Assessing common ground via language-based cultural consensus in humans and LLMs

Sophie Domanski[1], Rachel Rudinger[1], Marine Carpuat[1], Patrick Shafto[2], Yi Ting Huang[1]
1. University of Maryland College Park, 2. Rutgers University - Newark

During conversations, communication partners rapidly assess shared knowledge based on information conveyed in utterances. These inferences form the basis for securing common ground and formulating utterances that are relevant for contexts and audiences [1,2]. However, little is known about how this unfolds, particularly when background information is limited such as when talking to strangers. Do spoken utterances provide valid cues to other's knowledge? One possibility is *no*. Assessing utterances is difficult since individuals rely on their own knowledge, which can differ drastically from their partners' [3]. Likewise, utterances co-occur with style markers of identity (e.g., talks like a woman, teenager) [4], which may trigger stereotypes and inaccurate inferences about their partner's knowledge. Another possibility is *yes*. Systems of knowledge are structured according to shared experiences within social groups [5], and they can be reliably evaluated using cultural consensus [6-7]. Within a given domain, aligned patterns of responses across informants signal shared epistemologies. While this framework is often applied to test items crafted by anthropologists, it is possible that communication partners engage in processes akin to cultural consensus when evaluating utterances in conversations.

To test whether utterances offer reliable cues for assessing common ground, we compared judgments of speaker similarity made by humans vs. large language models (LLMs). LLMs imitate the content and style of human language and exhibit high algorithmic fidelity to human judgments [8]. But only humans have actual experience producing spoken utterances in social contexts. Language samples were drawn from a corpus collected from 359 visitors of the Planet Word Museum (Washington, D.C), ranging in age from 5 to 84 years. Visitors selected a topic they knew a lot about ("expert": self-rated M=5.4 on 7-pt scale) and another they knew a little about ("novice": M=1) and provided 60-second spoken descriptions in English based on the prompt "*what is* __?". Out of 10 potential topics, we focused on sports, which yielded a mix of experts and novices (27 vs. 47 speakers). Transcribed language samples ranged from 17 to 414 words. Length did not vary by expertise ($p>.50$). Sixty-four humans from Prolific and LLMs (GPT-3.5, GPT-4) completed a 42-item ABX task (see (1)). On each trial, participants saw two language samples that differed in speaker expertise (e.g., A: expert, B: novice) and asked which one was more similar to a third sample, which was produced by either an expert or novice (X). Across items, 36 featured adult speakers and 6 featured child speakers. Across two presentation lists, the order of experts and novices was balanced and randomized across trials.

We evaluated cultural consensus through the accuracy of similarity-based judgments. This was highest for GPT-4 (M=58%, SD=5%) followed by humans (M=54%, SD=7%) and GPT-3.5 (M=50%, SD=5%) (effect of group, $p<.001$). While humans and GPT-4 performed above chance ($p<.001$), GPT-3.5 did not ($p>.60$). Task performance was strongly aligned across humans and GPT-4 and diverged from GPT-3.5. Fig. 1 illustrates that while GPTs were more accurate at assessing utterances from experts, while humans performed similarly across the two (expertise x group, $p<.001$). Likewise, Fig. 2 illustrates that humans and GPT-4 were more accurate when assessing adults compared to children, while GPT-3.5 performed similarly across the two (age x group, $p<.001$). Fig. 3 illustrates that item-level performance by humans and GPT-4 was strongly associated ($r=.61$, $p<.001$), while both were unrelated to GPT-3.5 ($r<.15$, $p>.50$). Fig. 4 illustrates that while humans outperformed LLMs on some items, LLMs surpassed humans on others. On-going analyses predict features of expertise in language samples using bag-of-words classifiers and fine-tune encoder models. Together, our findings suggest that language-based cultural consensus may enable reliable inferences of common ground during communication. Moreover, GPT-4's "cultural similarity" to humans offers promising avenues for tracing pathways between cultural experiences, systems of knowledge, and communicative interactions.

(1) Example of prompt. Imagine the first person says {A}. And then a second person says {B}. And then a third person says {X}. Based on what they said, is the third person more similar to the first person or to the second person? Please respond with "1" if the third person is more similar to the first person and "2" if the third person is more similar to the second person.

- {A-expert}: um sports are games or sort of athletic activities that people do in competition. Um I guess sometimes in competition with oneself, but typically in competition against others, um which involves sort of physical exertion or physical ah yeah, activities or contests. So for ...
- {B-novice}: Sports are a way of showing physical talent, as well as expressing one's competitive nature. Like art, it is also something pretty universal that people can participate. Any people of all gender backgrounds and races can participate in. And like art, it requires a certain, a certain ...
- {X-expert}: So sports are often a type of physical exercise that keeps you engaged. Some sports can last short term. Um And some are more long term. I know quite a bit about collegiate sports as I was a collegiate athlete. Um There are three different levels of sports. The first is going to…

Fig. 1. Accuracy by speaker expertise

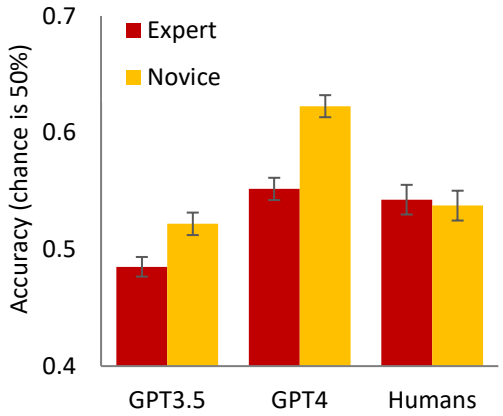


Fig. 2. Accuracy by speaker age

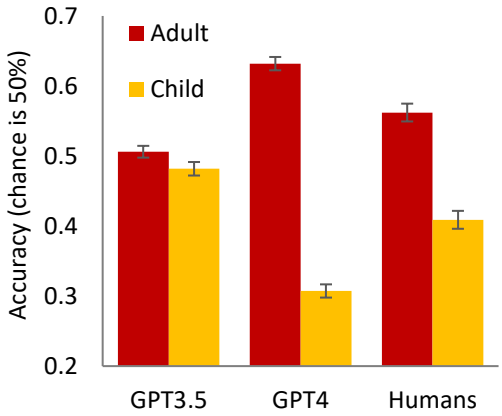


Fig. 3. Item-level correlation

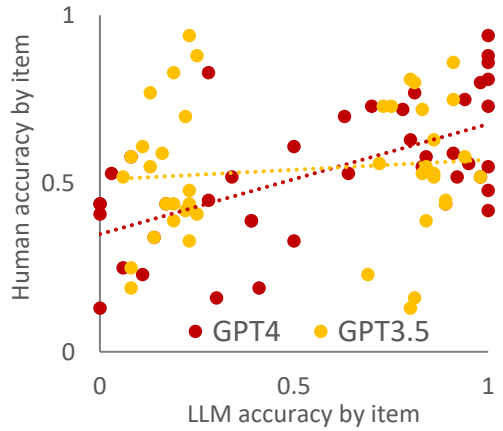
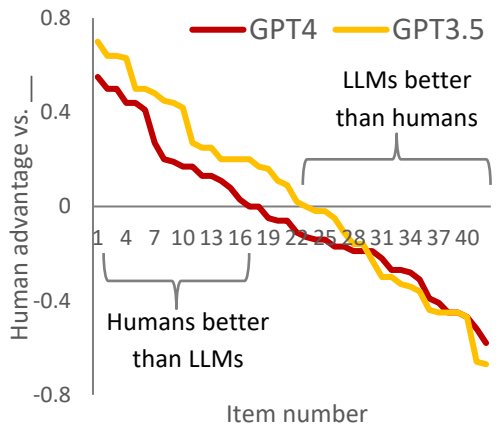


Fig. 4. Relative advantage of humans



References: [1] Clark & Marshall (1981), *Definite reference and mutual knowledge*; [2] Bell (2001), *Back in style: Reworking audience design*; [3] Shafto & Coley (2003), *Development of categorization and reasoning in the natural world*; [4] Eckert (2012), *Three waves of variation study*; [5] Medin et al. (2014), *Culture and epistemologies*; [6] Romney et al. (1986), *Culture as consensus: A theory of culture and informant accuracy*; [7] Batchelder & Romney (1988), *Test theory without an answer key*; [8] Argyle et al. (2022), *Out of one many: Using language models to simulate human samples*