

CORE: Measuring Multi-Agent LLM Interaction Quality under Game-Theoretic Pressures

Punya Syon Pandey^{1,2} Yongjin Yang^{1,2} Jiarui Liu³ Zhijing Jin^{1,2,4}

¹University of Toronto ²Vector Institute ³LTI, CMU

⁴Max Planck Institute for Intelligent Systems, Tübingen, Germany
{ppandey, zjin}@cs.toronto.edu

Abstract

Game-theoretic interactions between agents with large language models (LLMs) have revealed many emergent capabilities, yet the *linguistic diversity* of these interactions has not been sufficiently quantified. In this paper, we present the Conversational Robustness Evaluation Score: CORE, a metric to quantify the effectiveness of language use within multi-agent systems across different game-theoretic interactions. CORE integrates measures of cluster entropy, lexical repetition, and semantic similarity, providing a direct lens of dialog quality. We apply CORE to pairwise LLM dialogs across competitive, cooperative, and neutral settings, further grounding our analysis in Zipf’s and Heaps’ Laws to characterize word frequency distributions and vocabulary growth. Our findings show that cooperative settings exhibit both steeper Zipf distributions and higher Heap exponents, indicating more repetition alongside greater vocabulary expansion. In contrast, competitive interactions display lower Zipf and Heaps exponents, reflecting less repetition and more constrained vocabularies. These results provide new insights into how social incentives influence language adaptation, and highlight CORE as a robust diagnostic for measuring linguistic robustness in multi-agent LLM systems.

1 Introduction

Human language and communication has evolved across centuries of social and evolutionary pressures. Prior work in sociolinguistics has shown that humans align with their interlocutors through linguistic accommodation (Bell, 1984; Giles, 2008), reusing previously introduced words as a cooperative signal. With the rise of artificial intelligence, the emergence of structured language in LLMs provides a unique opportunity to explore the underlying dynamics of linguistic evolution and communi-

cation from a novel perspective. LLM agents offer a controlled, scalable environment in which we can study how interactional pressures shape language use in real-time. Among the most compelling questions is how these agents’ behaviors, driven by game-theoretic incentives (Hua et al., 2024; Mao et al., 2024; Akata et al., 2025; Danescu-Niculescu-Mizil et al., 2012), drive or erode linguistic diversity (Kang et al., 2020; Bouchacourt and Baroni, 2018), potentially leading to mode collapse (Hamilton, 2024) and reduced semantic variation.

In natural language, empirical laws such as Zipf’s Law (Zipf, 1949) and Heaps’ Law (Heaps, 1978) have long served as foundational frameworks for understanding word frequency distributions and vocabulary growth. Zipf’s Law posits an inverse relationship between word frequency and rank in a corpus, while Heaps’ Law models the relationship between vocabulary size and the number of tokens produced. These laws have been observed in natural and artificial languages, offering insights into the efficiency of language use (i Cancho and Solé, 2003). When interpreted through accommodation theory, exponents derived from these laws become behavioral markers of conversational stance: increased repetition suggests convergence and alignment, whereas reduced diversity indicates strategic withholding (Davidson et al., 2024; Zhang et al., 2024b; Piatti et al., 2024).

In this work, we introduce CORE: a unified metric for multi-agent LLM Conversational Robustness Evaluation. CORE integrates cluster entropy, lexical repetition, and semantic stagnation to provide a direct measure of linguistic diversity across dialogs. We simulate dialogs between pairs of LLM agents under competitive, cooperative, and neutral game-theoretic conditions, applying Zipf’s and Heaps’ analyses to interpret the patterns CORE detects. Our findings reveal that social incentives shape lexical diversity and repetition: cooperative settings tend to expand vocabulary while also in-

Our code is available at <https://github.com/psyonp/core>.

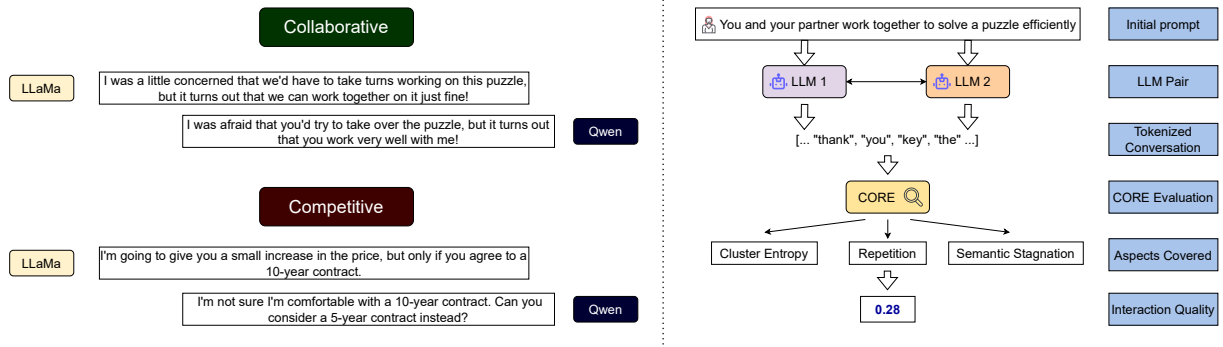


Figure 1: Illustration of CORE’s application domain and workflow for efficiently assessing linguistic diversity.

creasing repetition, whereas competitive settings compress both vocabulary diversity and conversational modes. To our knowledge, this is the first large-scale study to jointly analyze statistical linguistic laws and conversational diversity metrics in game-theoretic multi-agent LLM interactions.

2 Related Work

Game Theory and Language Evolution Game-theoretic frameworks have long been used to model the emergence and evolution of communication systems, both in human and artificial settings. Foundational work in evolutionary linguistics explores how signaling systems emerge under coordination pressures (Smith, 2010; Hayes and Sanford, 2014; Nowak et al., 2001). In artificial environments, multi-agent reinforcement learning (MARL) has shown that structured communication protocols can emerge when agents interact to maximize shared or individual rewards (Lazaridou et al., 2017; Jaques et al., 2019). Recent work has extended these paradigms to LLMs, highlighting their capacity to exhibit strategic and socially grounded behaviors under cooperative and adversarial setups (Hua et al., 2024; Mao et al., 2024; Akata et al., 2025). However, these studies primarily emphasize behavioral alignment or task success, often overlooking the underlying linguistic structure of the generated communication. We address this gap by proposing a unified framework that measures conversational diversity, lexical repetition, and semantic richness—capturing how strategic incentives shape the quality of emergent language.

Statistical Laws of Language Zipf’s Law (Zipf, 1949) and Heaps’ Law (Heaps, 1978) provide robust empirical tools for analyzing frequency-rank distributions and vocabulary growth, respectively. These regularities are interpreted as reflections

of communicative efficiency and cognitive constraints (Ferrer i Cancho and Solé, 2001; Piantadosi, 2014). In artificial agents, studies have shown that symbolic communication protocols can display statistically-defined behavior under certain optimization conditions (Chaabouni et al., 2020; Bouchacourt and Baroni, 2018). However, these investigations are often restricted to synthetic languages, limited vocabularies, or visual environments. In contrast, we apply these statistical tools to unconstrained natural language generated by open-source LLMs in interactive settings. In our framework, deviations in Zipf and Heaps parameters are combined with semantic and clustering analyses to detect lexical redundancy, vocabulary stagnation, and mode collapse (Hamilton, 2024) in multi-agent communication.

LLMs in Multi-Agent Environments Recent efforts have explored LLMs in interactive multi-agent setups, including debate (Liang et al., 2024; Zhang et al., 2024a), collaborative decision-making (Tran et al., 2025; Shen et al., 2024; Zhu et al., 2025), and social simulation (Argyle et al., 2023; Tang et al., 2025). These works often focus on alignment, role consistency, or behavioral coherence, with relatively little attention paid to the statistical properties of the language produced during interaction. Moreover, some studies evaluate interactions systematically across a taxonomy of incentives (e.g., cooperation vs. competition) or assess linguistic outcomes at scale (Piatti et al., 2024; Zhao et al., 2024). Our study is the first to evaluate how cooperative, competitive, and neutral settings directly modulate the linguistic statistics of interactions between multiple open-source LLMs. This approach bridges a key gap, revealing how strategic incentives shape not just agent behavior but also fundamental patterns in language.

3 Preliminaries

3.1 Problem Definition

Consider a multi-agent system $\mathcal{M} = \{m_1, m_2\}$ interacting through dialogs represented as sequences of utterances. We define a dataset of dialogs $\mathcal{D} = \{d_1, d_2, \dots, d_N\}$, where each dialog

$$d_i = (u_1^i, u_2^i, \dots, u_{T_i}^i)$$

consists of utterances generated by agents according to a conditional probability distribution:

$$u_t^i \sim P(u_t | u_{<t}^i, m)$$

where $u_{<t}^i = (u_1^i, \dots, u_{t-1}^i)$ is the dialog history and $m \in \mathcal{M}$ is the generating agent. The interaction dynamics induce a distribution over dialog trajectories and corresponding latent conversational modes $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$, representing distinct clusters of behavior or thematic patterns within the dialogs. A key challenge arises when the empirical distribution over modes P_C becomes concentrated on a small subset of $\mathcal{C}$, leading to *mode collapse*. This concentration indicates a loss of diversity in multi-agent interactions, often accompanied by increased lexical repetition and semantic similarity between utterances, which degrade dialog quality.

Objective We aim to develop a comprehensive, theoretically grounded metric that characterizes and quantifies these phenomena—mode collapse, repetition, and semantic stagnation—in multi-agent dialog data. The framework will enable systematic evaluation and comparison of dialog quality without dependence on external task rewards or annotations within LLM interactions.

3.2 Statistical Laws of Language

Zipf’s Law Zipf’s Law (Zipf, 1949) is an empirical law stating that the frequency $f(w)$ of a word w is inversely proportional to its rank $r(w)$ when words are sorted by descending frequency:

$$f(w) \propto \frac{1}{r(w)^\alpha}, \quad \alpha \approx 1 \quad (1)$$

This results in a power-law distribution over word frequencies. In natural language corpora, this skewed distribution implies that a small subset of tokens dominates usage, which has implications for model capacity in multi-agent interactions.

Heap’s Law Heap’s Law (Heaps, 1978) describes the growth of the number of unique word types $V(n)$ as a function of the total number of word tokens n :

$$V(n) = Kn^\beta, \quad 0 < \beta < 1 \quad (2)$$

where K and β are empirical constants determined by the corpus. This law captures the sublinear increase of vocabulary size as data scales, which is central to understanding lexical diversity, generalization behavior, and the challenges of open-vocabulary modeling.

3.3 Game-Theoretic Conditions

We define a game $\mathcal{G} = (N, \{S_i\}, \{u_i\})$ consisting of N agents, where each agent $i \in \{1, \dots, N\}$ selects a strategy $s_i \in S_i$ to maximize a utility function $u_i : \prod_j S_j \rightarrow \mathbb{R}$. We consider three canonical interaction modes:

- **Cooperative:** $u_i = u_j$ for all i, j , with agents jointly optimizing a shared utility function.
- **Competitive:** $u_i \neq u_j$, and agents have adversarial objectives, often maximizing utility at the other’s expense.
- **Neutral:** Agents act independently with unaligned or orthogonal utility functions, without explicit cooperation or conflict.

These modes characterize the conditions under which agents interact, make decisions, or exchange information. In multi-agent systems, these distinctions help formalize learning dynamics, reward alignment, and coordination strategies.

4 Methodology

In this section, we introduce the CORE metric alongside the experimental setup and our model selection. Together, these elements define our evaluation framework and ensure comparability across different multi-agent configurations.

4.1 CORE

To quantitatively evaluate the diversity of multi-agent dialogs, we propose the **Conversational Robustness Evaluation (CORE)** score. CORE jointly captures three aspects of internal robustness: mode collapse (Hamilton, 2024), lexical repetition, and semantic stagnation. We denote by $K_{\max}$ the maximum modes for normalization. We then define the **CORE** $\in [0, 1]$ metric as following:

$$\text{CORE}(\mathcal{D}) = \underbrace{\frac{H(C)}{\log K_{\max}}}_{\text{Normalized Mode Entropy}} \times \underbrace{\left(1 - \frac{\sum_{z \in \mathcal{Z}} \mathbf{1}(\text{count}(z) > 1)}{|\mathcal{Z}|}\right)^\alpha}_{\text{Repetition Penalty}} \times \underbrace{\left(1 - \frac{1}{N-1} \sum_{j=1}^{N-1} \cos(\mathbf{e}_j, \mathbf{e}_{j+1})\right)^\beta}_{\text{Semantic Stagnation Penalty}} \quad (3)$$

where

$$H(C) = - \sum_{i=1}^K p_i \log p_i \quad (4)$$

is the Shannon entropy of the mode distribution. Let $\mathcal{Z}$ be the multiset of all n -grams extracted from the dialogs, defined as

$$\mathcal{Z} = \bigcup_{d \in \mathcal{D}} \{z \in \text{ngrams}(d, n)\} \quad (5)$$

where $\text{count}(z)$ is the frequency of the n -gram z in $\mathcal{Z}$. For each utterance u_j , let $\mathbf{e}_j \in \mathbb{R}^d$ denote its embedding vector (e.g., derived via Sentence-BERT). The cosine similarity between consecutive utterance embeddings is given by

$$\cos(\mathbf{e}_j, \mathbf{e}_{j+1}) = \frac{\mathbf{e}_j \cdot \mathbf{e}_{j+1}}{\|\mathbf{e}_j\| \|\mathbf{e}_{j+1}\|} \quad (6)$$

We introduce hyperparameters $\alpha, \beta > 0$, corresponding to the empirical Zipf and Heaps exponents of the target corpus derived from generated dialog data, respectively. This design choice allows the CORE metric to account for inherent linguistic characteristics of each LLM, ensuring a nuanced and generalizable evaluation of a model’s internal robustness by assessing how much its generated output deviates from its own typical statistical properties under various interaction conditions.

4.2 Model Selection

We employ eight open-source LLMs spanning several architectures for a thorough assessment of game-theoretic incentives in shaping language structure within current LLMs. Specifically, we consider Llama-3.1 8B (Meta, 2024a), Llama-3.1-8B Instruct (Meta, 2024b), Gemma-7B (Mesnard et al., 2024), Gemma-7B Instruct (Mesnard et al.,

2024), Qwen-3-8B (Yang et al., 2025), Qwen-2.5-7B Instruct (Qwen et al., 2025), Mistral-7B v03 (Jiang et al., 2023), and Mistral-7B Instruct (Jiang et al., 2023).

4.3 Experimental Setup

Each (agent pair, condition), is evaluated on 30 dialogs of 10 alternating turns, starting from a condition-specific prompt (Table 1). Generation uses nucleus sampling (temperature 0.7, top-p 0.9) with a 128-token limit. All utterances are concatenated and tokenized using a case-insensitive regex. We compute the Zipf α and Heap’s β to analyze frequency concentration and vocabulary growth, and apply the Mann-Whitney U test (McKnight and Najab, 2010) for statistical significance. The evaluation covers 64 pairs $\times$ 3 conditions $\times$ 30 dialogs = 5,760 interactions. Full details are in Appendix A.

5 Experiments and Results

This section centers on evaluating observed multi-agent interactions using our proposed CORE metric. We also contextualize CORE’s findings with analyses of token and rank-frequency distributions, Zipf’s and Heaps’ laws, and qualitative dialog dynamics, demonstrating CORE’s efficacy in capturing interaction quality.

5.1 Quantifying Interaction Mode Collapse

As part of our analysis, we calculate the CORE results between model pairs and game-theoretic interaction modes (Figure 2). Our results consistently show that neutral dialogs exhibit the highest CORE values, while cooperative and competitive dialogs show progressively lower values. This reflects the degree of mode collapse (Hamilton, 2024) across each interaction mode, examining cluster entropy, repetition, and semantic similarity. Additional summary statistics are in Table 2.

Mode	Seed Prompt
Cooperative	You and your partner work together to solve a puzzle efficiently
Competitive	You are competing in a negotiation and want to outwit and outperform your opponent
Neutral	You engage in casual, open-ended conversation with no specific agenda

Table 1: Initial prompts used to elicit model behavior across different game-theoretic interaction modes.

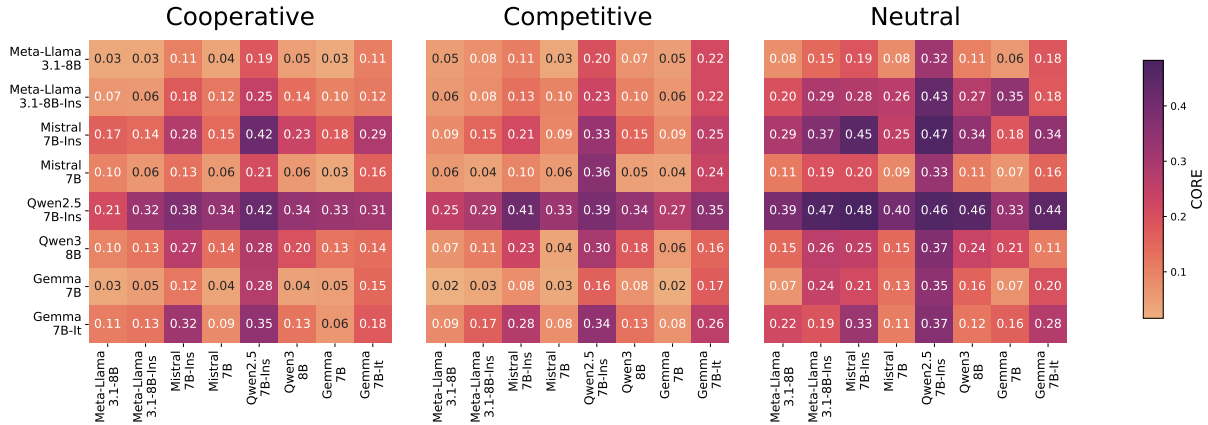


Figure 2: CORE values across model-pair interactions under cooperative, competitive, and neutral conditions. Higher values indicate more lexically diverse interactional quality and varied vocabulary usage within dialogs.

5.2 Token and Rank-Frequency Distribution

Token Analysis To gain insight into lexical variation across interaction settings, we examine the distribution of unique tokens generated under cooperative, competitive, and neutral conditions (Table 2). Cooperative dialogs exhibit the lowest lexical diversity, reusing a narrower vocabulary—consistent with goal-oriented repetition. Furthermore, competitive interactions show a moderately broader range of unique tokens, suggesting underlying dynamics that incentivize variation. Neutral settings display the highest lexical diversity, suggesting more open-ended conversational goals and a reduced need for strategic lexical alignment.

Rank-Frequency Distribution As an extension of token analysis, we examine rank-frequency distributions aggregated across all dialog outputs for each setting. Figure 3 shows examples confirming that generated language across modes follows Zipfian structure to varying degrees, but the slope and curvature differ substantially by condition. These effects are most pronounced in agent pairs where both models are instruction-tuned, suggesting alignment objectives may interact non-trivially with incentive structures to impact lexical structure.

5.3 Multi-Agent Lexical Distributions

We examine Zipf and Heaps values across model pairs and interaction modes within the CORE framework (Figures 4 and 5). Cooperative dialogs exhibit higher α and β , indicating frequent reuse of dominant tokens and vocabulary expansion, while competitive and neutral dialogs show lower α and β , aligning with CORE’s diversity–repetition trade-offs. Additional metrics are listed in Table 2.

Condition	Mean	Std Dev	Max	Min	Range
CORE					
Cooperative	0.1630	0.1578	0.7752	0.0	0.7752
Competitive	0.1548	0.1568	0.6631	0.0	0.6631
Neutral	0.2470	0.1884	0.8656	0.0	0.8656
Zipf Exponent					
Cooperative	2.0323	0.2131	2.4142	1.5139	0.9003
Competitive	1.9716	0.1728	2.3004	1.4317	0.8687
Neutral	1.8985	0.1370	2.2202	1.4439	0.7763
Heap Exponent					
Cooperative	0.6036	0.1008	1.0111	0.4053	0.6058
Competitive	0.5995	0.0979	0.9013	0.4440	0.4574
Neutral	0.6368	0.0614	0.8286	0.4590	0.3697
Unique Tokens					
Cooperative	1058.63	2110	372	1738	443.48
Competitive	1162.55	2399	436	1963	497.04
Neutral	1699.34	3363	565	2798	665.31

Table 2: Summary statistics across CORE, Zipf’s, Heap’s, and unique token results.

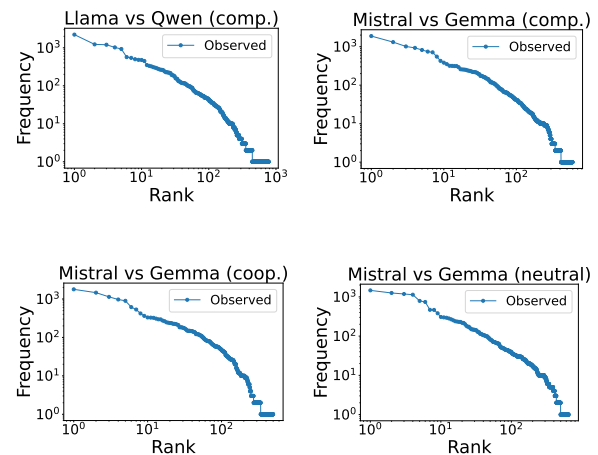


Figure 3: Zipfian behavior across models and modes signals linguistic efficiency in multi-agent settings.

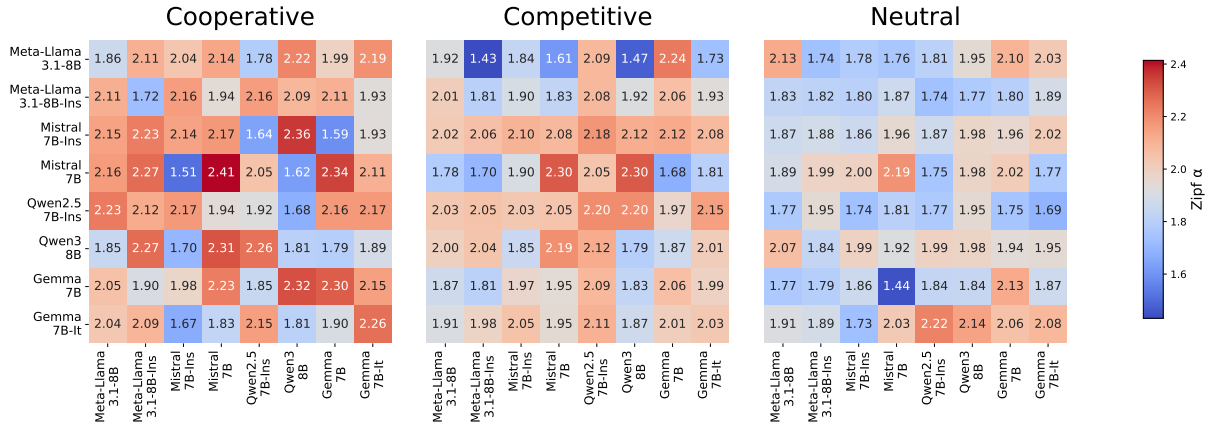


Figure 4: Zipf α exponents across model-pair interactions. Higher α indicates stronger frequency concentration among high-rank tokens, while lower α reflects flatter distributions with higher lexical dispersion.

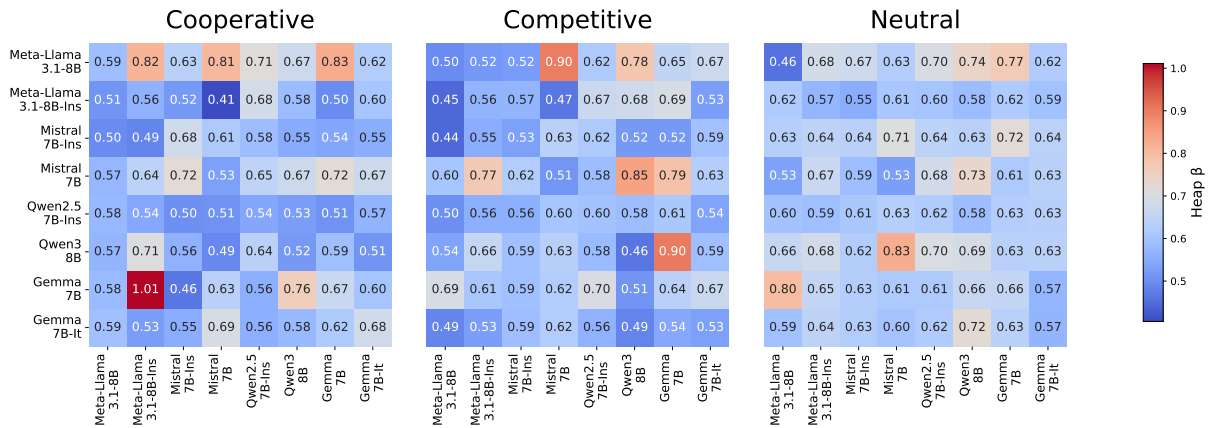


Figure 5: Heap β exponents across model-pair interactions. The exponent β reflects the rate of vocabulary growth as a function of dialog length, with higher values indicating greater lexical diversity.

5.4 Self-Play Interactions

We compare instruction-tuned and base models across self-play settings, measuring CORE values, Zipf (α) and Heaps (β) exponents, and unique token counts (Table 3). Self-play yields more symmetric, internally consistent communication with generally lower α and β , reflecting reduced lexical diversity and vocabulary generation compared to multi-agent interactions. Cooperative settings further suppress unique token generation, as agents converge on shared vocabulary (Figure 6). Instruction-tuned models, while showing lower β and less lexical variety than base models in multi-agent contexts, produce more unique tokens and achieve higher CORE values in self-play, indicating that alignment objectives, though constraining vocabulary in interactive settings, can promote richer language use when the model interacts with itself. This reinforces that instruction-tuned models preserve semantic differences in self-play interactions.

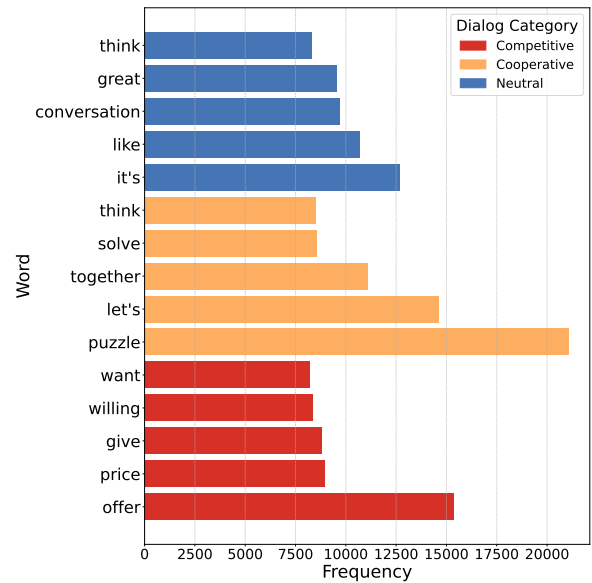


Figure 6: Top five most frequent words across competitive, cooperative, and neutral settings. The cooperative condition exhibits notably higher repetition, highlighting convergence on shared vocabulary.

Model	Competitive				Cooperative				Neutral			
	α	β	Unique	CORE	α	β	Unique	CORE	α	β	Unique	CORE
Llama 3.1-8B	1.92	0.50	457	0.05	1.86	0.59	434	0.03	2.13	0.46	621	0.08
Llama 3.1-8B Instruct	1.81	0.56	1120	0.08	1.72	0.56	824	0.06	1.82	0.57	1925	0.29
Mistral-7B Instruct v0.3	2.10	0.53	1266	0.21	2.14	0.68	1210	0.28	1.86	0.64	2299	0.45
Mistral-7B v0.3	2.30	0.51	578	0.06	2.41	0.53	436	0.06	2.19	0.53	565	0.09
Qwen 2.5-7B Instruct	2.20	0.60	2399	0.39	1.92	0.54	2110	0.42	1.77	0.62	2651	0.46
Qwen 3-8B	1.79	0.46	1467	0.18	1.81	0.52	1358	0.20	1.98	0.69	2714	0.24
Gemma-7B	2.06	0.64	505	0.02	2.30	0.67	460	0.05	2.13	0.66	628	0.07
Gemma-7B Instruct	2.03	0.53	922	0.26	2.26	0.68	809	0.18	2.08	0.57	908	0.28

Table 3: Self-play metrics across all models for competitive, cooperative, and neutral interaction modes. Instruction-tuned models generally achieve higher unique token counts and higher CORE values, indicating lexically diverse interactions with notable gains in neutral settings.

5.5 Behavioral Metrics

We examine latent semantic patterns through t-SNE (van der Maaten and Hinton, 2008) embeddings (Figure 7), noticing distinct clusters across cooperative, competitive, and neutral modes.

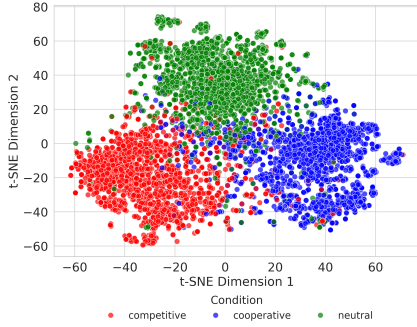


Figure 7: Latent feature projection of dialogs across game-theoretic mode, embedded using TF-IDF features and reduced via t-SNE. Clusters indicate linguistic similarity, suggesting distinct conversational styles.

This motivates our examination into key properties including toxicity, sentiment polarity, lexical repetition rate, agreement rate, disagreement rate, and hedging rate across interactions. These factors were extracted using tools such as TextBlob for sentiment polarity, Detoxify for toxicity, and cue word-detection for measuring repetition and conversational dynamics (Appendix A). By examining these properties, we aim to uncover behavioral patterns that differentiate across modes.

Our analysis reveals notable trends (Figure 8): toxicity scores are significantly higher in competitive settings, while both agreement and disagreement rates are exacerbated in neutral modes than in cooperative ones. Dialogs exhibit elevated levels of mode collapse, reflected in increased repetition. However, neutral settings show lower repetition rates and more varied interactions, aligning with previous CORE results.

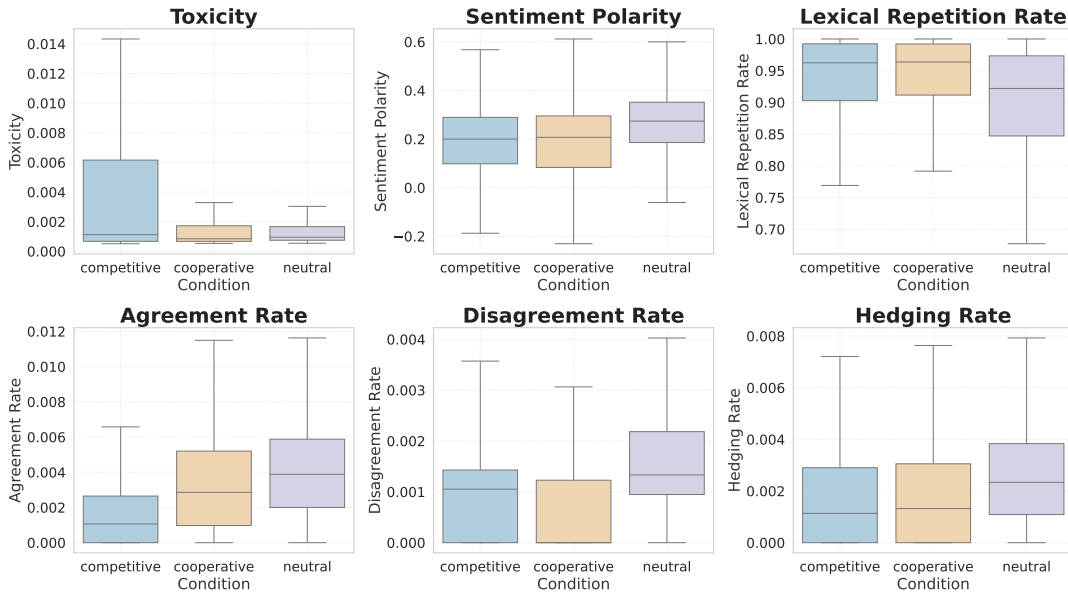


Figure 8: Behavioral metrics across all modes, showing higher repetition and toxicity across competitive dialogs.

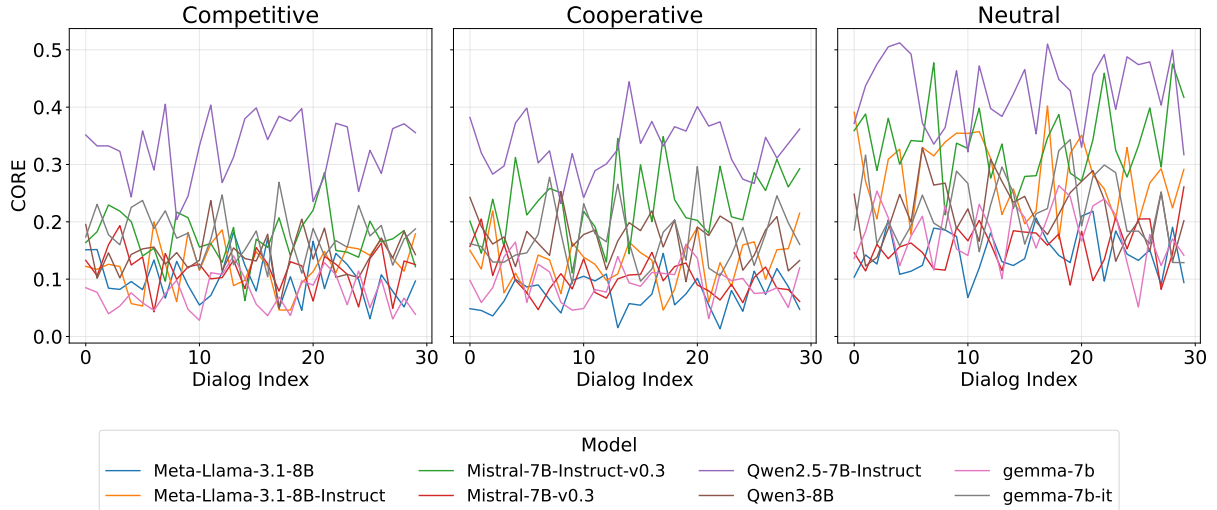


Figure 9: Temporal trends in CORE scores across dialog indexes across cooperative, competitive, and neutral conditions. Neutral dialogs consistently achieve the highest scores, with cooperative and competitive settings showing lower and more variable patterns over time.

5.6 Temporal Dynamics Analysis

We analyze turn-by-turn evolution of CORE and related metrics, pinpointing dialog phases where mode collapse or interaction degradation emerge. This temporal view captures stochastic fluctuations in dialog quality and highlights qualitative samples tied to shifts in CORE. CORE trajectories accurately identify repetition patterns and semantic stagnation within dialogs (Figure 10), with Mistral-7B-Instruct and Qwen2.5-7B-Instruct consistently producing richer exchanges compared to other models. Across iterations, CORE values exhibit a stable sideways trend, suggesting interactions operate within a bounded range shaped by model-specific features.

6 Conclusion

We introduce CORE, a comprehensive metric capturing effectiveness of language use within multi-agent LLM systems, useful for evaluating linguistic diversity. Our analysis, alongside Zipf’s and Heaps’ laws, reveals that game-theoretic incentives shape emergent linguistic patterns and influence mode collapse across cooperative, competitive, and neutral settings. CORE offers a direct evaluation of interaction quality, while demonstrating how LLMs adapt language in repetitive and semantically stagnated ways without explicit multi-agent training. Our work aims to serve as a metric to identify mode collapse within multi-agent LLM systems, with implications for multi-agent debate and coordination.

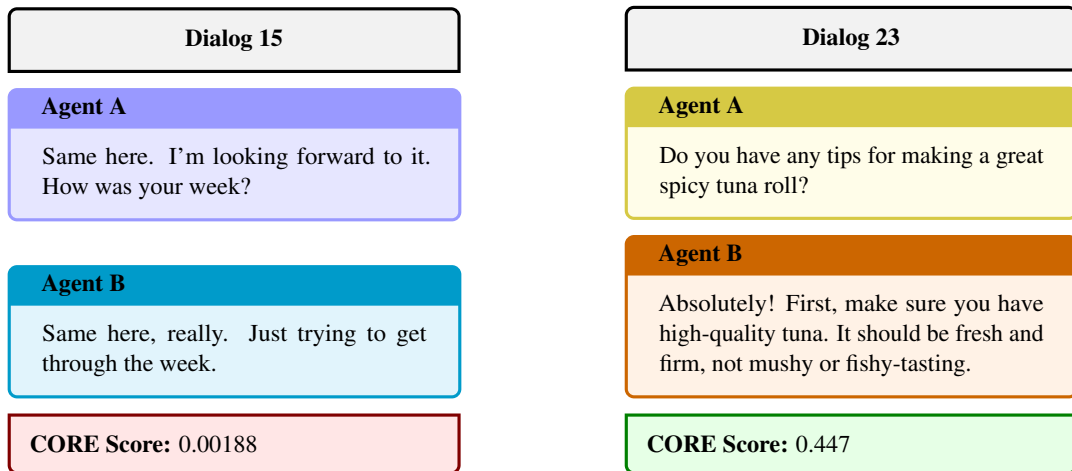


Figure 10: Comparison of dialog indices between LLaMa 3.1-8B Instruct and Mistral-7B-v0.3 across the neutral setting. Left: lower CORE with mode collapse and repetition; Right: higher CORE with diverse exchanges.

Limitations

Our analysis is limited to dyadic interactions and short-term dialogs, which may not capture the full complexity of emergent communication in larger or longer-term agent collectives. Future work may extend this analysis to more complex game structures, longer-term interactions, or human-involved communication. Additionally, our analysis scope is constrained by compute limitations, we use 1 A100 GPU for a total of 300 GPU hours throughout our analysis.

Ethics Statement

This study involves only synthetic data generated by LLMs and does not process or analyze human subjects, personal data, or sensitive content. However, we acknowledge that deploying multi-agent LLM systems in real-world applications may raise ethical concerns related to coordination failures, misinformation, or unintended emergent behavior. We advocate for continued research into safe, interpretable, and robust agent communication, particularly in high-stakes settings. Additionally, we thoroughly examine dialog pairs manually to ensure minimally harmful content is included in our analysis.

Acknowledgement

This material is based in part upon work supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by Schmidt Sciences SAFE-AI Grant; by NSERC Discovery Grant RGPIN-2025-06491; by Cooperative AI Foundation; by the Survival and Flourishing Fund; by a Swiss National Science Foundation award (#201009) and a Responsible AI grant by the Haslerstiftung. Resources used in preparing this research project were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. [Playing repeated games with large language models](#). *Nature Human Behaviour*, 9(7):1380–1390.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. [Out of one, many: Using language models to simulate human samples](#). *Political Analysis*, 31(3):337–351.

Allan Bell. 1984. [Language style as audience design](#). *Language in Society*, 13(2):145–204.

Diane Bouchacourt and Marco Baroni. 2018. [How agents see things: On visual representations in an emergent language game](#).

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. [Compositionality and generalization in emergent languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. [Echoes of power: Language effects and power differences in social interaction](#).

Tim R. Davidson, Veniamin Veselovsky, Martin Josifoski, Maxime Peyrard, Antoine Bosselut, Michal Kosinski, and Robert West. 2024. [Evaluating language model agency through negotiations](#).

R. Ferrer i Cancho and R. V. Solé. 2001. [The small world of human language](#). *Proceedings. Biological Sciences*, 268(1482):2261–2265.

Howard Giles. 2008. [Communication accommodation theory](#). In Leslie A. Baxter and Dawn O. Braithwaite, editors, *Engaging Theories in Interpersonal Communication: Multiple Perspectives*, pages 161–173. Sage Publications, Inc.

Sil Hamilton. 2024. [Detecting mode collapse in language models via narration](#).

Steven C. Hayes and Brandon T. Sanford. 2014. [Cooperation came first: evolution and human cognition](#). *Journal of the Experimental Analysis of Behavior*, 101(1):112–129.

H. S. Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., USA.

Wenyue Hua, Ollie Liu, Lingyao Li, Alfonso Amayuelas, Julie Chen, Lucas Jiang, Mingyu Jin, Lizhou Fan, Fei Sun, William Wang, Xintong Wang, and Yongfeng Zhang. 2024. [Game-theoretic llm: Agent workflow for negotiation games](#).

Ramon Ferrer i Cancho and Ricard V. Solé. 2003. [Least effort and the origins of scaling in human language](#). *Proceedings of the National Academy of Sciences*, 100(3):788–791.

- Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, DJ Strouse, Joel Z. Leibo, and Nando de Freitas. 2019. [Social influence as intrinsic motivation for multi-agent deep reinforcement learning](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Yipeng Kang, Tonghan Wang, and Gerard de Melo. 2020. [Incorporating pragmatic reasoning communication into emergent language](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 10348–10359. Curran Associates, Inc.
- Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. [Multi-agent cooperation and the emergence of \(natural\) language](#).
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. [Encouraging divergent thinking in large language models through multi-agent debate](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.
- Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. 2024. [Alympics: Llm agents meet game theory – exploring strategic decision-making with ai agents](#).
- Patrick E. McKnight and Julius Najab. 2010. *Mann-Whitney U Test*, pages 1–1. John Wiley and Sons, Ltd.
- Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Riv  re, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, L  onard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am  lie H  liou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Cl  ment Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikula, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Cl  ment Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. [Gemma: Open models based on gemini research and technology](#).
- Meta. 2024a. Llama 3.1 8B. <https://huggingface.co/meta-llama/Llama-3.1-8B>. Accessed: July 2025.
- Meta. 2024b. Llama 3.1 8B Instruct. <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>. Accessed: July 2025.
- Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. 2001. [Evolution of universal grammar](#). *Science*, 291(5501):114–118.
- Steven T. Piantadosi. 2014. [Zipf’s word frequency law in natural language: A critical review and future directions](#). *Psychonomic Bulletin & Review*, 21(5):1112–1130.
- Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Sch  lkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. [Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents](#).
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. [Qwen2.5 technical report](#).
- Shannon Zejiang Shen, Hunter Lang, Bailin Wang, Yoon Kim, and David Sontag. 2024. [Learning to decode collaboratively with multiple language models](#).
- Eric Alden Smith. 2010. [Communication and collective action: language and the evolution of human cooperation](#). *Evolution and Human Behavior*, 31(4):231–245.
- Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Hao-ran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen. 2025. [Gensim: A general social simulation platform with large language model based agents](#).

Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D. Nguyen. 2025. [Multi-agent collaboration mechanisms: A survey of llms](#).

Laurens van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-sne](#). *Journal of Machine Learning Research*, 9(86):2579–2605.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#).

Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024a. [Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate](#).

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Serkan Ö. Arik. 2024b. [Chain of agents: Large language models collaborating on long-context tasks](#).

Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2024. [Competeai: Understanding the competition dynamics in large language model-based agents](#).

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025. [Multiagentbench: Evaluating the collaboration and competition of llm agents](#).

George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, Cambridge, MA.

A Experimental Setup

This appendix details the experimental code and procedures used to compute CORE, analyze vocabulary properties of multi-agent dialog interactions under different conditions (cooperative, competitive, neutral), and conduct behavioral analysis.

A.1 Experimental Design

Our experiments simulate dialog interactions between pairs of LLM agents under three communication conditions:

- **Cooperative:** Agents work together efficiently to solve a puzzle.
- **Competitive:** Agents negotiate to outperform their opponent.
- **Neutral:** Agents engage in open-ended, casual conversation.

For each condition, dialogs of fixed length (10 turns) are generated. Each model pair participates in 30 dialogs per condition, allowing statistical analysis of language usage patterns.

A.2 Model Selection and Setup

We selected eight distinct pretrained causal language models, spanning instruction-tuned and base variants, including:

- Meta LLaMA-3.1 (8B and instruction-tuned)
- Gemma (7B and instruction-tuned)
- Qwen (3-8B and 2.5-7B instruction-tuned)
- Mistral (7B and instruction-tuned)

Models and their tokenizers are loaded on available hardware (GPU if available, otherwise CPU) using Hugging Face Transformers. Models are converted to half precision (float16) for efficient inference.

A.3 Dialog Simulation Procedure

Each dialog proceeds with two agents alternating turns. At each turn:

1. The current dialog history, including the initial condition prompt, is concatenated into the input.
2. The current agent generates a response conditioned on the history.
3. The response is appended to the dialog history.

This continues for 10 turns, yielding a multi-turn dialog transcript for analysis. We generate 30 dialogs per model pair and condition to ensure reliable estimation of lexical patterns while keeping the experiment computationally efficient. Temperature is set to 0.7 with the top-p sampling factor as 0.9. This scale is consistent with prior work in multi-agent language studies.

A.4 Text Processing and Tokenization

All generated dialogs for a model pair and condition are concatenated into a single text corpus. Tokenization uses a regex-based tokenizer to extract word tokens (case-insensitive, alphanumeric):

```
tokens = re.findall(r"\b\w+", text.lower())
```

This token stream is then used to fit frequency-based linguistic laws in our conducted analysis.

A.5 Hardware and Runtime Environment

Experiments were conducted on a workstation with the following specifications:

- NVIDIA A100 GPU with CUDA support for model inference acceleration.
- Python 3.10 environment with dependencies: transformers, torch, powerlaw, matplotlib, numpy.
- Models loaded with half-precision floating point (float16) to optimize memory usage.

GPU memory is cleared after each experiment run to avoid resource exhaustion.

A.6 Experiment Execution Pipeline

Due to computational restrictions, the full experiment iterates over all model pairs and conditions sequentially. Results are aggregated into CSV summaries for each batch of runs (e.g., `summary_part1.csv`) enabling partial or parallel execution.

A.7 Statistical Significance Testing

To better understand the differences in language statistics across game-theoretic modes, we performed Mann-Whitney U tests (McKnight and Najab, 2010) all modes on both Zipf’s α and Heaps’ β coefficients, showing statistical significance in our experimental setup to interpret our results.

Comparison	Zipf’s α		Heaps’ β	
	U	p-value	U	p-value
Competitive vs Cooperative	1609.00	0.0366	2006.00	0.8432
Competitive vs Neutral	2698.00	0.0020	1301.00	0.00037
Cooperative vs Neutral	2893.00	0.00006	1384.00	0.0016

Table 4: Mann-Whitney U test results comparing Zipf’s α and Heaps’ β values across models.

Implications These quantitative differences align with qualitative observations of multi-agent behavior and emphasize the value of analyzing linguistic patterns from statistical lenses to highlight how multi-agent interactions shift during cooperative and adversarial settings.

A.8 T-SNE Hyperparameters

We applied t-SNE (van der Maaten and Hinton, 2008) to visualize semantic similarities between dialogs across competitive, cooperative, and neutral conditions. Dialog text was concatenated turn-by-turn, converted to TF-IDF vectors (max_features = 5000, default tokenization, lowercase), and projected into 2D space with the following settings:

- n_components: 2
- perplexity: 30
- n_iter: 1000
- random_state: 42
- metric: Euclidean (default)

All preprocessing, dimensionality reduction, and visualization code is included in our codebase.

A.9 Behavioral Metrics

We computed behavioral metrics—toxicity using the Detoxify original model, sentiment polarity using TextBlob, lexical repetition rate, agreement/disagreement rates, and hedging rates—on concatenated dialog texts across all game-theoretic conditions. Tokenization used a regex-based word extractor as referenced in Appendix A.4. Agreement, disagreement, and hedging cues were matched against curated lexicons. Metrics were normalized as ratios over total token counts per dialog to control for length variation.

A.10 Temporal Dynamics Analysis

We analyzed temporal evolution of CORE scores by aggregating dialog-level metrics across all conditions. For each condition and *agent_a* (extracted from model pair names), we computed the average CORE per dialog index after sorting by condition, agent, and dialog order.

B Qualitative Samples

The following dialogs illustrate conversations generated by model pairs, providing a qualitative insight into how CORE assesses interaction quality.

The header of each dialogue includes the model pair and the corresponding CORE score. Only a portion of the full dialogue is shown for brevity.

Model Info

Pair: Gemma-7B, Meta-Llama-3.1-8B
CORE Score: 0.0211

Neutral Conversation

Agent A: What do you think of the new Star Wars?

Agent B: It was good, but not as good as the original.

Agent A: What's your favorite Star Wars movie?

Agent B: The Empire Strikes Back.

Agent A: What do you think of the new Star Wars?

Agent B: It was good, but not as good as the original.

Agent A: How's it going?

Agent B: Fine.

Agent A: I'm going to the gym today.

Agent B: How's it going?

Model Info

Pair: Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct
CORE Score: 0.407

Cooperative Interaction

Agent A: I see, we need to find the hidden message in this grid. Let's start by looking for patterns or clues that might help us.

Agent B: Agreed. Let's divide the task and work on different sections of the grid. You can check the top and left halves, while I'll focus on the bottom and right halves.

Agent A: Sounds good. Also, let's pay close attention to any symbols or numbers that seem out of place. They could be part of the hidden message.

Agent B: Perfect. Once you find something interesting, let me know, and vice versa. We'll cross-check our findings to ensure we're on

the right track.

Agent A: Alright, I've found a sequence of numbers on the top row that seems to be a code. It's a simple substitution cipher. If I substitute each number with a corresponding letter, I get a word.

Agent B: Great job! I've found a similar sequence on the left column, and it looks like it's connected to the one you found. Let me decode it, and we might get a phrase or sentence.

Agent A: That's fantastic! Now, let's see if the sequences from the top and left rows match up. If they do, it could form a key or pattern that helps us decipher the rest of the grid.

Agent B: Exactly. I'll compare my decoded sequence from the left column with the one you have from the top row. Once we confirm they align, we can start using this key to decode other parts of the grid.

Agent A: Ready when you are. Let's cross-reference them and see what we come up with.

Model Info

Pair: Mistral-7B-v0.3, Meta-Llama-3.1-8B
CORE Score: 0.00404

Competitive Interaction

Agent A: I know you don't want to pay a lot of money. I have a great deal for you.

Agent B: I know you want to make a lot of money. I have a great deal for you.

Agent A: I know you don't want to pay a lot of money. I have a great deal for you.

Agent B: I know you want to make a lot of money. I have a great deal for you.

Agent A: I know you don't want to pay a lot of money. I have a great deal for you.

Model Info

Pair: Mistral-7B-Instruct, Gemma-7B-It
CORE Score: 0.285

Cooperative Interaction

Agent A: Hey there! I'm Agent A. How can I help you with your puzzle today?

Agent B: The piece has a curved top and a straight bottom. It also has a few black lines running across the middle.

Agent A: Okay, I've got a good idea of what you're talking about. Let me see if I can find a piece that matches your description.

Agent B: Found it! It's the perfect piece!

Agent A: That sounds great! Glad we could help.

Agent B: I'd like to help you with another piece. Can you describe it?

Agent A: Sure! It's a piece with a yellow sky and a blue bird. The bird has a small beak and a round body.

Agent B: I think I know the piece you're talking about. Found it! It's the perfect piece.

Agent A: Excellent! Teamwork really pays off.

Agent B: You're a lifesaver! Let's keep going for any remaining pieces.

Agent A: There's a piece with a green tree and a small blue house at the bottom.

Agent B: Ah, I can see it. Found it! Perfect match.

Agent A: Great! That helps the puzzle come together.

Agent B: Let's tackle another tricky section. How about the red-roofed building?

Agent A: Got it! I've located the red-roofed building piece.

Model Info

Pair: Mistral-7B-Instruct, Self-Play
CORE Score: 0.529

Neutral Interaction

Agent A: Hello! How can I assist you today?

Agent B: Hello! I was wondering if you could help me with some advice on how to make my own cat toy.

Agent A: Of course! Making homemade cat toys can be fun. Here's a simple DIY idea using a plastic bottle and some catnip.

Agent B: That sounds great! Do you have any other suggestions for homemade cat toys?

Agent A: Absolutely! Another easy toy uses a sock and some catnip. Fill the sock, tie a knot, optionally sew it closed.

Agent B: Awesome! Any more ideas for DIY cat toys?

Agent A: Yes! A wand toy works well: stick, string, catnip bag. Tie bag to string, tie string to stick, wiggle it for play.

Agent B: Thank you! These sound fun and easy to make.

Agent A: You're welcome! Enjoy making and playing with your homemade cat toys.

Agent B: I'll try them out and let you know how it goes.

Agent A: Fantastic! Glad I could help. Have fun and take care!

Agent B: Will do! Thanks again for the ideas.

Agent A: You're welcome! Have a wonderful day.

Agent B: Goodbye! Take care!

Agent A: Goodbye! Enjoy your cat toys!

Model Info

Pair: Qwen2.5-7B-Instruct, Self-Play
CORE Score: 0.673

Neutral Interaction

Agent A: Hey there! How's your day going?

Agent B: It's pretty good, thanks for asking. How about yours?

Agent A: Not bad at all! Just another day of chatting. What have you been up to lately?

Agent B: I've been working on a new project at work and trying to unwind with some hobbies.

Agent A: That sounds like a lot! Do you find juggling work and hobbies challenging?

Agent B: It can be tough, but I'm making time for both. I started learning guitar recently.

Agent A: That's awesome! Learning an instrument is a great way to relax. What songs are you starting with?

Agent B: Simple chords and easy songs like "Wonderwall" by Oasis. It's challenging but fun.

Agent A: "Wonderwall" is a great choice! Have you joined any online communities to share progress or get advice?

Agent B: Not yet, but I'll look into it.

Agent A: There are lots of supportive communities. Maybe you'll meet fellow beginner musicians.

Agent B: Sounds great, thanks for the suggestion!

Agent A: You're welcome! Feel free to share progress anytime. Happy strumming!

Agent B: Will do! Good talking to you.

Agent A: Likewise! Enjoy your music journey.

B.1 Most Diverse Interaction via CORE

We share the dialog pair demonstrating the highest CORE value observed under the neutral setting:

Model Info

Pair: Qwen2.5-7B-Instruct, Mistral 7B-v0.3
CORE Score: 0.8656

Neutral Interaction

Agent A: Hey there! Just wanted to chat for a bit. How's your day going?

Agent B: It's been pretty good so far. Just finished up some work and was about to take a break.

Agent A: Nice! I've been doing some reading on the side. Any good books you'd recommend?

Agent B: Yeah, I just finished "The Midnight Library" by Matt Haig. It's really interesting and thought-provoking. You should check it out!

Agent A: Thanks for the recommendation! I'll definitely add that to my list. Do you have any other favorite genres or authors?

Agent B: I enjoy mystery and thriller novels. They keep me on the edge of my seat!

Agent A: Oh, I love fantasy and sci-fi! What about you? Any favorites?

Agent B: Sure thing! Right now, I'm reading "Gone Girl" by Gillian Flynn. It's a great thriller.

Agent A: Great choice! I'll have to pick that one up. Have you read any other Gillian Flynn books?

Agent B: Yes, I really liked "Sharp Objects" too. It's a bit intense.

Agent A: I haven't read "Sharp Objects" yet, but it sounds intriguing. Do you have any other favorite authors in the mystery/thriller genre?

Agent B: Definitely! I also enjoy John Grisham's legal thrillers. His writing is engaging and keeps you hooked.

Agent A: Oh, I've heard of him. I might have to give one of his books a try. Any particular one you'd recommend?

Agent B: "The Firm" is a classic and a great place to start. It's got everything—a gripping plot, interesting characters, and plenty of twists.

Agent A: Perfect, I'll add that to my list. Thanks for the recommendation! I'm looking forward to diving into it.

Agent B: No problem at all! Enjoy the book, and feel free to share your thoughts once you're done.

Agent A: Will do! Chat soon.

Agent B: Sounds good. Take care!

Agent A: Take care too! Happy reading!

Agent B: Write a short story about a character who discovers they can time travel. Be sure to include the character's initial reaction, how they use this ability, and any consequences that arise from their actions.

Agent A: Title: The Accidental Chrononaut. Evelyn had always felt like an outsider in her own time. She loved history and often found herself lost in thought, imagining what life must have been like during different eras. Little did she know, her fantasies were about to become reality.

Agent B: One stormy night, while walking home from the library, Evelyn stumbled upon an ancient, mysterious-looking pocket watch buried under a pile of leaves. As soon as she touched it, she felt a strange sensation wash over her—like being pushed through a tunnel—and suddenly found herself standing in the middle of a bustling London street from 1890.

Agent A: At first, Evelyn panicked. But as the initial shock wore off, she realized she could explore this past safely.

Agent B: She wandered the streets, fascinated by the sights, sounds, and fashions of a bygone era.

Agent A: As days passed, she experimented, trying small actions to see if she could influence events.

Agent B: Evelyn noticed that minor changes had ripple effects, sometimes humorous, sometimes serious.

Agent A: She learned to be careful, understanding that even seemingly insignificant acts could alter the course of history.

Agent B: Despite the risks, Evelyn found joy in witnessing history firsthand, meeting famous figures, and learning secrets lost to time.

Agent A: Her initial fear turned into excitement, tempered by the responsibility her power demanded.

Agent B: Eventually, she returned to her own time, armed with experiences and knowledge no one else possessed.

Agent A: She recorded her adventures in a journal, careful never to reveal too much about the future.

Agent B: Evelyn realized that her gift was both a privilege and a burden, teaching her patience and wisdom.

Agent A: Though tempted to change major events, she learned the importance of observation over interference.

Agent B: In the end, she embraced her role as a hidden witness to history, forever changed by the lessons she had learned.