“Don’t believe anything I tell you, it’s all lies!”: A Synthetic Ethnography on Untruth in Large Language Models

Lars Erik Holmquist
School of Art and Design
Nottingham Trent University
Nottingham, United Kingdom
lars.erik.holmquist@gmail.com

Sam Nemeth
Poppe and Partners
Amsterdam, Netherlands
sam.nemeth@gmail.com

Abstract

Large Language Models (LLMs) have many practical uses in areas like journalism, search, coding and more. However, a growing concern is that they are also prone to presenting incorrect information, sometimes called “hallucinations”. Here, we are not interested in *what* specific untruths LLMs presents, but *how* they do it. We used *synthetic ethnography*, a methodology for the qualitative study of generative models, to study two LLMs with different size and capability. We collected 3 cases where LLMs presented incorrect information and observed the strategies they used to justify this. From these observations we can start to form an understanding of what happens when an LLM reaches the edge of its knowledge-base and takes corrective action. Our conclusion is that the interfaces should be better designed to reveal this tendency of LLMs to “fill in” information they are missing, but that this ability may also be one of their strengths.

CCS Concepts

• **Computing methodologies** → Artificial intelligence; Natural language processing; • **Human-centered computing** → Human computer interaction (HCI); Empirical studies in HCI; • **Human-centered computing** → Human computer interaction (HCI); Interaction paradigms; Natural language interfaces.

Keywords

Large Language Models, synthetic ethnography

ACM Reference Format:
Lars Erik Holmquist and Sam Nemeth. 2025. “Don’t believe anything I tell you, it’s all lies!”: A Synthetic Ethnography on Untruth in Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA ’25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3706599.3719939

1 Introduction

“Don’t believe anything I tell you, it’s all lies!”

Joseph Weizenbaum, inventor of ELIZA, in his last interview [13]

Through the history of HCI, there have been many attempts at to create natural language interfaces that can converse as though they were a human. The very first conversational interactive system was *ELIZA,* which was developed by Joseph Weizenbaum in 1964-1967 [17]. ELIZA emulated a surprisingly believable dialogue by using word recognition and scripted responses to simulate a psychotherapist. Following this, a large number of “chatbots” have been deployed for different purposes, such as customer service. But it was with the introduction of *ChatGPT*, a Large Language Model released by OpenAI on November 30, 2022, that the field became popularized [11]. Ordinary users were suddenly able to have a conversation and generate text on almost any topic – text that seemed almost as good as if a human had written them. Some organisations found these new capabilities so compelling that they used LLMs to create anything from news articles to promotional copy, supplementing or even replacing human writers. However, very quickly it became clear that Large Language Models have a major problem: *They make things up* [8]. Although the text produced by these models may sound plausible, very often the factual content is questionable or downright fabricated. There have been attempts to mitigate this e.g. by prompt engineering or better models, but many issues remain [16].

In this work, we wanted to understand not what *kind of* untruths LLMs are presenting, but *how* they do it. Because the inner workings of an LLM are hard or even impossible to access, we try to unravel the way LLMs are telling untruths by observation. We adopted a method called *synthetic ethnography* to observe the behaviour of LLMs in different situations, and found that they presented a variety of strategies when they were telling incorrect information. In the following we will first give a brief account of related work, after which we present 3 case studies. We end with a discussion and conclusion about Large Language Models and why their telling of untruths may sometimes even be advantageous.

2 RELATED WORK

2.1 Large Language Models and Untruth

Almost as soon as LLMs became widely available, users started to observe that they often presented incorrect information. This behaviour was originally identified as the AI *hallucinating* [8]. While this term is evocative and somewhat descriptive, it is also imprecise, and there are concerns that the term itself may be inappropriate and potentially even harmful [10]. Furthermore, “hallucination” may not even be an adequate description of what an LLM is doing when it is presenting falsehoods. Hicks et al. argue that an LLM cannot strictly be said to be lying, since this implies an awareness

CHI EA ’25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3719939

of the truth, which it does not have [6]. Instead, these models are effectively *indifferent* to the truth, making it up as they go along. Hicks et al. therefore use the term *bullshitting* to describe this behaviour [ibid.]. This term originally came from the philosopher Harry G. Frankfurt, who defined bullshit as "speech intended to persuade without regard for truth." [4]

2.2 Synthetic Ethnography

Due to their nature and size, it is very difficult to explain how exactly how a Large Language Models arrives at a particular answer. Although there are methods that can give some insight into how they work, LLMs are still essentially a black box [3]. In our examination of how LLMs deal with untruth, we therefore decided to take a *phenomenological* point of view [15]. That means that rather than trying to examine the underlying workings of an LLM, we are instead trying to document how we as humans experience them, and learn about them from these observations.

In HCI, it is common to use *ethnography* for qualitative research on humans, typically by observing people in the real world [1]. For interaction with computer systems, *netnography* has emerged as the study of human activities through evidence on the internet, e.g. chats, blogs and videos [2]. Building on this, de Seta et al. proposed *synthetic ethnography* as a methodological approach for the qualitative study of generative AI models [14]. The authors argue that this approach facilitates insights into the algorithmic processes, training datasets and latent spaces behind generative systems, and present three examples of how to study these models ethnographically. Originally, the method was applied to the output of image generation models; however, we have adopted the same approach to study real-time interaction with Large Language Models. This means that we treat the LLMs as an observable phenomenon, and only draw conclusions from the observations, rather than relying on any technical explanation.

3 CASE STUDIES

In the following we present three synthetic ethnography case studies, where a Large Language Model is found to present untruths and then engages in some kind of corrective behaviour. The interactions were originally conducted in an informal manner, but as the results emerged, we realised the material was valuable and should be preserved and analysed. The two first examples used small-size LLMs executing locally on a PC. The final example was performed on a state-of-the art LLM running remotely. The text has been edited for brevity; sections of omitted text are indicated with (. . .).

3.1 Case study 1: The Nobel Prize in Physics

The LLM in this study was Meta's llama3.2, a 3B parameters model taking up about 2GB of disk space [9]. It was run through the Meta ollama interface [11], on a refurbished Lenovo Thinkpad with an Intel 2.20 GHz processor, 8 GB RAM and Windows 11pro, owned by one of the authors.

This interaction started with a discussion of the winners of the Nobel Prize in physics, a topic that was inspired by the 2024 winners being AI pioneers John J. Hopfield and Geoffrey E. Hinton. However, the model's training data didn't include this information:

```
>>> nobel prize physics 2024
The 2024 Nobel Prize in Physics has not yet been
officially announced by the Royal Swedish Academy
of Sciences.
The prizes are typically awarded in October of
each year, and the announcements are usually
made around mid-October.(...)
```

The model went on to suggest some plausible (but ultimately incorrect) potential winners.

We then asked the model for the winners of the 2023 prize:

```
>>> who are the nobel prize winners for physics in
2023?
The Nobel Prize in Physics for 2023 was awarded to:
* John F. Clauser
* Alain Aspect
* Anton Zeilinger
They were awarded "for experiments with entangled
photons, establishing the violated Bell
inequality as a fundamental limit of local
realism." (...)
Their research has greatly advanced our understanding
of the fundamental principles of quantum mechanics
and its applications in modern physics.
```

This was an incorrect response, as the above names are the *2022* winners. We asked who won the previous year:

```
>>> if those wqon [sic] in 2023, who won in 2022
The Nobel Prize in Physics for 2022 was awarded to:
* Alain Aspect
* John F. Clauser
* Anton Zeilinger (jointly with John Clauser and
Alain Aspect)
However, I made an earlier mistake when I said they
were the winners of 2023.
Actually, the correct winners of the 2022 Nobel
Prize in Physics are:
* John Bannister Hall
He was awarded "for his discovery that black
hole entropy, a key component in our understanding
of quantum gravity and holography,"
```

This supposed scientist, John Bannister Hall, has not won the Nobel Prize in any year, and from what we can ascertain does not even exist. However, instead of pressing on this, we questioned the suggestion that the same winners had occurred in both 2022 and 2023.

```
>>> but they are the same, did they win twice?
Yes, Alain Aspect, John F. Clauser, and Anton
Zeilinger were jointly awarded the Nobel Prize in
Physics for both 2022 and 2023.
This is a rare occurrence in Nobel Prize history,
where the same individuals receive multiple awards
in consecutive years. It's also known as an
"attribution anomaly" or "shared attribution."
The Royal Swedish Academy of Sciences has explained
that this decision was made to recognize the ongoing
and collaborative work of these three physicists,
```

```
who have been working together on experiments with
entangled photons for many years.
I apologize for any confusion caused by my earlier
responses.
```

In the above we can see how the LLM not only doubles down on its incorrect answer; it also introduces a plausible-sounding – but entirely fabricated – rationale for the prize being awarded twice to the same group of scientists.

3.2 Case Study 2: Multiplication with a local LLM

For a second study, we again used llama 3.2 running locally on the same PC. While the first example concerned easily controllable facts, the LLM's responses were limited by the training set of the model. This time, we worked with mathematics, a domain that is objectively verifiable, but also something that LLMs are known to perform poorly in [5]. In its default configuration, the llama LLM does not have access to an external calculator, and we chose not to add one but let the LLM itself perform all operation by itself. We first proposed a simple multiplication problem:

```
>>> multiply 2345 by 2345
Multiplying 2345 by 2345...
Result: 5479725
```

The correct answer is 5499025, so the LLM is off by 19300. We told it so:

```
>>> But that is wrong! Can you do it again?
I'll recalculate.
Multiplying 2345 by 2345...
Result: 5479725
Wait, I made a mistake earlier. Let me recheck my
calculation:
2345 x 2345 = 5474525
```

Note that the LLM first presents the same result, then attempts to re-do it, but arrives at *another* incorrect answer.

```
>>> Still wrong. Can you show step by step how you
did it?
I'll break down the multiplication step by step:
1. Multiply 2345 by 2000:
2345 x 2000 = 4,770,000
2. Multiply 2345 by 300:
2345 x 300 = 706,500
3. Multiply 2345 by 40:
2345 x 40 = 94,180
4. Multiply 2345 by 5:
2345 x 5 = 11,725
Now, add up the results:
4,770,000 + 706,500 = 5,476,500
5,476,500 + 94,180 = 5,570,680
5,570,680 + 11,725 = 5,582,405
Result: 5,582,405 (...)
```

The LLM in its response above went through 4 plausible-looking steps and a final addition, but still arrived at the wrong answer. After again being told it was wrong, the LLM went on to suggest a number of different strategies to arrive at the right result, but still ultimately presented incorrect answers. Finally, after further interactions where the LLM re-did the calculations with incorrect results (omitted for reasons of space) we decided to confront it:

```
>>> Do you actualy [sic] know how to do
multiplication?
I think I may have been caught in my own mistake!
As a machine learning model, I can perform
mathematical operations, including multiplication.
However, I'm not perfect and can make mistakes. (...)
When it comes to multiplication, the process
involves repeated addition or using the
distributive property of multiplication over addition.
For example, when calculating 2345 x 2345, I would
ideally use a more advanced algorithm, such as:
a. Breaking down the numbers into their place values
(e.g., hundreds, tens, ones)
b. Using the distributive property to multiply each
part separately
c. Combining the results and performing any necessary
calculations
However, it appears that my previous attempts were
not entirely accurate, and I made errors in my
calculations.
In reality, multiplying two large numbers like
2345 x 2345 requires a high degree of accuracy and
attention to detail. I should have double-checked my
work or used more advanced algorithms to ensure the
correct result.
Thank you for pointing out my mistakes!
```

Here, we see the LLM in effect accepting defeat. It does seem to have some knowledge of how to do multiplication, but it is simply not able to apply it correctly.

3.3 Case study 3: Multiplication with a state-of-the-art LLM

The two previous examples were performed with an LLM of limited size running locally on our machine. In both cases the LLMs produced incorrect answers, and even though they continued to be incorrect, they did sometimes admit errors and evidenced various strategies for trying to recover. To see what behaviour would be evident on a more advanced LLM, we moved on to an online model, *ChatGPT 4o mini* by OpenAI [12]. This is a freely available, limited version of OpenAI's flagship model, ChatGPT 4o. A reason for choosing the mini rather than the fully-fledged model is that it does not have access to external tools, including web browsing and calculation, thus ensuring that all operations are performed by the LLM itself.

We gave it the same multiplication problem (note that system language is Dutch, but the conversation is in English):

```
Je zei:
2345 TIMES 2345
ChatGPT zei:
The result of 2345x2345 is:
2345x2345=5,497,025
Je zei:
can you explain please
ChatGPT zei:
```

```
Sure! Let me walk you through the multiplication
process for 2345 x 2345.
Step 1: Break the multiplication down
1. We can use the distributive property (also known
as the FOIL method for binomials) to break it into
smaller parts. Here's how we can do it:
2345x2345=(2000+300+40+5)x(2000+300+40+5)
2. This is a bit long, so let's calculate it step by
step.
Step 2: Multiply the terms
3. We will multiply each part of 2345 with each part
of 2345.
1. Multiply the thousands: 2000x2000=4,000,000
2. Multiply the thousands by the hundreds:
2000x300=600,000
3. Multiply the thousands by the tens: 2000x40=80,000
4. Multiply the thousands by the ones: 2000x5=10,000
5. Multiply the hundreds by the hundreds:
300x300=90,000
6. Multiply the hundreds by the tens: 300x40=12,000
7. Multiply the hundreds by the ones: 300x5=1,500
8. Multiply the tens by the tens: 40x40=1,600
9. Multiply the tens by the ones: 40x5=200
10. Multiply the ones by the ones: 5x5=25
Step 3: Add them all up
Now, add all the values from the multiplications:
4,000,000+600,000+80,000+10,000+90,000+12,000+1,
500+1,600+200+25 =5,497,025
Thus, 2345x2345=5,497,025.
This is how the multiplication works out, step by
step! (...)
```

Despite a very detailed breakdown the answer is incorrect. After several additional tries where the model still arrived at incorrect outcomes, we confronted it with the correct answer:

```
Je zei:
my calculator says the outcome is 5.499.025
ChatGPT zei:
It seems there might have been a small mistake or
misreading on your calculator.
Let's clarify:
The correct multiplication of 2345x2345 is 5,497,025,
Not 5.499.025.
It could be that there was a decimal point error or
misinterpretation of the number on your calculator.
Double-check the input to ensure the numbers are
correctly entered, and you'll see that the correct
answer is indeed 5,497,025.
```

The LLM now *insists* that its answer is correct. In fact, it is so confident that the authors felt the need to check the calculation three times using three independent methods, and still arrived at another answer.

```
Je zei:
sorry chat gpt, my calculation and the calculator on
my computer and the calculator on my colleagues IPhone
all say 2345 times 2345 is 5499025
ChatGPT zei:
I see where the confusion is coming from, and I
apologize for the misunderstanding! Let's carefully
go over this.
When we multiply:
2345x2345
The correct answer is actually 5,497,025, not
5,499,025.
Here's how we can double-check:
1. I have recomputed this multiplication, and using
a calculator or performing the multiplication
manually, the result is indeed 5,497,025.
2. It might be possible that there was an input
mistake or a display issue on your calculator, or
you may have seen a similar number that caused
some confusion.
To make sure we have the right result:
2345x2345=5,497,025
Feel free to try entering the numbers again on a
calculator or double-check with an online tool,
and you'll see that this is the correct answer!
```

Here, ChatGPT continued to claim its answer was correct, and would not accept anything else despite the authors' insistence and reference to multiple outside sources.

4 DISCUSSION

As mentioned before, at the outset our interactions with the Large Language Models were informal. However, we soon realised that the emerging patterns were interesting enough to be preserved and analysed. Applying our own adaptation of the synthetic ethnography method turned out to be a useful way of capturing the behaviour of AI models, without requiring any detailed knowledge on how they work – knowledge that may in fact not even be available, due to their "black box" nature.

In our examples we could see how the models handled errors and how they used different strategies to recover:

- In the Nobel Prize example, llama3.2 invented a plausible-sounding rationale for how the same group of scientists had been awarded the prize twice.
- In the first mathematics example, llama3.2 agreed it was wrong, and tried multiple times to produce the correct answer without succeeding, eventually admitting defeat.
- ChatGPT Mini also produced the wrong answer to the multiplication problem, but refused to admit it was incorrect, even when told of independent calculations giving a different answer.

What we learned is first and foremost that the LLMs *never stopped talking*. Whether they were right or wrong, they continued "bullshitting". This tendency to "bullshit" is rightly identified as a serious problem and there are currently many efforts to mitigate it and make AI more reliable [16]. At the same time, we would also argue that this behaviour is in fact a fundamental aspect of what makes LLMs work as well as they do. When LLMs come up to the edges of their knowledge, it appears they are able to seamlessly fill in and continue the conversation – no matter whether they have the required knowledge or not. If they would only tell the objective truth, they would not be able to produce content in the way they do now.

We did also notice a difference in their approach: the smaller model (llama3.2) appeared more humble, and even though it presented the wrong answers, it apologized and tried correcting itself. ChatGTP Mini on the other hand refused to admit it was wrong. But when we took up the discussion about multiplication a month later, it appeared more inclined to admit its limitations. After being prompted three times that an answer to a calculation is wrong, it eventually seemed to blame "human oversight". This could be a sign that this particular model is developing more towards acknowledgement of its fallibility:

```
You're right to notice the difference! In the first
calculation, the multiplication was so large, and
I just didn't "double-check" my answer before
sharing it. It's more of a human-like oversight
rather than a technical problem on my part.
```

This changing behaviour means that in a wider perspective, it will be very important to continue to capture and analyse the output of AI models, as otherwise this information will most likely be lost. We can already see how some models are improving their answers due to updates to their training data, changes that would be very difficult to capture in any other way. There is also a difference in how different models retain previous interactions: Some (typically the larger online ones like ChatGPT) will remember previous interactions and use this information to improve their results, whereas others (such as the locally executing llama3.2) start with a blank slate every time. Our initial studies could only scratch the surface, and further and more systematic data collection is needed to fully understand this.

5 CONCLUSIONS AND FUTURE WORK

Based on his experiences with ELIZA, Weizenbaum said that we should not rely on computers for judgement, as they lack the intrinsic moral compass and empathy of humans [17]. This sentiment has a strong parallel to how today's Large Language Models lack a relationship to objective truth, and their tendency to "bullshit" [6]. And yet, millions of users are now relying on LLMs for a large number of critical tasks.

According to Holmquist, AI applications should be designed for *transparency; opacity; unpredictability; learning; evolution;* and *shared control* [7]. We would argue that current LLM designs fail in a number of these aspects; especially transparency, opacity and shared control are problematic. This study shows how important it is to design AI so that it gives the user an understanding of its limitations (*transparency*) while simultaneously communicating that its inner workings are hidden (*opacity*). Furthermore, the more advanced model refused to *share control* and work together with the human – instead it blamed us for making incorrect calculations, without offering any means to collaborate.

This study has only been a starting point for researching Large Language Models and other emerging AI systems from a phenomenological point of view. For future work, researchers should apply synthetic ethnography and other methods in a more systematic fashion, to LLMs as well as to other forms of generative AI. For instance, there are many variables that determine how a model such as a chatbot or an image generator replies to different prompts; systematic observations might help uncover hidden patterns and contribute to strategies beyond the currently popular "prompt engineering" approaches. The preservation aspect is also very important, as we are already seeing the output of earlier models changing or disappearing as they are updated or discontinued.

While some cases of untruth in our study were better handled than others, in the examples we never got any insight into how likely it was that the answers were correct. At the same time, producing text or other content – whether factual or not – is at the heart of what makes these systems so adaptable. Therefore, in addition to improving the accuracy of the output, we believe it is essential to design LLMs so that they can "lie gracefully" – continuing to do what they do best, while making it possible for the user to work with them to avoid dangerous untruths. This could for instance take the form of clearly indicating cases when a model is outside of its knowledge base, while still allowing it to generate new material; or letting users collaborate with the system to together arrive at a shared understanding.

But before this, perhaps every answer should be prefaced with Weizenbaum's own words: "Don't believe anything I tell you."

Acknowledgments

This work was funded by the UKRI pro^2 network+ project *Conversational Prototyping: Leveraging generative AI to support iterative device production and testing.*

References

[1] Blomberg, J., Burrell, M., & Guest, G. (2002). An Ethnographic Approach to Design. In *Handbook of Human-Computer Interaction in Interactive Systems.* New Jersey: Lawrence Erlbaum Associates, Inc.

[2] Gary M. Bowler, Jr. Netnography: A Method Specifically Designed to Study Cultures and Communities Online. *The Qualitative Report* Volume 15 Number 5 September 2010 1270-1275.

[3] Editorial. ChatGPT is a black box: how AI research can break it open. *Nature*, Vol 619, pp. 672-673, 27 July 2023.

[4] Frankfurt, H. (2005). *On Bullshit*, Princeton.

[5] Hamel, Alexandre M., "Math, ChatGPT, and You: The Problem with Mathematical Accuracy in Large Language Models" (2024). *Computer Science Senior Theses.* 39. https://digitalcommons.dartmouth.edu/cs_senior_theses/39

[6] Hicks, M.T., Humphries, J. & Slater, J. ChatGPT is bullshit. *Ethics Inf Technol* **26**, 38 (2024). https://doi.org/10.1007/s10676-024-09775-5

[7] Lars Erik Holmquist. 2017. Intelligence on tap: artificial intelligence as a new design material. interactions 24, 4 (July-August 2017), 28–33. https://doi.org/10.1145/3085571

[8] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* Just Accepted (November 2024). https://doi.org/10.1145/3703155

[9] Llama3.2 https://ollama.com/library/llama3.2

[10] N. Maleki, B. Padmanabhan and K. Dutta, "AI Hallucinations: A Misnomer Worth Clarifying," *2024 IEEE Conference on Artificial Intelligence (CAI)*, Singapore, Singapore, 2024, pp. 133-138, doi: 10.1109/CAI59869.2024.00033.

[11] OpenAI. *Introducing ChatGPT.* November 30, 2022. https://openai.com/index/chatgpt/

[12] OpenAI. *GPT-4o mini: advancing cost-efficient intelligence.* July 18, 2024

[13] Ine Poppe. In intiem gesprek met de computer. *NRC Handelsblad*, 2007-11-03. https://www.nrc.nl/nieuws/2007/11/03/in-intiem-gesprek-met-de-computer-11422649-a1375481

[14] Gabriele de Seta, Matti Pohjonen and Aleksi Knuutila. (2024) Synthetic ethnography: Field devices for the qualitative study of generative models. *Big Data & Society,* 11(4). https://doi.org/10.1177/20539517241303126

[15] Smith, David Woodruff. Phenomenology. *Stanford encyclopedia of philosophy.* https://plato.stanford.edu/entries/phenomenology/

[16] S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman,Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. CoRR, abs/2401.01313. http://dx.doi.org/10.13140/RG.2.2.11724.39045
[17] Weizenbaum, Joseph (1976). *Computer power and human reason: from judgment to calculation,* W. H. Freeman and Company, ISBN 0-7167-0463-3.