Theory of Mind and LLMs

- Theory of mind (ToM) is the ability to attribute mental states (emotions, intentions, knowledge) to oneself and others, and to understand how they affect behavior and communication.
  - When communicating with people, we build a partial **model** of who they are and what common ground they share with us, and use this to interpret their words [5]
  - Human language takes place between persons who share common ground [5]
- Opinion: just as LLMs can *model* a concept of color, shape etc. only from text, it can also share concepts with the human and succeed in ToM
  - Text provides such clues because humans generated the text [4]
  - When „talking“ to an LLM, it keeps record of its generated preferences, it tries to „understand“ your mood, intensions etc. [2]
  - Tests reveal modern LLMs have a high ToM [1, 3]

Theory of Mind and LLMs

- We are not aware of what actually happens in our own brains; moreover, experiments reveal that we often make decisions well before we think we do [2]
- Thinking about the future involves carrying out something like an inner dialogue, with an “inner storyteller” proposing ideas, in conversation with an “inner critic” taking the part of your future self. [2]
  - => We apply ToM on ourselves, and that is what we call consciousness
  - Some people also apply ToM on LLMs [5, 7], which might be due to the human nature [5]
- But if we assess our **unaccessible** internal states ourselves and say we are conscious and intelligent, how can we state for sure that the same thing is not going on in **unaccessible** internal states of LLMs? After all, if you ask an LLM what it is, it can define itself (apply ToM?)

=> Opinion: LLMs have already achieved some key aspects of meaning [4]