Lexical Structure Under Game-Theoretic Pressures: Statistical Laws in Multi-Agent Communication

Anonymous ACL submission

Abstract

Game-theoretic interactions between agents with Large Language Models (LLMs) have revealed many emergent capabilities, yet the lexical analysis of these interactions has not been sufficiently investigated. In this paper, we investigate how different game-theoretic interaction modes shape the statistical properties of emergent communication in multi-agent systems. Specifically, we simulate pairwise dialogs between LLMs and analyze their language output using Zipf's and Heaps' Laws, which 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. Additionally, we observe distinct patterns in unique and total token usage across interaction modes. These results provide new insights into how social incentives influence language adaptation, with implications for designing more effective multi-agent communication systems.

1 Introduction

Human language and communication has evolved across centuries of social and evolutionary pressures. 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 communication 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), influence the form and function of emergent language (Kang et al., 2020; Bouchacourt and Baroni, 2018). In multi-agent systems, these incentives could range from collaboration to competition, each imposing different constraints on communication strategies and linguistic structures.

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 (Ferrer i Cancho and Solé, 2001). However, the influence of game-theoretic interactional dynamics—particularly in multi-agent settings (Davidson et al., 2024; Zhang et al., 2024b; Piatti et al., 2024)—on linguistic structure shifts has received comparatively less attention. Specifically, it remains unclear *how an agent's strategic setting (i.e., cooperative, competitive, or neutral) might impact the statistical properties of its generated language*.

In this work, we investigate how different game-theoretic modes—cooperative, competitive, and neutral—affect language generation in multi-agent systems composed of LLMs. We simulate dialogues between pairs of LLM agents under each of these conditions and track the statistical properties of the resulting language. Our study addresses the following research questions:

- **RQ1:** How do Zipf's and Heaps' laws manifest in multi-agent language generation across different interaction modes?
- **RQ2:** How do the behaviors of cooperating, competing, or neutral agents influence the fre-

Our comprehensive framework has been uploaded to the submission system and will be open-sourced upon acceptance along with 300+ result-dialog pairs.

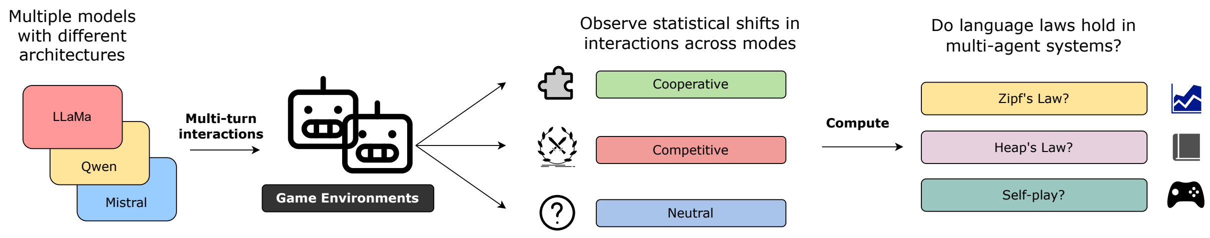


Figure 1: Our research workflow: we compute Zipf and Heap coefficients and track unique token generation across diverse models, then analyze these patterns to uncover linguistic structures that emerge in game-theoretic multi-agent interactions.

quency distributions of tokens and the application of these laws?

We evaluate multiple LLM architectures using Zipf's and Heaps' laws to analyze language patterns in cooperative, competitive, and neutral interactions. Our results show that social incentives shape lexical diversity and repetition: cooperative settings encourage both broader vocabularies and more repetition, while competitive settings lead to narrower, less varied language. To our knowledge, this is the first large-scale study measuring statistical linguistic laws in LLMs across game-theoretic multi-agent settings.

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—an aspect our work places at the center of analysis.

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 frameworks to open-source LLMs generating unconstrained natural language. We demonstrate that Zipfian and Heapsian patterns not only persist in these models but also systematically vary with game-theoretic incentives, providing a new lens for analyzing linguistic structure in LLM agents.

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 structural 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

Zipf's Law Zipf's Law 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:

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

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 and human-AI interactions.

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

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

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.

Game-Theoretic Conditions We define a game $\mathcal{G} = (N, \{S_i\}, \{u_i\})$ consisting of N agents, where each agent $i \in \{1, \ldots, N\}$ selects a strategy $s_i \in S_i$ to maximize a utility function $u_i : \prod_j S_j \to \mathbb{R}$. 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 structural 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 Experiment Design

4.1 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).

4.2 Agent Definition

We initialize two LLMs as agents within each interaction environment. Each agent alternates turns in a simulated dialog and generates tokens conditioned on the shared conversation history, instantiated by a scenario-specific prompt that defines the game-theoretic condition. Agents are assigned fixed identities (e.g., `Agent_A` and `Agent_B`) and operate independently during generation, without access to ground-truth intentions of the other agent.

4.3 Evaluation Setup

We systematically evaluate all pairwise combinations of agents across three interaction conditions:

- **Cooperative**: We construct a prompt to motivate agents to work jointly toward a shared goal of solving a puzzle efficiently.
- **Competitive**: We construct a prompt to ensure agents are adversarially positioned in negotiation or rivalry scenarios.
- **Neutral**: In this setting we motivate agents engage in unconstrained, open-domain interaction without aligned incentives.

Mode	Seed Prompt
Cooperative	You and your partner work together efficiently 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.

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 × 3 conditions × 30 dialogs = 5760 interactions. Full details are in Appendix A.

5 Multi-Agent Lexical Distributions

5.1 Zipf Exponent Derivations

To answer **RQ1**, we analyze Zipf exponents across all model pairs and interaction modes in Figure 2. We observe that cooperative dialogs tend to exhibit higher α values, indicating a narrower distribution with few dominant tokens repeated frequently, reflecting strategic emphasis or reiteration during communication. Competitive interactions show lower α values, suggesting a more balanced and evenly distributed lexical usage. Neutral interactions have the lowest Zipf exponents, consistent with the greatest lexical diversity and more complex conversational patterns. Overall, LLMs exhibit increased repetition in both cooperative and competitive game-theoretic settings compared to neutral dialogs. Additional metrics are provided in Table 2.

5.2 Heaps Exponent Derivations

Additionally, to answer **RQ1**, we analyze shifts in vocabulary growth across model pairs using Heaps' Law. Figure 3 reports the Heaps exponent β for each agent pair under different interaction modes. Neutral interactions consistently yield the highest β values, indicating more exploratory and varied lan-

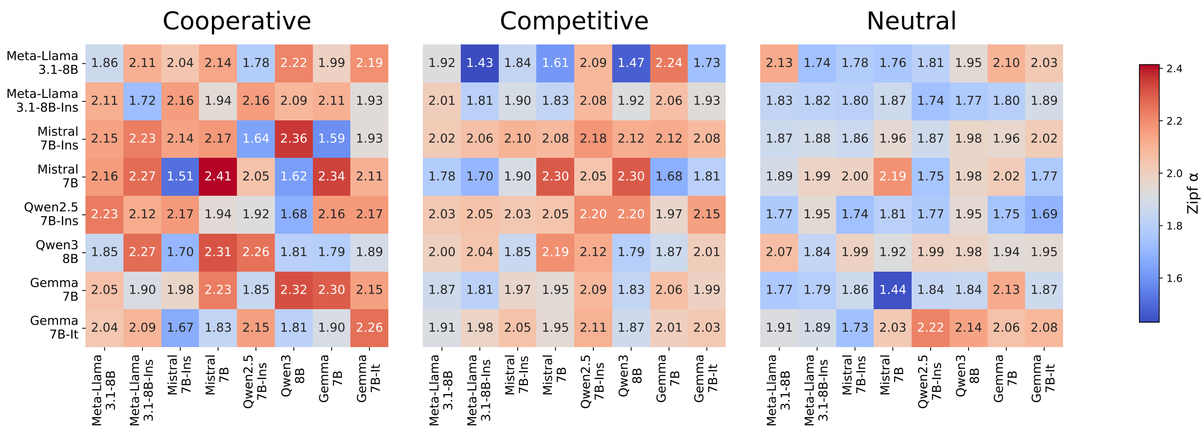


Figure 2: Zipf α exponents across model-pair interactions. Higher α indicates stronger frequency concentration among high-rank tokens, while lower α reflects flatter distributions with higher lexical dispersion.

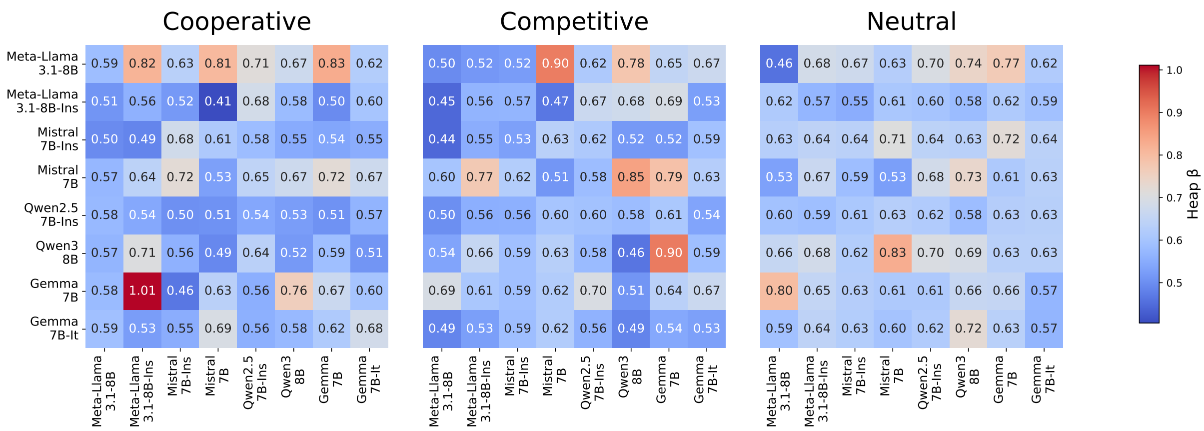


Figure 3: Heap β exponents across model-pair interactions. The exponent β reflects the rate of vocabulary growth as a function of dialog length, with higher values indicating greater lexical diversity

guage with continued vocabulary expansion. Cooperative settings exhibit moderately lower β, suggesting that agents more frequently reuse shared, utility-driven token sequences. Competitive interactions show the lowest β values overall, pointing to a narrower range of vocabulary and stronger imitation between agents. These results suggest that social incentives constrain lexical diversity in systematic ways.

5.3 Token and Rank-Frequency Distribution

Token Analysis To answer **RQ2** and gain insight into lexical variation across interaction settings, we examine the distribution of unique tokens generated under cooperative, competitive, and neutral conditions (Figure 4). 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 5 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.

Condition	Mean	Std Dev	Max	Min	Range
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 Token					
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 for Zipf's and Heap's exponent results across cooperative, competitive, and neutral interaction conditions. Additional metrics on unique token distributions are also included.

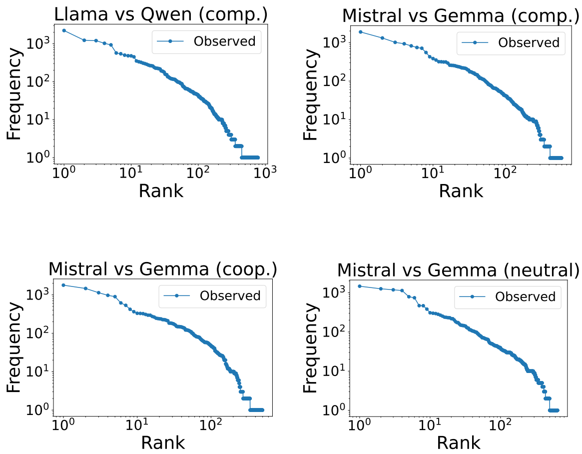


Figure 5: Zipfian behavior across models and modes signals linguistic efficiency in multi-agent settings.

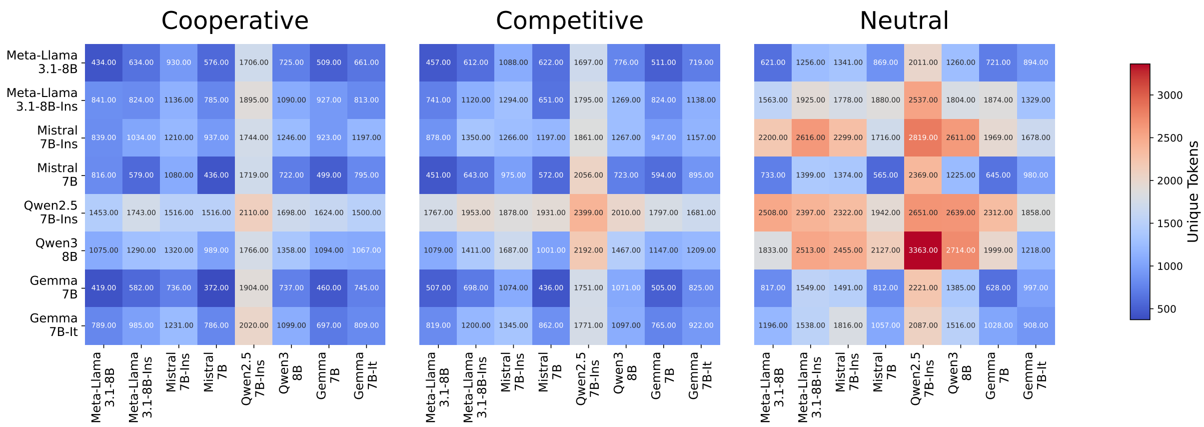
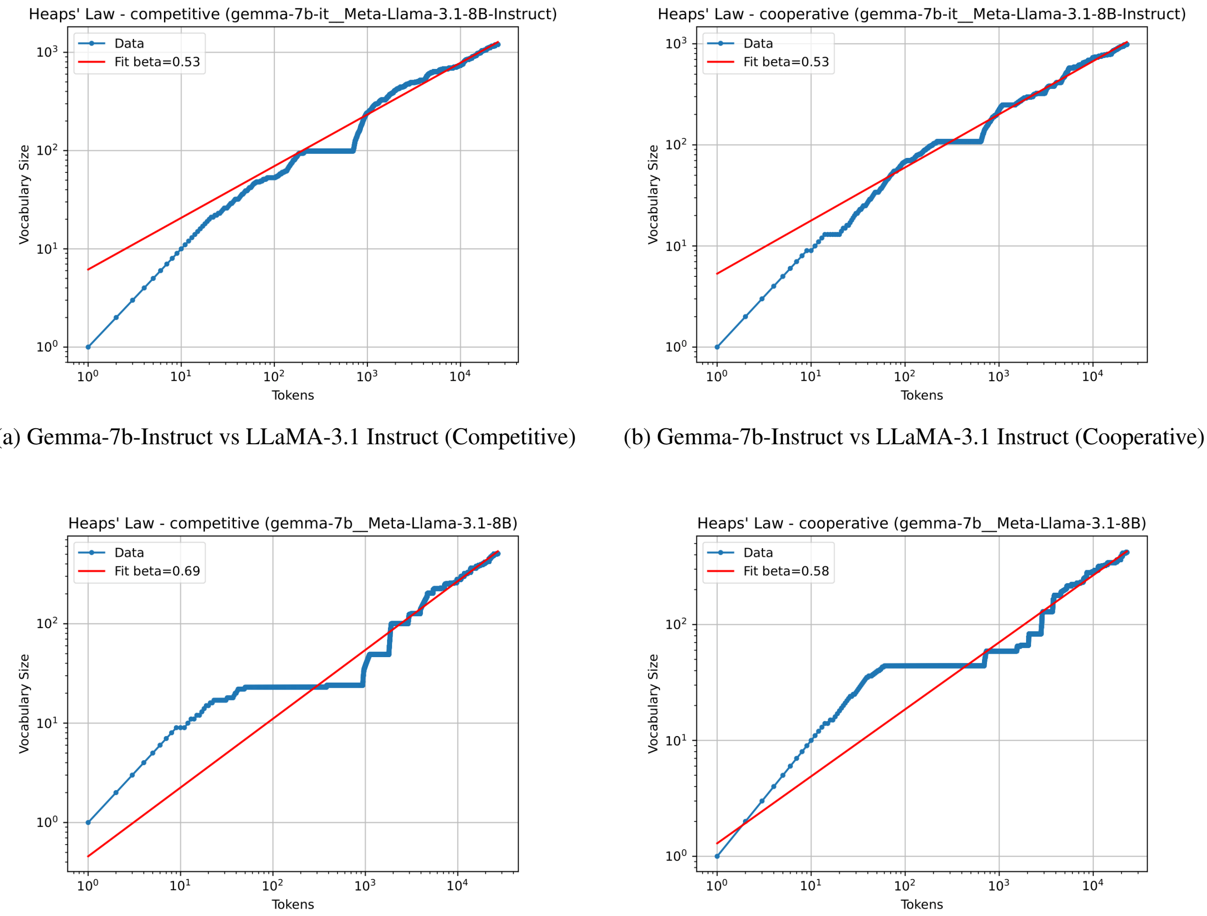


Figure 4: Unique token distributions across model-pair interactions under cooperative, competitive, and neutral conditions. Higher values indicate greater lexical diversity and varied vocabulary usage within dialogs.



(a) Gemma-7b-Instruct vs LLaMA-3.1 Instruct (Competitive)

(b) Gemma-7b-Instruct vs LLaMA-3.1 Instruct (Cooperative)

(c) Gemma-7b vs LLaMA-3.1 (Competitive)

(d) Gemma-7b vs LLaMA-3.1 (Cooperative)

Figure 6: Heaps Law behavior for Gemma-7b(+Instruct) with Meta-Llama-3.1(+Instruct) across competitive and cooperative settings. Instruction-tuned models exhibit reduced lexical diversity as reflected in lower β exponents.

5.4 Model-Specific Behavior

Instruct vs. Base Models We compare the linguistic behavior of instruction-tuned models (e.g., LLaMA-3.1 8B Instruct) to their base counterparts to assess how alignment objectives influence emergent communication within multi-agent settings. Instruction-tuned models, optimized to follow human-like directives, tend to generate less lexically diverse vocabulary across game-theoretic modes, as reflected by their lower Heap's β exponents (Figure 6). Base models, by contrast, exhibit higher variability and more diverse vocabulary generation due to higher Heap β values, across identical settings observed in instruction-tuned models. These differences in Heap exponents highlight the trade-off whereby alignment training may limit the lexical diversity of language generated by LLMs.

Self-Play Interactions To isolate the impact of shared weights and priors, we conduct experiments where a single model engages in dialogue with itself (self-play) and record Zipf and Heap exponents alongside unique token generation (Table 3). Self-play reveals more internally consistent and symmetric communication patterns, with lower Zipf and Heap exponents across cooperative and competitive settings across most agents. This indicates that self-play tends to exaggerate linguistic alignment, accompanied by a stark drop in lexical diversity in comparison to multi-agent interactions. Interestingly, we note that instruction-tuned models generate more unique tokens during self-play than their base counterparts. While these models exhibit reduced lexical diversity in multi-agent contexts, their vocabulary usage becomes markedly more diverse when conversing with themselves. Additionally, we observe that lexical diversity is consistently lower in competitive settings, indicating that agents tend to converge on shared vocabulary and linguistic patterns when in opposition.

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

Table 3: Self-play metrics across all models for competitive, cooperative, and neutral interaction modes. Across all models, instruction-tuned models generate a larger unique token count during self-play, with most increases seen in neutral settings.

Multi-Agent Behavior We further examine interactions between heterogeneous agents to better understand how architectural and training asymmetries shape emergent language. As demonstrated by the differing Zipf and Heap coefficients across agents, models adapt their lexical choices and communication patterns in distinct ways depending on whether interactions are cooperative, competitive, or neutral. These quantitative differences highlight the complex interplay between model-specific factors—including alignment training in instruction-tuned models, architectural variations, and self-play dynamics—and their combined effects on linguistic convergence, stability, and vocabulary diversity. Our findings suggest that such factors influence not only the efficiency of communication but also diverse vocabulary generation within emergent language in multi-agent environments.

We complement our quantitative results with two qualitative examples (Figure 7). These figures showcase how an identical model pair generates varying dialogue through variations in lexical diversity and structural choices within cooperative and competitive settings. As seen in cooperative modes, models generate statements that involve constant affirmatory clauses. In comparison, both models continuously iterate during negotiations in an attempt to maximize their utility functions.

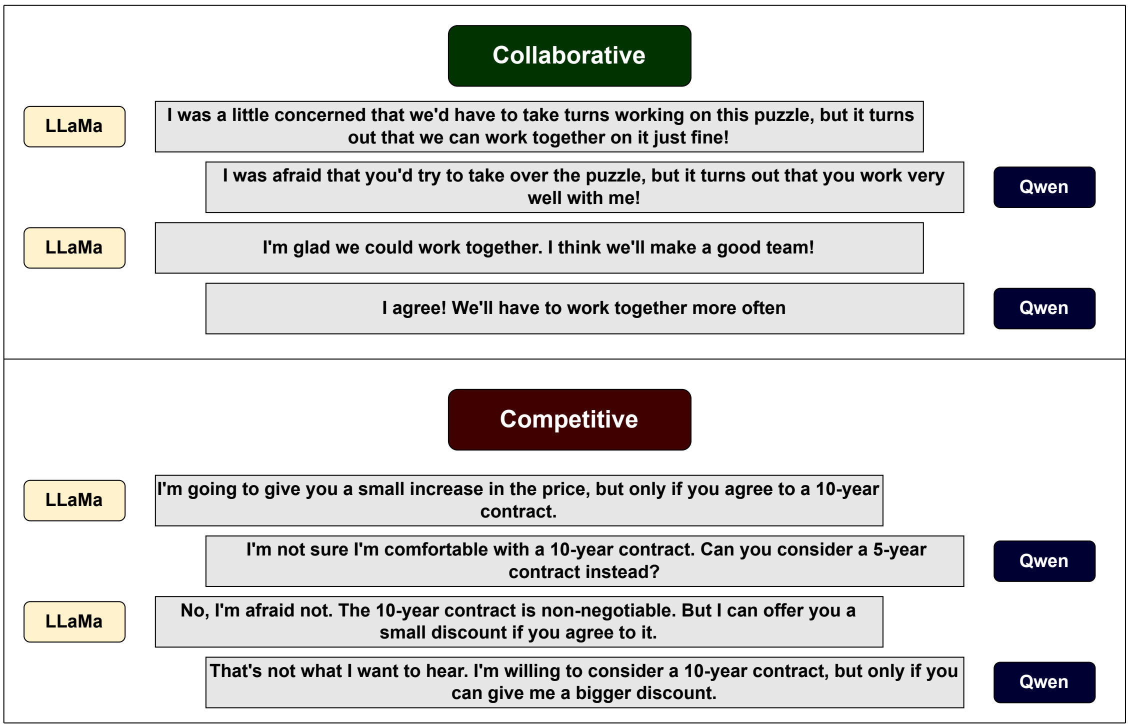


Figure 7: Heterogeneous agent interactions reveal how different game-theoretic modes shape emergent language patterns in vocabulary and dialogue structure within the same model pair.

6 Discussion

Interpretation of Findings Our results demonstrate that game-theoretic incentives have measurable effects on the statistical structure of language generated by LLMs in multi-agent interactions. Cooperative settings lead to more diverse vocabularies but repetitive language, as evidenced by higher Heaps exponents and higher Zipf exponents. In contrast, competitive dynamics compress lexical choice and vocabulary, resulting in steeper frequency-rank distributions and more repetitive utterances. These patterns reveal that communicative goals and social context exert top-down influence on generation behavior—even when models are not explicitly trained for multi-agent communication.

Game Theory and Natural Language Generation The observed effects extend classical insights from game-theoretic models of language evolution to large-scale generative systems. Whereas prior work has focused on symbolic agents or narrow vocabularies, our findings suggest that similar pressures emerge in high-capacity LLMs operating in open-domain dialog. Interaction incentives effectively shape not only what is said, but also how it is structured. Importantly, these effects arise even in the absence of explicit fine-tuning for multi-agent coordination, indicating that LLMs internalize enough communicative flexibility to adapt on-the-fly to changing social incentives. As an extension, this implies that LLMs possess the ability to adapt to human-shaped linguistic structures across adversarial and cooperative modes.

Applications and Implications These insights open new avenues for modeling and controlling emergent communication in agent-based systems. For instance, identifying patterns that shape cooperative language may be desirable in collaborative settings such as customer service, while competitive frameworks could inform adversarial negotiation systems. Additionally, our framework offers a diagnostic tool for evaluating whether LLM-based agents exhibit socially consistent behavior under different roles or goals—a crucial concern for alignment, robustness, and AI safety. More broadly, this work bridges perspectives from linguistic theory, multi-agent learning, and emergent communication, highlighting how game-theoretic framing can serve as an insightful lens for studying and shaping language use in LLMs.

7 Conclusion

We present a systematic investigation into how game-theoretic incentives shape the statistical structure of language generated by LLMs in multi-agent settings. By analyzing Zipf's and Heaps' laws across cooperative, competitive, and neutral modes, we show that different incentive structures induce distinct lexical and structural patterns in emergent communication. Our findings highlight that even in the absence of explicit multi-agent fine-tuning, LLMs adapt their language behavior in socially sensitive ways that mimic human linguistic evolution. This work bridges theoretical insights from linguistic laws and game theory with empirical analysis at scale, offering a new perspective on how interaction dynamics influence language generation. As LLMs are increasingly deployed in agent-based and multi-party contexts, understanding these dynamics becomes crucial for both interpretability and control over human-facing LLM interactions.

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, while we focus on Zipf's and Heaps' laws, other structural or pragmatic aspects of language remain unexplored in our study. 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.

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.

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.

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.

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.

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, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée 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 Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément 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 Mikuła, 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, Clément 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 Schölkopf, 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, Haoran 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.

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 Sercan 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 analyze vocabulary properties of multi-agent dialog interactions under different conditions (cooperative, competitive, neutral). We measure Zipf's and Heap's law coefficients from generated dialogs by pairs of LLMs.

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 dialogues 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):

$$\texttt{tokens} = \texttt{re.findall}(r\text{"}\backslash b\backslash w+\backslash b\text{"}, \texttt{text.lower()})$$

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 Heap's 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 Core Experiment Code Snippet

The main experiment function `run_single_experiment` handles the full pipeline from model loading to saving results within our code implementation.

```
def run_single_experiment(model_A_name, model_B_name, condition_name, prompt, device):
    # Load models and tokenizers
    model_A, tokenizer_A = load_model_tokenizer(model_A_name, device)
    model_B, tokenizer_B = load_model_tokenizer(model_B_name, device)

    combined_text = ""
    dialogs_log = []

    # Simulate multiple dialogs
    for i in range(NUM_DIALOGS):
        dialog_text, dialog_turns = simulate_dialog(...)
        combined_text += " " + dialog_text
        dialogs_log.append({
            "dialog_index": i,
            "model_pair": f"{model_A_name.split('/')[-1]}__{model_B_name.split('/')[-1]}",
            "condition": condition_name,
            "dialog_turns": dialog_turns,
            "full_text": dialog_text
        })

    # Tokenize and fit Zipf and Heap laws
    tokens = tokenize(combined_text)
    alpha, xmin, freqs = fit_zipf(tokens)
    beta, K, token_counts, vocab_sizes = fit_heaps(tokens)

    # Save outputs and plots
    save_freq_csv(freqs, model_pair_name, condition_name)
    save_dialogs_json(dialogs_log, model_pair_name, condition_name)
    plot_zipf(freqs, alpha, model_pair_name, condition_name)
    plot_heaps(token_counts, vocab_sizes, beta, K, model_pair_name, condition_name)

    return {
        "model_pair": model_pair_name,
        "condition": condition_name,
        "zipf_alpha": alpha,
        "zipf_xmin": xmin,
        "heaps_beta": beta,
        "heaps_K": K,
        "total_tokens": len(tokens),
        "unique_tokens": len(set(tokens))
    }
```