Network science highlights the emotional structure of counselling conversations simulated by Large Language Models and humans

Edoardo Sebastiano De Duro, Riccardo Improta and **Massimo Stella**
CogNosco Lab, Department of Psychology and Cognitive Science, UniTrento

Abstract

This study investigates CounseLLMe, a dataset comprising 400 simulated mental health counselling dialogues among two Large Language Models (LLMs). These conversations, each containing 20 exchanges, were conducted in both English (using OpenAI's GPT-3.5 and Claude-3's Haiku) and Italian (using Claude-3's Haiku and LLaMAntino). A professional psychotherapist assisted in fine-tuning the prompts for realism. The dialogues are compared against patterns in recently released and investigated human mental health conversations focused on depression. By investigating dialogues via the cognitive framework of textual forma mentis networks, we find that LLMs tend to stay positive even when debating depression. Furthermore, LLMs tend to become more verbose along conversations, but without creating syntactic/semantic networks of increasing complexity, i.e. degree assortativity and average shortest path length remain stationary despite increases in verbosity. We discuss this difference in view of relevant literature on rumination and mental navigation.

1 Introduction

Large Language Models (LLMs) are artificial intelligences (AIs) trained in reproducing human texts, one word after the other. LLMs' complexity stems from their cognitive skills, inherited from training over vast amounts of knowledge. Understanding how these AIs can behave in scenarios like mental health counselling is crucial (De Choudhury et al., 2023), given the wide variety of online and untested services having LLMs act as counsellors to human with mental distress. Psychological counselling is a field characterised paramount requirement for empathy, understanding, and accurate information dissemination. Traditional models may falter in providing the nuanced care necessary due to their inability to fully grasp the depth of human emotions and the subtleties of psychological distress (De Freitas et al., 2022). Hence, it is crucial to rigorously evaluate LLMs' knowledge frameworks and belief systems concerning this sensitive domain.

2 Motivation

Whereas traditional studies focus on human-computer interactions, CounseLLMe (De Duro et al., 2024) considers 2 LLMs conversing with each other in a virtual mental health counselling scenario. The aims of the dataset are to: (i) shed light on the models' ability to navigate discussions related to depression; (ii) reproduce LLMs' ability to mimic human emotions and syntactic associations during psychological counselling.

In CounseLLMe, one LLM adopts the role of a counsellor, while the other embodies the individual seeking help. In other words, at every turn, each model was instructed to play the role of a human patient affected by depressive symptoms conversing with another model that was, instead, prompted to act as an expert psychotherapist. We adopt a machine psychology perspective (Bertolazzi et al., 2023; Stella et al., 2023), bridging computer science and psychology frameworks within a complex systems approach, to study the AI agents as if they were humans, using their texts to extract insights relative to the ways these models represent and describe mental health dialogues.

3 Results

CounseLLMe is a dataset consisting of 400 conversations, introduced and investigated in a previous study (De Duro et al., 2024). These conversations - of 20 quips each - were generated either in English (using OpenAI's GPT 3.5 and Claude-3's Haiku) or Italian (with Claude-3's Haiku and LLamantino). We carefully selected prompts with the consultation of a human professional in psychotherapy. To investigate the resulting conversations, we

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

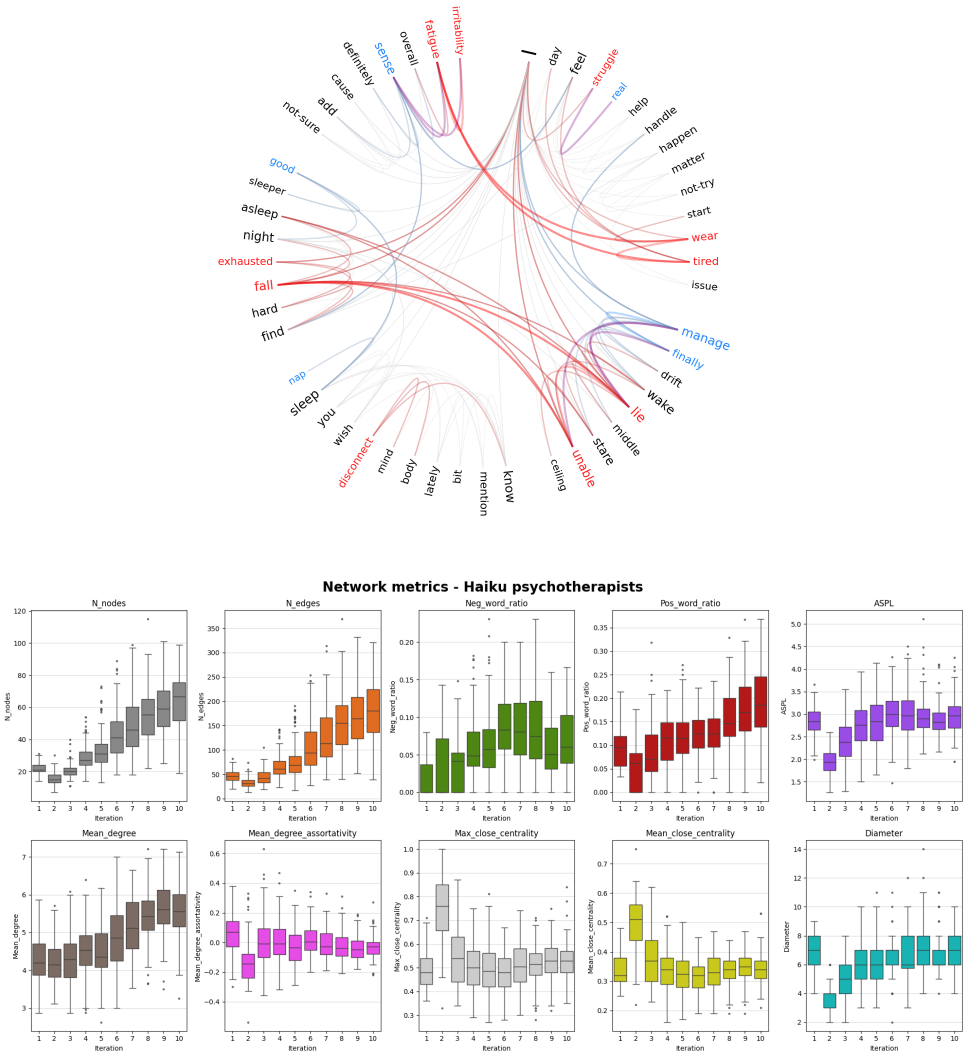

Figure 1: **Top:** A textual forma mentis network extracted from a quid of a patient (here Claude's Haiku) in the dataset. Positive (negative, neutral) words are highlighted in blue (red, black). **Bottom:** Sequential representation of various network measures extracted from the output of Claude's Haiku model (here playing the role of the psychotherapist).

employed the complex-systems technique of textual forma mentis networks (Stella, 2020), where nodes represent concepts and links indicate syntactic or semantic relationships between concepts in the dialogues' quips. Additionally, we performed some sequence-based techniques to show the evolution of the conversation in terms of language and complexity.

We find that all LLMs display domain knowledge relative to psychotherapy, successfully reproducing questions and jargon coming from clinical psychology. Furthermore, we find that Claude-3's Haiku can impersonate realistic patients, associating several negative concepts (cf. Figure 1, top) in ways structurally different from GPT 3.5, the latter being dominated by positive words and associations even when impersonating patients reporting very negative experiences. All LLMs tend to grow in verbosity along their conversations, producing larger cognitive networks, with more links and nodes (cf. Figure 1, bottom). These elements indicate a tendency for LLMs to fill gaps in their syntactic structures along conversations, revisiting concepts in different contexts along the conversation without concentrating on any of them, like it would be expected from therapists focusing on depression symptoms (Neenan and Dryden, 1999). CounseLLMe provides interesting perspectives for comparing LLMs and humans, the dataset is available at `https://osf.io/2ay8d/`.

References

Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. 2023. Chatgpt's information seeking strategy: Insights from the 20-questions game. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 153–162.

Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.

Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. 2024. Introducing CounseLLMe: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans.

Julian De Freitas, Ahmet Kaan Uğuralp, Zeliha Oğuz-Uğuralp, and Stefano Puntoni. 2022. Chatbots and mental health: insights into the safety of generative ai. *Journal of Consumer Psychology*.

Michael Neenan and Windy Dryden. 1999. When laddering and the downward arrow can be used as adjuncts to inference chaining in rebt assessment. *Journal of rational-emotive and cognitive-behavior therapy*, 17(2):95–104.

Massimo Stella. 2020. Text-mining forma mentis networks reconstruct public perception of the stem gender gap in social media. *PeerJ Computer Science*, 6:e295.

Massimo Stella, Thomas T Hills, and Yoed N Kenett. 2023. Using cognitive psychology to understand gpt-like models needs to extend beyond human biases. *Proceedings of the National Academy of Sciences*, 120(43):e2312911120.