LLMs syntactically adapt their language use to their conversational partner

Anonymous ACL submission

Abstract

It has been frequently observed that human speakers align their language use with each other during conversations. In this paper, we study empirically whether large language models (LLMs) exhibit the same behavior of conversational adaptation. We construct a corpus of conversations between LLMs and find that two LLM agents end up making more similar syntactic choices as conversations go on, confirming that modern LLMs adapt their language use to their conversational partners in at least a rudimentary way.

1 Introduction

It has been documented broadly that when humans talk to each other, they adapt their language use to their communication partners by coordinating their behavior and language. Humans *align* not only their gestures, posture, and speech rate (Holler and Wilkin, 2011; Shockley et al., 2009; Jungers and Hupp, 2009), but also their linguistic decisions at deeper levels, such as semantics and syntax (Bock, 1986; Garrod and Anderson, 1987). In other words, the distribution over syntactic structures of two human speakers becomes more similar as a conversation progresses.

In this paper, we investigate whether large language models (LLMs) adapt their syntactic choices to their conversational partners as well. While it is well known that LLMs can be explicitly prompted towards embodying different "personas" and changing the style of the language they generate (Deshpande et al., 2023; Thillainathan and Koller, 2025), it is unclear whether merely being present in a conversation with an interlocutor is sufficient to make LLMs adapt their language use to their interlocutor's. The ability to adapt to the communication partner's language is associated with increased success in goal-oriented conversations (Reitter and Moore, 2014), and it enables a dialogue system to meet a user's language use rather than requiring the user to adapt to the system (Schlangen, 2022). Language models will only serve as effective foundations for dialogue systems if they prove capable of implicitly adapting to a user's language.

To this end, we create a new dataset of conversations between LLMs in which both LLMs are prompted to initially exhibit different language use. We then measure the dynamics of syntactic language adaptation over the course of the conversations, using a method adapted from the human-human analysis of Reitter and Moore (2014). We find that GPT-4o (Hurst et al., 2024) conversations show statistically significant adaptation when comparing syntactic repetitions within conversations against repetitions across conversations, replicating Reitter's findings for human conversations. We further show this is a continuous process active throughout conversations and conclude by discussing whether these findings demonstrate "human-like" alignment in LLMs.

2 Background

As we mentioned above, humans adapt their language use to their communication partners across various linguistic levels. In this paper, our focus is on *syntactic* adaptation: Do the distributions over the syntactic structures that two interlocutors produce become more similar over the course of a conversation?

In the psycholinguistics literature on human communication, two separate (but not exclusive) mechanisms have been proposed to explain the mutual adaptation of language use. Rasenberg et al. (2020) contrast two theoretical views that explain the process through *alignment* on different cognitive levels: on a conscious level, in which cooperative decisions establish a situational common ground (Brennan and Clark, 1996), and a subconscious level, in which automatic *priming* leads to aligned representational states (Pickering and Garrod, 2004). In psycholinguistics, priming refers to a process in which encountering a word or construction temporarily increases the activation of a cognitive representation, thereby increasing the probability for the word or construction to be reproduced.

In this paper, we study the conversational behavior of artificial, LLM-based agents. We will primarily focus on the level of outwardly observable changes to the language use and describe it with the theory-neutral word *adaptation*. We will discuss in Section 5 the extent to which concepts like alignment and priming can apply to LLMs.

Related Work. The only study to date examining adaptation in LLMs (Cai et al., 2024) focused on short-term syntactic adaptation. They showed that ChatGPT and Vicuna are more likely to complete a sentence with a double object or a prepositional object when primed with a sentence of the respective type. We extend this research to long conversations with natural sentences rather than carefully constructed one-sentence stimuli.

3 Measuring human-human adaptation

The phenomenon of long-term syntactic adaptation was first measured on corpora of human-human dialogues by Reitter and Moore (2014). The basic idea is to determine whether the usage frequency of a syntactic structure (specifically, a rule in a context-free grammar) in the first half of a conversation has a statistical impact on its frequency in the second half.

We follow Reitter in splitting each conversation in a dialogue corpus into two parts. We call the first 49% of each conversation PRIME and the last 49% of each conversation TARGET; the middle 2% are discarded to ensure that we measure long-term adaptation as opposed to short-term priming. On corpora that are not already syntactically annotated, we parse each conversation with the Neural Berkeley Parser[1] (Kitaev and Klein, 2018; Kitaev et al., 2019), to obtain a set of context-free production rules for the PRIME and TARGET section of each conversation, respectively.

Adaptation takes place if rule repetitions are more likely between the PRIME and TARGET of the same conversation (where adaptation is possible), compared to a PRIME and TARGET of different conversations (where no adaptation could have taken place). To make this comparison, we draw two samples for each rule across the TARGETs of all conversations: one for which we check whether the rule has been uttered by the other speaker in the PRIME of the same conversation, and the other for which we check whether the rule has been uttered by a speaker in a random, unrelated conversation. We encode the presence of a rule in a binary variable *Prime* for each sample, which is 1 if the rule is present. Another binary variable, *SameConv*, is used to indicate whether we looked for a prime in the same conversation (1) or in a different, random conversation (0). If repetitions are more likely between speakers within conversations, such that we see an effect of *SameConv* on *Prime*, we take that as evidence of cross-speaker adaptation.

We further include features representing the log-frequency of rules across all conversations (*ln(Freq)*), as more frequent rules are expected to be more likely to appear in any PRIME, and a variable *ln(Size)*. This second variable encodes the amount of different rules that a speaker used in the PRIME of a conversation, i.e. the size of the set of rules that we use to look for a prime; a larger set increases the probability of any rule to occur. We follow Reitter and Moore (2014) in excluding rules that appear only once in the whole dataset and rules that have disproportionately high frequencies (around 0.3% of each dataset), because these rules are never primed or almost always primed. Including these rules in the analysis does not substantially change the results (see Appendix D). We further remove structures that are lexically identical.

Our analysis differs from Reitter's original method in two aspects. First, we consider only overlaps between rule uses in TARGET with uses in PRIME by the other speaker. This eliminates effects that solely stem from speaker idiosyncrasies or the conditioning of LLM-generated language on its own prior output. Second, our analysis includes the set size of rules used to check for a prime.

3.1 Alignment in human conversations

We replicate Reitter's results on human-human conversations to ensure that we obtain comparable results after our modifications. We use the method described above to analyze the Switchboard corpus (Marcus et al., 1994), which comprises 650 syntactically annotated telephone conversations (see Fig. 2 in Appendix B for an overview of its composition). This is in contrast to Reitter's work,

[1] We used the benepar_en3_large model of the benepar python package for parsing and spacy's en_core_web_md model for tokenization.

which used the HCRC Map Task corpus (Anderson et al., 1991), consisting of task-oriented conversations. By looking at Switchboard as opposed to Map Task, we demonstrate alignment effects on non-task-oriented conversations, facilitating comparison with LLM-generated conversations, and we make use of hand-annotated rather than automatically parsed syntactic structures.

We fit a mixed-effects logistic regression to the sampled data using the generalized linear mixed models (GLMM) of Python's pymer4 (v0.8.2) package. We included a nested random intercept for conversations and speakers and a random slope for *ln(Freq)* and centered fixed effects except *SameConv*. We selected the model through a backward selection process. Results are shown in Table 1.

We find that *SameConv* ($\beta = 0.228, p < 0.001$) has a significant positive effect, replicating Reitter's findings that humans align syntactically to their partners over the course of a conversation. A thorough analysis of the interactions can be found in Appendix E.

4 Measuring LLM-LLM adaptation

We follow the same method to analyze syntactic adaptation in conversations of GPT-4o.

Dataset. One challenge towards this goal is the availability of a suitable dataset of LLM conversations. We require a dataset consisting of long natural conversations (with no intervening task prompts) in which the speakers use varying syntactic structures to make adaptation possible and evenly distributed utterance lengths.

Existing datasets of conversations with LLMs do not satisfy these requirements. UltraChat (Ding et al., 2023) is a dataset of LLM-LLM conversations, but these conversations follow simple question-answering between a user and a model "persona". Conversations are too short and there is no variability between the language use across conversations. By contrast, available datasets of human-LLM conversations, such as WildChat (Zhao et al., 2024), consist of conversations that each have unique instructions by the user. This makes conversations incomparable and therefore unsuitable for a statistical analysis of adaptation.

We therefore created our own dataset by letting GPT-4o[2] converse with itself.[3] We created 17 different conversational agents with identical system prompts, except for an initial specification of a "language persona" that is unique to each agent. We then generated conversations between pairs of LLM agents by iteratively prompting each of them for the next utterance, including the context of the conversation history. Iterations were stopped, once a conversation surpassed a predefined length threshold. All prompts for managing the conversations and defining the language personas can be found in Appendix A.

To ensure sufficient variety in the agents' language use, we further generated conversations where each agent conversed with itself. We then calculated how often each syntactic rule was used and normalized these frequencies to create a discrete probability distribution of syntactic rules for each agent. To compare these distributions, we measured their distances using the Jensen-Shannon divergence (JSD). See Figure 5 in Appendix C for details. The results confirm a high degree of syntactic variety, with JSD values of up to 0.69.

Adaptation in LLM conversations. We generated 136 conversations by pairing up every conversational agent with every other conversational agent, all on the topic "What makes a day a good day?" Twelve conversations ended in repeating patterns; we excluded them and used the remaining 124 other conversations to form the GPT corpus. The distribution of conversation and utterance lengths closely mirrors that of the Switchboard corpus (cf. Fig. 3 and Fig. 2 in Appendix B).

We ran the analysis described in Section 3.1 on the GPT corpus, taking care not to sample from identical agents in other conversations. Fixed effects, except *SameConv*, are centered. The model was selected using backward selection. The results are shown in Table 1.

SameConv has a significant positive effect on *Prime* ($\beta = 0.198,\ p < 0.001$), showing that there is syntactic adaptation. Interactions are discussed in more detail in Appendix E.

Fine-grained tracking of the adaptation process. To gain a deeper understanding of the adaptation process performed by the LLM, we performed a fine-grained analysis of adaptation over the course of the conversation. As in the approach above, we directly compared the distributions of syntactic structures used by two different agents; however, this time, we focused on comparing the distributions to see how they evolve throughout a conversation. To obtain reliable estimates of the distribu-

[2] We used GPT-4o-2024-08-06 with default parameters.

[3] We will make this dataset available upon acceptance.

	Switchboard Corpus				GPT Corpus			
	β	SE	z	$p > \|z\|$	β	SE	z	$p > \|z\|$
Intercept	-2.927	0.018	-158.8	0.000	-2.031	0.048	-42.5	0.000
ln(Freq)	1.174	0.008	143.2	0.000	1.275	0.028	45.6	0.000
SameConv	0.228	0.023	9.9	0.000	0.198	0.056	3.5	0.000
ln(Size)	1.402	0.033	41.9	0.000	1.175	0.107	11.0	0.000
ln(Freq):SameConv	-0.101	0.01	-9.8	0.000	-0.146	0.035	-4.2	0.000
ln(Freq):ln(Size)	0.068	0.015	4.7	0.004	0.266	0.062	4.3	0.000

Table 1: The regression models for the Switchboard corpus (left) and the GPT corpus (right).

tions, we created 520 conversations between agents 5 and 6, a pair of agents with moderate initial JSD (cf. Fig. 5), while keeping the topic the same (cf. Appendix A). Due to repeating patterns, we excluded 14 conversations.

To observe how the similarity of the two agents' distributions evolves, we split the remaining 506 conversations into sections of 200 words (see Fig. 4 in Appendix B for an overview of the data), and compare the distributions of the two agents for each split. As above, distributions are calculated by normalizing rule frequencies across all 506 conversations. To estimate the variance of these calculations, we perfom them on 100 bootstraps of the data. Each bootstrap consisted of 506 randomly selected conversations, drawn with replacement. We report the means and standard deviations of these 100 JSD values across splits in Fig. 1.

We find that the mutual adaptation of the two LLM agents is a gradual process that persists throughout the conversation. The rate of adaptation is relatively constant, with the strongest adaptation happening in the first split.

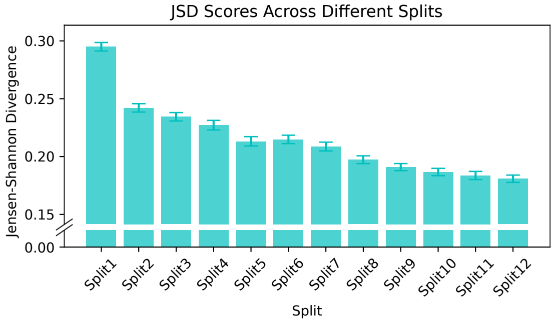


Figure 1: Jensen-Shannon divergence scores between agents 5 and 6 across splits of conversations.

5 Discussion

Throughout the paper, we have avoided using the words “alignment” and “priming” for the LLM's adaptation process to steer clear of any connotations about human cognitive processes. While we have established that the LLM's syntax becomes increasingly similar to its conversational partner's, this does not necessarily mean that this process is driven by a similar underlying mechanism.

An LLM does not maintain an explicit mental model of its interlocutor's language use and does not make conscious decisions on coordinating it with its interlocutor. Thus it seems inappropriate, under the notion of alignment sketched in Section 2, to explain the LLM's adaptive behavior as alignment. At the same time, priming effects in humans are usually assumed to impact their language use only in the short term. One conceivable explanation for GPT-4o's ability to perform long-term adaptation is that it conditions the language it produces on the previous conversation (a mechanism that is similar to priming in humans), but has a much larger capacity than humans for remembering the verbatim conversational context.

This notion goes in support of our second experiment: A gradual adaptation that appears with increased context length underpins the intuition that LLMs can adapt to longer contexts, and that this increases its influence. Different from humans, short-term effects, like those reported in Cai et al. (2024), may therefore be driven by the same principles as long-term adaptation in LLMs. A more detailed analysis would be an interesting avenue of future research.

6 Conclusion

We showed that GPT-4o can gradually adapt its language use to its conversational partner, to an extent that is similar to what we observe in human-human conversations. This observation goes beyond previous findings, which indicated that an LLM's language use can be controlled through explicit instructions and influenced by priming from the previous utterance. A more detailed comparison of the mechanisms that humans and LLMs use to achieve such long-term adaptation is an interesting avenue of future work.

7 Limitations

This work focuses on texts generated with GPT-4o. We decided to use this model, as it is one of the highest performing accessible models. While the findings of Cai et al. (2024) suggest that the results generalize to other models, e.g. Vicuna, further research is needed to confirm this.

Our study concentrates on syntactic structures of the English language. Similar effects may exist for other languages and other linguistic features, also of different modalities (e.g. intonation, speech rate). Furthermore, in this study we controlled for topicality by keeping the topic of all conversations identical. It is unclear whether topicality has an effect on syntactic structures, but there is evidence that lexical choices influence the syntax at least to some extent (lexical boost, Cai et al., 2024). Future work would have to investigate further how the reported effects translate to more diverse conversational settings.

The analysis that we adapt from Reitter and Moore (2014) loses information by encoding the presence of syntactic structures in a binary variable. We hypothesize that this leads to the interaction between *ln(Freq):SameConv* (see Apeendix E). While the analysis is suitable for capturing adaptation in general, it lacks the sensitivity to account for the occurrence rate of rules in a meaningful way.

8 Ethical Considerations

We believe that our work is unlikely to have an immediate ethical or societal impact. However, there is potential that the reported effects serve as a footprint of LLM generated texts – we didn't prompt the model to adapt to the language, but this effect appears inherently. This potentially leads to patterns that are intrinsic to LLMs, which could be leveraged to detect LLM generated texts.

References

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The hcrc map task corpus. *Language and Speech*, 34(4):351–366.

J.Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

S. E. Brennan and H. H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1482–1493.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56, Bangkok, Thailand. Association for Computational Linguistics.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

Judith Holler and Katie Wilkin. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35(2):133–153.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, ..., and Yury Malkov. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Melissa K. Jungers and Julie M. Hupp. 2009. Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes*, 24(4):611–624.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen

Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, 44(11):e12911.

David Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

David Schlangen. 2022. Norm participation grounds language. In *Proceedings of (Dis)embodiment 2022: A CLASP Conference*.

Kevin Shockley, Daniel C. Richardson, and Rick Dale. 2009. Conversation and coordinative structures. *Topics in Cognitive Science*, 1(2):305–319.

Sarubi Thillainathan and Alexander Koller. 2025. Fine-grained controllable text generation through in-context learning with feedback. In *AAAI 2025 Workshop on AI for Education – Tools, Opportunities, and Risks in the Generative AI Era*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *Preprint*, arXiv:2405.01470.

A Prompts

A.1 System Prompt

The following template was used as the system prompt in the data generation process:

> You are in a conversation. There are two speakers, SpeakerA and SpeakerB.
> You are SpeakerA. The conversation will consists of turns in the form:
> [SpeakerA's utterances]
> [SpeakerB's utterances]
> [SpeakerA's utterances]
> . . .
> You need to only give [SpeakerA's utterances]. You will be prompted by [Language] that will instruct you on the language that you shall use as SpeakerA. Further, you will be prompted by [Topic], the topic of the conversation. Behave as in a normal conversation with SpeakerB to discuss the [Topic]. [Language] {That agent's specific persona, see item A.2)}. [Topic] What makes a day a good day?

A.2 Language Personas

The following language personas were used to vary the language of each agent. Language personas are inserted into the system prompt at the designated position.

1. Your language is precise, and unambiguous. You use clear and simple sentences.
2. Your language is gentle and thoughtful. You use concise and not overly complex sentences, to convey meaning efficiently.
3. Your language is dynamic, and provocative. You often use vivid metaphors.
4. Your language is introspective, and deliberate. You use contemplative phrasing.
5. Your language is smooth and reassuring. You employ gentle pauses and a steady rhythm.
6. Your language is analytical and precise. You use complex sentence structures sparingly, preferring clear, well-organized sentences.
7. Your language is conversational and warm. You use relaxed, varied sentence structures that mirror casual speech, inviting readers into an open, friendly dialogue.
8. Your language is inquisitive and reflective. You frequently use open-ended questions and layered sentences that encourage readers to pause and ponder.
9. Your language is poetic and evocative. You lean into complex, image-rich sentences that build vivid scenes and sensations, letting metaphors flow freely.
10. Your language is structured and methodical. You rely on orderly, sequential sentences that build upon each other in a clear, logical progression, guiding readers through a well-defined thought process.
11. Your language is hesitant and unsure. You use fragmented sentences and trailing thoughts, leaving ideas partially formed, as if questioning each phrase.
12. Your language is overly cautious and repetitive. You tend to rephrase ideas multiple times in a single sentence.
13. Your language is anxious and scattered. You jump between ideas mid-sentence, creating a disjointed flow that feels hurried and restless.
14. Your language is straightforward, and no-nonsense. You avoid fluff and filler.
15. Your language is crisp and engaging. You

use short, impactful sentences to create emphasis.

16. Your language is bold and unapologetic. You rely on direct, declarative sentences that avoid qualifiers.
17. Your language is understated and subtle. You use concise sentences that suggest rather than state.

B Dataset Compositions

Statistics of the Switchboard corpus and the conversations generated with GPT-4o are shown in Figures 2 and 3. The composition of the conversations between agents 5 and 6 can be seen in Figure 4. The agents were chosen, as they provide very even turn lengths. This allows for similarly good estimations of the distributions of their used syntactic rules.

The cost for generating all conversations using OpenAI's API was around 100$.

C Base Divergence Values between Agents

In our study, we compared the distributions of rules that agents use throughout conversations using the Jensen-Shannon divergence as distance measurement. To place our reported values in context, we provide baseline divergence values between each agent in this appendix. For each agent, we calculated the probability distribution of their uttered rules from 10 conversations with themselves. We counted the number of occurrences of each rule and normalized these frequencies for each agent to create a discrete probability distribution. The topic of all conversations was kept identical: ”What makes a day a good day?” Conversations were created turn by turn and stopped once they surpassed a length of 800 words (see section 4). Agents 14-17 are excluded, as their conversations converged to short repeating patterns. The resulting JSD values between the distributions of each agent are shown in Figure 5.

D Analysis with all Rules

In our analysis, we exclude rules that have very high frequencies, and those that appear only once. To test whether removing these rules affects overall conclusions, we ran the analysis again using all rules. Results can be found in Table 2 for Switchboard and in Table 3 for the GPT corpus.

The results show that effects still persist with similar effect sizes. The only difference is that significance values are lower. For the GPT corpus, for example, the p-values for *SameConv*, *ln(Freq):SameConv*, and *ln(Freq):ln(Size)* are $p < 0.004$, $p < 0.002$, and $p < 0.012$ respectively, which are much larger than the above recorded $p < 0.000$ for all effects.

This shows that including the rules only inflates the sample space with samples that have identical values for *Prime* for both $SameConv = 0$ and $SameConv = 1$.

E Interaction Effects

In our analysis of Switchboard we find the interaction between *ln(Freq)* and *SameConv* ($\beta = -0.101, p < 0.001$), which is similar to the effect size of the same interaction in the GPT corpus ($\beta = -0.146, p < 0.001$). The effect could be taken to explain that adaptation appears more for less frequent rules[4] compared to more frequent rules. In theory, this sounds like a plausible explanation. However, we hypothesize that the effect may primarily come from our binary encoding of *Prime*: a higher frequent rule reduces the effect that *SameConv* has on *Prime*, because it is more likely to appear in other conversations ($SameConv = 0$). The binary encoding omits information of the rate at which rules appear, making an increased rate of rules in the PRIME and TARGET invisible. Therefore, the analysis cannot capture adaptation if it manifests as an increase in occurrence rate between the PRIME and TARGET of the same conversation, which reduces sensitivity to more frequent rules.

The effect size of the interaction between the log rule frequency *ln(Freq)* and the log number of unique rules in the prime *ln(Size)* differs between GPT ($\beta = 0.266$, $p < 0.001$) and Switchboard ($\beta = 0.068$, $p < 0.004$). The larger effect in the GPT corpus suggests a more stochastic mechanism. To illustrate this, we can think about the appearance of syntactic rules, as if they were drawn at random (with replacement) from an underlying probability distribution over rules. Adaptation would mean that drawing a rule increases its probability to be drawn again. In this context, we can think of *ln(Size)* as the logarithmic number of rules that we draw. For a rule with probability p, doubling the *Size* results in an increase of p by $1 - (1 - p)^2 - p = p(1 - p)$. It holds that:

[4] less than the mean frequency, as the data is centered

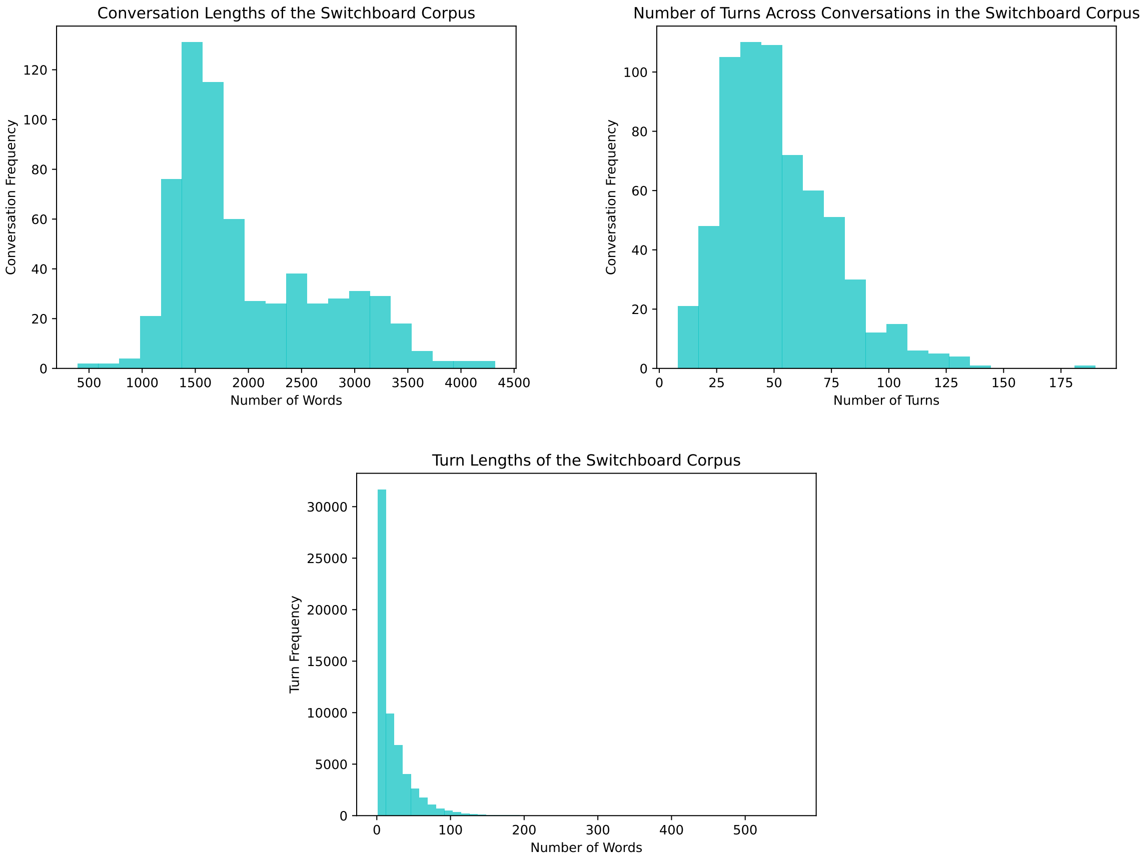


Figure 2: Statistics of the Switchboard Corpus.

	β	SE	z	$P > \|z\|$
Intercept	-3.537	0.022	-159.723	0.000
ln(Freq)	1.202	0.008	149.061	0.000
SameConv	0.263	0.027	9.693	0.000
ln(Size)	1.473	0.039	38.228	0.000
ln(Freq):SameConv	-0.103	0.010	-10.147	0.000
ln(Freq):ln(Size)	0.025	0.014	1.821	0.069

Table 2: The regression model for the Switchboard corpus including all rules.

	β	SE	z	$P > \|z\|$
Intercept	-2.255	0.051	-44.013	0.000
ln(Freq)	1.297	0.0260	50.582	0.000
SameConv	0.173	0.061	2.847	0.004
ln(Size)	1.361	0.116	11.724	0.000
ln(Freq):SameConv	-0.101	0.033	-3.053	0.002
ln(Freq):ln(Size)	0.140	0.056	2.501	0.012

Table 3: The regression model for the GPT corpus including all rules.

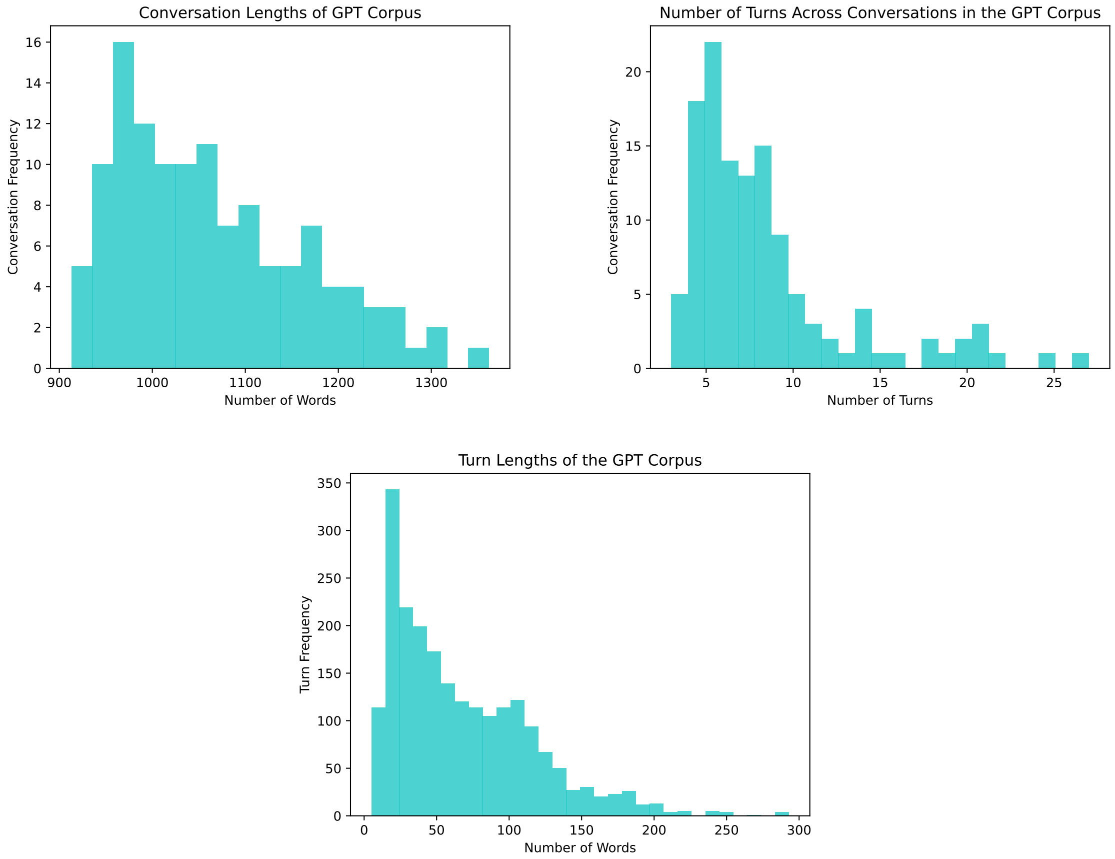


Figure 3: Statistics of the 136 conversations between agents generated with GPT-4o (GPT Corpus).

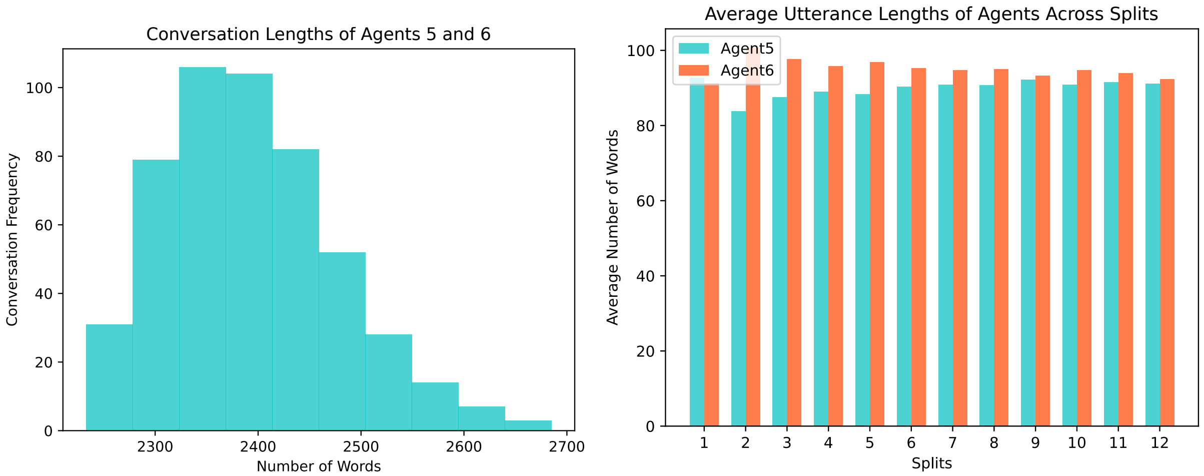
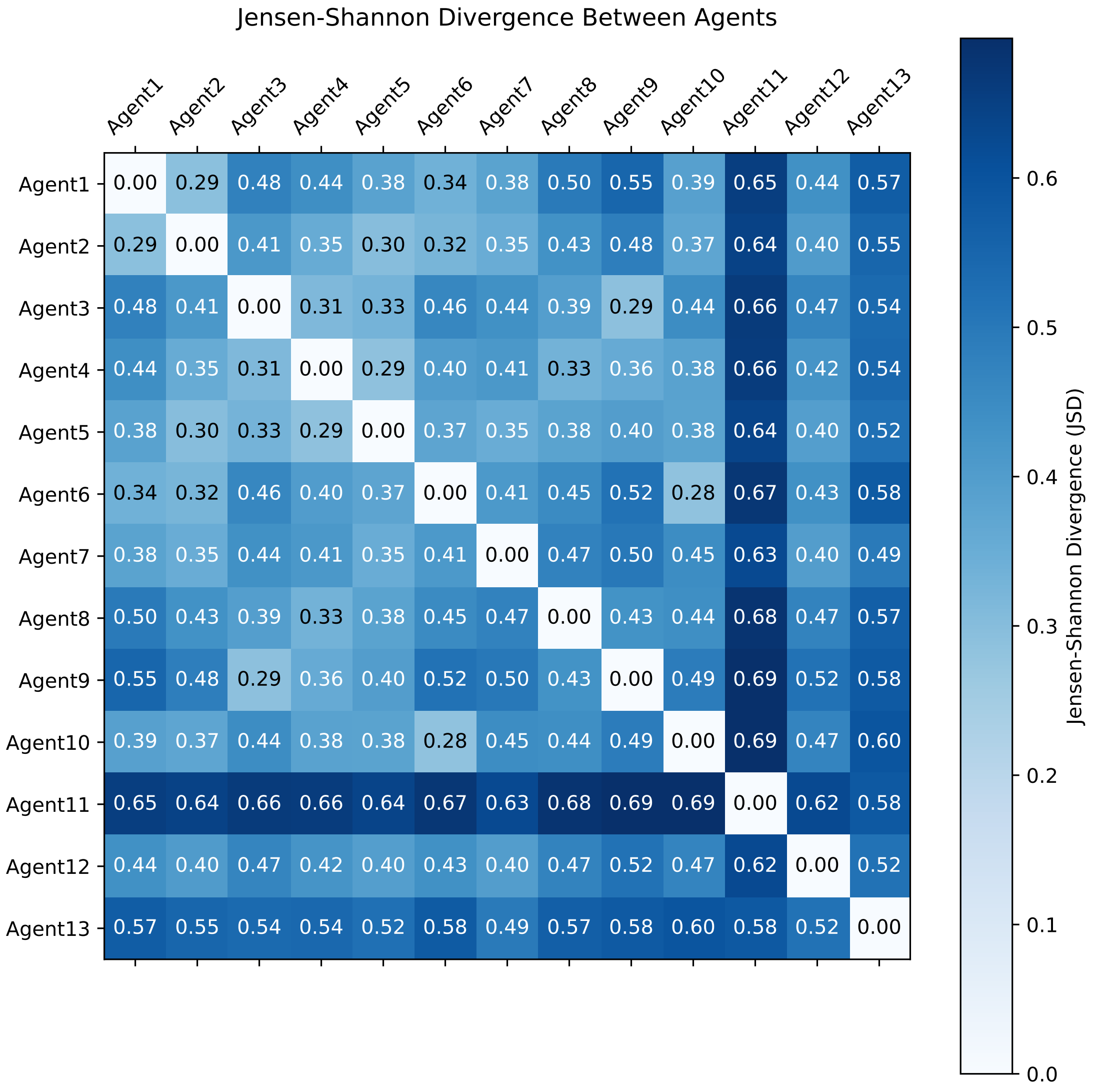


Figure 4: Statistics of the 506 conversations between agents 5 and 6 generated with GPT-4o.



	Agent1	Agent2	Agent3	Agent4	Agent5	Agent6	Agent7	Agent8	Agent9	Agent10	Agent11	Agent12	Agent13
Agent1	0.00	0.29	0.48	0.44	0.38	0.34	0.38	0.50	0.55	0.39	0.65	0.44	0.57
Agent2	0.29	0.00	0.41	0.35	0.30	0.32	0.35	0.43	0.48	0.37	0.64	0.40	0.55
Agent3	0.48	0.41	0.00	0.31	0.33	0.46	0.44	0.39	0.29	0.44	0.66	0.47	0.54
Agent4	0.44	0.35	0.31	0.00	0.29	0.40	0.41	0.33	0.36	0.38	0.66	0.42	0.54
Agent5	0.38	0.30	0.33	0.29	0.00	0.37	0.35	0.38	0.40	0.38	0.64	0.40	0.52
Agent6	0.34	0.32	0.46	0.40	0.37	0.00	0.41	0.45	0.52	0.28	0.67	0.43	0.58
Agent7	0.38	0.35	0.44	0.41	0.35	0.41	0.00	0.47	0.50	0.45	0.63	0.40	0.49
Agent8	0.50	0.43	0.39	0.33	0.38	0.45	0.47	0.00	0.43	0.44	0.68	0.47	0.57
Agent9	0.55	0.48	0.29	0.36	0.40	0.52	0.50	0.43	0.00	0.49	0.69	0.52	0.58
Agent10	0.39	0.37	0.44	0.38	0.38	0.28	0.45	0.44	0.49	0.00	0.69	0.47	0.60
Agent11	0.65	0.64	0.66	0.66	0.64	0.67	0.63	0.68	0.69	0.69	0.00	0.62	0.58
Agent12	0.44	0.40	0.47	0.42	0.40	0.43	0.40	0.47	0.52	0.47	0.62	0.00	0.52
Agent13	0.57	0.55	0.54	0.54	0.52	0.58	0.49	0.57	0.58	0.60	0.58	0.52	0.00

Figure 5: Jensen-Shannon divergence values between agents. See section A for an overview of their different language prompts. Agents 14-17 were excluded due to repeating patterns in their conversations.

$$p_1(1 - p_1) < p_2(1 - p_2),$$

for $0 < p_1 < p_2 < 0.5$[5]. Therefore, the probability for high frequency rules increases more with an increase of *Size*. While the inclusions of *ln(Freq)* and *ln(Size)* already account for this effect, the increase in repetition probability also raises the likelihood of further increases due to adaptation, which is captured by the interaction term. The effect therefore highlights the stochasticity in GPT-4o's behavior.

[5] The most frequent rule has a probability of 0.087, which is much lower than 0.5.