Syntactic Alignment in Conversations with Large Language Models: Do LLMs Adapt their Syntax Over the Long Term Similar to Humans?

Anonymous ACL submission

Abstract

This paper explores the effects of long-term syntactic alignment in Large Language Models (LLMs). Using OpenAI's GPT-4o, artificial conversations were generated, addressing a lack in existing research of long natural conversations with LLMs. A statistical analysis on syntactic structures present in these conversations reveals that syntactic alignment occurs in LLMs over extended periods. A second analysis further explores how the process of alignment evolves throughout a conversation, showing that LLMs progressively adjust their syntax, with the largest changes occurring early on. The results indicate that LLMs are not only influenced by the linear order in which tokens of their inputs appear, but also that its influence becomes continuously larger with increasing context lengths.

1 Introduction

Alignment in human language and communication is a widely studied process, in which people adapt to their communication partner by coordinating their behavior and language. These adaptation processes not only appear on a visual or auditory level, such as gestures, postures or the speech rate (Holler and Wilkin, 2011, Shockley et al., 2009, Jungers and Hupp, 2009), but also on more underlying levels, e.g. the semantics or syntax (Bock, 1986, Garrod and Anderson, 1987). Under these latter two aspects, artificial language generation has become almost indistinguishable from human language in recent years; Large Language Models (LLMs) are optimized to produce texts that seem as coherent as human language, yet their linguistic behavioral patterns haven't been studied much. Different from humans, LLMs work on a next-token-prediction task: They don't follow a conscious effort to convey meaning, but rather model a certain probability function to generate texts. Their concrete underlying workings are unknown, as they emerge from an optimization process on large amounts of data, making it unclear which patterns they have picked up on during that training.

Although they are never explicitly guided to exhibit behavior similar to humans, do Large Language Models nonetheless exhibit syntactic alignment in their text production, similar to us?

1.1 Priming and Alignment

Research on human adaptation processes in language and communication has covered many different aspects. For this reason the terminology in this field is quite varied; Adaptation processes have been found in different modalities (e.g. gestures, lexical phrases...) over different temporal distances or under different sequential relations. Instead of referring to all these variations with different expressions, Rasenberg et al., 2020 propose to simply distinguish them under the different aspects that they are analyzed on, avoiding confusion and redundancy in their terminology. The focus in this paper, for example, will be simply on syntactic adaptation (the sequential relation of words) over longer distances (longer time spans). Another differentation in their paper concerns the theoretical approaches used for explaining these effects. There are mainly to camps: One that explains alignment as a result of conscious, cooperative decisions made during communication (Brennan and Clark, 1996), and one that views alignment as an automatic, mechanistic process occurring across various linguistic levels. This latter perspective has its theoretical foundation in Pickering and Garrod, 2004's Interactive Alignment Model (IAM). Under this view, 'alignment' is used to refer to a process in which situational cognitive models of speakers approach each other, such that they develop shared representations on different levels. The process is driven by automatic repetitions, 'priming', in which encountering an utterance will activate a representation that increases the likelihood that a person produces an utterance

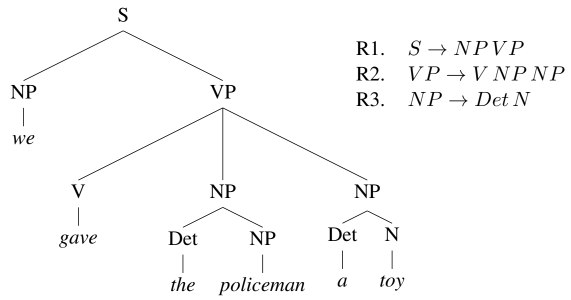

Figure 1: Example Phrase Structure Tree and Rules

that uses the same representation.

This paper focuses on alignment in LLMs, although the term is used in a more general sense, not suggesting any implicit mechanistic explanations.[1] Rather, it shall contrast adaptation that occurs in a more robust sense, in which language of different speakers generally becomes closer to one another over longer periods, from that of local short-term priming or repetitions. Nonetheless, the terms 'prime' and 'target' will be borrowed to refer to the first appearance of a linguistic structure and its subsequent repetition.

2 Methodology

To capture adaptation processes that occur over the long term we analyse effects across whole conversations. Syntactic adaptation occurs when the likelihood of syntactic structures become more likely, given that they have been used before. Reitter, 2008 introduced a method to measure such repeating effects for natural conversations, eliminating the need for controlled experimental setups.

2.1 Structural Annotations

In order to gain these syntactic structures, corpora must be structurally annotated. Annotations result in phrase structure trees (Figure 1), which can be interpreted in terms of their phrase structure rules, e.g. the subphrases that a phrase encompasses. Lexical and unary rules[2] are omitted, as they don't contribute to the syntactic structure of sentences.

2.2 Analysing Syntactic Alignment

Reitter's (2008) analysis examines whether the occurence of a syntactic rule is predictive for that same structure to have occured beforehand. Such a correlation is only plausible when there is syntactic alignment present in the data.

Conversations are split in half, removing a portion in the middle of around 20 words to eliminate short-term priming effects. From the remaining halves, the phrase structure rules are extracted, forming two sets of rules for each conversation. For simplicity, we call the first conversation half the PRIME and the second half the TARGET, analogous to repetitions in priming. For every rule in the set of TARGET we check whether that rule has been uttered by the other speaker in PRIME. The presence of a rule is encoded in a binary variable *Prime*. We want to compare whether the likelihood of a prime between speakers is larger within conversations compared to the general likelihood. To control for this, we sample from every rule in TARGET twice: Once to check for a prime in the same conversation and once to check for a prime in a random other conversation. This difference will be encoded in another binary variable *SameConversation*. Fitting a mixed effects logistic regression we check whether there is an effect of *SameConversation* on *Prime*. We would only expect a positive effect on *Prime* if there was syntactic alignment.

The analysis further includes the logarithmic frequency of rules across all conversations (*LogFrequency*) and the logarithmic size of PRIME (*LogSize*), where the size refers to the number of words uttered by the speaker sampled from. Rules that appear only once are excluded. A nested random intercept for the mixed effects logistic regression is included for speakers and conversations, as well as a random slope for *LogFrequency*.

3 Verifying the Analysis on Human Conversations

To verify the analysis, we first apply it to a dataset of human-human conversations, namely the Switchboard Corpus (Marcus et al., 1994). The corpus consists of a number of telephone conversations, out of which 650 annotated conversations were used. The analysis follows the method described in section 2.

The mixed effects logistic regression was applied using the generalized linear mixed models (GLMM) of python's pymer4 package. Outliers, rules with a frequency > 12000, were excluded. Results are shown in Table 1.

[1] The mechanistic behind alignment in LLMs, if existant, would definitely be much different from that of humans.

[2] phrases with a single child

	β	SE	OR	z	$P > \|z\|$
Intercept	-2.927	0.018	0.054	-158.847	0.000
ln(Frequency)	1.174	0.008	3.184	143.182	0.000
SameConversation	0.228	0.023	1.201	9.950	0.000
ln(Size)	1.402	0.033	3.804	41.921	0.000
ln(Frequency):SameConversation	-0.101	0.010	0.885	-9.757	0.000
ln(Frequency):ln(Size)	0.068	0.015	1.041	4.699	0.004

Table 1: Results of the GLMM on the samples drawn from the Switchboard Corpus. Fixed effects, except *SameConversation*, are centered. The model with the lowest AIC was taken ($\Delta AIC > 6$ compared to the second-best model).

3.1 Results: Switchboard

As would be expected, *ln(Frequency)* ($\beta = 1.174, p < 0.001$) and *ln(Size)* ($\beta = 1.402, p < 0.001$) largely increase the odds of a prime to be present in PRIME. *SameConversation* ($beta = 0.228, p < 0.001$) has a positive effect on *Prime*, supporting the findings of Reitter, 2008, although the reported values here are slightly lower ($\beta_{Reitter} = 1.064$, $OR_{Reitter} = 2.90$). The interaction between *ln(Frequency)* and *SameConversation* ($\beta = -0.101, p < 0.001$) indicates that alignment is stronger for less frequent syntactic structures. The low interaction of *ln(Frequency)* and *ln(Size)* ($\beta = 0.068, p < 0.004$) indicates that it is neglegible.

3.2 Preliminary Discussion

The results on the Switchboard Corpus show that the analysis captures alignment effects, verifying the results found in Reitter, 2008 on a different dataset and with a refined sampling method. The lower effect of *SameConversation* can be largely explained by this new strategy: By only sampling from the other speaker in PRIME, chances of finding a prime are expected to be reduced by more than half. Additionally, the Switchboard Corpus used here comprises normal conversations, while the Maptask Corpus used by Reitter, 2008 includes task-oriented dialoges (Anderson et al., 1991). Taken from Reitter's (2008) results, such goal driven conversations increase alignment effects.

4 Experiment 1: Long-Term Syntactic Alignment in LLMs

To analyze syntactic alignment of Large Language Models, the same analysis was run on conversations generated using OpenAI's GPT-4o. Existing datasets proved to be unsuitable for two main reasons: Datasets containing conversations between LLMs contain interactions that are too short to analyze long-term effects. Datasets of human-LLM interactions are much too diverse and far from natural conversations to be included.

4.1 Dataset Generation

For this process, we constructed a set of 'Agents'. Agents share a system prompt that instructs them on being in a conversation. Further, each agent is given a unique language prompt, varying their individual syntax, such that alignment effects can take place. Their responses are iteratively used to prompt the other agent, gradually building the context of a conversation. The process is run until a certain length threshold, a predefined number of words is surpassed.

Overall 17 different language instructions were used, resulting in 17 different agents. Those agents were cross-matched to create a total of 136 conversations with unique pairings, out of which 12 were excluded as they ended in repeating patterns. Conversations were generated using OpenAI's GPT-4o model. All conversations were prompted to have the same topic: "What makes a day a good day?". The remaining 124 conversations were used for the analysis.

4.2 Method

The same analysis was run again on the generated dataset. The sampling process was adapted to exclude sampling from identical agents of other conversations. Outliers, rules with a frequency > 2000, were excluded. The results are shown in Table 2.

4.3 Results

We find a significant effect of *SameConversation* on *Prime* ($\beta = 0.198$, $p < 0.001$). LLMs do in fact align their syntax. Again, *ln(Frequency)* and *ln(Size)* show strong positive effects, as

	β	SE	OR	z	$P > \|z\|$
Intercept	-2.031	0.048	0.131	-42.488	0.000
ln(Frequency)	1.275	0.028	3.580	45.582	0.000
SameConversation	0.198	0.056	1.219	3.538	0.000
ln(Size)	1.175	0.107	3.240	10.972	0.000
ln(Frequency):SameConversation	-0.146	0.035	0.864	-4.204	0.000
ln(Frequency):ln(Size)	0.266	0.062	1.305	4.297	0.000

Table 2: Results of the GLMM on the samples drawn from the 124 GPT-4o conversations. Fixed effects, except *SameConversation*, are centered. The model with the lowest AIC was taken ($\Delta AIC > 4$ compared to the second-best model).

is expected. Similar to human conversations, *ln(Frequency):SameConversation* has a negative effect on *Prime*, showing that alignment is stronger on less frequent rules. Deviating from it is the interaction between *ln(Frequency)* and *ln(Size)* ($\beta = 0.266, p < 0.001$).

4.4 Discussion

The difference in the effect of *ln(Frequency):ln(Size)* has an interesting implaction: Assuming there is alignment, whenever a rule occurs, it raises the likelihood of that rule to appear again. This means that increasing the sample set size not only boosts the occurrence of a rule, but also increases the probability of further occurences. This logic only applies in a setting in which occurences are actually driven by random processes, hence the results indicate the stochastic nature of LLMs, which is absent in human communication. The results show that LLMs exhibit syntactic alignment over long periods. LLMs must therefore adapt their syntax over the course of conversations aligning to the other agent or to the syntax present in their given contexts. The analysis is not designed to give any more insights of this process, beyond this simple statement. It is safe to say that the mechanism in LLMs differs from that of human, making the results highlight a shared feature between the language produced by LLMs and humans, rather than any shared mechanisms.

To further explore how LLMs align their syntax over the course of a conversation, we deploy a second experiment.

5 Experiment 2: Progression of Syntactic Alignment in LLMs

To explore the process of alignment throughout a conversation, we analyze the discrete distribution of used syntactic structures. We use the Jensen-Shannon Divergence (JSD) as distance measurement between a starting distribution and the evolving distributions over sections of the conversation.

5.1 Method

Conversations between two different agents are analyzed across different sections. For each section, the respective distribution of rules for both agents are extracted. Then, for each section, their Jensen-Shannon Divergence is calculated, resulting in a series of similarity measurements over the course of the conversation.

To estimate accurate rule distributions across sections, there needs to be sufficient data available. We ran a pre-analysis of variance in JSD values with variable amounts of data to determine a suitable section size (see Appendix). Following these results, a single conversation between two different agents was run for 520 trials, each containing an identical prompting. Out of these 520 conversations, 14 were excluded due to repeating patterns. The remaining 506 conversations were used for the analysis.

5.2 Results

Figure 2 shows that the probability distributions of agents become progressively closer across splits. Most alignment happens primarily between the first and second splits. Additionally, the divergence score for the first split is already much lower than the expected 0.37 between Agents 7 and 8 **??**. This suggests that agents align their language predominantly on just a few initial exchanges. Overall, these findings align with the results from section 4.

5.3 Discussion

The process of alignment in LLMs follows a continuous pattern - syntactic structures gradually converge towards the distribution given in their input

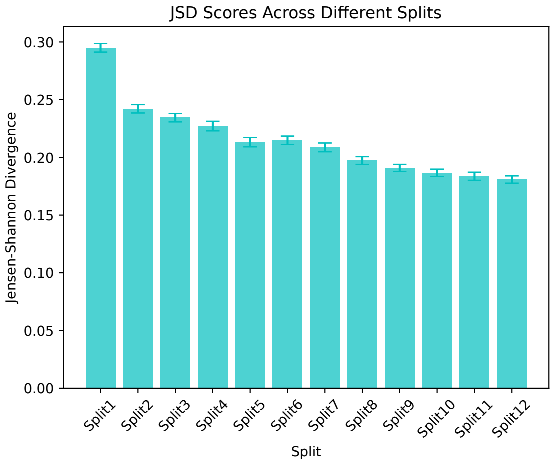

Figure 2: Jensen-Shannon Divergence scores for each split of the conversation. Values are averaged over 100 bootstraps of 506 conversations. Errorbars show their standard deviation.

context. The effects are largest at the beginning and fall off quickly. This behaviour is expected for continuous convergence of two approximating values: The further they are apart, the larger the change. It should be noted however, that the JSD naturally follows a non-linear decay (Lin, 1991). Further analysis is required to determine the exact decay rate for the alignment process.

Nonetheless, we see clear evidence that LLMs not only adapt their syntax to the context that they are given, but also that this process correlates with the amount of syntactic structures that are available in their input.

6 Conclusion

We have shown that LLMs adapt to the syntax of their input in a scenario of aritifical conversations. Unlike humans, who exhibit variability far from stochastic behavior (Kilgarriff, 2005), the findings support that LLMs operate in a much more deterministic way. While short-term and long-term alignment effects in humans are thought to arise from distinct mechanisms (Reitter and Moore, 2014, see discussion in Rasenberg et al., 2020), the results underpin the intuition that short-term effects, like those reported in Cai et al., 2024, are driven by the same principles as long-term alignment in LLMs; LLMs are not only influenced by the semantics of their input vectors, but also by the order in which these token embeddings appear.

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 G. Cai, Xufeng Duan, David A. Haslett, Shuqi Wang, and Martin J. Pickering. 2024. Do large language models resemble humans in language use? *Preprint*, arXiv:2303.08014.

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.

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.

J. Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

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. 2008. Context effects in language production: Models of syntactic priming in dialogue corpora.

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

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

A Example Appendix

This is an appendix.