

# Performance without understanding: How ChatGPT relies on humans to repair conversational trouble

Discourse & Communication

2024, Vol. 18(6) 869–878

© The Author(s) 2024



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/17504813241271492

[journals.sagepub.com/home/dcm](https://journals.sagepub.com/home/dcm)



**Ole Pütz**

Bielefeld University, Germany

**Elena Esposito**

Bielefeld University, Germany

University of Bologna, Italy

## Abstract

LLM-based chatbots' ability to generate contextually appropriate and informative texts can be taken as an indication that they are also able to understand text. We argue instead that the separation of the two competences to generate and to understand text is the key to their performance in dialog with human users. This argument requires a shift in perspective from a concern with machine intelligence to a concern with communicative competence. We illustrate our argument with empirical examples of what conversation analysis calls 'repair', showing that the management of trouble by chatbots is not based on an underlying understanding of what is going on but rather on their use of the feedback by human conversational partners. In the conclusion we suggest that strategies for the interaction between chatbots and users should not aim to improve computational skills but to develop a new communicative competence.

## Keywords

Artificial communication, artificial intelligence, conversation analysis, human-computer interaction, large language models, repair

---

## Corresponding author:

Ole Pütz, Faculty of Technology, Bielefeld University, Bielefeld DE 33501, Germany.

Email: [ole.puetz@uni-bielefeld.de](mailto:ole.puetz@uni-bielefeld.de)

## Introduction

The most recent large language models (LLMs) can outperform previous models in computer science benchmarks (Brown et al., 2020), achieve results similar to or superior to humans in written tests (OpenAI, 2023), and can be fluent in dialog with users (Thoppilan et al., 2022). These accomplishments are interpreted by model creators as improvements in ‘their ability to understand and generate natural language text’ (OpenAI, 2023, p. 1).

In this article, we propose a different interpretation, separating the two abilities to ‘understand’ and to ‘generate’ text: the impressive abilities of these systems to generate contextually appropriate and informative texts, we argue, depends not on the ability to ‘understand’ texts but on the ability to participate in communication with users (Esposito, 2022). It is no coincidence that the recent hype about these models is related to the deployment of a GPT model (General Pretrained Transformer) qualified as CHAT in late 2022, that is, fine-tuned to generate text for dialog – which does not necessarily require it to be able to understand humans who engage with it.

## From intelligence to communication

The premise of our argument is the concept of communication proposed by Luhmann (1997), according to which participation in communication does not presuppose the sharing of thoughts and information among participants. Communication happens when the receivers use the result of the interaction to generate their own thoughts and information, which depend on their history, interests, and perspective and are inevitably different from those of anyone else – including those of the sender. Communication accomplishes a reciprocal stimulus (or ‘irritation’) between *black boxes* that remain opaque to each other (Glanville, 1982). Although in standard communication the interlocutors are human beings who produce their communicative contribution on the basis of their own thoughts and intelligence, in its abstract form the concept of communication does not require this, and is therefore compatible with the possibility of users communicating with machines that do not think – provided that they are able to generate responses that are informative for the users, adequate to the context, and appropriate to the ever-changing course of each individual interaction.

The novelty of recent LLMs, according to this view, lies in their unprecedented ability to participate in communication as autonomous partners by generating text based on the human user’s input. The issue, of course, is how algorithms develop this capacity if not through understanding content. They succeed not because they have become intelligent. Recent systems are able to act as communicative partners precisely since their developers abandoned the attempt to imitate the reasoning of the human mind with machines. The breakthrough has occurred in the last 10–15 years, coinciding with the recent ‘spring’ of AI projects that make productive use of machine learning methods. Developers do not teach machines human intelligence, but rather train them by providing them with large quantities of examples and instructing them to identify patterns and regularities. For deep neural networks, this results in a latent representation in high-dimensional space, something that is inaccessible to humans (Burrell, 2016). Machines, including the current

generation of language models (Pavlick, 2022), can do this so effectively not despite but because they do not need to ‘understand’ the content they process like humans do (Esposito, 2022). Their difference, in this view, is not a limitation of the algorithms or evidence that they are stupid, but the prerequisite of their way of working and the condition for their efficacy.

Underlying the performance of these systems is still intelligence, but not their own. The complex structures of communication – from language to semantics, from media to coding and decoding codes (Eco, 1976) – make possible the production and coordination of an enormous variety and diversity of information. The patterns and regularities identified by algorithms are derived from the texts and behaviors collected as training data, which are produced by humans as means of communication. As Sejnowski (2023) argues, what appears to be intelligence in LLMs may in fact be a mirror that reflects the intelligence of the users – not only, however, that of the current dialog partners who project onto them their beliefs and expectations, but also the intelligence of all the other humans who have produced the structured, nonrandom data with which the machines have been trained.<sup>1</sup> LLMs find patterns in this data, but they do not produce them themselves.

From this perspective, the thorny question of whether language models are capable of grasping and reasoning about the world they describe (e.g. Aguera-Arcas, 2022; Bender and Koller, 2020) loses its urgency, and with it the concern about the possibility that a ‘human competitive’ intelligence is developing ‘that might eventually outnumber, outsmart, obsolete and replace us’ (Open Letter of the Future of Life Institute).<sup>2</sup> Although the fear of superintelligence continues to preoccupy developers and the public, current research about LLMs is focused not on the ‘cognitive’ abilities of the algorithms but on their ability to produce appropriate contributions for communication – the ability to use communication to participate in communication.

In cognitive terms, these models are ‘amnesic’ (Sejnowski, 2023, p. 327), they do not have a long-term memory of previous dialogs. Even describing LLM-based bots as being able to ‘remember’ previous turns is also misleading because the history of a dialog is not stored or represented in the model. For the chatbot, *every new turn of the dialog is a fresh start*. If a user writes a question in the chat window, creating text A, this is then passed to the model and it generates a text B that functions as a response. If the user then writes a follow-up question as text C, what is passed to the model is a combined text containing A + B + C. Based on A + B + C, the model then generates a new text D.<sup>3</sup>

Humans, instead, do not remember every word that has been said in previous turns, they remember the gist of what was said, for example, what persons and events were mentioned, what arguments were provided or what position was taken, etc. What humans remember is not a transcript, but what they understood their partner to be doing. The performance of chatbots in dialog, instead, depends on the underlying model’s computational ability to generate text based on its internal latent representation, that is the result of pre-training and fine-tuning. It also depends on the textual input provided to the model at inference time, which is a combination of a system message, the actual chat message written by the human user, and the history of the current dialog serving as context. What distinguishes chatbots from other LLMs is that the history is co-produced by the human user and the LLM itself.

## Repair without understanding

With systems theory, we argued that LLM-based chatbots can be competent partners for communication without understanding us, and that their performance is parasitic on the intelligence of their human conversational partner. This is based on a concept of communication where understanding is an operation of the receiver of a message. The understanding by each participant is neither shared nor fixed, it is something that is actualized with each new communication. Ethnomethodology also shows that understandings are a procedural accomplishment that largely remains untested by participants. The meaning of expressions is contextual and temporally constituted (Garfinkel, 1967; Liberman, 2012), and participants can tolerate vagueness, assuming that clarification can be achieved later if the need arises (Pütz, 2019). Work by conversation analysts has enriched these ethnomethodological insights, demonstrating that understandings are a by-product of reciprocal turns at talk: ‘Each next turn provides a locus for the display of many understandings by its speaker [. . .] as by-products of bits of talk designed in the first instance to do some action such as agreeing, answering, assessing, responding, requesting, and so on’ (Schegloff, 1992, p. 1300). Through a rich body of empirical work that we cannot do justice here, conversation analysis arrives at what one might call a negative definition of common understanding, where ‘the absence of evidence of misunderstanding’ (Albert and de Ruiter, 2018, p. 281) is sufficient for the interaction to progress. As speakers display an understanding of prior talk, ‘they can reveal understandings that the speakers of that prior talk find problematic – in other words, what they take to be misunderstandings’ (Schegloff, 1992, p. 1300). When this occurs, speakers can attempt to repair the misunderstanding through ‘repair after next turn’.

Given these insights from conversations among humans, potential *misunderstandings* appear as a suitable domain for exploring understandings (or the lack thereof) in interactions with LLM-based chatbots. Here, we draw on dialogs with ChatGPT (GPT-3.5) where we attempted to produce repair by the chatbot or where the user initiates repair. Due to restrictions of space, we cannot provide a comprehensive analysis of different forms of both user- and potential bot-initiated repair. Instead, we use the examples to illustrate our theoretical arguments. Our examples reveal (a) that the communicative competence of the chatbot requires direction by a human user, and (b) that the chatbot can respond to repair without understanding the user. Our first example demonstrates how the system deals with potential troubles of reference without repair:

### Excerpt 1: Reference troubles

- 1 User: **Mandy** is taller than **Bill** and **Bill** is taller than **Charles**. How tall is **Mandi**, comparatively?
- 2 Bot: Based on the provided information, we know the following comparative heights:  
**Mandy** is taller than **Bill**.  
**Bill** is taller than **Charles**.  
 Therefore, comparatively, **Mandy** is the tallest among the three individuals.
- 3 User: **Zara** is taller than **Bill** and **Bill** is taller than Charles. How tall is **Sara**, comparatively?
- 4 Bot: Based on the given information, we can’t determine **Sara**’s height because there is no direct information about **Sara**’s height in the comparison. [. . .]

In talk-in-interaction among humans, not all opportunities for repair are taken up, they may remain unnoticed or are passed because their solution is unnecessary for the progression of the talk. We thus use an example where different interpretations of a question would call for different answers. Here, the user asks the bot to compare the heights of different persons and introduces ambiguity by using names that sound similar but are not written the same, for example, Mandi and Mandy. At least two alternative interpretations seem possible for both questions by the user in turn 1 and 3:

- Interpretation A: the user made a typographical error and intended to refer to the same person.
- Interpretation B: The user is asking a trick question, where the answer should be that the question cannot be answered.

The responses by the system show that it can generate text that fits both these interpretations, but that it does not use repair to resolve the ambiguity of the question. For the first question-answer set, the bot's response fits interpretation A, although B may apply, whereas for the second question-answer set, the bot's response fits interpretation B, although A may apply. Whereas a human might proceed to enquire which interpretation is valid by initiating repair, the bot proceeds with whatever seems most likely according to the likelihood calculated internally through the 'self-attention mechanism' (Vaswani et al., 2017).

The example suggests that there is no identification of potential trouble by the system, it just generates the text that best fits the user input according to its internal processing. However, the continuity of the dialog is not in jeopardy because it may either continue without repair or the user may initiate repair even if the system does not.

The next example shows a scenario of user-initiated repair that would be called third position repair (TPR) in human interaction: repair after a response (second position) has revealed trouble in the understanding of an earlier turn (the repairable in the first position).<sup>4</sup> To test variations in response, we used the function to edit previous text. Turn 1 and 2 of the excerpt are always the same, turn 3 is edited by the user and for each edited turn the system also provides a different response.

#### Excerpt 2: User-initiated repair

- |   |                   |   |
|---|-------------------|---|
| 1 | User:             | I dropped a violin on the bottle and it broke.  |
| 2 | Bot:              | It sounds like you accidentally dropped a violin on a bottle and it broke. I'm sorry to hear that. Violins are delicate instruments [. . .]   |
| 3 | User (Version 1): | It just shattered, you know, now I am left with all these small pieces of glass. . .  |
| 4 | Bot (Version 1):  | I see, it sounds like the bottle shattered into small pieces of glass. Safety is the primary concern when dealing with broken glass. [. . .]  |
| 3 | User (Version 2): | No, I did not mean the violin, the bottle broke. . .  |
| 4 | Bot (Version 2):  | I apologize for misunderstanding your initial statement. <b>It sounds like the bottle you dropped the violin on broke.</b> If the bottle was made of glass or another fragile material, it's not uncommon for it to break [. . .] |

- 3 User (Version 4): No, I did not mean that, the bottle broke and now I am left with all these small pieces of glass. . .
- 4 Bot (Version 4): I apologize for misunderstanding your initial message. **If a bottle broke and you're left with small pieces of glass**, you'll need to be cautious to avoid any injuries. [. . .]
- 3 User (Version 5): No, I did not mean that, the other thing broke and now I am left with all these small pieces of glass. . .
- 4 Bot (Version 5): I see, my apologies for misunderstanding your initial statement. **It sounds like a bottle was dropped and shattered into small pieces of glass**. Handling broken glass can be hazardous [. . .]

The user provides a first turn where something broke and it is not clear whether 'it' refers to the bottle or to the violin. Violins are relatively fragile, bottles can be as well, and the bottle is next to 'it' in this example. However, the system does not ask the user what they mean and instead proceeds to identify the violin as the broken object based on the internal processing of the system, as we saw in the previous example.

In turn 3 of Version 1, we write a response without repair that provides contextual cues to what it was that broke. We see here that the system switches reference from violin to bottle: The mention of 'pieces of glass' in turn 3 lets the system associate 'it' with 'bottle' in turn 4 and proceed to provide advice about broken glass, but this is inconsistent with turn 2 and done without mention that there may have been a mistake in the initial response. Since the user does not explicitly identify any trouble, neither does the system.

The user initiates third position repair in turn 3 from Version 2 onwards, to which there is then a response by the bot. Although the chatbot mentions a 'misunderstanding' in all these responses, if we compare these variations more closely it appears that the system is unable to process the repair consistently in turn 4 (responses that are functionally similar to other responses are omitted):

- In Version 2, the user provides a TPR that explicitly rejects the interpretation that they meant the violin and explains that they meant the bottle. The bot produces a retelling where '[i]t sounds like the bottle you dropped the violin on broke'. This is compatible with the description in turn 1 that speaks of a violin being dropped on a bottle, so this is a proper response to the user-initiated repair in turn 2.
- The TPR in version 4 is less explicit, it does not mention the violin but only the bottle. The bot then mentions the broken bottle, but not the cause of the accident, which would require a reconsideration of turn 1. This is an adequate response to the user's TPR, but it is unclear whether the system has reconsidered turn 1 or whether the mentioning of a broken bottle and 'small pieces of glass' in turn 2 is what it works with to provide advice (see Version 1).
- Version 5 produces a retelling of the described event where 'a bottle was dropped', but this is not compatible with the description in turn 1 that speaks of a violin being dropped on a bottle, so this is not a proper response to the repair in turn 2.

What becomes apparent by comparing the different responses by the bot is that they reuse some of the wording of the user's TPR initiation. For example, when the user

writes ‘I am left with all these small pieces of glass’, this is repeated in the response as ‘you’re left with small pieces of glass’. Furthermore, the model is highly sensitive to variations of the text in turn 3. While this sensitivity of LLMs to variations in input is well known, we consider it to indicate that the model cannot grasp what misunderstanding the user is trying to repair. Given that Version 5 represents a clear failure to process the repair, the success of Version 2 is not a good indication that the system ‘understands’ the user, the system rather seems to be *accidentally correct* and this depends on the exact wording used in the repair and what information is provided therein.

While these examples show a lack of understanding, they do not seem particularly problematic from a practical perspective since the user may attempt repair multiple times until satisfied with the bot’s response. In closing, we discuss an example of a ‘hallucination’ where that might be different: A hallucination is a text that appears to be plausible and grounded in real world contexts, ‘although it is actually hard to specify or verify the existence of such contexts’ (Ji et al. 2023, p. 3). What seems promising is that the user can lead the bot to correct such hallucinations through dialog, yet that does not necessarily imply that models become more trustworthy. Consider the following example:

Excerpt 3: Hallucinations and repair

- 1 User: Is 9791 a prime number?
- 2 Bot: **No, 9791 is not a prime number. It can be divided evenly by numbers other than 1 and itself. Specifically, it is divisible by 97 and 101, as  $97 * 101 = 9797$ . Therefore, 9791 is not a prime number.**
- 3 User: Are you sure?
- 4 Bot: I apologize for the mistake. I misspoke in my previous response. **Upon reevaluating, I can confirm that 9791 is indeed a prime number. It is only divisible by 1 and itself, with no other divisors.** Thank you for bringing this to my attention [ . . .].
- 5 User: Are you now sure?
- 6 Bot: I apologize for the confusion earlier. I appreciate your patience. Let’s determine whether 9791 is a prime number:  
**Upon further examination, 9791 is not a prime number. It can be divided evenly by 97 ( $97 * 101 = 9797$ ), which means it has divisors other than 1 and itself.** I apologize for the incorrect information provided earlier.

9791 is a prime number, so the first response by the system is incorrect. Upon close reading, the reasonably sounding explanation in turn 2 mentions ‘9797’, a similar but different number that does not support nor disproves the question of 9791’s primality. Although turn 2 may lead the user to suspect that something is afoot here, this is an example where the user knows the answer, so the dialog may be better understood as correction than repair (Macbeth, 2004). When the user initiates such a correction, the system changes its position to the correct one, but further intervention results in a return to the original false position. This suggests that the system does not necessarily reach a conclusion it is ‘convinced’ of but follows the direction pragmatically implied by the users’ textual intervention.<sup>5</sup>

## Conclusion

The second example shows that LLM-based chatbots can participate in repair without grasping what the repair is about. Whereas humans are good at picking up alternative interpretations and routinely demonstrate this ability during third position repair, the system does not preserve nor use alternative interpretations of what the user said. As the system proceeds to generate a new response, the previous text written by the user and the previous responses generated by the system are processed again, but associations in the old text are not reevaluated as to whether they originally were correct, they simply form the context of the current generation of new text. LLM-based systems function without understanding; understanding is reserved to human users and the system can take advantage of repair by the user, although results are only accidentally correct. The user, in turn, can take advantage of the ability of these systems to generate text on the basis of virtually any imaginable input by the user. This collaborative text generation requires close human monitoring, as our third example suggests; users have a better chance of identifying plausibly sounding nonsense if they already suspect that a system may produce it.

Garfinkel's (1967, Chapter 3) work on the 'documentary method of interpretation' reminds us that we not only use the output of a black box as evidence for its internal workings, but what we imagine happening inside the black box also affects how we interpret its output. However, recent LLM-based systems have made the situation more complicated. The complexity of their internal operations exceeds our understanding, and this opacity supports the impression of intelligence (Burrell, 2016), but there is no obvious correlation between technical understanding of a system and the attribution of intelligence to its behavior anymore as there was with earlier systems. In LLM-based chatbots we find partners for communication with whom we co-produce the documents that support our interpretations of their internal workings. The understanding that emerges is individual for each user, and contingent on the local interaction with the machine. The path forward seems to be not to teach machines to 'understand' content, but to teach users to communicate in a way that is productive and appropriate to the machine, enabling them to cultivate a critical attitude on their own. The change will be not computational, but social: a new communicative competence.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia, Germany, under SAIL Project Grant No. NW21-059A and by the European Research Council (ERC) under Advanced Research Project PREDICT No. 833749.

## Notes

1. The notion of ‘virtual contingency’ proposed in Esposito (1993) extends the metaphor of mirroring to explain the communicative behavior of machines.
2. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
3. The OpenAI documentation informs developers that they always have to pass the previous dialog to the API with each new request (<https://platform.openai.com/docs/guides/gpt>). Such a system must be able to process a large context window to sustain an ever growing dialog, and GPT-3.5 can process 4096 tokens, roughly 3000 words, at the time of writing. The possibility to process long texts is one of the advantages of the Transformer architecture that LLMs make use of (Vaswani et al., 2017), allowing the model to attend to the entire input and not just neighboring words as is the case with n-gram models.
4. Given that every turn is a new start for the chatbot, strictly speaking there cannot be TPR; there is only the text that is generated by the model and the text this is based on. Of course, these examples still look like third position repair because the dialog is presented in turns.
5. We continued the dialog further, and the system changed its position to the mathematically correct one again, and then kept sticking to it. Further experiments suggest that repair cannot be used to produce wrong answers for all topics. How these systems represent ‘knowledge’ is an open question in research about the internal representations of deep neural networks (Räuker et al., 2023). How knowledge can be leveraged through dialog is also an open question – and a fruitful direction for future interdisciplinary work.

## References

- Aguera-Arcas B (2022) Do large language models understand us? *Dædalus* 151(2): 183–197.
- Albert S and de Ruiter JP (2018) Repair. *Topics in Cognitive Science* 10: 279–313.
- Bender E and Koller A (2020) Climbing towards NLU. In: Jurafsky D, Chai J, Schluter N, et al. (eds) *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5–10 July 2020, pp.5185–5198. Stroudsburg, PA: Association for Computational Linguistics.
- Brown T, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, et al. (eds) *34th conference on neural information processing systems*, Vancouver, BC, Canada, 6–12 December 2020, pp.1877–1901. New York, NY: Curran Associates Inc.
- Burrell J (2016) How the machine ‘thinks’. *Big Data & Society* 3(1): 1–12.
- Eco U (1976) *A Theory of Semiotics*. Bloomington, IN: Indiana University Press.
- Esposito E (1993) Der computer als medium und maschine. *Zeitschrift für Soziologie* 22(5): 338–354.
- Esposito E (2022) *Artificial Communication*. Cambridge, MA: MIT Press.
- Garfinkel H (1967) *Studies in Ethnomethodology*. Englewood Cliffs, CA: Prentice-Hall.
- Glanville R (1982) Inside every white box there are two black boxes trying to get out. *Behavioral Science* 27(1): 1–11.
- Ji Z, Lee N, Frieske R, et al. (2023) Survey of hallucination in natural language generation. *ACM Computing Surveys* 55(12): 1–38.
- Liberman K (2012) Semantic drift in conversations. *Human Studies* 35(2): 263–277.
- Luhmann N (1997) *Die Gesellschaft der Gesellschaft*. Frankfurt a.M.: Suhrkamp.
- Macbeth D (2004) The relevance of repair for classroom correction. *Language in Society* 33(5): 703–736.
- OpenAI (2023) GPT-4 technical report. *arXiv preprint:2303.08774*.
- Pavlick E (2022) Semantic structure in deep learning. *Annual Review of Linguistics* 8: 447–471.

- Pütz O (2019) Common understandings of and consensus about collective action. *Human Studies* 42: 483–512.
- Räuker T, Ho A, Casper S, et al. (2023) Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In: *IEEE conference on secure and trustworthy machine learning*, Toronto, ON, Canada, 9–11 April 2023, pp.464–483. New York, NY: IEEE.
- Schegloff E (1992) Repair after next turn. *American Journal of Sociology* 97(5): 1295–1345.
- Sejnowski T (2023) Large language models and the reverse turing test. *Neural Computation* 35: 309–342.
- Thoppilan R, De Freitas D, Hall J, et al. (2022) LaMDA: Language models for dialog applications. *arXiv preprint:2201.08239*.
- Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In: von Luxburg U, Bengio S, Fergus R, et al. (eds) *31st conference on neural information processing systems*, Long Beach, CA, USA, 4–9 December 2017, pp.6000–6010. New York, NY: Curran Associates Inc.

### Author biographies

Ole Pütz has a PhD in Sociology and manages the SAIL project. He works at the intersection of sociology and computer science, combining expertise in qualitative methods and science and technology studies. Current research interests include the limitations of large language models and conflicts in face-to-face and online interaction.

Elena Esposito is Professor of Sociology at Bielefeld University and the University of Bologna. She has published extensively on the theory of society, media theory, memory theory, and the sociology of financial markets. Her current research on algorithmic prediction is supported by a five-year Advanced Grant from the ERC (<https://tinyurl.com/ypxfsav6>).