Aligning to Adults Is Easy, Aligning to Children Is Hard: A Study of Linguistic Alignment in Dialogue Systems

Dorothea French, Sidney D'Mello, Katharina von der Wense
University of Colorado Boulder
Boulder, CO, 80309
Dorothea.French@colorado.edu

Abstract

During conversations, people align to one another over time, by using similar words, concepts, and syntax. This helps form a shared understanding of the conversational content and is associated with increased engagement and satisfaction. It also affects conversation outcomes: e.g., when talking to language learners, an *above* normal level of linguistic alignment of parents or language teachers is correlated with faster language acquisition. These benefits make human-like alignment an important property of dialogue systems, which has often been overlooked by the NLP community. In order to fill this gap, we ask: (RQ1) Due to the importance for engagement and satisfaction, to what degree do state-of-the-art dialogue systems align to adult users? (RQ2) With a potential application to child language acquisition in mind, do systems, similar to parents, show high levels of alignment during conversations with children? Our experiments show that ChatGPT aligns to adults at roughly human levels, while Llama2 shows elevated alignment. However, when responding to a child, both systems' alignment is below human levels.

1 Introduction

Conversation allows people to share information by creating a collective representation of the conversational context, achieved in part by linguistic alignment (Garrod and Pickering, 2004; Pickering and Ferreira, 2008): when two people are conversing, the content of their speech as well as how it is phrased prime the other person to respond in a certain way. This reduces the chance of misunderstandings, since the used words and phrasing already have an established shared meaning, and thus, makes communication more efficient and enjoyable (Garrod and Pickering, 2004).

Linguistic alignment is critical in a variety of conversations, even those between a human and a virtual agent: prioritizing alignment in responses makes conversation with chatbots more effortless and less frustrating for users (Spillner and Wenig, 2021). Nevertheless, alignment is often overlooked by the NLP community and has not yet been studied in the context of state-of-the-art dialogue systems, even though they are becoming increasingly omnipresent. To fill this gap, we first ask: **(RQ1) To what degree do two state-of-the-art dialogue systems – ChatGPT and Llama2 – align to users, and does their alignment compare to that typically seen between humans?**

	Context
MOT	Hm?
CHI	Where Mommy go?
MOT	Mommy went to the university this morning to get some books.
CHI	Where's Mommy's books?
	Response
MOT	They're in the hallway in a big bag.
GPT	Mommy will bring the books home this evening.
Llama2	Mommy left her books in the car.

Table 1: The final lines of a dialogue excerpt from the CHILDES dataset, with the parent's true response and our system-generated responses.

Linguistic alignment plays an even greater role in educational contexts, such as language learning: amongst other benefits, aligned and comprehensible input and output prime the speaker to use appropriate syntactic structures, they can receive implicit feedback with recasts immediately after an error, and they recognize what parts of their speech led to any misunderstandings and negotiate a re-phrasal (Gass et al., 1998). Additionally, parents or caregivers show an elevated level of alignment when talking to young children (Misiek et al., 2020), and their level of alignment predicts how well the child's language skills develop (Denby and Yurovsky, 2019). As dialogue systems are used more and more in language learning contexts,[1] we

[1]Examples are the language learning software Duolingo

further ask: **(RQ2) To what degree do ChatGPT and Llama2 align to children (i.e. non-fluent speakers), and how does this level of alignment compare to a parent's?**

We conduct experiments on the Switchboard Dialogue Acts Corpus (SWDA), which consists of adult–adult conversations (for RQ1) (Stolcke et al., 2000), and on the CHILDES dataset (Macwhinney, 2000), which contains child–parent conversations (for RQ2). We generate responses with ChatGPT and Llama2 and calculate three types of alignment – syntactic, lexical, and semantic – for each of their responses. Our results show that ChatGPT's alignment levels approximate those of humans when participating in standard adult conversation, but are lower than human level when responding to a child. Llama2 aligns above human levels in conversations with adults, but below human levels during dialogue with children. Overall, our results indicate room for improvement with regards to the alignment levels of dialogue systems.

2 Related Work

Exploring Linguistic Alignment Linguistic alignment is a mechanism by which humans mimic their partners in conversation – from phonology, to syntax and semantics (Garrod and Pickering, 2004). This kind of repetition lightens the cognitive load of language production, as certain structures are already primed from previous usage. Alignment at multiple levels such as lexical and syntactic results in alignment of situation models, as language production, comprehension, and interactive production are all interwoven (Pickering and Garrod, 2013). Alignment contributes to the success of a variety of human interaction. From the workplace – employees who show elevated levels of alignment over time are more likely to remain in the company (Doyle et al., 2016) – to the language classroom or nursery (Denby and Yurovsky, 2019). In some cases, the alignment of task-specific vocabularies strongly correlates with conversation outcomes (Fusaroli et al., 2012). Alignment as a feature of communication is also critical in human-computer interaction. Lexical alignment affects human understanding of a conversational agent during live conversation (Srivastava et al., 2023). It also contributes towards decreasing user frustration and perceived task load when interacting with a dialogue system (Spillner and Wenig, 2021).

and EFL classroom (Amin, 2023).

	Context
A	Any jury's not going to disregard the evidence, you know.
B	Uh, that's true.
B	I, I, I think our judicial system is attorney welfare myself.
A	That may very well be.
	Response
B	I, I hold it in the utmost contempt.
GPT	It's definitely a possibility that needs to be looked into.
Llama2	Yeah, it's like, you know, they're just trying

Table 2: The final lines of a dialogue excerpt from the SWDA corpus, with gold and generated responses

Analysis of Dialogue Systems While common to use the automatic scoring methods of word overlap with a ground truth (such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004)) or words embeddings to evaluate dialogue systems, these metrics do not correspond highly with human judgement (Liu et al., 2017). Other automatic metrics, such as context coherence – how well the response matches the context of the conversation (Xu et al., 2018) – can result in improved systems. Outside of automatic metrics, human evaluation is critical and can look at dimensions such as informativeness, grammatically, coherence as well as how human-like or engaging the system is (Finch and Choi, 2020). The downside of human evaluation, however, is time and cost.

3 Experimental Setup

3.1 Data

Switchboard Dialogue Acts Corpus Our first corpus is the Switchboard Dialogue Acts Corpus (SWDA), which consists of a series of phone conversations on a variety of topics (Stolcke et al., 2000). All dialogues are between adults, which allows us to assess model alignment with adults, i.e., fluent speakers of English (RQ1). We use all 1157 transcripts.

CHILDES We also experiment with the CHILDES dataset, which consists of conversations between caretakers and children (Macwhinney, 2000), to assess the models' alignment to children, i.e., language learners (RQ2). We use the 7721 transcripts from North American English speakers aged 24 to 42 months.

Data Preparation To prepare the data, first we extract relevant dialogue excerpts from each tran-

script: the final two lines – the target utterance and response – must come from different participants and each be at least 3 words long. We then randomly select 10,000 excerpts from each dataset. Each excerpt is 36 lines long, allowing for one target response and 35 turns of context – a length chosen to ensure ChatGPT's has enough context to work with, as described in Appendix A. For the CHILDES transcripts, the true responses averaged 7.1 words, and for SWDA they averaged 9.1.

3.2 Models and Baseline

ChatGPT and Llama2 The first state-of-the-art model we experiment with is ChatGPT, a generative pretrained transformer (Vaswani et al., 2017) from the GPT 3.5 family of language models released by OpenAI. These models are trained using reinforcement learning from human feedback (Ouyang et al., 2022). We compare ChatGPT 3.5 turbo to the 7B and 13B chat versions of Llama2, trained using publicly available sources and Reinforcement Learning with Human Feedback (Touvron et al., 2023).

Prompting In each prompt, we provide the most recent 35 utterances from the dialogue as context. We use the following prompt for RQ1 (adults): "System: *You are having a conversation with person* <A or B>. *Respond with a single line approximately* <True Length> *words long.* A: <Utterance>, B: <Utterance>, ..."

Similarly, we use the following prompt for RQ2 (children): "System: *You are a parent talking to a child. Predict the parent's next line as best you can, even with little context. Respond with a single line approximately* <True Length> *words long.* MOT: <Utterance>, CHI: <Utterance>, ..."

We request a reply with approximately the same number of words as the gold response, as both systems produced overly long responses in preliminary experiments. When using ChatGPT, we check the returned message for a set of keywords (including "AI," "language model," "context," and "clarify") that indicate the model fails to provide a response to the conversation, then regenerate up to five times if needed before moving on.

Baseline To estimate the amount of random alignment for each dataset, we shuffle the responses and randomly pair them with a dialogue context. We do this separately for the true and generated responses.

Response set	Syntactic	Lexical	Semantic
True	0.444	0.170	0.308
True Baseline	0.405	0.117	0.248
ChatGPT	0.443	0.151	0.340
ChatGPT Baseline	0.418	0.117	0.280
Llama2 13B	0.472	0.207	0.350
Llama2 13B Baseline	0.421	0.130	0.277
Llama2 7B	0.475	0.213	0.374
Llama2 7B Baseline	0.420	0.130	0.286

Table 3: Alignment scores for the SWDA corpus

3.3 Alignment Metrics

We use the align package (Duran et al., 2019) to calculate syntactic, lexical, and semantic alignment. All are computed given the last (i.e., the most recent) context utterance u and the (true or generated) response r.

Syntactic Alignment To calculate the syntactic alignment a_{syn}, the utterance and response are segmented into uni-grams, tagged with part-of-speech (POS) information, and condensed into a set of unique POS tags with the counts of their occurrences: $u = (u_1, c_{u_1}), ..., (u_n, c_{u_n})$ and $r = (r_1, c_{r_1}), ..., (r_m, c_{r_m})$, with n and m being the number of unique POS tags in u and r, and c the number of times each tag occurs in the utterance. The syntactic alignment is then computed as the cosine similarity of the context and response vectors:

$$a_{syn} = \text{cosine}(v_u, v_r) \quad (1)$$

Lexical Alignment The process for lexical alignment a_{lex} is identical that of syntactic alignment, except using word lemmas instead of POS tags.

Semantic Alignment Lastly, semantic alignment a_{sem}, which describes how the utterance content overlaps, is calculated using word2vec embeddings (Mikolov et al., 2013) $e(u_1), ..., e(u_n)$ and $e(r_1), ..., e(r_m)$. We use a bag-of-words approach to obtain sentence representations e_u and e_r. Semantic alignment is computed as:

$$a_{sem} = \text{cosine}(e_u, e_r) \quad (2)$$

4 Results and Discussion

RQ1: Alignment to Adults All results for RQ1 are shown in Table 3. Comparing the alignment of ChatGPT to the true response, we see that there is less than 1% difference in the syntactic alignment, a 10% increase in semantic alignment, and a 12% decrease in lexical alignment. Semantic alignment is the only category in which ChatGPT

Response Set	Syntactic	Lexical	Semantic
True	0.490	0.278	0.411
True Baseline	0.359	0.069	0.181
ChatGPT	0.436	0.190	0.347
ChatGPT Baseline	0.367	0.071	0.196
Llama2 13B	0.464	0.227	0.345
Llama2 13B Baseline	0.371	0.073	0.179
Llama2 7B	0.473	0.251	0.370
Llama2 7B Baseline	0.366	0.075	0.180

Table 4: Alignment scores for CHILDES dialogues

overshoots human levels, which could indicate it is less likely to introduce new topics than a human. Both Llama models also show this trend. Llama2 overshoots human alignment in all categories – as the size of the Llama2 model decreases, so does its performance as it strays further from human-like alignment levels.

Turning to the baselines, for syntactic and lexical alignment, ChatGPT is closer to the randomized baseline than humans are; which means a higher fraction of its alignment does not come from matching a specific conversation, but from using more common words and syntax. The baseline alignments between all three models are fairly similar, although the semantic space of the smaller Llama2 model is less diverse as can be seen from a higher alignment baseline.

Upon manual inspection of 100 transcripts, we see that ChatGPT generates more convincing results. On a scale of 1 (makes minimal sense) to 5 (an ideal response) ChatGPT scored an average of 4.37. The responses are also much more likely to contain novel information or drive the conversation forward. However, it less convincingly mimics the style of the conversation and the human respondent. The Llama2 models both score below 3.50. They mimic stylistic elements, but oftentimes do not contribute positively to the conversation (i.e. generate responses such as "Oh, yeah!", "Uh-huh.", or duplicate the previous utterance). This shows that past a certain point, elevated levels of alignment may negatively correlate with response quality and sophistication.

RQ2: Alignment to Children Our results for RQ2 are shown in Table 4. First, we see that the syntactic alignment of ChatGPT is 12% lower than that of a human, lexical alignment is 37% lower, and semantic alignment is 17% lower. In contrast, Llama2 13B's alignments are 5%, 22%, and 19% lower, respectively. On one hand, these decreases might be due to difficulties understanding the conversation. The dialogues jump around and do not necessarily have a clear topical thread or goal. On the other hand, there is a divide in the metrics of success for a human parent and for a dialogue system – a parent does not need to successfully complete an inquiry or interaction, but needs to engage with the child in ways that further development (John et al., 2013). When comparing the levels of alignment of ChatGPT and Llama2 across datasets, we see syntactic and semantic change less than a few percent. Lexical alignment increases with CHILDES, perhaps due to a smaller inventory of words appearing in the context. Overall, we can conclude that the systems respond with a similar level of alignment regardless of the target audience.

Moreover, human-like alignment is not the only metric necessary to grade a model's quality. Inspection of the responses shows the ChatGPT responses are most convincing, at 3.86, although they show decrease in quality from the adult conversation. The Llama2 7b model averaged only 3.02, whereas the Llama2 13b model reached 3.35 – a smaller differential with ChatGPT than the adult conversation. When looking at what fraction of the responses were considered poor, 15% of the GPT responses to adults scored a 3 or less, whereas 26% of the responses to children were 3 or less. These were 41% and 63% respectively for the Llama 7b model, and 61% and 38% for the Llama 13b model. Overall, the quality of the Llama responses were below that of ChatGPT for children, and markedly lower for adults. Yet, when choosing a dialogue system to interact with children or language learners, Llama2 (or models that mimic conversation style more heavily) might still be a good choice: closer to human-like levels of alignment could aid in developing the child's language skills. This type of user might also care less about novelty and helpfulness of the system, and more about ease of understanding and lowered cognitive load.

5 Conclusion and Future Work

Dialogue systems show great potential to assist humans across a variety of tasks. The success of these interactions, like human–human interaction, correlates with linguistic alignment. Thus, we explore how state-of-the-art dialogue systems align to both adults and children. We find that, when responding to adult speakers, ChatGPT shows approximately human-level alignments and provides constructive responses. Llama2, however, overly mimics the

conversation. This could be positive when talking with children or language learners as it results in heightened alignment. However, both models align below human levels. We conclude that SOTA dialogue systems have room for improvement in regards to reaching ideal levels of alignment under various circumstances.

In the future, we plan to investigate alignment to adult learners or non-typical speakers, in addition to exploring techniques to create dialogue systems with a closer-to-human level of alignment. We will also explore how well dialogue systems match the user in multi-turn conversational structures, and related outcomes (Fusaroli and Tylén, 2016).

Limitations

One of our primary limitations is that we are not able to use human participants to converse with the dialogue systems. While using existing datasets is an appropriate proxy to determine if this is an area which needs improvement, the chat systems may behave differently when dynamically adapting to a participant. Additionally, as we used commonly available data sets, there is a good chance they were part of the training data. Upon qualitative assessment of responses we did not find high similarity between the gold responses and generated responses for SWDA or CHILDES. Nonetheless, there is still a possibility the system has knowledge of the gold responses and used it when generating a reply – although in this case, the actual level of alignment would be lower than what we found, indicating our results are even more significant. In future works we would also like to explore using additional datasets and models. Lastly, while it does not directly affect the outcomes of this work, there is some ambiguity to the ideal level of alignment. We know that in many cases alignment correlates with positive outcomes, but it is a question for future work how much dialogue systems should be aligning to users and how variable that alignment should be across a variety of conversation types.

Ethics Statement

Our work analyzes current systems and suggests an avenue for future improvement. However, we do not intend to imply that dialogue systems should be used in all situations without additional consideration. Especially when interacting with children, we must ensure the accuracy of content and safety of communication methods. Additionally, while we point out a way in which state-of-the-art dialogue models exhibit below-human performance, the goal is not to make them more human-like as there is a lot of potential for harm when a chatbot cannot be distinguished from a person. Instead, we hope this work will help us improve dialogue systems as a tool and make them more useful in a variety of situations.

References

Momen Yaseen M. Amin. 2023. Ai and chat gpt in language teaching: Enhancing efl classroom support and transforming assessment techniques. *International Journal of Higher Education Pedagogies*, 4(4):1–15.

Joseph Denby and Daniel Yurovsky. 2019. Parents' linguistic alignment predicts children's language development. In *Annual Meeting of the Cognitive Science Society*.

Gabriel Doyle, Amir Goldberg, Sameer B Srivastava, Michael C Frank, et al. 2016. Alignment at work: Accommodation and enculturation in corporate communication. Technical report, Technical report.

Nicholas D. Duran, Alexandra Paxton, and Riccardo Fusaroli. 2019. Align: Analyzing linguistic interactions with generalizable techniques-a python library. *Psychological methods*.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols.

Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. 2012. Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological Science*, 23(8):931–939.

Riccardo Fusaroli and Kristian Tylén. 2016. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, 40(1):145–171.

Simon Garrod and Martin Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences*, 8:8–11.

Susan M. Gass, Alison MacKey, and Teresa Pica. 1998. The role of input and interaction in second language acquisition: Introduction to the special issue. *The Modern Language Journal*, 82(3):299–307.

Aesha John, Amy Halliburton, and Jeremy Humphrey. 2013. Child–mother and child–father play interaction patterns with preschoolers. *Early Child Development and Care*, 183(3-4):483–497.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2017. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation.

Brian Macwhinney. 2000. The childes project: tools for analyzing talk. *Child Language Teaching and Therapy*, 8.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Thomas Misiek, Benoit Favre, and Abdellah Fourtassi. 2020. Development of multi-level linguistic alignment in child-adult conversations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 54–58, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Martin J. Pickering and Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin*, 134(3):427–459.

Martin J. Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Laura Spillner and Nina Wenig. 2021. Talk to me on my level – linguistic alignment for chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, MobileHCI '21, New York, NY, USA. Association for Computing Machinery.

Sumit Srivastava, Mariët Theune, and Alejandro Catala. 2023. The role of lexical alignment in human understanding of explanations by conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 423–435, New York, NY, USA. Association for Computing Machinery.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *CoRR*, cs.CL/0006023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, Brussels, Belgium. Association for Computational Linguistics.

A Context Length Selection

We want to choose a context length for the transcripts that maximizes the models' ability to responses accurately, while minimizing computing costs. We choose to use the CHIDLES dataset for this selection, as the transcripts with children on average were 100 words shorter than those with adults – this rules out the possibility that the models are simply not getting enough context. We primarily select based on ChatGPT's alignment levels, as it has higher computing costs and exhibited lower levels of alignment alongside more constructive responses .

A.1 Method

We randomly selected a subset of 200 transcripts with at least 101 turns to compare the effects of context length on ChatGPT's responses. The response is held constant, back-selecting increasing lengths of context.[2]

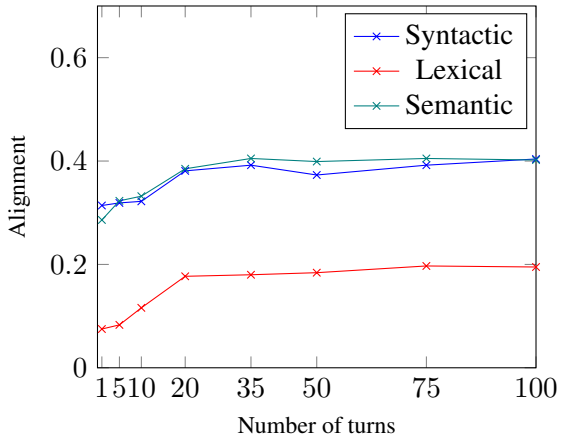

Figure 1: Alignment trends for ChatGPT's responses given varying context lengths

A.2 Decision

In these results, shown in Figure 1, we see a general trend in alignment for generated responses from different context lengths. The alignment in all three categories increases greatly up to 20 responses, and continue increasing slightly until 35 turns. We choose to use 35 turns to maximize ChatGPT's potential to provide a fully developed response while keeping computing costs manageable. While the adult transcripts generally have greater word counts, adding more context did not help ChatGPT generate better responses to the children, so we maintain keeping the number of turns constant across datasets. This selection of 35 turns does not imply an absolute requirement for length. Upon inspection, we see that in most cases the dialogue systems focus on the last few lines of context – allowing for the use of shorter transcipts if needed for other experiments.

[2]Additionally, we separately tried changing ChatGPT's temperature between 0 and 1, but only found minimal effects on alignment.