How do we understand machines that talk to us?
Communication with Large Language Models
Project leader: Ingrid Lossius Falkum (IFIKK/ILN)

Introduction

In the near future, most of us will likely find ourselves spending a considerable proportion of our working life and free time interacting with Large Language Models (LLMs). A great deal of scholarly work is currently being devoted to the abilities of LLM (e.g., investigating to what extent LLMs may change grant writing,[1] or reflect the cognitive organisation of human grammar; Frank 2023). We propose to study a pressing question, which has received much less attention: **how do people interact with LLMs?** In particular, do we interpret the 'utterances' of LLMs in the same way as we understand each other?

Large Language Models (LLMs) and their immediate precursors have been under development for several years but burst into public consciousness in late 2022 with the public release of OpenAI's ChatGPT. LLMs are neural networks trained with very large amounts of text so that they learn a probability distribution over sequences of words. This allows them to generate text, given a prompt, by choosing a highly likely successor word at each step (Bender et al. 2021; Wolfram 2023). The interface for ChatGPT and similar LLMs is conversation-like. Users give instructions by typing in text in English or other natural languages. The LLM responds with text in natural language which is (in general) grammatical, relevant and coherent.

Much work in human communication takes Gricean inferentialism (Grice 1957, 1967) as foundational to communication. On this view, the interpretation of verbal utterances (whether spoken or written) is an inferential process for working out the speaker's communicative intention – what the speaker aimed to convey (e.g. Bach & Harnish 1979; Sperber & Wilson 1986; Korta & Perry 2006; Tomasello 2008). To do this we combine information about the context and background knowledge with the sentence structure derived by syntactic parsing (Allott 2023). But this implies a puzzle: How can we engage in what appear to be conversational exchanges with LLMs despite the fact that LLMs lack communicative intentions?

Research background and aims

Unlike classic AI systems which were underpinned by a finite set of explicit rules, it is not clear what internal representations LLMs have. The question is currently being explored by linguists and philosophers (Mallory forthcoming; draft). Recent research indicates that, regardless of remarkable superficial similarities (noted by Piantadosi 2023), LLMs' representation of language is different from human linguistic knowledge (Katzir 2023). Whether they are correctly characterised as "performing natural language understanding" (Bender et al. 2021) is the subject of intense controversy and ongoing research: Bender and colleagues urge scepticism, but others take the 'pragmatic abilities' of LLMs as an empirical question (e.g. Barattieri di San Pietro et al. 2023). However, we are not aware of any serious scholarly claim that LLMs possess intentions (or at least are perceived by us humans as such), or whether to communicate or to do anything at all. So what happens when we humans understand some text or speech produced by an LLM? Do we represent that LLM as having a communicative intention? These questions are of vital importance: various professional and education settings rely on LLMs, and the way such interactions are processed is central to their impact on our cognitive environments.

Joseph Weisenbaum, creator of the rudimentary early chatbot ELIZA, noted that users of the system felt that it understood them, and that they attributed background knowledge and rationality to it which "manifest themselves inferentially in the interpretations [they] make[] of the offered responses." (Weisenbaum 1967: 474). The eponymous 'ELIZA effect' is thus "the susceptibility of people to read far more understanding than is warranted into strings of symbols – especially words – strung together by computers" (Hofstadter 1995: 157; cf. Reeves & Nass 1996; Dennett 1987; Marchesi et al. 2019).

There is striking anecdotal evidence that some people take the much more human-like utterances of LLMs as indicating that they share key properties with us. In 2022, Blake Lemoine, a software engineer at Google working on an LLM called LaMDA, claimed that it was sentient and that he "consider[ed it] to be his 'colleague'

[1] https://www.nature.com/articles/d41586-023-03238-5

and a 'person,' even if not a human." (De Cosmo 2022). A very recent development is voice mode, which takes spoken prompts as input and reads aloud the text generated by the LLM (OpenAI, 2023). Lilian Weng, who works on AI safety at OpenAI, tweeted:

> "Just had a quite emotional, personal conversation w/ ChatGPT in voice mode, talking about stress, work-life balance. Interestingly I felt heard & warm. Never tried therapy before but this is probably it? Try it especially if you usually just use it as a productivity tool."[2]

A related issue concerns the fact that LLMs are not designed to generate reliable information, given that their main task is to predict the next word in a sentence. One question, then, is to what extent knowing that the interactional partner is not human, but a chatbot, affects our general susceptibility to false information (Pantazi et al. 2018) and our ability to ward it off (Sperber et al. 2010)? And how does this "content unreliability" of LLMs affect vulnerable users like children and old people? These are some of the questions that this project aims to address.

Methods

The project will be divided into three interdependent subprojects, and will combine theoretical investigation and experimental studies. On the theoretical side the project will employ the closely-related methodologies of empirically informed philosophy of language, philosophy of mind, philosophy of AI and theoretical linguistics: conceptual analysis, identification and evaluation of assumptions, reflection on cases and examples, theory construction, and evaluation and revision of theories in response to evidence. On the experimental side, the project will make use of the techniques of experimental pragmatics and psycholinguistics: controlled pre-registered experiments which test participants' performance using on- and off-line measures including eye-tracking, timing measures, selections on Likert scales and among items such as written statements or pictures, with rigorous statistical testing of results.

Subproject 1: How do we attribute meanings to LLMs?

A key focus for linguistic pragmatics is inference about speaker meaning where that goes beyond what the speaker states and/or the linguistically encoded meaning of the sentence uttered, since in such cases successful communication clearly requires context-sensitive inference (Allott 2023). This includes utterances where part of what is meant is implicated or presupposed and cases of figurative speech such as metaphor, metonymy and irony.

After some 50 years of theoretical pragmatics and more than 20 of experimental pragmatics, a lot is known about how we interpret utterances, down to quite fine details about the time-course of processing and the developmental trajectories of these abilities from infancy to adulthood (e.g Breheny, Ferguson & Katsos 2013; Köder & Falkum 2020; Noveck 2018; Pouscoulous & Tomasello 2020; Ronderos & Falkum 2023). However, some fundamental theoretical questions are still open. In particular there is ongoing debate about whether inference and recovery of speaker's intentions are necessary for communication (Millikan 1984; Recanati 2002; Breheny 2006; Kissine and Klein 2013; Carston 2007, Allott 2008; Mazzarella 2014; 2015; see Allott 2023 for discussion). This subproject's theoretical work will consider how fundamental questions apply to interactions with LLMs: Is inferentialism about communication correct for our interactions with LLMs? Is a unified account possible of the interpretation of human and LLM utterances or are they understood in fundamentally different ways? In the experimental strand we will investigate whether we attribute figurative meanings and implicatures to LLMs, or interpret them more literally than we would human speakers, using a selection of eye-tracking and picture selection tasks (for an example of an experimental set-up, see Ronderos, Mathisen, Noveck & Falkum, under revision).

Subproject 2: How do we evaluate information produced by LLMs?

Recently, researchers in pragmatics, drawing on work in philosophy on testimony and lies (e.g. Fricker 2012; Stokke 2018), have looked at the impact on hearers' assessment of speaker trustworthiness of whether

[2] Retrieved from https://twitter.com/lilianweng/status/1706544602906530000 /

falsehoods are asserted, implicated or presupposed (Mazzarella et al 2018; Mazzarella & Pouscoulous 2020; Hall & Mazzarella 2023; Miller & Kissine 2023). In this subproject, theoretical work centres on the epistemology of testimony from LLMs: Is it ever rational to believe what an LLM (seems to) tell you? Are there any good arguments for modulating your belief depending on how the LLM (seems to) present information: assertion, implicature or presupposition? In the experimental strand of this subproject we investigate whether the degree to which we trust LLMs is impacted by the way in which (mis)information is communicated (e.g. whether asserted, implicated or presupposed), as it is for human speakers, adapting the tasks by Mazzarella et al. (2018), where they found that for human communication, implicating is taken to be less committal than asserting and presupposing. We will test both whether this holds for interactions with LLMs and whether utterances made by LLMs are taken as equally committal as utterances of human beings.

Subproject 3: How do children perceive their interactions with LLMs?

Contemporary theories of pragmatic development emphasise the foundational role that children's early pragmatic capacity plays in the emergence of pre-linguistic gestures and language (Clark, 2018; Csibra & Gergely, 2011; Sperber, 1994; Tomasello, 2003, 2008). In this subproject we will investigate how children apply this early, arguably quite sophisticated, pragmatic ability in their interactions with LLMs. Do they interact with and trust them similarly to human agents? One set of experimental tasks will use a word learning paradigm to investigate children's assessment of the reliability of human vs. digital speakers. This task takes as its starting point research showing that young children track speakers' mental states when learning new words and are sensitive to the prior accuracy and conventionality of speakers' labelling (Diesendruck, Carmel, & Markson, 2010; Koenig & Harris, 2005; Koenig & Woodward, 2010). Several studies have shown that children prefer to learn novel labels for unfamiliar objects from speakers who have a history of accuracy in their labelling of familiar objects over speakers who have proven themselves to be inaccurate. We propose to use an adaptation of this word learning paradigm with 3-7-year-old children (with whom this paradigm has been used successfully in the past) to investigate whether inaccurate digital speakers might be more severely penalised by children than human speakers. We will also develop follow-up experiments, taking our lead from the results we obtain on the word learning experiment.

Relevance and quality of the project

The project's central question is a pressing concern for society. People and organisations are struggling to come to terms with LLMs; far-reaching and revolutionary changes are widely foreseen (Dell'Acqua et al. 2023; Eloundou et al. 2023). Even at this early stage, it is clear that many of us will in the near future find ourselves spending a considerable proportion of our working life and free time interacting with LLMs through 'conversational' interfaces. Yet very little is known about how we interact conversationally with LLMs or even what it is to do so. Investigating LLMs will also open new perspectives on some of the most fundamental questions of philosophy and linguistics: What is content, what events or states possess it, and in virtue of what? Is speaker meaning a function of speaker intentions? Is communication inferential? When are we warranted in learning from utterances?

This project will break new ground by being the first systematic investigation of how we interact conversationally with LLMs. This requires expertise in linguistics, philosophy, and computer science since it raises questions that can only be addressed by combining the intellectual resources of linguistic pragmatics with philosophy of language and epistemology, as well as the technical knowledge of how LLMs actually work. It is also a question that calls out for both theoretical and empirical investigation. The project therefore leverages a key strength of HF by bringing together researchers within HF from across these domains and methods.

The project will conduct research of the highest international quality. The PI has experience running ERC and RCN research projects, and all the project participants are members of RCN projects and leading researchers in their fields. The project team combines expertise on theoretical and experimental pragmatics, philosophy of language, epistemology, and computer dialogue models.

The project team and the added value of the collaboration

The project team comprises researchers from three of HF's departments and partners from the Norwegian and international research sectors. The project leader, **Ingrid Lossius Falkum**, who has a joint position at ILN and IFIKK, is a leading expert on pragmatics and currently PI of two experimental projects (ERC and RCN funded) both based at UiO. She has recently written popular articles with Pierre Lison on whether AI and LLMs can understand speech (Lison & Falkum 2020; Falkum & Lison 2023). **Nicholas Allott**, ILOS, is an expert on pragmatics, with a particular interest in fundamental questions about the (post-)Gricean inferential model of communication. **Joanna Pollock**, IFIKK, has expertise in epistemology and philosophy of language, has worked on the importance of pragmatics for understanding the nature of knowledge communication and is PI of an RCN project on testimony. **Pierre Lison** is Chief Research Scientist at Norsk Regnesentral (Norwegian Computing Center), and Associate Professor at the Language Technology Group, UiO, and is an expert on (computer) natural language processing, machine learning and conversational interfaces. He currently leads two RCN research projects, one of which seeks to represent the dialogue state of conversational domains. His expert knowledge of the technical aspects of LLMs will be vital to the development of both the theoretical and empirical work of this project. **Mikhail Kissine** is professor of linguistics at the Université libre de Bruxelles, and has extensive expertise in experimental and theoretical pragmatics. He has published on pragmatic processing and communication, children and adults, both typical and atypical, and he has also worked on susceptibility to misinformation. He is a leading pragmatic theorist and advocate of non-inferentialist accounts of communication. Thus, the linguists on this project represent both sides of the debate noted above on inference and speaker's intentions in communication; we envisage productive 'adversarial collaboration' on this issue (in the spirit of Kahneman & Klein 2009).

Academic and societal impact of the project; utilisation of results

The impact of AI on society is one of today's most pressing questions, and is one for which there is an urgent need for insights from the humanities. Humanities perspectives are needed to properly understand the effects of the AI technological revolution on our human cognitive environment, and to politically ensure that the rapid development follows a safe and ethically sound route, mitigating the many risks involved. This project, which will provide novel, much-needed insights into the effects of our conversational interactions with LLMs will be a crucial contribution to this.

The project will result in several articles in high-impact academic journals (e.g., Cognition, Mind & Language, Synthese, Journal of Memory and Language, Child Development, etc.) as well as opinion pieces in national and international newspapers and journals (e.g., Aftenposten, Morgenbladet, Klassekampen, The Conversation, etc). The project funding will be used to establish an interdisciplinary research group on AI, language and philosophy at HF. Students will be encouraged to participate in events organised by the research group and we will announce MA scholarships for students interested in writing MA theses related to the project theme. We envisage a close collaboration with the BA Honours programme at HF, where AI has been the common interdisciplinary theme since the study programme was launched in 2019.

The project will serve as a pilot for larger funding applications to upcoming thematic calls by RCN and EU on the societal impact of AI, to be developed by the project group. Since the topic of the project is also of considerable public interest, we will aim to organise public debates on some of the questions to be investigated, in addition to the academic workshops that are planned during the project period. The project leader recently participated in an NRK debate on the impact of LLMs on Norwegian language, which attracted a large audience.[3] The project also has considerable potential for innovation, given the likelihood of a large and increasing need for knowledge about the effects of LLMs, both in the public and private sector.

References

Allott, N. (2008). Pragmatics and rationality. PhD thesis, University of London. / Allott, N. (2023). Encapsulation, inference and utterance interpretation. *Inquiry*. DOI:10.1080/0020174X.2023.2267084 / Bach, K. & Harnish, R. M. (1979). *Linguistic*

[3] https://radio.nrk.no/podkast/spraakteigen/l_1fd7fca0-4309-4624-97fc-a04309c62488

Communication and Speech Acts. MIT Press. / Barattieri di San Pietro, C., Frau, F., Mangiaterra, V. & Bambini, V. (2023). The pragmatic profile of ChatGPT: assessing the communicative skills of a conversational agent. *Psyarxiv Preprint*. / Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the dangers of stochastic parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York. / Breheny, R. (2006). Communication and folk psychology. *Mind & Language*, 21(1), 74–107. / Breheny, R., Ferguson, H. J. & Katsos, N. (2013). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition*, 126(3), 423–440. DOI: 10.1016/j.cognition.2012.11.012 / Carston, R. (2007). How many pragmatic systems are there? In M.-J. Frapolli (Ed.), *Saying, Meaning, Referring: Essays on the Philosophy of Francois Recanati* (pp. 18–48). Palgrave. / De Cosmo, L. (2022). Google engineer claims AI chatbot is sentient: Why that matters. *Scientific American*, *July 12, 2022*. / Dell'Acqua, F., McFowland, E., … Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, *24-013*. / Dennett, D. C. (1987). *The Intentional Stance*. MIT Press. / Eloundou, T., Manning, S., … Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint*. Falkum, I. L. & Lison, P. (2023). Er prateroboten ChatGPT en klok samtale-partner eller papegøye? *forskersonen.no*, 18. februar 2023. / Frank, M. C. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Science*. DOI 10.1016/j.tics.2023.08.007 /Fricker, E. (2012). Stating and insinuating. *Aristotelian Society Supplementary Volume, 86*(1), 61-94. DOI:10.1111/j.1467-8349.2012.00208.x / Grice, P. (1957). Meaning. *The Philosophical Review, 66*, 377–388. / Grice, P. (1967[1989]). Logic and conversation: William James lectures. In *Studies in the Way of Words* (1989) (pp. 1–143). Harvard University Press. / Hall, A. & Mazzarella, D. (2023). Pragmatic inference, levels of meaning and speaker accountability. *Journal of Pragmatics, 205*, 92-110. DOI: 10.1016/j.pragma.2022.12.007 / Hofstadter, D. R. (1995). *Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books. / Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist, 64*(6), 515-526. doi:10.1037/a0016755 / Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi. *Lingbuzz Preprint*, 007190. / Kissine, M. & Klein, O. (2013). Models of communication, epistemic trust, and epistemic vigilance. In J. P. Forgas, O. Vincze & J. László (Eds.), *Social Cognition and Communication* (pp. 139–154). Psychology Press. DOI:10.4324/9780203744628 / Köder, F. & Falkum, I. L. (2020). Children's metonymy comprehension: Evidence from eye-tracking and picture selection. *Journal of Pragmatics, 156*, 191-205. DOI: 10.1016/j.pragma.2019.07.007 / Korta, K. & Perry, J. (2006). Pragmatics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. / Lison, P. & Falkum, I. L. (2020). Kan kunstig intelligens «forstå» språk? *Aftenposten*, 19.11.2020/ Mallory, F. (draft). Teleosemantics for neural word embeddings. https://fintanmallory.files.wordpress.com/ / Mallory, F. (forth.). What do large language models model? In R. Sterken & H. Cappelen (Eds.), *Communicating With AI*. OUP. / Marchesi, S., Ghiglino, D., … Wykowska, A. (2019). Do we adopt the intentional stance toward humanoid robots? *Frontiers in Psychology, 10*, 450. DOI: 10.3389/fpsyg.2019.00450 / Mazzarella, D. (2014). Is inference necessary to pragmatics? *Belgian Journal of Linguistics, 28*, 71-95. doi:10.1075/bjl.28.04maz / Mazzarella, D. (2015). Inferential pragmatics and epistemic vigilance. PhD thesis, UCL / Mazzarella, D. & Pouscoulous, N. (2020). Pragmatics and epistemic vigilance: A developmental perspective. *Mind & Language*, 1–22. DOI: 10.1111/mila.12287 / Mazzarella, D., Reinecke, R., Noveck, I. & Mercier, H. (2018). Saying, presupposing and implicating: How pragmatics modulates commitment. *Journal of Pragmatics, 133*, 15–27. DOI: 10.1016/j.pragma.2018.05.009 / Miller, E. & Kissine, M. (2023). Suggestibility to presupposed contents. [Lecture]. XPrag 2023, Paris. / Millikan, R. G. 1984. *Language, Thought and Other Biological Categories*. MIT Press. / Noveck, I. A. (2018). *Experimental Pragmatics: The Making of a Cognitive Science*. CUP. / OpenAI. (2023). ChatGPT can now see, hear, and speak. / Pantazi, M., Kissine, M., & Klein, O. (2018). The power of the truth bias: False information affects memory and judgment even in the absence of distraction. *Social Cognition, 36*(2), 167-198. DOI:10.1521/soco.2018.36.2.167 / Piantadosi, S. (2023). Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint*, 7180. / Pouscoulous, N. & Tomasello, M. (2020). Early birds: Metaphor understanding in 3-year-olds. Journal of Pragmatics, 156, 160–167. DOI: 10.1016/j.pragma.2019.05.021 / Recanati, F. 2002. Does linguistic communication rest on inference? *Mind & Language 17* (1&2): 105–126. DOI:10.1111/1468-0017.00191 / Reeves, B. & Nass, C. I. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. CSLI Publications / Ronderos, C. R. & Falkum, I. L. (2023). Suppression of literal meaning in single and extended metaphors. *Front Psychol, 14*, 1135129. DOI: 10.3389/fpsyg.2023.1135129 / Ronderos, Mathisen, Noveck & Falkum, under revision. Straight enough: Deriving imprecise interpretations of absolute adjectives. / Sperber, D. & Wilson, D. (1986). *Relevance: Communication and Cognition.* Blackwell. / Stokke, A. (2018). *Lying and insincerity*. Oxford University Press. / Tomasello, M. 2008. Origins of Human Communication. MIT Press. Weizenbaum, J. (1967). Contextual understanding by computers. *Communications of the ACM, 10*(8), 474–480. / Wolfram, S. (2023). *What is ChatGPT Doing And Why Does it Work?* Wolfram Media.