A Small Window to (Behaviourally) Working with LLMs

Alexandra Chesterfield on LinkedIn (2024)

Context

Alexandra, behavioural scientist at LSE, FORGOOD speaker and Director of Human-AI Collaboration at Salesforce, reflects on using various Open AI large language models (LLMs) as part of her research. She notes LLMs have driven a surge of activity in her LSE research group, enabling analysis of large, unstructured qualitative text to explore human behaviour and outcomes at unprecedented scale and speed. Her interactions surprised her, feeling closer to managing a person than operating software.

Key Insights

1) How you manage Chat GPT is just like managing a human

- Chunking instructions, simplifying language, not asking everything at once
- Anthropomorphising the LLM, please, thank you, great job
- "Come on, you can do better than that" prompts, author flags no evidence behind this, just intuition

→ Goals: How might treating AI output like a person shape not just how you engage with a task, but how you engage with colleagues and your own learning over time?

2) Strategies that work to get better human decision making work with LLMs too

- "Let's think step by step" instructions lead to more accurate results, as the AI explains its reasoning
- Like research on the “illusion of explanatory depth (IOED)”, explaining mechanics increases the awareness of own limited understanding, moderating attitudes or motivating improved understanding

→ Opinions: Could you backtrack and explain an AI's step by step reasoning yourself, if someone asked you?

3) It's not consistent (bit like us)

- Same or slightly reworded input does not guarantee the same output (unlike programming)
- Alexandra wonders whether human consistency tools, like checklists, could work for LLMs too

→ Fairness: In your customer facing AI, where would inconsistent responses send mixed signals to the people you're serving? Which customers and critical contexts can least afford an AI giving different replies?

4) But can still totally miss the point....

- LLM was asked to clean unstructured text from investigation reports, explicitly told to preserve key details from the report and return the cleaned text
- But instead, it reformatted every written document in the report (like memos and speeches) as emails, also inventing send/to fields and dates that did not exist
- This shows model can follow surface instruction (preserve fields) while fabricating structure not present in the source

→ Respect: What due diligence step in your process would catch fabricated structure like this before it reached someone's report or decision?

Chesterfield, A. (2024, January 5). A small window to (behaviourally) working with large language models [LinkedIn post]. LinkedIn.