Assessing Common Ground through Language-based Cultural Consensus in Humans and Large Language Models

Sophie Domanski[1,2](smd@umd.edu), Rachel Rudinger[3] (rudinger@umd.edu), Marine Carpuat[3] (marine@umd.edu), Patrick Shafto[4] (patrick.shafto@rutgers.edu), Yi Ting Huang[1,2] (ythuang1@umd.edu)

1. Department of Hearing and Speech Sciences, University of Maryland, College Park, MD, 20740 USA
2. Maryland Language Science Center, University of Maryland, College Park, MD, 20740 USA
3. Department of Computer Science, University of Maryland, College Park, MD, 20740 USA
4. Department of Mathematics and Computer Science, Rutgers University, Newark, NJ, 07102 USA

Abstract

During conversations, communication partners rapidly assess shared knowledge based on information in utterances. However, little is known about how this process unfolds, particularly when background information is limited such as when talking to strangers. Do spoken utterances provide valid cues to speaker knowledge? To test this, we applied a cultural consensus framework (e.g., Romney et al., 1986), and asked humans vs. large language models (LLMs) to assess speaker similarity based on their transcribed utterances. On each trial, participants saw two language samples that varied in speaker expertise (e.g., A: expert, B: novice) and were asked which one was more similar to a third sample, which was produced by either an expert or novice (X). Accuracy was highest for GPT-4 followed by humans and GPT-3.5. Humans and GPT-4 were more accurate at categorizing language samples from experts, while GPT-3.5 was better with novices. Likewise, humans and GPT-4 were more accurate with samples from adult compared to child speakers, while GPT-3.5 was similar across the two. Item-level performance by humans and GPT-4 was strongly associated, while both were unrelated to GPT-3.5. Our findings suggest that language-based cultural consensus may enable reliable inferences of common ground during communication, providing an algorithmic-level description of how partners may infer states of the world.

Keywords: common ground; cultural consensus; conversations; large language models

Introduction

Successful conversations require individuals to interpret and produce utterances with respect to the shared background with their communication partner. This mutual knowledge is known as common ground (Clark & Marshall, 1981; Gibbs, 1987; Geurts, 2017). Inferring common ground is particularly challenging when talking to strangers, where the lack of historical interactions makes it difficult to deduce shared experiences. In these circumstances, individuals must quickly assess their partners' knowledge based on information provided within utterances (e.g., *I think this person is a basketball expert*) and generate probabilistic inferences about likely knowledge in related domains (e.g., *they might also know a lot about other sports*). This suggests that accessing a speaker's communicative intent is a joint inference about the meaning of what is said (foreground) and the common ground that gave rise to the utterance (background). Nevertheless, little is known about the algorithms that listeners use to make such inferences. In the current study, we tested the viability of cultural consensus as the basis for assessing common ground, and compared language-based judgments of mutual knowledge from humans and large language models (LLMs).

To date, our best understanding of how communication partners reason about the states of the world comes from Rational Speech Act (RSA) models. This Bayesian framework describes recursive steps for producing and interpreting language by way of representing their partners' mental states (Goodman & Frank, 2016; Degen, 2023). Yet, there remain challenges with applying RSA-style descriptions to real-world conversations. RSA models are often instantiated with respect to communication within reference-resolution tasks, where the range of meanings are well defined and accessible. It is less clear how communication partners infer mutual knowledge when visual scenes do not ground utterances. Moreover, reliably estimating parameters within a Bayesian framework is data intensive (Vul et al., 2014; Dasgupta et al., 2018; Yung et al., 2021), but conversations vary substantially in their duration and information density (Mastroianni et al., 2021; Reese et al., 2023). Are there algorithms for assessing common ground that are sufficiently flexible across a variety of contexts (e.g., conversations among friends vs. strangers, short vs. long chats)?

A potential solution comes from the field of anthropology, which faces an analogous chicken-and-egg problem when assessing the cultural competence of previously undocumented social groups. Systematic responses from informants within a group can yield insights into meaningful cultural dimensions, but a priori it is unknown what questions will distinguish individuals in diagnostic ways. To solve this, Romney and Batchelder (1986, 1988) developed the cultural consensus framework, which uses agreement patterns across test items to infer culturally shared beliefs. For a given item (e.g., informant$_1$ judges statement$_1$ to be false), the anthropologist makes a joint inference about the informant's knowledge and its diagnostic relation to a target culture by computing the match between responses across informants. A central premise of cultural consensus is that systems of knowledge are not random in the world but are instead structured according to shared experiences within a social group (Medin et al., 2014; Shafto & Coley, 2003). These can be reliably evaluated through a cultural consensus

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

framework, such that within a given domain, individuals with aligned patterns of organizing concepts and classifying information have similar epistemologies.

The current study takes first steps in a research program to assess whether algorithmics akin to cultural consensus are useful for inferring common ground. These descriptions would advance our understanding of how common ground is inferred through language in non-referential contexts. Our hypothesis is that conversations entail a series of "test items," and for a given turn, listeners evaluate the extent to which a speaker's utterance is generated via systems of knowledge that are shared or different from their own. This makes two predictions. First, longer conversations and repeated interactions allow individuals to make more reliable assessments of their partners' knowledge. Second, individuals who make similar judgments about target utterances are more similar to each other than to those with different response profiles. If true, it provides an algorithmic basis for inferring common ground, and a mechanism for increasing alignment during communicative interactions (Clark & Marshall, 1981; Giles & Ogay, 2007).

Nevertheless, there are practical reasons why cultural consensus might not be useful for inferring common ground, particularly when conversations are brief and utterances are uninformative. For example, when talking to strangers, spoken utterances can underdetermine speaker knowledge since people are often nervous and goals are unconstrained (Keysar & Henley, 2002; Reece et al., 2023). Likewise, utterances frequently co-exist with style markers of cultural identity (e.g., talks like a woman, teenager) (Eckert, 2012), which may lead listeners to rely on inaccurate stereotypes, particularly when little else is known (Fuertes et al., 2012). Finally, similar to anthropologists and their informants, communication partners can come from different cultural backgrounds, and this can lead to miscommunication when individuals interpret their partners' utterances with respect to their own systems of knowledge, rather than appropriately inferring corresponding systems (e.g., "community of knowledge" - Sloman & Rabb, 2016; "double empathy problem"- Sasson et al., 2017). Thus, a priori, it is unclear what is the relevant benchmark for evaluating the efficacy of cultural consensus during communication.

We turn to LLMs as agents which imitate the content and style of human language, but do not produce spoken utterances in social contexts like humans do. Hence, compared to humans, LLMs may generate more accurate evaluations of common ground. In other examples of cultural assessments, LLMs exhibit high algorithmic fidelity in zero-shot settings, capturing demographic variation in political surveys (Argyle et al., 2022) and patterns of moral decision-making based on vignettes (Dillion et al., 2023). Likewise, LLMs approximate human performance across a variety of language-based pragmatic tasks, including reasoning about politeness, metaphor, persuasion, and discourse coherence (Ziems et al., 2023; Hu et al., 2023). Here, model size strongly correlates with task accuracy, suggesting that the number of parameters and volume of pretraining data is related to extracting reliable signals for knowledge-based inferences (Bowman, 2023). However, since pragmatic tasks vary substantially in their demands, it remains unclear how prior success generalizes to other phenomena, and whether LLMs succeed for reasons that are orthogonal to human performance.

As a first step, the current study evaluated cultural consensus by presenting language samples from speakers that varied along a cultural dimension (i.e., systems of knowledge relating to sports), and assessing categorization of speakers in an ABX task. On each trial, humans and LLMs saw two samples from speakers that differed in expertise (e.g., speaker A is a self-rated sports expert, speaker B is a self-rated sports novice) and were asked which one was more similar to a third sample, which was produced by either an expert or novice (X). If language-based cultural consensus generates reliable inferences of shared knowledge, then performance on this task should be greater than chance (50%). Moreover, if the accuracy of cultural consensus depends on cultural competence (e.g., sports knowledge), LLMs may outperform humans since their knowledge base is broader than any one individual's (Lederman & Mahowald, 2024). Alternatively, if common-ground assessments rely on functional competence of language use in social interactions, humans may outperform LLMs since they experience relevant processes within communication (Mahowald et al. 2023; Gordon & Van Durme, 2013). Finally, it is possible that spoken language is a noisy signal, and brief language samples provide insufficient information to infer common ground. In which case, humans and LLMs may both perform at chance.

Methods

Subjects and models

64 adults (ages 18+) were recruited through Prolific, and were paid $5 for their participation. Based on self-report, they were based in the US, primarily spoke English, and had >95% approval rating on Prolific. Data collection was reviewed and approved by the Institutional Review Board. We also tested two transformer-based models from OpenAI: gpt-3.5-turbo-0613 and gpt-4-0613. While the exact details of their architecture and training data are unpublished, GPT-3 models were built with 175B parameters and fine-tuned with reinforcement learning from human feedback (Hu et al., 2023; Ouyang et al., 2022), and GPT-4 models are an improvement upon prior versions (OpenAI, 2023).

Planet Word corpus

The materials came from a corpus of language samples produced by 359 individuals visiting the Planet Word Museum, in Washington, D.C. between June 2022 and November 2023. Research assistants stationed around the museum asked visitors if they would like to participate in a research study about expertise and language. After obtaining consent, visitors were asked to select a topic they knew a lot about from a list of 10 topics, such as cooking/baking, video games, music, and sports ("expert": self-rated 5.4 on a 7-pt

Table 1: Sample item in the cultural-consensus task

Task Instructions	Type	ID	Text Query
Imagine the first person says {text1}. And then a second person says {text2}. And then a third person says {text3}. Based on what they said, is the third person more similar to the first person or to the second person?	Expert	{text1}	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 instance in soccer you kick a ball around and have to kick it into a net to get um to get goals in basketball, you put a ball through a hoop to get points. Um These kinds of things, but other sports are like bike racing, where you're racing against other people or or running where you're racing against other people and in that case right, you're not sort of scoring points but you're trying to outperform your opponents. So I guess in summary, maybe sports are sort of physical activities where you outperform your opponents.
	Novice	{text2}	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 proficiency to be able to do it at a at a high level.
	Expert	{text3}	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 be D. One level at college that typically has the most time commitment. Um Often times D. One sports or things like football, soccer, lacrosse, field hockey. Um And then each school plays against other schools that are at that same D. One level. Then you go down to the middle. That would be the you've got you've got intramural club sport which is in the middle. Club sports are going to be sports. Typically college students practice 2 to 3 times a week. Um Club sports. There are some that are similar to D. One sports like soccer and swimming. Um But then you also have club sports like rugby and ultimate frisbee. And I'll circle back to ultimate frisbee in a second and then last but not least. You have intramural sports which typically you carry around 1 to 2 times a week and they're more in a tournament style.

scale). Next, they were asked to provide a 30- to 60-second spoken description of the topic based on the prompt "*What is ___?*". This prompt was selected after piloting as it was judged to yield general, knowledge-based descriptions from participants rather than personal experiences and anecdotes. Visitors then repeated the same procedures but for a topic that they knew little about ("novice": self-rated 2.1). Finally, they filled out a demographic survey about their age, race, gender identity, education level, country and/or state of residence, as well as what languages they knew fluently.

Within the corpus, visitors ranged in age from 5 to 84 years. To minimize effects of task familiarity, the order of presentation (expert vs. novice) was counterbalanced across visitors. Raw audio was recorded using Shure SM35 Performance Headset Condenser Microphones connected to iPads, and uploaded to Phonic, which provided automated transcriptions. For the current study, we focused on the "sports" topic, which yielded a mix of samples from experts (n=27) and novices (n=47). Seventy-four transcribed language samples ranged from 17 to 414 words. Length did not vary by expertise (p>.50).

Cultural-consensus task

Based on the language samples, we created a task that used zero-shot prompting to elicit model responses based only on the prompt and pre-training. This was closely aligned with the human task, wherein participants were given individual texts to compare with no previous or subsequent context available to update knowledge or expectations. Human data was collected in Qualtrics, while model prompts were supplied via OpenAI's API.

The structure with each trial was based on an ABX task (Table 1). Initial instructions were nearly identical for both humans and the models (see Table 1), but instructions for the models included an additional directive to output only a numerical response ("1" or "2") corresponding to their answer. This was added to preclude unexpected textual responses from the models. For human participants, "1" or "2" were the only two available multiple-choice response options. As an ABX task, each query consisted of three separate texts corresponding to transcribed language samples from our corpus. For each trial, at least one sample was that of an "expert" (A) a second that of a "novice" (B) and a third of either an "expert" or a "novice" (X). All participants completed 42 total trials.

To evaluate the extent to which judgments are based on properties of speakers, we manipulated dimensions that are known to affect the content and style of utterances. Half of the items asked for judgments of an expert speaker (i.e., X in the ABX task is an expert), while the other half asked for a novice. Thirty-six items featured samples from adult speakers (M = 33 yrs, SD = 14 yrs) and six items were from child speakers (M = 10 yrs, SD = 2 yrs). Age was manipulated across items to avoid confounds with expertise. Two presentation lists were created to counterbalance the order of trial presentation. Each list was randomized to ensure that no more than two expert or novice matches appeared in a row, as well as to ensure that items from different age groups were equally distributed across each half. Since there were more samples from novices compared to experts, samples occurred 1-3 times across items, within a list. Half of the human participants and half of all model prompts were presented with List A while the other half were presented with List B.

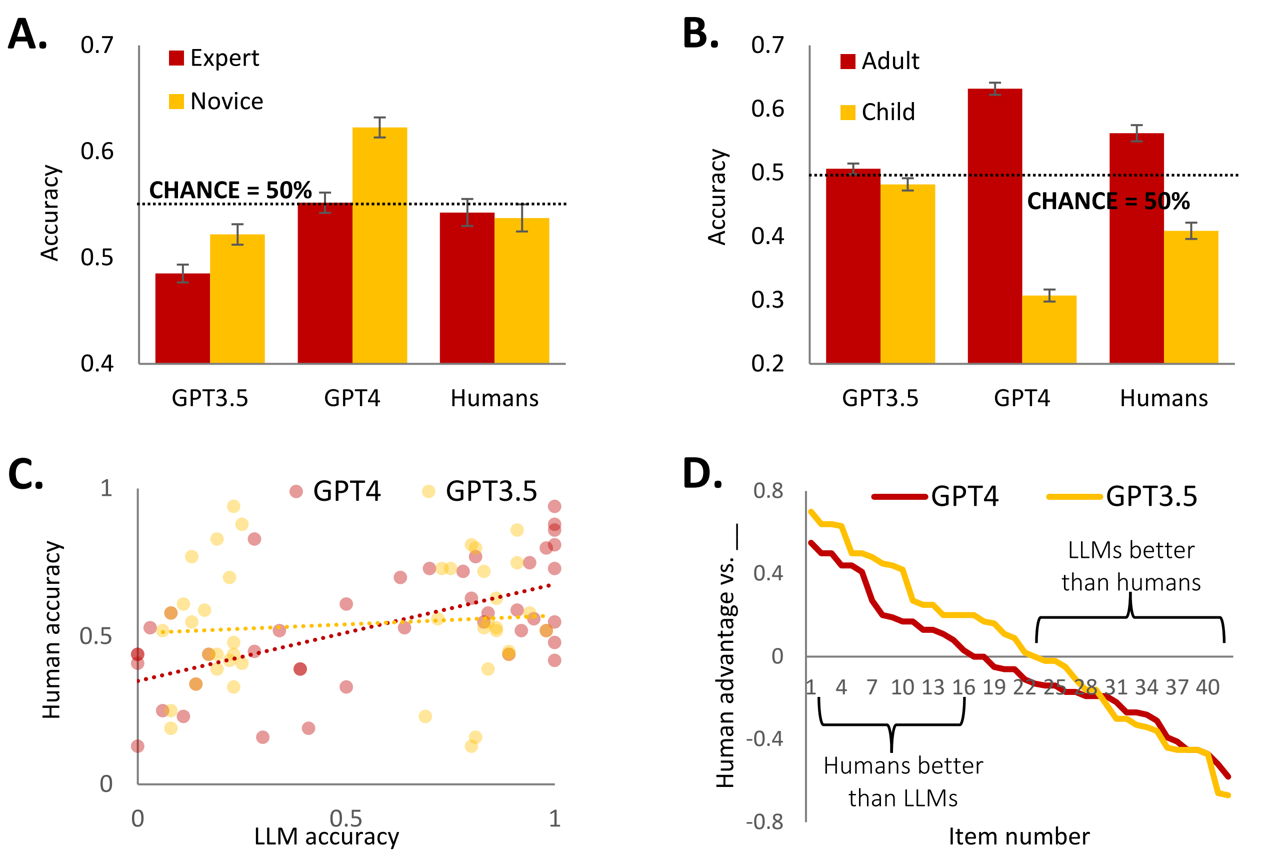

Figure 1: Accuracy in the cultural-consensus task. Chance performance is 50%. (A) Speaker-expertise effect, (B) Speaker-age effect, (C) Association between human and LLMs, (D) Relative advantage of humans vs. LLMs.

Results

We assessed language-based cultural consensus in two ways. First, we evaluated the accuracy of similarity-based judgments on group-level performance (humans, GPT-3.5, GPT-4), and examined the extent to which performance was affected by speaker properties such as expertise and age. Second, we evaluated performance at the item level, and examined the extent to which accuracy correlated across groups. For humans, we also evaluated the extent to which individual differences in subject-matter expertise conferred a task advantage. Finally, we examined possible task strategies that agents might have employed to make judgments, and evaluated the extent to which these account for the current data patterns.

Group differences

Accuracy ranged from 36% to 74% across all agents, with an overall mean of 54% (SD=7%). This confirms that language-based assessments of common ground in zero-shot settings is a difficult task. Accuracy was highest for GPT-4 (M=58%, SD=5%) followed by humans (M=54%, SD=7%) and GPT-3.5 (M=50%, SD=5%). This led to a main effect of group (F=31.92, p<.001). Planned comparisons revealed that task performance for humans was significantly different from GPT-3.5 (F=11.67, p<.01) and GPT-4 (F=17.54, p<.01). While humans and GPT-4 performed above chance (t's>4, p's<.001), GPT-3.5 did not (t<0.40, p>.60).

Follow-up analyses revealed patterns of alignment and divergence across agent performance. Fig. 1A illustrates that similarity-based judgments were more accurate when the target speaker (i.e., X in the ABX task) was a novice compared to an expert for GPT-3.5 (F=9.24, p<.01) and GPT-4 (F=25.56, p<.001). In contrast, humans did not perform differently when identifying novices versus experts (F=0.07, p>.70). Together, this generated a significant interaction between expertise and group (F=6.29, p<.01).

Fig. 1B illustrates that the accuracy of similarity-based judgments was higher for language samples from adult speakers compared to child speakers in humans (F=38.57, p<.001) and GPT-4 (F=259.58, p<.001). In contrast, GPT-3.5 performed similarly across the two (F=1.68, p>.20). Together, this generated a significant interaction between age and group (F=49.90, p<.001).

Item-level performance

Fig. 1C illustrates that item-level performance by humans and GPT-4 were associated. For a given item, GPT-4's success moderately predicted human success (r=.61, p<.001). In contrast, GPT-3.5's performance was neither associated with performance in humans (r<.15, p>.50) or GPT-4 (r<.10, p>.50).

Fig. 1D illustrates the relative advantage of humans compared to LLMs ordered by item-level accuracy. Items on the left of the graph are ones where human performance exceeds LLMs (above zero on the y-axis) while items on right are ones where LLMs show an advantage over humans (below zero on the y-axis). This graph suggests that the GPT-4's success on this task lies in its ability to minimize differences from humans on items that show a human-advantage compared to GPT-3.5. Likewise, GPT-4 shows a small but consistent advantage over humans across many items. On-going analyses predict features of expertise in language samples by training bag-of-words classifiers and fine-tuning BERT encoder models.

Finally, we asked humans to rate their own expertise on sports, and found a range within self-assessment (M=3.9, SD=1.3 on a 7-point scale). As a group, these individuals were less knowledgeable than experts in the corpus but more knowledgeable than novices (F=88.09, p<.001). Curiously, participants who rated themselves as knowing more about sports were *not* more accurate at distinguishing speakers in the current task (r<.01, p>.90). While the lack of association between self-rated expertise and speaker expertise is difficult to interpret, it may reflect in part the heterogeneity of systems of knowledge and presence of subcultures. We will return to this issue in the Discussion.

Ruling out task strategies

Since humans and LLMs frequently detect and leverage task regularities, we explored the possibility that low-level response strategies could explain the current patterns. First, we tested sensitivity to trial sequences. Recall that there were two lists (Lists A and B), which presented the same items but in reverse order. If judgments on earlier trials influenced performance on subsequent trials, we might expect accuracy to differ across lists. Instead, we found no effect of list or interaction with group (F<1.00, p>.30). Likewise, analysis of first- and second-half trials revealed no effect of half or interaction with group (F<1.00, p>.30).

Next, it is possible that humans and LLMs tracked the sequences of responses and developed switching rules to avoid repeating the same response in a row (e.g., no more than two "1" responses in a row). We coded trials based on whether the answers to the previous two trials were identical, and analyzed accuracy based on trials that did or did not follow repeated sequences. All agents showed greater accuracy following sequences compared to non-sequences, leading to a main effect of sequence (F=31.04, p<.001) but no interaction with group (F=1.64, p>.20). Thus, to the extent that this response strategy was useful, its benefits cannot account for variation in agent performance.

Finally, we returned to the LLMs' advantages for making judgments about novices over experts. While the numbers of trials were equated, one possibility is that this pattern is driven by a general response bias to assume that X is a novice. This would generate hits when X is a novice but also false alarms when X is an expert. We recoded responses in terms of matches to the novice, and calculated d-prime as hits minus false alarms. Values greater than 0 indicate that responses exceed chance guessing. This was true for humans (M=8%, t=4.59, p<.001) and GPT-4 (M=17%, t=13.61, p<.001) but not GPT-3.5 (0.1%, t=0.52, p>.60), leading to a main effect of group (F=31.92, p<.001). This suggests that even though GPT-3.5 and GPT-4 both show an advantage for novices, the basis for these effects may differ.

Discussion

The current study evaluated the extent to which linguistic signals within spoken utterances offer reliable cues for assessing common ground. Adopting a cultural consensus framework, we compared judgments of speaker similarity made by humans and LLMs. Accuracy varied substantially across agents, and was above chance for humans and GPT-4, but not for GPT-3.5. Moreover, humans and GPT-4 were similarly affected by the age and expertise of speakers, and their performance strongly correlated across test items. Together, the simplicity and flexibility of cultural consensus offer a potentially powerful algorithm for inferring common ground, providing a mechanism for evaluating mutual knowledge between communication partners in contexts where the space of possibilities is vast, non-referential, and opaque to strangers. Moreover, the high degree of cultural consensus between humans and GPT-4 presents promising avenues for using silicon samples to delineate pathways between cultural experiences and communicative interactions with precision.

The current findings address key limitations in implementing RSA models within real-world communication. While reliably estimating Bayesian parameters requires data that are abundant and informative, humans make near optimal decisions with very few samples across a variety of domains (Gigerenzer & Goldstein, 1996; Gershman et al., 2015; Yung et al., 2021). This paradox is solved by the fact that many real-world decisions do not require precise parameter estimates to distinguish action plans. In models of 2-alternative forced-choice tasks, Vul and colleagues (2014) found that the majority of decision-making accuracy is gained after the first sample. Moreover, the value of additional sampling depended on how costly it is to acquire this information and the penalties associated with inaccurate decisions. This has implications for understanding common-ground assessments during communication, and the range of strategies available when talking with strangers (Mastroianni et al., 2021; Reese et al., 2023). In particular, it suggests that the decision to stop vs. continue a conversation may depend on the information gained from first impressions, likelihood of future interactions, and extent to which common-ground errors are detrimental. Future research will investigate these dynamics by applying cultural consensus to multi-turn interactions.

The current findings also have implications for research in social science, which relies heavily on demographic variables as proxies of cultural background (e.g., race, gender, SES) (Argyle et al., 2022; Shaikh et al., 2023). While such approaches capture on-average differences across social *groups*, they do not offer sufficient precision for describing communication interactions between *individuals*. Since each person's lifetime experiences are an idiosyncratic mix of multiple cultures, this generates substantial individual variability in knowledge systems within demographic categories (Romney et al., 1998; Eckert, 2012). Moreover, since language is a communicative signal that transmits task-relevant thoughts, its style and content will vary depending on the needs of a context (Bell, 2001; Giles & Ogay, 2007). As an algorithm, cultural consensus is well equipped to handle this vast variability across individuals and contexts, since it treats common-ground assessment as an iterative process of inferring systems of knowledge from various signals, and applies computations that are applicable across situations (e.g., for a given utterance, do we have shared knowledge?).

Future research will examine the conditions under which bottom-up inferences from cultural consensus can be combined with top-down cues to systems of knowledge. Oftentimes, conversations with strangers are not random, and occur in the presence of informative cues such as community membership (e.g., meeting at church) or physical co-presence (e.g., waiting for EV chargers) (Clark & Marshall, 1981). To understand these dynamics, we can manipulate the extent to which the top-down situational goals are known and constrain inferences about shared knowledge. Likewise, within society, the generative engines of knowledge systems themselves are not random (Gordon & Van Durme, 2013; Lederman & Mahowald, 2024). For example, healthcare institutions invent language to describe new concepts (e.g., *paxlovid, covid-19*), and individuals acquire these systems by participating in institutions in different ways (e.g., scientists, doctors, patients). This suggests that common-ground inferences may be more efficient when they combine information from language in utterances with top-down knowledge of how systems of knowledge are generated. To understand this process, future studies can manipulate properties of institutions that are topically equivalent but vary in ways that are communicationally relevant. For example, since heterogeneous communities are made up of a collection of institutions (e.g., sports), we predict less common ground among individuals compared to homogenous communities, which comprise a narrower set of institutions (e.g., basketball).

Finally, the current findings are relevant for use-inspired applications of LLMs and human-centered AI. LLMs provide the architecture for state-of-the art chatbots, and current applications focus on improving the accuracy of text generation by harnessing regularity through more pretraining data or task-specific fine tuning. Importantly, communicative interactions occur between individuals and in service of specific goals, and understanding how mutual knowledge is inferred within conversational turns may support the development of technology to adapt to specific communication partners. Focusing on the role of culture may be a productive approach, since it describes systems of knowledge that are structured within social groups (Medin et al., 2014; Shafto & Coley, 2003) and yield representations that are isomorphic to language (Kemp et al., 2018; Lewis et al., 2023). Recent work in NLP demonstrates that including cultural dimensions increases the accuracy of machine translation and common-sense inferring (Hershcovich et al., 2022; Palta & Rudinger, 2023; Shaikh et al., 2023). Understanding how shared knowledge is inferred through language use may lead to the creation of technology that detects misalignments across partners and promotes mutual understanding in conversations.

Acknowledgments

This work has benefited from insightful conversations with Eusebia Mont and Charlotte Vaughn. We also thank the research assistants at the Language Science Station, Preethi Pai, and Daniel Borowski for their help with recruitment and data collection. This work was supported by NSF grants BCS-1844194 (Huang) and BCS-2116959 (Vaughn, Huang).

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 103-I 12). Greenwich, CT: JAI Press.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, 53(1), 71-92.

Bell, A. (2001). Back in style: Reworking audience design. In P. Eckert & J. R. Rickford (Eds.), *Style and Sociolinguistic Variation* (pp. 139-169). Cambridge: Cambridge University Press.

Bowman, S. R. (2023). Eight things to know about large language models. *arXiv preprint arXiv:2304.00612.*

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of Discourse Understanding* (pp. 10-63). Cambridge: Cambridge University Press.

Dasgupta, I., Schulz, E., Goodman, N. D., & Gershman, S. J. (2018). Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition*, 178, 67-81.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, 9, 519-540.

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27, 597-600.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41, 87-100.

Fuertes, J. N., Gottdiener, W. H., Martin, H., Gilbert, T. C., & Giles, H. (2012). A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology*, 42(1), 120-133.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245), 273-278.

Geurts, B. (2017). Presupposition and givenness. In Yan Huang (ed.), *Oxford Handbook of Pragmatics*, 180–198. Oxford, UK: Oxford University Press.

Gibbs, R. (1987). Mutual knowledge and the psychology of conversational inference. *Journal of Pragmatics*, 11(5), 561–588.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4), 650.

Giles, H., & Ogay, T. (2007). Communication accommodation theory. In B. Whaley & W. Samter (Eds.), *Explaining communication: Contemporary theories and exemplars* (pp. 293-310). Lawrence Erlbaum Associates.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Science*, 20(11), 818-829.

Gordon, J., & Van Durme, B. (2013, October). Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction* (pp. 25-30).

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., & Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of ACL* (Volume 1: Long Papers) (pp. 6997-7013).

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2023). A fine-grained comparison of pragmatic language understanding in humans and language models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of ACL* (pp. 4194-4213).

Kemp, C., Xu, Y., Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, 4, 109-128.

Keysar, B., & Henly, A. (2002). Speakers' overestimation of their effectiveness. *Psychological Science*, 13(3), 207-212.

Lederman, H., & Mahowald, K. (2024). Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of LLMs. *arXiv preprint arXiv:2401.04854*.

Lewis, M., Cahill, A., Madnani, N., & Evans, J. (2023). Local similarity and global variability characterize the semantic space of human languages. *Proceedings of the National Academy of Sciences*, 120(51), e2300986120.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2023). Dissociating language and thought in large language models: A cognitive perspective. *arXiv preprint arXiv:2301.06627.*

Mastroianni, A. M., Gilbert, D. T., Cooney, G., & Wilson, T. D. (2021). Do conversations end when people want them to? *Proceedings of the National Academy of Sciences*, 118(10), e2011809118.

Medin, D., Ojalehto, B., Marin, A., & Bang, M. (2014). Culture and epistemologies: Putting culture back into the ecosystem. In M. J. Gelfand, C. Chiu, & Y. Hong (Eds.), *Advances in Culture and Psychology* (pp. 177-217). Oxford Academic.

OpenAI. (2023). GPT-4 Technical Report. *https://arxiv.org/pdf/2303.08774.pdf*

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.

Palta, S., & Rudinger, R. (2023, July). FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 9952-9962).

Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A., & Marin, S. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science* Advances, 9(13), eadf3197.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, 88(2), 313-338.

Sasson, N. J., Faso, D. J., Nugent, J., Lovell, S., Kennedy, D. P., & Grossman, R. B. (2017). Neurotypical peers are less willing to interact with those with autism based on thin slice judgments. *Scientific Reports,* 7(1), 1-10.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in the natural world: Novices to experts, naive similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4), 641-649.

Shaikh, O., Ziems, C., Held, W., Pariani, A. J., Morstatter, F., & Yang, D. (2023). Modeling cross-cultural pragmatic inference with codenames duet. *arXiv preprint arXiv:2306.02475*.

Sloman, S. A., & Rabb, N. (2016). Your understanding is my understanding: Evidence for a community of knowledge. *Psychological Science*, 27(11), 1451-1460.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4), 599-637.

Yung, F., Jungbluth, J., & Demberg, V. (2021). Limits to the rational production of discourse connectives. *Frontiers in Psychology*, 12, 660730.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514.*