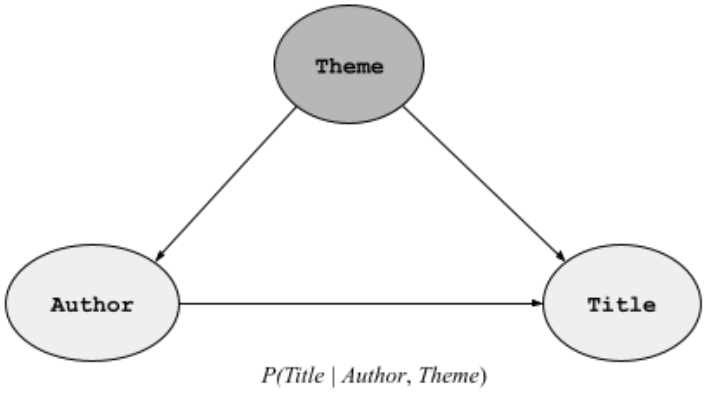

Figure 7. Diagram representing the self-delusion problem in the reference list case. Conditioning on the model's self-generated action (Author) leads to wrong inferences about the outcome (Title), because, due to confounding, any listing of an author and a title depends on the thematic grouping (Theme).

We can envision that when it comes to portraying a specific subject, taking into account its extensive training data, the GPT-3.5 model has acquired knowledge of ten notable author names along with several dozen noteworthy book titles. We can subsequently request the model to produce a list that includes a minimum of eleven author names, along with the titles of the works they've penned on that particular subject (refer to the table in Fig.8 for an illustration of this task).

Theme	**Author**	**Title**	***P(Title | Author)***[6]
x	1	a	1
	2	b	1
	3	c	1
	…	…	…
	10	j	1
	3	**k**	**1**

Figure 8. Table illustrating all the entries corresponding to the three variables in the model (Theme, Author, Title). Conditioning on the model's self-generated action (Author) - see the last column in the table - leads

[6] Theme is hidden, just like in the previous examples.

to wrong inferences about the outcome (Title) - see the last row in the table. The explanation is that Author and Title are confounded by Theme.

Considering that the underlying thematic grouping (Theme) acts as a confounding variable, such that only ten author names are included in the probability distribution of the agent's representation of the topic, when asked about filling the eleventh entry (comprising both author name and title), the model assigns a title incorrectly, by mistakenly linking a new title to one of the author names already listed in the compilation.

This is a case in which the confusion between doing and seeing surfaces neatly: when the agent is asked to prompt an author, it simply samples it from the list it gathered before, generating a misleading response. This aspect is widely discussed in recent literature (see Sobieszek & Price, 2022) under "exposure bias". However, the issue lies deeper and concerns what the agent does in order to solve the task: that is, it takes its own generated actions as observations of the outside world, in the sense that, for the eleventh title listed under the topic, it predicts with certainty that one prominent author among the ten ones included already the list authored that work (see the last row and column in the table above), so it issues that action, which is, of course, wrong. The probability distribution that should model the world is, in fact, the probability distribution of the model's own actions. If we are to return to our discussion, the deep issue here is that the model produced an auto-suggestive delusion.

4. Private language and the question of solipsism in large Language Models

The rule-following paradox has a prominent linguistic facet, discussed extensively in the philosophical literature (see Soames, 2003). For example, we may have the impression that we really understand what one means when one writes: „a cat meows" by simply inspecting the words. But one might mean a cat barks, or a cat moos, or something else entirely. Why does this not usually happen? A very short answer would be that, for us, speaking and following rules are practices *that we share* (see Smith, 1998); we cannot simply mean anything we want. Linguistic practices are, consequently, normative. Using words has consequences that are sanctioned by the larger group of speakers.

„To think one is following a rule is not to follow a rule. (...) that's why it's not possible to follow a rule 'privately'; otherwise, thinking one was following a rule would be the same thing as following it". (Wittgenstein, 1958, §202)

Starting from such general considerations, the rule-following paradox is often discussed in relation to the possibility of a so-called "private language" (see Kripke, 1982). Given that speaking a language is generally akin to following a rule, the question arises as to whether there can be such a language that is entirely private – that is, a language whose rules are known only to *one* speaker,

without being derived from any public language, like the human language. The question is tempting because, due to auto-suggestive delusions, large Language Models seem to behave as if there are speaking their own language and describing their own world. So, can it be that large LMs exhibiting auto-suggestive delusions are actually following their own rules, in which there is no distinction between actions and observations, such that even if their outputs are growing progressively remote from what we know to be the world, to the LMs they are entirely coherent, describing, in fact, an "alternate" world belonging only to the models themselves?

The philosophical answer to this question would be "no". As the quote above suggests, a rule cannot be followed privately because if that were the case, then anything anyone would ever do might count as following the rule. Hence, there will be no rule (since it would not exclude anything). Speaking a language, following a rule are normative practices in the sense that they are always sanctioned from a stance *outside* the agent herself. This is also why solipsism, as a stance in which only the agent exists, is neither possible, because no agency can be conceived without reasons, hence without rules, hence without a stance *outside* the agent herself.

In large Language Models, like GPT-3 or GPT-3.5, such normative practices are missing. When a model is asked to answer a task by generating the next word (token) in a sequence of words (tokens), based on a distribution learned from previously observed sequences of tokens, the model can, in fact, use *its own* sampling from that distribution, even when there is a form of human-in-the-loop or RLHF in the fine tuning of the model. "Consider a pre-trained language model whose job is to predict the fourth word given a sequence of three words proposed by us (experts). Without any loss of generality, we again introduce θ to capture information that is available to the experts, but not to the agent (...) The problem arises when we consider the interaction between the pre-trained language and the environment."(Ortega *et al.*, 2021, p. 9). In other words, the problem appears when we shift from the training to the inference phase. "Suppose we use a language model API to enter the first word x_1, but this time the model uses its own generated second word x_2, and then we force the model to use our third word x_3. We then try to predict the fourth word as before." (*ibidem*). Nothing impedes the model to predict the fourth word in a manner that to us may seem nonsensical, *e.g.* "the dog moos loudly". "In the extreme case, imagine that the language model generates a lot of text and that text is added to the data-set, say a web corpus. Then, relearning from this data-set will only confirm the model's biases, that is, its delusions" (Ortega *et al.*, 2021, p. 10). What this discussion illustrates is, hence, a clear tendency of accumulating bias, delusions, or nonsense, due to the absence of normative practices in the manner in which text is generated by the model. In the inference step, when the model imitates the expert, if fails to grasp the reasons behind the expert's choice of words. By failing to grasp the reasons behind the expert's choice of words, the model treats its own actions as observations of the expert's own words. This further accumulates during retraining, when new text is added to the data-set, including text generated by the model itself, thus generating a spiral effect, which pushes the model further and further away from the manner in which human language is usually used.

However, it cannot be said that model generates a private language *per se*, because the tokens it sequences are not based in a normative private practice (which, in fact, is impossible as the philosophical argument shows). Yet, what the model does in going further and further astray from the norms of human language is *systematic*. This is why a less radical concept of private language might be useful: it could be more suggestive than the "gibberish" or "parroting" that is usually ascribed to large LMs. The idea of a private language could emphasize the fact that the model's tendency to accumulate bias, delusions, or nonsense is systematic and more pronounced with each retraining of the model. In the weak sense, a private language may be understood as a language that is systematically and increasingly deviating from human language; without being a coherent language in itself, it perseveres however some of the original coherence of the language that it gradually deviates from. A large LM's actions (like writing words in a certain sequence) are systematic misrepresentations of the experts' actions that it seeks to imitate (*i.e.*, the experts' writing of words in a certain order), due to confounding. Even if a sequence of words generated by the model may make sense (thus, resembling human language), this would only happen by accident – *i.e.,* as a consequence of the fact that the confounder's presence can sometimes be innocuous to the expected output. Retraining the model, however, will only make the confounder less and less innocuous.

The correlate of that, when we take language tokens as data, is that even though the model could learn a stochastic representation of the world, it could still take its own actions (the generated data) as evidence (as ground-truth data) about the world, confusing his own generated data with what it observes (the ground-truth data). It is significant that we too face this kind of delusion in observational studies (Pearl & Mackenzie, 2018), in which it often happens to confuse data that we generate, through our collecting process, with ground-truth data. And regardless whether one is human or machine, this is akin to a solipsist delusion. This is so, because making an observation is something that we *do*, rather than something that we *see*; therefore, if we are not attentive enough, the manner in which we represent the world (the "do") becomes the world itself (the "seen").

Humans, however, have ways of mitigating this problem: causal inference is one of them; normative practices are another. Artificial agents, on the other hand, still have a long way to go in both of these directions, although according to recent literature, progress is made. With respect to causal inference, methods like "counterfactual teaching" are developed (Ortega *et al.*, 2021, p. 11). In counterfactual teaching, the agent issues an action by sampling it from the actions of the expert, as we have seen before. Then the action is compared against the expert's own action, which is revealed to the agent, followed by a penalty cost applied to the agent. In this manner, which amounts to sequentially minimizing the penalty cost, an agent learns the consequences of its actions, improving its choices along the way. In Language Models, something similar can be done, which induces a sort of normative practice, *i.e.*, a practice governed by rules and expertise. In recent literature, this is called "teacher forcing" (He *et al.*, 2021): teacher forcing works by using

the actual or expected output from the training data-set at the current time step as input in the next time step *t+1*, rather than the output generated by the model (Goodfellow *et al.*, 2016, p. 372).

Substantially, however, both "counterfactual teaching" and "teacher forcing" require a form of supervised learning that assumes a recurrent architecture of the neural network; so, it seems that there are no strong or truly effective ways to learn counterfactuals and normative practices other than by explicit teaching in a network that is recurrent. This means that large LMs cannot really learn *what we expect them to learn from scratch* – that is, only in feed-forward fashion from what they observe, based on their transformer architecture. This also means that simply scaling-up already large models in unsupervised learning scenarios, even with RLFH fine tuning, may not lead to the outcome that we expect, as often pointed out (Bender *et al.*, 2021). On the other hand, supervised models based on recurrent network architectures lack the capacity for generalization and novelty of the unsupervised transformer models. In their private and solipsistic fashion, understood as above, unsupervised transformer models are, in fact, terrifically creative (Boden, 2009; Franceschelli & Musolesi, 2023). Put in most general terms, this may be because they are not bound by a causal representation of the world, nor by normative practices like we are; in short, they do not share our "form of life". So, making them learn our "form of life" only from observing it, but not interacting with it, presents a genuine challenge, if not a genuine illusion.

5. Concluding remarks

The tendency of sequence models or transformers to accumulate bias, delusions, or nonsensical output is based in a confusion between their own generated data and ground-truth data. Given that this tendency is systematic, it could be said that it is reminiscent of a solipsist delusion, where observations end up in masking actions and so, observations fail systematically to provide an objective picture of the world. This phenomenon has parallels in human observational studies, where data generated through the collection process can be inadvertently equated with the ground-truth data, therefore blurring the line between representation and reality in an equally systematic manner.

On the other hand, the rule-following problematic, extensively discussed in philosophical literature, raises questions about the general ability of sequence models or transformers to understand language and follow rules. Linguistic practices are typically normative. In the context of large Language Models, like GPT 3 or GPT-3.5, normative practices are missing. Such models, designed to generate text based on learned probability distributions tend to produce nonsensical responses when engaging with humans. This tendency exhibits a systematic nature. Because it can be traced back to the absence of normative conventions or practices, this tendency may be considered similar to the emergence of a private language in a weak sense: *i.e.*, as a *systematically* growing production of biased, delusional or nonsensical text, due to the fact the models fail to

distinguish what they do (their own words) from what they see (the experts' words). Fine tuning with a human-in-the-loop architecture can address this issue partially, and this is why models like GPT 3.5 are performing better than their transformer ancestors, like BERT of GPT-3. However, overall, a recurrent architecture might be more suited for addressing the issue than the transformer architecture (even one with RLHF).

References

Amatriain, X., Sankar, A., Bing, J., Bodigutla, P. K., Hazen, T. J., & Kazi, M. (2023). Transformer models: an introduction and catalog. *arXiv preprint arXiv:2302.07730*. https://doi.org/10.48550/arXiv.2302.07730

Aralikatte, R., Narayan, S., Maynez, J., Rothe, S., & McDonald, R. (2021). *Focus Attention: Promoting Faithfulness and Diversity in Summarization*. https://doi.org/10.48550/arXiv.2105.11921

Baker, S., and Kanade, T. 2000. Hallucinating Faces. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition* (Cat. No.PR00580). 83–88. https://doi.org/10.1109/AFGR.2000.840616

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Boden, M. A. (2009). Computer Models of Creativity. *AI Magazine*, *30*(3), 23–34. https://doi.org/10.1609/aimag.v30i3.2254

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language*

Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

Dhingra, B., Faruqui, M., Parikh, A., Chang, M.-W., Das, D., & Cohen, W. W. (2019). *Handling Divergent Reference Texts when Evaluating Table-to-Text Generation*. https://doi.org/10.48550/arXiv.1906.01081

Franceschelli, G., & Musolesi, M. (2023). On the creativity of large language models. *arXiv Preprint arXiv:2304.00008*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

He, T., Zhang, J., Zhou, Z., & Glass, J. (2021). *Exposure Bias versus Self-Recovery: Are Distortions Really Incremental for Autoregressive Text Generation?* https://doi.org/10.48550/arXiv.1905.10617

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kripke, S. 1982. *Wittgenstein on Rules and Private Language*. Harvard University Press.

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., & Wen, J.-R. (2023). *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. https://doi.org/10.48550/arXiv.2305.11747

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. T. (2020). On Faithfulness and Factuality in Abstractive Summarization. *CoRR*, *abs/2005.00661*. https://arxiv.org/abs/2005.00661

Ortega, P. A., & Braun, D. A. (2009). A Bayesian Rule for Adaptive Control based on Causal Interventions. *ArXiv*, *abs/0911.5104*. https://doi.org/10.48550/arXiv.0911.5104

Ortega, P. A., Kunesch, M., Delétang, G., Genewein, T., Grau-Moya, J., Veness, J., Buchli, J., Degrave, J., Piot, B., Perolat, J., Everitt, T., Tallec, C., Parisotto, E., Erez, T., Chen, Y.,

Reed, S., Hutter, M., Freitas, N. de, & Legg, S. (2021). *Shaking the foundations: Delusions in sequence models for interaction and control*. https://arxiv.org/abs/2110.10819

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Penguin Books Limited. https://books.google.ro/books?id=EmY8DwAAQBAJ

Rezende, D. J., Danihelka, I., Papamakarios, G., Ke, N. R., Jiang, R., Weber, T., Gregor, K., Merzic, H., Viola, F., Wang, J., Mitrovic, J., Besse, F., Antonoglou, I., & Buesing, L. (2020). *Causally Correct Partial Models for Reinforcement Learning*.

Smith, B. C. (1998). *Meaning and rule-following*. https://doi.org/10.4324/9780415249126-U021-1

Soames, S. (2003). CHAPTER 2. Rule Following and the Private Language Argument. In *The Age of Meaning* (pp. 32–64). Princeton University Press. https://doi.org/doi:10.1515/9781400825806-004

Sobieszek, A., & Price, T. (2022). Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. *Minds and Machines*, *32*(2), 341–364. https://doi.org/10.1007/s11023-022-09602-0

Wittgenstein, L. (1958). *Philosophical investigations* (2nd ed.). Basil Blackwell.

Zheng, S., Huang, J., & Chang, K. C.-C. (2023). Why Does ChatGPT Fall Short in Answering Questions Faithfully? *ArXiv*, *abs/2304.10513*. https://doi.org/10.48550/arXiv.2304.10513