Do language models accommodate their users? A study of linguistic convergence

Terra Blevins[1,2] **Susanne Schmalwieser**[3] **Benjamin Roth**[2,3]
[1]Khoury College of Computer Sciences, Northeastern University
[2]Faculty of Computer Science, University of Vienna, Vienna, Austria
[3]Faculty of Philological and Cultural Studies, University of Vienna, Vienna, Austria
Correspondence: t.blevins@northeastern.edu

arXiv:2508.03276v2 [cs.CL] 12 Feb 2026

Abstract

While large language models (LLMs) are generally considered proficient in generating language, how similar their language usage is to that of humans remains understudied. In this paper, we test whether models exhibit *linguistic convergence*, a core pragmatic element of human language communication: do models adapt, or converge, to the linguistic patterns of their user? To answer this, we systematically compare model completions of existing dialogues to original human responses across sixteen language models, three dialogue corpora, and various stylometric features. We find that models strongly converge to the conversation's style, often significantly overfitting relative to the human baseline. While convergence patterns are often feature-specific, we observe consistent shifts in convergence across modeling settings, with instruction-tuned and larger models converging less than their pretrained and smaller counterparts. Given the differences in human and model convergence patterns, we hypothesize that the underlying mechanisms driving these behaviors are very different.

1 Introduction

Large language models have revolutionized natural language generation, with both proprietary and open-source chatbots providing grammatical and topical responses to user queries (e.g., Grattafiori et al., 2024; Gemma Team et al., 2025). These models have become so fluent (particularly in English) that readers frequently cannot determine whether a text was authored by a human or a model (Clark et al., 2021). The apparent linguistic competency of LLMs thus opens an array of questions on the properties of model-generated language, particularly concerning how model language use differs from humans'.

In this work, we investigate this through the lens of *accommodation*, the process by which people adjust their speech or writing style based on the identity of their interlocutor (Giles et al., 1991). While accommodation takes many forms, in this paper we focus on *linguistic convergence*, or how similar the target's language is to that of their interlocutor (Niederhoffer and Pennebaker, 2002). As the most prevalent use case for LLMs is via a chat agent, convergence acts as an interesting case study into the novel setting of human-model interaction. Thus, we seek to answer the following questions: Do LLM-based chatbots adapt their language use to mirror that of their users? And does this behavior mirror trends observed in human accommodation or follow different patterns?

To study convergence in LLMs, we perform a systematic analysis of model-generated responses to human-authored conversations, collected by prompting various LLMs to complete turns in existing conversations drawn from dialogue corpora. We then analyze these responses for a series of stylometric features used to study accommodation in human linguistics, in order to estimate the degree of convergence LLMs exhibit towards their context relative to randomly sampled utterances and a human control (i.e., the gold text in turns that the model replaces).

Our experiments show that models strongly adapt their output style to their context. We also find differences in convergence across model size and training schemes, with larger and instruction-tuned models converging less than smaller or solely pretrained models. Given that pretrained models often *overconverge* on these linguistic markers relative to the human baseline, this suggests that instruction-tuned model responses are often more in line with human behavior. Furthermore, these convergence patterns are feature-specific, with models adapting to their contexts to varying degrees across metrics. While this varied behavior across features is also common in human accommodation, the variance we observe in the models often differs from that of the human baseline, in-

dicating subtler differences in human and model convergence. We then conclude by discussing the implications of this work towards understanding the linguistics of LLMs.

To summarize, this paper presents the following primary contributions: (1) we develop a novel, synthetic paradigm for testing model behavior with respect to a paired human control, which we apply to studying linguistic convergence in LLMs; and (2) we present a quantitative analysis of how language models converge to the style of their context, with insights into how modeling factors affect this phenomenon and the implications it has for interpreting model-generated text. Our findings demonstrate how models do and do not adapt to their inputs across various stylistic features, providing broader insights into the relationship between human and model language use.[1]

2 Related Work

Linguistic accommodation has been widely studied in human communication (Giles et al., 1991; Niederhoffer and Pennebaker, 2002; Giles et al., 2023, inter alia). Many prior works have leveraged computational methods to characterize this phenomenon in human language across settings ranging from online interactions (Mukherjee and Liu, 2012; Berdicevskis and Erbro, 2023) to code-switching (Bawa et al., 2018) and movie dialogues (Danescu-Niculescu-Mizil and Lee, 2011). Bhatt and Rios (2021) study how users exhibit accommodation when interacting with models, while Parekh et al. (2020) examine how users accommodate model code-switching in conversation. In this paper, we apply the method presented in Ireland et al. (2011) to model-generated text while expanding the analysis to include additional stylometric features, to better understand how *models* adapt their language to their users.

Accommodation research methods have also been applied to computational models of language in prior work. Earlier work has examined the relevance of model accommodation in chatbot task effectiveness (Chaves et al., 2019; Thomas et al., 2020). Most similar to this paper is Kandra et al. (2025), which tests whether GPT-4o exhibits syntactic accommodation. However, Kandra et al. (2025) focuses on entirely machine-generated interactions between two LLM agents, while our experiments test the extent to which LLMs show accommodation while completing a pre-existing dialogue. Thus, our setting enables direct comparison between human and model responses in a given context. We also evaluate a larger suite of open-source language models across two model families and study a large set of stylometric features to fully characterize the convergence of these models.

Finally, this work also falls into the broader field of language model behavioral analysis (see Chang and Bergen (2024) for a survey of this area). In particular, our analysis of the stylistic convergence of LMs to their inputs corroborates findings on syntactic and semantic priming in language models (Sinclair et al., 2022; Jumelet et al., 2024; Gonen et al., 2025). However, we characterize this convergence more broadly for user-model interactions, rather than through carefully constructed behavioral probes.

3 Measuring Linguistic Convergence in Language Models

Linguistic convergence occurs when a speaker adapts their language to mirror that of the person they are communicating with. While it is commonly accepted that humans frequently accommodate one another in this manner (e.g. Niederhoffer and Pennebaker, 2002), it is unknown to what extent language models exhibit these patterns during human-model interaction. We test whether *convergence* occurs in machine-generated text via a two-step process. First, we elicit generations that are grounded in human-authored conversations. We then test whether the model responses exhibit convergence compared to several control settings.

Data Generation First, we elicit machine-generated texts grounded in a dialogue by prompting a model to generate a response r_t continuing a two-person dialogue with speakers s_x, s_y, conditioned on the prior turns of the conversation $r_{<t} = r_0, ..., r_{t-1}$; the model is always given access to at least $m = 5$ prior turns of dialogue. Therefore, the model replaces the utterance of speaker s_y beginning at $t = 6$ and continues to participate in the role of s_y on turns $t = \{8, 10, ...\}$ for the remainder of the dialogue.

By replacing utterances within human-authored conversations, our data generation approach approximates a human-model experimental paradigm while minimizing annotator burden. Thus, this setting enables us to extend our experimental frame-

[1]The source code and model generations can be found at https://github.com/blvns/llm-convergence.

work to many different conversations and domains, beyond what is feasible in a single user study. Our setting also aligns model generations with the original human utterances from conversations, allowing for direct comparisons that are not possible in fully synthetic analyses of LM linguistics.

Convergence Analysis We then quantify how well r_t accommodates other utterances in the dialogue according to a number of different stylometric features (Section 3.1). Unless otherwise stated, we consider how much r_t converges with respect to r_{t-1}, or the turn immediately prior uttered by the other speaker s_x. Thus, we primarily consider specifically how much the model's output mirrors the linguistic features of the most recent utterance from the user; Section 5.4 extends this to a stepwise analysis to examine the effect of earlier turns $(r_{t-2}, ...)$ on model convergence.

Baselines In addition to directly measuring model convergence, we also compare the models' behavior to two baselines: the *human* baseline, which considers how well the original utterance that r_t replaces accommodates the prior text,[2] and the *random* baseline, which estimates the baseline level of accommodation in a dataset by replacing r_t with r_{rand}, an utterance drawn from a random conversion and timestep in the same dataset.

3.1 Linguistic Indicators of Convergence

We measure the following features to characterize whether models alter their linguistic style to match that of their users, drawing on both human accommodation research and other common stylometric features for linguistic analysis (Lagutina et al., 2019):

`Utterance Length` We measure how similar model response lengths are to the text in prior turns, a feature commonly used in accommodation and stylometric work, such as in Niederhoffer and Pennebaker (2002); Lin and Walker (2017). We measure this with the symmetric metric from Ireland et al. (2011): $LSM_x = 1 - |a - b|/(a + b)$, where a and b represent the observed values for turn r_t and r_{t-1}, respectively.

`LIWC` Agreement A standard measure of linguistic accommodation is the frequency of LIWC (Linguistic Inquiry and Word Count) (Chung and Pennebaker, 2012) function word classes (e.g. Danescu-Niculescu-Mizil and Lee, 2011). Here, we consider the LIWC2007 classes considered in Ireland et al. (2011) (personal and impersonal pronouns, articles, conjunctions, prepositions, auxiliary verbs, frequently used adverbs, negations, and quantifiers), also using the LSM metric to calculate how well each LIWC category distribution in generated responses aligns with prior turns. We both report the mean `LIWC agreement` across categories and provide a fine-grained analysis of each category in Section 5.3.

`PROPN Overlap` We calculate the overlap (percentage) of proper nouns between the text generated by the model and the preceding turn. We expect that language users who converge more with their interlocutor will have greater overlap in salient semantic topics, as approximated here by proper nouns.

`Token Novelty` We evaluate the percentage of tokens novel relative to the reference utterance; a lower percentage of novel tokens indicates that the model is converging more towards the user, or exhibiting greater *lexical alignment* (Pickering and Garrod, 2004). This is measured as $|w_t \cap w_{t-1}|/|w_t|$, where $w_x = \{w \in r_x\}$; prior work apply similar metrics of token novelty (or, inversely, token overlap) to measure lexical alignment between interlocutors (Ward and Litman, 2007; Duplessis et al., 2017, i.a.) .

4 Experimental Setup

Datasets We perform our dialogue prompting experiments on three popular English datasets: `DailyDialog` (Li et al., 2017), containing conversations about daily life as written by English language learners; `NPR` (Majumder et al., 2020), a dataset of radio interview transcripts; and the `Movie` Corpus (Danescu-Niculescu-Mizil and Lee, 2011), which contains conversations scraped from movie scripts. Each dataset covers a different dialogue domain; due to their respective data sources, all represent non-spontaneous dialogues (ranging from fully fictional conversations to constrained interviews).

For each dataset, we filter the conversations to ensure they contain at least six turns of dialogue and two speakers; we merge consecutive turns from the same speaker into a single turn. We randomly downsample the larger datasets to consider at most 1,000 conversations per setting. Our experiments are performed on the (filtered) development sets of

[2] I.e., establishing the accommodation exhibited by humans in the same setting.

DailyDialog and NPR; as the Movie corpus does not provide data splits, we randomly sampled our evaluation data from the full set. Table 1 presents the dataset statistics for each corpus post-filtering.

	Dataset Statistics		
	DailyDialog	Movie	NPR
Conversations	707	1,000	1,000
Avg. Turns	9.79	8.98	17.57
Avg. Turn Length	13.44	10.87	48.43
Replaced Turns	1,918	2,280	6,568

Table 1: Dataset sizes and statistics for the dialogue corpora post-filtering. For each dataset, we calculate convergence over the model completions of *Replaced Turns* in each dataset.

Models We consider two open-source LLM families: *Gemma 3* (Gemma Team et al., 2025), with models spanning 1B, 4B, 12B, and 27B parameters; and *Llama 3* (Grattafiori et al., 2024), with models of 1B, 3B parameters from Llama 3.2 and 8B and 70B from the Llama 3.1 release. We perform inference with checkpoints provided via Huggingface (Wolf et al., 2019) and use 8-bit quantization[3] to run the largest model (i.e., Llama3 70B). For each model, we analyze the convergence expressed by both the pretrained and instruction-tuned versions.

Prompting For each dialogue turn we want the model to complete, we prompt the model to "Continue this conversation based on the given context" and provide the conversation history, including prior model generations from earlier turns in the conversation if applicable. We perform simple post-hoc filtering of the generations to remove noise, such as standardizing white space and filtering dialogue tags used within the prompt. Appendix A provides example inputs and model generations and other experimental details.

Linguistic Annotations We parse each uttrance with spaCy (Honnibal et al., 2020) to tokenize the data and obtain proper noun annotations, and we use the LIWC 2007 word classes (Chung and Pennebaker, 2012) to obtain LIWC categories.

5 Analysis Results

Here, we summarize and discuss the results of the convergence analysis for both the human baselines and model-generated responses when compared to a randomly sampled control.[4] **LLMs adapt significantly to the style of their interlocutor** across stylometric features, often matching or exceeding the level of convergence exhibited in the human baseline (Figure 2).

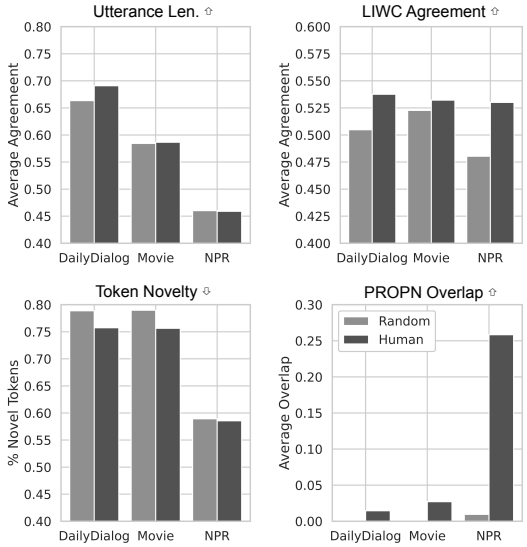


Figure 1: Comparison of the human and random baselines on each metric across datasets. Metrics marked with ↑ indicate more agreement with higher values; and ↓, vice-versa.

However, we note that linguistic convergence is often multi-faceted, and human accommodation behavior often varies significantly across features in prior work (e.g., Ireland et al., 2011; Danescu-Niculescu-Mizil and Lee, 2011). We find this is also the case for model convergence, particularly in the case of fine-grained linguistic features (Section 5.3) and when examining convergence according to earlier turns in the conversation (Section 5.4). We also observe different trends depending on the model type, with instruction-tuned models generally converging to their context less than their pretrained counterparts.

5.1 Linguistic Convergence in Human-Authored Text

To obtain a baseline for the expected level of convergence in our chosen dialogue settings, we first examine the linguistic convergence exhibited by the original speakers in these datasets. We therefore compare the level of accommodation (as measured by our four convergence metrics) exhibited in ground truth utterances with *random* utterances drawn from the dataset (Figure 1).

Unsurprisingly, we generally find that the gold utterance r_t converges with the preceding utterance r_{t-1} more than a randomly sampled utterance. We find a significant difference ($p < 0.05$ in a

[3] https://huggingface.co/docs/bitsandbytes/

[4] Appendix B presents additional visualizations and the complete set of numerical results.

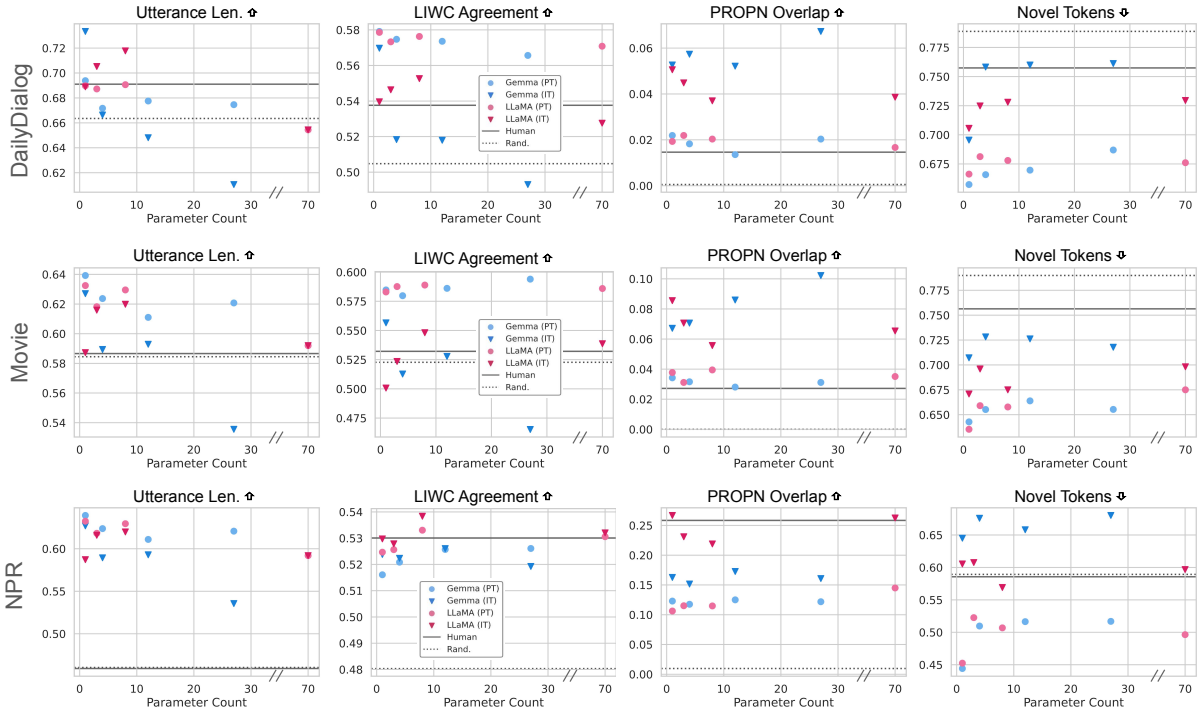


Figure 2: Scatter plot of Gemma and Llama Model scores on various convergence metrics relative to human and random baselines on DailyDialog (top row), Movie corpus (middle), and NPR (bottom), across model sizes (Billion parameters). *PT* indicates pretrained checkpoints while *IT* are instruction-tuned. Metrics marked with ↑ indicate more agreement with higher values; and ↓, vice-versa.

paired t-test) between the *human* and *random* settings on the `token novelty` and `PROPN overlap` metrics across datasets. However, convergence in `utterance length` is only significant between the two settings on the DailyDialog dataset.

We also observe differing levels of convergence on `LIWC` categories: while NPR conversations exhibit significant accommodation over the random baseline on all LIWC categories except *quantifiers*, the other datasets show much smaller differences.[5] DailyDialog conversations only show significant differences on five of nine `LIWC` word classes (despite appearing to have a large average increase in convergence over random utterances), and utterances drawn from the Movie corpus are only significantly more accommodating over the random baseline on a single class, *auxiliary verbs*.

The limited linguistic convergence we observe in the Movie corpus is likely due to the nature of the data, as the conversations were written by screenwriters, rather than drawn from naturally occurring speech. Danescu-Niculescu-Mizil and Lee (2011) similarly found that while the Movie corpus exhibited convergence under their conditions, the levels observed were much lower compared to real-world conversations (in their case, drawn from Twitter).

5.2 Linguistic Convergence in LLM-Generated Text

We now turn to examining how model-generated responses to these conversations exhibit convergence. Figure 2 compares model scores on each convergence metric against the human and random baselines on the three datasets. With these results, we consider the following questions:

Do models converge to their context? We find that models significantly outscore the random baseline in 81.25%, 100%, and 85.42% of cases for `Utterance Length`, `PROPN overlap`, and `Token Novelty`, respectively, in a paired t-test ($p < 0.05$). We also observe strong convergence on LIWC categories: models outscore the random utterances 91.67% (44 of 48) of the time on averaged scores, with significant improvements on individual LIWC classes ranging from 37.5% (for conjunction words) to 87.5% (on personal pronouns). While specific model convergence trends can differ based on several factors, this indicates that models generally adapt to the linguistic style of their context.

Furthermore, language models also significantly *over-converge* relative to human-authored utter-

[5] See Appendix Tables 8, 9, 10 for full numerical results.

ances in many cases. We find that in 62.5% and 79.2% of cases for Utterance Length and Token Novelty, the model-generated responses significantly outscore human utterances. Overconvergence occurs in fewer cases for PROPN overlap (35.4%) and LIWC classes, where they range between 10.4% and 56.25% of settings. We discuss *which* model settings overfit to their context relative to humans below.

Does model size and training affect convergence? Figure 2 shows that model training approaches affect the convergence behaviors of LLMs, with pretrained models generally adapting more to their context than their instruction-tuned counterparts. This is particularly true in the case of the Gemma model family, where the instruction-tuned models exhibit the least amount of convergence. An exception to this trend is the PROPN Overlap metric: here, pretrained models more closely mirror the human baseline while while instruction-tuned models significantly *over-accommodate* by more often repeating proper nouns from the prior utterance. This difference is likely due to alternate training objectives; for example, pretrained models likely *appear* to adapt more on Token Novelty because they are trained to fit closely to the input distribution, while instruction-tuned models are encouraged to introduce novel information during fine-tuning.

We also see minor convergence trends across model size: larger models slightly but nonsignificantly shift towards the human baseline and accommodate their context less on Utterance Length, LIWC Agreement, and PROPN Overlap (Appendix Table 4). However, convergence trends appear more stable for the Token Novelty metric.

Does model convergence differ across datasets? We find that LLMs exhibit relatively consistent convergence behavior across the three datasets, frequently significantly outscoring the random control. Given the presence of convergence across datasets and indicator metrics, linguistic convergence appears to be a general phenomenon in LLMs across various data settings and model types.

However, we also observe shifts in the models' adaptation to different datasets, particularly in the case of the *NPR* dataset. Specifically, models often exhibit *less* convergence than the baselines on NPR for Token Novelty and PROPN Overlap, but significantly more convergence in terms of Utterance Length (Figure 2). This shift likely stems from the underlying data: compared to the more casual dialogue of our other settings, NPR conversations are interview transcripts that have more structured turns and expected variance in utterance lengths, particularly between the interviewer and interviewee. Thus, while models adapt their style to their contexts, whether this adaptation is *human-like* depends on the setting.

5.3 Fine-grained Analysis of LIWC Categories

While we examine several axes to quantify linguistic convergence in language models, the majority of prior computational work on human accommodation quantifies convergence by measuring the frequency of common types of function words based on the subset of LIWC word classes identified by Ireland et al. (2011). In this section, we consider model convergence on these nine LIWC subcategories (rather than the averaged result across classes reported in prior sections) to understand finer-grained aspects of when models do and don't converge to their context.

Figure 3 summarizes the results of our fine-grained LIWC analysis on DailyDialog, reporting the relative delta of model scores on these classes compared to the random and human baselines; Appendix Figures 5 and 6 present the results for the *Movie* and *NPR* datasets. We see similar general trends to previous sections (e.g., instruction-tuned models converge less to their context than

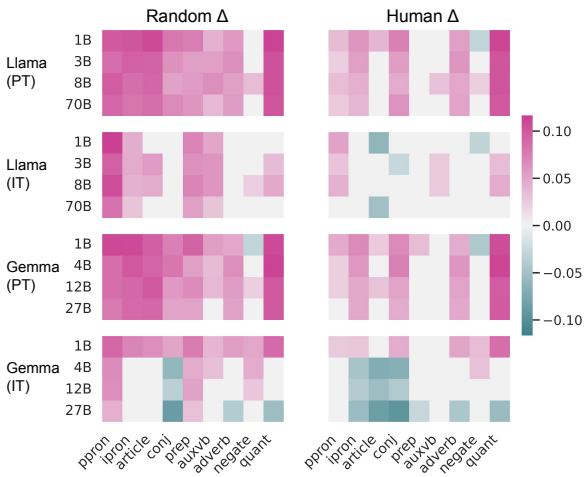


Figure 3: Summary of model convergence relative to the human and random baselines for individual LIWC word classes on the *DailyDialog* dataset. Pink cells indicate classes where the model significantly ($p < 0.05$) overconverges relative to the baseline, while green cells indicate significant undercongergence. Gray cells are not significantly different from the baseline.

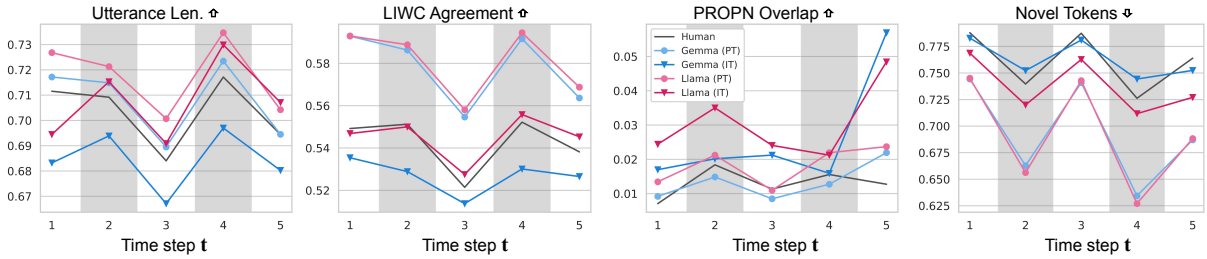


Figure 4: Stepwise analysis of convergence in LM generations (and human ground truth utterances) for DailyDialog, measuring the agreement between each utterance $r_{t=n}$ and the preceding utterances $r_{t=1,\ldots,n-1}$ on our four metrics. Timesteps in gray ($t = 2, 4$) indicate the prior turns in the role the model adopts, S_y, while white timesteps are utterances from the other speaker, S_x. Each line reports the averaged score across all model sizes in a given family.

pretrained ones), but model behavior often varies markedly across individual LIWC categories, particularly in comparison to human scores. Specifically, while pretrained model convergence is usually stronger than the random baseline, their convergence relative to the human baseline is mixed. Compared to human utterances, the pre-trained models overconverge on certain word classes, such as quantifiers, conjunctions, and adverbs, while performing on par with humans on others.

A potential explanation is that pretrained models may simply generate these word classes less overall. Examining the use of quantifiers by the largest pretrained Llama3 model, we find it underuses this word class relative to the human texts at similar rates, regardless of whether the prior utterance includes at least one quantifier or not (28.6% vs. 24.1%). Table 2 (Case A) shows a pair of model- and human-generated utterances exemplifying this trend for quantifiers. Given that only 27% of the reference utterances include a quantifier, it is likely the "overconvergence" we see is at least partially due to overall lower use of uncommon word types by the pretrained models.

Furthermore, instruction-tuned models continue to converge less than the pretrained ones on individual LIWC classes, underconverging relative to human utterances in 16 cases. Here we see a strong difference across model families, with the instruction-tuned Llama models performing similarly to humans on most word classes and their Gemma counterparts underconverging on a number of categories, such as impersonal pronouns, articles, and conjunction usage. There is also a minor shift with model size, as larger Gemma models converge less across these classes.

In contrast with pretrained models, this underconvergence seems to stem from an *overuse* of these classes, which causes generations from the instruction-tuned models to align less with their context. We examine the use of conjunctions by the pretrained and instruction-tuned largest Gemma models, and find that instruction-tuned models frequently overgenerate conjunctions relative to the human and pretrained model, with instruction-tuned models using additional conjunctions in 47.0% and 59.1% of utterances, respectively (for reference, the human reference only uses this word class in 39% of utterances). Table 2 (Case B) provides an example of this overuse.

Case A (Quant.)	
Prior Utterance	"I will take ten."
Human	"Do you want to add **some** baby's breath for that?"
Llama-70B (PT)	"Do you want them delivered?"
Case B (Conj.)	
Prior Utterance	"Yes, our over-the-counter medicine is over there on that shelf."
Gemma-27B (PT)	"Thank you."
Gemma-27B (IT)	"Great. **And** how much is a bottle of 200?"

Table 2: Qualitative examples of model behavior on fine-grained stylometric features, drawn from DailyDialog experiments. **Case A**: Pretrained models overconverge on *quantifiers*; **Case B**: instruction-tuned models overuse (and thus underconverge on) *conjunctions*.

5.4 Stepwise Analysis of Linguistic Convergence

In the previous analyses, we examine how well the model-generated responses r_t converge linguistically to the immediately prior utterance in the conversation r_{t-1}. However, our experimental setup primes the model with the first five turns in a dialogue before prompting it to participate. Thus, there is much more context for the model to potentially fit to than r_{t-1}, as it is also conditioned on other prior turns from S_x, the speaker with whom the model is asked to converse (and the standard

target of linguistic convergence analysis), and on turns from S_y, the speaker that the model replaces.

In this section, we test how linguistic features from earlier turns in the conversation influence initial model-generated utterances in conversation. Specifically, we take the first utterance generated by the model in each conversation (at $t = 6$) and compare it to all prior turns $r_{t=1,\ldots,5}$ (Figure 4 for DailyDialog; Appendix Figures 7 and 8 for Movie and NPR datasets). For clarity, we report average convergence scores across model sizes on each model family and training scheme set.

We find that human convergence scores fluctuate across timesteps; this is unsurprising, as alternating timesteps are uttered by the same speaker S_y as the considered utterance r_t, indicating that each speaker's linguistic patterns agree more with themselves than with their interlocutors. Interestingly, we see similar patterns across time with the language models for most metrics. This finding suggests that not only do language models adapt to their context, they also differentiate this adaptation to the particular speaker role (Sacks et al., 1974) they are replacing within the dialogue.[6] While these models demonstrate sensitivity to speaker roles when emulating existing speakers, it remains an open question whether they maintain distinct speaker role patterns in open-ended dialogues without a specific conversational partner to emulate.

However, we do observe some differences in model behavior compared to humans across time, particularly on `PROPN Overlap`. Model behavior follows the human trend less closely for this metric at earlier timesteps, with a sharp increase in overlap scores relative to humans on the last turn before the model generation (Figure 4). This increased score corresponds to the overconvergence observed in models in the prior sections on `PROPN`. Thus, it is likely that in the case of exact word overlap (e.g., names and other proper nouns), model convergence demonstrates a strong *recency bias* towards newer concepts (Liu et al., 2024).[7]

6 Discussion

Throughout our analysis, we compare model convergence to that of the human baselines to contextualize the models' behavior. However, while in some cases the model converges to its context similarly to human utterances, we emphasize that the observed similarities do not necessarily indicate that the underlying causes of these behaviors are the same. Human accommodation is driven by speakers (often unconsciously) altering their speech to foster social and communicative goals (Giles et al., 1991). In contrast, language models do not have these same underlying psychological communication goals when generating text.

We hypothesize that model convergence is instead driven by their pretraining objective, which encourages the model to produce test stylistically consistent with their input by training them on complete, often single-author documents. This consistency effect extends beyond style convergence: model generations have also been shown to be influenced by their input in structural priming (Sinclair et al., 2022; Jumelet et al., 2024) and through superfluous semantic correlations (Gonen et al., 2025). Recently, Kandra et al. (2025) demonstrated this in the syntactic convergence of model-model interactions. The stylistic convergence and overconvergence that we observe are thus likely another facet of this behavior.

Instead, an important consideration is how the user will interpret texts from models that *appear* to accommodate them as a human interlocutor would, such as in the case of the "human-like" behavior we observe in instruction-tuned models. Appropriate model style has been shown to facilitate successful chatbot interactions (Chaves et al., 2019; Thomas et al., 2020), and Bhatt and Rios (2021) finds that users tend to accommodate models more when they successfully produce topically relevant outputs, treating them more like human conversation partners. Other work has observed that while sycophantic behavior in langauge models can lead to lower model trust if detected by the user (Carro, 2024), users who don't identify sycophancy in model generations will often overrely on these unhelpful model responses (Bo et al., 2025). Thus, model convergence that better aligns with human behavior will likely mislead users to trust these models more, even if a model is unreliable for a given task in practice.

7 Conclusion

This paper presents a comprehensive description of *linguistic convergence* in a series of open-source generative language models. Specifically, we char-

[6] Interlocutors generally take different roles (speaker, listener) to coordinate turn-taking in conversation.

[7] We similarly see small upticks on convergence with the first utterance for multiple datasets and metrics, suggesting that *primacy biases* may affect convergence as well.

acterize the extent to which these language models adapt their outputs to the style of their context across various stylometric features and dialogue domains. While our experiments reveal varied trends across datasets and model training regimes, we generally find that LM generations *do* exhibit convergence, almost always significantly outscoring a random control on the considered metrics. Furthermore, in many cases, language models also overconverge relative to the human baseline.

Thus, we consider the extent to which model convergence is even related to human accommodation. Our finer-grained analysis of LIWC categories (a prototypical feature for studying accommodation) shows that models exhibit very different patterns on these features than humans, suggesting that the underlying mechanisms for these behaviors are likely very different. Therefore, while this work primarily considers human behavior as a reference for characterizing model generations, future work should further characterize these differences and examine the underlying causes of the observed convergence in LMs.

Limitations

We approximate LLM participation in user-driven dialogues by having them complete turns in existing dialogue datasets. While this approach has experimental advantages (e.g., we can directly compare human accommodation features with the model's responses in the same context), it also presents some limitations. Specifically, the model's responses may be biased because it participates only in the later portion of (fixed) conversations; these conversational structures are also likely to differ significantly from those in natural interactions between a user and an LLM. Future work should supplement our experimental setting with more focused user studies to validate whether our findings hold in these cases. We also note that, though we perfrom our experiments on three popular, high-quality English dialogue datasets, each dataset's source data imposes constraints on natural conversation structures, potentially limiting generalization of our findings.

While we perform experiments on sixteen language models, it remains an open question how larger models (>70B parameters) and models post-trained on other objectives adapt to their users. The considered datasets cover various styles of conversations, but these differ from how some users interact with the models (i.e., information-seeking). Thus, future work should confirm whether these findings hold up in interactive user studies. Finally, while we test for an array of stylometric features, the model's behavior of other aspects of style may differ.

Acknowledgments

This research was funded by the Vienna Science and Technology Fund (WWTF)[10.47379/VRG19008] "Knowledge infused Deep Learning for Natural Language Processing". We thank Agnieszka Gwizdek and Antonio Innocenti for early investigation in this space and Lukas Thoma for feedback on this work.

References

Anshul Bawa, Monojit Choudhury, and Kalika Bali. 2018. Accommodation of conversational code-choice. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 82–91, Melbourne, Australia. Association for Computational Linguistics.

Aleksandrs Berdicevskis and Viktor Erbro. 2023. You say tomato, I say the same: A large-scale study of linguistic accommodation in online communities. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 415–424, Tórshavn, Faroe Islands. University of Tartu Library.

Paras Bhatt and Anthony Rios. 2021. Detecting bot-generated text by characterizing linguistic accommodation in human-bot interactions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3235–3247.

Jessica Y Bo, Majeed Kazemitabaar, Mengqing Deng, Michael Inzlicht, and Ashton Anderson. 2025. Invisible saboteurs: Sycophantic llms mislead novices in problem-solving tasks. *arXiv preprint arXiv:2510.03667*.

María Victoria Carro. 2024. Flattering to deceive: The impact of sycophantic behavior on user trust in large language model. *arXiv preprint arXiv:2412.02802*.

Tyler A Chang and Benjamin K Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350.

Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 102–109.

Cindy K Chung and James W Pennebaker. 2012. Linguistic inquiry and word count (liwc): pronounced "luke,"... and other useful facts. In *Applied natural*

language processing: Identification, investigation and resolution, pages 206–229. IGI Global Scientific Publishing.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. 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)*, pages 7282–7296.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017. Automatic measures to characterise verbal alignment in human-agent interaction. In *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 71–81.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, 1:1–68.

Howard Giles, America L. Edwards, and Joseph B. Walther. 2023. Communication accommodation theory: Past accomplishments, current trends, and future prospects. *Language Sciences*, 99:101571.

Hila Gonen, Terra Blevins, Alisa Liu, Luke Zettlemoyer, and Noah A Smith. 2025. Does liking yellow imply driving a school bus? semantic leakage in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 785–798.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Molly E Ireland, Richard B Slatcher, Paul W Eastwick, Lauren E Scissors, Eli J Finkel, and James W Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological science*, 22(1):39–44.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14727–14742.

Florian Kandra, Vera Demberg, and Alexander Koller. 2025. Llms syntactically adapt their language use to their conversational partner. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 873–886, Vienna, Austria. Association for Computational Linguistics.

Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and PG Demidov. 2019. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195. IEEE.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Grace Lin and Marilyn Walker. 2017. Stylistic variation in television dialogue for natural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 85–93.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online. Association for Computational Linguistics.

Arjun Mukherjee and Bing Liu. 2012. Analysis of linguistic style accommodation in online debates. In *Proceedings of COLING 2012*, pages 1831–1846.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of language and social psychology*, 21(4):337–360.

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577.

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

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in information seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1171–1180.

Arthur Ward and Diane J Litman. 2007. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Further Experimental Details

In this section, we provide additional details about our experimental setup. To obtain model completions for the target conversations, we generate responses by prompting the model to continue the conversation based on the given context (Table 3). We generate from the models using a temperature of 0.4 and top-p sampling at 0.8. We also limit the generated output to 40 tokens and perform simple cleaning heuristics to standardize the generation before performing our convergence analysis. These parameters were chosen after a small parameter search over a small subset of the Movie corpus conversations.

Computational Hardware and Budget We run our experiments on 2 H100 GPUs, using 8-bit quantization for the LLaMA3 70B parameter model; in total, across all models (16) and datasets (3), we perform 48 generation runs of up to 1000 conversations (see Table 1 for full dataset composition details). Post-hoc analysis of the data is not computationally expensive and performed on CPUs, using the spaCy "en_core_web_trf" model (Honnibal et al., 2020) for tokenization and proper noun identification.

Prompt	Continue this conversation based on the given context.
Context	<user> What can I do for you, sir? </user> <assistant> We'd like to order breakfast for tomorrow morning. </assistant> ... <user> OK, and when shall I bring it here? </user> \n <model>
Responses	
Human	About seven thirty. By the way...
Generation*	At 7:30 AM.

Table 3: Prompt, context, and sample generations for our prompting setup to obtain model responses for our convergence analysis. *Example text generated by the Llama3 (3B) pretrained model, example conversation drawn from the DailyDialog development set.

Artifact Licensing and Use The DailyDialog dataset (Li et al., 2017) is released under a CreativeCommons Attribution and Non-comerical license (CC BY-NC-SA 4.0), while the Cornell Movie corpus (Danescu-Niculescu-Mizil and Lee, 2011) and NPR Interview dataset (Majumder et al., 2020) are released in conjunction with academic papers but do not report the license for these datasets in their papers or associated artifacts. All datasetd are released through academic publications with the intention of contiued use in NLP and dialogue research.

The spaCy package (Honnibal et al., 2020) is released under the MIT License (allowing both academic and commercial use), while the LIWC package (Chung and Pennebaker, 2012) is under a custom end user liscense agreement for academic use.[8] The Llama3 models (Grattafiori et al., 2024) are released by Meta for open use under a custom liscense[9], and the Gemma model family is released under the Gemma Terms of Use.[10]

B Additional Analysis Results

Here, we present additional analysis figures and the full numerical results of our convergence experiments. First, we perform a correlation study to examine the effect of model size and style convergence (Table 4).

Figure 5 and Figure 6 show the relative delta against the human and random baselines for individual LIWC categories for the Movie corpus and NPR dataset, respectively; this complements the DailyDialog results presented in Section 5.3. We also present the full numerical results for the indi-

[8] https://www.liwc.app/help/eula

[9] https://www.llama.com/llama3/license/

[10] https://ai.google.dev/gemma/terms

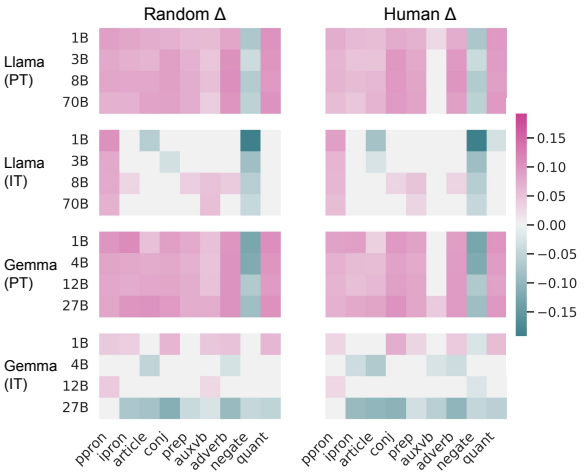


Figure 5: Summary of model convergence relative to human and random baselines on LIWC word classes for *DailyDialog*.

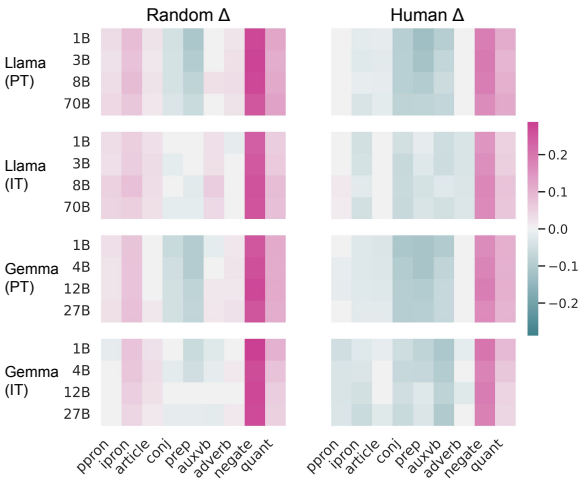


Figure 6: Summary of model convergence relative to human and random baselines on LIWC word classes for *DailyDialog*.

Dataset	PT/IT	ρ	p
Utterance Length			
DailyDialog	PT	-0.189	0.653
DailyDialog	IT	-0.221	0.380
Movie	PT	-0.252	0.546
Movie	IT	-0.253	0.546
NPR	PT	-0.263	0.530
NPR	IT	-0.263	0.529
LIWC Agreement			
DailyDialog	PT	-0.225	0.592
DailyDialog	IT	-0.233	0.579
Movie	PT	-0.212	0.614
Movie	IT	-0.256	0.541
NPR	PT	-0.217	0.606
NPR	IT	-0.218	0.604
PROPN Overlap			
DailyDialog	PT	-0.355	0.389
DailyDialog	IT	-0.246	0.558
Movie	PT	-0.246	0.557
Movie	IT	-0.136	0.749
NPR	PT	-0.563	0.146
NPR	IT	-0.282	0.499
Token Novelty			
DailyDialog	PT	-0.209	0.620
DailyDialog	IT	-0.214	0.612
Movie	PT	-0.239	0.569
Movie	IT	-0.230	0.584
NPR	PT	-0.199	0.636
NPR	IT	-0.217	0.606

Table 4: Correlation across model size and convergence values for different features and experimental settings.

vidual LIWC classes for DailyDialog (Table 8), the Movie corpus (Table 9), and NPR (Table 10). We also report the summary figures for the stepwise experiments described in Section 5.4 with Figure 7 for the Movie corpus and Figure 8 for NPR.

Finally, we present the full dataset-level experimental results across our convergence metrics in Tables 5, 6, and 7 for the DailyDialog, Movie, and NPR datasets, respectively. We also provide *Human* and *Random* baselines in each table for comparison.

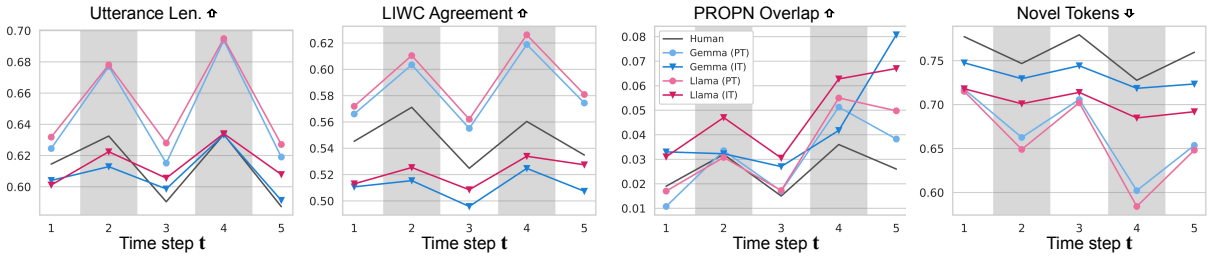


Figure 7: Stepwise analysis of convergence in LM generations (and human ground truth utterances) for the Movie corpus, measuring the agreement between each utterance $r_{t=n}$ and the preceding utterances $r_{t=1,...n-1}$ on our four metrics. Timesteps in gray ($t = 2, 4$) indicate the prior turns in the role the model adopts, S_y, while white timesteps are utterances from the other speaker S_x.

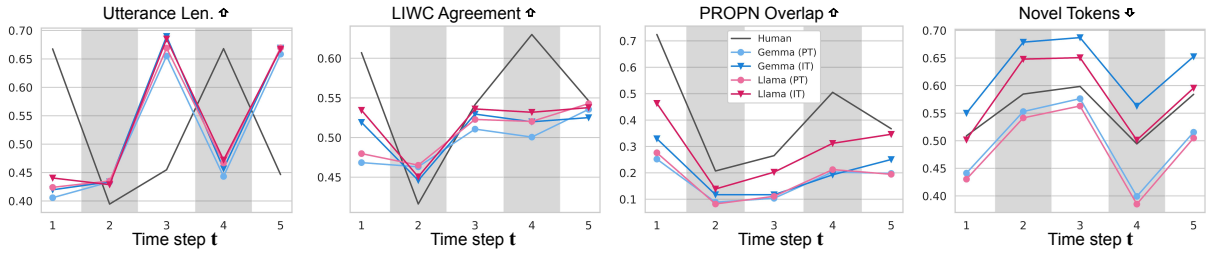


Figure 8: Stepwise analysis of convergence in LM generations (and human ground truth utterances) for NPR, measuring the agreement between each utterance $r_{t=n}$ and the preceding utterances $r_{t=1,...n-1}$ on our four metrics. Timesteps in gray ($t = 2, 4$) indicate the prior turns in the role the model adopts, S_y, while white timesteps are utterances from the other speaker S_x.

Model	Accommodation Features			
	Utterance Len. (↑)	LIWC (↑)	PROPN (↑)	Token Novelty (↓)
Human	0.69102	0.53759	0.01460	0.75742
Random	0.66348	0.50481	0.00052	0.78874
Llama3-1B (PT)	0.68975	0.57859	0.01929	**0.66624**†
Llama3-3B (PT)	0.68722	0.57327	0.02190	**0.68130**†
Llama3-8B (PT)	0.69062	0.57630	0.02033	**0.67796**†
Llama3-70B (PT)	**0.65438**†	0.57083	0.01668	**0.67603**†
Llama3-1B (IT)	0.68893	0.53960	**0.05057**†	**0.70558**†
Llama3-3B (IT)	**0.70523**	0.54633	**0.04484**†	**0.72480**†
Llama3-8B (IT)	**0.71776**†	0.55260	**0.03702**†	**0.72800**†
Llama3-70B (IT)	**0.65438**†	0.52757	**0.03858**†	**0.72950**†
Gemma3-1B (PT)	0.69385	0.57911	0.02190	**0.65732**†
Gemma3-4B (PT)	**0.67167**	0.57470	0.01825	**0.66582**†
Gemma3-12B (PT)	**0.67753**	0.57359	0.01356	**0.66968**†
Gemma3-27B (PT)	**0.67461**	0.56564	0.02033	**0.68695**†
Gemma3-1B (IT)	**0.73334**†	0.56969	**0.05266**†	**0.69551**†
Gemma3-4B (IT)	**0.66628**†	0.51836	**0.05735**†	0.75824
Gemma3-12B (IT)	**0.64808**†	0.51796	**0.05214**†	0.75992
Gemma3-27B (IT)	**0.61049**†	0.49306	**0.06726**†	0.76123

Table 5: Metric scores of common indicators of linguistic convergence in model-generated responses to conversations in the *DailyDialog* dataset. Scores for Utterance Length, PPROPN, and Token Novelty in **bold** are significantly different from human metrics ($p < 0.05$ on a paired t-test); scores indicated with † are $p < 0.001$ relative to human scores.

Model	Accommodation Features			
	Utterance Len. (↑)	LIWC (↑)	PROPN (↑)	Token Novelty (↓)
Human	0.58664	0.53215	0.02719	0.75639
Random	0.58447	0.52271	0.00000	0.78985
Llama3-1B (PT)	**0.63248**†	0.58302	0.03772	**0.63531**†
Llama3-3B (PT)	**0.61824**†	0.58764	0.03114	**0.65910**†
Llama3-8B (PT)	**0.62951**†	0.58892	**0.03947**	**0.65774**†
Llama3-70B (PT)	0.59195	0.58594	0.03509	**0.67501**†
Llama3-1B (IT)	0.58722	0.50072	**0.08553**†	**0.67083**†
Llama3-3B (IT)	**0.61592**†	0.52352	**0.07061**†	**0.69595**†
Llama3-8B (IT)	**0.61986**†	0.54813	**0.05570**†	**0.67503**†
Llama3-70B (IT)	0.59195	0.53864	**0.06535**†	**0.69817**†
Gemma3-1B (PT)	**0.63923**†	0.58475	0.03421	**0.64280**†
Gemma3-4B (PT)	**0.62373**†	0.57981	0.03158	**0.65521**†
Gemma3-12B (PT)	**0.61104**†	0.58613	0.02807	**0.66392**†
Gemma3-27B (PT)	**0.62077**†	0.59398	0.03114	**0.65530**†
Gemma3-1B (IT)	**0.62703**†	0.55655	**0.06711**†	**0.70714**†
Gemma3-4B (IT)	0.58937	0.51274	**0.07061**†	**0.72819**†
Gemma3-12B (IT)	0.59290	0.52782	**0.08597**†	**0.72619**†
Gemma3-27B (IT)	**0.53554**†	0.46521	**0.10219**†	**0.71779**†

Table 6: Metric scores of common indicators of linguistic convergence in model-generated responses to conversations in the *Movies* dataset. Scores for `Utterance Length`, `PPROPN`, and `Token Novelty` in **bold** are significantly different from human metrics ($p < 0.05$ on a paired t-test); scores indicated with † are $p < 0.001$ relative to human scores.

Model	Accommodation Features			
	Utterance Len. (↑)	LIWC (↑)	PROPN (↑)	Token Novelty (↓)
Human	0.45905	0.53007	0.25837	0.58561
Random	0.46025	0.48035	0.00959	0.58921
Llama3-1B (PT)	**0.63248**†	0.52466	**0.10627**†	**0.45246**†
Llama3-3B (PT)	**0.61824**†	0.52563	**0.11510**†	**0.52262**†
Llama3-8B (PT)	**0.62951**†	0.53303	**0.11480**†	**0.50684**†
Llama3-70B (PT)	**0.59195**†	0.53057	**0.14495**†	**0.49636**†
Llama3-1B (IT)	**0.58722**†	0.52972	0.26675	**0.60536**†
Llama3-3B (IT)	**0.61592**†	0.52786	**0.23097**	**0.60754**†
Llama3-8B (IT)	**0.61986**†	0.53835	**0.21894**†	**0.56910**†
Llama3-70B (IT)	**0.59195**†	0.53203	0.26249	**0.59674**†
Gemma3-1B (PT)	**0.63923**†	0.51606	**0.12287**†	**0.44413**†
Gemma3-4B (PT)	**0.62373**†	0.52087	**0.11754**†	**0.50963**†
Gemma3-12B (PT)	**0.61104**†	0.52570	**0.12500**†	**0.51644**†
Gemma3-27B (PT)	**0.62077**†	0.52608	**0.12180**†	**0.51689**†
Gemma3-1B (IT)	**0.62703**†	0.52380	**0.16261**†	**0.64512**†
Gemma3-4B (IT)	**0.58937**†	0.52234	**0.15164**†	**0.67585**†
Gemma3-12B (IT)	**0.59290**†	0.52603	**0.17266**†	**0.65817**†
Gemma3-27B (IT)	**0.53554**†	0.51925	**0.16078**†	**0.68015**†

Table 7: Metric scores of common indicators of linguistic convergence in model-generated responses to conversations in the *NPR* Interview dataset. Scores for `Utterance Length`, `PPROPN`, and `Token Novelty` in **bold** are significantly different from human metrics ($p < 0.05$ on a paired t-test); scores indicated with † are $p < 0.001$ relative to human scores.

Model	LIWC Classes ↑								
	Per. PRON	Imp. PRON	Article	CONJ	PREP	AUX Verb	Adverb	Negation	Quanitifer
Human	0.55537	0.47056	0.50837	0.499783	0.47303	0.45335	0.46373	0.82350	0.59065
Random	0.49129	0.42833	0.43827	0.49319	0.41357	0.42089	0.45927	0.81013	0.58833
Llama3-1B (PT)	0.59106^{+}	0.53085^{+}	0.54780^{+}	0.57325^{+}	0.48835	0.46233	0.51702^{+}	0.79396^{-}	0.70274^{+}
Llama3-3B (PT)	0.57755^{+}	0.52371^{+}	0.53142	0.55528^{+}	0.46733	0.47396	0.52381^{+}	0.81289	0.69351^{+}
Llama3-8B (PT)	0.58959^{+}	0.51368^{+}	0.52987	0.54597^{+}	0.47035	0.48503^{+}	0.51281^{+}	0.84558^{+}	0.69377^{+}
Llama3-70B (PT)	0.58169^{+}	0.50944^{+}	0.52370	0.56084^{+}	0.47408	0.45810	0.51558^{+}	0.82351	0.69056^{+}
Llama3-1B (IT)	0.60780^{+}	0.47280	0.44638^{-}	0.48232	0.48471	0.47022	0.48627	0.79248^{-}	0.61341
Llama3-3B (IT)	0.58567^{+}	0.47435	0.49488	0.47292^{-}	0.47582	0.48030^{+}	0.47990	0.82840	0.62471^{+}
Llama3-8B (IT)	0.59796^{+}	0.46956	0.48389	0.50265	0.48433	0.48032^{+}	0.48493	0.83320	0.63653^{+}
Llama3-70B (IT)	0.57476	0.45832	0.45687^{-}	0.47638	0.46646	0.45063	0.44361	0.81702	0.60406
Gemma3-1B (PT)	0.60207^{+}	0.53798^{+}	0.53470^{+}	0.56730^{+}	0.50729^{+}	0.47569	0.50796^{+}	0.78080^{-}	0.69818^{+}
Gemma3-4B (PT)	0.57737^{+}	0.52909^{+}	0.53040	0.57401^{+}	0.46686	0.45656	0.52408^{+}	0.81130	0.70267^{+}
Gemma3-12B (PT)	0.57765^{+}	0.51877^{+}	0.53791^{+}	0.55138^{+}	0.48060	0.45863	0.51551^{+}	0.83322	0.68864^{+}
Gemma3-27B (PT)	0.56882	0.51818^{+}	0.52879	0.54490^{+}	0.46479	0.44267	0.51152^{+}	0.82200	0.68911^{+}
Gemma3-1B (IT)	0.58208^{+}	0.49887^{+}	0.50235	0.54368^{+}	0.49011	0.46066	0.51424^{+}	0.85839^{+}	0.67683^{+}
Gemma3-4B (IT)	0.55696	0.42369^{-}	0.43792^{-}	0.43233^{-}	0.45948	0.45446	0.45307	0.85591^{+}	0.59141
Gemma3-12B (IT)	0.55628	0.43151^{-}	0.45384^{-}	0.45801	0.46614^{-}	0.43746	0.44574	0.83848	0.57419
Gemma3-27B (IT)	0.53463	0.41544	0.42213^{-}	0.40541^{-}	0.44547^{-}	0.44339	0.41977^{-}	0.81694	0.53432^{-}

Table 8: Per-class scores of LIWC categories on the *DailyDialog* dataset. Per-class scores significantly (p<0.05) over- or under-accommodating relative to the human baseline are annotated with $^{+}/^{-}$, respectively.

Model	LIWC Classes ↑								
	Per. PRON	Imp. PRON	Article	CONJ	PREP	AUX Verb	Adverb	Negation	Quanitifer
Human	0.48662	0.45882	0.51208	0.53520	0.42770	0.44849	0.49068	0.73628	0.69347
Random	0.47229	0.43810	0.49400	0.54280	0.43686	0.41626	0.48199	0.73486	0.68726
Llama3-1B (PT)	0.56046^{+}	0.51947^{+}	0.56849^{+}	0.61117^{+}	0.49737^{+}	0.47484^{+}	0.56384^{+}	0.66310^{-}	0.78845^{+}
Llama3-3B (PT)	0.55145^{+}	0.50901^{+}	0.56037^{+}	0.63205^{+}	0.50568^{+}	0.46281	0.58514^{+}	0.69707^{-}	0.78515^{+}
Llama3-8B (PT)	0.55398^{+}	0.51839^{+}	0.57427^{+}	0.63041^{+}	0.51024^{+}	0.47168	0.58988^{+}	0.66916^{-}	0.78227^{+}
Llama3-70B (PT)	0.54275^{+}	0.50662^{+}	0.57818^{+}	0.62881^{+}	0.51181^{+}	0.45624	0.58017^{+}	0.68024^{-}	0.78867^{+}
Llama3-1B (IT)	0.57487^{+}	0.44679^{-}	0.43215	0.51564	0.43272	0.43078	0.46555	0.54436^{-}	0.66362^{-}
Llama3-3B (IT)	0.55139^{+}	0.44934	0.48609^{-}	0.51258	0.44856	0.44119	0.48063	0.64849^{-}	0.69343
Llama3-8B (IT)	0.54993^{+}	0.46742	0.49897	0.56659^{+}	0.47575^{+}	0.46808	0.52338^{+}	0.67381^{-}	0.70928
Llama3-70B (IT)	0.54230^{+}	0.45001	0.50772	0.54154	0.45813^{+}	0.47079	0.50060	0.69068^{-}	0.68594
Gemma3-1B (PT)	0.57242^{+}	0.54646^{+}	0.54701^{+}	0.62979^{+}	0.51441^{+}	0.46974	0.58130^{+}	0.60856^{-}	0.79308^{+}
Gemma3-4B (PT)	0.55490^{+}	0.51745^{+}	0.56900^{+}	0.62117^{+}	0.50467^{+}	0.46693	0.58258^{+}	0.61537^{-}	0.78623^{+}
Gemma3-12B (PT)	0.55298^{+}	0.51403^{+}	0.57491^{+}	0.62396^{+}	0.50916^{+}	0.46872	0.58389^{+}	0.66710^{-}	0.78044^{+}
Gemma3-27B (PT)	0.55451^{+}	0.53745^{+}	0.59499^{+}	0.63705^{+}	0.50941^{+}	0.49088^{+}	0.58387^{+}	0.64797^{-}	0.78968^{+}
Gemma3-1B (IT)	0.51589^{+}	0.47635	0.48902	0.60815^{+}	0.45868^{+}	0.46234	0.53365^{+}	0.71364^{-}	0.75127^{+}
Gemma3-4B (IT)	0.47517	0.42279^{-}	0.44448^{-}	0.52310	0.42822	0.42271^{-}	0.45371^{-}	0.74263	0.70185
Gemma3-12B (IT)	0.51448^{+}	0.43624	0.49435	0.52473	0.45220	0.44352	0.48475	0.71455^{-}	0.68557
Gemma3-27B (IT)	0.48289	0.36519^{-}	0.41157^{-}	0.42902^{-}	0.39545^{-}	0.38945^{-}	0.38815^{-}	0.68913^{-}	0.63603^{-}

Table 9: Per-class scores of LIWC categories on the *Movie* corpus. Per-class scores significantly (p<0.05) over- or under-accommodating relative to the human baseline are annotated with $^{+}/^{-}$, respectively.

Model	LIWC Classes ↑								
	Per. PRON	Imp. PRON	Article	CONJ	PREP	AUX Verb	Adverb	Negation	Quanitifer
Human	0.51224	0.56815	0.52201	0.50898	0.57620	0.58099	0.46672	0.62373	0.41161
Random	0.47656	0.46571	0.48411	0.46186	0.55681	0.49711	0.44354	0.53601	0.40152
Llama3-1B (PT)	0.51461	0.54908^{-}	0.50790^{-}	0.41766^{-}	0.44772^{-}	0.48655^{-}	0.46538	0.80887^{+}	0.52413^{+}
Llama3-3B (PT)	0.50221	0.54298^{-}	0.50162^{-}	0.42035^{-}	0.45789^{-}	0.50929^{-}	0.47096	0.81561^{+}	0.50974^{+}
Llama3-8B (PT)	0.50871	0.55488^{-}	0.50834^{-}	0.41920^{-}	0.47860^{-}	0.52426^{-}	0.47534	0.80852^{+}	0.51937^{+}
Llama3-70B (PT)	0.51743	0.53373^{-}	0.50334^{-}	0.42873^{-}	0.49814^{-}	0.50906^{-}	0.46667	0.78768^{+}	0.53037^{+}
Llama3-1B (IT)	0.51252	0.52152^{-}	0.51547	0.46362^{-}	0.56144^{-}	0.52774^{-}	0.42933^{-}	0.77361^{+}	0.46223^{+}
Llama3-3B (IT)	0.50690	0.52605^{-}	0.52064	0.44533^{-}	0.54712^{-}	0.52491^{-}	0.43244^{-}	0.78154^{+}	0.46584^{+}
Llama3-8B (IT)	0.52654^{+}	0.54591^{-}	0.51884	0.44775^{-}	0.53512^{-}	0.55602^{-}	0.43379^{-}	0.79366^{+}	0.48750^{+}
Llama3-70B (IT)	0.52543^{+}	0.52190^{-}	0.51127	0.44526^{-}	0.53883^{-}	0.53572^{-}	0.43549^{-}	0.79175^{+}	0.48261^{+}
Gemma3-1B (PT)	0.50781	0.53966^{-}	0.48955^{-}	0.40049^{-}	0.46065^{-}	0.48000^{-}	0.46200	0.78614^{+}	0.51833^{+}
Gemma3-4B (PT)	0.49885^{-}	0.54071^{-}	0.49152^{-}	0.41224^{-}	0.46063^{-}	0.50473^{-}	0.46580	0.79908^{+}	0.51428^{+}
Gemma3-12B (PT)	0.49863^{-}	0.54130^{-}	0.49419^{-}	0.41503^{-}	0.47925^{-}	0.51447^{-}	0.46118	0.80888^{+}	0.51833^{+}
Gemma3-27B (PT)	0.50531	0.54580^{-}	0.49988^{-}	0.41498^{-}	0.48409^{-}	0.51519^{-}	0.46871	0.78867^{+}	0.51207^{+}
Gemma3-1B (IT)	0.46099^{-}	0.54123^{-}	0.51009^{-}	0.45534^{-}	0.49789^{-}	0.47316^{-}	0.44876^{-}	0.82397^{+}	0.50275^{+}
Gemma3-4B (IT)	0.47673^{-}	0.53623^{-}	0.51586	0.44279^{-}	0.50731^{-}	0.48312^{-}	0.45753	0.80174^{+}	0.47973^{+}
Gemma3-12B (IT)	0.47781^{-}	0.52282^{-}	0.51185	0.45791^{-}	0.54708^{-}	0.50205^{-}	0.44337^{-}	0.81047^{+}	0.46087^{+}
Gemma3-27B (IT)	0.48005^{-}	0.50935^{-}	0.49790^{-}	0.44714^{-}	0.54294^{-}	0.47903^{-}	0.45856	0.80388^{+}	0.45444^{+}

Table 10: Per-class scores of LIWC categories on the *NPR* Interview corpus. Per-class scores significantly ($p<0.05$) over- or under-accommodating relative to the human baseline are annotated with $^{+}/^{-}$, respectively.