ORIGINAL ARTICLE

# Can language models trained on written monologue learn to predict spoken dialogue?

**Muhammad Umair[1*] | Julia Beret Mertens[2*] | Lena Warnke[2] | Jan P. de Ruiter[1,2]**

[1]Department of Computer Science, Tufts University, 177 College Ave., Medford, MA, USA, 02155

[2]Department of Psychology, Tufts University, 490 Boston Ave., Medford, MA, USA, 02155

**Correspondence**
Julia Mertens
Department of Psychology, Tufts University, 490 Boston Ave., Medford, MA, USA, 02155
Email: julia.mertens0@gmail.com

**Funding information**

Transformer-based Large Language Models (LLMs) have recently increased in popularity, in part due their impressive performance on a number of language tasks. While LLMs can produce human-like writing, the extent to which these models can learn to predict *spoken* language in natural interaction remains unclear. This is a non-trivial question, as spoken and written language differ in syntax, pragmatics, and norms that interlocutors follow. Previous work suggests that while LLMs may develop an understanding of linguistic rules based on statistical regularities, they fail to acquire the knowledge required for language use. This implies that LLMs may not learn the normative structure underlying interactive spoken language, but may instead only model superficial regularities in speech. In this paper, we aim to evaluate LLMs as models of spoken dialogue. Specifically, we investigate whether LLMs can learn that the *identity* of a speaker in spoken dialogue influences what is likely to be said. To answer this question, we first fine-tuned two variants of a specific LLM (GPT-2) on transcripts of natural spoken dialogue in English. Then we used these models to compute surprisal values for two-turn sequences with the same first-turn but different second-turn speakers and compared the output to human behavioral data. While the predictability of words in all fine-tuned models was influenced by speaker identity information, the models did not replicate humans' use of this information. Our findings suggest that although LLMs may learn to generate text conforming to normative linguistic structure, they do not (yet) faithfully replicate human behavior in natural conversation.

**KEYWORDS**
Generative pre-trained Transformers; Natural Language Processing; Language in Interaction

*The two lead authors contributed equally to this work.

# 1 | INTRODUCTION

Informal spoken conversation is one of the most ubiquitous ways through which we communicate with each other. In such conversations, participants alternate between speaker and listener roles, the assignment of which is locally managed by a well-documented set of rules (Levinson, 1983; Sacks et al., 1974). It is therefore tempting for dialogue researchers to model, analyze, and automate spoken dialogue with recently developed and highly effective transformer-based *Large Language Models* (LLMs).

LLMs have provided a breakthrough in modeling sequential dependencies in language (Vaswani et al., 2017), enabling these models to achieve human-like performance on various language tasks and quickly gain widespread popularity. For example, models such as GPT-4 Omni, Meta's Llama, and Google's Gemini now boast multi-modal processing capabilities as well as the ability to use voice to engage in back and forth conversations (Achiam et al., 2023; Touvron et al., 2023; Reid et al., 2024). These LLMs are increasingly being utilized in a diverse range of applications, such as assistance in academic and scientific work (Lund and Wang, 2023; Kung et al., 2023), influencing the media landscape (Cheng, 2024), and serving as programming assistants (e.g., OpenAI's Codex) (Finnie-Ansley et al., 2022). Modern LLMs are also continuously refined through various techniques (e.g., Reinforcement Learning from Human Feedback) to better align their responses with human preferences (Kirk et al., 2023). For example, LLMs are now capable of producing articles that are indistinguishable from those produced by humans (Kreps et al., 2022; Dou et al., 2022). Despite some evidence to the contrary, e.g., that LLMs simply mirror the intelligence of the interviewer (Sejnowski, 2023), these improvements have fueled speculation that LLMs might pass the Turing test i.e., their ability to generate human-like language implies an underlying intelligent thought process indistinguishable from that of humans (Mahowald et al., 2024).

Previous research into the ability of LLMs to replicate human language processing has yielded mixed results. On the one hand, current state-of-the-art language model outputs correlate with human neural data during language comprehension tasks. LLMs and human brains seem to predict words similarly from the preceding context: brain activity to specific words, as measured by a variety of neuroimaging techniques, is correlated with LLM-generated word surprisal (i.e. the probability of a word given its context) (Caucheteux and King, 2022; Michaelov et al., 2024; Caucheteux et al., 2021). In some cases, LLM generated lexical predictions more closely match human brain activity than predictions made by humans (Michaelov et al., 2022). These findings suggest that predictive processes underlying human language comprehension may be more reliant on the surface-level statistics of language than previously thought. When compared with human behavioral data, however, LLM performance diverges from that of humans. For example, LLMs with lower perplexity, a metric indicating a better fit to training data, actually provided a worse fit to human reading times (Oh et al., 2022; Oh and Schuler, 2023), suggesting that LLM surprisal estimates differ from human-like expectations. Furthermore, while LLMs rely on superficial statistical patterns in language, humans additionally draw on social norms, and reason about others' mental states, when producing language. LLMs fail to grasp reasoning functions and instead learn the statistical features of logical reasoning problems (Zhang et al., 2023; Mahowald et al., 2013). Similarly, LLMs do not perform as well as humans when asked to reason about the mental states of others, indicating that statistical learning from language may not be sufficient for belief attribution (Trott et al., 2023).

The extent to which LLMs replicate human language processing has been mostly studied in the context of monologic sentence comprehension experiments with models that, we assume, are trained primarily on written language (Dingemanse and Liesenfeld, 2022). These experimental contexts and the data on which LLMs are trained differ significantly from the way in which language occurs in natural conversation. It is therefore unclear how well LLMs generalize to spoken dialogue. In this context, it is useful to distinguish between *spoken-first* and *written-first* language. Spoken-first language is generated by a speaker (and then potentially transcribed), while written-first language is generated by an author (and then potentially converted to speech). These two modalities have different affordances. Writers have time to construct and revise their statements, while speakers have limited time to plan and produce their turns. Additionally, writers can rely on the fact that readers can re-read statements, while speakers must consider that listeners retain limited information. Speakers can also receive immediate feedback from listeners, whereas writers receive limited and delayed feedback from readers. Comparing written-first to spoken-first dialogue highlights key differences. For example, there is evidence that written responses are shorter and more diverse than spoken responses (Drieman, 1962). A comparative analysis of movie subtitles (written-first language) with natural dialogue (spoken-first language) found that written language has less frequent and more negative verbal feedback signals than spoken-first language (Pilan et al., 2024). This difference matters when fine-tuning large language models: when two conversational agents trained on subtitles were asked to interact, they produced feedback in rates and valence that matched the subtitle corpora instead of the pattern found in human interaction (Pilan et al., 2024). Even when trained on purely spoken dialogue, LLMs are able to mimic some paralinguistic features, such as silence and laughter, but lack the ability to consistently produce semantically coherent speech (Nguyen et al., 2023). Taken together, this research presents an important unanswered question: how well can LLMs predict language in the context of spoken dialogue?

One critical aspect of spoken conversations is the ability of listeners to identify who is speaking. We very rapidly incorporate speaker identity into the construction of meaning and the prediction of upcoming words (Van Berkum et al., 2008; Warnke and de Ruiter, 2023). While engaging in a conversation, humans draw on those lexical predictions to anticipate when a turn will end (De Ruiter et al., 2006; Magyari and De Ruiter, 2012), critically allowing them to begin their turn at the appropriate time. Warnke (2024) explicitly investigated listeners' use of speaker identity and preceding context to predict the end of an incoming turn. The author manipulated the plausibility of conversational turns by changing the speaker identity while keeping the linguistic content the same. Participants listened to two-turn sequences and pressed a button when they believed the second turn was going to end. The study found that participants took longer to anticipate the end of the turn in the conditions in which the speaker identity was manipulated compared to the congruent condition, suggesting that listeners use speaker identity to predict upcoming turns and their endings.

In the current study, we leverage the design and stimuli from Warnke (2024) to assess whether LLMs can replicate the human ability to use information about who is speaking to predict upcoming language in conversation. This comparison is crucial as it provides a benchmark for evaluating LLM performance against established human behaviors in dialogue processing. Specifically, we seek to determine whether LLMs produce higher surprisal values for turns spoken by incongruent speakers compared to congruent speakers, indicating an understanding of spoken dialogue structure. Additionally, we explore the impact of fine-tuning dataset size on model performance, and the influence of speaker representation (implicit vs. explicit) on LLM output. This comprehensive approach aims to provide insights into the capabilities and limitations of LLMs in predicting spoken dialogue, thereby bridging the gap between written and spoken language processing.

# 2 | METHODS

## 2.1 | Modeling

### 2.1.1 | Generative Pre-trained Transformer Models

Transformers provide a breakthrough in capturing *long-range* dependencies in language, achieving human-like performance on a variety of language tasks (Vaswani et al., 2017). This is largely enabled by their ability to use *attention* - a mechanism to recognize the relative importance of words in a *context* when predicting upcoming words. Attention, which can be of different types (e.g., multi-headed dot-product attention), has allowed transformers to use context more effectively and surpass the previous state-of-the-art (e.g., Recurrent Neural Networks) on text-based language modeling tasks (Karita et al., 2019). Additionally, word and positional embeddings are crucial components of transformers. Word embeddings convert words into multidimensional vectors, capturing their meanings based on context and relationships with other semantically similar words. Positional embeddings encode the position of each word in a sequence, allowing the model to maintain the order of words, which is vital for understanding syntax and meaning. Together, these embeddings enable transformers to process and interpret sequences of text effectively, capturing both the meaning of individual words and their arrangement within a sentence. While transformers were originally proposed as sequence-to-sequence models consisting of an encoder and decoder, modern LLMs typically use only the decoder component of the original architecture (Achiam et al., 2023). A defining characteristic of state-of-the-art LLMs is their autoregressive nature, meaning they only consider preceding words in the sequence when predicting the next word, similar to human language processing (Levinson and Torreira, 2015).

Due to their popularity and ability to harness vast amounts of data, there have been frequent advances in LLMs that have significantly improved their performance across tasks compared to earlier variants (Yang et al., 2024b). For instance, Google's LaMDA and Meta's Llama models have substantially increased in size and capability compared to earlier models (Touvron et al., 2023; Cohen et al., 2022), while OpenAI's GPT-4 has significantly more parameters than GPT-3 (Li et al., 2021). Some newer models (e.g., OpenAI's GPT-4) now include multi-modal processing capabilities, accepting image and text inputs and producing text output. These typically operate on large context windows, such as 8,192 tokens for GPT-4 (Achiam et al., 2023) compared to 1,024 for GPT-2 (Radford et al., 2019), that allows them to capture much longer range dependencies in the input (Guo et al., 2022). These improvements have allowed state-of-the-art LLMs to outperform their predecessors in almost all text-based language tasks.

Despite these advancements, we use OpenAI's GPT-2, a model with significantly fewer parameters and a smaller pre-training corpus compared to state-of-the-art LLMs (Radford et al., 2019) for several reasons. First, using a smaller model provides foundational information critical for understanding more complex models. Previous research indicates that larger and more complex models do not always lead to better performance across tasks (Gholami, 2024), and that larger training datasets can lead to diminishing returns (Shumailov et al., 2023). While larger LLMs may better learn formal competence (i.e., knowledge of linguistic rules, patterns, and norms), they often fail to achieve functional competence (i.e., the ability to use language in interaction), which depends on a host of non-linguistic capabilities that LLMs—regardless of size—struggle to achieve (Mahowald et al., 2024). Furthermore, state-of-the-art LLMs continue to hallucinate, reason poorly, and propagate bias when performing complex tasks (Achiam et al., 2023). Some studies even find that larger LLMs result in worse fits to human behavior (Oh et al., 2022). Thus, while it is possible that novel architectures and a greater number of parameters could enhance LLMs' functional competence—and by extension, their ability to model spoken language—there is also evidence suggesting that larger models do not necessarily lead to improvements in these areas.

Second, many state-of-the-art models, such as recent variants of GPT, are proprietary and not open-source, limiting their direct use in research (Liesenfeld and Dingemanse, 2024). These models do not provide direct probability or surprisal estimates for words; instead, such estimates must be measured indirectly by sampling sentence completions and analyzing the resultant distributions. In contrast, GPT-2 is open-source and integrated into popular machine learning libraries (e.g., huggingface (Wolf et al., 2020)), making it a practical choice for our study.

GPT-2 also requires fewer computational resources, making it more accessible for researchers without access to extensive computational infrastructure (Sathish et al., 2024). This allows for broader participation in research and easier replication of our results. We also make our methods[1] and data[2] open-source to and invite researchers to scale up this

[1]Implementation of the LLMs used in this work can be found here: https://github.com/mumair01/GPT-Monologue-to-Dialogue

[2]Fine-tuned models, inference results, and additional project data can be found here: https://osf.io/fxn8y/?view_only=9baf4033a2cb49cfaf107f9a753ab445

work to more complex models (e.g., Meta's LLama) as they become available.

Further, most LLMs do not explicitly encode speaker identities, which are key when predicting words in spoken language. Instead, they treat speaker identity labels in transcripts as any other token. This implicit speaker representation may cause GPT-2 to struggle to learn that speaker identities are not words, but qualities that are present and relevant over the course of an entire turn. Therefore, we use two variants of GPT-2 in this work: one with implicit (GPT-2) and another with explicit (TurnGPT) speaker representations (Ekstedt and Skantze, 2020). TurnGPT augments GPT-2 by adding a third type of embedding, in addition to word and positional embeddings, to explicitly represent the speaker of each word in an input sequence. It was originally designed to predict Transition Relevance Places (TRPs) i.e., points in a turn where interlocutors may, but do not need to, start speaking. Since TurnGPT requires additional special tokens to represent speaker identities, it must be fine-tuned to accurately use speaker identity information. This implies that there was no pre-trained only (or null) version of TurnGPT. See Appendix A for a detailed explanation of the fine-tuning and inference procedures used in this work.

### 2.1.2 | Fine-tuning

Transformer-based models have demonstrated high performance when learning new tasks due to their capacity for transfer learning. Under this paradigm, models are first pre-trained on large datasets with data-rich tasks (e.g., next-word prediction) in an unsupervised fashion. This pre-training allows the model to gain general-purpose domain knowledge, which can then be enhanced and applied to specific tasks by fine-tuning on smaller, task-specific datasets (Raffel et al., 2020; Brown et al., 2020). This process of pre-training and fine-tuning enables a language model to achieve state-of-the-art performance on numerous language benchmarks (VM et al., 2024).

Since we aim to investigate whether LLMs can generalize to natural spoken dialogue, we fine-tuned our models using transcripts of naturalistic conversations from the In Conversation Corpus (ICC)[3]. Each conversation in the ICC is approximately 25 minutes long and features a pair of undergraduate students. Participants sat in two sound-proofed rooms separated by a glass window, communicated using a microphone and headset, and were recorded on separate channels for complete sound isolation. In half of the conversations, the participants were recruited separately and were strangers, while in the other half, they were recruited together and knew each other.

We selected the ICC over publicly available dialogue corpora to maximize the naturalism and diversity of the turn-taking behaviors present in the fine-tuning data. While some open-source datasets are widely studied, well-annotated, and diligently transcribed, limitations of the data collection strategies affect the range and naturalism of behaviors exhibited during the interaction (Reece et al., 2023). For example, researchers often provide interlocutors with topics to elicit specific behaviors or to encourage more fluent conversation, which can limit the range of speech produced during the conversation.

To ensure linguistic consistency, we filtered the ICC and only selected conversations spoken in American English. The language – or, more precisely, the interactive style associated with culture – can affect some aspects of turn-taking. The culture-invariant components of turn-taking include the mechanisms speakers and listeners can use to take or pass on turns (Stivers et al., 2009), as well as general patterns in the timing of turns (Schegloff, 1982). However, the specific manifestation, frequency, and appropriateness of different conversational behaviors can change from culture to culture. For example, approximately 21% of Korean turns were continuers (e.g. "mhm"), in contrast to only 9% of English turns (Dingemanse and Liesenfeld, 2022). In addition, Korean backchannels were more often produced in overlap with a

[3]While the ICC is not publicly available due to restrictions imposed by the Tufts University's IRB regulations, it has been used in previously published research (Mertens, 2022; Warnke, 2022), and its protocol was reviewed and approved by the Tufts University's IRB before data collection. The Human Interaction Lab is actively working to meet IRB regulations to make the corpus publicly accessible in the future.

concurrent turn.

Our use of the ICC, a dialogue corpus, for fine-tuning LLMs requires qualification. First, the lack of transparency in state-of-the-art LLM training data raises concerns about data contamination and appropriate sources for fine-tuning (Balloccu et al., 2024). We assume most pre-training data for LLMs is *written-first* monologue data. Therefore, we fine-tune our model with *spoken-first* dialogue data for our dialogue-based task. If this decreases accuracy, it highlights the challenge of using monologue-trained models for dialogue tasks. The model may not be accustomed to dialogue structures, so adding dialogue data doesn't necessarily improve predictions in dialogue contexts (Yang et al., 2024a; Sun et al., 2024). Second, since the fine-tuning data were spoken in American English, the LLMs in this study may predict specific turns (e.g., backchannels) at different rates than if they were fine-tuned on other data. However, the stimuli in this study are simple, two-turn sequences without backchannels or timing information, so we assume the exact language will not affect our results. Fine-tuning an LLM on a limited set of data from an under-resourced language (of which there are many (Besacier et al., 2014)) might result in an LLM that can replicate words but not the interaction style necessary for effective communication in those languages. Future research should investigate how the languages and cultures in training data affect LLM behavior.

| | Five Conversations | | | Twenty-eight Conversations | | |
|---|---|---|---|---|---|---|
| | Turns | Words | Words/Turn | Turns | Words | Words/Turn |
| Speaker 1 | 1,975 | 12,924 | 6.54 | 10,691 | 84,901 | 7.94 |
| Speaker 2 | 1,916 | 15,884 | 8.04 | 10,412 | 85,076 | 7.96 |
| Speaker 1 / Speaker 2 | 1.03 | 0.81 | 0.81 | 1.03 | 1.00 | 1.00 |

**TABLE 1** Distribution of words and turns by speaker in the fine-tuning datasets (five vs. twenty-eight conversations). Note that while there are only two speaker labels (Speaker 1 and Speaker 2), each conversation features a unique set of participants.

Creating accurate and detailed verbatim transcripts of spoken dialogue is a notoriously painstaking and time consuming process (Tilley, 2003). Therefore, we investigated the amount of natural language data required for fine-tuning by using two datasets: one containing five conversations and another containing twenty-eight conversations from the ICC. We use 'Five' and 'Twenty-eight' to refer to these datasets in tables and figures. We transcribed an additional fourteen conversations for use as a validation set during fine-tuning. Note that the identity labels of Speakers 1 and 2 was effectively randomized between conversations (the first participant speaking in a conversation was labeled Speaker 1), and that each conversation featured a unique pair of interlocutors i.e., each interlocutor participated in exactly one conversation. To ensure that the two fine-tuning datasets did not substantially differ (which might affect model output), we analyzed the number of words, amount of turn-taking, and distribution of speaker transitions and holds in the datasets. First, to compare our datasets, we extracted all words from the training datasets to compare the frequency of words . Word frequencies were similar between the two datasets. The only notable difference between the two groups was the vocabulary size: the twenty-eight conversation dataset (3,886 words) was approximately 2.5 times larger than the five-conversation dataset (1,542 words). However, the words unique to the twenty-eight corpus (e.g., "exaggerate", "shady", "biography") composed only 8.90% of the total words spoken by interlocutors (see Appendix A for further details).

Further, we compared the number of words and turns by speaker for each dataset to determine whether the models would learn coincidental differences in the turns produced by each speaker. Table 1 shows that both speakers contributed approximately equal number of turns in both datasets. However, in the five conversation set, Speaker 2 produced 1.1 more words per turn on average than Speaker 1. In contrast, Speaker 1 and 2 produced roughly the same number of words in the twenty-eight conversation set.

| Datasets | Transitions | | | Holds | | |
|---|---|---|---|---|---|---|
| | % of Total Turn-Pairs | SP1 → SP2 | SP2 → SP1 | % of Total Turn-Pairs | SP1 → SP1 | SP2 → SP2 |
| Five | 81.42% | 1582 (40.71%) | 1582 (40.71%) | 20.32% | 390 (10.97%) | 332 (9.34%) |
| Twenty-eight | 79.83% | 8413 (39.92%) | 8413 (39.92%) | 22.26% | 2261 (11.85%) | 1988 (10.42%) |

**TABLE 2** Distribution of turn-pairs that have speaker transitions and or holds in each fine-tuning dataset (five versus twenty-eight conversations). The percentages reflect the percentage of total turn-pairs within a particular dataset.

We also investigated the ratio of speaker transitions to speaker holds in both ICC datasets. As shown in Table 2, most turns involved speaker transitions, while only 20% were speaker continuations. Although identifying speaker transitions in the corpus is straightforward, detecting continuations — when a speaker resumes after a meaningful contribution known as a *Turn Construction Unit* (TCU) — is much more complex. In the ICC, speaker continuations are based on silence thresholds between consecutive turns by the same speaker.

Despite minor differences between the training datasets, we are confident that they were sufficiently similar for fine-tuning our models. Note that we did not exclude any words (e.g., stop-words) from the fine-tuning datasets since we do not make any assumptions about the contribution of specific words to speaker-specific patterns. After training, we had five total models: GPT-2 and TurnGPT, both trained on the five and twenty-eight conversation datasets, and the pre-trained (or null) GPT-2 model.

## 2.2 | Human Experimental Data

For comparing the performance of the LLM models with that of human conversationalists, we used the stimuli and experimental setup from Warnke (2024). This study investigated whether listeners in dialogue can predict the speech act (or *illocution*, not to be confused with *sentence type*) of an upcoming turn. To assess the degree to which participants were able to predict the next turn, the authors leveraged the well-established task of turn-end anticipation (De Ruiter et al., 2006; Riest et al., 2015; Wesseling et al., 2006). In this task, participants listen to fragments of conversation and have to anticipate, either by button press or minimal vocalizations, when the turn they are listening to is going to end. In the study by Warnke (2024), the participants had to listen to two consecutive turns, and indicate per button press the end of the second turn. The authors' motivation for using this task was that it has been shown to require on-the-fly language prediction processes in human listeners, and the temporal difference between the actual turn-end and the anticipated turn-end gives a reliable estimate of the predictability (for human listeners) of the content of the turn (De Ruiter et al., 2006; Riest et al., 2015; Magyari and De Ruiter, 2008, 2012; Magyari et al., 2014; Magyari, 2022; Levinson, 2016).

The results showed that listeners more accurately predicted the ends of turns spoken by the "correct" speaker as compared to the "incorrect" speaker. This suggests that listeners use speaker identity representations to anticipate upcoming turns, as has been demonstrated in prior dialogue research (Warnke and de Ruiter, 2023; Metzing and Brennan, 2003), as well as the sentence comprehension literature (Van Berkum et al., 2008). We used these experimental stimuli to investigate whether LLMs can accurately emulate human dialogue behavior. In this section, we describe how we leverage the methods and measures from Warnke (2024) in the current study.

### 2.2.1 | Stimuli

Warnke (2024) found that listeners use both the preceding context and identity of the speaker of the current turn to predict upcoming speech. In their study, participants listened to two-turn sequences and pressed a button at the moment that they anticipated the second turn to end. This task has been shown to be sensitive to anticipatory processing in conversation (De Ruiter et al., 2006, see also discussion and references above). Stimuli belonged to one of six conditions in a two (speaker) by three (congruence) design, depending on the second turn in the two-turn sequence. The second turn differed in the identity of the speaker (same vs. different) and the plausibility of the turn by that speaker (congruent, incongruent, and violative). Congruent second turns were relatively plausible, i.e. spoken by the “correct" speaker. Incongruent second turns were not plausible given the preceding turn context and speaker identity. Specifically, they contained the same words as the congruent stimuli, except that they were spoken by the “wrong" speaker, which rendered them implausible.

***Congruent***

| | |
|---|---|
| **SP1:** Why'd you turn off the AC?<br>**SP1:** I'm hot | **SP1:** Why'd you turn off the AC?<br>**SP2:** Sam did |

***Incongruent***

| | |
|---|---|
| **SP1:** Why'd you turn off the AC?<br>**SP1:** Sam did | **SP1:** Why'd you turn off the AC?<br>**SP2:** I'm hot |

***Violative***

| | |
|---|---|
| **SP1:** Why'd you turn off the AC?<br>**SP1:** Yup | **SP1:** Why'd you turn off the AC?<br>**SP2:** Sounds nice |

**FIGURE 1** Example of congruent, incongruent, and violative stimuli used by Warnke (2024).

Figure 1 displays one six-stimulus group. All stimuli in the group had the same first turn “Why’d you turn off the AC?". “I’m hot" was congruent in the same-speaker condition (spoken by Speaker 1) and incongruent in the different-speaker condition (spoken by Speaker 2). Similarly, “Sam did” was congruent in the different-speaker condition (spoken by Speaker 2) and incongruent in the same-speaker condition (Speaker 1). The turns “Yup” and “Sounds nice,” in Figure 1 were violative since they were implausible regardless of the speaker. The advantage of this experimental design is that the congruence of the second turn changed while the linguistic content remained the same, thereby isolating the effect of speaker identity. The authors conducted an online plausibility norming study in which participants were asked to listen to each stimulus, and to rate how plausible it is that they would hear it in a conversation. Ratings were collected on a scale of 1 to 6 (1 for highly implausible and 6 for highly plausible), with 20 ratings per stimulus. The results confirmed that stimuli in the congruent condition were rated as more plausible ($M = 5.12$) than stimuli in both the incongruent ($M = 3.83$) and the violative conditions ($M = 2.11$). A Bayesian linear mixed effects regression revealed that the plausibility ratings were infinitely more likely under a model with condition as a fixed factor and random intercepts for both participants and items.

A variety of factors unrelated to congruence can affect the probability of words. Some first turns can highly constrain the second turn, while others allow for many possible responses. For example, "Do you mind helping me with my homework?" strongly projects either acceptance or rejection, whereas "You haven’t been answering any of my emails" could lead to various responses, including apologies, excuses, or denials. To control for this effect, the same first turn was used for every sequence in the same stimulus group. Additionally, words vary in frequency, with more frequent words (e.g., "I’m hot") being less surprising than infrequent words (e.g., "Sam did"). Therefore, the same second turn is used in both the congruent and incongruent conditions, with only the speaker identity changing. Finally, longer turns contain more information and generally result in lower probabilities overall. To minimize the effect of stimulus length, the second turn contains two syllables, resulting in turns of one or, at most, two words.

### 2.2.2 | Measures

In the current paper, we draw on model estimated surprisal values to compare model behavior to human behavioral data. Surprisal is a measure derived from the probability distribution produced by language models. According to surprisal theory, the difficulty of processing a word corresponds to its surprisal based on the context within which it appears; suprisal is therefore hypothesized to correlate with the cognitive load experienced by a comprehender (Hale, 2001; Levy, 2008). Although surprisal is a strong predictor of other metrics of processing difficulty, it is important to distinguish that it represents model-assigned probabilities, not direct measures of cognitive effort. Previous work, however, has shown surprisal to be an accurate predictor of cognitive load (Wilcox et al., 2020). Therefore, we analyze surprisal in this paper, defining it as the negative log probability of an event (Shannon, 1948). The less probable an event is, the more surprising it is, and the more information it contains.

$$Surprisal = -\log P(t_i \mid t_1, \ldots, t_{i-1}) \tag{1}$$

$$Surprisal^{word}_{secondTurn} = \sum_{i=1}^{N} -\log P(w_i^2 \mid w_1^2, \ldots, w_{i-1}^2, w_1^1, \ldots, w_K^1) \tag{2}$$

We calculate surprisal for the stimuli from Warnke (2024) based on the surprisal of individual words within turns. Formally, let $t_i \in V$ be a token that is defined in the vocabulary $V$ of a language model. Equation 1 shows the surprisal for a single token, which can be a word $w_i$, given all the previous words $(w_1, \ldots, w_{i-1})$ in a sequence. For a given sequence of words $S$, Equation 2 defines the of the second turn $Surprisal^{Word}_{secondTurn}$ in a two-turn stimulus (see Section 2.2.1) where the first turn has K words, denoted $w_1^1, \ldots, w_K^1$, and the second turn has N words, denoted $w_1^2, \ldots, w_N^2$. Here, the superscript represents the turn number and the subscript represents the position of a word in that turn. The $Surprisal^{Word}_{secondTurn}$ is then the sum of the negative log probability for each word in the second turn given all previous words in the second turn and the entire first turn. Note that the second turn in our stimuli can contain at most two words, N $\in \{1, 2\}$.

Finally, we compared these surprisal values to the data and analysis from Warnke (2024). In that experiment, the duration between the end of a turn and the button press was calculated into a variable called *bias* (calculated in milliseconds). To avoid confusion with the machine learning literature, where bias represents a systematic error, we refer to bias as *offset response time* (ORT) for the remainder of this work. A positive ORT indicates that participants pressed the button after the end of the turn, while a negative ORT indicates that participants pressed the button before the end of the turn. Results from Warnke (2024) show that ORT values are shortest for congruent turns, slightly longer for incongruent turns, and longest for violative turns. In other words, participants were more accurate at anticipating the end of the speaker's turn when the turn was congruent, as confirmed by offline plausibility judgements. The authors interpreted these results as demonstrating an effect of predictability: the more predictable a turn given the preceding context, the more accurate participants are at estimating its precise ending. This conclusion falls in line with prior literature showing that the predictability of the words in a turn affects turn-end anticipation timing (Riest et al., 2015; Magyari and De Ruiter, 2012). Though we do not have direct access to human predictability measures (e.g. cloze norms) for these conversational turns, we draw on the well-documented relationship between a turn's ORT and its linguistic content's predictability: the more predictable a turn's words, the shorter the ORT. We also draw on the relationship between plausibility and predictability: though plausibility and predictability are distinct constructs, implausible words and events are less predictable than plausible ones (Matsuki et al., 2011). Given these findings, we infer that ORT at least partially reflects language prediction processes in humans. Given that participants respond earlier to more predictable turns, a LLM that replicates human-like predictions should provide higher surprisal to turns with delayed participant responses.

### 2.2.3 | Analysis Plan

In this section, we highlight the various statistical analyses used to produce the results in Section 3. To estimate the random and fixed effects on LLM-produced surprisal values, we use *mixed effects regression* (Baayen et al., 2008), which accounts for hierarchical relationships within the data. Each stimulus group (as described in Section 2.2.1) contains six stimuli with the same first turn but a second turn with different speaker identity and congruence conditions. If the first turn (e.g., "Do you like my wonderful painting?") strongly projects a specific second turn (e.g., “Yes"), both humans and language models will be very surprised when the second turn does not match the first, regardless of whether the second turn is incongruent or violative. To account for these non-independent relationships, we included a random intercept per stimulus group to account for any baseline differences in surprisal. Where necessary, we performed follow-up, post-hoc t-tests to determine the source of the main effects. For example, a main effect of congruence could be due to a difference between the violation and congruent condition, the violation and incongruent condition, and/or the incongruent and congruent conditions. Without explicitly testing for these differences, the source of the effect remains ambiguous.

Further, to identify which predictors improved the regression performance, we created multiple regression models using the same surprisal data and compared them using *likelihood ratio tests*. Likelihood ratio tests allow us to compare the results from two statistical models, one with and the other without a target factor. If the data are more probable under model with the target factor, or if the model with the target factor has a statistically significantly better fit to the data, then the likelihood ratio test suggests that the factor improves the model.

We conducted all tests using both frequentist and Bayesian statistics[4]. Frequentist statistics provides easily computable, concrete thresholds for statistical significance based on p-values. In contrast, *Bayes Factors* evaluate the strength of the evidence for one hypothesis over another. Bayes Factors (specified as $BF_{10}$) indicate evidence for the alternative hypothesis ($H_1$) over the null hypothesis ($H_0$). We interpret Bayes Factors using evidence categories from Wetzels et al. (2011), adapted from Jeffreys (1939). Frequentist and Bayesian statistics often show the same directionality, but can differ in their estimated strength of the effect.

Finally, we use R syntax to describe regression models (such as in Table 3). The variable to the left of the ∼ indicates the outcome or dependent variable, in this case the surprisal of the LLMs. The ∼ indicates that the outcome variable is regressed on all the variables to the right. Random intercepts for the stimulus group are represented by (1 | Group). Congruence refers to the three-level categorical variable representing whether the second turn was congruent, incongruent, or violative, and Speaker refers to the two-level categorical variable indicating that the speaker of the second turn was the same (speaker hold) or different (speaker switch) than the speaker of the first turn.

For conciseness and clarity, we present the Bayesian results from the best regression models as determined using likelihood ratio rests in Section 3. Detailed results from all regression models are in their respective appendices. For brevity, we refer to regression models as *RMs* in the remainder of this text.

## 3 | RESULTS

### 3.1 | Effect of Congruence and Speaker

The predictability of a turn in natural spoken dialogue depends on the identity of the speaker. If LLMs learn to model the underlying structure of language in a similar way to humans, then we expect LLMs to be more surprised when the “wrong" speaker produces a turn compared to when the “right" speaker produces the same turn. In the human data, Warnke

[4]All Bayesian and frequentist statistics were conducted using the R packages lme4, lmerTest, BayesFactor, brms, and bayesTestR (Bates et al., 2015; Bürkner, 2017)

(2024) found only a main effect of congruence with no interaction between speaker and congruence, and no main effect of speaker. Therefore, we hypothesize that LLMs will find that:

**Hypothesis 1** *Incongruent second turns are more surprising than congruent second turns.*

**Hypothesis 2** *There is no main effect of speaker and no interaction effect between speaker and congruence on second-turn probabilities.*

To test these hypotheses, we used mixed-effects regression to model both the experimental effects and the random effect of stimulus group. As described in Section 2.1.2, we fine-tuned each LLM (TurnGPT and GPT2) on each dataset (five and twenty-eight conversations). Next, we identified which predictors improved the regression performance by creating and comparing five RMs for each LLM using the same surprisal data. We used likelihood ratio tests and Bayes factors to determine under which RM the data are most likely, and performed follow-up t-tests (both frequentist and Bayesian) to determine the source of the main effects where necessary (see Appendix B). Description 3 defines the best RM for each LLM, which includes main effects of speaker and congruence, along with interaction between speaker and congruence. In contrast, Warnke (2024) found that there was a main effect of speaker and congruence in humans but did not find any interaction effects (See Section 2.2.1).

$$Surprisal \sim Speaker * Congruence + (1|Group) \tag{3}$$

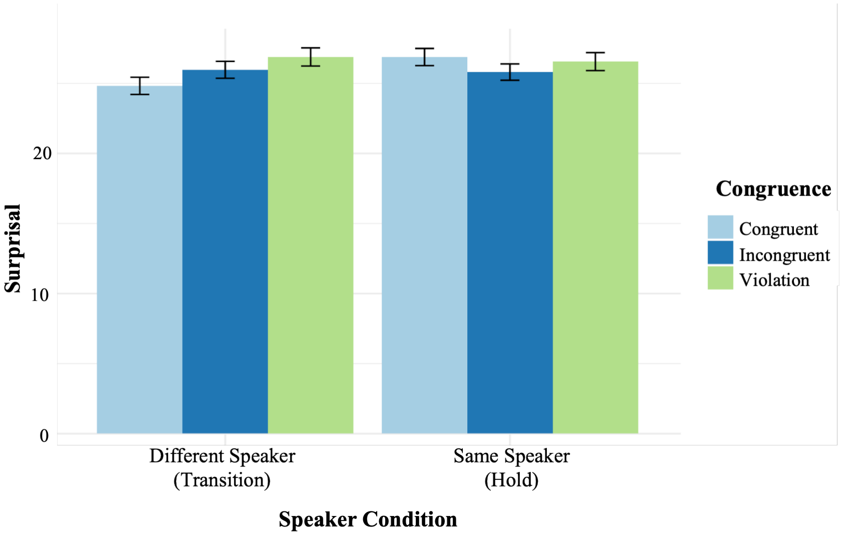


**FIGURE 2** Surprisal across congruence and speaker conditions for GPT-2 fine-tuned on twenty-eight conversations. The results indicate that the model aligns with Hypothesis 1 in the different speaker condition, but not in the same speaker condition.

As visualized in Figure 2, GPT-2 fine-tuned on twenty-eight conversations produced statistically significant differences in surprisal between congruent and incongruent conditions[5]. Contradicting Hypothesis 2, there was a main effect of speaker identity, where the same-speaker condition ($M = 26.35$) was more surprising than the different-speaker condition ($M = 25.40$) (see Table 6). We also found interaction effects between speaker and congruence that provided mixed support for Hypothesis 1. Specifically, we found substantial evidence that the incongruent stimuli ($M = 25.97$) were more surprising than the congruent stimuli ($M = 24.82$, $BF_{10} = 3.30$) within the different-speaker condition, supporting Hypothesis 1. However, we found anecdotal evidence for the opposite conclusion within the same-speaker condition: the congruent stimuli ($M = 26.88$) were *more* surprising than the incongruent stimuli ($M = 25.81$), which contradicts Hypothesis 1. Appendix B provides a more detailed description of these results.

## 3.2 | Effect of Amount of Fine-tuning Data

We investigated whether the amount of data used for fine-tuning LLMs (described in Section 2.1.1) affected surprisal values (see Figure 3). We suspected that fine-tuning LLMs on more conversations would result in surprisal values that more closely matched human responses and that RMs would find an interaction effect between the amount of fine-tuning and the effect of congruence. Specifically, we hypothesized:

**Hypothesis 3** *The difference in surprisal values for the incongruent (more surprising) and congruent (less surprising) stimuli will increase for the fine-tuned LLMs compared to the pre-trained-only model.*

**Hypothesis 4** *Increasing the amount of fine-tuning will result in diminishing returns.*

To explore potential interaction effects between amount of fine-tuning and congruence, we first concatenated the data i.e., used all of the surprisal values produced by GPT-2 models trained on no (pre-trained only), five, and twenty-eight conversations (see Figure 3). We excluded TurnGPT from this analysis since it must be fine-tuned on speaker identity information before it can be used to produce meaningful surprisal values. Next, we created a categorical predictor indicating the dataset used to fine-tune the language model and created five mixed-effects RMs (described in Table 7) that added predictors to the best RM (see description 3) identified in Section 3.1. Note that we created one frequentist and one Bayesian RM since we concatenated the output from each LLM. See Appendix C for a more detailed description of these models.

$$Surprisal \sim Speaker * Congruence * Dataset + (1|Group) \tag{4}$$

We found decisive evidence that the data (surprisal values) were most likely under a mixed effects model (see description 4) with a three-way interaction between speaker (same vs. different), congruence (congruent, incongruent, violative), and amount of data used for fine-tuning (five vs. twenty-eight conversations). The data were 17 times more likely under this model than the next most likely model. While the Bayesian and frequentist RMs had the same directionality for all effects, the frequentist likelihood ratio tests found that the best model included only a main effect of fine-tuning amount and no interaction effects with fine-tuning amount. Since frequentist statistics are less robust against low samples sizes than Bayesian statistics, this effect is likely due to the low sample sizes within each combination of factors. Therefore, we present results from the RM described above (see description 4), which included three-way interaction effects.

[5] We present results from GPT-2 fine-tuned on twenty-eight conversations for simplicity. All LLMs showed the same effect and had the same best RM (see description 3).

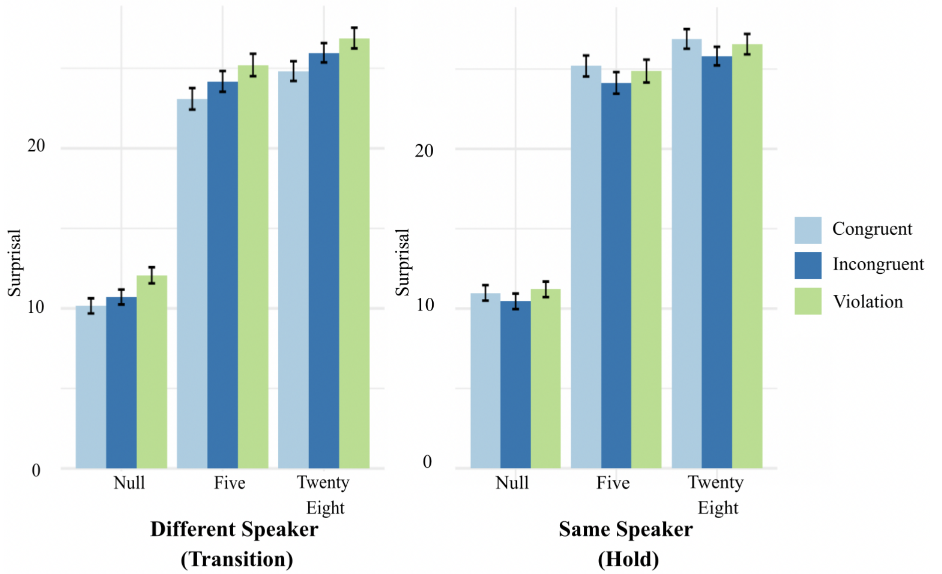


**FIGURE 3** Surprisal values for GPT-2 models fine-tuned on different amounts of data: none (null), five conversations, and twenty-eight conversations. The figure demonstrates that the baseline surprisal increases at a decreasing rate as the amount of fine-tuning data increases.

Interestingly, GPT-2 fine-tuned on five ($\beta_5 = 12.94$) and twenty-eight ($\beta_{28} = 14.67$) conversations produced overall higher surprisal values than the null (pre-trained-only) GPT-2 model. In addition, the fine-tuned GPT-2 models produced slightly higher surprisal values than the null (pre-trained-only) model for the incongruent ($\beta_5 = 0.53, \beta_{28} = 0.59$) and violation ($\beta_5 = 0.20, \beta_{28} = 0.16$) conditions. This supports Hypothesis 3, since fine-tuning models resulted in a larger increase in surprisal for the incongruent stimuli compared to the congruent stimuli. Additionally, the difference in surprisal between the models fine-tuned on twenty-eight and five conversations was much smaller than the difference in surprisal between the models fine-tuned on five conversations and the null (pre-trained-only) model, which supports Hypothesis 4. However, this regression also found a number of unexpected results. Specifically, GPT-2 fine-tuned on five ($\beta_5 = 12.94$) and twenty-eight ($\beta_{28} = 14.67$) conversations produced much higher surprisal values compared to the null (pre-trained-only) model and was more surprised by the same speaker stimuli ($\beta_5 = 1.26, \beta_{28} = 1.23$).

## 3.3 | Explicit Versus Implicit Speaker Representation

As described in Section 2.1.1, GPT-2 encodes words and their relative positions while TurnGPT additionally explicitly adds embeddings that encode speaker identity. It may be that providing speaker identity information to GPT-2 – similar to how humans hear the voice (and therefore can assess the identity) of their interlocutor in every word – would influence the models' ability to be appropriately surprised in the context of spoken language.

**Hypothesis 5** *Models with explicit speaker representation will more strongly distinguish between congruence conditions compared to models with implicit speaker representation.*

To investigate the effect of speaker representation on the LLMs' ability to model spoken dialogue, we first concatenated data (surprisal) produced by GPT-2 and TurnGPT, both trained on twenty-eight conversations, assuming that models

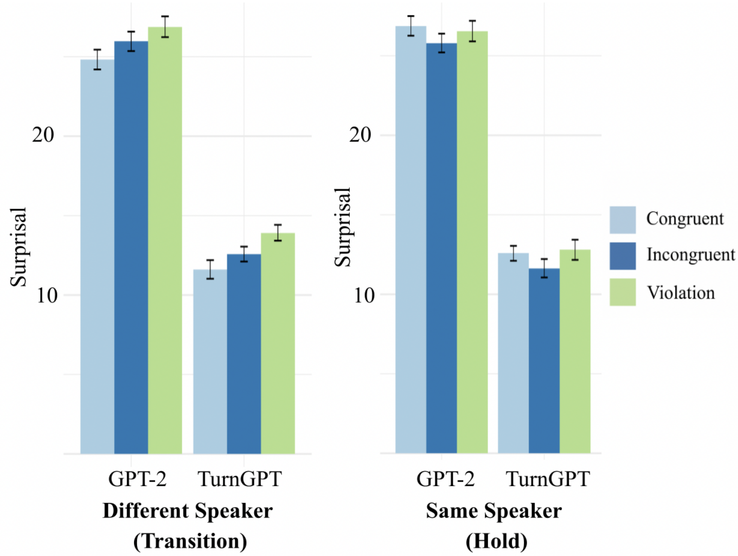


**FIGURE 4** Effect of speaker representations (GPT-2 vs. TurnGPT fine-tuned on twenty-eight conversations) on surprisal for different and same-speaker stimuli. Note that the baseline surprisal values significantly differ based on the model type (TurnGPT vs. GPT-2).

fine-tuned on a greater number of conversations will more closely match human behavior. Next, we created a categorical predictor indicating the model type (implicit vs. explicit) and created five mixed-effects RMs (described in Table 10) that added predictors to the best RM (see description 3) identified in Section 3.1. As with the analyses conducted in Section 3.2 (to explore the effect of fine-tuning amount), Bayesian analysis found decisive evidence ($BF_{10}$ = 293.82) that the data were most likely under the model that included all two-way interaction effects and a three-way interaction effect between speaker, congruence, and model type (see description 5).

$$Surprisal \sim Speaker * Congruence * Model + (1|Group) \qquad (5)$$

However, frequentist likelihood ratio tests suggested that the interaction effects did not improve model performance (RM 11 in Table 10). Given that the Bayesian and frequentist coefficients pointed in the same direction, and that frequentist analyses are less robust against lower sample sizes, we present the results from the best RM as determined by Bayesian analyses in this section (see Appendix D for a description of all other RMs). As shown in Figure 4, TurnGPT produced surprisal values that were less affected by incongruence values than GPT-2 ($\beta$ = -0.18), contradicting Hypothesis 5. Interestingly and unexpectedly, TurnGPT produced much lower surprisal values overall ($\beta$ = -13.21) and was less surprised by the same speaker condition ($\beta$ = -1.09).

## 3.4 | Predicting End-of-Turn Response Times

In Sections 3.1, 3.2, and 3.3, we analyzed patterns in surprisal values produced by LLMs to turns that ranged in their speaker and congruence, as judged by humans. We found that the LLMs produced expected surprisal patterns in the different-speaker condition, but unexpected surprisal patterns in the same-speaker condition. A stronger test of language model performance is to directly compare model surprisal values with the human behavioral data from Warnke (2024). In their study, the authors calculated offset response time (ORT) as the duration between the end of the second turn and participants' button press (See Section 2.2.1), and found that ORT was dependent on congruence: ORT was largest for the violation condition and shortest for the congruent condition. Here, we investigate whether the model-estimated surprisal values predict human ORTs.

**Hypothesis 6** *Turns with higher ORTs (indicating later end-of-turn anticipation by humans) will exhibit higher surprisal values.*

To investigate this hypothesis, we generated a baseline model (Equation 6), and determined whether the data were more likely under the model that included surprisal as an additional predictor (Equation 7). Below, we present the results of TurnGPT trained on twenty-eight conversations. We chose to use only TurnGPT for the current analysis because it explicitly represents speaker identity, therefore capturing information that humans also have access to.

$$ORT \sim Speaker * Congruence + (1|Group) + (1|Participant) \tag{6}$$

$$ORT \sim Speaker * Congruence + Surprisal + (1|Group) + (1|Participant) \tag{7}$$

We found strong evidence that the data were more likely under the model that included surprisal as a predictor ($BF_{10}$ = 11.27) – but in the opposite direction as stated in Hypothesis 6. Surprisal was *negatively* associated with ORT ($\beta$ = -0.04, $t$ = -3.32, $p < 0.01$). This effect indicates that human participants responded earlier to turns with words that TurnGPT found more surprising.

To understand these surprising results, we examined individual stimuli qualitatively. This stimulus-by-stimulus approach can generate potential explanations and hypotheses to explore in future work. We find that factors other than surprisal, such as turn construction, may influence ORT in different ways than LLM-produced surprisal. Note that this strategy has severe limitations, including the fact that word frequency strongly affects surprisal values: common words in a violative condition may be less surprising than rare words in a congruent condition. We include the results of this analysis in Appendix E.

It is also important to consider that the surprisal (see Equation 2) used in this task is based on the predictability of individual words within the turns. This method captures local dependencies and provides a detailed view of word-by-word predictability, aligning with traditional LLM training objectives (Radford et al., 2019; Brown et al., 2020). Surprisal is therefore a *direct* function of the predictability of words. In contrast, in the experimental data we analyze here, humans were asked to predict the end of turn through a button press task. Our comparison of ORT and surprisal rests on the assumption that the timing of the button press is dependent on the predictability of the words in the turn. In order to bypass this assumption, we conducted a follow-up analysis in which we calculated the surprisal of the end of the turn and then compared these values to human ORT data, effectively mimicking the experimental task in our models. We calculated surprisal based on the probability of the *end of turn* (EOT) token, which is used by LLMs internally to indicate end of turns, *after* all the words in both the first and second turns of the two-turn stimulus. This method considers the turn as a whole and its completion, addressing potential biases from incomplete fragments and aligning more closely with the task of predicting turn endings. Using this method, we find no relationship between turn-end surprisal and ORT: the model-estimated end-of-turn surprisal had no relationship with human end-of-turn estimation timing. This suggests that our new model also does not predict spoken dialogue in the same way that humans do. See Appendix F for the results of our experiments based on the alternative surprisal formulation.

## 4 | DISCUSSION

Notwithstanding the success story of LLMs, these models are predominantly pre-trained on written monologue. This raises the question of whether LLMs are able to model the unique dynamics of spoken dialogue, the oldest and most ubiquitous way humans communicate with each other. In the current paper, we investigate whether LLMs learn the normative structure underlying interactive spoken language, or whether they instead replicate superficial statistical regularities of language.

An utterance's message depends on who is saying it, so a crucial aspect of spoken conversation is listeners' ability to identify who is speaking. Humans use their knowledge of speaker identity during language comprehension to predict upcoming language in conversation. We therefore specifically investigated the ability of LLMs to accurately incorporate speaker identity information in their predictions. First, we fine-tuned several variants of GPT-2 on transcripts of natural conversations containing speaker identity information. We then obtained model-estimated surprisal values for conversational turns from Warnke (2024)'s experimental stimuli. We investigated whether our models could differentiate between experimental conditions based on congruence, and then compared model surprisal to human behavioral data from the same experiment. Below, we briefly summarize our findings, and then discuss their implications for the use of LLMs in spoken dialogue research.

Our analyses show that all fine-tuned LLMs found incongruent turns more surprising than congruent turns in sequences with speaker transitions, but not in sequences with speaker holds. Our models showed a main effect of speaker: turns with speaker holds were more surprising to the models than turns with speaker transitions. Lastly, we found an interaction effect: incongruent and violation conditions (turns that were unexpected independent of speaker identity) were deemed *less* surprising in the same-speaker condition than in the different-speaker condition. These results suggest that our models do not take speaker identity information into account when differentiating between turn congruence in the same way that humans do.

Given that humans take into account speaker identity in their linguistic predictions, we explored whether a model with an explicit representation of speaker identity, TurnGPT, would better predict language in dialogue. We found that surprisal values were much lower overall for TurnGPT. Further, while this model produced lower surprisal values for the same-speaker condition as compared to GPT-2, it found that the same-speaker stimuli were more surprising than the different-speaker stimuli. This indicates that even when speaker identity is explicitly represented, the model still does not replicate human behavioral data.

An additional goal in the current paper was to investigate the effect of fine-tuning dataset size on model performance. We found that models trained on five and twenty-eight conversations produced higher surprisal values than the null (pre-trained only) GPT-2 model. We found a bigger difference in surprisal between the models fine-tuned on five vs. twenty-eight conversations than models fine-tuned on five conversations vs. the null model. This suggests that a smaller amount of data may be sufficient for fine-tuning our models.

Lastly, we directly investigated the relationship between model surprisal and human ORTs for the stimuli from Warnke (2024)'s end-of-turn prediction task. We found that model surprisal was negatively correlated with ORT in the corresponding human data: turns that the models predicted to have high surprisal were associated with faster human responses. This somewhat surprising finding suggests that our models do not replicate human dialogue processing.

Taken together, our results show that LLMs that are fine-tuned on dialogue data with speaker identity information generally do not exhibit human-like performance in spoken dialogue. Only when there was a speaker transition, the fine-tuned language models were able to use speaker identity to predict the probability of words in a pattern similar to that of human participants. We would like to note that evaluating LLM ability to use speaker identity information constitutes a relatively weak test of understanding conversational structure. Conversational interaction consist of complex sequential relations that are much more-open ended, context dependent, and less contrastive than our test of speaker identity use to predict upcoming words (Sidnell and Enfield, 2012; Levinson, 2013). If LLMs are not able to take into account speaker identity, there is little reason to think that LLMs would grasp other more complex features of conversational structure. The most principled way to address this problem would be to pre-train LLMs using data from naturally occurring spoken dialogue. At present, the availability of transcribed spoken dialogue data is several orders of magnitudes lower than for written monologue data, but incremental progress can perhaps be achieved by adapting the models and/or the fine-tuning regimes to improve the models' awareness of speaker identity in other, more principled and effective ways than we could in this study.

In the current study, we compare GPT-2 to human data from a behavioral experiment with relatively high ecological validity. That said, our comparison of model output to data from this particular experiment has several limitations. First, the experiment consists of an *overhearer* paradigm, meaning participants listened to conversations rather than actively participating in them. Second, the experimental data and the fine-tuning data differ in the proportions of speaker holds and transitions. In the experimental stimuli, turn-taking was evenly split: half of the second turns were spoken by the same speaker as the first, and half were spoken by a different speaker. In contrast, in the naturally occurring conversations used for fine-tuning, 80% of turns involved a speaker transition. This imbalance may explain why the model makes more human-like predictions in sequences with speaker changes but performs less accurately in sequences with speaker holds. As a result, our findings might reflect the model's sensitivity to this imbalance, rather than its capacity to make human-like prediction.

It is important to note that the experimental design in this work involves a key trade-off: the experimental stimuli were designed to isolate the effect of speaker identify on word probabilities (e.g, by controlling for turn length, speaker transition rations, etc.), which inherently differentiates them from the naturally occurring conversations used for fine-tuning. One factor is that, while the fine-tuning data reflects the uneven distribution of speaker holds and transitions typical in real dialogue, the testing stimuli balanced speaker transitions as is common and necessary in experimental design (Warnke, 2024). To address this imbalance and better understand its effect on LLMs' ability to make human-like predictions, we recommend the following steps for future research. Training and evaluation data should have matching ratios of speaker transitions, ideally reflecting those in natural conversations rather than the artificially balanced experimental designs. Additionally, data used to fine-tune the models should be transcribed more granularly such that that successive TCUs spoken by the same speaker would appear as speaker holds rather than as a single turn spoken by one speaker. Capturing accurate speaker transition ratios from these more detailed natural transcripts could also inform the design of experimental stimuli. Achieving closer alignment between training and evaluation data will be crucial for validating these findings in future research.

A further limitation of the current study rests in its assumptions. Specifically, we compare lexical level model-estimated surprisal to turn-end anticipation as measured by ORT in humans. Given that prior research has shown a relationship between lexical predictability and turn-end anticipation timing, we assume that turn-end anticipation timing indexes lexical predictions in humans. We then draw on this assumption to evaluate the model's performance compared to humans. One limitation of our approach is that turn-end estimation as measured by a button press is an indirect measure of human lexical predictability. Future work could bypass this, and more directly measure offline human predictability judgments (e.g. cloze norms) in addition to on-line behavioral (or neural) measures of turn predictability to compare to model-estimated surprisal. This would provide stronger evidence for evaluating LLMs' ability to capture predictability as humans do in a spoken dialogue setting.

One unexpected finding of the current paper is that model-estimated surprisal values were negatively associated with human ORTs: turns with shorter ORTs were more surprising to the model. To investigate this relationship, we conducted a qualitative analysis of individual stimuli. While this stimulus-by-stimulus approach has severe limitations, it can generate tentative explanations and hypotheses for explaining our results and for future research. In our analysis, we found examples of stimuli with high ORT but low model-estimated surprisal. In these stimuli, participants could have understood the short second turn to project upcoming talk, and thus waited to indicate the end of the turn. Listeners generally assume cooperativity in conversation (Warnke and de Ruiter, 2023), thus perceiving an incongruent stimulus as an incomplete fragment. It is worth noting that our experiment only consisted of relatively short second turns with only one or two words. Further research should investigate the relationship between surprisal, perceived turn completeness, and the incorporation of speaker identity in dialogue prediction using turns that vary more in their length and complexity, providing a more ecologically valid environment.

Another unexpected finding of our study is that fine-tuned LLMs showed an increase in surprisal compared to the null models. One explanation for these higher baseline surprisal values might be the differences in the distribution of the pre-training and fine-tuning data (See Section 2.1.2). Specifically, the fine-tuning data from the ICC included speech particles and unique terminologies absent from the written-first language data used in pre-training. These elements, such as transcribed word cutoffs and stutters, may cause the fine-tuned models to predict words at a lower probability, reflecting a more nuanced understanding of spoken dialogue. Despite these potential differences, the ICC data provides a richer context for understanding conversational dynamics, essential for modeling spoken dialogue. Average surprisal (i.e., perplexity) does not necessarily indicate worse predictions; instead, speakers can purposely increase surprisal to create a more uniform information density (Jaeger and Levy, 2006) or to perform specific actions, such as telling jokes (Xie et al., 2021). Experimental evidence shows that models with higher perplexity can better model human language comprehension (Oh and Schuler, 2023; Kuribayashi et al., 2021). Therefore, our use of the ICC, a corpus of unscripted dialogue, provides valuable insights into the use of LLMs in a spoken language context.

Though the limitations discussed above somewhat impact the generalizability of the work presented here, we do not think that they substantially undermine our conclusion that LLMs trained on written monologue do not replicate the unique dynamics of spoken dialogue. It is worth noting that our study investigated only models trained on English language using English experimental material. Languages and cultures vary in their dynamics of dialogue. Australian Aboriginal people, for example, are comfortable with longer silences between turns in conversation (Mushin and Gardner, 2009), whereas speakers of English consider longer pauses to be indicative of a communication problem (Jefferson, 1989). In Japanese talk, backchannels are far more frequent compared to American English (White, 1989), and across languages and cultures, interruptions can signify different communicative intentions (Murata, 1994). Given the cross-cultural variation of talk in dialogue, it would be interesting and important to replicate the current work in other languages to investigate which dialogic dimensions LLMs can and cannot learn. Future research should also investigate the ability of LLMs to predict spoken language with state-of-the-art models as they become available. Taken together, our findings suggest that the fact that LLMs show impressive human-like performance in written language, does not (yet) mean that they are suitable for employment in embodied interactive agents, dialogue systems, or the scientific analysis of spoken conversation.

## 5 | RISKS AND ETHICAL CONSIDERATIONS

Despite the major advancements in language modeling provided by LLMs, they are accompanied by a number of inherent risks. The data used to train LLMs generally contain ableist, racist, or misogynistic worldviews, which means that they tend to absorb and amplify harmful stereotypes (Bender et al., 2021). Off-the-shelf models, for example, have been found to exhibit considerable anti-queer bias (Felkner et al., 2023). While this bias can be reduced by fine-tuning LLMs on data written directly by members of particular marginalized communities, most widely available LLMs are not fine-tuned to mitigate these stereotypes. Biased language output from LLMs can be particularly harmful due to our natural tendency to infer coherence and communicative intent originating from a real person from language (Bender et al., 2021; Warnke and de Ruiter, 2023), even when it is generated by machines (Nass et al., 1994; Weizenbaum, 1976). Because LLMs are not individuals, are not 'intelligent' (Pasquinelli, 2020), and simply replicate statistical dependencies, language generated by them cannot contain any communicative intent. Our propensity to interpret language as communicative acts that convey intent can therefore lead to a flawed interpretation of meaning from LLMs' biased output. Furthermore, because LLMs vary in their degree of openness, they lack computational reproducibility (Liesenfeld et al., 2023). It is especially important to keep these harms in risks in mind since most state-of-the art LLMs are not truly open-source and are only available through public facing interfaces (Liesenfeld and Dingemanse, 2024). We would like to note that no AI-tools were used to assist in the writing of or analysis conducted in this work.

## REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023) Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Baayen, R. H., Davidson, D. J. and Bates, D. M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, **59**, 390–412.

Balloccu, S., Schmidtová, P., Lango, M. and Dusek, O. (2024) Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. Y. Graham and M. Purver), 67–93. St. Julian's, Malta: Association for Computational Linguistics. URL: `https://aclanthology.org/2024.eacl-long.5`.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Besacier, L., Barnard, E., Karpov, A. and Schultz, T. (2014) Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, **56**, 85–100.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020) Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

Bürkner, P.-C. (2017) brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, **80**, 1–28.

Caucheteux, C., Gramfort, A. and King, J.-R. (2021) Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning*, 1336–1348. PMLR.

Caucheteux, C. and King, J.-R. (2022) Brains and algorithms partially converge in natural language processing. *Communications biology*, **5**, 134.

Cheng, S. (2024) When journalism meets ai: Risk or opportunity? *Digital Government: Research and Practice*.

Cohen, A. D., Roberts, A., Molina, A., Butryna, A., Jin, A., Kulshreshtha, A., Hutchinson, B., Zevenbergen, B., Aguera-Arcas, B. H., ching Chang, C., Cui, C., Du, C., Adiwardana, D. D. F., Chen, D., Lepikhin, D. D., Chi, E. H., Hoffman-John, E., Cheng, H.-T., Lee, H., Krivokon, I., Qin, J., Hall, J., Fenton, J., Soraker, J., Meier-Hellstern, K., Olson, K., Aroyo, L. M., Bosma, M. P., Pickett, M. J., Menegali, M. A., Croak, M., Díaz, M., Lamm, M., Krikun, M., Morris, M. R., Shazeer, N., Le, Q. V., Bernstein, R., Rajakumar, R., Kurzweil, R., Thoppilan, R., Zheng, S., Bos, T., Duke, T., Doshi, T., Zhao, V. Y., Prabhakaran, V., Rusch, W., Li, Y., Huang, Y., Zhou, Y., Xu, Y. and Chen, Z. (2022) Lamda: Language models for dialog applications. In *arXiv*.

De Ruiter, J.-P., Mitterer, H. and Enfield, N. J. (2006) Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 515–535.

Dingemanse, M. and Liesenfeld, A. (2022) From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5614–5633.

Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A. and Choi, Y. (2022) Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. S. Muresan, P. Nakov and A. Villavicencio), 7250–7274. Dublin, Ireland: Association for Computational Linguistics. URL: https://aclanthology.org/2022.acl-long.501.

Drieman, G. H. (1962) Differences between written and spoken language: An exploratory study. *Acta Psychologica*, **20**, 36–57.

Ekstedt, E. and Skantze, G. (2020) TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (eds. T. Cohn, Y. He and Y. Liu), 2981–2990. Online: Association for Computational Linguistics. URL: https://aclanthology.org/2020.findings-emnlp.268.

— (2022) Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*, 5190–5194.

Falcon, W. and The PyTorch Lightning team (2019) PyTorch Lightning. URL: https://github.com/Lightning-AI/lightning.

Felkner, V., Chang, H.-C. H., Jang, E. and May, J. (2023) WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds. A. Rogers, J. Boyd-Graber and N. Okazaki), 9126–9140. Toronto, Canada: Association for Computational Linguistics. URL: https://aclanthology.org/2023.acl-long.507.

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A. and Prather, J. (2022) The robots are coming: Exploring the implications of openai codex on introductory programming. In *Proceedings of the 24th Australasian Computing Education Conference*, 10–19.

Gholami, S. (2024) Do generative large language models need billions of parameters? In *Redefining Security With Cyber AI*, 37–55. IGI Global.

Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.-H. and Yang, Y. (2022) LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022* (eds. M. Carpuat, M.-C. de Marneffe and I. V. Meza Ruiz), 724–736. Seattle, United States: Association for Computational Linguistics. URL: https://aclanthology.org/2022.findings-naacl.55.

Hale, J. (2001) A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, 1–8. USA: Association for Computational Linguistics. URL: https://doi.org/10.3115/1073336.1073357.

Jaeger, T. and Levy, R. (2006) Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, **19**.

Jefferson, G. (1989) Notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In *Conversation: An interdisciplinary perspective* (eds. D. Roger and P. Bull), chap. 8, 166–196. Clevendon: Multilingual Matters.

Jeffreys, H. (1939) *Theory of Probability*. Oxford, England: Clarendon Press.

Jiang, B., Ekstedt, E. and Skantze, G. (2023) Response-conditioned turn-taking prediction. In *Findings of the Association for Computational Linguistics: ACL 2023* (eds. A. Rogers, J. Boyd-Graber and N. Okazaki), 12241–12248. Toronto, Canada: Association for Computational Linguistics. URL: https://aclanthology.org/2023.findings-acl.776.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X. et al. (2019) A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 449–456. IEEE.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E. and Raileanu, R. (2023) Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Kreps, S., McCain, R. M. and Brundage, M. (2022) All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, **9**, 104–117.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. et al. (2023) Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, **2**, e0000198.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M. and Inui, K. (2021) Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (eds. C. Zong, F. Xia, W. Li and R. Navigli), 5203–5217. Online: Association for Computational Linguistics. URL: https://aclanthology.org/2021.acl-long.405.

Levinson, S. C. (1983) *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

— (2013) Recursion in pragmatics. *Language*, 149–162.

— (2016) Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences*, **20**, 6–14.

Levinson, S. C. and Torreira, F. (2015) Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, **6**, 731.

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

Li, J., Tang, T., Zhao, W. X. and Wen, J.-R. (2021) Pretrained language model for text generation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (ed. Z.-H. Zhou), 4492–4499. International Joint Conferences on Artificial Intelligence Organization. URL: https://doi.org/10.24963/ijcai.2021/612. Survey Track.

Liesenfeld, A. and Dingemanse, M. (2024) Rethinking open source generative ai: open-washing and the eu ai act. In *Seventh Annual ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2024)*. ACM.

Liesenfeld, A., Lopez, A. and Dingemanse, M. (2023) Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, 1–6.

Lund, B. D. and Wang, T. (2023) Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News*, **40**, 26–29.

Magyari, L. (2022) Predictions in conversation. *A Life in Cognition: Studies in Cognitive Science in Honor of Csaba Pléh*, 59–75.

Magyari, L., Bastiaansen, M. C., De Ruiter, J. P. and Levinson, S. C. (2014) Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience*, **26**, 2530–2539.

Magyari, L. and De Ruiter, J. P. (2008) Timing in conversation: the anticipation of turn endings. In *12th Workshop on the Semantics and Pragmatics Dialogue*, 139–146. King's college.

— (2012) Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in psychology*, **3**, 376.

Mahowald, K., Fedorenko, E., Piantadosi, S. T. and Gibson, E. (2013) Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, **126**, 313–318.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B. and Fedorenko, E. (2024) Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, **28**, 517–540. URL: https://www.sciencedirect.com/science/article/pii/S1364661324000275.

Matsuki, K., Chow, T., Hare, M., Elman, J. L., Scheepers, C. and McRae, K. (2011) Event-based plausibility immediately influences on-line language comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **37**, 913.

Metzing, C. and Brennan, S. E. (2003) When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, **49**, 201–213.

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K. and Coulson, S. (2024) Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of language*, **5**, 107–135.

Michaelov, J. A., Coulson, S. and Bergen, B. K. (2022) So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*, **15**, 1033–1042.

Murata, K. (1994) Intrusive or co-operative? a cross-cultural study of interruption. *Journal of pragmatics*, **21**, 385–400.

Mushin, I. and Gardner, R. (2009) Silence is talk: Conversational silence in australian aboriginal talk-in-interaction. *Journal of pragmatics*, **41**, 2033–2052.

Nass, C., Steuer, J. and Tauber, E. R. (1994) Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72–78.

Nguyen, T. A., Kharitonov, E., Copet, J., Adi, Y., Hsu, W.-N., Elkahky, A., Tomasello, P., Algayres, R., Sagot, B., Mohamed, A. et al. (2023) Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, **11**, 250–266.

Oh, B.-D., Clark, C. and Schuler, W. (2022) Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, **5**.

Oh, B.-D. and Schuler, W. (2023) Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, **11**, 336–350.

Pasquinelli, M. (2020) How a machine learns and fails–a grammar of error for artificial intelligence. spheres.

Pilan, I., Prévot, L., Buschmeier, H. and Lison, P. (2024) Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (eds. T. Kawahara, V. Demberg, S. Ultes, K. Inoue, S. Mehri, D. Howcroft and K. Komatani), 440–457. Kyoto, Japan: Association for Computational Linguistics. URL: https://aclanthology.org/2024.sigdial-1.38.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019) Language models are unsupervised multitask learners. *OpenAI blog*, **1**, 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. J. (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, **21**, 5485–5551.

Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A. and Marin, S. (2023) The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, **9**, eadf3197. URL: https://www.science.org/doi/abs/10.1126/sciadv.adf3197.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.-b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J. et al. (2024) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Riest, C., Jorschick, A. B. and De Ruiter, J.-P. (2015) Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, **6**, 89.

Sacks, H., Schegloff, E. A. and Jefferson, G. (1974) A simplest systematics for the organization of turn-taking for conversation. *Language*, **50**, 696–735. URL: http://www.jstor.org/stable/412243.

Sathish, V., Lin, H., Kamath, A. K. and Nyayachavadi, A. (2024) Llempower: Understanding disparities in the control and access of large language models. *arXiv preprint arXiv:2404.09356*.

Schegloff, E. A. (1982) Discourse as an interactional achievement: Some uses of ‘uh huh’and other things that come between sentences. *Analyzing discourse: Text and talk*, **71**, 71–93.

Sejnowski, T. J. (2023) Large language models and the reverse turing test. *Neural computation*, **35**, 309–342.

Shannon, C. E. (1948) A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N. and Anderson, R. (2023) The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Sidnell, J. and Enfield, N. J. (2012) Language diversity and social action: A third locus of linguistic relativity. *Current Anthropology*, **53**, 302–333.

Skantze, G. (2017) Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 220–230.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J. P., Yoon, K.-E. et al. (2009) Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, **106**, 10587–10592.

Sun, J., Mei, C., Wei, L., Zheng, K., Liu, N., Cui, M. and Li, T. (2024) Dial-insight: Fine-tuning large language models with high-quality domain-specific data preventing capability collapse. *arXiv preprint arXiv:2403.09167*.

Tilley, S. A. (2003) “challenging” research practices: Turning a critical lens on the work of transcription. *Qualitative inquiry*, **9**, 750–773.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (2023) Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Trott, S., Jones, C., Chang, T., Michaelov, J. and Bergen, B. (2023) Do large language models know what humans know? *Cognitive Science*, **47**, e13309.

Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M. and Hagoort, P. (2008) The neural integration of speaker and message. *Journal of cognitive neuroscience*, **20**, 580–591.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I. (2017) Attention is all you need. In *Advances in Neural Information Processing Systems* (eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett), vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

VM, K., Warrier, H., Gupta, Y. et al. (2024) Fine tuning llm for enterprise: Practical guidelines and recommendations. *arXiv preprint arXiv:2404.10779*.

Warnke, L. (2024) *The Communication of Social Meaning in Conversation*. Ph.D. thesis, Tufts University.

Warnke, L. and de Ruiter, J. P. (2023) Top-down effect of dialogue coherence on perceived speaker identity. *Scientific Reports*, **13**, 3458. URL: https://doi.org/10.1038/s41598-023-30435-z.

Weizenbaum, J. (1976) *Computer Power and Human Reason: From Judgment to Calculation*. USA: W. H. Freeman & Co.

Wesseling, W., van Son, R., Pols, L. C. et al. (2006) On the sufficiency and redundancy of pitch for trp projection. In *Interspeech*. Citeseer.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J. and Wagenmakers, E.-J. (2011) Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, **6**, 291–298.

White, S. (1989) Backchannels across cultures: A study of americans and japanese1. *Language in society*, **18**, 59–76.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P. and Levy, R. P. (2020) On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 1707–1713.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A. (2020) Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds. Q. Liu and D. Schlangen), 38–45. Online: Association for Computational Linguistics. URL: https://aclanthology.org/2020.emnlp-demos.6.

Xie, Y., Li, J. and Pu, P. (2021) Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (eds. C. Zong, F. Xia, W. Li and R. Navigli), 33–39. Online: Association for Computational Linguistics. URL: https://aclanthology.org/2021.acl-short.6.

Yang, H., Zhang, Y., Xu, J., Lu, H., Heng, P.-A. and Lam, W. (2024a) Unveiling the generalization power of fine-tuned large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (eds. K. Duh, H. Gomez and S. Bethard), 884–899. Mexico City, Mexico: Association for Computational Linguistics. URL: https://aclanthology.org/2024.naacl-long.51.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B. and Hu, X. (2024b) Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, **18**. URL: https://doi.org/10.1145/3649506.

Zhang, H., Li, L. H., Meng, T., Chang, K.-W. and Van den Broeck, G. (2023) On the paradox of learning to reason from data. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*. URL: http://starai.cs.ucla.edu/papers/ZhangIJCAI23.pdf.

# A | FINE-TUNING PROCEDURES

This appendix provides a detailed overview of the data, data transformations, and fine-tuning procedures used in this study. First, since verbatim transcripts of conversation are resource intensive to produce, we investigate the effect of the amount of data available for fine-tuning on the LLM-surprisal values (see Section 3.2). To ensure that any difference was due to the amount of fine-tuning and not unexpected differences between corpora, we examined the frequency of words in the two datasets. Most words had very similar frequencies across both datasets. As shown in Figure A1, the frequency of almost all words changed by less than 0.5% in either direction. The frequency of "know" changed the most between datasets; it composed 2.79% of the words produced in the five-conversation dataset, and 2.17% of the words produced in the twenty-eight conversation dataset. 74.31% of the unique words in the five-conversation corpus had a lower frequency in the twenty-eight conversation corpus; 25.68% of the words had a higher frequency in the twenty-eight conversation corpus. 60.32% of all unique words in the twenty-eight conversation corpus were not present in the five-conversation corpus, but these words only made up 8.9% of all the spoken words. We determined that the differences in these data were negligible for the purposes of this study. Therefore, differences in the outcome of models can be attributed to the amount of fine-tuning data and not the vocabulary used in the corpora.

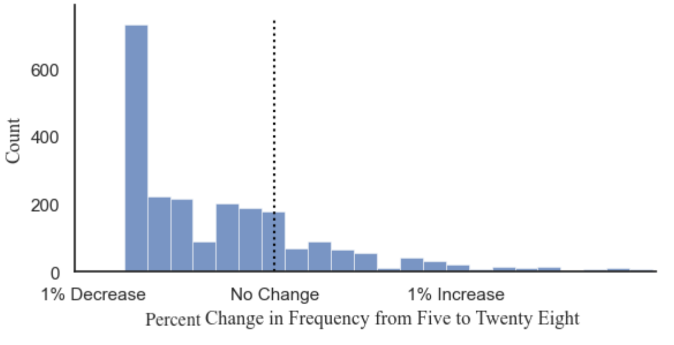


**FIGURE A1** Percent change in word frequencies after adding twenty-three conversations to the five-conversation dataset. Values to the left of the vertical dotted line (negative values) indicate that the five-conversation dataset had a higher word frequency than the twenty-eight-conversation dataset.

Next, we prepared the GPT-2 and TurnGPT models and the five- and twenty-eight-conversation datasets for the fine-tuning process (see Section 2.1.2). We used the pre-trained GPT-2 model from the transformers library (Wolf et al., 2020) and implemented TurnGPT on top of this base model using PyTorch Lightning (Falcon and The PyTorch Lightning team, 2019) as the main implementation framework. TurnGPT requires additional tokens to represent speaker identities and must be fine-tuned to use this information accurately. Due to resource constraints, we used the smallest GPT-2 model with 117M parameters, 12 layers, 12 heads, and 768 hidden units as the pre-trained model in both cases.

To ensure accurate surprisal calculations, we performed data preprocessing on both the ICC used for fine-tuning and the experimental stimuli from Warnke (2024) (see Figure A2). In line with prior work, we added additional tokens to GPT-2 to indicate speaker turns, along with start- and end-of-sequence tokens (e.g., Ekstedt and Skantze 2020) in all cases. These additional tokens were unnecessary for TurnGPT, as its tokenizer already assigns explicit speaker identities to each turn in a sequence. The preprocessing steps differed depending on the surprisal calculation method (see Section 2.2.2). For the word-only surprisal method (see Equation 2 ), no extra tokens were required. However, for the end-of-turn (EOT) token surprisal method (see Equation 8), we inserted an explicit EOT token after each turn in the fine-tuning data and after the first turn in the two-turn stimuli during inference. This step is crucial for teaching the model to recognize turn boundaries, which is essential for accurate turn-taking predictions in dialogue systems (Skantze, 2017; Ekstedt and Skantze, 2022; Jiang et al., 2023).

We fine-tuned each model on the next-word prediction task, which is a common practice in training language models. This objective enables models to learn the probability distribution of words, improving their ability to generate coherent and contextually appropriate text (Radford et al., 2019; Brown et al., 2020). As a result, we created two versions of each model (GPT-2 and TurnGPT) fine-tuned on both the five and twenty-eight conversation datasets, resulting in ten total fine-tuned language models. On average, it took approximately 2 hours to fine-tune each model using NVIDIA's T4 GPUs on a high performance cluster.

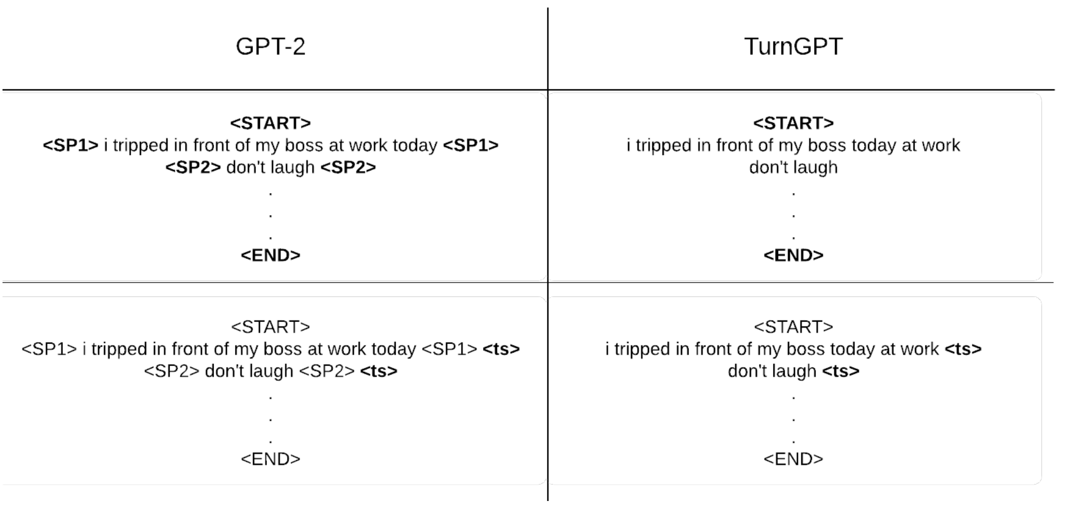


**FIGURE A2** Example of data preprocessing applied to the ICC dataset and stimuli from *Anonymous* (2024). The top row shows preprocessing for the word-based surprisal method, while the bottom row shows preprocessing for the end-of-turn (EOT) surprisal method. Differences appear between surprisal methods (rows) and models (columns). For GPT-2, the EOT token (<ts>) and speaker labels (<SP1>, <SP2>) are added, along with start and end tokens (<START>, <END>) to indicate complete sequences. TurnGPT, however, encodes speaker information internally and does not require explicit speaker labels. Formatting was kept consistent across fine-tuning and inference data.

# B | FULL CONGRUENCE AND SPEAKER REGRESSION RESULTS

This appendix provides a detailed overview of the analysis and results obtained in Section 3.1. In that section, we investigated whether the fine-tuned LLMs, described in Section 2.1.1, would find incongruent stimuli more surprising than congruent stimuli (Hypothesis 1) and whether there would be any main effects of speaker or interaction effects between speaker and congruence (Hypothesis 2). As shown in Figure B1, we created five RMs for *each* language model (for a total of 25 Bayesian and frequentist models) to help determine whether speaker identity (same vs. different), congruence (congruent, incongruent, and violative), and any interaction effects between the two influence LLM produced surprisal values for the second turn in each stimulus (see Section 2.2.1). We compared the frequentist RMs using likelihood ratio tests and Bayesian RMs using Bayes Factors to identify the best RMs.

| Regression Model | Regression Equation |
|---|---|
| Model 1 | Surprisal ∼ (1 \| Group) |
| Model 2 | Surprisal ∼ Speaker + (1 \| Group) |
| Model 3 | Surprisal ∼ Congruence + (1 \| Group) |
| Model 4 | Surprisal ∼ Speaker + Congruence + (1 \| Group) |
| Model 5 | Surprisal ∼ Speaker * Congruence + (1 \| Group) |

**TABLE 3** Regression models created for each language model using the lmer (Bates et al., 2015) (frequentist) and brms (Bürkner, 2017) (Bayesian) packages in R (Bates et al., 2015; Bürkner, 2017).

Table 4 shows the Bayes Factors comparing RMs 2-5 to RM 1 in Table 3 for the pre-trained (Null) model, GPT-2, and TurnGPT fine-tuned on five and twenty-eight conversations. For each LLM, the most likely RM was Model 5 (as described in Table 3), which included main effects of and interaction effects between speaker and congruence. Note that the RMs for an LLM can be compared by dividing the Bayes Factor of one by the other. For example, the best RM (Model 5 in Table 4) for the null GPT-2 ($BF_{10} = 7.64E + 08$) was 1,800 times more likely than the next best RM ($BF_{10} = 4.18E + 05$).

| | GPT-2 | | | Turn GPT | |
|---|---|---|---|---|---|
| Regression Model | Null | Five | Twenty-Eight | Five | Twenty-Eight |
| Model 2 | 0.544 | 17.37 | 7.62 | 2.5 | 2.44 |
| Model 3 | 7.84E+05 | 852.29 | 243.05 | 7.89E+03 | 3.28E+06 |
| Model 4 | 4.18E+05 | 1.57E+04 | 1.89E+03 | 2.03E+04 | 8.40E+06 |
| Model 5 | 7.64E+08 | 1.35E+10 | 5.19E+08 | 4.05E+07 | 4.77E+11 |

**TABLE 4** Bayes Factors for regression models (as described in Table 3) investigating the effect of predictors on surprisal. The denominator for the Bayes Factors was Model 1 in Table 3.

Additionally, to compare the variance explained by the RMs in Table 3, we performed frequentist likelihood ratio tests, which replicated the pattern of Bayes Factors described above. Table 5 presents the coefficients for all RMs described in Table 3 for the pre-trained only (null) GPT-2 as well as GPT-2 and TurnGPT fine-tuned on each dataset (five and twenty-eight conversations). Figure B1 provides a visualization of the coefficients and 95% CI for all models described in Table 5. It shows that all fine-tuned models found that the main effects of speaker and congruence, along with their interaction effects, were all statistically significant predictors of surprisal.

| | GPT-2 | | | Turn GPT | |
|---|---|---|---|---|---|
| | Null | Five | Twenty-eight | Five | Twenty-Eight |
| Intercept | 10.16 (9.68 - 10.64)** | 23.1 (22.42 - 23.77)** | 24.82 (24.21 - 25.44)** | 12.15 (11.51 - 12.8)** | 11.61 (11.08 - 12.15)** |
| Same | 0.82 (0.21 - 1.43)* | 2.09 (1.39 - 2.8)** | 2.06 (1.31 - 2.81)** | 0.87 (0.03 - 1.71)* | 0.96 (0.28 - 1.65)* |
| Incongruent | 0.55 (-0.06 - 1.16) | 1.08 (0.38 - 1.79)* | 1.15 (0.4 - 1.89)* | 0.88 (0.03 - 1.72)* | 0.97 (0.28 - 1.66)* |
| Violation | 1.91 (1.3 - 2.52)** | 2.11 (1.41 - 2.82)** | 2.06 (1.32 - 2.81)** | 2.26 (1.42 - 3.10)** | 2.31 (1.62 - 3.00)** |
| Same *Incongruent | -1.09 (-1.95 - -0.23) | -2.15 (-3.15 - -1.15)** | -2.24 (-3.3 - -1.18)** | -1.73 (-2.92 - -0.53)* | -1.92 (-2.89 - -0.94)** |
| Same *Violation | -1.68 (-2.55 - -0.82)** | -2.43 (-3.43 - -1.43)** | -2.39 (-3.45 - -1.33)** | -2.15 (-3.34 - -0.95)** | -2.09 (-3.06 - -1.11)** |
| $R^2$ | 0.24 | 0.47 | 0.29 | 0.17 | 0.22 |

**TABLE 5** Frequentist regression results for most predictive model, Model 5 in Table 3, for all language models. 95% confidence intervals presented in parentheses. * = p-value under 0.05, ** = p-value under 0.01.

| | GPT-2 | | | | | | Turn GPT | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Null | | Five | | Twenty-eight | | Five | | Twenty-eight | |
| | Same | Different | Same | Different | Same | Different | Same | Different | Same | Different |
| Congruent Mean | 10.98 | 10.16 | 25.19 | 23.09 | 26.88 | 24.82 | 13.02 | 12.15 | 12.58 | 11.61 |
| Incongruent Mean | 10.45 | 10.71 | 24.13 | 24.18 | 25.81 | 25.97 | 12.18 | 13.03 | 11.64 | 12.58 |
| $t$ | 1.50 | -1.62 | 2.19 | -2.27 | 2.50 | -2.60 | 1.87 | -1.93 | 2.49 | -2.56 |
| $p$ | 0.14 | 0.11 | 0.03* | 0.02* | 0.01* | 0.01* | 0.06 | 0.05* | 0.01* | 0.01* |
| $BF_{10}$ | 0.40 | 0.48 | 1.33 | 1.59 | 2.58 | 3.30 | 0.72 | 0.80 | 2.57 | 2.98 |

**TABLE 6** T-tests comparing surprisal of (in)congruent stimuli for each language model, split by speaker condition. Most fine-tuned models found that congruent stimuli were more surprising than incongruent stimuli in the same-speaker conditions. However, they found that congruent stimuli were less surprising than incongruent stimuli in the different-speaker condition.

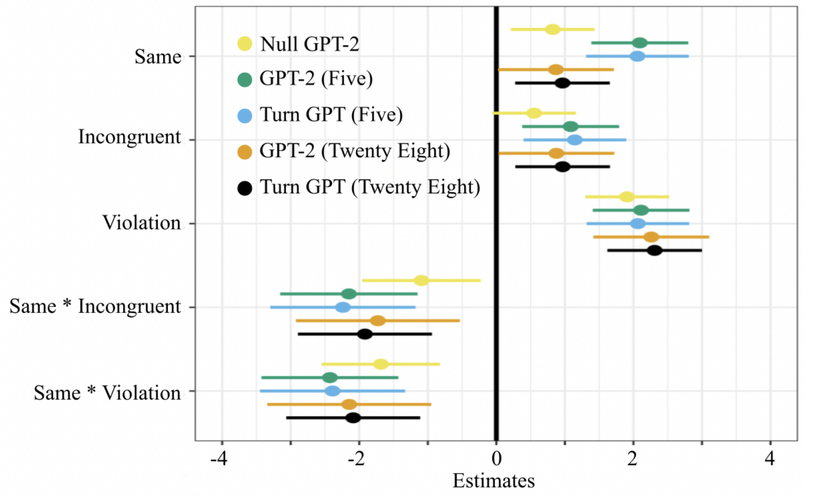


**FIGURE B1** The regression coefficients for all fixed effects in regression model 5 (see Table 3) for all language models.

To investigate the statistically significant interaction effects between speaker and congruence highlighted above, we performed a series of follow-up t-tests (both Bayesian and frequentist). These tests compared the incongruent and congruent stimuli within each speaker condition (same vs. different) for each language model (GPT-2 and TurnGPT) fine-tuned on both datasets (five and twenty-eight conversations). Table 6 shows that most of the fine-tuned models found statistically significant differences between the congruent and incongruent stimuli. The Bayes Factors indicated anecdotal evidence for most of the statistically significant effects. Interestingly, the directionality of the effects differed between the same ($surprisal_{congruent} > surprisal_{incongruent}$) and different ($surprisal_{congruent} < surprisal_{incongruent}$) speaker conditions. Nether the Bayesian nor frequentist t-tests found differences between the incongruent and congruent conditions for the null (pretrained-only) model.

## C | REGRESSIONS INVESTIGATING AMOUNT OF FINETUNING

This appendix provides a detailed overview of the results presented in Section 3.2, where we investigated whether the amount of natural conversational data used for fine-tuning LLMs affected surprisal values. To do so, we concatenated the surprisal data produced by GPT-2 models trained on no (pre-trained-only), five, and twenty-eight conversations (see Section 2.1.2). We created a categorical predictor indicating the dataset used to fine-tune the models and created five RMs, as described in table 7, that added predictors to the best RM (see Equation 3) from Section 3.1.

| Regression Model | Regression Equation |
|---|---|
| Model 6 | Surprisal ~ Speaker * Congruence + Dataset + (1 \| Group) |
| Model 7 | Surprisal ~ Speaker * Congruence + Speaker * Dataset + (1 \| Group) |
| Model 8 | Surprisal ~ Speaker * Congruence + Congruence * Dataset + (1 \| Group) |
| Model 9 | Surprisal ~ Speaker * Congruence + Congruence * Dataset + Speaker * Dataset + (1 \| Group) |
| Model 10 | Surprisal ~ Congruence * Dataset * Speaker + (1 \| Group) |

**TABLE 7** Regression models created using the lmer (Bates et al., 2015) (frequentist) and brms (Bürkner, 2017) (Bayesian) packages in R to explore the effect of the amount of fine-tuning data on the surprisal values produced by GPT-2.

| Regression Model | Bayes Factor |
|---|---|
| Model 7 | 9.75 |
| Model 8 | 0.57 |
| Model 9 | 5.45 |
| Model 10 | 169.91 |

**TABLE 8** Bayes Factors for regression models (as described in Table 7) investigating the effect of training amount on surprisal patterns. The data were so unlikely under the null model (that did not contain training amount as a predictor) that the resulting Bayes Factors were too large to compute. Therefore, in this table, the denominator for the Bayes Factors was the model that contained the baseline model (random intercept for stimulus group, main effects of congruence and speaker, and an interaction effect between congruence and speaker) and a main effect for training amount (Model 6 in Table 7).

We found that the only statistically significant interaction effects between fine-tuning amount and other factors were the interaction effects between training amount and speaker identity. Additionally, we compared the frequentist RMs using likelihood ratio tests (see Table 9) and Bayesian RMs using Bayes Factors (see Table 8). The likelihood ratio tests found no statistically significant difference in model performance when eliminating all interaction effects between the dataset size and the other predictor. In contrast, we found decisive evidence that the best model ($BF_{10} = 169.91$) contained all the interaction effects (Model 10 in Table 7).

| | Estimate | t | p |
|---|---|---|---|
| (Intercept) | 10.16 (9.56 - 10.76) | 33.33 | <0.01** |
| Five | 12.94 (12.2 - 13.67) | 34.43 | <0.01** |
| Twenty-Eight | 14.67 (13.93 - 15.4) | 39.04 | <0.01** |
| Incongruent | 0.55 (-0.18 - 1.28) | 1.47 | 0.14 |
| Violation | 1.91 (1.17 - 2.64) | 5.08 | <0.01** |
| Same Speaker | 0.82 (0.09 - 1.56) | 2.19 | 0.03* |
| Five * Incongruent | 0.53 (-0.51 - 1.57) | 0.99 | 0.32 |
| Twenty-Eight * Incongruent | 0.59 (-0.44 - 1.63) | 1.12 | 0.26 |
| Five * Violation | 0.20 (-0.83 - 1.24) | 0.38 | 0.70 |
| Twenty-Eight * Violation | 0.16 (-0.88 - 1.19) | 0.29 | 0.77 |
| Five * Same Speaker | 1.26 (0.23 - 2.30) | 2.38 | 0.02* |
| Twenty-Eight * Same Speaker | 1.23 (0.20 - 2.27) | 2.33 | 0.02* |
| Incongruent * Same Speaker | -1.10 (-2.14 - -0.06) | -2.06 | 0.04* |
| Violation * Same Speaker | -1.68 (-2.72 - -0.65) | -3.17 | <0.01* |
| Five * Incongruent * Same Speaker | -1.05 (-2.52 - 0.42) | -1.4 | 0.16 |
| Twenty-Eight * Incongruent * Same Speaker | -1.15 (-2.61 - 0.32) | -1.52 | 0.13 |
| Five * Violation * Same Speaker | -0.74 (-2.21 - 0.73) | -0.98 | 0.33 |
| Twenty-Eight * Violation * Same Speaker | -0.71 (-2.17 - 0.76) | -0.94 | 0.35 |

**TABLE 9** Coefficients for frequentist regression including all two- and three-way interactions. 95% confidence intervals presented in parentheses. * = p-value under 0.05, ** = p-value under 0.01.

## D | REGRESSION MODELS ANALYZING SPEAKER REPRESENTATIONS

This appendix provides a detailed overview of the results presented in Section 3.3, where we investigated the effect of speaker representation (implicit vs. explicit) on second-turn surprisal values generated by LLMs. To do so, we concatenated the data from the GPT-2 and TurnGPT models fine-tuned on twenty-eight conversations. The frequentist regressions (Table 11) found that TurnGPT produced statistically significantly lower surprisal values. It also found that TurnGPT was statistically significantly less surprised by the stimuli in the same-speaker condition.

| Regression Model | Regression Equation |
|---|---|
| Model 11 | Surprisal ∼ Speaker * Congruence + Model + (1 ∣ Group) |
| Model 12 | Surprisal ∼ Speaker * Congruence + Speaker * Model + (1 ∣ Group) |
| Model 13 | Surprisal ∼ Speaker * Congruence + Congruence * Model + (1 ∣ Group) |
| Model 14 | Surprisal ∼ Speaker * Congruence + Congruence * Model + Speaker * Model + (1 ∣ Group) |
| Model 15 | Surprisal ∼ Speaker * Congruence * Model + (1 ∣ Group) |

**TABLE 10** Regression models created using the lmer Bates et al. (2015) (frequentist) and brms Bürkner (2017) (Bayesian) packages in R to explore the effect model type (GPT-2 or TurnGPT, both fine-tuned on twenty-eight conversations).

| | Estimate | t | p |
|---|---|---|---|
| (Intercept) | 24.83 (24.25 - 25.4) | 84.15 | <0.01** |
| TurnGPT | -13.21 (-13.94 - -12.48) | -35.27 | <0.01** |
| Incongruent | 1.15 (0.41 - 1.88) | 3.06 | <0.01* |
| Violative | 2.06 (1.33 - 2.80) | 5.51 | <0.01** |
| Same Speaker | 2.06 (1.33 - 2.79) | 5.50 | <0.01** |
| TurnGPT * Incongruent | -0.18 (-1.21 - 0.86) | -0.34 | 0.74 |
| TurnGPT * Violation | 0.25 (-0.79 - 1.28) | 0.46 | 0.64 |
| TurnGPT * Same Speaker | -1.09 (-2.13 - -0.06) | -2.07 | 0.04* |
| Incongruent * Same Speaker | -2.24 (-3.27 - -1.2) | -4.22 | <0.01** |
| Violation * Same Speaker | -2.39 (-3.42 - -1.35) | -4.51 | <0.01** |
| TurnGPT * Incongruent * Same Speaker | 0.32 (-1.15 - 1.78) | 0.42 | 0.67 |
| TurnGPT * Violation * Same Speaker | 0.30 (-1.16 - 1.77) | 0.40 | 0.69 |

**TABLE 11** Results for most complex regression model analyzing how speaker representations predict surprisal (Model 15 in Table 10). 95% confidence intervals presented in parentheses. * = p-value under 0.05, ** = p-value under 0.01.

We compared frequentist models using likelihood ratio tests and Bayesian models using Bayes Factors. The likelihood ratio test found no statistically significant difference in model performance when eliminating all interaction effects between the dataset size and the other predictor. In contrast, as Table 12 shows, we found decisive evidence that the best model contained all the interaction effects (RM 15 in Table 10).

| Regression Model | Bayes Factor |
|---|---|
| Model 12 | 51.35 |
| Model 13 | 1.67 |
| Model 14 | 86.25 |
| Model 15 | 293.82 |

**TABLE 12** Bayes Factors for regression models (described in Table 10) investigating the effect of embedding type (TurnGPT vs. GPT-2 embedding) on surprisal patterns. The data were so unlikely under the null model (that did not contain model type as a predictor) that the resulting Bayes Factors were too large to compute. Therefore, the denominator for these Bayes Factors is the model that contained the baseline model (random intercept for stimulus group, main effects of congruence and speaker, and an interaction effect between congruence and speaker) and a main effect for model type (Model 11 in Table 10).

# E | ANALYSIS OF INDIVIDUAL STIMULI

In Section 3.4, we investigated whether LLM-produced surprisal values predicted human offset response times (ORTs). We found that human participants responded earlier to turns with words that TurnGPT found more surprising, contradicting Hypothesis 6. To understand these surprising results, we explored individual stimuli. First, since multiple participants responded to the same stimulus, we calculated the median ORT for each stimulus. Then, we calculated the z-scores for surprisal and ORT. Hypothesis 6 stated that the z-scores for surprisal and median ORT would be similar to each other. Below, we present example stimuli where TurnGPT produced a high surprisal, which either did or did not match ORTs.

***Excerpt 1**: Low ORT, high surprisal (unexpected pattern)*

```
*SP1:  I'd like to meet your girlfriend
*SP2:  Sure when
```

***Excerpt 2**: High ORT, high surprisal (predicted pattern)*

```
*SP1:  I'd like to meet your girlfriend
*SP1:  Sure when
```

Excerpts 1 and 2 come from the same stimulus pair. In both, TurnGPT produced similarly high word surprisal values, with z-scores of approximately 2.08. However, median ORT for Excerpt 1 was extremely low, with a z-score of -4.00, while ORT for Excerpt 2 was high, with a z-score of 2.78. In Excerpt 1, participants may perceive “Sure" as a sufficient response to the first turn. As a result, participants may have indicated the end of the turn after “sure", without waiting to hear “when". In contrast, “sure" would not complete the turn in Excerpt 2.

***Excerpt 3**: High ORT, low surprisal (unexpected pattern)*

```
*SP1:  Where have you been
*SP1:  Maybe
```

***Excerpt 4**: Low ORT, low surprisal (expected pattern)*

```
*SP1:  Do you think you'll make it to my presentation tomorrow
*SP1:  That's true
```

In Excerpt 3, median ORT was high (z-score of 3.86) but surprisal was low (z-score of -1.30). This example illustrates another phenomenon: participants may have understood “maybe" to project upcoming talk and therefore waited to indicate the end of the turn. At the same time, the word “maybe" is a frequent word, resulting in a low surprisal values. In contrast, Excerpt 4 had both a low word surprisal (z-score of -1.58) and a low ORT (z-score of -2.36). This may be because “that's true" is a phrase that is both common and typically ends a turn.

# F | END-OF-TURN BASED SURPRISAL FORMULATION

In this paper, we investigate whether LLM-produced surprisal values mimic human ORTs. In the main text, we analyzed a formulation of surprisal (see Equation 2) based on the predictability of individual words within the turns. We compared this word-based surprisal to human-produced *ORT* (Section 3.4). Humans were asked to predict the end of turn through a button press task. ORT is the difference between the actual end of the turn and the moment the participants press the button, an indirect measure of the predictability of words in the turn (see Section 2.2). While there is an extensive literature to support the relationship between ORT and word predictability, this relationship is indirect. Therefore, in this appendix, we report the results of our study when analyzing surprisal based on the predictability of end-of-turn token after the second turn. Specifically, Equation 8 builds on the formalism presented in Section 2.2.2 and presents an alternative method to capture model surprisal that may more directly link to ORT.

$$Surprisal^{EoT}_{secondTurn} = -\log P(t_{EOT} \mid w^2_1, \ldots, w^2_N, w^1_1, \ldots, w^1_K) \tag{8}$$

This method calculates the probability of the *end of turn* (EoT) token *after* all the words in both the first and second turns of the two-turn stimulus. LLMs use this EoT token internally to explicitly indicate whether the model believes a turn has ended. This method considers the turn as a whole and its completion, addressing potential biases from incomplete fragments. To ensure clarity, we refer to our original formulation of surprisal as $Surprisal^{word}_{secondTurn}$ and the alternative formulation as $Surprisal^{EoT}_{secondTurn}$. Refer to Appendix A for a comprehensive explanation of the data preprocessing and fine-tuning procedures used to train the models for each surprisal method.

When performing the same RMs on $Surprisal^{EoT}_{secondTurn}$ as in Section 3.4, we found very strong evidence for the null hypothesis; the data were more likely under the model that did *not* include $Surprisal^{EoT}_{secondTurn}$ as a predictor ($BF_{10}$ = 0.03). Surprisal had a near-zero relationship with ORT ($\beta$ = 0.01, 95% CI = -0.01 - 0.03).

Since our analysis of Hypothesis 6 differed on the method used to calculate surprisal (see Equations 2 and 8), we now replicate our study based on $Surprisal^{EoT}_{secondTurn}$ and present the results below. We find that, while there were some differences based on the method used to calculate surprisal, the results point to the same conclusion: LLMs are not able to replicate human behavioral data from Warnke (2024).

## F.1 | Effect of Congruence and Speaker

Similar to our analysis in Section 3.1, we first predict $Surprisal^{EoT}_{secondTurn}$ using the same regression Equation 3. We hypothesized (Hypothesis 1) that surprisal for the incongruent stimuli would be higher compared to that for congruent stimuli.

We found decisive evidence that the best model included the main effects of speaker and congruence, as well as their interaction ($BF_{10}$ = 1.54e+04). Compared to congruent stimuli, both incongruent ($\beta$ = -0.43, 95% CI = -0.73 to -0.13) and violative ($\beta$ = -0.34, 95% CI = -0.64 to -0.04) stimuli had lower $Surprisal^{EoT}_{secondTurn}$. This finding contradicts Hypothesis 1. Additionally, stimuli in the same-speaker condition were less surprising ($\beta$ = -0.59). Finally, we observed interaction effects: within the same-speaker stimuli, both incongruent ($\beta$ = 0.88, 95% CI = 0.61 to 1.46) and violative ($\beta$ = 1.04, 95% CI = 0.61 to 1.46) stimuli were more surprising than congruent stimuli, supporting Hypothesis 1.

Interestingly, the effects of congruence and speaker identity on $Surprisal^{EoT}_{secondTurn}$ are almost opposite to their effects on $Surprisal^{word}_{secondTurn}$. While $Surprisal^{word}_{secondTurn}$ matched Hypothesis 1 in the different-speaker condition but not in the same-speaker condition, $Surprisal^{EoT}_{secondTurn}$ matched Hypothesis 1 in the same-speaker condition but not in the

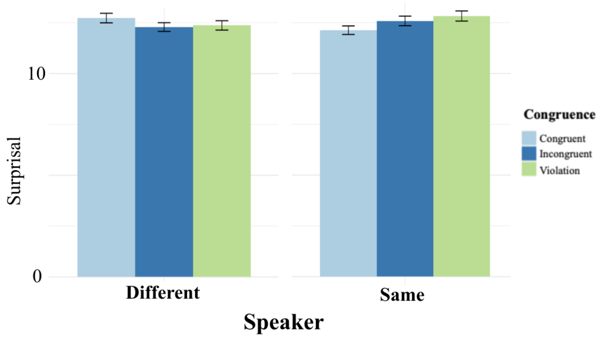


**FIGURE F1** $Surprisal^{EoT}_{secondTurn}$ across congruence and speaker conditions for GPT-2 fine-tuned on twenty-eight conversations. The results indicate that the model aligns with Hypothesis 1 in the same speaker condition, but not in the different speaker condition.

different-speaker condition. Despite these differences, neither type of surprisal matched the patterns produced in human studies across both speaker conditions.

One possible explanation for this reversal of results is that the model has different expectations regarding the length of the turn, i.e. when it would end, depending on who is speaking. Our stimuli contained very short turns with only one or two syllables, whereas the training data contained turns of varying length. It is possible that shorter turns occurred less frequently after a speaker switch compared to when the same speaker continued speaking, and that the model was therefore more surprised when turns ended, even if they were congruent. Further exploration and analysis is needed to investigate this.

## F.2 | Effect of Amount of Fine-tuning

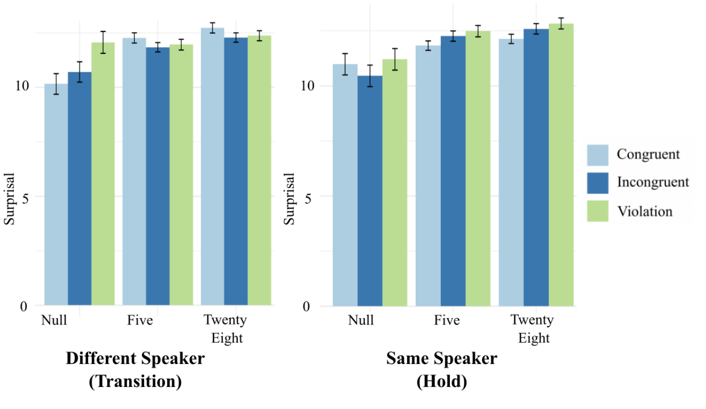


**FIGURE F2** EOT Surprisal for the null GPT2, the GPT2 trained on five conversations and the GPT2 trained on twenty-eight conversations.

In this section, we replicate the analyses from Section 3.2 and Appendix C to examine how the amount of fine-tuning data influences $Surprisal^{EoT}_{secondTurn}$. Similar to our findings with $Surprisal^{Word}_{secondTurn}$, we found decisive evidence that the data was most accurately modeled by RMs that included all main and interaction effects (See Equation4 and Table 13).

| Regression Model | Bayes Factor |
|---|---|
| Model 7 | 2.44e+49 |
| Model 8 | 2.76e+53 |
| Model 9 | 1.49e+53 |
| Model 10 | 3.19e+62 |

**TABLE 13** Bayes Factors for regression models (as described in Table 7) investigating the effect of training amount on $Surprisal^{EoT}_{secondTurn}$ patterns. The denominator for the Bayes Factors was the model that contained the baseline model (random intercept for stimulus group, main effects of congruence and speaker, and an interaction effect between congruence and speaker) and a main effect for training amount.

Fine-tuning the LLM increased baseline surprisal values: GPT-2 fine-tuned on five or twenty-eight conversations produced higher $Surprisal^{EoT}_{secondTurn}$ values than the null model. Additionally, training the models resulted in different patterns of surprisal based on speaker and congruence conditions. Models trained on five and twenty-eight conversations produced lower $Surprisal^{EoT}_{secondTurn}$ values for incongruent and violation stimuli compared to congruent stimuli in the different-speaker condition.

As shown by Figure F2, the five-conversation model did show higher baseline $Surprisal^{EoT}_{secondTurn}$ than the null model. However, the magnitude of this difference is at least three times smaller for $Surprisal^{EoT}_{secondTurn}$ than for $Surprisal^{word}_{secondTurn}$ (Section 3.2, Appendix C). Further, while the fine-tuned LLMs showed similar patterns of $Surprisal^{word}_{secondTurn}$ as the null model, fine-tuned models had different results than null models for $Surprisal^{EoT}_{secondTurn}$. However, both surprisal measures found diminishing returns as the amount of fine-tuning increased, with small, if any differences in the values for the five and twenty-eight GPT2 models.

## F.3 | Explicit versus Implicit Speaker Representations

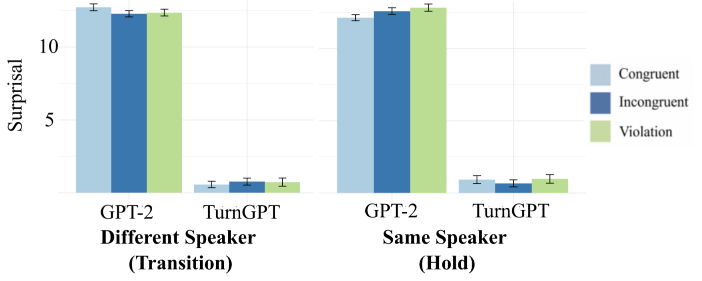


**FIGURE F3** EOT Surprisal for the TurnGPT and GPT2 trained on twenty-eight conversations.

In this Section, we replicate the analyses from Section 3.3 and Appendix D i.e., we analyze the effect of speaker representations on $Surprisal^{EoT}_{secondTurn}$. The results (see Tables 15 and 16) indicate that the data ($Surprisal^{EoT}_{secondTurn}$)

| | Estimate | t | p |
|---|---|---|---|
| (Intercept) | 10.16 (9.82 - 10.50) | 58.69 | <0.01** |
| Five | 2.09 (1.60 - 2.58) | 8.32 | <0.01** |
| Twenty-eight | 2.55 (2.06 - 3.04) | 10.16 | <0.01** |
| Incongruent | 0.55 (0.08 - 1.02) | 2.30 | 0.02* |
| Violation | 1.91 (1.44 - 2.38) | 7.97 | <0.01** |
| Same Speaker | 0.82 (0.36 - 1.29) | 3.44 | <0.01** |
| Five * Incongruent | -0.96 (-1.64 - -0.28) | -2.74 | 0.01* |
| Twenty-eight * Incongruent | -0.97 (-1.66 - -0.29) | -2.78 | 0.01* |
| Five * Violation | -2.20 (-2.89 - -1.52) | -6.28 | <0.01** |
| Twenty-eight * Violation | -2.25 (-2.94 - -1.57) | -6.43 | <0.01** |
| Five * Same Speaker | -1.25 (-1.94 - -0.57) | -3.58 | <0.01** |
| Twenty-eight * Same Speaker | -1.41 (-2.09 - -0.72) | -4.02 | <0.01** |
| Incongruent * Same Speaker | -1.08 (-1.74 - -0.42) | -3.19 | <0.01** |
| Violation * Same Speaker | -1.68 (-2.35 - -1.02) | -4.98 | <0.01** |
| Five * Incongruent * Same Speaker | 1.90 (0.93 - 2.87) | 3.84 | <0.01** |
| Twenty-eight * Incongruent * Same Speaker | 1.94 (0.97 - 2.91) | 3.92 | <0.01** |
| Five * Violation * Same Speaker | 2.65 (1.68 - 3.61) | 5.34 | <0.01** |
| Twenty-eight * Violation * Same Speaker | 2.74 (1.77 - 3.70) | 5.53 | <0.01** |

**TABLE 14** Coefficients for frequentist regression including all two- and three-way interactions. 95% confidence intervals presented in parentheses. * = p-value under 0.05, ** = p-value under 0.01.

were most likely under the regression model (RM) that included all main and interaction effects (see 5). We found decisive evidence that this model was more likely than the next best RM ($BF_{10}$ = 500). TurnGPT produced lower surprisal values ($\beta$ = -12.15, 95% CI = -12.49 to -11.81) compared to GPT-2. Furthermore, the relationship between surprisal and congruence condition depended on the type of speaker representations. Specifically, for TurnGPT, the incongruent ($\beta$ = 0.64, 95% CI = 0.17 to 1.12) and violation ($\beta$ = 0.52, 95% CI = 0.05 to 1.00) conditions had even higher surprisal compared to the congruent condition. Additionally, $Surprisal^{EoT}_{secondTurn}$ values were also higher in the same-speaker condition for TurnGPT.

Surprisal was lower for TurnGPT than for GPT-2, regardless of the method of calculating surprisal. However, we found that the pattern of surprisal across conditions differed between TurnGPT and GPT-2 – but only when analyzing $Surprisal^{EoT}_{secondTurn}$. This may be due to GPT-2 producing different patterns when producing $Surprisal^{EoT}_{secondTurn}$ and $Surprisal^{word}_{secondTurn}$. For the different-speaker condition, GPT-2 found the incongruent condition to have higher $Surprisal^{word}_{secondTurn}$, but lower $Surprisal^{EoT}_{secondTurn}$ than the congruent condition. For the same-speaker condition, GPT-2 found the opposite: the incongruent condition had lower $Surprisal^{word}_{secondTurn}$, but higher $Surprisal^{EoT}_{secondTurn}$. However, future research is needed to more deeply understand the root causes of these differences.

| Regression Model | Bayes Factor |
|---|---|
| Model 12 | 0.48 |
| Model 13 | 0.18 |
| Model 14 | 0.08 |
| Model 15 | 582.71 |

**TABLE 15** Bayes Factors for regression models (described in Table 10) investigating the effect of embedding type (TurnGPT vs. GPT-2 embedding) on surprisal patterns. The data were so unlikely under the null model (that did not contain model type as a predictor) that the resulting Bayes Factors were too large to compute. Therefore, the denominator for these Bayes Factors is the model that contained the baseline model (random intercept for stimulus group, main effects of congruence and speaker, and an interaction effect between congruence and speaker) and a main effect for model type (Model 11 in Table 10).

| | Estimate | t | p |
|---|---|---|---|
| (Intercept) | 12.72 (12.46 - 12.98) | 95.37 | <0.01** |
| TurnGPT | -12.15 (-12.49 - -11.81) | -70.13 | <0.01** |
| Incongruent | -0.45 (-0.79 - -0.10) | -2.52 | 0.01* |
| Violative | -0.36 (-0.70 - -0.01) | -2.03 | 0.04* |
| Same Speaker | -0.60 (-0.94 - -0.26) | -3.40 | <0.01** |
| TurnGPT * Incongruent | 0.64 (0.17 - 1.11) | 2.65 | 0.01* |
| TurnGPT * Violation | 0.52 (0.05 - 0.99) | 2.15 | 0.03* |
| TurnGPT * Same Speaker | 0.95 (0.48 - 1.43) | 3.95 | <0.01** |
| Incongruent * Same Speaker | 0.91 (0.43 - 1.40) | 3.65 | <0.01** |
| Violation * Same Speaker | 1.07 (0.58 - 1.55) | 4.28 | <0.01** |
| TurnGPT * Incongruent * Same Speaker | -1.36 (-2.03 - -0.70) | -3.98 | <0.01** |
| TurnGPT * Violation * Same Speaker | -1.18 (-1.85 - -0.51) | -3.45 | <0.01** |

**TABLE 16** Results for most complex regression model analyzing how speaker representations predict $Surprisal^{EoT}_{secondTurn}$ (Model 15 in Table 10). 95% confidence intervals presented in parentheses. * = p-value under 0.05, ** = p-value under 0.01.

## F.4 | Analysis of Individual Stimuli

When performing the same RMs as in Section 3.4 on $Surprisal^{EoT}_{secondTurn}$, we found very strong evidence for the null hypothesis, that the data were more likely under the model that did *not* include $Surprisal^{EoT}_{secondTurn}$ as a predictor. To generate potential hypotheses to explain this finding, we present the same analysis of individual stimuli as in Appendix E, but based on the $Surprisal^{EoT}_{secondTurn}$. Specifically, we examined stimuli where $Surprisal^{EoT}_{secondTurn}$ z-scores were opposite of median ORT z-scores. Stimuli that matched Hypothesis 6 had $Surprisal^{EoT}_{secondTurn}$ z-scores in the same direction and magnitude as its ORT.

First, we explored stimuli that did not match our hypothesis. In Excerpt 5, TurnGPT produced a high $Surprisal^{EoT}_{secondTurn}$, but human produced low ORTs. In Excerpt 6, TurnGPT produced a low $Surprisal^{EoT}_{secondTurn}$, but ORTs were high.

In Excerpt 7, both $Surprisal^{EoT}_{secondTurn}$ (z-score of 3.96) and ORT (z-score of 3.85) were high, while $Surprisal^{word}_{secondTurn}$

***Excerpt 5**: Low ORT, high $Surprisal^{EoT}_{secondTurn}$ (unexpected pattern)*

```
*SP1:  I got you a present
*SP2:  Stay safe
```

***Excerpt 6**: High ORT, low $Surprisal^{EoT}_{secondTurn}$ (unexpected pattern)*

```
*SP1:  Do you mind helping with my homework
*SP2:  Please
```

***Excerpt 7**: High ORT, high $Surprisal^{EoT}_{secondTurn}$ (expected pattern)*

```
*SP1:  Where have you been
*SP1:  Maybe
```

***Excerpt 8**: Low ORT, low $Surprisal^{EoT}_{secondTurn}$ (expected pattern)*

```
*SP1:  Where have you been
*SP2:  Nowhere
```

(z-score of -1.29) was low. In Excerpt 8, both $Surprisal^{EoT}_{secondTurn}$ (z-score of -0.49) and ORT (z-score of -0.96) were somewhat lower, while $Surprisal^{word}_{secondTurn}$ was above average (z-score of 0.59). Exactly why $Surprisal^{EoT}_{secondTurn}$ differs from $Surprisal^{word}_{secondTurn}$, and exactly when each measure corresponds with human ORT, is still unclear and should be investigated in future work.