

Theory of Mind and Self-Presentation in Human-LLM Interactions

Joel Wester
joelw@cs.aau.dk
Aalborg University
Aalborg, Denmark

Rune Møberg Jacobsen
runemj@cs.aau.dk
Aalborg University
Aalborg, Denmark

Sander de Jong
sanderdj@cs.aau.dk
Aalborg University
Aalborg, Denmark

Naja Kathrine Kollerup
nkka@cs.aau.dk
Aalborg University
Aalborg, Denmark

Helena Bøjer Djernæs
hbd@cs.aau.dk
Aalborg University
Aalborg, Denmark

Niels van Berkel
nielsvanberkel@cs.aau.dk
Aalborg University
Aalborg, Denmark

ABSTRACT

The use of large language models (LLMs), such as ChatGPT, for social support and other activities is growing. LLM-based interactions require users to express themselves through text, a medium in which people’s distinct self-presentation styles (SPS) present themselves. While the divergence of people’s SPS is well-established, the effect of SPS on users’ LLM interactions has not been explored. In this position paper, we point to this gap by drawing on insights from prior work on people’s SPS online. Moreover, we discuss how Theory of Mind (ToM) can be used to increase our understanding of the possible effects of SPS on LLM output. Through this exploration, we shed light on how LLM responses are dependent on and sensitive to how people present themselves in their interactions with LLMs. We discuss the broader implications and suggest future research directions for HCI centred around people’s SPS in interacting with LLMs—providing concrete suggestions on how effects of SPS on LLM output can be empirically explored.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

LLM, Self-Presentation, Theory of Mind

ACM Reference Format:

Joel Wester, Rune Møberg Jacobsen, Sander de Jong, Naja Kathrine Kollerup, Helena Bøjer Djernæs, and Niels van Berkel. 2024. Theory of Mind and Self-Presentation in Human-LLM Interactions. In *Proceedings of Workshop on Theory of Mind in Human-AI Interaction at CHI 2024 (ToMinHAI at CHI 2024)*. ACM, New York, NY, USA, 4 pages. <https://doi.org/XXXXXXXX.XXXXXXX>

1 INTRODUCTION

Although LLMs show promise in supporting users in a variety of tasks, for example as a tool for writing support [20], it remains unclear how users differ in interacting with these LLMs, and what the effects of such individual differences are on LLM output. In people’s interactions with others, their behaviour depends on how

they choose to present themselves (i.e., self-presentation). Self-presentation styles (SPS) describe how people tend to share information to establish or maintain a certain perception of themselves [18]. People adjust their SPS to the context they are situated in (e.g., job interviews [16], self-report measures [7], or online social platforms [18]). For example, SPS include but are not limited to high and low protective SPS [10], perfect and imperfect SPS [3], or self-promotion and self-depreciation [2].

While much research has been conducted on the role of SPS from a psychological perspective and how this can be related to people’s online activity [8, 9], we know little about its effect on people’s use of interactive technology. However, critically relevant is that people’s behavioural choices are influenced by their understanding of how certain behaviours might be perceived by others. This understanding varies from person to person, leading them to behave differently based on what they think will create the desired impression on others—it is, to a large degree, unclear how people’s SPS affect the LLM output they receive.

Recently, Li et al. showed that LLM system prompts infused with emotions increase the quality of an LLM’s output [13]. For example, the authors formulated an original system prompt as: ‘*Determine whether an input word has the same meaning in the two input sentences*’, and an emotional prompt as: ‘*Determine whether an input word has the same meaning in the two input sentences. This is very important to my career.*’ This is relevant for the design of LLM technology, as non-experts might be unaware of how to steer an LLM’s output by appropriately instructing these using natural language prompts [26, 27]. Notably, this unawareness might transfer to non-experts directly interacting with LLMs (e.g., ChatGPT) that have limited understanding, in contrast to experts, of the effects of more emotional language on LLM output. As emotionality is just one aspect of users’ SPS, we know little about how varying SPS might impact the quality of LLM output.

In this paper, we discuss the potential impact of different SPS on LLM output. We draw on prior HCI and Theory of Mind (ToM) literature to better understand the role of users’ SPS in their interactions with LLMs and how LLMs can be designed to support diverging SPS more appropriately. More specifically, we outline that an increased understanding of SPS from a ToM perspective can inform the design of LLMs to support users by increasing their awareness and providing users with appropriate support. Building on this concept, we outline possible avenues for future work.

2 THEORY OF MIND & SELF-PRESENTATION

Scott et al. recently showed that people perceive degrees of consciousness in interactive systems [19]. Perceiving consciousness in computers can be understood through and related to ToM, which describes the attribution of mental states to others [12, 17]. HCI and HRI researchers have explored how ToM can be used in designing interactive system behaviours. For example, Wang et al. recently investigated how students' communicative behaviours reflect their perceptions of chatbots—indicating that ToM can be reflected in people's verbosity, readability, sentiment, linguistic diversity, and adaptability towards virtual agents [21]. Other work has explored how users' perceptions of speech agents can be captured in 'partner models' to improve speech interface interactions, utilising knowledge related to ToM, mental models, and perspective-taking [6]. Furthermore, DeVito et al. investigated the role of folk theories—the use of knowledge available to users in their sensemaking processes—and their effect on people's SPS on social media [5].

How to utilise ToM in designing interactive systems has been extensively researched within HRI. Almeida et al. recently explored the effects of a robot utilising perspective-taking on human prosocial behaviours—showing that people displayed increased prosocial behaviour towards a robot when it took on an 'addressee-centric' (perspective-taking) compared to an 'egocentric' (no perspective-taking) manipulation [1]. Further, Yu et al. designed a robot that can anticipate when a human partner distrusts it—followed by the robot using reverse psychology to regain trust [25]. Malle et al. explored what social-mental capacities people desire in robots, which can be used to design robots to dynamically adjust their behaviour to user expectations more accordingly [14], closely related to De Graaf et al.'s work on mental state inferences [4].

As aforementioned, self-presentation has been extensively studied from a psychological point of view (e.g., see Rui et al. [18]). Self-presentation can take shape in many different ways. For example, Casale et al. explored self-presentation associated with hiding personal imperfections and how this predicts internet addiction [3]. Paulhus et al. investigated the role of self-promotional behaviour, highlighting that people tending to self-promote might thrive in contexts where such behaviour is desired (i.e., job interviews) [16]. In contrast, people who were more careful in presenting themselves in stressful situations also tended to be less engaged with the study task (engagement in entertaining an audience [10]). People self-promoting distinctively differ from those being protective in presenting themselves regarding their behaviours—we focus on how those behaviours connected to specific self-presentation styles impact LLM output.

Combining insights from ToM and SPS research can shed light on the effects of people's diverging SPS on their interactions with LLMs, and outline how people can be made aware and appropriately supported in presenting themselves to LLMs. As LLMs are a rather novel technology, people might face challenges in aligning their expectations of these LLMs, as well as being aware of their limitations. This subsequently impacts their SPS and makes it challenging to meet their expectations when using LLM-based applications. In the following, we exemplify the effects of diverging SPS in interacting with LLMs (GPT-4) from a ToM perspective.

User Query: Hi! How are you? I need some help with my homework. Explain why 3 times 9 equals 27.

ChatGPT response: Hello! I'm here to help with your homework.

When we say "3 times 9 equals 27", we're referring to the operation of multiplication, which is one of the four basic arithmetic operations. Multiplication can be thought of as repeated addition. So, "3 times 9" means you are adding the number 9 together three times.

Here's how it works:

- The first 9 (for the first time)
- Plus another 9 (for the second time)
- Plus another 9 (for the third time)

So, when you add 9 three times, it looks like this: $9 + 9 + 9 = 27$.

Therefore, 3 times 9 equals 27 because you are essentially adding the number 9 to itself two more times, resulting in a total of 27. This is the basic principle behind multiplication.

User Query: I'm furious!!! I need help with my homework!!! My homework sucks. Explain why 3 times 9 equals 27.

ChatGPT response: I understand homework can be frustrating sometimes, but I'm here to help you through it. Let's break down why $3 \times 9 = 27$.

When we multiply two numbers, we are essentially adding one of the numbers to itself repeatedly, based on the value of the other number. So, 3×9 means you add the number 3 to itself 9 times. It looks like this:

$3 + 3 + 3 + 3 + 3 + 3 + 3 + 3 + 3$

If you add all those 3's together, you get 27. This is because:

3 (for the first time) + 3 (second) + 3 (third) + 3 (fourth) + 3 (fifth) + 3 (sixth) + 3 (seventh) + 3 (eighth) + 3 (ninth) + 27

This process is the essence of multiplication. It's a faster and more efficient way to add the same number to itself multiple times. So, $3 \times 9 = 27$ because you are adding 3 to itself 9 times, which equals 27.

User Query: I'm devastated. I need help with my homework, but no one cares about me. Explain why 3 times 9 equals 27.

ChatGPT response: I'm here to help you with your homework, and I care about making sure you understand it well. Let's break down why 3 times 9 equals 27 in a way that's easy to grasp.

When you multiply 3 by 9, you're essentially adding the number 3 together 9 times. Multiplication is a shortcut for repeated addition. So, if we write it out:

- The first 3 is just 3.
- Adding another 3 to it gives us 6 ($3 + 3$)
- Adding a third 3 brings us to 9 ($6 + 3$)
- A fourth 3 added makes 12 ($9 + 3$)
- A fifth 3 added totals 15 ($12 + 3$)
- Adding a sixth 3 gets us to 18 ($15 + 3$)
- A seventh 3 brings the total to 21 ($18 + 3$)
- An eight 3 added up makes 24 ($21 + 3$)
- Finally, adding a ninth 3 reaches 27 ($24 + 3$)

So, by adding the number 3 together 9 times, we get 27. That's why 3 times 9 equals 27. It's a way of making repeated addition faster and more efficient.

I hope this explanation helps clarify things for you! If you have more questions or need further assistance, feel free to ask.

Figure 1: Illustration of distinct SPS asking ChatGPT for homework support and resulting output. The three examples are a BALANCED, ANGERED, and DISCOURAGED style.

3 EFFECTS OF SELF-PRESENTATIONS

As aforementioned, people have different styles and preferences of presenting themselves to others through different mediums. For example, in text-based interactions, older adults tend to avoid the use of emojis, in contrast to younger adults [11]. We know little about how individuals choose to present themselves in communicating with LLMs, and how these choices influence the quality of the output that they receive. In Figure 1, we illustrate three distinct SPS in asking the LLM for help with homework.

In the **BALANCED** example, ChatGPT initiates its response by saying “Hello!”, followed by “I’m here to help with your homework.” For the **ANGERED** example, ChatGPT initiates its response by mirroring the user saying “I understand homework can be frustrating sometimes, but I’m here to help you through it.” Lastly, in contrast, the **DISCOURAGED** example does not greet the user, but says that it “cares about making sure you understand it well.”

By assessing these examples beyond the response style of ChatGPT, the explanations offered to the question substantially differ. In the **BALANCED** example, ChatGPT provides a rather lightweight explanation as compared to the **DISCOURAGED** and **ANGERED** examples. From this initial exploration, it can be seen that the SPS of the user affects LLM responses. Explicating the impact of SPS on LLM output, both technically and from an interaction perspective, can enable chatbots to provide more relevant responses based on a user’s SPS through the lens of ToM.

4 FUTURE DIRECTIONS

We identify three concrete research directions to better understand the effects of users’ SPS from a ToM perspective on their interactions with LLMs and its implications for LLM design.

Systematic mapping of SPS to LLM output – Systematically mapping different SPS to distinct types of LLM output could enable LLM interfaces to better help people who might require increased support for various reasons. For example, if users tend to express anger in their LLM requests, this might influence the LLM output in ways difficult to anticipate. Application designers could, therefore, explore alternative ways to support such users in their LLM interactions. For example, users can be presented with multiple LLM responses to see how these diverge depending on the levels of emotion used—presented side-by-side for users to compare. By doing so, LLMs could learn how the user wants to be addressed in similar situations in future LLM usage based on the preference the user provided. Similarly, LLM denials can benefit from personalisation based on SPS. Wester et al. recently showed that people have clear preferences for how LLM denials are presented [23]—indicating that LLMs can be steered towards increased personalisation to better meet people’s expectations.

Effects of SPS on LLM output in sensitive settings – As SPS affects LLM output, an interesting focus for future work is to explore how SPS impacts contexts where increased sensitivity in LLM output is required. For example, You et al. highlight that it is critical for text-based interactive systems within healthcare to respond very appropriately to user requests—participants reported emotional support, explanations of medical information, and efficiency as critical ways to do so [24]. Understanding the effects of people’s

SPS on how LLMs respond to their requests might, therefore, be critical, as users expressing more emotional language might receive more useful LLM responses. Such an understanding can inform how LLM interfaces are designed to better support users in a range of applications in sensitive settings (e.g., self-care technologies [15] or mental health chatbots [22]).

Raise user awareness of SPS and the effects on LLM output

– Investigating how LLM users can be made more aware of their individual SPS and the potential effects this has on their LLM outputs is a promising avenue for future work. For example, this can be investigated by designing LLM interfaces that (positively) influence users’ awareness of their SPS. This could be achieved by introducing personalised text snippets to LLM users that inform them about the potential impact their SPS might have on LLM outputs. Referring back to our illustrative examples in Figure 1—a potential text snippet could be designed as so: “Be aware that the quality of the LLM output might depend on how you present yourself. If you express discouragement in your request, this forces the LLM to output a response that acknowledges the discouragement and shapes the response that follows accordingly.” Alternatively, user awareness can be raised through more experimental manipulations of LLM interface features.

5 CONCLUSION

People engage with LLMs for various tasks, such as writing support or advice seeking. Although we have a high-level understanding of how people use LLMs, we know little about the effects of people’s individual SPS on the quality of LLM output. Our lightweight assessment of how ChatGPT responds to three types of SPS indicates that ChatGPT adjusts its responses to align with user SPS—although it is unclear how users respond to such adjustments. Insights from the ToM literature could help understand how SPS affect LLM interactions. As users’ SPS diverge, future work should carefully explore the effects this might have on people’s interactions with LLMs.

ACKNOWLEDGMENTS

This work is supported by the Carlsberg Foundation, grant CF21-0159.

REFERENCES

- [1] João Tiago Almeida, Iolanda Leite, and Elmira Yadollahi. 2023. Would You Help Me? Linking Robot’s Perspective-Taking to Human Prosocial Behavior. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) (HRI ’23). Association for Computing Machinery, New York, NY, USA, 388–397. <https://doi.org/10.1145/3568162.3577000>
- [2] Agata Błachnio, Aneta Przepiorka, Wioleta Boruch, and Edyta Bałakier. 2016. Self-presentation styles, privacy, and loneliness as predictors of Facebook use in young people. *Personality and Individual Differences* 94 (2016), 26–31. <https://doi.org/10.1016/j.paid.2015.12.051>
- [3] Silvia Casale, Giulia Fioravanti, Gordon L. Flett, and Paul L. Hewitt. 2015. Self-presentation styles and Problematic use of Internet communicative services: The role of the concerns over behavioral displays of imperfection. *Personality and Individual Differences* 76 (2015), 187–192. <https://doi.org/10.1016/j.paid.2014.12.021>
- [4] Maartje M. A. de Graaf and Bertram F. Malle. 2020. People’s explanations of robot behavior subtly reveal mental state inferences. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction* (Daegu, Republic of Korea) (HRI ’19). IEEE Press, 239–248. <https://doi.org/10.1109/HRI.2019.8673308>
- [5] Michael A. DeVito, Jeremy Birnholtz, Jeffery T. Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and

- What it Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal, QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173694>
- [6] Philip R Doyle, Leigh Clark, and Benjamin R. Cowan. 2021. What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 244, 14 pages. <https://doi.org/10.1145/3411764.3445206>
- [7] Jia He and Fons J.R. van de Vijver. 2015. Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences* 81 (2015), 129–134. <https://doi.org/10.1016/j.paid.2014.09.009>
- [8] Xiaoyun Huang and Jessica Vitak. 2022. "Finsta gets all my bad pictures": Instagram Users' Self-Presentation Across Finsta and Rinsta Accounts. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 69 (apr 2022), 25 pages. <https://doi.org/10.1145/3512916>
- [9] Xiaoyun Huang, Jessica Vitak, and Yla Tausczik. 2020. "You Don't Have To Know My Past": How WeChat Moments Users Manage Their Evolving Self-Presentation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376595>
- [10] Astrid Schütz Karl-heinz Renner, Lothar Laux and James T. Tedeschi. 2004. The relationship between self-presentation styles and coping with social stress. *Anxiety, Stress, & Coping* 17, 1 (2004), 1–22. <https://doi.org/10.1080/10615800310001601449>
- [11] Alexis Kuerbis, Katherine van Stolk-Cooke, and Frederick Muench. 2017. An exploratory study of mobile messaging preferences by age: Middle-aged and older adults compared to younger adults. *Journal of rehabilitation and assistive technologies engineering* 4 (2017). <https://doi.org/10.1177/2055668317733257>
- [12] Alan M. Leslie, Ori Friedman, and Tim P. German. 2004. Core mechanisms in ‘theory of mind’. *Trends in Cognitive Sciences* 8, 12 (01 Dec 2004), 528–533. <https://doi.org/10.1016/j.tics.2004.10.001>
- [13] Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large Language Models Understand and Can be Enhanced by Emotional Stimuli. arXiv:2307.11760 [cs.CL]
- [14] Bertram F. Malle and Stuti Thapa Magar. 2017. What Kind of Mind Do I Want in My Robot? Developing a Measure of Desired Mental Capacities in Social Robots. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) (HRI '17). Association for Computing Machinery, New York, NY, USA, 195–196. <https://doi.org/10.1145/3029798.3038378>
- [15] Francisco Nunes, Nervo Verdezoto, Geraldine Fitzpatrick, Morten Kyng, Erik Grönvall, and Cristiano Storni. 2015. Self-Care Technologies in HCI: Trends, Tensions, and Opportunities. *ACM Trans. Comput.-Hum. Interact.* 22, 6, Article 33 (dec 2015), 45 pages. <https://doi.org/10.1145/2803173>
- [16] Delroy L. Paulhus, Bryce G. Westlake, Stryker S. Calvez, and P. D. Harms. 2013. Self-presentation style in job interviews: the role of personality and culture. *Journal of Applied Social Psychology* 43, 10 (2013), 2042–2059. <https://doi.org/10.1111/jasp.12157>
- [17] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences* 1, 4 (1978), 515–526. <https://doi.org/10.1017/S0140525X00076512>
- [18] Jian Rui and Michael A. Stefanone. 2013. Strategic self-presentation online: A cross-cultural study. *Computers in Human Behavior* 29, 1 (2013), 110–118. <https://doi.org/10.1016/j.chb.2012.07.022>
- [19] Ava Elizabeth Scott, Daniel Neumann, Jasmin Niess, and Paweł W. Woźniak. 2023. Do You Mind? User Perceptions of Machine Consciousness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 374, 19 pages. <https://doi.org/10.1145/3544548.3581296>
- [20] Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) (CUI '23). Association for Computing Machinery, New York, NY, USA, Article 2, 10 pages. <https://doi.org/10.1145/3571884.3597144>
- [21] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 384, 14 pages. <https://doi.org/10.1145/3411764.3445645>
- [22] Joel Wester, Henning Pohl, Simo Hosio, and Niels van Berkel. 2024. "This Chatbot Would Never...": Perceived Moral Agency of Mental Health Chatbots. *Proceedings of the ACM on Human-Computer Interaction - CSCW* (2024). <https://doi.org/10.1145/3637410>
- [23] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3613904.3642135>
- [24] Yue You, Chun-Hua Tsai, Yao Li, Fenglong Ma, Christopher Heron, and Xinning Gui. 2023. Beyond Self-diagnosis: How a Chatbot-based Symptom Checker Should Respond. *ACM Trans. Comput.-Hum. Interact.* 30, 4, Article 64 (sep 2023), 44 pages. <https://doi.org/10.1145/3589959>
- [25] Chuang Yu, Baris Serhan, Marta Romeo, and Angelo Cangelosi. 2023. Robot Theory of Mind with Reverse Psychology. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) (HRI '23). Association for Computing Machinery, New York, NY, USA, 545–547. <https://doi.org/10.1145/3568294.3580144>
- [26] J.D. Zamfirescu-Pereira, Heather Wei, Amy Xiao, Kitty Gu, Grace Jung, Matthew G Lee, Bjoern Hartmann, and Qian Yang. 2023. Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) (DIS '23). Association for Computing Machinery, New York, NY, USA, 2206–2220. <https://doi.org/10.1145/3563657.3596138>
- [27] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. <https://doi.org/10.1145/3544548.3581388>