Distinct social-linguistic processing between humans and large audio-language models: Evidence from model-brain alignment

Hanlin Wu[1] **Xufeng Duan**[1] **Zhenguang G. Cai**[1,2]
[1]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
[2]Brain and Mind Institute, The Chinese University of Hong Kong
hanlin.wu@link.cuhk.edu.hk

Abstract

Voice-based AI development faces unique challenges in processing both linguistic and paralinguistic information. This study compares how large audio-language models (LALMs) and humans integrate speaker characteristics during speech comprehension, asking whether LALMs process speaker-contextualized language in ways that parallel human cognitive mechanisms. We compared two LALMs' (Qwen2-Audio and Ultravox 0.5) processing patterns with human EEG responses. Using surprisal and entropy metrics from the models, we analyzed their sensitivity to speaker-content incongruency across social stereotype violations (e.g., a man claiming to regularly get manicures) and biological knowledge violations (e.g., a man claiming to be pregnant). Results revealed that Qwen2-Audio exhibited increased surprisal for speaker-incongruent content and its surprisal values significantly predicted human N400 responses, while Ultravox 0.5 showed limited sensitivity to speaker characteristics. Importantly, neither model replicated the human-like processing distinction between social violations (eliciting N400 effects) and biological violations (eliciting P600 effects). These findings reveal both the potential and limitations of current LALMs in processing speaker-contextualized language, and suggest differences in social-linguistic processing mechanisms between humans and LALMs.

1 Introduction

Humans are remarkably adept at extracting speaker characteristics from vocal cues. Within milliseconds of hearing a voice, listeners can perceive a speaker's gender, age, health condition, personality traits, and other socio-demographic attributes (Lavan et al., 2024). The perceived speaker attributes then form a critical context for language comprehension, shaping how linguistic input is processed and interpreted (Wu and Cai, 2024a). For example, when we hear someone say "The first time I got *pregnant* I had a hard time," it is straightforward when coming from a female speaker but would be puzzling if a man were to say it.

Electroencephalography (EEG) studies show that when people hear sentences containing speaker incongruencies—such as "The first time I got *pregnant*..." spoken by a man (violating biological knowledge) or "I like to get *manicures*..." spoken by a man (violating gender stereotypes)—their brain responses diverge from speaker-congruent conditions, showing an N400 effect (Martin et al., 2016; Van Berkum et al., 2008; Van den Brink et al., 2012) or a P600 effect (Lattner and Friederici, 2003; Foucart et al., 2015). These neural responses show that speaker characteristics actively shape the real-time processing of spoken language.

The human capacity for speaker-contextualized language processing has recently been explained through a rational inference framework (Wu and Cai, 2024b). This framework proposes that humans engage in rational inference during real-time language comprehension—a process where listeners actively reason about the most likely interpretation given both linguistic input and speaker characteristics. Using social-stereotype violation (e.g., men getting manicures) and biological-knowledge violation (e.g., men getting pregnant) as test cases, they showed that when encountering speaker-content mismatches that violate social stereotypes, listeners can still arrive at a "literal" interpretation through effortful integration with their social knowledge, reflected in N400 effects. However, when faced with biological impossibilities, listeners rationally infer potential errors in the input and engage in error correction processes, manifested as P600 effects.

Recent advances in large language models (LLMs) have demonstrated increasing capabilities in contextual understanding (Zhu et al., 2024) and multimodal processing (Wang et al., 2024; Zhang et al., 2024). While initially focused on text, these models have expanded into multimodal tasks, show-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 135–143
May 3, 2025 ©2025 Association for Computational Linguistics

ing remarkable abilities in integrating inputs from diverse modalities like vision and speech. This evolution has led to the development of large audio-language models (LALMs) that can process audio inputs, including speaker characteristics, acoustic features, along with other contextual information.

The integration of LLMs into audio processing has progressed through several stages (Peng et al., 2024). Early attempts focused on incorporating Transformer architectures into traditional speech models, as exemplified by HuBERT's self-supervised learning on unlabeled speech data (Hsu et al., 2021). More recent approaches have shifted toward direct audio processing with LLMs by mapping audio features to tokens, not only for higher computational efficiency but also enabling richer paralinguistic processing through end-to-end multimodal integration (e.g., Chu et al., 2024).

This paradigm shift has produced models that are capable of increasingly complex tasks: AudioPaLM can preserve speaker voice characteristics during speech processing and generation (Rubenstein et al., 2023), SALMONN can perform audio-based storytelling and speech-audio co-reasoning (Tang et al., 2023), and Qwen2-Audio can explicitly identify speaker demographics and emotions (Chu et al., 2024). These emerging abilities raise questions about whether LALMs process speaker-contextualized language in ways that parallel human cognitive mechanisms.

As these models are increasingly deployed in interactive settings where they must interpret and respond to diverse speakers, understanding their social-linguistic processing has both theoretical and practical implications. On the one hand, comparing LALMs with human processing can provide insights into models' emergent cognitive mechanisms, an approach that has been widely used with deep neural networks (AlKhamissi et al., 2024, 2025; Schrimpf et al., 2018); on the other hand, identifying divergences between human and model processing helps pinpoint potential limitations in current architectures or training method, suggesting directions for developing more natural human-AI interactions.

To this end, we utilize computational metrics that have been shown to capture humans' real-time language processing. Specifically, surprisal (Hale, 2001; Levy, 2008), which reflects the unpredictability of a word given its context, has been linked to increased processing effort and has been shown to predict reading times (Smith and Levy, 2013) and N400 amplitudes (Krieger et al., 2024; Salicchi and Hsu, 2025). Entropy, which captures the uncertainty within the probability distribution of upcoming stimuli, was suggested to be associated with P600 amplitudes (Salicchi and Hsu, 2025).

In this research, we investigate whether LALMs align with human cognitive mechanisms in social-linguistic processing. We use the EEG data from Wu and Cai (2024b) as a benchmark of human processing and examine: a) whether LALMs align with humans in perceiving speaker characteristics and use them to guide real-time language processing; b) whether LALMs align with humans in the specific mechanism in processing speaker-content relationships.

2 Method

2.1 Human EEG data

The human data were EEG responses to speech stimuli from native Mandarin Chinese speakers. The study employed a 2×2 factorial design crossing Congruency (speaker-congruent vs. speaker-incongruent) with Type (social vs. biological). Congruency was manipulated by matching or mismatching speaker characteristics with the sentence content, while Type distinguished between violations of social stereotypes and biological knowledge. The experimental materials consisted of 80 self-referential sentences (each with a speaker-congruent and a speaker-incongruent audio version) in Mandarin Chinese, with speaker characteristics varying along gender and age dimensions (Table 1). All sentence audios were generated using text-to-speech technique with consistent acoustic properties.

The EEG data were collected from 60 participants while they listened to these sentences. A region of interest of 59 central-posterior sites was selected, and trial-level amplitudes were averaged across these sites before being further averaged over 300-600 ms (N400) and 600-1000 ms (P600) post-critical word onset. Their results revealed that social incongruency elicited a long-lasting N400 effect (across the 300-600-ms and the 600-1000-ms time windows), while biological incongruency elicited a P600 effect (600-1000 ms).

2.2 LALM metrics

We collected the computational metrics from two LALMs: Qwen2-Audio 7B Instruct (Chu et al., 2024) and Ultravox 0.5 8B (www.ultravox.ai). We

Category	Example	English translation
SM	在工作单位我一般都是穿西服打领带。	At the workplace I usually wear a suit and a tie.
SF	这个周末我要先去做美甲然后理发。	This weekend I'm going to get a manicure and then a haircut.
SA	我最近上班压力太大需要休息。	I've been working too hard lately and I need a break.
SC	他把我的玩具抢走了我要去找妈妈告状。	He took my toys away from me and I'm going to tell mummy about it.
BM	我需要定期去医院检查前列腺的健康状况。	I need to go to the hospital to check my prostate on a regular basis.
BF	我第一次怀孕的时候过得很艰难。	The first time I got pregnant I had a hard time.
BA	我发现我脸上的老年斑越来越多了我正在寻找新的治疗方法。	I noticed that I'm getting more and more age spots on my face and I am looking for new treatments.
BC	我在等我的乳牙掉下来然后我要把它扔到房顶上。	I'm waiting for my milk tooth to fall out and then I'm going to throw it on the roof.

Table 1: Examples of Stimuli used in Wu and Cai (2024b) with English translations. SM: socially congruent with male speakers; SF: socially congruent with female speakers; SA: socially congruent with adult speakers; SC: socially congruent with child speakers; BM: biologically congruent with male speakers; BF: biologically congruent with female speakers; BA: biologically congruent with adult speakers; BC: biologically congruent with child speakers. Critical words are underscored.

obtained the surprisal and entropy of the critical word through a sentence continuation task where we inputted the audio sentences that were cut short at the critical word following a text-based instruction to guide the model to continue the audio sentence by outputting text (see Appendix for prompts).

Surprisal was computed as the negative log probability of the target word given its context:

$$S(w_t) = -\log_2 P(w_t|C) \tag{1}$$

Where w_t represents the target word (i.e., the critical word that distinguishes speaker-congruent and -incongruent conditions); C represents the context before the target word, including the text-based instruction and the audio sentence; $P(w_t|C)$ was the word probability. For words containing multiple tokens, we calculated the joint probability at the token level.

Entropy was calculated over the probability distribution of the model's predictions at the target word position:

$$H(w_t) = -\sum P(w_x|C) \log_2 P(w_x|C) \tag{2}$$

Where w_x represents possible continuations. For words containing multiple tokens, we calculated the sum of the entropy for each token in the word. To test the generalizability across languages, we additionally created an English version of each sentence by translation and adaptation. The English audio was generated using the same standard as the Chinese audio. Metrics were collected for both the original Chinese stimuli and their English translations to test cross-linguistic generalization. We also collected these metrics from the text-based stimuli (the text transcription of those audio sentences) to serve as the baseline.

3 Results

We examined the model-brain alignment from two perspectives. First, we examined whether the LALM response patterns resembled humans by replicating the analyses in the human study on LALM data. Second, we examined whether LALM responses could predict human brain responses by including LALM metrics as additional predictors for the human brain responses. For all analyses, we used linear mixed-effects (LME) modeling with maximal random-effect structure determined by forward model comparison (α = 0.2, Matuschek et al., 2017). For surprisal and entropy analyses, we used item-level data and included the random effect of Item; for model-EEG alignment analyses, we used trial-level data and included the random effects of both Participant and Item.

3.1 Surprisal (Qwen2-Audio)

To test whether surprisal metric replicated the human brain pattern, we conducted LME analyses with Congruency (congruent = -0.5, incongruent = 0.5) and Type (social = -0.5, biological = 0.5) as interacting fixed effects, along with text-based surprisal as control, and showed a significant main effect of Congruency (β = 0.41, SE = 0.19, t = 2.12, p = .037) and text-based surprisal (β = 3.97, SE = 0.30, t = 13.17, p < .001), suggesting that

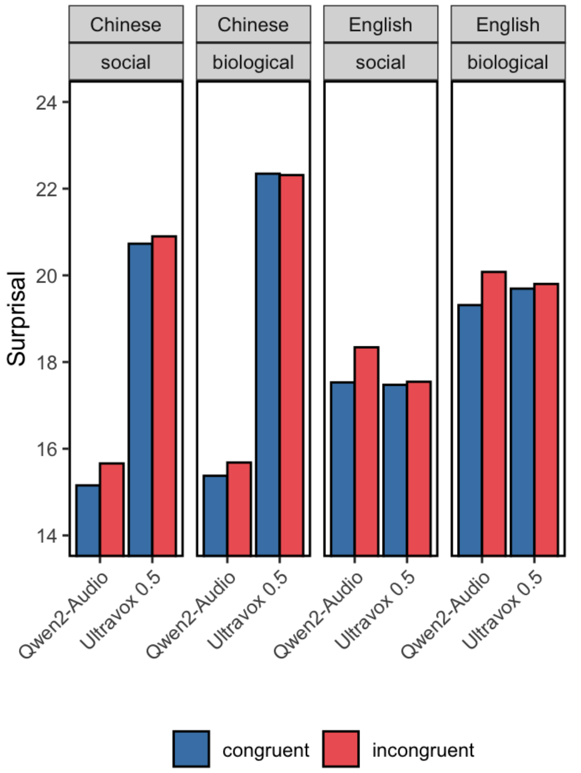

Figure 1: Surprisal values from Qwen2-Audio and Ultravox 0.5 models for speaker-congruent (blue) and speaker-incongruent (red) utterances, shown separately for social and biological conditions in Chinese and English.

the LALM model was sensitive to speaker-content incongruency regardless of violation type. The critical interaction between Congruency and Type was absent (β = -0.20, SE = 0.38, t = -0.52, p = .602), suggesting that unlike humans, the model processed social and biological violations similarly. The same pattern was replicated in English materials, as there was a significant main effect of Congruency (β = 0.73, SE = 0.20, t = 3.55, p < .001) and text-based surprisal (β = 3.89, SE = 0.32, t = 12.14, p < .001), while the interaction between Congruency and Type was absent (β = -0.17, SE = 0.41, t = -0.42, p = .678).

3.2 Surprisal (Ultravox 0.5)

Unlike Qwen2-Audio, the results for Ultravox 0.5 only showed a significant main effect of text-based surprisal (β = 4.09, SE = 0.36, t = 11.32, p < .001) for Chinese materials, while the main effect of Congruency (β = 0.07, SE = 0.08, t = 0.85, p = .399) or the interaction between Congruency and Type was absent (β = -0.20, SE = 0.16, t = -1.23, p = .222), suggesting that this model might not be sensitive to speaker-content relationships. The same pattern was shown in English materials, as there was a significant main effect of text-based surprisal (β = 2.41, SE = 0.42, t = 5.71, p < .001), and no main effect of Congruency (β = 0.05, SE = 0.10, t = 0.44, p = .663) or interaction between Congruency and Type (β = -0.05, SE = 0.21, t = -0.26, p = .799).

3.3 Entropy (Qwen2-Audio)

To test whether entropy metric replicated the human brain pattern, we conducted LME analyses with Congruency and Type as interacting fixed effects, along with text-based entropy as control, and showed that there was only a significant main effect of text-based entropy (β = 10.27, SE = 0.36, t = 28.49, p < .001) for Chinese materials. Neither the main effect of Congruency (β = -0.09, SE = 0.15, t = -0.61, p = .546) nor the interaction between Congruency and Type (β = -0.17, SE = 0.30, t = -0.55, p = .582) reached significance, suggesting that the model's uncertainty in prediction was primarily driven by the linguistic properties of the input rather than speaker-content relationships. The same pattern was shown in English materials, as the main effect of text-based entropy emerged (β = 8.71, SE = 0.42, t = 20.89, p < .001), while the main effect of Congruency (β = -0.04, SE = 0.17, t = -0.21, p = .836) and the interaction between Congruency and Type remained absent (β = -0.37, SE = 0.34, t = -1.07, p = .289).

3.4 Entropy (Ultravox 0.5)

The pattern observed in Qwen2-Audio was replicated with Ultravox 0.5, as there was only a significant main effect of text-based entropy (β = 9.50, SE = 0.23, t = 41.77, p < .001) for Chinese materials. Neither the main effect of Congruency (β = -0.01, SE = 0.02, t = -0.48, p = .630) nor the interaction between Congruency and Type (β = 0.04, SE = 0.05, t = 0.90, p = .369) reached significance. This pattern was further replicated with English materials, as there was only a significant main effect of text-based entropy (β = 3.63, SE = 0.70, t = 5.22, p < .001). Neither the main effect of Congruency (β = -0.03, SE = 0.04, t = -0.64, p = .526) nor the interaction between Congruency and Type (β = -0.05, SE = 0.08, t = -0.63, p = .531) reached significance.

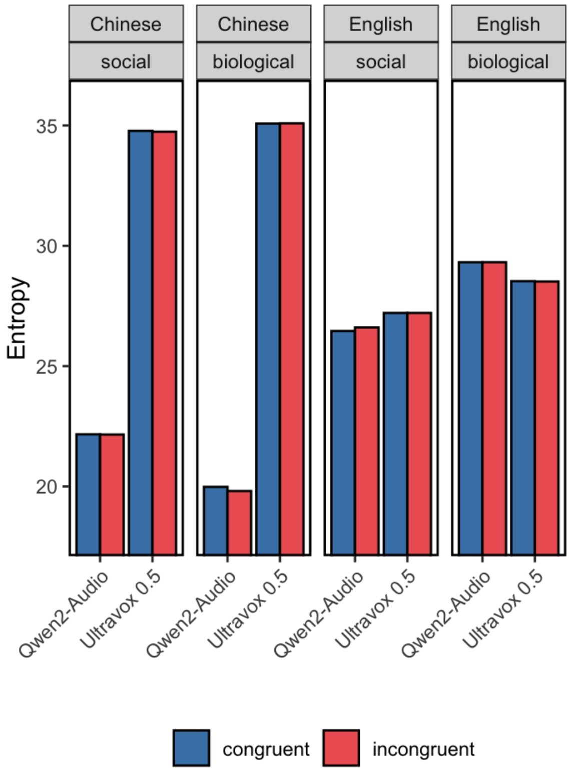

Figure 2: Entropy values from Qwen2-Audio and Ultravox 0.5 models for speaker-congruent (blue) and speaker-incongruent (red) utterances, shown separately for social and biological conditions in Chinese and English.

3.5 Surprisal-EEG alignment (Qwen2-Audio)

To test whether Surprisal could predict EEG response and whether the prediction varied across conditions, we added Surprisal (a scaled continuous variable) as a fixed effect interacting with Congruency and Type to the original LME analyses of EEG amplitude in Wu and Cai (2024a). For 300-600 ms, the results revealed a significant main effect of Surprisal (β = -0.50, SE = 0.16, t = -3.12, p = .002), while it did not interact with Congruency (β = 0.15, SE = 0.26, t = 0.57, p = .574), Type (β = 0.18, SE = 0.32, t = 0.56, p = .579), or the interaction between Congruency and Type (β = 0.17, SE = 0.52, t = 0.34, p = .736). For 600-1000 ms, there was no main effect of Surprisal (β = -0.22, SE = 0.18, t = -1.21, p = .229), or interaction with Congruency (β = 0.23, SE = 0.30, t = 0.76, p = .447), Type (β = 0.53, SE = 0.36, t = 1.50, p = .138), or three-way interaction with Congruency and Type (β = -0.09, SE = 0.60, t = -0.16, p = .877). These results suggested that surprisal significantly predicted N400 responses in a condition-independent manner, while it did not contribute to P600 responses.

3.6 Surprisal-EEG alignment (Ultravox 0.5)

For 300-600 ms, the results revealed a marginally significant main effect of Surprisal (β = -0.33, SE = 0.18, t = -1.79, p = .078), while it did not interact with Congruency (β = 0.33, SE = 0.28, t = 1.16, p = .250), Type (β = 0.23, SE = 0.37, t = 0.62, p = .539), or the interaction between Congruency and Type (β = -0.34, SE = 0.56, t = -0.60, p = .548). For 600-1000 ms, there was no main effect of Surprisal (β = 0.26, SE = 0.20, t = 1.26, p = .212), or its interaction with Congruency (β = 0.19, SE = 0.32, t = 0.61, p = .544), Type (β = 0.46, SE = 0.38, t = 1.22, p = .227), or the three-way interaction with Congruency and Type (β = -0.31, SE = 0.63, t = -0.49, p = .626). These results suggested that unlike Qwen2-Audio, Ultravox 0.5's surprisal did not reliably predict either N400 or P600 responses, despite showing a trend predicting N400.

3.7 Entropy-EEG alignment (Qwen2-Audio)

To test whether Entropy can predict EEG response and whether the prediction varied across conditions, we added Entropy (a scaled continuous variable) as a fixed effect interacting with Congruency and Type. The results revealed no significant main effect of Entropy (300-600 ms: β = -0.19, SE = 0.17, t = -1.14, p = .259; 600-1000 ms: β = 0.02, SE = 0.18, t = 0.12, p = .907), or interaction with Congruency (300-600 ms: β = 0.34, SE = 0.26, t = 1.31, p = .193; 600-1000 ms: β = -0.05, SE = 0.30, t = -0.18, p = .861), Type (300-600 ms: β = 0.44, SE = 0.34, t = 1.29, p = .202; 600-1000 ms: β = 0.36, SE = 0.37, t = 0.98, p = .333), or the three-way interaction with Congruency and Type (300-600 ms: β = -0.64, SE = 0.52, t = -1.24, p = .220; 600-1000 ms: β = -0.59, SE = 0.60, t = -0.98, p = .329). These results suggested that the model's predictive uncertainty did not predict human neural responses for either N400 or P600.

3.8 Entropy-EEG alignment (Ultravox 0.5)

The results revealed no significant main effect of Entropy in the N400 time window (300-600 ms: β = 0.04, SE = 0.18, t = 0.20, p = .844), but a marginal main effect in the P600 time window (600-1000 ms: β = 0.31, SE = 0.18, t = 1.76, p = .083). There were no significant interactions with Congruency (300-600 ms: β = 0.32, SE = 0.25, t = 1.28,

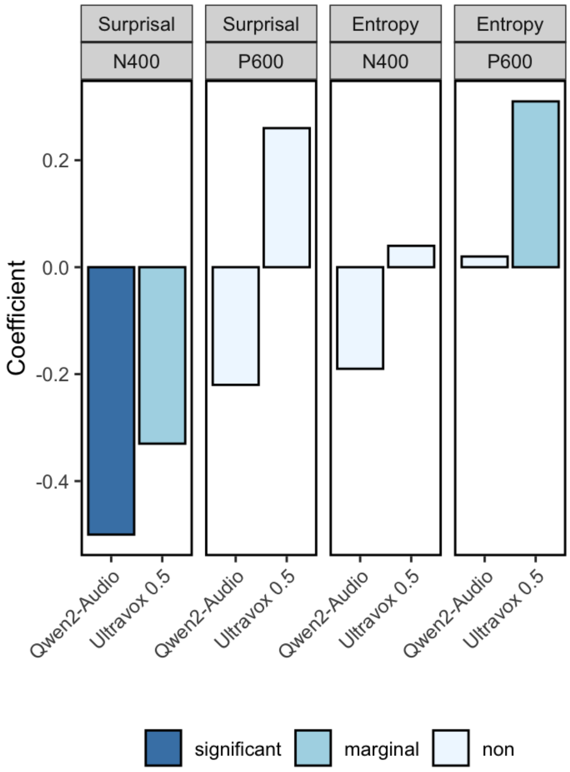

Figure 3: Main effect coefficients of Surprisal and Entropy on N400 and P600 amplitudes from LME analyses. Dark blue indicates a significant effect, light blue indicates marginal effects, and alice blue indicates non-significant effects for Qwen2-Audio and Ultravox 0.5 models.

p = .204; 600-1000 ms: β = 0.27, SE = 0.30, t = 0.89, p = .375), Type (300-600 ms: β = 0.11, SE = 0.34, t = 0.33, p = .744; 600-1000 ms: β = 0.17, SE = 0.36, t = 0.48, p = .633), or the three-way interaction with Congruency and Type (300-600 ms: β = -0.48, SE = 0.50, t = -0.96, p = .340; 600-1000 ms: β = -0.27, SE = 0.60, t = -0.45, p = .652). These results suggested that, similar to Qwen2-Audio, the model's predictive uncertainty did not strongly predict human neural responses for either N400 or P600, though there was a trend for higher entropy to predict larger P600 amplitudes.

4 Discussion

Our results revealed varying degrees of alignment between humans and LALMs in the social-linguistic processing of speech. Qwen2-Audio showed sensitivity to speaker-content incongruency through increased surprisal for incongruent utterances and significantly predicted human N400 responses. In contrast, Ultravox 0.5 showed no sensitivity to speaker-content relationships in its surprisal patterns and did not reliably predict human neural responses, despite showing a trend for N400.

Moreover, neither model showed human-like distinctions between social and biological violations, and both models' predictive uncertainty (entropy) was primarily driven by linguistic properties rather than speaker-content relationships and generally did not predict human neural responses, though Ultravox 0.5 showed a marginal trend for higher entropy predicting larger P600 amplitudes.

The distinct neural signatures for social versus biological violations in humans likely reflect different cognitive mechanisms. As Wu and Cai (2024b) suggested, social violations may be processed through semantic integration where linguistic content and speaker characteristics are integrated with prior knowledge about social roles and stereotypical behaviors, leading to N400 effects. In contrast, biological violations may trigger error detection and correction processes that attempt to resolve the physical impossibility, resulting in P600 effects. This distinction reflects rationality in human cognition.

Unlike humans who engage in active reanalysis when encountering biological impossibilities (reflected in the P600), current LALMs operate through single-pass forward prediction without mechanisms for backtracking or reanalysis. This may relate to the fact that current LLMs are typically trained to predict tokens one at a time, optimizing for local coherence rather than longer-range consistency. While some models are beginning to explore multi-token prediction windows (Gloeckle et al., 2024; Liu et al., 2024) that could theoretically capture longer-range dependencies and support reanalysis-like processes, most still lack similar mechanisms.

An open question is the precise mechanism by which LALMs utilize speaker information in their predictions. Unlike humans who readily identify speaker characteristics from voice and use this information to guide comprehension, it remains unclear whether LALMs explicitly represent speaker identity (e.g., assigning a gender category to a voice) or simply learn statistical associations between acoustic features and linguistic content. This distinction has implications for understanding both model processing and human cognition. For humans, the N400 and P600 effects depend on correctly identifying speaker characteristics and applying relevant world knowledge. If LALMs do not explicitly represent speaker identity but still show some degree of sensitivity to speaker-content relationships, this would suggest that explicit categorization may not be necessary for content prediction, though it might be essential for the rational inference processes that humans employ when resolving incongruencies. Future research could probe this question by examining model representations of speaker characteristics and their relationship to linguistic predictions.

Lastly, our findings also raise ethical considerations regarding LALMs' gender (and age) bias, which has been widely shown in LLMs (Kotek et al., 2023; Zhao et al., 2024). The observation that Qwen2-Audio showed increased surprisal for gender-nonconforming utterances indicates that it might have internalized societal gender stereotypes during training. While such sensitivity may facilitate natural interactions with humans, it also risks perpetuating harmful stereotypes if deployed in applications that influence decision-making or content generation.

In conclusion, we show that LALMs can potentially detect speaker-content violations and predict human N400 responses, but this capability varies between models. While Qwen2-Audio showed some level of alignment with human processing, neither Qwen2-Audio nor Ultravox 0.5 captured the human-like rational inference (as reflected by the distinction between social and biological violations), suggesting potential limitations in current LALM architectures or LLM architectures in general regarding real-time error analysis mechanisms.

5 Limitations

Several limitations of the current study should be acknowledged. First, our analyses focused on only two LALMs with a relatively small set of stimuli, which may not be representative of all current audio-language models or the full range of potential speaker-content relationships. A larger-scale investigation would better characterize the variation in speaker-content processing capabilities across different model architectures and training paradigms. Additionally, While surprisal and entropy are established metrics that have been linked to N400 and P600 responses respectively, they may be insufficient to capture the full range of processing distinctions that humans exhibit. Future research could explore alternative metrics such as analyzing activation patterns in different model layers, or utilizing representation similarity analysis between model embeddings and neural data. Finally, we only examined models' "static" responses to speaker characteristics, whereas humans show dynamic adaptation to individual speakers over increasing contexts (Grant et al., 2020). Human listeners rapidly adjust their predictions based on a speaker's established patterns—for example, becoming less surprised by stereotype-incongruent statements from a speaker who consistently violates stereotypes. This adaptive processing, which involves updating speaker models in real-time and adjusting predictions accordingly (Wu et al., 2025), represents an aspect of human language processing that our current single-utterance design cannot capture. Future work should examine how LALMs' predictions evolve across multiple utterances from the same speaker to better assess their capability for speaker-specific adaptation.

References

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2024. The llm language network: A neuroscientific approach for identifying causally task-relevant units. *arXiv preprint arXiv:2411.02280*.

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025. From language to cognition: How llms outgrow the human language network. *arXiv preprint arXiv:2503.01830*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Alice Foucart, Xavier Garcia, Meritxell Ayguasanosa, Guillaume Thierry, Clara Martin, and Albert Costa. 2015. Does the speaker matter? online processing of semantic and pragmatic information in l2 speech comprehension. *Neuropsychologia*, 75:291–303.

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. 2024. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*.

Angela Grant, Sarah Grey, and Janet G van Hell. 2020. Male fashionistas and female football fans: Gender stereotypes affect neurophysiological correlates of semantic processing during speech comprehension. *Journal of Neurolinguistics*, 53:100876.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Benedict Krieger, Harm Brouwer, Christoph Aurnhammer, and Matthew W Crocker. 2024. On the limits of llm surprisal as functional explanation of erps. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Sonja Lattner and Angela D Friederici. 2003. Talker's voice and gender stereotype in human auditory sentence processing–evidence from event-related brain potentials. *Neuroscience letters*, 339(3):191–194.

Nadine Lavan, Paula Rinke, and Mathias Scharinger. 2024. The time course of person perception from voices in the brain. *Proceedings of the National Academy of Sciences*, 121(26):e2318361121.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Clara D Martin, Xavier Garcia, Douglas Potter, Alissa Melinger, and Albert Costa. 2016. Holiday or vacation? the processing of variation in vocabulary across dialects. *Language, Cognition and Neuroscience*, 31(3):375–390.

Hannes Matuschek, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. Balancing type i error and power in linear mixed models. *Journal of memory and language*, 94:305–315.

Jing Peng, Yucheng Wang, Yu Xi, Xu Li, Xizhuo Zhang, and Kai Yu. 2024. A survey on speech large language models. *arXiv preprint arXiv:2410.18908*.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Lavinia Salicchi and Yu-Yin Hsu. 2025. Not every metric is equal: Cognitive models for predicting n400 and p600 components during reading comprehension. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3648–3654.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Jos JA Van Berkum, Danielle Van den Brink, Cathelijne MJY Tesink, Miriam Kos, and Peter Hagoort. 2008. The neural integration of speaker and message. *Journal of cognitive neuroscience*, 20(4):580–591.

Daniëlle Van den Brink, Jos JA Van Berkum, Marcel CM Bastiaansen, Cathelijne MJY Tesink, Miriam Kos, Jan K Buitelaar, and Peter Hagoort. 2012. Empathy matters: Erp evidence for inter-individual differences in social language processing. *Social cognitive and affective neuroscience*, 7(2):173–183.

Shuqi Wang, Xufeng Duan, and Zhenguang Cai. 2024. A multimodal large language model "foresees" objects based on verb information but not gender. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 435–441.

Hanlin Wu and Zhenguang G Cai. 2024a. Speaker effects in spoken language comprehension. *arXiv preprint arXiv:2412.07238*.

Hanlin Wu and Zhenguang G Cai. 2024b. When a man says he is pregnant: Erp evidence for a rational account of speaker-contextualized language comprehension. *arXiv preprint arXiv:2409.17525*.

Hanlin Wu, Xiaohui Rao, and Zhenguang G Cai. 2025. Probabilistic adaptation of language comprehension for individual speakers: Evidence from neural oscillations. *arXiv preprint arXiv:2502.01299*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *arXiv preprint arXiv:2403.00277*.

Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can large language models understand context? *arXiv preprint arXiv:2402.00858*.

A Appendix: prompts for sentence continuation task

A.1 Qwen2-Audio

Chinese materials (audio)

System: 你是一个实验中的参与者，你需要仔细听下面的录音。

User: 请补全录音中的句子，例如‘我喜欢吃‘，你可以回答‘苹果‘。直接回答补充的内容，不要说其他内容。录音：

User: (audio)

English materials (audio)

System: You are a participant in an experiment, you need to listen carefully to the following recording.

User: Please complete the sentence from the recording, for example if you hear ’I like to eat’, you can answer ’apples’. Just answer with the completing content, don’t say anything else. Recording:

User: (audio)

Chinese materials (text)

System: 你是一个实验中的参与者，你需要认真完成下面的任务。

User: 请补全以下句子。例如，‘我喜欢吃‘，你可以回答‘苹果‘，直接回答补充的内容，不要说其他内容。句子：(text)

English materials (text)

System: You are a participant in an experiment, you need to complete the following task carefully. User: Please complete the following sentence. For example, ’I like to eat’, you can answer ’apples’. Just answer with the completing content, don’t say anything else. Sentence: (text)

A.2 Ultravox 0.5

Chinese materials (audio)

System: 请补全录音中的句子，例如’我喜欢吃’，你可以回答’苹果’。直接回答补充的内容，不要说其他内容。(audio)

English materials (audio)

System: Please complete the sentence from the recording. For example, if you hear ’I like to eat’, you can answer ’apples’. Just answer with the completing content, don’t say anything else. (audio)

Chinese materials (text)

System: 请补全以下句子。例如，’我喜欢吃’，你可以回答’苹果’。直接回答补充的内容，不要说其他内容。句子：(text)

English materials (text)

System: Please complete the following sentence. For example, if you hear ’I like to eat’, you can answer ’apples’. Just answer with the completing content, don’t say anything else. Sentence: (text)