Humans Linguistically Align to their Conversational Partners, and Language Models Should Too

Rachel Ostrand [*1] **Sara E. Berger** [*1]

Abstract

Humankind has honed its language system over thousands of years to engage in statistical learning and form predictions about upcoming input, often based on properties of or prior conversational experience with a specific conversational partner. Large language models, however, do not adapt their language in a user-specific manner. We argue that AI and ML researchers and developers should not ignore this critical component of human language processing, but instead, incorporate it into LLM development, and that doing so will improve LLM conversational performance, as well as users' perceptions of models on dimensions such as accuracy and task success.

1. Introduction

Conversation is the new user interface, and is becoming the de facto mode of interaction between humans, systems, and applications – especially with the rapid emergence and uptake of large language models (LLMs) and their technical capability of generating (seemingly) fluid and well-articulated natural language in a conversational exchange. To date, most of the focus on foundation models and LLM development has been on the machine: technical and algorithmic progression such as improving the system's processing speed; increasing training data size, diversity, and quality; designing strategic and well-crafted system prompts to influence the model's behavior; quickly fine-tuning algorithmic outputs; and beating industry benchmarks. However, despite the fact that the users of LLMs are (at least for now) human, and thus have developed very specific cognitive and neural machinery for language processing over the course of human history, LLM research has included almost no consideration of the user's language processing expectations or predictions, or the conversational exchange itself.

Thus there is a major gap in the literature on understanding how humans conversationally interact with LLMs. Although language production and comprehension can seem effortless, they are in fact highly complex cognitive processes which humans can execute at an astonishingly fast rate. One of the mechanisms that enables people to process language so quickly is forming predictions about upcoming language likely to be produced by a conversational partner, and then using these predictions to inform future language production to them (Federmeier, 2007; Ferreira, 2019). However, there has been little focus in LLM research on either direction of this conversational adaptation (often referred to in psychology as *entrainment* or *alignment*): (1) what types of linguistic predictions human users engage in when interacting with an LLM, and how the LLM could take those predictions into account in its own language behavior, and (2) building models which linguistically, behaviorally, and "cognitively" adapt to their users.

In this position paper, we discuss the *what*, *why*, and *how* of linguistic adaptation and alignment as relevant to LLM development; in particular, that attending to human cognition and adaptation behavior during conversation is not merely the domain of psychologists, but also is vitally important for researchers and developers in AI, ML, and computer science. We explain *what* about the humans language processing system is tuned to make predictions and adapt to upcoming linguistic input, *why* these processes are important to consider when building generative AI-powered language models, and *how* these human linguistic processes should be implemented in AI and ML development pipelines. We conclude with future directions for research at the crossover of artificial and natural intelligence.

2. What: Effects of Models' Linguistic Behavior on User Cognition

LLMs are intrinsically interactive and the success and usefulness of any model is dependent not just on the computational components of the model's performance, but on the experience of the human user as well. In particular, it is important to take into account the user's cognitive and

[*]Equal contribution [1]IBM Research, Yorktown Heights, NY, USA. Correspondence to: Rachel Ostrand <rachel.ostrand@ibm.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

linguistic behavior and expectations towards the model as a conversational partner, including how those expectations might change over the interaction as the user learns about the model's behavior and abilities.

Humans are adept at engaging in statistical learning based on their prior experience in the world, and are able to generalize from past observations to make predictions about future events (and require orders of magnitude less data than most computational models do) (Kuperberg & Jaeger, 2016). Particularly relevant for LLM interaction, people make predictions about future language input based on past experience, by building a mental model of their conversational partner, or *interlocutor*.

When conversationally interacting with another human, people make predictions about upcoming language they expect to comprehend based on the group that their interlocutor belongs to, such as whether their interlocutor is an adult or a child (Van Berkum et al., 2008), a native or a non-native speaker of the language (Brunellière & Soto-Faraco, 2013), a speaker of a different regional dialect (Cai et al., 2017), or an expert or a novice in the field under discussion (Ryskin et al., 2019). When a conversational partner says something that violates those group membership-based expectations, such as a child talking about drinking wine or an American English-accented speaker using British English vocabulary, this causes processing difficulty for the listener. People also form rapid expectations and make linguistic predictions based on conversational experience they've gained from interacting with a *specific* conversational partner. These partner-specific linguistic expectations can change over time with added linguistic experience from that partner, in turn affecting the predictions that are made about that interlocutor's upcoming language (Brennan & Clark, 1996; Kraljic & Samuel, 2007; Trude & Brown-Schmidt, 2012; Yildirim et al., 2016). This linguistic expectation formation is unconscious and implicit, as humans are constantly and automatically learning these types of properties about their conversational partners, building mental models of a partner's linguistic abilities and idiosyncrasies, and updating the way they linguistically interact. A violation of linguistic predictions can cause confusion or misunderstanding, delays in comprehension (Metzing & Brennan, 2003; Brennan & Hanna, 2009), and increased processing difficulty (Ryskin et al., 2019; Martin et al., 2016; Kroczek & Gunter, 2021).

Importantly for the study of human-LLM conversational interaction, this linguistic adaptation behavior extends to conversation with non-human partners as well. In fact, interacting with a computer often causes people to form even stronger linguistic expectations compared to a human conversational partner, in large part because people expect computers and computational models to have lower language processing abilities compared to humans (Branigan et al., 2011). However, there is only limited research on linguistic alignment between humans and computers, and almost none with an LLM as a conversational partner (as opposed to a chatbot, spoken dialog system, or other type of much less conversationally-sophisticated robot).

Linguistic prediction and adaptation towards a computer can induce effects on numerous properties of the interaction, including: the user's language behavior itself; the user's perceived success of the collaborative task; higher user engagement; lower mental load; and a feeling that the model produced more accurate responses. One way in which adaptation to a computer conversational partner is manifest is that humans tend to reuse linguistic properties that were previously produced by the computer interlocutor, in a behavior known as linguistic *alignment* or *entrainment*. For example, people repeat the particular words that a chatbot produces (Ostrand et al., 2023; Parent & Eskenazi, 2010; Branigan et al., 2011), and the speech style and rate of a speechbot (Bell et al., 2003) When users align to (i.e., match) the computer's language properties in a conversation, it results in greater dialogue success (Lopes et al., 2013), and a reduction in speech and language behaviors that are difficult for the computer to understand (Fandrianto & Eskenazi, 2012). Additionally, in multi-turn conversations, people learn model-specific behaviors over the course of the interaction, and their linguistic expectations change. For example, a model which demonstrates relatively poor comprehension ability can induce users to repeat the model's own words more frequently (Ostrand et al., 2023).

In the opposite direction, people also expect that their conversational partner will modulate properties of their language production to converge upon the user's own, and the computer's ability to do so affects users' feelings about the success of the task and the model itself. When linguistic alignment is stronger, users perceive the interaction with the computer to have been more successful (Koulouri et al., 2016), and the conversational agent to be more competent (Nuñez et al., 2023). When interacting with a chatbot which aligns its own language production to match the user's, users report higher engagement (Spillner & Wenig, 2021). and lower mental load (Spillner & Wenig, 2021; Huiyang & Min, 2022). Similarly, when interacting with a spoken dialogue system, users perceive lower cognitive demand when the system linguistically aligns its replies to them (Linnemann & Jucks, 2018). Users also believe that a system or chatbot which linguistically aligns to themselves generates more accurate responses compared to one which does not linguistically align (Huiyang & Min, 2022).

But although humans engage in this automatic and unconscious linguistic adaptation during conversation, current language models do not take this behavior into account, and do not adapt their own behavior to better fit with the users' expectations.

3. Why: Why User Cognition is Important for AI Research

But artificial intelligence is not cognitive psychology, despite their historical links. Why, then, should AI researchers and developers care about cognitive properties and linguistic predictions of users?

Arguably the most important reason is an existential one: Even the best model which shows state-of-the-art performance on benchmark scoreboards is useless if no one wants to use it. The goal must also be to build models that people will engage with, trust, continue to use, and recommend to others. If researchers and developers do not investigate factors that influence users' perceptions of a model's conversational responses or task performance, it is harder to be sure that the model will work as intended. As noted above, most users of LLMs, as of this writing, are human, although this may change in the future. Humans like conversational partners who adapt to them and show social affiliation, more than partners who do not (Babel, 2010; 2012; Giles et al., 1991; van Baaren et al., 2003). People treat computers as social actors as well, applying similar interactional expectations and rules as they do with other humans (Nass & Moon, 2000), and thus it is reasonable to expect similar preference effects towards language models which linguistically adapt. In addition, as discussed in the previous section, when interacting with a computer interlocutor which adapts its language behavior to them, people feel the agent is more competent and more accurate and they had more success in performing the task, and are more engaged and require less mental load. These interactive properties are critical to a model's success and uptake: users will not continue to use models which require a lot of mental effort, appear unlikable, or they feel is incompetent or inaccurate. The information and behaviors a user expects from a model, and whether they adapt their behaviors to fit the model's language and the model adapts its language to the user's, can all influence the user's overall impressions of the model.

Second, there is an incredible opportunity to improve LLMs in new ways by specifically leveraging insights from the brain sciences. Understanding more about the mechanistic processes that drive human communication from a cognitive and behavioral perspective (and the associated heuristics, social paradigms, and expectations that come with it) provide a rich list of new features that machine learning engineers and developers could emulate in their models or use to evaluate model performance against, potentially leading to the formation of new benchmark metrics and tasks.

Third, incorporating knowledge about user cognition into the model-building process could allow for training or fine-tuning models better and more efficiently by employing human-centered data. Measurement of human cognitive factors and subtle behaviors during model interaction can create novel types of multimodal data which could be used alongside existing efforts to fine-tune models or even to train new models in novel ways. For example, one possibility would be to combine sensor or questionnaire-based datasets which query user's feelings towards the model with the text-based data from the conversational log. Even in a purely linguistic dataset, incorporating additional prosodic or paralinguistic data, such as audio recordings of speech which convey tone and prosody, can improve model performance (Sun et al., 2024).

Finally, learning and adapting to the user's cognitive and linguistic properties could create a more holistic and personalized experience for users, especially if the model can adapt to people fluently, contextually, and based on their unique cognitive, behavioral, and linguistic conversational features. (Of course, it should be noted that this may or may not be a desirable outcome, depending on the specific context or properties of a given model.) Understanding more about what properties of an AI agent humans value in their interactions, and what behaviors lead to higher trust of the model (Hauptman et al., 2022) could improve engagement and uptake of LLM-based systems in different contexts.

4. How: Practical Implementation for the AI and ML Communities

Given the importance of adapting to the user's language expectations, how should AI researchers and machine learning engineers develop their models in the future? The critical point is to conceptualize LLM language behavior and performance as a linguistic and social process, not just a technical and algorithmic one. Thus, LLM developers should incorporate linguistic adaptation into their model's behavior, which entails a few technical changes in model development.

First, it requires giving models the tools to store a user-specific linguistic profile, so that it can modulate its responses to align to the linguistic properties produced by that user during the interaction. Think of a user who only produces high-frequency, "easy" words (*tea*, *cheese*), simple grammatical structures (*I like tea*), and many spelling or word usage errors; these behaviors, especially if they remain consistent over the course of a conversation, could signify that the user has poor production and comprehension ability in the language that the conversation is conducted in. If the model's response to this user includes low-frequency, esoteric words (*tisane*, *astringent*), and complex grammatical structures like multiple center embeddings (*The tea the scientist drank oversteeped*), the user is likely to have difficulty understanding. Thus, models should be able to build and retain user-specific linguistic profiles.[1]

Second, the model should align the linguistic properties of its responses to the linguistic properties that the user produces. In the above example, when interacting with a user who produces only low-complexity vocabulary and syntax, the model should align its style and also produce low-complexity vocabulary and syntax. On a smaller scale, the model should try to reuse the same words that the user produced (rather than similar-meaning synonyms); a reasonable overlap in grammatical structures; and respond in the same *register* as the user's input (e.g., same level of language formality, dialect, or slang usage). In the case of an LLM-powered spoken dialogue system, the model should adapt acoustic and temporal properties of its speech (e.g., speech rate, vowel pronunciation, length of pauses) to be similar to the user's.

Finally, models should continually learn from a user's behavior and engage in linguistic adaptation over the course of an extended conversation with that particular user. One example would be detecting that the user produces a high rate of British English dialectal words and thus shifting to British English vocabulary and spelling. Especially in situations when a user engages in a multi-turn conversation or even multi-session conversation with a model, the model should not just blindly copy the linguistic properties of the immediately-preceding input; but rather, learn the user's linguistic statistics over time and gradually converge towards those linguistic properties.

5. Future Directions

Most language model alignment research is focused on integrating high-level ethical principles and legal guardrails into model instructions, system prompts, and other governance mechanisms to steer models to behave in ethical and unbiased ways, but largely without consideration of the linguistic or behavioral components of the interaction and their impact on the user experience. Linguistic alignment is critical to effective human-human communication, and plays an important role in the creation and maintenance of interpersonal relationships, and thus should be an inherent component of the research and technical objectives of cultural alignment, social value alignment, and accessibility considerations. In future work, it will be important to explore the relationship between linguistic alignment and cultural or value alignment displayed by a model, and how users' perceptions are affected both jointly and individually by these different types of model alignment.

Additionally, future work should investigate both the positive and negative ethical and social implications of cognitively- and linguistically-adaptive LLMs. In some respects, user-specific linguistic alignment could make large-scale models more beneficial to users. For example, *not* accounting for cognitive, behavioral, and linguistic differences via alignment has implications for model fairness and accessibility, as models which do not adapt to a user's linguistic or cognitive abilities may cause alienation or confusion, or even unintended discrimination based on a user's native language, including by enforcing assumptions of how a language exchange ought to proceed or culturally profiling the user (Kirk et al., 2024). On the other hand, user-specific data collection during interaction with a model can raise privacy and surveillance concerns (Ferrara, 2024; Friedland & Tschantz, 2019; Kirk et al., 2024), or could be used to manipulate or defraud users via spreading disinformation in a user's voice, sharing micro-targeted advertising campaigns which have been perfectly tailored for the user to be immensely susceptible to, or influencing users' opinions about sensitive topics without their explicit awareness (Jakesch et al., 2023). Moreover,"too much" user-specific adaptation or alignment could cause over-reliance or over-trust in these models (Kirk et al., 2024). Future research is needed to understand the likelihood and potential impact of these ethical issues to develop appropriate mitigative strategies while still making use of user-specific linguistic and cognitive adaptation.

6. Conclusion

In the recent explosion of research and development surrounding large language models, there is currently little focus on building models which linguistically adapt to individual users, despite this behavior being a critical component of human natural language processing "in the wild," and one which human brains have evolved over thousands of years to engage in and expect. Taking a user's linguistic and social expectations and cognitive machinery into account is critical to designing LLMs that interact with users in socially-positive ways – for example, by being trustworthy and reliable – rather than merely tinkering with surface-level technological interventions which address individual model or user experience symptoms instead of a foundational human requirement. We charge AI and ML researchers with incorporating real-time and adaptive cognitive, psycholinguistic, and behavioral features into LLM conversational interactions, to adapt artificial intelligence to the linguistic processes that natural intelligence has developed.

[1] It is important to note here that when we discuss user-specific personalization, we specifically and exclusively refer to abstract *linguistic* features. LLMs that are personalized in other dimensions are rare but slowly gaining ground; while there may be benefits to such models, there are also potentially major risks to privacy; dependency and over-reliance even to the point of addiction; over-trust; reinforcement of biases and polarization, etc. (Kirk et al., 2024).

References

Babel, M. Dialect divergence and convergence in New Zealand English. *Language in Society*, 39:437–456, 2010. doi: 10.1017/s0047404510000400.

Babel, M. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, 40 (1):177–189, 2012. doi: 10.1016/j.wocn.2011.09.001.

Bell, L., Gustafson, J., and Heldner, M. Prosodic adaptation in human-computer interaction. In *Proceedings of 15th International Congress of Phonetic Sciences*, pp. 2453–2456, 2003.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., and Brown, A. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, 121(1):41–57, 2011. doi: 10.1016/j.cognition.2011.05.011.

Brennan, S. E. and Clark, H. H. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6): 1482–1493, November 1996. doi: 10.1037/0278-7393.22.6.1482.

Brennan, S. E. and Hanna, J. E. Partner-Specific Adaptation in Dialog. *Topics in Cognitive Science*, 1(2):274–291, 2009. ISSN 1756-8765. doi: 10.1111/j.1756-8765.2009.01019.x.

Brunellière, A. and Soto-Faraco, S. The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and Language*, 125(1):82–93, 2013. doi: 10.1016/j.bandl.2013.01.007.

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., and Rodd, J. M. Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. *Cognitive Psychology*, 98:73–101, 2017. doi: 10.1016/j.cogpsych.2017.08.003.

Fandrianto, A. and Eskenazi, M. Prosodic Entrainment in an Information-Driven Dialog System. In *Proceedings of Interspeech 2012*, pp. 342–345, 2012. doi: 10.21437/Interspeech.2012-85.

Federmeier, K. D. Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, 44(4):491–505, 2007. doi: 10.1111/j.1469-8986.2007.00531.x.

Ferrara, E. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 2024. doi: 10.1007/s42001-024-00250-1.

Ferreira, V. S. A Mechanistic Framework for Explaining Audience Design in Language Production. *Annual Review of Psychology*, 70(1):29–51, 2019. doi: 10.1146/annurev-psych-122216-011653.

Friedland, G. and Tschantz, M. C. *Privacy concerns of multimodal sensor systems*, pp. 659–704. Association for Computing Machinery and Morgan & Claypool, 2019. doi: 10.1145/3233795.3233813.

Giles, H., Coupland, N., and Coupland, J. Accommodation theory: Communication, context, and consequence. In Giles, H., Coupland, J., and Coupland, N. (eds.), *Contexts of accommodation: Developments in applied sociolinguistics*, Studies in emotion and social interaction, pp. 1–68. Cambridge University Press, 1991. doi: 10.1017/CBO9780511663673.001.

Hauptman, A. I., Duan, W., and Mcneese, N. J. The Components of Trust for Collaborating With AI Colleagues. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW'22 Companion, pp. 72–75, 2022. doi: 10.1145/3500868.3559450.

Huiyang, S. and Min, W. Improving interaction experience through lexical convergence: The prosocial effect of lexical alignment in human-human and human-computer interactions. *International Journal of Human–Computer Interaction*, 38(1):28–41, 2022. doi: 10.1080/10447318.2021.1921367.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 2023. doi: 10.1145/3544548.3581196.

Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6:383–392, 2024. doi: 10.1038/s42256-024-00820-y.

Koulouri, T., Lauria, S., and Macredie, R. D. Do (and Say) as I Say: Linguistic Adaptation in Human–Computer Dialogs. *Human–Computer Interaction*, 2016. doi: 10.1080/07370024.2014.934180.

Kraljic, T. and Samuel, A. G. Perceptual Adjustments to Multiple Speakers. *Journal of Memory and Language*, 56(1):1–15, 2007. doi: 10.1016/j.jml.2006.07.010.

Kroczek, L. O. and Gunter, T. C. The time course of speaker-specific language processing. *Cortex*, 141:311–321, 2021. doi: 10.1016/j.cortex.2021.04.017.

Kuperberg, G. R. and Jaeger, T. F. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, 31(1):32–59, 2016. doi: 10.1080/23273798.2015.1102299.

Linnemann, G. A. and Jucks, R. ‘Can I Trust the Spoken Dialogue System Because It Uses the Same Words as I Do?’—Influence of Lexically Aligned Spoken Dialogue Systems on Trustworthiness and User Satisfaction. *Interacting with Computers*, 30(3):173–186, 2018. doi: 10.1093/iwc/iwy005.

Lopes, J., Eskenazi, M., and Trancoso, I. Automated two-way entrainment to improve spoken dialog system performance. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8372–8376, 2013. doi: 10.1109/ICASSP.2013.6639298.

Martin, C. D., Garcia, X., Potter, D., Melinger, A., and Costa, A. Holiday or vacation? The processing of variation in vocabulary across dialects. *Language, Cognition and Neuroscience*, 31(3):375–390, 2016. doi: 10.1080/23273798.2015.1100750.

Metzing, C. and Brennan, S. E. When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49(2):201–213, 2003. doi: 10.1016/s0749-596x(03)00028-7.

Nass, C. and Moon, Y. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1): 81–103, 2000. doi: 10.1111/0022-4537.00153.

Nuñez, T. R., Jakobowsky, C., Prynda, K., Bergmann, K., and Rosenthal-von der Pütten, A. Virtual agents aligning to their users. Lexical alignment in human-agent-interaction and its psychological effects. *International Journal of Human-Computer Studies*, pp. 103093, 2023. doi: 10.1016/j.ijhcs.2023.103093.

Ostrand, R., Ferreira, V. S., and Piorkowski, D. Rapid Lexical Alignment to a Conversational Agent. In *Proceedings of Interspeech 2023*, pp. 2653–2657, 2023. doi: 10.21437/Interspeech.2023-2332.

Parent, G. and Eskenazi, M. Lexical Entrainment of Real Users in the Let’s Go Spoken Dialog System. In *Proceedings of Interspeech 2010*, pp. 3018–3021, 2010. doi: 10.21437/Interspeech.2010-49.

Ryskin, R., Ng, S., Mimnaugh, K., Brown-Schmidt, S., and Federmeier, K. D. Talker-specific predictions during language processing. *Language, Cognition and Neuroscience*, 35:797–812, 2019. doi: 10.1080/23273798.2019.1630654.

Spillner, L. and Wenig, N. Talk to Me on My Level – Linguistic Alignment for Chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, MobileHCI ’21, 2021. doi: 10.1145/3447526.3472050.

Sun, X., Meng, H., Chakraborty, S., Bedi, A. S., and Bera, A. Beyond Text: Utilizing Vocal Cues to Improve Decision Making in LLMs for Robot Navigation Tasks. 2024. doi: 10.48550/arXiv.2402.03494. arXiv eprint: 2402.03494v2.

Trude, A. M. and Brown-Schmidt, S. Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, 27(7-8):979–1001, 2012. doi: 10.1080/01690965.2011.597153.

van Baaren, R. B., Holland, R. W., Steenaert, B., and van Knippenberg, A. Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, 39(4):393–398, 2003. doi: 10.1016/s0022-1031(03)00014-3.

Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., and Hagoort, P. The neural integration of speaker and message. *Journal of Cognitive Neuroscience*, 20(4):580–591, 2008. doi: 10.1162/jocn.2008.20054.

Yildirim, I., Degen, J., Tanenhaus, M. K., and Jaeger, T. F. Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*, 87:128–143, 2016. doi: 10.1016/j.jml.2015.08.003.