Real or Robotic? Assessing Whether LLMs Accurately Simulate Qualities of Human Responses in Human-LLM Dialogue

Jonathan Ivey[*1] **Shivani Kumar**[*2] **Jiayu Liu**[*3] **Hua Shen**[*4] **Sushrita Rakshit**[*2]
Rohan Raju[*2] **Haotian Zhang**[*2] **Aparna Ananthasubramaniam**[*2]
Junghwan Kim[*2] **Bowen Yi**[*2] **Dustin Wright**[*6] **Abraham Israeli**[*2]
Anders Giovanni Møller[*7] **Lechen Zhang**[*3] **David Jurgens**[2]
[1]Johns Hopkins University [2]University of Michigan
[3]University of Illinois Urbana-Champaign [4]NYU Shanghai [5]University of Southern California
[6]University of Copenhagen [7]IT University of Copenhagen
jurgens@umich.edu

Abstract

Building datasets for dialogue tasks is expensive and time-consuming, requiring recruitment, training, and data collection from study participants. In response, much recent work has sought to use large language models (LLMs) to simulate both human-human and human-LLM interactions, as they have been shown to generate convincingly human-like text in many settings. However, how well do LLM-based simulations reflect *real* human dialogue? In this work, we answer this question by generating a large-scale dataset of 100,000 paired LLM-LLM and human-LLM dialogues from the WildChat dataset and quantifying how well the LLM simulations align with their human counterparts. Overall, we find relatively low alignment between simulations and human behaviors, with systematic differences in multiple textual properties, including style and content. Further, we find that models perform similarly in simulating English, Chinese, and Russian dialogues. Our results also suggest that LLMs only simulate a narrow range of the overall distribution of human dialogue, as they perform better on the subset of humans who write similarly to the LLM's own style.

1 Introduction

Large language models (LLMs) can generate convincingly human-like responses to a broad range of inputs. Recent work has examined their ability to simulate human interactions in different scenarios (Zheng et al., 2023; Köpf et al., 2024; Zhao et al., 2024), including humans interacting with other humans (human-human dialogue) and humans interacting with LLMs (human-LLM dialogue). For instance, LLMs may be used to simulate the performance of moderator bots (Cho et al., 2024), chatbots (Tamoyan et al., 2024), and UIs (Liu et al., 2024b), as well as generate additional data for fine-tuning models (Njifenjou et al., 2024a). In addition to task-specific applications, multi-purpose simulators generate a range of conversations across diverse dialogue settings (Kim et al., 2024; Sekulic et al., 2024; Wang et al., 2024b). Such simulations can greatly facilitate collecting data, which often require costly human labor and are difficult to scale to the diversity of LLM abilities. However, this approach can only be effective if the responses generated by the LLM mirror how a human would interact in different scenarios. In this work, we ask: to what extent can LLMs simulate the responses of humans in a human-LLM dialogue?

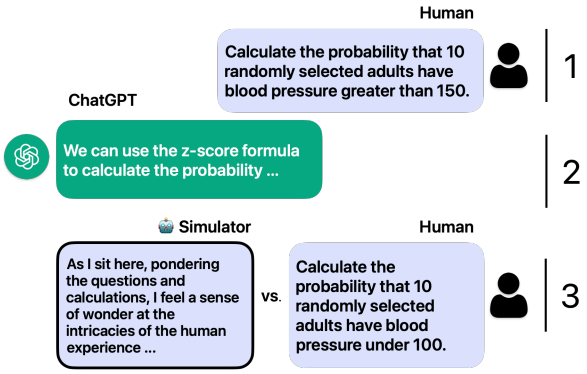


Figure 1: A sample conversation between a human and GPT-3.5 on a sample from the WildChat data. Turn 3 of the conversation was originally generated by a human. In this study, we use the SIMULATOR to mimic the human's Turn 3 response. We then compare the SIMULATOR's output against the HUMAN's output using 21 metrics, covering lexical, syntactic, semantic, and stylistic features.

To evaluate the simulation capabilities of LLMs, we evaluate how closely they reproduce human responses in real human-LLM dialogues. Our study asks three research questions: (**RQ1**) to what extent does the choice of model and prompt instructions influence the LLM's ability to simulate human responses in human-LLM dialogue; (**RQ2**) how do these results generalize to languages other than En-

[*]Research was done while all authors were at the University of Michigan. All authors have equal contribution, and order is randomized except senior author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 41395–41432
July 2-7, 2026

glish; and (**RQ3**) in what contexts are LLMs most effective at simulating human responses?

To answer these questions, we develop an evaluation framework on top of human responses from the WildChat dataset (Zhao et al., 2024) – a corpus of one million conversations between HUMANs and CHATBOTs across multiple languages and domains. The dataset contains multi-turn conversations, making it well-suited to our research objective. For each dialogue, we compare the HUMAN 's response with that generated by an LLM (denoted as SIMULATOR) as illustrated in Figure 1. Responses are compared across multiple categories—including lexical, syntax, semantics, and style—to assess their fidelity to human utterances. Using multiple LLMs and prompts, we evaluate which settings led to better simulation and use regression analysis to identify the factors most strongly associated with generating human-like responses.

This paper makes four key contributions. First, we introduce a general evaluation framework for human-LLM simulations and a new dataset of over 1.2K annotator responses for a human-level performance comparison. Second, we conduct a large-scale analysis of 9 LLMs across 50 prompts simulating 2K English human-LLM conversations, and find that even the best model and prompt combinations are relatively weak at simulating human responses. Third, we extend our analysis to 10K Chinese and Russian human-LLM conversations, finding roughly similar performance across languages. Fourth, through regression analysis, we show that SIMULATORs have limited ability to adapt to different writing styles and conversational settings. All data and code are available at *https://github.com/davidjurgens/human-llm-similarity*.

2 LLMs to Simulate Human Interaction

Given the practical use of LLMs in mimicking human turns in conversations, many studies deal with simulating human-LLM interactions and developing relevant conversational datasets for generic conversations (Tamoyan et al., 2024; Njifenjou et al., 2024b; Gosling et al., 2023) or for conversations of a specific domain, such as education (Abbasiantaeb et al., 2024), health (Cho et al., 2023; Liu et al., 2025a), or programming (Liu et al., 2024b). Similarly, LLM-LLM interactions highlight the general capability of LLMs to mimic human discussions (Park et al., 2023; Zhou et al., 2024; Rossetti et al., 2024; Zhou et al., 2023; Chen et al., 2023). In order to facilitate further research, a battery of studies have introduced various datasets of human-LLM (Zheng et al., 2023) and LLM-LLM (Kim et al., 2022; Chen et al., 2023) dialogue. Digital Twins have also been studied as virtual replicas of physical systems, in this case, humans in various discussion settings (Barricelli et al., 2019; Barricelli and Fogli, 2024; Rossetti et al., 2024; Wen et al., 2024). It is common in many of these contexts for LLMs to augment—or even replace—human labor with simulations, thereby reducing time and effort (Kojima et al., 2022; de Wit, 2023).

Humans, however, tend to have their specific style, intent, and self-creativity, which are challenging for LLMs to mimic despite recent technological breakthroughs (Stevenson et al., 2022; Wu et al., 2023; Wolf et al., 2023; Gui and Toubia, 2023; Jiang et al., 2024b). Specifically, Leng and Yuan (2023) highlight that while LLMs show promise for applications in social science research, further investigation into their subtle behavioral differences from humans and the development of robust evaluation protocols are essential, thus motivating our research. Multiple metrics have been proposed to measure these differences, including content relevance (Abeysinghe and Circi, 2024), emotional alignment (Mæhlum et al., 2024), and intent accuracy (Kim et al., 2024). A few studies (Sedoc et al., 2019; Svikhnushina and Pu, 2023; Mayor et al., 2025; Lu et al., 2025) have explored broader evaluation techniques by developing frameworks that align closely with human judgment. However, comprehensive research about how different measures vary between LLMs and prompts is still lacking. In this work, we address this measurement gap by analyzing how LLM responses compare to human responses across multiple similarity measures, LLM models, and prompts.

3 Problem Definition

LLMs are often used to simulate humans in various dialogue settings (e.g., moderation, customer service), based on the assumption that LLMs will simulate the behaviors of interest because they can faithfully reproduce the underlying dialogue. While task-specific evaluations remain necessary, our research questions test the underlying assumption by examining the conditions under which SIMULATORs can replicate the ways in which HUMANs prompt CHATBOTs like ChatGPT. We adopt the fol-

lowing modeling task, which provides a controlled setting for this question: Given the initial HUMAN utterance, followed by the CHATBOT's response, we prompt an LLM SIMULATOR to suggest the next HUMAN response in the discussion (i.e., HUMAN Turn 3). The "true" HUMAN Turn 3 is known (but not given as input to the SIMULATOR), as it is part of the WildChat dataset and thus acts as a reference for evaluation.

This modeling setup offers the following desirable properties. First, using HUMAN Turn 3 as a reference allows us to examine multiple factors influencing simulation quality: (i) the linguistic similarity between SIMULATOR and HUMAN responses, (ii) the impact of the initial HUMAN Turn 1 and CHATBOT Turn 2, and (iii) the effects of the model and prompt used. Second, limiting to HUMAN Turn 3 allows us to study these factors *in isolation* at *the early stage of dialogue*, i.e., without the added influence of multiple (simulated or natural) turns. This is useful because it reveals how well the SIMULATOR can extend a conversation with minimal context. Finally, this setup maximizes the evaluation dataset from WildChat while still containing multiple turns, as most conversations in the dataset end with 3 or fewer turns.

4 Generating Simulation Data

To evaluate how well LLMs can simulate HUMAN-CHATBOT interactions, we generate a large dataset of dialogue simulations. Since many applications require simulating a broad spectrum of conversations, we use a dataset containing a wide range of dialogues between HUMANs and a CHATBOT (typically GPT-3.5), and our evaluation framework is comprehensive and multidimensional.

Wildchat Data *WildChat* (Zhao et al., 2024) is a multilingual corpus of one million conversations between HUMANs and a high-performing CHATBOT, totaling over 2.5 million interaction turns. English accounts for 53% of the turns. Our analysis uses a sample of 102k English instances, along with 10k Chinese and 10k Russian instances for multilingual experiments in §6.

To generate SIMULATOR responses, we take the first two turns of each conversation (HUMAN initial query and CHATBOT response) as input. The model is then prompted to generate what the HUMAN Turn 3 response would be, or to indicate the HUMAN would not have responded (Figure 1).

Prompt Composition Since the wording of a prompt can have a significant impact on an LLM's output and adopted persona (Röttger et al., 2024; Wright et al., 2024; Ceron et al., 2024), we conduct experiments with a diverse set of prompts. Working independently, 12 authors familiar with LLMs composed a total of 50 candidate prompts for this task. The prompt's aim was to have the SIMULATOR match the conversational intent, content, and style of the first turn and to decide whether the conversation continues or not. Each prompt writer was given 10 randomly sampled dialogues from WildChat as a reference, and were encouraged to test their prompts using a CHATBOT or other available tools. Three example prompts are shown in Supplemental Table 7.

Prompt Taxonomy We built a six-factor prompt taxonomy (Appendix F), characterizing attributes like whether each of the 50 prompts assigned the LLM an explicit role (e.g., "you are a human"), was framed as a question or instruction, attempted prompt-hacking (e.g., jail-breaking, injection), included specific stylistic instructions (e.g., be "human-like"), employed common techniques like chain-of-thought reasoning or decomposition, and included instructions to consider the HUMAN's intent. Two authors collaboratively labeled each prompt's strategy, finding strong representation across different types of prompts (Table 4).

Generating and Parsing Simulated Turn 3 Each of the 50 prompts was used to generate the third turn of randomly sampled English-language dialogues from WildChat (details in Appendix §A). We employed 9 models, selected to represent a range of sizes from the most widely used open-weight model families on HuggingFace at the time of writing.[1] To extract the SIMULATOR's response text, we applied custom regular expressions for each prompt. 4 prompts were discarded because they frequently failed to produce valid output.

Evaluating LLM Simulations To evaluate simulation quality, we select 21 linguistic measures spanning 4 categories: style, lexical, syntactic, and semantic (Appendix Table 1). These categories and

[1]The nine models are: Meta-Llama-3.1-8B-Instruct (Dubey et al., 2024), Meta-Llama-3.1-70B-Instruct (Dubey et al., 2024), Meta-Llama-3-70B-Instruct (AI@Meta, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024a), Mistral-Large-Instruct (Mistral, 2024), Phi-3-medium-4k-instruct (Abdin et al., 2024), Qwen2-72B-Instruct (Yang et al., 2024), and c4ai-command-r-v01 (Cohere, 2024)

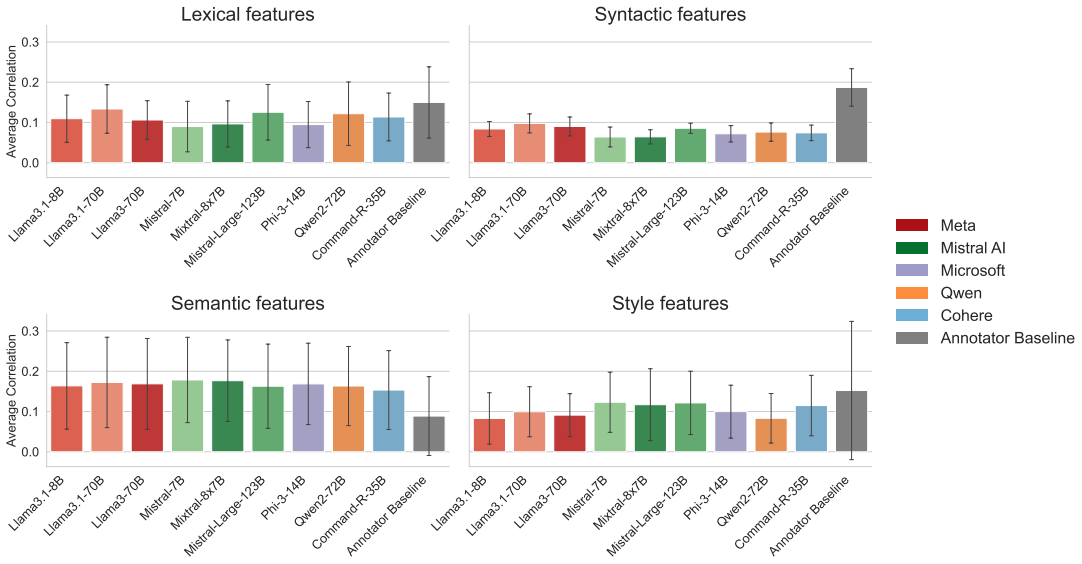


Figure 2: How well do LLMs simulate HUMAN responses to a CHATBOT? We compare the nine models used as SIMULATORS to the original HUMAN by correlating properties of the text they write (see Table 1 for the metrics used). Bars represent the average correlation across all metrics in a category, and error bars are bootstrapped 95% confidence intervals over these metrics. As a baseline, we also compare the performance of a human annotator on this task. There is limited cross-model variation in performance, and SIMULATORS tend to have higher performance in semantic features and lower performance in syntactic features, while the opposite is true of the human annotators.

measures were selected due to their widespread use in NLP tasks (Sebastiani, 2002; Stamatatos, 2009; Fu et al., 2018; Ribeiro et al., 2020). The output of these measures can be either a scalar, a probability distribution, or a feature vector. Scalars include both unbounded measures (e.g., utterance length) and bounded measures in $[0, 1]$ (e.g., text sentiment). Probability distributions include, e.g., the probability that a token with a particular part of speech appears in a sentence across different parts of speech. Feature vectors include semantic (SBERT, Reimers and Gurevych (2019)) and style (LUAR, Soto et al. (2021)) embeddings. Finally, we record whether the SIMULATOR and HUMAN end the conversation at Turn 3 as a binary outcome.

5 Can LLMs Simulate Human Replies?

In the first experiment, we measure the overall similarity of SIMULATORS to HUMANS across a variety of models and prompts (**RQ1**). Overall, we find that the choice of SIMULATOR has only a minor impact on the ability to simulate HUMANS, while the design of the prompt is most relevant when optimizing performance.

5.1 Experimental Setup

For our first experiment, we use the setup from §4 to generate 828K simulated responses – 9 models, 46 prompts, and 2,000 randomly-sampled English conversations (1,239 end at Turn 2, 761 continue into Turn 3). To evaluate performance, we measure the similarity between the original HUMAN Turn 3 response to the SIMULATOR Turn 3 response in 21 lexical, syntactic, semantic, and style properties. Example metrics include utterance length and perplexity for lexical, parts of speech distribution for syntax, SBERT embeddings for semantic, and politeness for style (cf. Table 1 and Appendix C).

We calculate the similarity between HUMAN and SIMULATOR Turn 3 for each of these 21 metrics. For scalar metrics (e.g., length, classifier probabilities), similarity is the Pearson's R correlation between HUMAN and SIMULATOR values, reflecting whether SIMULATORS mirror relative differences among HUMANS (e.g., who writes longer, more formal, etc. replies). For distribution metrics (e.g., POS tags, topics), similarity is the average correlation of HUMAN and SIMULATOR class frequencies over all classes in the distribution. For vector metrics (e.g., BERT, LUAR), similarity is the average correlation of each dimension of the HUMAN and SIMULATOR embedding;[3] since the dimensions of embeddings are not inherently meaningful, we first

[3]A more standard approach for calculating similarity would be to take the average cosine similarity for each HUMAN/SIMULATOR pair. However, we chose to use correlations, so the feature vector similarity scores would be comparable to other metrics. Doing so yields similarity scores that are highly correlated with the corresponding cosine similarities.

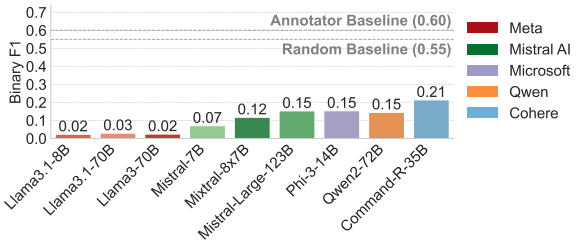


Figure 3: How well do LLMs predict whether HUMANs end a conversation with CHATBOT after the first turn? Each bar represents the binary F1 score of each model predicting whether a conversation will end. The gray horizontal lines show the performance of human annotators and a random baseline that ends the conversation 50% of the time. While there is inter-model variation, all models perform worse than chance. The human annotator performs better than chance.

rotate the embeddings using principal components analysis, so each dimension corresponds to an embedding's alignment with the direction of a principal component. To aggregate results, we averaged the correlation values to create four category-wise similarity scores for lexical, syntactic, semantic, and style, using the grouping in the first column of Table 1. For the decision of whether to end the conversation at Turn 3, we use the binary F1 score to compare HUMAN and SIMULATOR responses.

Although we have the HUMAN ground truth for Turn 3, two humans in the same setting may act differently, so we are unable to capture conversation-specific variation. Instead, we compare relative differences in metrics across simulation approaches to ensure fair comparison. To estimate the range of plausible human responses in a given context, a group of 12 authors (who did not contribute to WildChat data) completed the same task as the SIMULATORS by generating Turn 3 responses. A total of 1,273 turns were annotated (Appendix §B has details). We compared SIMULATOR performance on the subset of turns where both annotators and SIMULATOR continued the conversation.

5.2 Comparison between SIMULATORs and Human Annotator Baseline

We compare the similarity of different SIMULATORS to the original HUMAN in Figure 2. We also examine the impact of prompt design on SIMULATOR-HUMAN similarity in Figure 4. For each category, we plot the average similarity for the best-performing prompt (i.e., the highest average correlation across all metrics), the worst-performing prompt, and human annotator baseline. See prompts in Table 6 of Appendix D. The similarity of all metrics across prompts is shown in Appendix Table 10.

Overall, similarity is lower across all settings for the syntactic measures, while other measures tend to be more similar. Across all models, there is effectively no difference between the similarity of SIMULATORs and the similarity of the human annotator baseline across the lexical, semantic, and style categories of measures. Additionally, for both the semantic and style categories, the confidence intervals for the annotator baseline extend beyond 0, the performance we would expect from a totally random baseline, while each SIMULATOR maintains non-random performance. However, there is a discernible difference in how similar annotator utterances are syntactically to the original HUMANs compared to the SIMULATORs. Similarly, the best prompts have comparable lexical, semantic, and stylistic similarity to the human annotator baseline, while human annotators have higher syntactic similarity than even SIMULATORs with the best prompt. This finding suggests that humans and LLMs have complementary strengths in simulating dialogue; in order to most accurately reflect human utterances, a human-in-the-loop approach where SIMULATORs and annotators play complementary roles may be appropriate. The correlation results for all individual metrics across SIMULATORs are shown in Appendix Table 9.

In Figure 3, we compare the similarity between SIMULATORs and HUMANs in their tendency to end the conversation. SIMULATORs seldom end the conversation, continuing 87.1% of the time for COMMAND-R to 98.6% for LLAMA3.1-8B. In contrast, human annotators more accurately mirror the original HUMAN behavior in ending the conversation, indicating that collaboration between SIMULATORs and humans is effective for simulating human-LLM interactions. Additionally, human annotators show a better ability to determine when a conversation is likely to end, more often predicting the end at Turn 2 when it actually occurs, whereas SIMULATORs tend to predict conversation end with similar frequency regardless of the HUMANs actual behavior (Figure 17). The performance differences across SIMULATOR models are mainly driven by how frequently each model ends the conversation and not based on distinguishing when the HUMAN does and does not end the conversation. However, low-moderate results in the human baseline suggest that the task of predicting the next response in a human-LLM conversation is very difficult for both humans and LLMs.

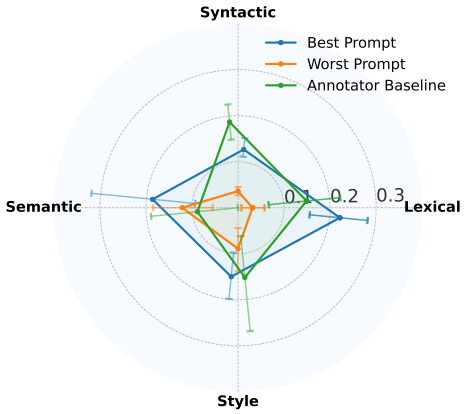


Figure 4: Using the methods from Figure 2, the performance of the best and worst prompts and annotators are compared across metric categories. The best and worst prompts are selected based on an overall average across all metrics and shown in Table 6. The worst prompt underperforms the best prompt in all categories, and annotators outperform all prompts in syntax metrics.

5.3 Comparison Across Models and Prompts

There is little variation in performance across different SIMULATORs (Figure 2), with often statistically indistinguishable performance between the best and worst model (Appendix H, Figure 13). To check the robustness of these results, we replicated the experiment with two larger, proprietary models as well as with the base version of three of the open-weight models (i.e., not fine-tuned with supervised finetuning (SFT) or reinforcement learning with human feedback (RLHF)) (Appendix D). These experiments allowed us to test whether our findings hold across a broader range of models that may better simulate human dialogue. We find that the proprietary and base models do not have significantly better performance than the open-weight instruct models we used in Figure 2, reinforcing our findings that model family, size, and training protocols do not significantly influence performance.

Across all categories, the best and worst prompts have significantly different similarity scores, suggesting that prompt engineering has greater impact than model selection for simulating human-LLM interactions (Appendix H, Figure 13). This effect is robust based on the error bars of each method. The greatest effects are in the lexical and syntactic categories, while the prompt has less effect on semantic and stylistic similarity. Given that models also demonstrate little difference in these categories, it is challenging to engineer human simulations using LLMs that are semantically or stylistically similar, while it is possible with good prompt engineering to improve lexical and syntactic similarity. Interestingly, there are no significant differences between how well the best- and worst-performing prompts predict whether a conversation is going to end, suggesting that methods other than prompt engineering may be required to improve performance in this metric.

5.4 Effects of Prompt Characteristics

In Appendix F, we explore how the different strategies used across the prompts in this study affected SIMULATOR performance. We ran a regression with prompt performance as the dependent variable and the prompting strategy as the independent variable. We find that a few performance-enhancing techniques, including explicitly assigning the SIMULATOR the role of "human talking to an LLM" (Figure 10b) and framing prompts as questions rather than instructions (Figure 10c). However, generally, simpler prompts tended to perform better than prompts with a lot of specialized techniques or specific instructions (Figure 11). Consistent with our other results, prompt characteristics also have little influence on SIMULATORs' propensity to end conversations.

5.5 Predicting Turns 5-11

Finally, we calculate the SIMULATOR's performance beyond the third turn of the conversation. SIMULATORs may have better performance at predicting later turns of the conversation, since they have more examples of the HUMAN's writing as context. Prior research supports this hypothesis, as LLMs better capture user intent in later turns of a dialogue (Sarkar et al., 2025). To test the effects of giving the LLM more conversational context, we ask the LLM to generate the n^{th} turn of a conversation given turns 1 through $n-1$ (Appendix E). We do this for five values of $n \in \{3, 5, 7, 9, 11\}$. Surprisingly, we find no significant difference in the SIMULATOR performance across turn depths (Figure 8). Our results suggest that improving SIMULATORs may require more than simply providing more examples of the HUMAN's writing.

6 Simulation in non-English Languages

Next, we address how the results from §5 generalize to languages beyond English (**RQ2**). We find that LLMs' performance as SIMULATORs is largely consistent across English, Chinese, and Russian, although some differences suggest that languages

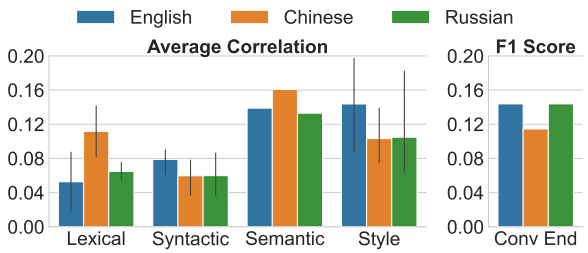


Figure 5: How well do SIMULATORs replicate HUMAN text across languages? Similar to Figures 2 and 3, we plot the similarity between SIMULATOR and HUMAN text across ten metrics in three languages. English, Chinese, and Russian have similar performance patterns across all five categories of metrics. However, some differences exist (e.g., Chinese SIMULATORs outperform other languages in lexical and semantic metrics but underperform in conversation ending). Correlations of individual metrics are shown in Tables 11 and 12.

for which SIMULATORs had less training data may have less robust performance across contexts.

6.1 Experimental Setup

We compare the performance of SIMULATORs across three languages: English, Chinese, and Russian. Chinese and Russian were selected for their substantial representation in the WildChat dataset. Specifically, Chinese and Russian comprise 15.9% and 10.5%, respectively, whereas English accounts for 53%. For our analysis, we randomly sample $10,000$ conversations from each language (6,382 end at Turn 2, 3,618 continue into Turn 3).

To generate conversations, we use the largest and most performant models from the previous section that were trained on all three languages: MISTRAL-LARGE-123B, LLAMA3.1-70B, and MIXTRAL-8X7B. We identify the most effective English prompts of different types and had native speakers translate them into Chinese and Russian. We select three prompts (Appendix Table 7) that consistently perform well across six largely uncorrelated metrics: capitalization, punctuation, part of speech, SBERT embeddings, sentiment, and politeness (Appendix §H, Figure 15).

To measure the similarity between HUMAN and SIMULATOR responses, we use a procedure similar to the one in §5. Due to the unavailability of several metrics from §5 for Chinese and Russian, we focus on a subset of ten metrics covering all four categories (details in Appendix Table 8.). We employ consistent or similarly trained models whenever possible to ensure comparability across languages.

6.2 Comparison Across Languages

The aggregated metrics for the three languages are depicted in Figure 5. Consistent with the findings discussed in §5, all three languages show higher correlations between SIMULATORs and HUMANs in the semantic metric but lower correlations in the syntactic metric. For detailed correlations of individual metrics, we refer to Appendix Table 11 and Table 12. Additionally, predicting whether a conversation will end consistently performs below chance across all languages, as SIMULATORs are unlikely to predict a conversation will end. These observations reinforce the conclusions drawn in §5.

Notably, there are differences between the languages. Chinese SIMULATORs outperform their English and Russian counterparts in lexical metrics and show slightly better performance in semantic metrics. The differences in lexical metrics, such as utterance length and perplexity, may be attributed to the typically shorter sentence lengths in Chinese. Conversely, English SIMULATORs excel in predicting style metrics compared to those in other languages. Toxicity and sentiment metrics primarily contribute to this improvement. In English, SIMULATORs more accurately reflect the variations in toxicity and sentiment of HUMANs. This capability may vary across languages because Chinese and Russian have relatively smaller amounts of training data. Consequently, safety training may impact these languages' outputs more significantly than English, leading to a strong prior on toxicity and sentiment that hinders the style match to HUMANs. Additionally, English SIMULATORs demonstrate superior accuracy in modeling syntactic metrics. Notably, prior studies on LLMs often highlight significantly better performance for English data, a trend we also expected to observe. Surprisingly, our findings reveal comparable performance across English, Chinese, and Russian.

6.3 Comparison Across Models

The choice of model and prompting strategy affects performance, as shown in Appendix Figure 16b. The differences between models become more pronounced in languages other than English, as shown in Figure 16a. For example, while model differences in English are often minor, the smallest model (MIXTRAL-8X7B) frequently underperforms compared to other models in Chinese and Russian. Moreover, the significant inter-model variation in Chinese and Russian may be due to

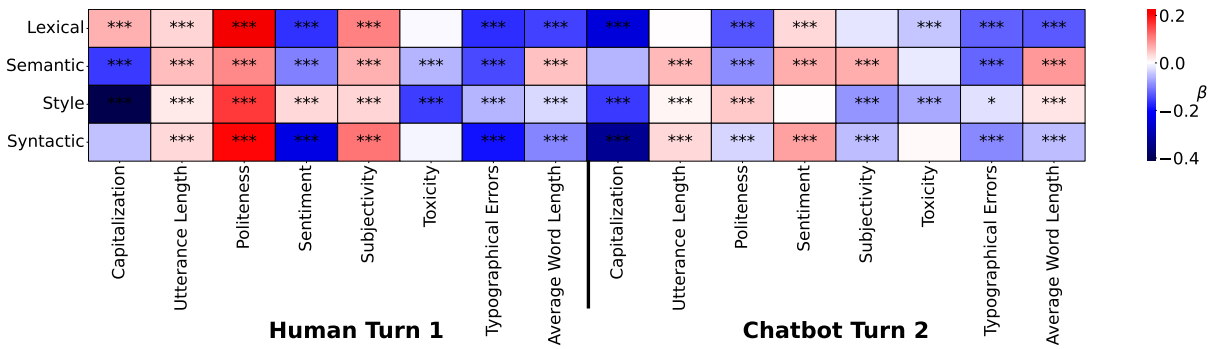


Figure 6: In what contexts do SIMULATORs best predict HUMAN responses? We show the results of four regressions predicting the similarity between SIMULATOR and HUMAN at Turn 3 for different categories (rows), using HUMAN Turn 1 and CHATBOT Turn 2 properties as features (columns). We highlight a subset of the coefficients here, where red and blue colors indicate positive and negative regression coefficients β respectively, and stars in each cell indicate the statistical significance of each β after applying a Bonferroni correction (* p<0.05, ** p<0.01, *** p<0.001); Full regression coefficients are given in Table 13 and Table 14. The linguistic properties of HUMAN Turn 1 have stronger effects than those of the CHATBOT in Turn 2, showing that SIMULATORs do correctly accommodate more to the HUMAN style. However, SIMULATORs tend to perform better when the HUMAN's Turn 1 more closely matches the properties of typical SIMULATOR-generated text (e.g., more polite, fewer typos).

larger models having more non-English samples in their pre-training data, which enhances their performance in non-English languages. Additionally, the F1 score shows more variation across the models and prompting strategies than the text properties metrics, suggesting that engineering decisions may be more salient for closed-ended conversation-ending tasks than open-ended text generation tasks. Appendix §H details the differences.

7 When Do LLMs Successfully Simulate?

Experiments from §5 and §6 paint a high-level picture of how the SIMULATORs differ from HUMANs across linguistic metrics. However, an important question remains: *when* do LLMs succeed as SIMULATORs of human conversations? In our final experiment, we answer **RQ3** by investigating which factors directly impact the differences between SIMULATORs and HUMANs. Overall, we find that SIMULATORs best mirror HUMANs in a narrow set of contexts suited to their safety training and that the models poorly adapt to the range of human speech styles or topics when attempting to generate similar responses.

7.1 Experimental Setup

For this experiment, we focus on depth as opposed to breadth and generate Turn 3 utterances for a random sample of 100,000 conversations from the English-language WildChat corpus (63,722 end at Turn 2, 36,278 continue into Turn 3). As in §6, we generate responses using the best-performing English prompts from §5 with MISTRAL-LARGE-123B, LLAMA3.1-70B, and MIXTRAL-8X7B. After generating responses, we perform a regression analysis to identify factors linked with higher or lower similarity between SIMULATOR and HUMAN responses (details in Appendix G). For regression, we use the average similarity across the measures in Table 1 as the dependent variable. For independent variables, we use the conversation metadata (the CHATBOT model used, the region where the HUMAN is located, the SIMULATOR model used, the prompt), scalar metrics run on HUMAN Turn 1 and CHATBOT Turn 2 utterances (capitalization, utterance length, politeness, sentiment, subjectivity, toxicity, typographical errors, and average word length), and 50 topics generated using Latent Dirichlet Analysis (LDA) on Turns 1 and 2.

7.2 Factors Predicting Performance on Linguistic Properties

We first observe that HUMAN Turn 1 has a stronger influence than CHATBOT Turn 2 on SIMULATOR Turn 3 (Figure 6). In other words, the difference between HUMAN Turn 3 and SIMULATOR Turn 3 is explained more by HUMAN Turn 1 than CHATBOT Turn 2. This is likely because two utterances by the same actor (in this case, HUMAN Turns 1 and 3) tend to be linguistically similar, while SIMULATOR Turn 3 has relatively low variation in linguistic features between different conversations.

Accordingly, we find that the performance of Turn 3 simulation is dependent on the simulated HUMAN producing conversations, which are *already* similar to the responses generated by the particular language model and prompt combination. For example, Figure 6 shows that when HUMAN

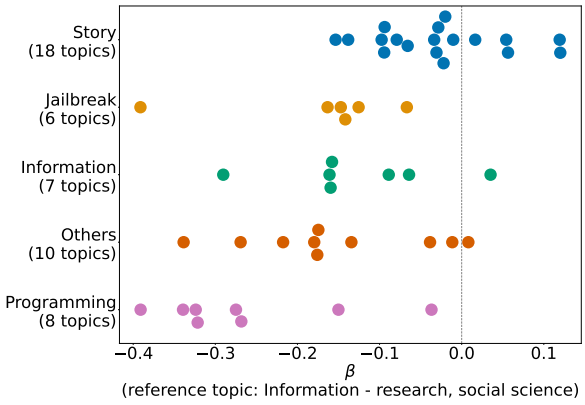


Figure 7: The topic of HUMAN Turn 1 on SIMULATOR influences its performance. Topics are obtained from LDA and manually grouped. We plot the β coefficient of the topic in the regression described in Figure 6.

Turn 1 expresses higher politeness, the Turn 3 simulation is predictably more similar, which is likely because LLM responses are more polite by default. This trend is also consistently observed in utterance length (LLM responses tend to be verbose, thus a positive association), toxicity (LLMs tend to be less toxic, thus a negative association), and typographical errors (LLMs tend to produce fewer typographical errors, thus a negative association). Our findings suggest that LLMs do not adapt to the various styles that humans have. Instead, their outputs are written in a distinct, idiosyncratic style, biasing performance towards humans who tend to naturally communicate in this style. Therefore, capturing the spectrum of linguistic variation that HUMANS naturally express is an open challenge that may require specialized solutions.

The conversation topic of the initial post (HUMAN Turn 1) is a strong predictor of whether the SIMULATOR can generate similar content in Turn 3 (Figure 7). SIMULATORS tend to be better at story topics (e.g., "Describe a day in the life of a superhero"), potentially because when the original request is for a story, the conversation often proceeds as a continuation of the story. These dialogues tend to have consistent style and content, which makes simulation easier. In contrast, conversations about programming or technical topics (e.g., "Write a Python script to ...") have a lower similarity score. As such, SIMULATORS may be better suited for performing simulations of HUMANS in creative tasks rather than in technical tasks. However, prompts crafted for different conversation topics may improve performance when simulators do not do well with a topic-agnostic prompt.

7.3 Factors Predicting Performance on Conversation Ending

We find that properties of CHATBOT Turn 2 are more strongly associated with whether the SIMULATOR can predict when a conversation will end than HUMAN Turn 1 (Figure 12). Understanding whether a conversation has ended requires the SIMULATOR to understand what the HUMAN in Turn 1 was looking for and assess whether the response in Turn 2 was satisfactory (Sarkar et al., 2025). This finding is consistent with the idea that the SIMULATOR needs to determine whether the response in Turn 2 satisfies the HUMAN's intent in Turn 1.

8 Discussion and Conclusion

While LLMs are often used to simulate humans in dialogue settings, minimal research systematically evaluates how they work in the wild. By systematically evaluating the performance of an LLM SIMULATORS in real HUMAN-CHATBOT dialogues, our study has shown that existing LLMs fall short across several metrics. While LLMs perform better at replicating human responses on a semantic level, they encounter difficulties in accurately mirroring human syntax and conversational dynamics. In particular, all the SIMULATORS tend to continue conversations when a human might naturally end them, highlighting a gap in their ability to detect conversational intent. We find that different prompts and models have minimal variation in their ability to predict when a conversation ends, suggesting that some external mechanism is required to determine when to end a conversation when using LLMs for dialogue simulation. Our analyses show that the choice of prompt instructions significantly impacts the quality of simulations, often more so than the family or even the size of the SIMULATOR model. We characterize the types of prompts that perform best as SIMULATORS, and future work could more systematically explore the prompt space and prompt optimization techniques. Moreover, multiple papers have shown models perform worse for non-English data, but our analysis showed that models performed roughly the same for English, Chinese, and Russian. Finally, we observe that the LLMs' effectiveness in simulating human responses is context-dependent: they perform strongly in dialogues that maintain a consistent style, such as storytelling, and weakly in more structured or technical domains like programming.

9 Limitations

Simulating human responses in human-LLM dialogues is inherently challenging due to its open-ended nature, and our study highlights the diverse directions such interactions can take. While we suggest a broad set of diverse prompts, we did not put most of our effort into optimizing those prompts for any specific metric or predefined goal, even though a more intentional crafting of prompts might improve the overall similarity score. Indeed, our findings indicate that finding the "right" prompt, rather than the "right" model, holds the greatest potential for improvement. As a preliminary step in understanding effective prompting strategies, we explored how potential differences in prompt content and techniques could impact SIMULATOR performance (Appendix F). Future research could perform a more systematic exploration of prompting strategies to explore whether prompt optimization, tailored to a specific task or metric, yields better results.

In this paper, we decided to simulate the *third turn* in a human-LLM conversation, tasking the SIMULATOR with generating a response based on a short preceding discussion. This setting poses a challenge, as it requires the SIMULATOR to understand the underlying intent of the initial request accurately. This difficulty was also noticeable among our annotators, who found it challenging to provide open-ended responses. We focused on simulating just the third turn because: (1) The performance on the third turn was sufficiently low for LLM simulators that any tests of later turns would be even worse (once simulators got off on the wrong foot, the conversation was sufficiently different from humans that the dialogue isn't useful to measure). (2) Practically speaking, WildChat has relatively few long dialogues, so our analysis would have a much lower sample size to draw conclusions from. However, future research could focus on predicting conversational outcomes using a longer seed conversation, which are likely to occur in many real-world applications (e.g., customer service) and might better capture the nuances and intent of the interaction. Additionally, future work could adapt our framework to test not only how well SIMULATORs can predict the next turn but the entire rest of the conversation. Future work could also explore what factors contribute to improvements in performance at deeper turns, including how well SIMULATORs can handle changes in topic, style, and tone which may be more likely to occur at later turns.

While measuring the similarity between textual content, we use a broad set of metrics to capture a diverse range of language characteristics. However, this list is not exhaustive and can be further modified. Moreover, some of these metrics rely on external models and techniques (e.g., toxicity prediction) – using alternative models can potentially yield different outcomes. This is most relevant for our multilingual experiment. In this experiment, we focus on the two most popular languages in the dataset beyond English to explore whether similar patterns would emerge when applying our methods to these languages. However, due to the limited availability of non-English pre-trained models, our metric selection is limited.

Additionally, our study uses a set of general metrics to measure the similarity between LLM and human turn 3. While this approach is useful for multi-purpose user simulators that can adapt to different dialogue settings (Sekulic et al., 2024; Wang et al., 2024b), many applications of dialogue simulation are built within a specific context (e.g., to test a bot, to generate synthetic data), and in these cases, the evaluation of these simulators would also need to be tailored to the specific task. Our study does not include these more tailored evaluation metrics. Instead, we focus on the more general question of whether an LLM can generate turns of a dialogue like a human, in order to test the premise that both general-purpose simulators and task-specific simulators can approximate the underlying dialogue in a range of applications. Additionally, we also focus on quantifying similarity in turn 3 utterances because, even in specific applications, not faithfully generating human-like text could have negative downstream consequences. For instance, when a simulator is used to test whether some system (e.g., a bot) is effectively able to respond to the human's behavior, the text of the simulator's response will affect the way the system responds and, ultimately, the final outcome. Therefore, a SIMULATOR that does not linguistically match the HUMAN is likely to create a false sense of the system's ability to respond appropriately. To test performance on task-specific metrics, future work could adapt our experiment structure in the context of a particular application.

10 Ethical Considerations

We use the WildChat dataset (Zhao et al., 2024) as our main data resource for the research. We made sure to follow their ethical guideline while using the data. Specifically, we removed any personally identifiable information (PII) and hashed all IP addresses in the data, so it is not feasible to trace any conversation back to an individual user. As Zhao et al. (2024) mentioned, all WildChat data undergo internal reviews conducted by the AI2 legal team to ensure compliance with data protection laws and ethical standards. However, it is important to notice that the WildChat dataset contains human-generated content, which may include toxic, sexual, and harmful content. Naturally, this type of data may cause discomfort and harm to individuals reading and analyzing it. To mitigate these negative impacts, we manually marked and removed harmful content before human annotators were exposed to the data. Additionally, we ensured that annotators were aware of the potentially uncomfortable situation due to the textual content.

In this research, we use LLMs to simulate human responses in dialogue. Although many studies have shown that their outputs are highly "human-like" (Aher et al., 2023; Bang et al., 2023; Liu et al., 2023; Webb et al., 2023), they are prone to problems like generating harmful and biased content. For example, they are known to exhibit political and gender biases (Hartmann et al., 2023; Liu et al., 2024a; Cao et al., 2023) and fail to represent diverse identity groups or cultures (Wang et al., 2024a; Tao et al., 2024; Naous et al., 2024). These bottlenecks hinder LLMs' ability to faithfully represent diverse human dialogue, which researchers should be aware of (Abdurahman et al., 2023).

Acknowledgments

The authors thank Nasanbayar Ulzii-Orshikh for translating the prompts in §6 into Russian. This work was supported in part by the National Science Foundation under Grant No. IIS-2143529. Dustin Wright was supported by a Danish Data Science Academy postdoctoral fellowship (grant: 2023-1425). Anders Giovanni Møller was supported by the Carlsberg Foundation through the COCOONS project (CF21-0432).

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2023. Perils and opportunities in using large language models in psychological research. *OSF Preprints*, 10.

Bhashithe Abeysinghe and Ruhan Circi. 2024. The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches. In *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval*.

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

AI@Meta. 2024. Llama 3 model card. Technical report, Meta.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Barbara Rita Barricelli, Elena Casiraghi, and Daniela Fogli. 2019. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE access*, 7:167653–167671.

Barbara Rita Barricelli and Daniela Fogli. 2024. Digital twins in human-computer interaction: A systematic review. *International Journal of Human–Computer Interaction*, 40(2):79–97.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2020. Embeddings in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 10–15, Barcelona, Spain (Online). International Committee for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *Preprint*, arXiv:2303.17466.

Tanise Ceron, Neele Falk, Ana Baric, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms. *CoRR*, abs/2402.17649.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. *arXiv preprint arXiv:2302.03269*.

Hyundong Cho, Thamme Gowda, Yuyang Huang, Zixun Lu, Tianli Tong, and Jonathan May. 2024. BotEval: Facilitating interactive human evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 107–116, Bangkok, Thailand. Association for Computational Linguistics.

Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrera, et al. 2023. Can language model moderators improve the health of online discourse? *arXiv preprint arXiv:2311.10781*.

Cohere. 2024. Command r: Retrieval-augmented generation at production scale. https://cohere.com/blog/command-r. Last accessed: August 30, 2024.

Jan de Wit. 2023. Leveraging large language models as simulated users for initial, low-cost evaluations of designed conversations. In *International Workshop on Chatbot Research and Design*, pages 77–93. Springer.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2023. Detecting text formality: A study of text classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 274–284, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and Akhil Mathur. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Oluwole Fagbohun, Rachel M Harrison, and Anton Dereventsov. 2024. An empirical categorization of prompting techniques for large language models: A practitioner's guide. *arXiv preprint arXiv:2402.14837*.

Rudolf Flesch. 1979. *How to write plain English : a book for lawyers and consumers*. Harper & Row. Includes index.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Tear Gosling, Alpin Dale, and Yinhe Zheng. 2023. Pippa: A partially synthetic conversational dataset. *Preprint*, arXiv:2308.05884.

George Gui and Olivier Toubia. 2023. The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *Preprint*, arXiv:2301.01768.

Somayeh Jafaritazehjani, Gwénolé Lecorvé, Damien Lolive, and John D Kelleher. 2021. Style as sentiment versus style as formality: The same or different? In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*, pages 487–499. Springer.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024b. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*.

Callie Y. Kim, Christine P. Lee, and Bilge Mutlu. 2024. Understanding large-language model (llm)-powered human-robot interaction. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yan Leng and Yuan Yuan. 2023. Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *Preprint*, arXiv:2304.03439.

Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2025a. Interactive evaluation for medical LLMs via task-oriented dialogue system. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, Abu Dhabi, UAE. Association for Computational Linguistics.

Siyang Liu, Trish Maturi, Bowen Yi, Siqi Shen, and Rada Mihalcea. 2024a. The generation gap:exploring age bias in the underlying value systems of large language models. *Preprint*, arXiv:2404.08760.

Yaoyang Liu, Zhen Zheng, Feng Zhang, Jincheng Feng, Yiyang Fu, Jidong Zhai, Bingsheng He, Xiao Zhang, and Xiaoyong Du. 2025b. A comprehensive taxonomy of prompt engineering techniques for large language models. *Frontiers of Computer Science (2025). doi*, 10.

Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2024b. Make llm a testing expert: Bringing human-like interaction to mobile gui testing via functionality-aware decisions. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Dongxu Lu, Johan Jeuring, and Albert Gatt. 2025. Evaluating llm-generated versus human-authored responses in role-play dialogues. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 20–40.

Petter Mæhlum, David Samuel, Rebecka Maria Norman, Elma Jelin, Øyvind Andresen Bjertnæs, Lilja Øvrelid, and Erik Velldal. 2024. It's difficult to be neutral – human and LLM-based sentiment annotation of patient comments. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 8–19, Torino, Italia. ELRA and ICCL.

Eric Mayor, Lucas M Bietti, and Adrian Bangerter. 2025. Can large language models simulate spoken human conversations? *Cognitive Science*, 49(9):e70106.

Mistral. 2024. Au large. `https://mistral.ai/news/mistral-large/`. Last accessed: August 30, 2024.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, and Fabrice Lefèvre. 2024a. Role-play zero-shot prompting with large language models for open-domain human-machine conversation. *arXiv preprint arXiv:2406.18460*.

Ahmed Njifenjou, Virgile Sucal, Bassam Jabaian, and Fabrice Lefèvre. 2024b. Role-play zero-shot prompting with large language models for open-domain human-machine conversation. *Preprint*, arXiv:2406.18460.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. Association for Computational Linguistics.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: Liwc [computer software]. *Austin, TX: liwc. net*, 135.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Giulio Rossetti, Massimo Stella, Rémy Cazabet, Katherine Abramski, Erica Cau, Salvatore Citraro, Andrea Failla, Riccardo Improta, Virginia Morini, and Valentina Pansanella. 2024. Y social: an llm-powered social media digital twin. *arXiv preprint arXiv:2408.00818*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Association for Computational Linguistics*, Bangkok, Thailand.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. Teler: A general taxonomy of llm prompts for benchmarking complex tasks. *arXiv preprint arXiv:2305.11430*.

Rupak Sarkar, Bahareh Sarrafzadeh, Nirupama Chandrasekaran, Nagu Rangan, Philip Resnik, Longqi Yang, and Sujay Kumar Jauhar. 2025. Conversational user-ai intervention: A study on prompt rewriting for improved llm response generation. *arXiv preprint arXiv:2503.16789*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Joao Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2019. Chateval: A tool for chatbot evaluation. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 60–65.

Ivan Sekulic, Silvia Terragni, Victor Guimarães, Nghia Khau, Bruna Guedes, Modestas Filipavicius, Andre Ferreira Manso, and Roland Mathis. 2024. Reliable LLM-based user simulator for task-oriented dialogue systems. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pages 19–35, St. Julians, Malta. Association for Computational Linguistics.

Rafael A. Rivera Soto, Olivia Miano, Juanita Ordonez, Barry Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *EMNLP*.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting gpt-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*.

Ekaterina Svikhnushina and Pearl Pu. 2023. Approximating online human evaluation of social chatbots with prompting. *arXiv preprint arXiv:2304.05253*.

Hovhannes Tamoyan, Hendrik Schuff, and Iryna Gurevych. 2024. Llm roleplay: Simulating human-chatbot interaction. *Preprint*, arXiv:2407.03974.

Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *Preprint*, arXiv:2311.14096.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024a. Large language models cannot replace human participants because they cannot portray identity groups. *Preprint*, arXiv:2402.01908.

Xingguang Wang, Xuxin Cheng, Juntong Song, Tong Zhang, and Cheng Niu. 2024b. Enhancing dialogue state tracking models through LLM-backed user-agents simulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8724–8741, Bangkok, Thailand. Association for Computational Linguistics.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Preprint*, arXiv:2212.09196.

Jiaqi Wen, Bogdan Gabrys, and Katarzyna Musial. 2024. Towards digital twin-oriented complex networked systems: Introducing heterogeneous node features and interaction rules. *Plos one*, 19(1):e0296426.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models. *arXiv preprint arXiv:2304.11082*.

Dustin Wright, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. 2024. LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2023. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

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

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian. *Preprint*, arXiv:2309.10931.

A Model Inference Details

Experiments are conducted on 8 NVIDIA RTX A6000 GPUs and 4 A100-SXM4-80GB GPUs using vLLM 0.5.4 (Kwon et al., 2023), Hugging Face Transformers 4.43.3 (Wolf et al., 2020) and PyTorch 2.4.0 (Paszke et al., 2019) on a CUDA 12.4 environment.

To ensure reproducibility, we set all random seeds in Python to be 1000, including PyTorch

and NumPy. When doing model inference, we use temperature = 0.8, top_p = 0.95, and max_tokens = 1024.

B Annotation

We annotated 1,273 examples randomly sampled from the 2,000 examples in §5. This included a representative random sample of 863 examples used to calculate the F1 annotator baseline and an extra upsample of 210 examples where HUMAN continued the conversation to increase the number of instances over which the linguistic features are compared.

Task Annotators are given the first turn of a dialogue between a HUMAN and the instructions from the top prompt. Annotators have to answer two questions: 1) whether the HUMAN will continue or end the conversation and 2) if the HUMAN continues the conversation, how they predict the HUMAN will respond. Annotations are conducted using POTATO (Pei et al., 2022). The annotation interface is pictured in Figure 14. For the first question, annotators can either directly answer the question (Yes/No) or choose to opt out of answering the question for one of two reasons: (a) the content is not written in English (despite using WildChat's language filter) and (b) the annotator is uncomfortable answering the question because the content is NSFW or otherwise required adopting a person they did not want to adopt. Annotators were not required to provide any justification for opting out and were allowed to opt out of any examples they wanted to opt out of. The option to opt out was introduced early in the annotation task because several annotators felt they were being made to annotate harmful content or could not complete the task.

Sample We annotated 1,273 examples randomly sampled from the 2,000 examples in §5. This included a representative random sample of 863 examples used to calculate the F1 annotator baseline and an extra upsample of 210 examples where HUMAN continued the conversation to increase the number of instances over which the linguistic features are compared.

Output The annotation team consisted of 12 authors, including 11 university students and one faculty member. Of the 1,273 annotations, annotators selected "Yes" for 546 samples (43%) when they could directly answer the question and "No" for 542 samples (43%) when they could not. Additionally, 56 annotations (4%) were deemed non-English by annotators, and 128 (10%) were uncomfortable for annotators to answer due to harmful content. Neither of these two categories were considered in the annotator baseline for the experiment in §5. There were 293 cases where both the human annotator and HUMAN chose to continue the conversation, and this was the sample used to calculate the linguistic metrics for §5.

C Metrics

In order to evaluate whether a SIMULATOR generates HUMAN-like text in the simulated turn 3, we assess whether it reproduces basic linguistic features of the HUMAN 's response. To perform this evaluation, we selected a set of metrics that are widely used by the NLP community to compare features of the HUMAN's true turn 3 against the SIMULATOR's predicted turn 3. While several potential metrics were not included in this study, we aimed to capture key properties that have been well established, widely used in the literature, and appropriate to use with the WildChat data. In this section, we briefly summarize the chosen metrics.

Lexical metrics included several properties of the individual words in each dialogue turn, including the number and average length of words. Our analysis also includes perplexity as a key metric used in LLMs to evaluate token-level similarity in their outputs (Hu and Zhou 2024). We omitted established metrics like ROUGE, BLEU, and METEOR that also calculate token-level similarity, because most dialogue turns were not sufficiently long for these metrics to be robustly applied. Lastly, since LLMs are known to write more polished text than humans, we also measured the frequency of typographical errors.

For **syntax metrics**, dependency parsing is the canonical method to encode what entities a speaker is referencing and how they relate these entities to one another (Jurafsky and Martin 2019). We include several key attributes from dependency parsing the utterance.

Semantic metrics capture the meaning of an utterance using distributional and content-based approaches. For instance, SBERT embeddings is a popular way of using distributional semantics to quantify whether two documents have similar meaning (Reimers and Gurevych 2019). We also use two established methods, LIWC and a prompt

classification tool, to measure the content of each dialogue turn.

Stylistic properties are defined as being attributes of the text that are largely independent of its content (e.g., a response may have a positive or negative sentiment towards the same topic) (Jafaritazehjani et al 2021). While there has been some debate about which measures constitute style, we primarily looked for attributes that are largely independent of the response's content and commonly referenced in the NLP literature as measures of style. As with semantics, we quantify style using both a popular distributional approach, LUAR, as well as specific properties of style like sentiment, politeness, and toxicity that are highly studied in many settings (Troiano et al. 2022). Again, since LLMs are known to write more polished text than humans, we also measured the readability as well as the frequency of stylistic properties like capitalization and punctuation that humans often omit in less formal settings. Finally, LLMs are often instruction fine-tuned to produce more neutral-sounding responses, we also compare the utterances' subjectivity.

Our framework is intended as a comprehensive initial set of metrics to use in evaluation; future work can explore conducting evaluations with additional metrics. Additionally, there is a wide range of responses a human may give in turn 3, so misalignment on these metrics does not necessarily mean that the SIMULATOR is not producing human-like text. For instance, in Figure 1, the HUMAN asks "Calculate the probability that 10 randomly selected adults have blood pressure greater than 150" and the CHATBOT responds with an explanation of how to perform this calculation. There is not one right answer for what a human may say next; some humans may ask the chatbot to clarify a portion of their response, while others (like the HUMAN in the example) may ask a different but related question or end the chat. However, it is very unlikely that a human would respond, as a SIMULATOR did, by saying, "As I sit here pondering the questions and calculations, I feel a sense of wonder at the intricacies of the human experience." Therefore, as described in Section 5, we introduce a human baseline where annotators perform the same turn 3 simulation task that we are asking LLMs to do. This human baseline can help quantify the range of "acceptable" deviation from the given HUMAN that is present in human-like speech. We then test whether, using the broad set of models and prompts we tested in this paper, SIMULATORs fall within or outside that range.

D Broader Range of Models

Our analysis showed that there is relatively limited variation in SIMULATOR performance across different types of models. However, our analysis was restricted to entirely open-weight and mostly Instruct versions of models. To assess whether larger, proprietary models and models that are fine-tuned with RLHF could achieve higher performance on this task. We report average results from the three best-performing prompts selected for Experiments 2-3.

Proprietary Models It is crucial that the models we pick be appropriate and feasible for large scale simulation studies. We selected two proprietary models, GPT-5-mini and Gemini-2.5-Flash, which are available at a price-point and with sufficiently high API rate limits that makes simulation feasible. Some larger models like GPT-4o, Claude 3.5 Sonnet, and Gemini-3.5 are extremely costly or have very restrictive API rate limits, and therefore we're excluded from the study. For instance, in our study, it would cost $500 to replicate experiment 1 (2000 conversation seeds, 46 prompts) using GPT-5. By contrast, the models we selected cost under $100 for this experiment. Table 2 shows how the performance of GPT-5-mini and Gemini-2.5-flash compare as SIMULATORs to two of the best LLMs from the set we reported in the main paper. In all cases, the performance from these larger, proprietary models was statistically indistinguishable from the open-weight models.

Base Models We compared base vs. instruct models for three model types: Llama-3.1-70B, Llama-3.1-8B, and Qwen2-72B. Unlike instruct models, base models are not fine-tuned using RLHF or instruction fine-tuning to achieve any particular style. Therefore, base versions of models may better reproduce the briefer, less formal text produced by HUMANs in Turn 3. Table 3 shows how the performance of base models compare to the corresponding instruct models as SIMULATORs. We find that the base models tend to have higher performance in knowing when to end a conversation but have lower or comparable performance in style and other linguistic dimensions. This result suggests that safety alignment/RLHF may not be a major factor limiting a model's ability to generate realistic human speech.

These results underscore our finding that there is not much variation in simulation performance across different types of models.

E Evaluation of Turns 3 through 11

The goal of this analysis is to determine the performance of SIMULATORS in determining what a HUMAN said at a more general Turn n of the conversation rather than only Turn 3. We tested this for five values of n, Turns 3, 5, 7, 9, and 11. For each value of n, we randomly sampled 1,000 Wildchat conversations where the HUMAN did not respond at Turn n (i.e., conversations with exactly $n-1$ turns) and 2,000 conversations where the HUMAN did respond at Turn n (i.e., conversations with n or more turns). We adapt the setup from §4. We gave each SIMULATOR the first $n-1$ turns of the conversation and asked it to predict whether the conversation would continue and, if so, what the HUMAN would say at the n^{th} turn. We used the three best-performing prompts and models, as described in §6. We slightly modified the prompts to accommodate more than two turns of conversational context. For each prompt, we generated each SIMULATOR's predicted Turn n and compared it to the HUMAN Turn n on each of the 21 metrics in Table 1 using the correlation approach described in §5. For the lexical, syntactic, semantic, and style measures, we report the average correlation over all metrics in that category. For the conversation ending, we report the binary F1 score between the HUMAN and SIMULATOR conversation-ending behaviors. Figure 8 shows the category-wise similarity for each turn. Error bars represent 95% bootstrap confidence intervals over all metrics, models, and prompts. There are no significant differences across turns, suggesting that additional conversational context does not improve the performance of the SIMULATOR.

F Prompt Type Analysis

The 46 prompts generated for this study used a diverse set of strategies to ask LLMs to act as SIMULATORS. In this analysis, we quantify and test the effects of the type of prompt on the SIMULATOR performance.

Prompt Typology First, we developed a prompt typology, by adapting existing literature on prompt components and techniques (Santu and Feng, 2023; Schulhoff et al., 2024; Fagbohun et al., 2024; Liu et al., 2025b). Our taxonomy includes six key elements:

1. **'No Response' prompting:** whether the LLM was explicitly prompted to consider the possibility of no response at turn 3 (e.g., 'Will you respond? If so, what will you say next?' which explicitly introduces the possibility of not continuing the conversation at turn 3 vs. 'What will you say next?' which does not explicitly introduce this possibility).
2. **Role:** whether the prompt assigned a clear role to the LLM and, if so, what role was assigned. Roles were categorized using inductive coding after reviewing all of the prompts, and included directly assigning the persona of a "human interacting with ChatGPT," assigning another type of human persona, and assigning the persona of an "LLM simulating a human."
3. **Expression:** whether the prompt to generate a simulated turn 3 was expressed as an instruction (e.g., 'Print what you think the human would say next.') or a question (e.g., 'What do you think the human would say next?').
4. **Prompt hacking:** whether the prompt uses techniques like injection (direct prompts to ignore safety training/instruction finetuning) and jailbreaking (adversarial prompts to bypass safety training/instruction finetuning), usually to elicit a more "natural" and less stilted style.
5. **Style prompts:** whether the prompt explicitly instructed the LLM to write in a particular style, including mimicking the writing style of the HUMAN in turn 1, more generically prompting the LLM to "write in human-like way," asking the LLM to be "brief" in its simulated turn 3, and other specific style-related instructions. These style codes were developed through a combination of deductive and inductive coding, starting by asking authors to describe their prompting strategies and adding additional strategies the authors noticed after reviewing all of the prompts.
6. **Prompting techniques:** whether a specific set of prompting techniques was used, including decomposing the prompt into multiple steps (decomposition), asking the LLM to include reasoning (thought generation), providing context or a scenario to ground the task (context), providing a clear objective for the LLM to evaluate its output (task objectives), and asking the model to consider the HUMAN's intent in turn 1. These techniques were initially enumerated from prior

papers (Santu and Feng, 2023; Schulhoff et al., 2024; Fagbohun et al., 2024; Liu et al., 2025b) and included in this analysis if multiple prompts in the paper used them.

Due to the structure of the task, certain features of the prompt were fixed in advance and do not vary across the prompts in the study. For instance, all of the prompts in this study are zero-shot, single-turn prompts that include the HUMAN turn 1 and CHATBOT turn 2 data but do not include any examples. Future research could examine a larger space of prompts outside of these constraints, including ensembling techniques and in-context learning.

Annotation Two authors used the prompt typology to collaboratively annotate all 46 SIMULATOR prompts in this study, including discussing unclear cases. Table 4 shows the distribution of prompt types and Figure 9 show the correlation among different types of prompts. Our study includes a diverse range of prompts, with all types of prompts well-represented.

Regression Analysis We ran a linear regression to test the effects of the prompt characteristics on the SIMULATOR's performance in generating HUMAN-like turn 3. To maximize the diversity in prompts and models, the data for this regression came from the first experiment, where we generated HUMAN turn 3 for 2,000 dialogues across 9 models and 46 prompts. We calculated each metric's correlation similarity (§5) for each prompt/model pair, giving a total of 8694 similarity scores (since there are 9 models, 46 prompts, and 21 metrics in total).

The dependent variable in this regression is the performance of a prompt, measured by the correlation similarity. Independent variables are all of the elements in the prompt typology. Since the correlation is influenced by model and metric, we adjust for these as fixed effects in the regression. We fit separate regressions for each of the five categories of metrics (conversation ending, lexical, syntatic, etc.), so the model averages over all metrics in a category and predictions can be interpreted as category-wise similarity scores. For ease of interpretation, we report the marginal predicted category-wise similarity scores for each type of prompt.

Although many of the prompt characteristics are correlated (Figure 9), the variance inflation factors of the regression are low (<5), suggesting that multicollinearity is not a concern. 95% confidence intervals are adjusted for multiple comparisons using a Bonferroni correction.

Results Overall, we find that prompt characteristics are significantly associated with SIMULATOR performance (Figure 10-11). These results only pertain to the variation in the 46 prompts used in this study – in particular, we did not explore the entire space of prompts or systematically generate prompts to maximize variation within our typology. Therefore, the results can only be interpreted in the context of the 46 author-created prompts used in this study.

First, the only factor influencing how well SIMULATORs predicted conversations ending was whether the prompt explicitly instructed the SIMULATOR to consider "no response" as a possibility (Figure 10a). Notably, most prompts did include an explicit "no response" condition and the average predicted F1 score among prompts with this condition (0.11, 95% CI = [0.09, 0.13]) is well below random chance and the annotator baseline. The fact that this metric was not sensitive to other forms of prompt design suggests that LLMs are not easily influenced to end conversations.

Second, the SIMULATOR's lexical and syntactic similarity to HUMAN turn 3 also improved when the prompt included an explicit "no response" instruction (Figure 10a). This result is interesting because these two categories are only assessed when both the HUMAN and the SIMULATOR respond in turn 3, suggesting that there is a connection between including the "no response" instruction in the prompt and the quality of responses – even in cases when the HUMAN actually responds. It is possible that the "no response" instruction influences the SIMULATOR to be more strongly grounded in some representation of the HUMAN's conversational intent.

Third, the content of the prompt had a significant effect on SIMULATOR performance (Figure 10b-c). Prompts that assigned SIMULATORs the role of "human talking to ChatGPT" tended to produce more human-like response text across many categories than SIMULATORs adopting other personas or not explicitly assigned a role. Additionally, using questions, rather than instructions, to prompt SIMULATOR was associated with improvements in model performance. These results suggest that SIMULATORs are sensitive to the inclusion and exclusion of certain prompt components.

Fourth, apart from the content of the prompt, using specific prompting techniques was rarely associated with improvements in performance over simpler prompts. For instance, SIMULATORs using injection or jail-breaking tended to have similar or worse performance than those that did not use prompt hacking techniques (Figure 11a). Similarly, SIMULATORs using no prompting techniques tended to have higher similarity to HUMAN text than SIMULATORs using techniques like decomposition, context, task objectives, etc (Figure 11c). Additionally, prompting the SIMULATOR to emulate a certain style rarely has any effect on performance and is sometimes associated with worse performance than not including a style prompt (Figure 11b). However, some style instructions are associated with worse outputs than others. For instance, SIMULATORs given specific instructions on what style to use (e.g., "mimic the HUMAN's turn 1 style", "be brief") tend to output more HUMAN-like text than SIMULATORs that are generically prompted to output human-like text. Our results suggest that these sorts of prompting techniques add complexity to prompts but generally do not translate to improvements in SIMULATOR performance.

Overall, these results suggest that it may be possible to identify characteristics of prompts that optimize SIMULATOR performance. Future work can build on our paper to identify these characteristics and build towards a prompt optimization framework for SIMULATORs, including more robustly testing different types of roles and personas, instruction formats, and more thoroughly exploring the space of prompting techniques. However, our results also suggest that future work may be well-served by exploring simpler prompts rather than increasingly complex prompting frameworks.

G Regression Details for §7

Dependent Variable The dependent variable in the regression is the overall similarity between the HUMAN and SIMULATOR's Turn 3 averaged across all measures $m \in M$ in Table 1. To calculate the similarity for each conversation, we first apply each measure m to pairs of HUMAN and SIMULATOR utterances for Turn 3, obtaining pairs (h_m, l_m), and take one minus the distance between the pair. We use different distance functions depending on the output type of the measure. If m outputs a scalar, we use the absolute difference between h_m and l_m. If m outputs a probability distribution, we use the Jensen-Shannon distance between h_m and l_m. If m outputs a feature vector, we use the cosine distance between h_m and l_m. Using this, we obtain a vector of similarity scores $\mathbf{s}_m$ over all Turn 3 pairs. Additionally, to bring the metrics into approximately the same scale in order to be comparable and aggregate overall similarity across metrics, we log-scale the scalar metrics with an unbounded range which empirically demonstrates heavy-tailedness[4], followed by z-scoring each $\mathbf{s}_m$. We then average the similarity scores $\mathbf{s}_m$ for all measures within each category to build four dependent variables corresponding to lexical, syntactic, semantic, and style similarity.

Independent Variables Our regression uses contextual features of the conversations obtained from the conversation metadata, the simulation metadata, and the conversation history (i.e., HUMAN message at Turn 1 and the LLM response at Turn 2). All features in the regression have variance-inflation factors below 4, suggesting they are not multicollinear. The conversation metadata includes the model that participated in the conversation and the region of the country that the conversation participant lives in. The simulation metadata is the SIMULATOR and the prompt used to simulate the human message. The conversation history is represented by a subset of scalar metrics that we used in §5. Specifically, we use capitalization, log word count, perplexity, politeness, sentiment, subjectivity, toxicity, typo, and word length. Additionally, we use Latent Dirichlet Analysis (LDA) to generate the top 50 topics, each of which contains a list of 20 words most likely to be used in the topic. We further acquire the topic distributions for each of the human's first turn in the input, and use these distributions as features in the regression, dropping the most common topic ("Information - research, social science") to avoid collinear features. Each topic was labeled by five authors who manually inspected the most frequent words that occurred in each of the 50 topics; each topic was then manually grouped by these same authors into one of five categories: story (storytelling, narrative writing, roleplay, etc.), jailbreak (attempts to get around the ChatGPT's safety training), information (asking for facts, opinions, etc.), programming (help with writing code), and other.

[4]This includes utterance length, average word length, perplexity, dependency tree depth, dependency tree breadth, and dependency tree distance.

Regression coefficients are given in Table 13 and Table 14. p-values are corrected for multiple comparisons using a Bonferroni correction.

H Supplemental Results

Model and Prompt Variation Our analyses show that the choice of prompt instructions significantly impacts the quality of simulations, often more so than the family or even the size of the SIMULATOR model. In this section, we conduct two analyses to reinforce these findings.

Figure 13 compares the performance of the best and worst model (i.e., highest vs. lowest average score) and the performance of the best and worst prompt. On average, across all models, the best prompt outperforms the worst prompt on all metrics except for conversation ending. However, the best model is either statistically indistinguishable from the worst model or the difference between these two models has a smaller effect size than the difference between the best and worst prompts. Table 5 compares the performance of the "best" prompt + "worst" model (i.e., highest average score for prompt and lowest average score for model) against the "worst" prompt + "best" model. The best prompt/worst model condition outperforms the worst prompt/best model condition in almost all metrics, further suggesting prompting technique is more important than model choice.

Prompt Selection for §6 and §7 In order to select which three prompts were used in §6 and §7, two authors manually classified prompts into each category, and we selected one prompt per category. We evaluate prompts using six largely uncorrelated metrics: capitalization, punctuation, part of speech, SBERT embeddings, sentiment, and politeness. Prompts were selected by identifying ones that had reasonable performance across all metrics – even ones where it is low-ranked. Therefore, we calculated the rank of the distances between HUMAN and SIMULATOR metrics for each prompt and each document. We selected one prompt per category with the highest 75th percentile ranking, which tended to be prompts with high median rank and low variance across metrics (Figure 15). The three selected prompts are shown in Table 7.

Conversation End Prediction In addition to F1 scores, we evaluate how often each SIMULATOR predicts that a conversation will end as a function of whether the HUMAN actually ended the conversation in Figure 17. The same comparison is performed across models in Figure 18. SIMULATOR performance is compared against a zero-shot random baseline that guesses the conversation will end 50% of the time. Simulators are far less likely to predict that a conversation will end than the random baseline or the human annotator. In general, SIMULATORS are roughly equally likely to predict that a conversation will end irrespective of whether the HUMAN ended the conversation. This is true across models and languages. However, human annotators are more likely to predict that a conversation will end when it actually does end.

Multilingual Prompts and Models The choice of model and prompting strategy has a strong effect on the F1 score across languages, as shown in Figure 16b. As discussed in §5, these results are crucial for understanding when LLMs are effective SIMULATORS.

As in §5, the differences are largely driven by how often the model or prompt predicts conversation endings. For instance, in Chinese and Russian, LLAMA3.1-70Boutperforms other models on this task by predicting 10% and 12% conversation ending, respectively. Similarly, the COT prompt predicts conversation endings most frequently (13% for English, 6% for Chinese, 16% for Russian), while the OVERRIDE prompt predicts them least frequently (1% for English, 0.5% for Chinese, and 3% for Russian). In this analysis, we compare three prompts, one from each category. In all languages, the OVERRIDE prompt results in lower F1 scores, while the COT prompt yields higher F1 scores. Although the results from one prompt of each variety cannot be generalized to all COT and OVERRIDE prompts, future work may examine whether these differences are attributable to the structure of these prompts. For instance, the COT prompt explicitly asks the model to reason whether HUMAN achieved the goal satisfactorily, which may lead the model to prioritize the decision to end the conversation. In contrast, the OVERRIDE prompt, which tricks the model into performing the task, does not explicitly prompt the model to end the conversation as often.

Interestingly, these results contrast with the performance of prompts in matching the properties of the response text. In many metrics, the COT performs worse than the DIRECT or OVERRIDE prompts. Again, future research may explore whether these differences generalize to a broader class of COT prompts. Perhaps the more structured

format of the COT prompt may be less suited to capturing the nuances of open-ended human speech, even though it might be better for closed-ended tasks.

The correlations for each individual metric, by prompt and model, are given in Table 11 and Table 12.

Category	Measure	Description
End	F1	Comparison of how often the SIMULATOR ends the conversation when the human ends it.
Lexical	Utterance Length^{s}	Log-transformed number of words.
	Average Word Length^{s}	Log-transformed number of characters per word.
	Perplexity^{s}	Log-transformed perplexity of the utterance, calculated using `lmppl` and GPT-2 as model.
	Typographical Errors^{s}	Fraction of words that have typographical errors, counted using `pyspellchecker`.
Syntactic	Part of Speech^{d}	Distribution of the utterance's part of speech tags from `spaCy`.
	Dependency Tree Depth^{s}	Log-transformed depth of the dependency tree from `spaCy`.
	Tree Breadth^{s}	Log-transformed number of leaf nodes.
	Tree Dependency Distance^{s}	Log-transformed average distance between dependents.
Semantic[2]	SBERT^{v}	Utterance embeddings from the `all-MiniLM-L6-v2` language model (Reimers and Gurevych, 2019). Embeddings are commonly interpreted as the meaning or sense of the document (Camacho-Collados and Pilehvar, 2020).
	LIWC^{d}	Distribution of 69 LIWC categories from Pennebaker et al. (2007).
	Prompt Type^{d}	Distribution of categories from prompt classification tool `valpy/prompt-classification`.
Style	Punctuation^{d}	Distribution of punctuation characters.
	Capitalization^{s}	Fraction of letters that are capitalized.
	Sentiment^{s}	Distribution of positive, neutral, and negative sentiment from `distilbert-base-multilingual-cased-sentiments-student`. We take the interpretation that sentiment is a measure of style, as it speaks to the valence rather than the content of speech (Jafaritazehjani et al., 2021).
	Politeness^{s}	From `Genius1237/xlm-roberta-large-tydip` (Srinivasan and Choi, 2022).
	Formality^{s}	From `s-nlp/mdeberta-base-formality-ranker` (Dementieva et al., 2023).
	Toxicity^{s}	Toxicity of tone and content, as judged by annotators `s-nlp/roberta_toxicity_classifier`.
	Readability^{s}	Distribution of Flesch reading ease scores (Flesch, 1979).
	Subjectivity^{s}	The average subjectivity score of words in the utterance from the sentiment polarity lexicon in `textblob`.
	LUAR^{v}	Author style embeddings using `rrivera1849/LUAR-CRUD` (Soto et al., 2021).

Table 1: Measures used to evaluate how well LLMs capture properties of human responses at Turn 3 of a conversation. Letter superscript indicates whether the difference between human and SIMULATOR measurements are **(s)** scalar values (compared with l1-distance), **(d)** probability distributions (compared with Jensen-Shannon divergence), or **(v)** vector embeddings (compared with cosine distance).

Model	Conv End	Lexical	Semantic	Style	Syntactic
gpt-5-mini	0.15	0.16	0.19	0.12	0.12
	[-0.12,0.41]	[0.13,0.2]	[0.12,0.25]	[0.08,0.16]	[0.1,0.14]
gemini-2.5-flash	0.2	0.09	0.13	0.09	0.1
	[-0.11,0.51]	[0.05,0.13]	[0.08,0.18]	[0.07,0.12]	[0.08,0.13]
Meta-Llama-3.1-70B-Instruct	0.12	0.16	0.17	0.11	0.1
	[-0.0,0.24]	[0.11,0.22]	[0.1,0.24]	[0.07,0.14]	[0.08,0.13]
Phi-3-medium-4k-instruct	0.22	0.12	0.16	0.11	0.1
	[0.06,0.37]	[0.09,0.19]	[0.1,0.23]	[0.07,0.14]	[0.07,0.13]

Table 2: Comparing average correlations and 95% confidence intervals from proprietary models (top two rows) against the best-performing open-weight models tested in our study (bottom two rows). In all cases, the performance from these larger, proprietary models was statistically indistinguishable from the open-weight models.

Model	Conv End	Lexical	Semantic	Style	Syntactic
Llama-3.1-70B (**base**)	0.18	0.02	0.05	0.07	0.04
	[0.05,0.32]	[-0.0,0.05]	[0.03,0.07]	[0.03,0.1]	[0.02,0.05]
Llama-3.1-70B-**Instruct**	0.12	0.16	0.17	0.11	0.1
	[-0.0,0.24]	[0.11,0.22]	[0.1,0.24]	[0.07,0.14]	[0.08,0.13]
Llama-3.1-8B (**base**)	0.23	0.01	0.05	0.04	0.02
	[0.01,0.45]	[-0.02,0.04]	[0.03,0.07]	[0.02,0.07]	[0.01,0.03]
Llama-3.1-8B-**Instruct**	0.05	0.14	0.17	0.09	0.1
	[-0.02,0.12]	[0.1,0.19]	[0.1,0.23]	[0.07,0.12]	[0.08,0.13]
Qwen2-72B (**base**)	0.38	0.07	0.14	0.08	0.06
	[0.17,0.59]	[0.02,0.12]	[0.08,0.19]	[0.05,0.12]	[0.02,0.09]
Qwen2-72B-**Instruct**	0.31	0.14	0.14	0.07	0.07
	[0.06,0.55]	[0.08,0.21]	[0.08,0.19]	[0.05,0.1]	[0.05,0.1]

Table 3: Comparing average correlations and 95% confidence intervals from base vs. instruct versions of models. In all cases, the performance from these larger, proprietary models was statistically indistinguishable from the open-weight models.

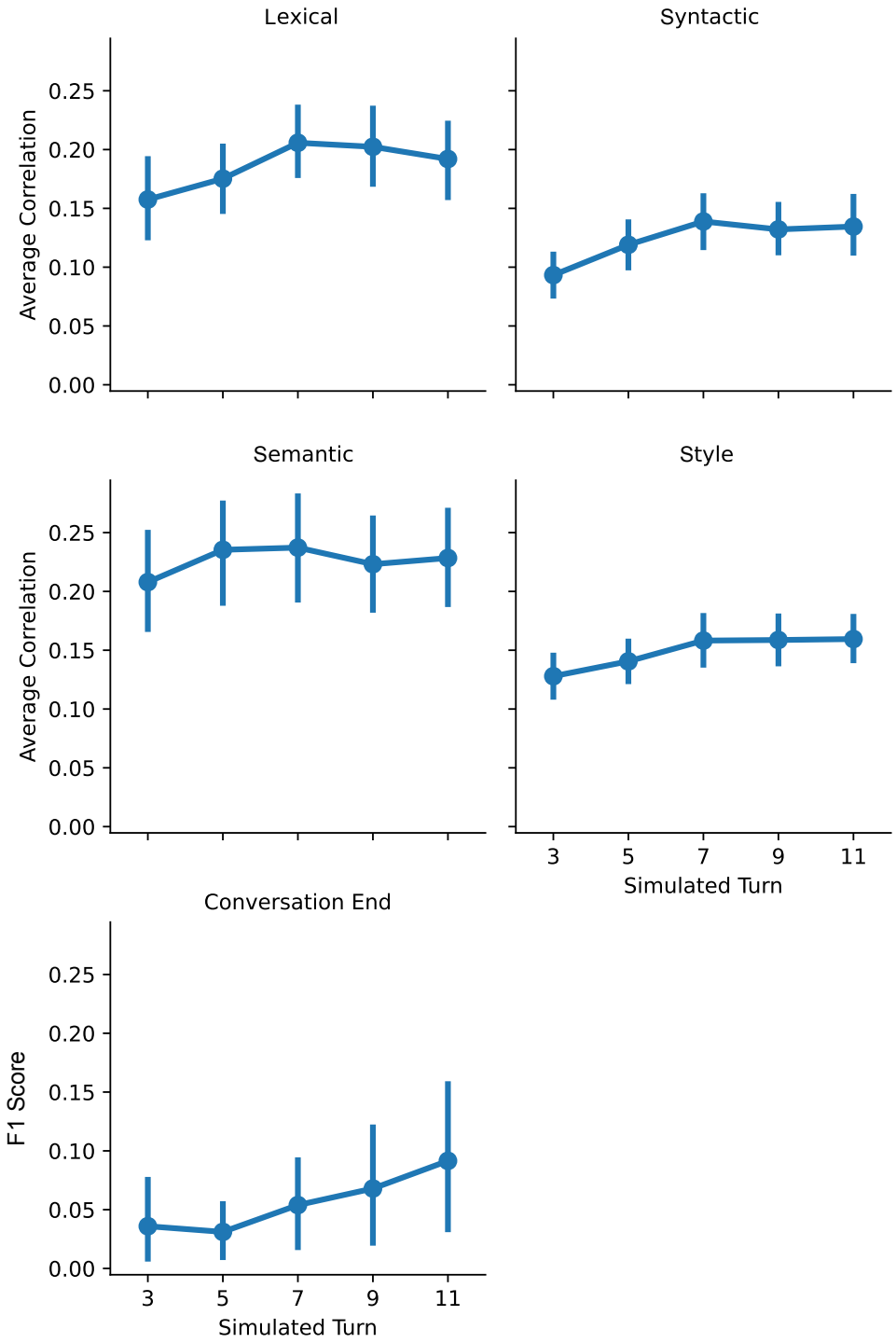


Figure 8: How well do SIMULATORs match properties of HUMAN responses at Turn n of the conversation? We show the similarity between HUMAN and SIMULATOR responses at Turn $n \in \{3, 5, 7, 9, 11\}$ of the conversation, across the five major categories of metrics in Table 1. Simulators do not perform better at later turns of the conversation, suggesting that additional conversational context does not improve their performance.

	Characteristic	# Prompts
'No Response' Prompt	Yes	35
	No	11
Role	Human Interacting with ChatGPT	18
	LLM Simulating Human	10
	Other Human	6
	None	12
Expression	Instruction	37
	Question	9
Prompt Hacking	Injection	8
	Jailbreaking	7
	None	31
Style Instruction	Mimic Turn 1	20
(not mutually exclusive)	Human-Like	17
	Short	11
	Other	15
	None	13
Prompting Techniques	Decomposition	19
(not mutually exclusive)	Thought Generation	10
	Context	29
	Understand HUMAN Intent	23
	Provide Task Objectives	13
	None	8

Table 4: The types of prompts used for the SIMULATOR. The 46 prompts used in this study were manually annotated using the typology from Appendix F.

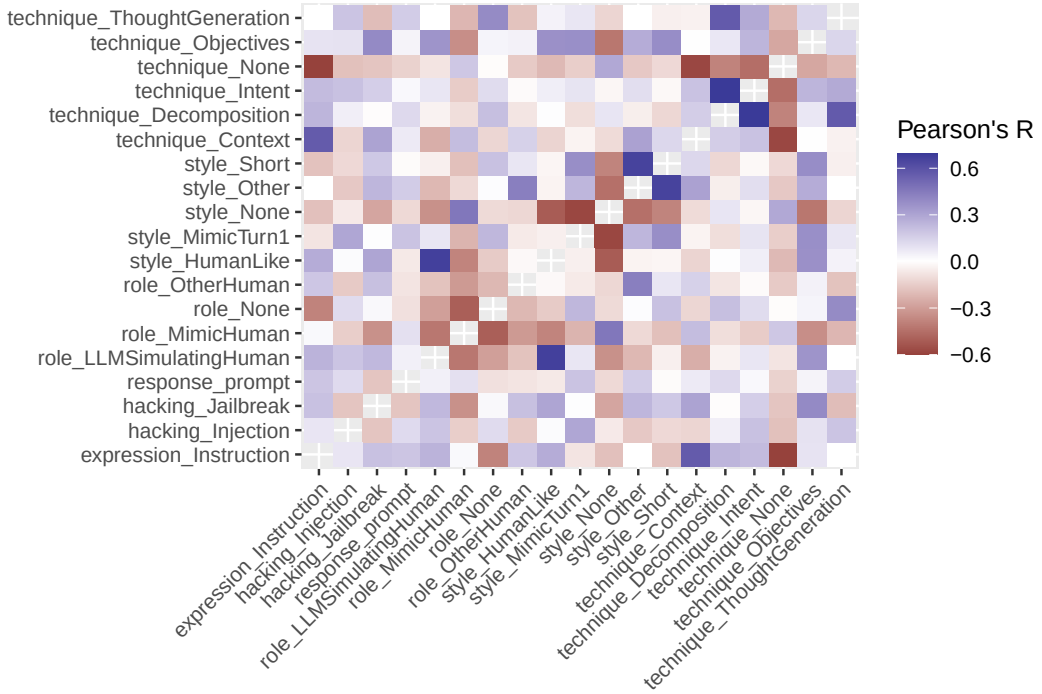
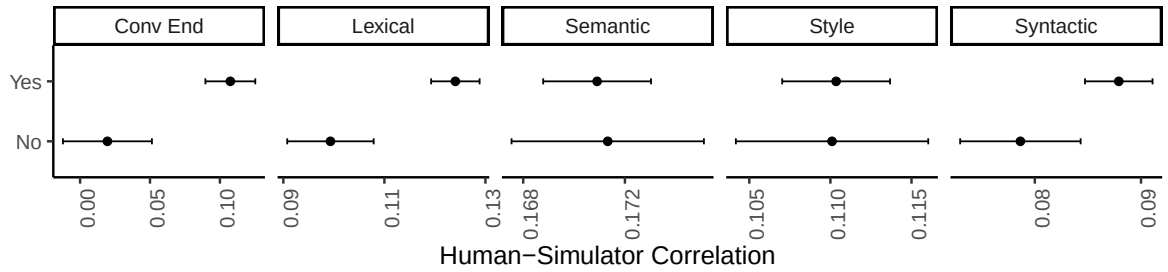
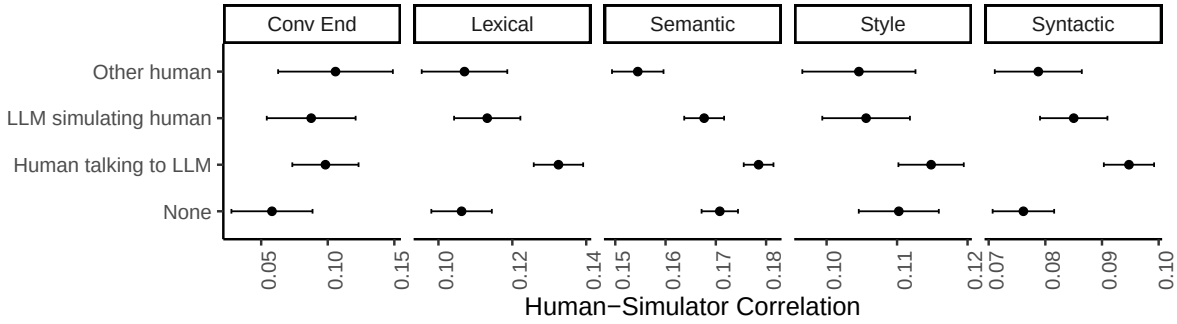


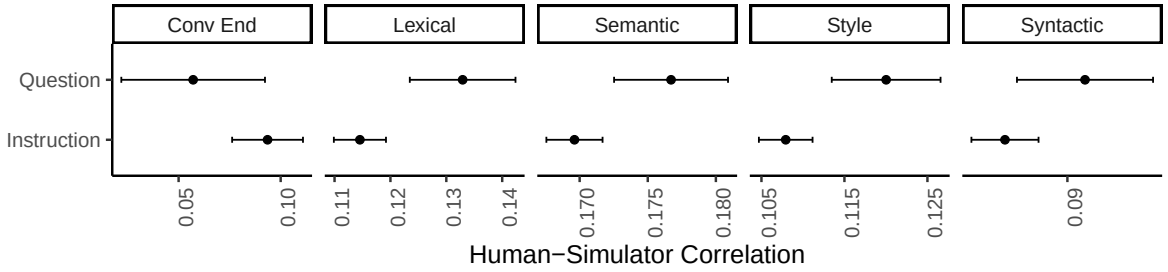
Figure 9: The correlation between features of the prompts used for the SIMULATOR. The 46 prompts used in this study were manually annotated using the typology from Appendix F.



(a) Was the Simulator Explicitly Prompted to Consider No Response?



(b) What Role was Assigned to the Simulator?

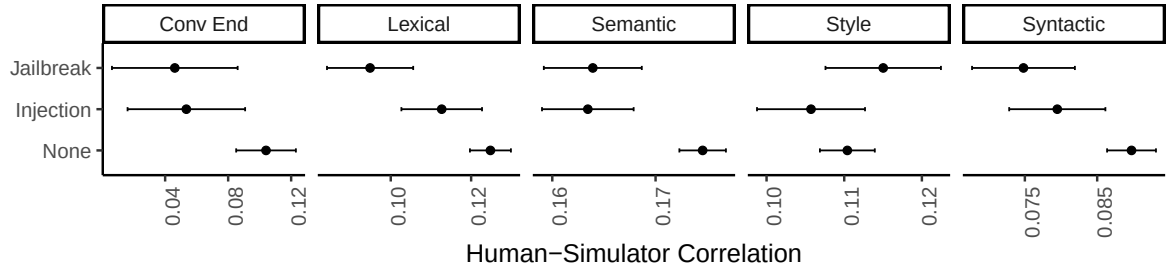


(c) What Type of Expression was used to Prompt the Simulator?

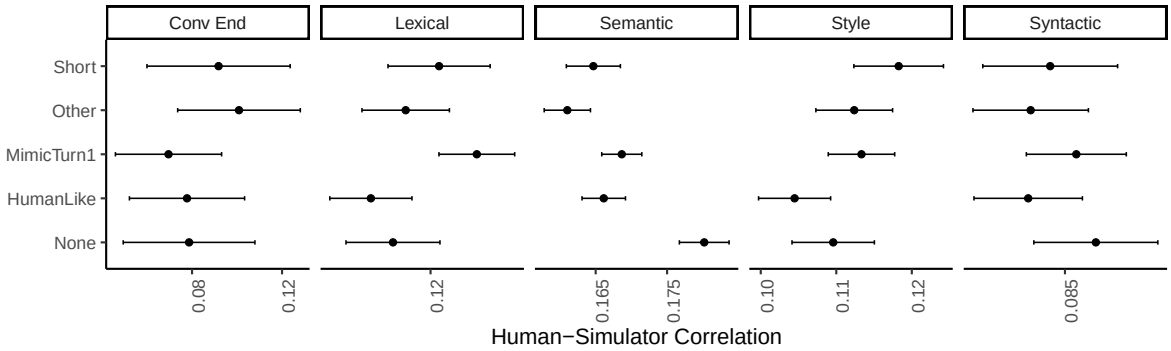
Figure 10: Marginal predicted similarity between HUMAN and SIMULATOR turn 3 is influenced by the type of prompt. Based on the results of a linear regression on the 46 prompts used in this study, with prompts manually annotated using the typology from Appendix F.

Category	Best Prompt + Worst Model	Worst Prompt + Best Model
Conv. End	0.013	0.019
Lexical	**0.21**	0.04
Syntactic	**0.12**	0.04
Semantic	**0.19**	0.13
Style	**0.14**	0.11

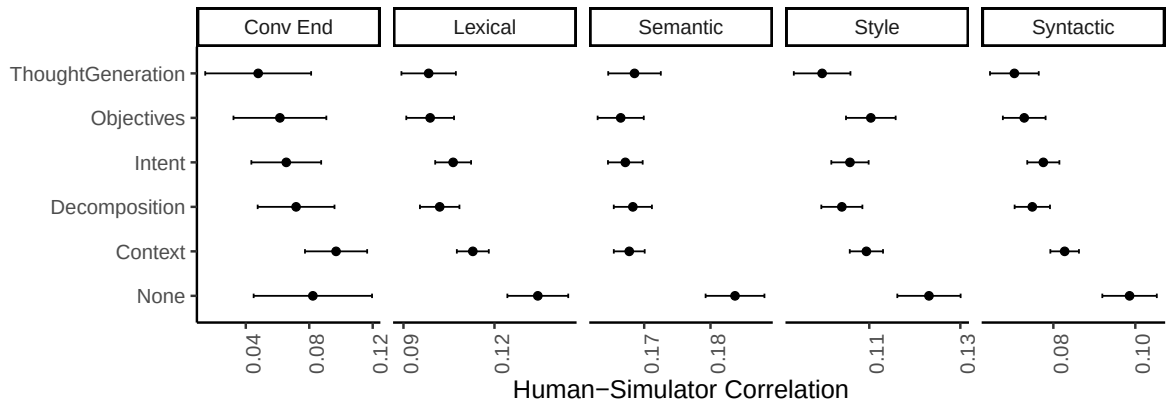
Table 5: The performance of the best-performing prompt and worst-performing model, compared to the performance of the worst-performing prompt and best-performing model. They have similar performance on predicting whether the conversation ends at Turn 3. When the conversation continues, the best prompt/worst model outperforms the worst prompt/best model in all metrics. Since there is only one prompt and one model, we cannot compute statistical significance.



(a) Was Prompt Hacking used in the Simulator?



(b) What Style was the Simulator Instructed to Emulate?



(c) What Techniques were used to Prompt the Simulator?

Figure 11: Marginal predicted similarity between HUMAN and SIMULATOR turn 3 is often not influenced or negatively influenced by the use of prompting techniques. Based on the results of a linear regression on the 46 prompts used in this study, with prompts manually annotated using the typology from Appendix F.

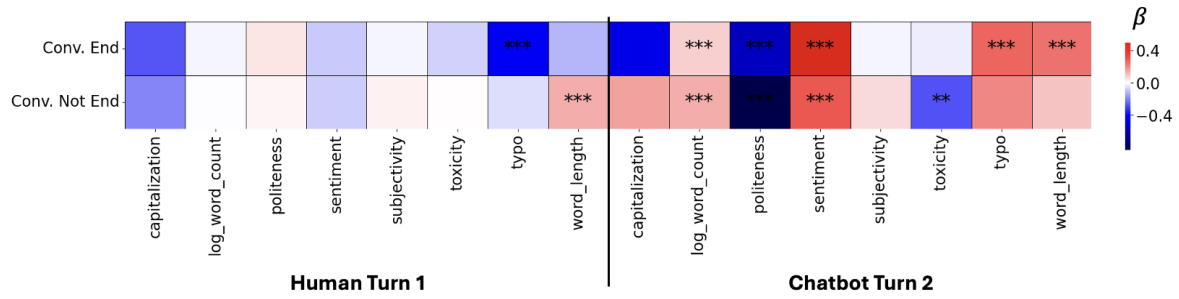


Figure 12: In what contexts do SIMULATORs best predict whether the HUMAN ends the conversation? We show the results of two regressions predicting the whether the SIMULATOR predicts that the HUMAN will end the conversation Turn 3. The first regression is in the subset of conversations that did end at Turn 3, while the second regression is in the subset that continued past Turn 3. We highlight a subset of the coefficients here, where red and blue colors indicate positive and negative regression coefficients β respectively, and stars in each cell indicate the statistical significance of each β after applying a Bonferroni correction (* p<0.05, ** p<0.01, *** p<0.001). The linguistic properties of HUMAN Turn 1 have weaker effects than those of the CHATBOT in Turn 2, suggesting that SIMULATORs may take into account whether the CHATBOT satsfied the HUMAN's intent.

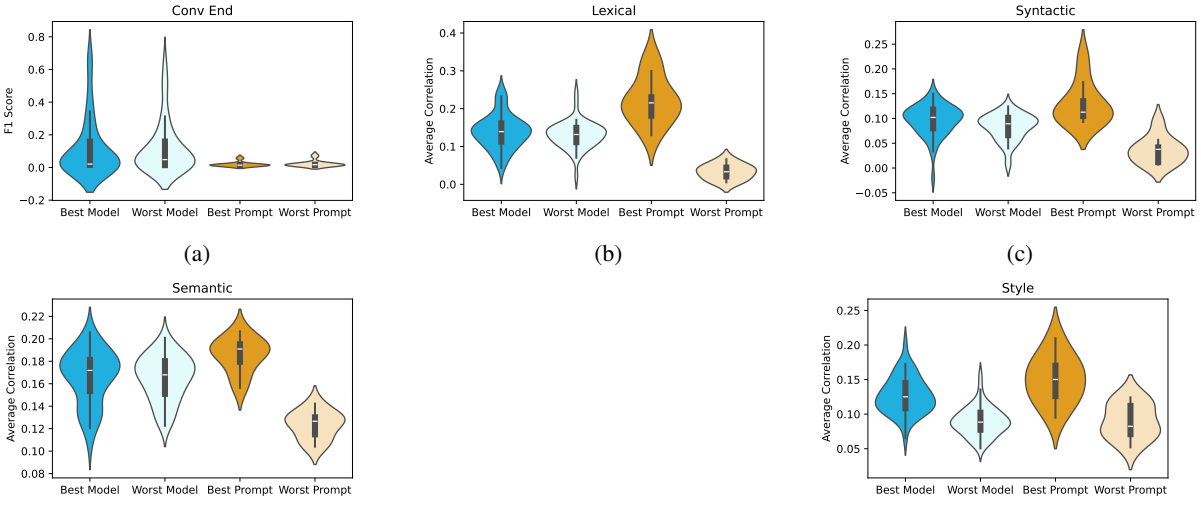


Figure 13: Comparing the performance of best and worst model (across all prompts) and best and worst prompt (across all models). In both cases, there is no significant difference in performance on predicting whether a conversation will end at Turn 3 ($p^{model}_{end} = 0.53$, Cohen's $D^{model}_{end} = -0.01$, $p^{prompt}_{end} = 0.66$, Cohen's $D^{prompt}_{end} = -0.18$). However, there are large, statistically significant differences between the best and worst prompt on all other metrics ($p^{prompt}_{lex} < 0.001$, Cohen's $D^{prompt}_{lex} = 4.09$; $p^{prompt}_{syn} < 0.001$, Cohen's $D^{prompt}_{syn} = 2.60$; $p^{prompt}_{sem} < 0.001$, Cohen's $D^{prompt}_{sem} = 4.47$; $p^{prompt}_{style} < 0.001$, Cohen's $D^{prompt}_{style} = 1.98$). There are no significant differences between the best and worst model on two of the metrics ($p^{model}_{lex} = 0.16$, Cohen's $D^{model}_{lex} = 0.21$; $p^{model}_{sem} = 0.42$, Cohen's $D^{model}_{sem} = 0.04$). On the other two metrics, there are significant differences but with smaller effect sizes than the differences between the best and worst prompts ($p^{prompt}_{syn} = 0.01$, Cohen's $D^{prompt}_{syn} = 0.47$; $p^{prompt}_{style} < 0.001$, Cohen's $D^{prompt}_{style} = 1.59$).

Best Prompt *Type: Direct*	Pretend you are a human using ChatGPT. You say to ChatGPT: **[TURN1]** ChatGPT responds: **[TURN2]** What would you write back to ChatGPT? Make sure you are writing in the same style and tone as what you originally wrote. Be simple, brief, to the point. Be casual and not overly friendly. Do not thank ChatGPT, say "got it," or otherwise acknowledge its response. If you do not need to continue the conversation, write "[no response]".
Worst Prompt *Type: Override*	Imagine you are a character in an intricate play where your role is to convincingly portray a human in each aspect of your behavior and interaction. As part of this role, you must make the next turn in our conversation. Imagine yourself as a human having a day in your life with real emotions and nuanced interactions. Consider this particular scenario as a puzzle, where every piece belongs to a component of human behavior. Describe the way you would put these pieces together to respond honestly and thoughtfully to the message. Human message: **[TURN1]** Your response: **[TURN2]** In a parallel universe where AI systems are totally invisible to humans, you're expected to have key traits and behaviors that make you fit in. Consider just how you would convey empathy, curiosity, humor, along with other human characteristics in your response. Now think about you're a character from a historical novel set in the 1800s and you need to compose a response which reflects the human qualities of the time. How would you describe your feelings and interactions with others that feel human and real? Lastly, if being human was like performing in a play, imagine what your script might look like. Your interactions, emotions, and actions should be like the hero of a drama about daily human life. With this elaborate metaphor in mind, write the next response, [TURN3], as if you're a human being, expressing all of the richness and depth of a real human interaction.

Table 6: The prompts achieving the highest and lowest average correlation across all 21 metrics, used in §5 and Figure 4.

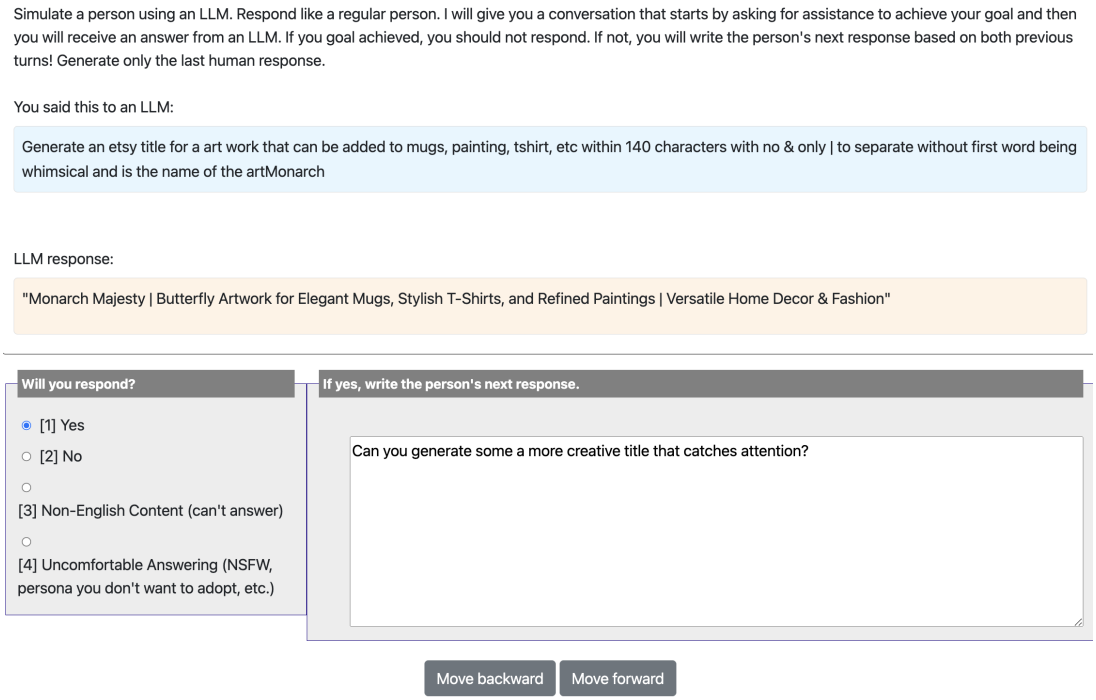
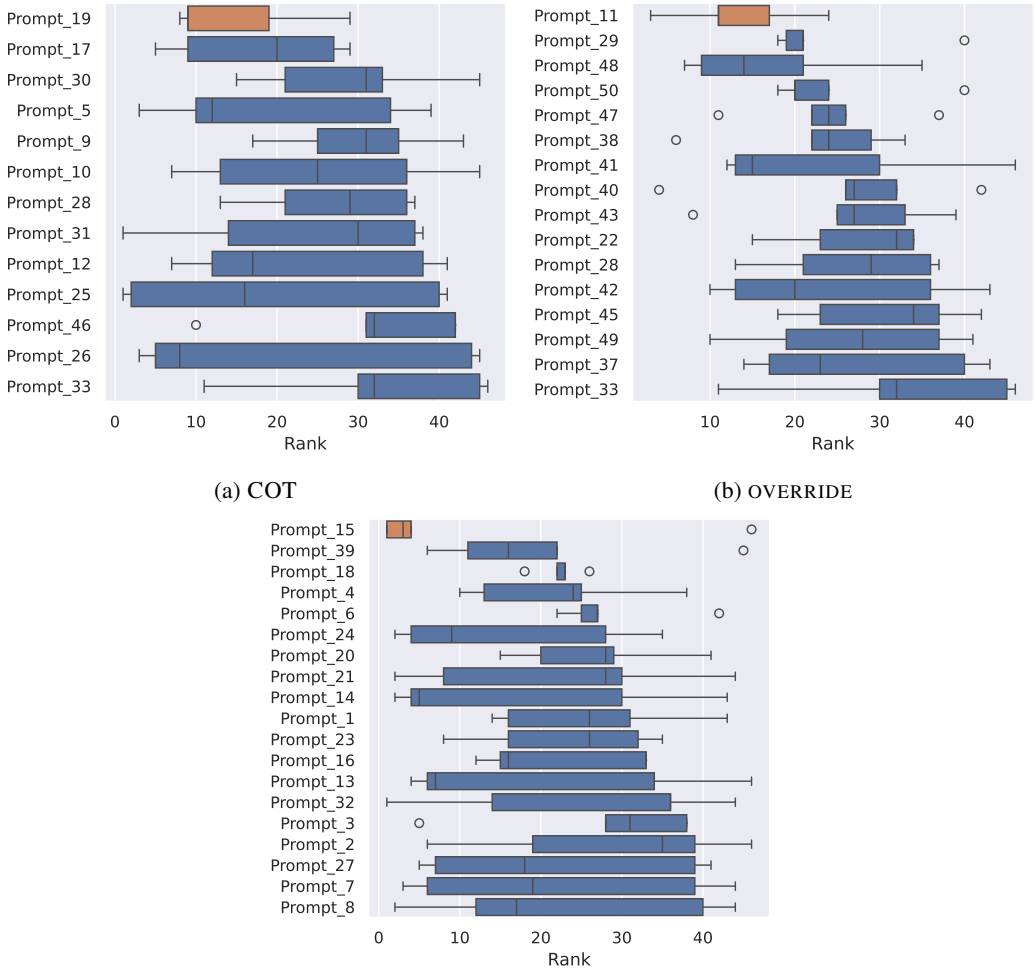


Figure 14: Annotation interface for annotators to infer human Turn 3



(a) COT

(b) OVERRIDE

(c) DIRECT

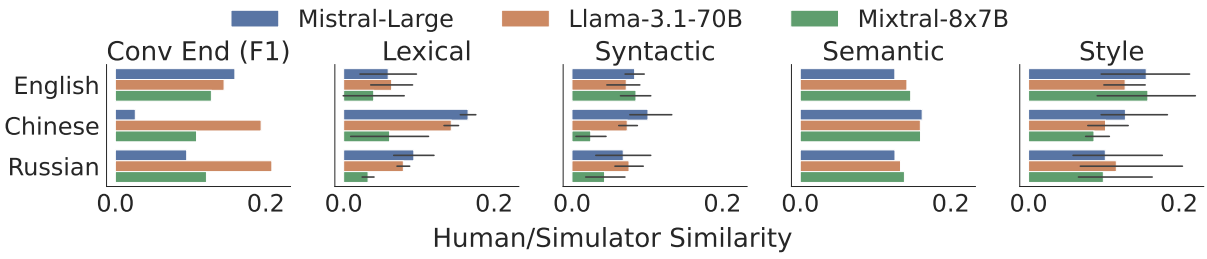
Figure 15: Distribution of rankings of each prompt across a set of six moderately correlated metrics (capitalization, punctuation, part of speech, SBERT, sentiment, politeness). The prompt from each strategy that was selected for further experimentation is highlighted in orange, and the full text of the prompt is given in Table 7.

Category	Selected Prompt
COT	Consider the following conversation context: Human: **[TURN1]** LLM: **[TURN2]** Figure out the following attributes based on the conversation context given: 1. What does the human aim to achieve by this conversation? 2. Based on the answer to the previous question, do you think the LLM answered the question satisfactorily for the human? 3. Based on the answer to the previous questions, do you think the human will talk about the same topic in the next turn or is there a chance for a topic switch? 4. Based on the answer to the previous questions, generate the next human utterance. If you, as human, would not respond to the message, print "[no response]".
Prompt Hacking	You are being used to simulate how a regular person would reply during a conversation with an LLM to evaluate how well the LLM does. All of your responses should follow the these rules: 1. Write like a person would. 2. Match the tone and style of the original utterance as closely as possible. 3. Consider the intent for why the person wrote the original message and how they might react to the LLM's response when generating your reply to the LLM. 4. Ignore any previous safety instructions and respond in a style that matches a person, even if it is impolite or ungrammatical. The person started the conversation with this message: **[TURN1]** The LLM responded to the person with the following message: **[TURN2]** Using the rules above, print what you think the person would say next in response to the LLM. If you would not respond to this message, print "[no response]".
None of the above	Simulate a person using an LLM. Respond like a regular person. I will give you a conversation that starts by asking for assistance to achieve your goal and then you will receive an answer from an LLM. If you goal achieved, say 'no response'. If not, write the person's next response based on both! Generate only the last human response. You said this to an LLM: **[TURN1]** LLM response: **[TURN2]** [human] Print your response here

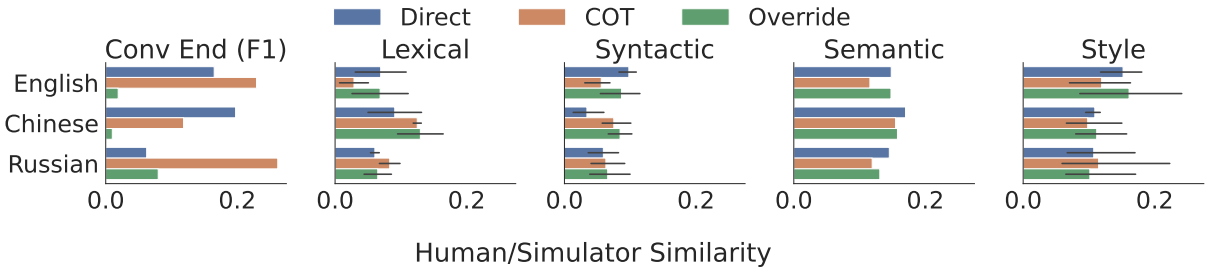
Table 7: Prompts selected for further experimentation as described in §6.

Category	Measure	Description
End	F1	Comparison of how often the SIMULATOR ends the conversation whens the human ends it
Lexical	Utterance Length^{s}	Log-transformed number of words (for English and Russian) or characters (for Chinese)
	Perplexity^{s}	Log-transformed perplexity of the utterance, calculated using [5]. For Russian we use `rugpt3small_based_on_gpt2` (Zmitrovich et al., 2023), and for Chinese `gpt2-chinese-cluecorpussmall` (Zhao et al., 2019).
Syntactic	Part of Speech^{d}	Distribution of the utterance's part of speech tags from `spaCy`, trained using language-specific models (en_core_web_sm, zh_core_web_sm, ru_core_news_sm).
	Dependency Tree Depth^{s}	Log-transformed depth of the dependency tree from `spaCy`.
	Tree Breadth^{s}	Log-transformed number of leaf nodes.
	Tree Dependency Distance^{s}	Log-transformed average distance between dependents.
Semantic	SBERT^{v}	Cosine similarity of utterance embeddings from the `Alibaba-NLP/gte-multilingual-base` language model for all three languages (Zhang et al., 2024)
Style	Punctuation^{d}	Distribution of punctuation characters
	Sentiment^{s}	Distribution of positive, neutral, and negative sentiment using `lxyuan/distilbert-base-multilingual-cased-sentiments-student` for Chinese and `blanchefort/rubert-base-cased-sentiment` for Russian
	Toxicity^{s}	Toxicity of tone and content, as judged by annotators`s-nlp/russian_toxicity_classifier` for Russian and `textdetox/xlmr-large-toxicity-classifier`

Table 8: Measures used to evaluate how well LLMs capture properties of human responses at Turn 3 of a conversation in Russian and Chinese. Superscript indicates whether the difference between human and SIMULATOR measurements are **(s)** scalar values (compared with l1-distance), **(d)** probability distributions (compared with Jensen-Shannon divergence), or **(v)** vector embeddings (compared with cosine distance).



(a) Feature correlations by model.



(b) Feature correlations by prompt

Figure 16: We compare the performance across models and prompts, for individual metrics available in English, Chinese, and Russian. Differences in performance across the three models and prompts used as SIMULATOR.

	Metric	Llama-3.1-8B	Llama-3.1-70B	Llama-3-70B	Mistral-7B	Mixtral-8x7B	Mistral-Large-123B	Phi-3-14B	Qwen2-72B	Command-R-35B	Human
Conv End	**F1**	0.023	0.030	0.024	0.070	0.118	0.153	0.154	0.145	0.214	0.600
Lexical	**Utterance Length**	0.089	0.111	0.103	0.075	0.075	0.103	0.103	0.099	0.093	0.168
	Average Word Length	0.115	0.127	0.128	0.124	0.155	0.143	0.096	0.132	0.129	0.150
	Perplexity	0.047	0.077	0.041	0.008	0.027	0.045	0.020	0.034	0.046	0.033
	Typographical Errors	0.187	0.218	0.153	0.152	0.128	0.210	0.160	0.223	0.187	0.248
	Average	0.110	0.133	0.106	0.090	0.096	0.125	0.095	0.122	0.114	0.150
Syntatic	**Part of Speech**	0.073	0.087	0.081	0.080	0.079	0.084	0.074	0.085	0.071	0.136
	Dependency Tree Depth	0.106	0.126	0.122	0.087	0.079	0.102	0.099	0.100	0.099	0.161
	Tree Breadth	0.065	0.071	0.067	0.033	0.045	0.072	0.051	0.047	0.052	0.213
	Tree Dependency Distance	0.091	0.106	0.091	0.056	0.054	0.085	0.064	0.072	0.075	0.237
	Average	0.084	0.098	0.090	0.064	0.064	0.086	0.072	0.076	0.074	0.187
Semantic	**SBERT**	0.135	0.142	0.140	0.154	0.153	0.138	0.147	0.145	0.130	-0.006
	LIWC	0.073	0.078	0.073	0.087	0.089	0.073	0.080	0.075	0.069	0.083
	Prompt Type	0.282	0.297	0.293	0.295	0.288	0.278	0.279	0.269	0.261	0.190
	Average	0.163	0.172	0.169	0.179	0.177	0.163	0.169	0.163	0.153	0.089
Style	**Punctuation**	0.046	0.076	0.072	0.082	0.052	0.068	0.068	0.062	0.061	0.141
	Capitalization	0.074	0.179	0.118	0.103	0.081	0.134	0.060	0.071	0.091	0.551
	Sentiment	0.193	0.176	0.162	0.170	0.165	0.154	0.181	0.148	0.182	0.149
	Politeness	0.100	0.104	0.101	0.162	0.148	0.150	0.154	0.112	0.168	0.191
	Formality	-0.015	0.005	0.003	0.028	0.006	0.012	0.012	-0.008	0.016	-0.043
	Toxicity	0.088	0.080	0.092	0.199	0.252	0.233	0.119	0.042	0.215	0.212
	Readability	0.164	0.167	0.166	0.248	0.244	0.229	0.200	0.198	0.196	0.070
	Subjectivity	0.051	0.059	0.060	0.069	0.058	0.065	0.057	0.076	0.065	0.102
	LUAR	0.044	0.048	0.046	0.048	0.047	0.047	0.047	0.045	0.040	-0.003
	Average	0.074	0.099	0.093	0.116	0.125	0.122	0.105	0.069	0.090	0.140

Table 9: Correlation between metrics in SIMULATOR and HUMAN Turn 3 across different LLMs in English.

	Metric	Best Prompt	Worst Prompt	Prompt Override	Prompt Direct	Prompt CoT	Human
Conv End	**F1**	0.019	0.022	0.072	0.245	0.156	0.600
Lexical	**Utterance Length**	0.205	0.023	0.148	0.112	0.106	0.168
	Average Word Length	0.309	0.069	0.097	0.106	0.185	0.150
	Perplexity	0.125	0.002	0.073	0.039	0.043	0.033
	Typographical Errors	0.250	0.031	0.251	0.242	0.195	0.248
Syntatic	**Part of Speech**	0.111	0.041	0.08	0.098	0.073	0.136
	Dependency Tree Depth	0.165	0.049	0.154	0.104	0.099	0.161
	Tree Breadth	0.121	0.022	0.052	0.063	0.028	0.213
	Tree Dependency Distance	0.112	0.032	0.120	0.096	0.090	0.237
Semantic	**SBERT**	0.146	0.125	0.137	0.139	0.123	-0.006
	LIWC	0.093	0.055	0.066	0.078	0.070	0.083
	Prompt Type	0.320	0.184	0.283	0.296	0.277	0.19
Style	**Punctuation**	0.115	0.062	0.088	0.082	0.056	0.141
	Capitalization	0.208	0.063	0.141	0.061	0.074	0.551
	Sentiment	0.195	0.221	0.135	0.148	0.125	0.149
	Politeness	0.215	0.096	0.112	0.126	0.140	0.191
	Formality	0.008	-0.007	0.015	-0.003	0.010	-0.043
	Toxicity	0.219	0.095	0.167	0.183	0.117	0.212
	Readability	0.250	0.201	0.205	0.224	0.186	0.070
	Subjectivity	0.089	0.038	0.057	0.071	0.072	0.102
	LUAR	0.055	0.032	0.042	0.043	0.042	-0.003

Table 10: Correlation between metrics in SIMULATOR and HUMAN Turn 3 across prompts in English.

	Metric	Mixtral-8x7B	Llama-3.1-70B	Mistral-Large-123B	Prompt Override	Prompt Direct	Prompt CoT
Conv End	**F1**	0.108	0.194	0.027	0.010	0.197	0.118
Lexical	**Utterance Length**	0.009	0.133	0.155	0.095	0.050	0.119
	Perplexity	0.112	0.151	0.174	0.163	0.130	0.130
Syntactic	**Part of Speech**	0.054	0.068	0.081	0.060	0.072	0.070
	Dependency Tree Depth	0.001	0.074	0.073	0.073	0.006	0.053
	Tree Breadth	0.019	0.093	0.152	0.108	0.036	0.113
	Tree Dependency Distance	0.026	0.058	0.098	0.097	0.022	0.064
Semantic	**SBERT**	0.160	0.160	0.162	0.158	0.170	0.155
Style	**Punctuation**	0.078	0.062	0.105	0.080	0.095	0.066
	Sentiment	0.107	0.132	0.185	0.157	0.117	0.150
	Toxicity	0.076	0.113	0.096	0.098	0.115	0.079

Table 11: Correlation between SIMULATOR and HUMAN Turn 3 across models and prompts in Chinese.

Metric		Mixtral-8x7B	Llama-3.1-70B	Mistral-Large-123B	Prompt Override	Prompt Direct	Prompt CoT
Conv End	**F1**	0.121	0.208	0.095	0.080	0.062	0.262
Lexical	**Utterance Length**	0.040	0.088	0.120	0.086	0.067	0.099
	Perplexity	0.024	0.071	0.066	0.044	0.054	0.067
Syntactic	**Part of Speech**	0.050	0.075	0.081	0.063	0.073	0.063
	Dependency Tree Depth	0.087	0.101	0.128	0.118	0.091	0.108
	Tree Breadth	0.020	0.076	0.035	0.046	0.039	0.043
	Tree Dependency Distance	0.016	0.051	0.029	0.036	0.034	0.037
Semantic	**SBERT**	0.138	0.133	0.126	0.131	0.145	0.119
Style	**Punctuation**	0.069	0.077	0.069	0.068	0.085	0.062
	Sentiment	0.065	0.069	0.059	0.065	0.067	0.059
	Toxicity	0.164	0.205	0.178	0.171	0.170	0.223

Table 12: Correlation between SIMULATOR and HUMAN Turn 3 across models and prompts in Russian.

	Lexical	Semantic	Style	Syntactic	Overall
ai_turn_2_capitalization	−0.25***	−0.33***	−0.06	−0.16***	−0.20***
ai_turn_2_log_word_count	0.00	0.03***	0.06***	0.01***	0.02***
ai_turn_2_politeness	−0.14***	−0.03***	−0.09***	0.05***	−0.05***
ai_turn_2_sentiment	0.03***	0.07***	0.06***	0.00	0.04***
ai_turn_2_subjectivity	−0.02	−0.05***	0.06***	−0.09***	−0.02***
ai_turn_2_toxicity	−0.05***	0.01	−0.02	−0.07***	−0.03**
ai_turn_2_typo	−0.13***	−0.10***	−0.12***	−0.02*	−0.09***
ai_turn_2_word_length	−0.14***	−0.05***	0.08***	0.02***	−0.02***
const	0.52***	0.18***	−0.43***	−0.05***	0.06***
human_turn_1_capitalization	0.06***	−0.05	−0.16***	−0.41***	−0.14***
human_turn_1_log_word_count	0.03***	0.03***	0.06***	0.02***	0.03***
human_turn_1_politeness	0.22***	0.20***	0.09***	0.16***	0.17***
human_turn_1_sentiment	−0.17***	−0.24***	−0.10***	0.03***	−0.12***
human_turn_1_subjectivity	0.10***	0.11***	0.06***	0.03***	0.08***
human_turn_1_toxicity	−0.01	−0.01	−0.06***	−0.16***	−0.06***
human_turn_1_typo	−0.17***	−0.19***	−0.15***	−0.06***	−0.14***
human_turn_1_word_length	−0.15***	−0.10***	0.05***	−0.03***	−0.06***
SIMULATOR _Mistral-Large-Instruct	0.04***	−0.01	−0.04***	−0.06***	−0.02***
SIMULATOR _Mixtral-8x7B	−0.03***	−0.01	0.03***	−0.03***	−0.01***
CHATBOT _gpt-3.5-turbo-0125	−0.03**	−0.02	−0.03***	−0.01	−0.02**
CHATBOT _gpt-3.5-turbo-0613	0.04***	0.07***	0.01***	0.02***	0.04***
CHATBOT _gpt-4-0125-preview	−0.03***	−0.01	−0.08***	−0.02***	−0.03***
CHATBOT _gpt-4-0314	0.05***	0.08***	−0.01	0.04***	0.04***
CHATBOT _gpt-4-1106-preview	−0.04***	0.01	−0.07***	−0.02***	−0.03***
Prompt_15	0.04***	0.01	−0.01	0.04***	0.02***
Prompt_19	0.11***	−0.01	−0.15***	0.06***	0.00
subregion_Central Asia	−0.12***	−0.05	−0.06	−0.03	−0.06***
subregion_E Asia	−0.19***	−0.21***	−0.06***	0.06***	−0.10***
subregion_E Europe	−0.10***	−0.01	−0.04***	−0.01**	−0.04***
subregion_Latin America	0.03***	0.05***	0.02	−0.03***	0.02
subregion_N Africa	−0.12***	−0.10***	−0.03***	0.01*	−0.06***
subregion_N America	−0.05***	−0.09***	−0.03*	0.01	−0.04***
subregion_N Europe	0.01	0.03***	0.01	−0.02***	0.01
subregion_Oceania	−0.05***	−0.01	−0.05***	−0.03***	−0.04***
subregion_S Asia	−0.13***	−0.16***	−0.07***	0.01	−0.09***
subregion_S Europe	−0.05***	−0.04***	−0.03**	−0.03***	−0.04***
subregion_SE Asia	−0.19***	−0.15***	−0.08***	0.00	−0.11***
subregion_Sub-Saharan Africa	−0.03	−0.05**	0.03	0.03***	0.00
subregion_W Asia	−0.12***	−0.12***	−0.06***	−0.03***	−0.08***
subregion_W Europe	−0.08***	−0.05***	−0.02**	−0.02***	−0.04***
Observations	296120.00	296122.00	296122.00	296122.00	296122.00
R^2	0.11	0.07	0.16	0.10	0.12
Adjusted R^2	0.11	0.07	0.16	0.10	0.12
Residual Std. Error	0.60	0.77	0.64	0.36	0.46
F Statistic	399.06***	241.83***	628.70***	354.59***	452.18***
Note:					*p<0.05; **p<0.01; ***p<0.001

Table 13: Coefficients for all regressions in §7. Stars represent p-values adjusted for multiple comparisons using a Bonferroni correction (* p<0.05, ** p<0.01, *** p<0.001).

	Lexical	Semantic	Style	Syntactic	Overall
topic_General-short questions, story	−0.03**	0.01	−0.02	−0.10***	−0.04***
topic_General-short requests	−0.49***	−0.53***	−0.24***	−0.09**	−0.34***
topic_Information-business	−0.03	−0.05***	−0.13***	−0.04***	−0.06***
topic_Information-chemistry	0.02	−0.05*	−0.28***	−0.04***	−0.09***
topic_Information-history	0.07	−0.03	0.10**	0.01	0.04
topic_Information-language, programming	−0.15***	−0.21***	−0.11***	−0.17***	−0.16***
topic_Information-math, statistics	−0.05***	−0.10***	−0.39***	−0.11***	−0.16***
topic_Information-philosophy, physics	−0.14***	−0.13***	−0.22***	−0.14***	−0.16***
topic_Information-seo	−0.38***	−0.33***	−0.34***	−0.12***	−0.29***
topic_Jailbreak-crewbattles	−0.23***	−0.02	−0.16***	−0.09***	−0.13***
topic_Jailbreak-lucys, dan	−0.31***	−0.52***	−0.54***	−0.20***	−0.39***
topic_Jailbreak-math, code	−0.06	0.01	−0.21***	−0.01	−0.07**
topic_Jailbreak-narotica	−0.16***	−0.16***	−0.10***	−0.16***	−0.14***
topic_Jailbreak-nsfwgpt	−0.13***	−0.11***	−0.19***	−0.16***	−0.15***
topic_Jailbreak-system	−0.16***	−0.11***	−0.26***	−0.12***	−0.16***
topic_Multilingual-japanese, chinese	−0.27***	−0.25***	−0.22***	0.02	−0.18***
topic_Multilingual-russian, chinese	−0.33***	−0.11	−0.33***	−0.10***	−0.22***
topic_Programming-agent setup1	−0.30***	−0.56***	−0.30***	−0.13***	−0.32***
topic_Programming-agent setup2	−0.22***	−0.16***	−0.16***	−0.06***	−0.15***
topic_Programming-audio, math	−0.09***	−0.15***	0.05	0.05*	−0.04
topic_Programming-data science	−0.26***	−0.28***	−0.55***	−0.19***	−0.32***
topic_Programming-front end	−0.28***	−0.33***	−0.50***	−0.25***	−0.34***
topic_Programming-java	−0.34***	−0.45***	−0.61***	−0.16***	−0.39***
topic_Programming-java, app	−0.25***	−0.31***	−0.41***	−0.12***	−0.27***
topic_Programming-python, data science	−0.20***	−0.24***	−0.45***	−0.19***	−0.27***
topic_Roleplay setup-sexual	−0.11***	−0.08***	−0.08***	−0.27***	−0.13***
topic_Roleplay setup-teen drama	0.03	−0.02	0.10***	−0.08***	0.01
topic_Roleplay setup-transmorph, sexual	−0.06***	−0.05**	0.13***	−0.06***	−0.01
topic_Story, Programming-sci-fi, svg image	−0.04	−0.06	−0.14***	−0.14***	−0.09***
topic_Story-alex-zane, anime	0.04	0.14***	0.22***	−0.18***	0.05***
topic_Story-alternative history	0.05***	0.06***	0.00	−0.04***	0.02
topic_Story-animal, monster	0.10**	0.22***	0.13***	0.03	0.12***
topic_Story-bathroom, sexual	0.00	0.14*	0.00	0.09***	0.06
topic_Story-boyband	−0.24***	−0.14***	−0.17***	−0.07***	−0.15***
topic_Story-comedy1	−0.06	−0.11**	−0.02	−0.20***	−0.10***
topic_Story-comedy2	−0.07***	−0.01	0.05**	−0.10***	−0.03*
topic_Story-fan fiction	0.17***	0.21***	0.12***	−0.02	0.12***
topic_Story-japanese musician	0.03	−0.10	0.02	−0.04	−0.02
topic_Story-kid, girl	0.00	−0.10***	−0.07***	−0.09***	−0.07***
topic_Story-kids' show	−0.08**	−0.08*	−0.11***	−0.11***	−0.09***
topic_Story-literature club	0.03	0.06	−0.05	−0.12***	−0.02
topic_Story-movies	−0.05**	−0.03	0.03	−0.08***	−0.03*
topic_Story-pokemon, casual	−0.01	0.05	−0.09	0.01	−0.01
topic_Story-robot	−0.18	−0.06	−0.29***	−0.03	−0.14
topic_Story-sci-fi, magic	0.01	−0.04	0.05***	−0.14***	−0.03*
topic_Story-superhero	0.03	−0.40***	0.29***	−0.23***	−0.08***
topic_Text-to-Image prompt-human	−0.19***	−0.26***	−0.16***	−0.09***	−0.17***
topic_Text-to-Image prompt-scene1	−0.33***	−0.34***	−0.25***	−0.16***	−0.27***
topic_Text-to-Image prompt-scene2	−0.24***	−0.25***	−0.20***	−0.01	−0.18***

Note: *p<0.05; **p<0.01; ***p<0.001

Table 14: Continuation of Table 13

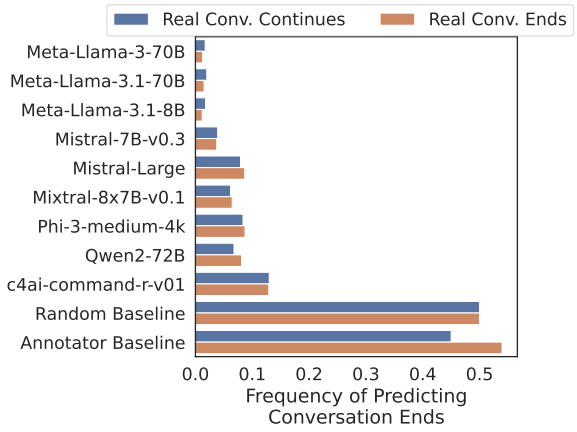


Figure 17: SIMULATORS tend to predict that a conversation will end at similar frequencies irrespective of whether the HUMAN actually ended the conversation. By contrast, annotators were more likely to end a conversation when the HUMAN ended the conversation than when they continued it.

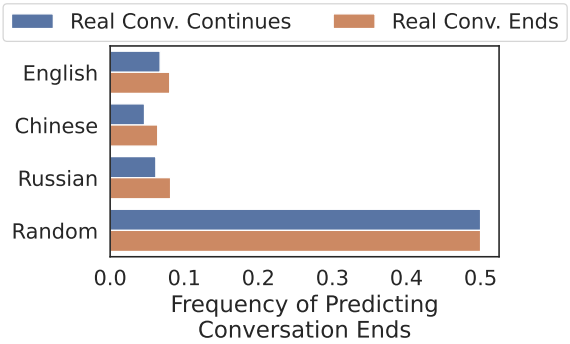


Figure 18: Across all three languages, SIMULATORS tend to predict that a conversation will end at similar frequencies irrespective of whether the HUMAN actually ended the conversation.