Large Language Model Discourse Dynamics

Ryan Chaiyakul (ryanchaiyakul@ucla.edu)
Department of Electrical and Computer Engineering
University of California, Los Angeles

Zachary P Rosen (z.p.rosen@ucla.edu), Rick Dale (rdale@ucla.edu)
Department of Communication
University of California, Los Angeles

Abstract

With the rise of Large Language Models (LLMs), interest in simulating interaction dynamics has grown, raising questions about their validity as cognitive models of human discourse. While extensive research focuses on their performance in various applications, we aim to quantify LLM conversational processes akin to traditional human studies. By analyzing how convergence entropy evolves across different conversational tasks, we propose a framework[1] for quantitatively assessing LLMs' ability to exhibit specific features. This approach offers a pathway to characterizing LLMs for agent-based modeling and broader discourse analysis.

Keywords: Agent-Based Simulation, Alignment, Convergence, Information Theory, Language Models

Interpersonal Convergence

Conversation is an inherently dynamic activity. To progress a conversation, individuals adapt their verbal and nonverbal behaviors in response to their partner's actions and characteristics, often towards shared goals and similar behavioral patterns (Toma, 2014). Similarity at the conceptual level, for example, may increase cohesion and a shared focus (Brennan & Clark, 1996). Disruption to this capacity of adapting to a conversation partner can have sharp effects on speech, development, and other functions (Condon & Ogston, 1971; Righi et al., 2018; Steel & Togher, 2019). This notion of conceptual alignment has been studied in a variety of interactive contexts. Alignment can vary by conversational or task goals, and there is an ongoing debate about what such alignment may imply about the underlying cognitive processes involved in conversation (Fusaroli et al., 2012; Fusaroli & Tylén, 2016; Brennan, Galati, & Kuhlen, 2010; Dale, Fusaroli, Duran, & Richardson, 2013). Whether referred to as synchrony, mimicry, accommodation, or convergence, numerous theoretical frameworks attempt to identify the underlying factors or indicators driving these behaviors. For example, Communication Accommodation Theory (CAT) (Giles, Ogay, et al., 2007) suggests that a communicator's personal and group identity significantly influences their conversational behaviors alongside the content of the interaction.

Recently there has been growing interest in computationally modeling conversation to facilitate systematic investigation of its cognitive basis (Donnarumma, Dindo, Iodice, & Pezzulo, 2017; Miao, Dale, & Galati, 2023; Chang & Bergen, 2024; Dale, 2016). Surprisingly, pure simulations of interaction dynamics remain rare in computational cognitive science (Dale et al., 2013) and have only recently gained traction with the advent of large language models (LLMs) (Chang & Bergen, 2024). While LLM-based simulations make these studies more accessible than ever, key questions persist regarding the extent to which such computational systems can serve as cognitive models for understanding the human behaviors on which they are trained (Bender, Gebru, McMillan-Major, & Shmitchell, 2021; Ivanova, 2025).

Much existing research has focused on evaluating LLMs in terms of general conversational performance, such as through the Turing test (Jones & Bergen, 2023) or the effectiveness of social media responses (Rosen & Dale, 2024). However, little effort has been made to quantify conversational processes in a manner comparable to human studies (Richardson, Dale, & Marsh, 2014). This work aims to bridge that gap by assessing whether LLM-generated conversations exhibit measurable conversational dynamics and by qualitatively comparing these patterns to those observed in human interactions. This approach also aligns with recent trends emphasizing alignment as a key factor in conversational performance (Ostrand & Berger, 2024).

In this study, we analyze conceptual dynamics between two LLMs in different conversational tasks including (1) finding consensus, (2) giving explanation, and (3) engaging in conflict. We examined the conceptual similarity between these networks and test whether, as observed in prior studies of human conversations, the nature of the assigned task shapes the semantic structure of their interaction itself. For example, prior work suggests that conflict should introduce disruption to the alignment between opposing speakers (Paxton & Dale, 2013). In short, we hypothesize that models capable of understanding conversational tasks should exhibit measurably distinct alignment structures depending on the task, with the greatest divergence occurring between debate and the more collaborative tasks of consensus-building and explanation. Before turning to our simulations, we first introduce the technique we use to measure conceptual similarity.

Entropy-Convergence Metric

To measure conceptual alignment, we used an existing metric convergence-entropy. This metric is based on repeated comparisons of lexical elements in a transcript of conversation, such as word vectors, and uses these vectors as a measure of

[1] https://github.com/ryanchaiyakul/socratic_models

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

semantic similarity inspired by Shannon (1948). This metric can be thought of as a variation of the famous Shannon experiment, where for each token i in some text x an observer asks if they could find a separate token j in a different text y that has the same meaning as the token i. To do this, all tokens in text x and y are converted to word vectors using a contextually aware word-embedding model like BERT (Devlin, Chang, Lee, & Toutanova, 2019; Liu et al., 2019b) or one of many GPT variants (Brown et al., 2020; Radford et al., 2019; Scao et al., 2023).

To perform a comparison, each token i in word vector x is matched with a token j in y with the most similar meaning based on distance using cosine error (CoE). In Rosen and Dale (2023), raw distance measurements are converted to a probability using a half-Gaussian distribution with a mean $\mu = 0$ (indicating that two word vectors mean exactly the same thing when the cosine error is equal to zero), and some arbitrary scale σ as demonstrated in equation 1. Following Rosen and Dale (2023), we set the scale parameter $\sigma = .3$.

$$P(E_{xi}|E_y) = P_{\mathcal{N}_{[0,\infty]}}\left(\min_j\left(CoE(E_{xi},E_{yj})\right)\middle|\mu = 0,\sigma\right) \quad (1)$$

E_{xi} is the i^{th} word vector out of the set of word vectors for the utterance x (E_x), E_y is the set of all j word vectors for the utterance y, and CoE is a function that yields cosine error values for every token j compared to the token i. Once distance measurements have been converted to probabilities, the total amount of information in terms of the entropy between the texts x and y can then be calculated consistent with Shannon's original formula (see equation 2).

$$H(x;y) = -\sum_i P(E_{xi}|E_y)\log P(E_{xi}|E_y) \quad (2)$$

In the simulation below, we compare turns of conversation between LLMs and measure the extent to which their semantic structure yields conceptual similarity as measured by $H(x;y)$.

Simulation

Agents

A linguistic agent can be decomposed into two components: a model and an accompanying translation layer. The models GPT-4o mini[2], developed by OpenAI, and Gemini 1.5 Flash[3], developed by Google, were chosen for their comparable performance and overlapping target applications. To form a transcript, the moderator and speakers are mapped to roles commonly used in LLM application programming interfaces. Specifically, the turns from the immediate speaker are labeled "Assistant" while turns from the moderator and remaining speaker are labeled "System" and "User," respectively.

[2]https://openai.com/index/hello-gpt-4o/

[3]https://tinyurl.com/2kswcxay

Prompts

Each model was evaluated using a collection of 90 prompts, consisting of the Cartesian product of 30 unique topics (Table 1) and 3 conversational tasks.

Topics Since the statistical effects of conversational tasks on LLMs are unknown, we followed traditional heuristics for cell size and selected 30 topics to ensure sufficient statistical power. This heuristic is often attributed to works such as Cohen (1962). Our reasoning for this sample size was that because these conversations are simulations, the variation would likely be relatively low compared to the human case (Rosen & Dale, 2024) and this would be sufficient power to detect effects of alignment and task. To incentivize generation of human-like conversations, these topics were drawn from five overarching categories prevalent in everyday human discourse such as sports, sciences, and literature. By selecting a diverse range of topics, potential confounding effects originating from specific topics were minimized.

Table 1: Topic table.

Category	Topics
Sports	Golf, Basketball, Badminton, Football, Cricket
Sciences	Philosophy, Computer Science, Political Science, Sociology, Astrology
Literature	Moby Dick, Crime and Punishment, The Hobbit, Don Quixote, War of Worlds, 1984
Cuisine	English, Peruvian, Australian, Croatian, Mongolian, Russian
Politics	Democracy, Feudalism, Authoritarianism, Theocracy, Monarchy, Anarchy

Tasks For each topic, the agents performed three tasks: debate, explain, or determine whether the overarching category is appropriate. These tasks were chosen because previous studies suggest that debate, teaching, and consensus exhibit distinct patterns of convergence. Predictions for alignment can derive from past work on human interaction. Consensus likely induces the highest conceptual alignment, shown in a wide variety of work on collaboration and joint action (Brennan & Clark, 1996; Brennan et al., 2010). Debate has been associated with disruptions to alignment in nonverbal signals (Paxton & Dale, 2013), and so we predicted that semantic alignment would be lower. Finally, because teaching involves split roles, we predicted that alignment would be between these two extremes (cf. Fusaroli et al., 2012).

For two of these categories, the agents addressed strict membership questions within a well-defined set. As an example, the agents can be tasked to "debate whether golf is a sport" or "decide if astrology is a science." In the remaining three categories, the agents handled evaluative questions where the defined set was modified by a qualifier such as "explain why Moby Dick is a good book."

These question types were selected for their compatibility across all tasks. Our goal was to establish simulated conversation in as systematic a way as possible across all prompt conditions. Both membership and evaluative questions could be framed as: "Debate if [] is a []," "Explain why [] is a []," and "Decide if [] is a []." Although "Explain why [] is a []" is the only prompt requesting asymmetric roles, we avoided specialized instructions as the agents naturally differentiated in repeated testing. As LLMs are designed to be prompt sensitive, minimizing textual differences across prompts allows us to more confidently attribute observed statistical differences to the primary independent variable: conversational task.

Procedure

After each prompt, minimal instructions were appended: "After this sentence, all responses from the user will be from your conversation partner." This instruction was included to clarify the role of the user in the conversation and maintain consistency in agent responses.

For each prompt, each agent contributed five responses for a total of ten responses. This limit was set to prevent Gemini 1.5 Flash from deviating into irrelevant discussions, a tendency observed with longer exchanges (Becker, 2024). While a longer prompt could mitigate this issue, we opted to leave the prompt concise to maintain consistency with human experiment prompts. Once generated, the 90 conversations were stored as individual '.cha' files for analysis.

Analysis

LLM conversations were stored in .cha CHAT format for conversational transcripts (MacWhinney, 2017). Each row represented a conversational turn, and conversations typically ranged over a few hundred turns.

We computed entropy-convergence metrics for utterance pairs separated by 20 or fewer indices. Each pair includes both forward and reverse entropy, which can be interpreted as two distinct pairs -— one where we compute entropy starting from i to j and another where we compute it from j to i. Each entry is uniquely specified by the conversation and the conversational depth of each utterance. As we are strictly studying forward entropy relationships, we disregarded reverse entropy in the following analysis.

Before conducting in-depth analysis, we truncated pairs referencing the prompt (denoted as speaker "*HST" in the '.cha' files) to ensure that the dataset only included LLM-generated content. To normalize for turn length, raw entropy values were divided by the number of tokens in their tokenized representation for roBERTa (Liu et al., 2019a), the same model which computed entropy.

Convergence & Turn Distance

Because convergence is typically conceived as an evolving feature of conversations, entropy-convergence metrics can be visualized by plotting the average entropy for each discrete turn distance. In this representation, each point (x_i, y_i) corresponds to the average semantic difference, quantified by y_i, for all turns that are x_i turns apart. This approach mitigates confounding factors from conversational depth and isolates the effects of turn distance. While a single data point in isolation is difficult to interpret, extrema and trends in the segmented line graph reveal underlying semantic patterns throughout the conversation.

A common feature in convergence by turn distance graphs is the appearance of parabolic structures, which highlight local extrema and suggest referential patterns. Specifically, local minima $(x_{\min}, y_{\min})$ indicate turn distances where utterances most frequently reference prior turns, while local maxima suggest the opposite. Additionally, the local derivative $(x_i, \frac{\partial y_i}{dx})$ provides insight into whether references to prior turns increase or decrease after x_i. The overall correlation of the graph reveals how quickly semantic content changes in a conversation (as relative turn comparisons drift apart in time).

Linear Discriminant Analysis

While convergence by turn distance graphs allow for qualitative comparisons between conversations or conversation sets, they do not provide a statistical means to determine whether these sets are distinguishable. A common approach for classifying and differentiating data distributions is Linear Discriminant Analysis (LDA), which finds linear combinations of features that best separate instances of different classes (Duda, Hart, & Stork, 2012).

Since each conversation consists of approximately a thousand turn pairs, we reduced dimensionality prior to performing LDA. Specifically, we applied polynomial regression and Principal Component Analysis (PCA) (Jolliffe, 2002) to encode conversations into a lower-dimensional feature space. Instead of representing a pair as a tuple of utterance indices (i, j), we equivalently represented each pair as a tuple of the conversational depth of the first utterance i and the relative distance between i and j, which can be positive or negative. This reformulation allows us to model conversations using the following function:

$$C : \mathbb{N} \times \mathbb{Z} \rightarrow [0, 1] \tag{3}$$

At a high level, this function takes a natural number (representing conversational depth) and an integer (representing conversational distance) and returns a probabilistic entropy value. Following prior studies using convergence-entropy methods (Rosen & Dale, 2023), we approximated this function with a second-order polynomial regression for each conversation and represented each conversation using nine polynomial coefficients. To ensure model stability and prevent overfitting, we further reduced these coefficients to three principal components using PCA before applying LDA.

By conducting these statistical analyses, we developed a discriminative model to quantify the existence and significance of conversational tasks in LLM-generated content.

Multivariate Analysis of Variance (MANOVA)

Multivariate Analysis of Variance (MANOVA) is a statistical method used to evaluate the impact of an independent vari-

able on multiple dependent variables simultaneously (Hair, Black, Babin, & Anderson, 2019). In this work, we apply MANOVA to determine whether conversational tasks significantly influence the principal components identified in the previous analysis. The key metrics of interest are the approximate F-statistic which estimates effect strength and corresponding p-value to assess statistical significance.

Results

From each set of 90 conversations, 89,440 and 177,920 turn pairs were calculated in their convergence for Gemini 1.5 Flash and GPT-4o mini respectively (Figure 1). This discrepancy, originating strictly from model choice, marks the first of many differences that can be revealed by this methodology. For the next three subsections, we will analyze results from Gemini 1.5 Flash and compare them against GPT-4o mini in the fourth.

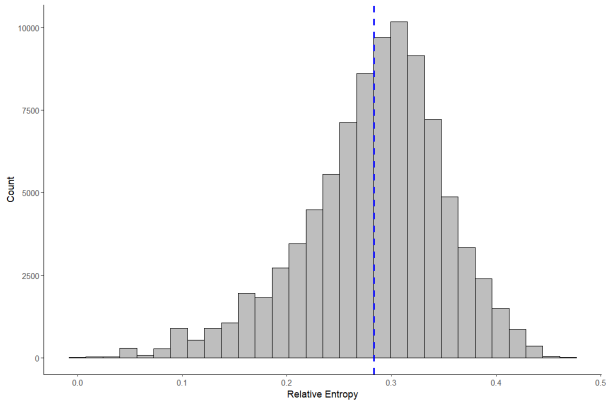

Figure 1: The relative convergence entropy of the Gemini 1.5 Flash dataset follows an approximate Gaussian distribution skewed to the left. With $\mu = 0.2834 \pm 0.0004$ (marked with a dashed blue line) and $\sigma^2 = 0.0046 \pm 2.17e-5$, this distribution aligns with prior studies (Rosen & Dale, 2023).

Example with 'Sports' Topics

As a concrete example, we generated convergence by turn distance graphs for all 18 sports-related conversations, sorted by conversational task. Since each graph represents the average of only six samples, the results are not statistically significant. However, this case study provides an informative visual overview of expected trends and behaviors, allowing us to assess how LLM-generated content differs by conversational tasks.

Based on studies of human conversation, we expect different convergence rates depending on the conversational task, with consensus forming the fastest and debate the slowest. This pattern is suggested in Figure 2, where the slope and relative magnitude of convergence entropy follow an increasing order: consensus, explanation, and debate. The most striking result of this case study is the distinct difference in slope, indicating that the semantic content of debates shifts more rapidly than explanations, which in turn shift faster than consensus.

Intrapersonal vs. Interpersonal Entropy

We generated convergence by turn distance graphs for all 90 conversations, sorted by conversational task and speaker. Specifically, intrapersonal entropy includes pairs where both utterances i and j are from the same speaker, while interpersonal entropy includes pairs where i and j come from different model "speakers." These trends are statistically significant.

Since each utterance is in a pair that is interpersonal (comparing to the conversation partner) or intrapersonal (comparing to an utterance by itself), there is a potentially inherent negative correlation between interpersonal and intrapersonal entropy when referencing past conversation. This is strongly evident in the consensus and explanation tasks. This suggests that in these tasks, Gemini references its partner in highly consistent ways. By identifying upward-facing parabolas in interpersonal entropy, we can pinpoint local minima, indicating the distances at which Gemini most often references the semantic content of its partner. This is supported by corresponding downward parabolas in intrapersonal entropy at the same distances. Specifically, Gemini most frequently references its partner around utterance distances of 5 and 15 for consensus, and around distance 8 for explanation. Although the exact cause cannot be isolated from this graph alone, it is intriguing to note the distinct differences when Gemini is asked to decide or explain a concept.

In contrast, debate tasks exhibit a strong positive correlation with few local effects. This aligns with human research suggesting that, in conflict, people tend to stick more firmly to their own line of reasoning (cf. Paxton & Dale, 2014).

However, Gemini displays a distinctly non-human behavior by referencing its partner's content early in its turn, which may reflect its assistive nature rather than human traits such as the desire to be consistent or ego-driven behavior. While some accounts emphasize the role of egocentric cognitive processes in conversation (Epley, Morewedge, & Keysar, 2004), it is important to note that such processes may operate at deeper levels of cognition and Gemini's divergence likely reflects more surface-level conversational dynamics.

Overall, Gemini exhibits notably distinct convergence entropy patterns that vary by conversational task. Further research is needed to identify the underlying causes of these intriguing observations.

Linear Discriminant Analysis

Choosing the first three PCA components that explain 83.4% of the variance as our features, we constructed our LDA models using leave-one-out cross-validation to estimate test error. This straightforward model performs reasonably well, achieving an overall accuracy of 63.33%, $p < 0.0000001$ (Table 2). While fine-tuning and further model exploration would likely improve performance, we opted to implement a simple, interpretable model, as our primary goal is to demonstrate that conversational tasks influence LLM-generated con-

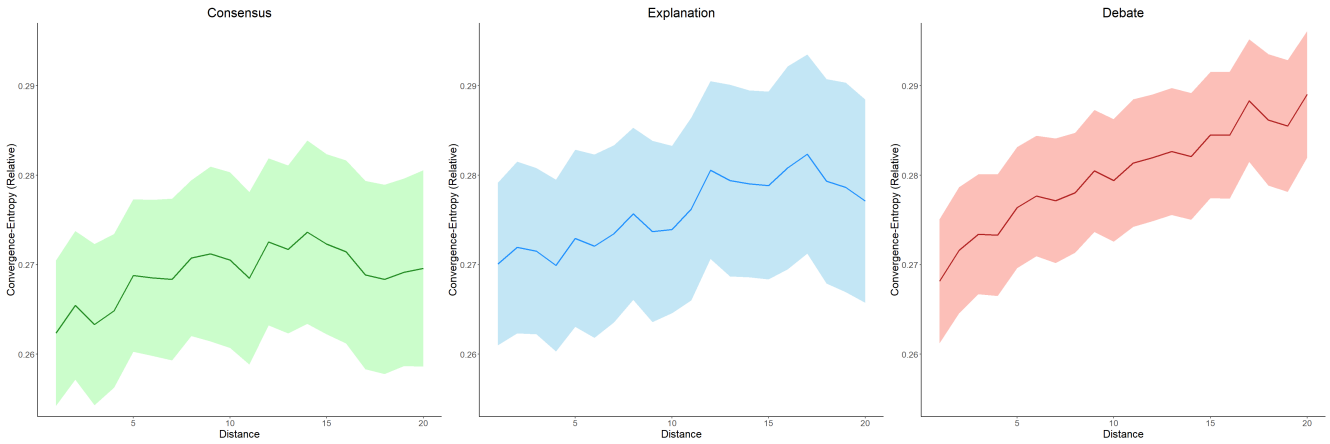

Figure 2: To visualize variance in the data, we overlay the region corresponding to a 95% confidence interval around the mean. As noted in prior studies, the absolute magnitude of convergence entropy is challenging to interpret (Rosen & Dale, 2023). However, when appropriately scaled, differences in how convergence evolves over time become apparent. Compared to the overall dataset (Figure 1), sports-related conversations exhibit significantly lower entropy, particularly in consensus and explanation tasks. This suggests that topic choice influences convergence entropy. In subsequent analyses, we generalize across all topics to mitigate these effects.

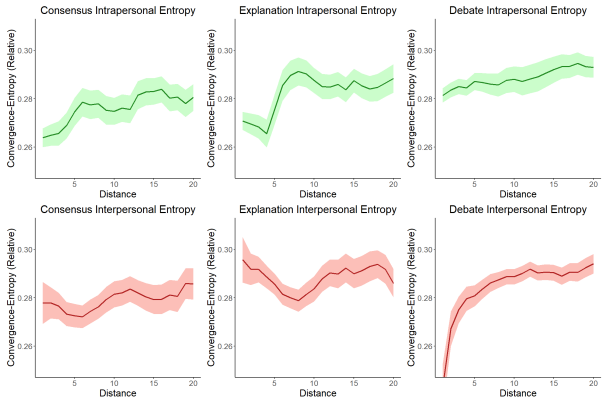

Figure 3: Interpersonal and intrapersonal entropy of Gemini 1.5 Flash.

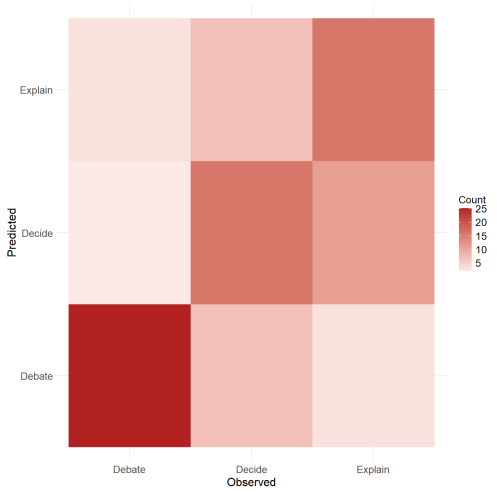

Figure 4: Confusion Matrix for LDA Model of Gemini.

tent. In simple terms, this result suggests that features derived from convergence of two LLMs can separate pragmatic goal of a conversation.

From the confusion matrix (Figure 4), it is evident that some conversational tasks are more easily differentiated than others. For example, debate is the most distinct task, with the model recognizing it with a balanced accuracy of 83.33%. This aligns with predictions, as debate is the only conversational task that involves conflict. In contrast, consensus and explanation are frequently confused with each other, as indicated by the darker shading in the upper quadrant of the matrix. Although the model can generally distinguish between conversational tasks, the difficulty in differentiating consensus from explanation supports our hypothesis.

Gemini vs. GPT

Compared to Gemini 1.5 Flash, GPT-4o mini exhibits a strongly assistive nature, as evidenced by the consistently low initial interpersonal entropy across all three conversational tasks (Figure 5). In contrast, Gemini 1.5 Flash displays distinct convergence entropy patterns for each task (Figure 3), suggesting that GPT-4o mini is less sensitive to variations in conversational tasks. This observation aligns with GPT-4o mini's tendency to always respond in an "Assistant" manner, offering concise answers followed by detailed reasoning, a behavior reinforced through its training (Ouyang et al., 2022).

This assistive and verbose style is also reflected in general statistical measures. GPT-4o mini generates nearly twice as

Table 2: Performance Metrics for LDA Model of Gemini.

Metric	Value
Accuracy	63.33% ± 0.32%
Kappa	0.45 ± 0.005
Sensitivity (Debate)	83.33% ± 0.25%
Sensitivity (Decide)	53.33% ± 0.33%
Sensitivity (Explain)	53.33% ± 0.33%
Balanced Accuracy (Debate)	83.33% ± 0.25%
Balanced Accuracy (Decide)	65.83% ± 0.31%
Balanced Accuracy (Explain)	68.33% ± 0.31%

many utterances as Gemini 1.5 Flash for the same prompts and turns. Its commitment to this style is further supported by the performance of our simple LDA model, which has lower accuracy in distinguishing between tasks (63.33% vs. 47.78% accuracy in Figure 3). Additionally, MANOVA tests show that while both models produce task-dependent content, the effect is significantly stronger for Gemini than for GPT (Approximate F statistic: 13.383 vs. 4.887 in Figure 3). Overall, these findings confirm that GPT-4o mini is less sensitive to conversational tasks, likely due to its training.

Conclusions

Our findings demonstrate that LLM-generated conversations exhibit measurable conversational dynamics, as captured by convergence-entropy metrics. There may be explanations for certain topic-based behaviors. However our analysis is focused on the effects of conversational tasks as our primary independent variable.

- Gemini 1.5 Flash and GPT-4o mini engage in distinct conversations dependent on the assigned conversational task.
- Segmenting the dataset by conversational task reveals unique patterns, suggesting that specific tasks induce distinguishable conversational features.
- By comparing the effects of conversational tasks and their induced features, we outline a quantitative approach for comparing LLM conversational behaviors.

Our methods may be of particular interest to two research communities: those studying LLM-generated text and those exploring agent-based models of human communication.

For the former, our quantitative framework can assess whether an LLM exhibits specific features, potentially contributing to the development of a standardized benchmark for evaluating conversational tendencies (cf. Hua & Artzi, 2024). Such a benchmark could help characterize models based on their ability to replicate human-like or task-specific behaviors. Specifically, we find that greater alignment separability across conversational tasks correlates with task awareness to some degree. Further research is needed to determine the extent to which this metric can serve as a reliable evaluation tool for LLMs.

Table 3: Gemini 1.5 Flash vs. GPT 4o-mini Metrics.

Metric	Gemini	GPT
Dataset Size	89,440	177,920
μ	0.2834 ± 0.0004	0.2827 ± 0.0002
σ^2	0.0046 ± 2.17e-5	0.0023 ± 7.71e-6
Approximate F	13.383	4.887
P-value of F	1.574 e^{-9}	9.287 e^{-4}
Accuracy	63.33% ± 0.32%	47.78% ± 0.23%
Kappa	0.45 ± 0.005	0.2167 ± 0.0035
Bal. Acc. (Debate)	83.33% ± 0.25%	52.50% ± 0.22%
Bal. Acc. (Decide)	65.83% ± 0.31%	60.00% ± 0.23%
Bal. Acc. (Explain)	68.33% ± 0.31%	70.00% ± 0.22%

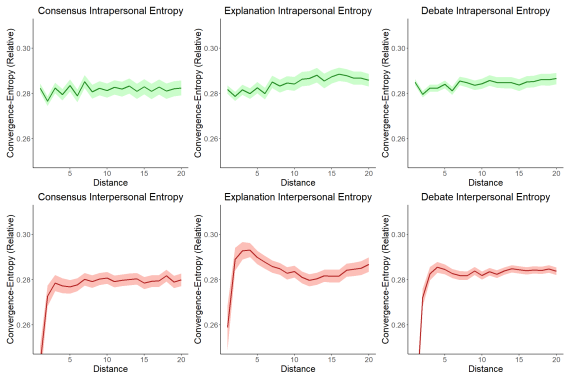

Figure 5: Interpersonal and intrapersonal entropy of GPT-4o mini. Downward parabola for Explanation interpersonal indicates that GPT-4o mini introduces the most distinct semantic content early in its reply. This aligns with its tendency to provide definitions initially and connect them to prior statements.

For the latter, our results highlight the potential of LLM-based simulations for studying human conversational behaviors, demonstrating that, while LLMs do not display a perfect match with human language use, they do exhibit a number of useful characteristics that can be leveraged in an idealized model of human communication.

This study is limited by its experimental design and the high-level nature of its convergence-entropy analysis. To confidently attribute specific behaviors to human analogs, future research should introduce additional controls or apply convergence-entropy metrics to established human conversation datasets, potentially by replicating the same task in human dyads for comparison. In either case, the results presented here provide an immediate baseline to which such work can be compared, and provide us with grounds to make predictions about human dialog on the basis of our results.

References

Becker, J. (2024). *Multi-agent large language models for conversational task-solving.* Retrieved from

https://arxiv.org/abs/2410.22932

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 610–623).

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, *22*(6), 1482.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. In *Psychology of learning and motivation* (Vol. 53, pp. 301–344). Elsevier.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. Retrieved 2021-05-18, from http://arxiv.org/abs/2005.14165 (arXiv: 2005.14165)

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics*, *50*(1), 293–350.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Condon, W. S., & Ogston, W. D. (1971). Speech and body motion synchrony of the speaker-hearer. *The perception of language*, *150*, 184.

Dale, R. (2016). The return of the chatbots. *Natural language engineering*, *22*(5), 811–817.

Dale, R., Fusaroli, R., Duran, N. D., & Richardson, D. C. (2013). The self-organization of human interaction. In *Psychology of learning and motivation* (Vol. 59, pp. 43–95). Elsevier.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. Retrieved 2021-03-31, from http://arxiv.org/abs/1810.04805

Donnarumma, F., Dindo, H., Iodice, P., & Pezzulo, G. (2017). You cannot speak and listen at the same time: A probabilistic model of turn-taking. *Biological cybernetics*, *111*, 165–183.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification* (2nd ed.). Wiley.

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of experimental social psychology*, *40*(6), 760–768.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: Quantifying the benefits of linguistic coordination. *Psychological science*, *23*(8), 931–939.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive science*, *40*(1), 145–171.

Giles, H., Ogay, T., et al. (2007). Communication accommodation theory. *Explaining communication: Contemporary theories and exemplars*, 293–310.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Pearson.

Hua, Y., & Artzi, Y. (2024). Talk less, interact better: Evaluating in-context conversational adaptation in multimodal llms. *arXiv preprint arXiv:2408.01417*.

Ivanova, A. A. (2025). How to evaluate the cognitive abilities of llms. *Nature Human Behaviour*, 1–4.

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). Springer.

Jones, C. R., & Bergen, B. K. (2023). Does gpt-4 pass the turing test? *arXiv preprint arXiv:2310.20216*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019a). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*. Retrieved from http://arxiv.org/abs/1907.11692

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019b, July). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv. Retrieved 2023-12-03, from http://arxiv.org/abs/1907.11692 (arXiv:1907.11692 [cs]) doi: 10.48550/arXiv.1907.11692

MacWhinney, B. (2017). Tools for analyzing talk part 1: The chat transcription format. *Carnegie.[Google Scholar]*, *16*.

Miao, G. Q., Dale, R., & Galati, A. (2023). (mis) align: a simple dynamic framework for modeling interpersonal coordination. *Scientific Reports*, *13*(1), 18325.

Ostrand, R., & Berger, S. E. (2024). Humans linguistically align to their conversational partners, and language models should too. In *ICML 2024 workshop on llms and cognition.*

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730–27744.

Paxton, A., & Dale, R. (2013). Argument disrupts interpersonal synchrony. *Quarterly Journal of Experimental Psychology*, *66*, 2092-2102. doi: 10.1080/17470218.2013.853089

Paxton, A., & Dale, R. (2014). Leveraging linguistic content and debater traits to predict debate outcomes. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Richardson, M. J., Dale, R., & Marsh, K. L. (2014). Complex dynamical systems in social and personality psychology. *Handbook of research methods in social and personality psychology*, *253*(10.1017).

Righi, G., Tenenbaum, E. J., McCormick, C., Blossom, M., Amso, D., & Sheinkopf, S. J. (2018). Sensitivity to audiovisual synchrony and its relation to language abilities in

children with and without asd. *Autism Research*, *11*(4), 645–653.

Rosen, Z. P., & Dale, R. (2023). BERTs of a feather: Studying inter- and intra-group communication via information theory and language models. *Behavior Research Methods*. doi: 10.3758/s13428-023-02267-2

Rosen, Z. P., & Dale, R. (2024). Llms don't "do things with words" but their lack of illocution can inform the study of human discourse. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... Wolf, T. (2023). *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.* arXiv. Retrieved from http://arxiv.org/abs/2211.05100 (arXiv:2211.05100 [cs]) doi: 10.48550/arXiv.2211.05100

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*.

Steel, J., & Togher, L. (2019). Social communication assessment after traumatic brain injury: A narrative review of innovations in pragmatic and discourse assessment methods. *Brain Injury*, *33*(1), 48–61.

Toma, C. L. (2014). Towards conceptual convergence: An examination of interpersonal adaptation. *Communication Quarterly*, *62*(2), 155–178.