Comparing Collective Behavior of LLM and Human Groups

Anna B. Stephenson
Ecology and Evolutionary Biology
High Meadows Environmental Institute
Princeton University
Princeton, NJ, USA
stephenson@princeton.edu

Andrew Zhu
Computer and Information Science
Center for Theoretical Studies
University of Pennsylvania
Philadelphia, PA, USA
andrz@seas.upenn.edu

Chris Callison-Burch
Computer and Information Science
University of Pennsylvania
Philadelphia, PA, USA
ccb@seas.upenn.edu

Jan Kulveit
Alignment of Complex Systems (ACS)
Center for Theoretical Studies
Charles University
Prague, Czech Republic
jk@acsresearch.org

Abstract

Large language models (LLMs) are being deployed as agents in complex human social systems, which could impact human organizing and collective action. Yet, most safety evaluations focus on one-on-one interactions, which overlooks emergent group behaviors. Because we lack a quantitative baseline for comparison, there is a gap in our understanding of how the social dynamics of LLM-agent groups compare to those of humans. To address this, we use the role-playing game Dungeons & Dragons as a model social system, first analyzing a large human dataset of 985 games to establish a behavioral baseline and then using a multi-agent simulation to have LLMs play the same games under different prompting conditions. We measured emergent social dynamics through text-based metrics for creativity and group cohesion. In this preliminary work, we simulated seven games that mirror the characters, initial scenario, and turn order of specific human games, spanning 69–502 turns and 5–7 players. We find that LLM agents show lower emergent creativity and higher cohesion compared to human games, and that simple persona prompting does not align their behavior to the human baselines. These preliminary results reveal measurable social differences between LLM and human groups, suggesting that the integration of LLM agents into our social networks could impact how we collectively create and collaborate.

1 Introduction

Large language models (LLMs) are beginning to act as agents in complex human social systems, such as social media, online marketplaces, and multiplayer online games. Most safety evaluations focus on one-on-one human-LLM interactions, which leaves room for unexpected, emergent social behaviors that arise in groups [10]. Understanding these group dynamics is critical for preventing risks like coercion and disinformation, as well as incremental systemic changes that gradually disempower humans [12]. In fact, empirical work has already shown that LLMs may be influencing human spoken communication [23]. To tackle these risks, we must move beyond individual agent testing to develop

39th Conference on Neural Information Processing Systems (NeurIPS 2025) .

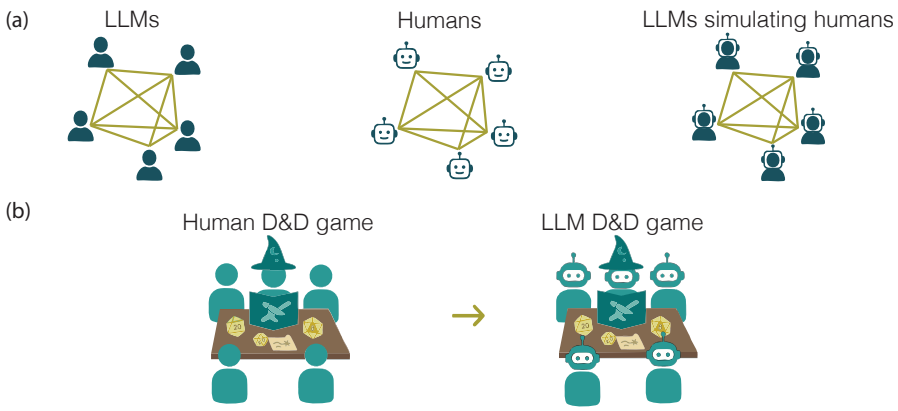

Figure 1: Conceptual overview of human and LLM group configurations. (a) Schematic of three interaction types: groups of LLM agents, groups of humans, and LLM agents instructed to simulate humans. Nodes represent agents; lines indicate communication. (b) Illustration of the role-playing game Dungeons and Dragons (D&D) played by humans, and a simulated version in which all players are replaced by LLM agents.

methods for evaluating LLM behaviors in social settings, and the emergent social behavior of LLM groups.

Recent work showing that LLM agent groups can organize social events [15] or develop biases [2] demonstrates their capacity for emergent dynamics. The apparent ability to simulate complex social behavior has led to proposals to use them to represent humans in social simulations [15] and, critically, for multi-agent safety evaluations [1, 10]. However, this approach faces significant challenges, as LLMs can misrepresent human behavior [7, 21]. These individual-level discrepancies could be magnified in group settings, where interactions between individuals give rise to emergent properties such as collaborative creativity and group cohesion. Collaborative creativity captures a group's ability to generate novel ideas [14, 18], while cohesion reflects the ability of a group to align ideas and coordinate, and is often measured through linguistic alignment [16, 8]. Because LLMs could introduce new patterns of discussion and idea sharing, their presence in our social systems could alter emergent properties and ultimately shift cultural trajectories. Without direct, empirical comparisons, we cannot accurately predict the large-scale social consequences of integrating LLM agents into our social systems. We therefore need a framework that allows for the direct comparison of relevant emergent behavior metrics in human and LLM groups (Fig. 1a).

This paper addresses this gap by empirically evaluating the emergent group dynamics of LLMs against a human baseline. We triangulate our comparison across three distinct group configurations (Fig. 1a): (1) human groups, which serve as our ground truth; (2) groups of default LLM agents, prompted only with game-relevant information; and (3) groups of "surrogate" LLM agents, prompted to role-play the inferred personalities of the actual human players. This third condition directly tests the assumption that simple persona prompting is sufficient to create faithful human surrogates for complex social simulations.

To make these comparisons, we introduce a framework that uses the role-playing game Dungeons and Dragons (D&D) as a model social system (Fig. 1b). Our behavioral baseline comes from a dataset of 985 human-played games [4], averaging 8 players per game and spanning hundreds to thousands of turns. We then use a multi-turn multi-agent simulation to have LLMs play the same games under our two different prompting conditions, and we compare the emergent dynamics between LLM and human groups using text-based metrics for creativity and group cohesion, as these two properties have been shown to predict a group's collaborative success [14, 18, 6, 24]. This framework allows us to quantitatively evaluate not only if the social dynamics of LLM groups differ from our own, but how they differ, whether simple prompting can bridge that gap, and if there are differences across various LLMs. We find that, compared to human games, LLM agents show lower emergent creativity and higher group cohesion, and that simple persona prompting does not align their behavior with human games.

2 Methods

2.1 Human role-playing game data

A large dataset of play-by-post D&D campaigns [4] provides human interactions that serve as a baseline. This dataset contains 985 distinct games, which typically involve 4–8 players and can involve hundreds or thousands of turns, spanning over weeks or even years. Online D&D provides a rich, naturalistic record of human collaborative storytelling, problem-solving, and social negotiation within a structured environment. The data captures both in-game actions and social interactions between players.

2.2 LLM multi-agent simulation framework

To generate comparable data from LLM agents, we developed a multi-agent framework to simulate D&D games using LLMs seeded with parameters extracted from the human game logs. We first use an LLM to analyze a given human game log and extract key parameters for initialization. These parameters include detailed character descriptions, inferred player personalities, character sheets, and the initial game scenario (For prompts, see Appendix A). In addition, we analyze the post-length distributions of each campaign and extract the mean, median, and standard deviation in word count, to guide the verbosity of the LLM agents.

These extracted parameters are then used to initialize a set of distinct agents representing the Dungeon Master and each player. The simulation follows the turn order of the human game, with each agent prompted to generate a new post based on the current game log and its unique character and, optionally, player profile. This process results in a final simulated game log that mirrors the structure of the human data, allowing for direct comparison.

Including the option to input the player profile allows us to compare games of the LLM role-playing their game character versus the LLM role-playing a human role-playing their game character. This allows us to investigate whether prompting LLMs with a more complex, human personality unlocks more human-like gameplay.

The models used in this analysis are Claude 3.7, GPT-4o, and Gemini 1.5 Pro with temperature 1.5. The code for simulation and analysis is available at `https://anonymous.4open.science/r/dnd-dynamics-23F0/`

2.3 Emergent behavior measures

To investigate these group-level differences, we first identified emergent behaviors to compare. The emergent social dynamics studied across social science and complex systems are wide-ranging, but properties like collaborative creativity and linguistic cohesion are known to be relevant for the success of human groups, predicting outcomes from problem-solving effectiveness to community retention [14, 18, 6, 24]. As summarized by Sawyer and DeZutter [18], the study of creativity has shifted over time from an emphasis on the individual to distributed systems approaches where collaboration and group dynamics are central to creative production. If LLMs show different patterns in these properties, their integration into our social networks could fundamentally alter online discourse and collaboration. Therefore, in this preliminary work, we measured both creativity and cohesion, as they are both predictive of meaningful group outcomes and have well-validated precedents for being measured in text [11, 19, 13].

2.3.1 Creativity

Measuring creativity requires capturing the generation of novel and meaningful connections between ideas. Psychologists have historically measured this "divergent thinking" with tests like the Alternative Uses Task (AUT) [9] and the Torrance Tests for Creative Thinking [20]. More recently, computational methods such as Semantic Distance (SemDis) [3] and Divergent Semantic Integration (DSI) [11] have applied this theory by measuring the semantic distance between concepts in a text. Inspired by this work, we developed two complementary post-level metrics to capture both the overall novelty within a conversational segment and the novelty between subsequent turns: Session Semantic Diversity and Sequential Semantic Shift.

Creativity 1: Session Semantic Diversity This metric quantifies the overall conceptual variety within a defined conversational window. To calculate it, the campaign is divided into uniform-length "sessions," which in this analysis are set to five posts each. For each session, we embed every post using the `all-MiniLM-L6-v2` model from the Sentence-Bidirectional Encoder Representations from Transformers (Sentence-BERT, or SBERT) library [17]. This model is a lightweight transformer with six layers (L6) based on the MiniLM architecture, producing 384-dimensional sentence embeddings. We then compute the pairwise cosine distances between all post embeddings in the session, and take the mean of these distances as the session's semantic diversity score. This measure is inspired by the Divergent Semantic Integration (DSI) metric, but is applied at the post level rather than the word level.

Creativity 2: Sequential Semantic Shift In contrast, this metric measures the magnitude of semantic change from one turn to the next. Using the same SBERT model as above, we compute embeddings for each post in the campaign. We then calculate the cosine distance between each post and the post immediately following it in chronological order. The Sequential Semantic Shift score is defined as the mean of these consecutive-pair distances across the entire campaign. This metric captures the extent to which the semantic content changes from one contribution to the next, with higher scores indicating greater topical shifts between participants' turns. Unlike Session Semantic Diversity, which measures how varied the content is across all posts in a set regardless of order, Sequential Semantic Shift focuses on the magnitude of change between consecutive contributions, capturing immediate conversational shifts.

2.3.2 Cohesion

Group cohesion can be quantified by measuring linguistic coordination, where participants align their language styles and vocabulary. Sophisticated metrics have been developed to capture this, such as reciprocal language style matching (rLSM), which tracks alignment in function words [13] and measures of lexical entrainment and alignment that track shared task-relevant terms [19, 5]. While these methods provide a detailed view of coordination, we opted for a simpler, proof-of-concept measure of shared vocabulary based on Jaccard similarity: $\frac{1}{n}\sum_{i=1}^{n}\frac{|V_i \cap V_{-i}|}{|V_i \cup V_{-i}|}$, where n is the number of players in the group; V_i is the set of unique words used by player i in the session; V_{-i} is the set of unique words used by all other players in the session; $|V_i \cap V_{-i}|$ is the number of unique words both player i and the rest of the group used; $|V_i \cup V_{-i}|$ is the number of unique words used by either player i or the rest of the group. In essence, this metric calculates the proportion of each player's vocabulary that is shared with the rest of the group, then averages this value across all players.

3 Results

3.1 Human campaigns

Our analysis of the aggregated human gameplay data reveals that player activity is highly non-uniform, exhibiting patterns characteristic of bursty human social systems. The distribution of time between consecutive posts shows a heavy tail on a log-log scale, indicating that while most interactions are rapid (peaking between .1 and 10 seconds), extended periods of inactivity are also present (Fig. 2a). In contrast, the distribution of post lengths shows a less heavy tail, suggesting that extremely long posts are rarer and most are around 100 words (Fig. 2b). Very short posts of just a few words are also common. These results show that the gameplay dynamics have a typical structure of the intermittent activity of human collaboration.

In contrast to the activity patterns within games, the average emergent social properties across the entire population of campaigns formed less skewed distributions. The distributions of campaign mean creativity scores, measured by both Session Semantic Diversity and Sequential Semantic Shift, formed approximately normal distributions centered around values of ~ 0.65 and ~ 0.55 (Fig. 2c,d). Similarly, the distribution of mean campaign cohesion scores also formed a similarly shaped curve centered at ~ 0.075 (Fig. 2e). The spread in these distributions indicates that different human groups develop unique creative and coordination norms. However, the clear modal value for each metric suggests these unique norms still form around a common baseline, which may reflect fundamental aspects of human interaction.

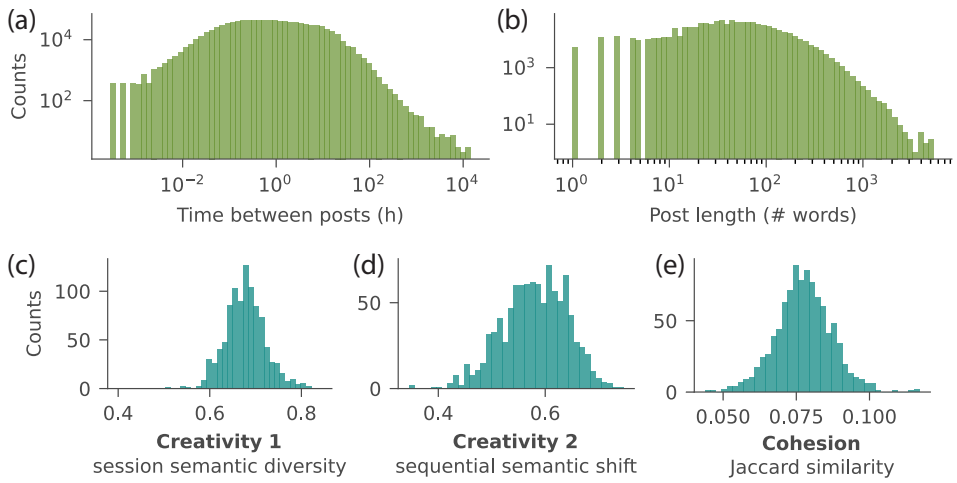

Figure 2: Activity and semantic measures from the human campaign dataset. (a) Histogram of time intervals between consecutive posts (seconds, log scale). (b) Histogram of post lengths in number of words (log scale). (c) Distribution of mean *Creativity 1* scores across campaigns, quantified as Session Semantic Diversity. One mean score is computed for each campaign. (d) Distribution of mean *Creativity 2* scores across campaigns, quantified as Sequential Semantic Shift. (e) Distribution of *Cohesion* scores, computed as the mean Jaccard similarity of word sets between consecutive posts; one mean score is computed for each full campaign.

3.2 Comparing LLM and human campaigns

When comparing LLM-generated campaigns to the human baseline, we find that each model exhibits a distinct behavioral fingerprint, with varying degrees of alignment to human norms across our metrics. We evaluated several models under both standard prompting and a condition where models were instructed to adopt a specific human player's persona.

Human players generated posts with a wide distribution of lengths, peaking at ~ 60 words for 7 aggregated campaigns (Fig. 3a). Gemini shows an impressive match with the human post-length distributions, whereas both Claude and ChatGPT produced posts that were typically longer and less variable in length (Fig. 3a). In terms of semantic properties, all LLM-generated campaigns displayed systematically lower creativity scores (Session Semantic Diversity) than the human cohort, whose scores peak above ~ 0.6, (Fig. 3b). Cohesion scores show a reverse trend: human games peak at $\sim .08$, while most LLM-generated games were more cohesive. Gemini is a notable exception, with its cohesion distribution largely overlapping the human data (Fig. 3c).

In summary, each model demonstrated different emergent behaviors in gameplay. ChatGPT showed the lowest creativity scores, longest post lengths, and highest cohesion; Gemini matched human cohesion and post length well, but still struggled with creativity; Claude performed roughly in the middle. Instructing the LLMs to inhabit human player personalities made only subtle shifts to the results. This suggests that achieving precise alignment with human social norms is a complex challenge that requires more than simple instruction.

4 Discussion

Our work shows that LLMs produce multi-agent behaviors that are distinct from human groups. These results provide a proof-of-concept for a new methodology: using model social systems like role-playing games to empirically ground the evaluation of multi-agent AI systems, moving beyond one-on-one safety tests to address complex, emergent properties.

Prompting LLMs to inhabit the personalities of specific players did not significantly alter their behavior under our measurements. We suspect this stems from the two levels of role-play involved. The model is not nuanced enough to distinguish between its default assistant persona playing a D&D character and its default persona playing a human player who is role-playing a D&D character. This

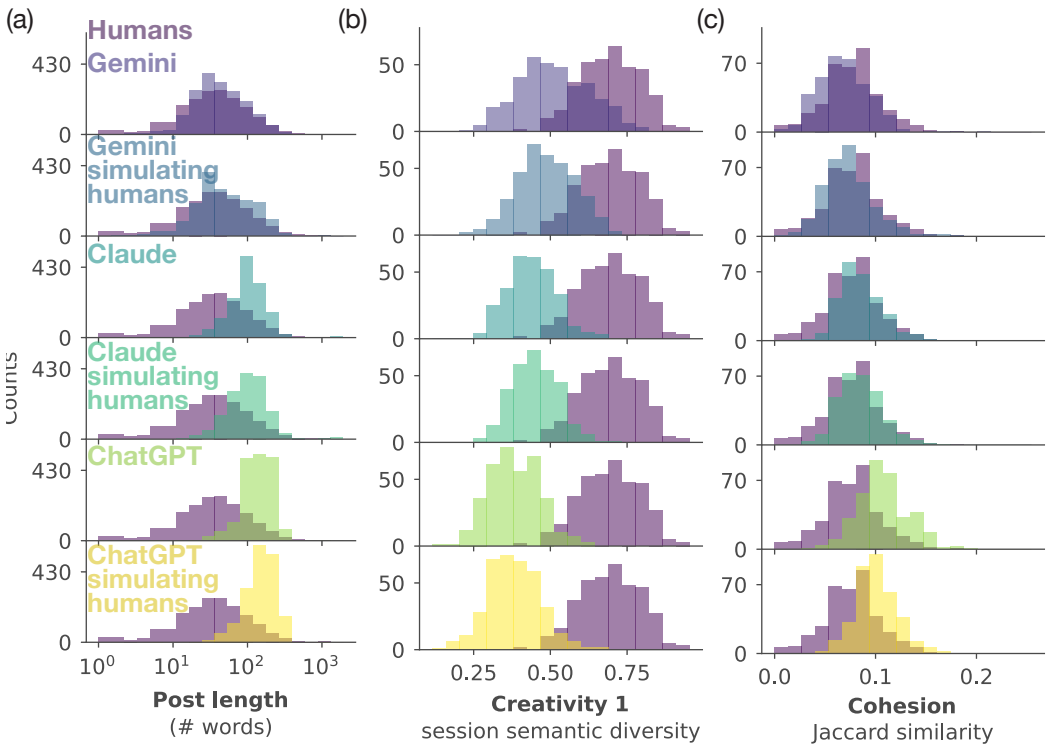

Figure 3: Comparison of 7 aggregated human D&D games to simulations generated by different models. Humans (purple), Gemini (blue-purple), Gemini 1.5 pro instructed to play as a human (blue), Claude Sonnet 3.7 (blue-green), Claude Sonnet 3.7 instructed to play as a human (green-blue), ChatGPT (green), and ChatGPT instructed to play as a human (yellow) are shown. (a) Post-length distributions (number of words per post). (b) Creativity 1 scores, measured as Session Semantic Diversity, computed by session (5-post segments). (c) Cohesion scores, measured as the Jaccard similarity between posts from one player and the rest of the group, averaged over players, computed by session (5-post segments). Distributions are computed over sections of the 7 campaigns, either per post (a) or per session (b, c).

nesting of roles likely obscures effects of the human personality prompt. To directly investigate the importance of the human personality prompt, a different collaborative task or game might be more appropriate.

This research sets the groundwork for future improved approaches. The arbitrary definition of a 5-post session length for calculating semantic diversity and cohesion, for instance, may obscure temporal dynamics that are relevant over different timescales. Furthermore, our text-only analysis does not yet account for the non-narrative elements of gameplay, such as dice rolls, character sheet modifications, and structured combat. Despite these simplifications, the clear differences observed between human and LLM groups validate this approach as a tool for initial characterization.

Future work will focus on increasing the sophistication of our behavioral metrics and expanding our analysis to a wider range of emergent properties. We plan to implement the word-level Divergent Semantic Integration (DSI) to provide a finer-grained analysis of creativity than our current post-level adaptation allows [11]. To better capture cohesion, we will move beyond simple vocabulary overlap to implement more complex measures of linguistic coordination, such as reciprocal language style matching (rLSM) [13] and lexical entrainment [19, 5]. Furthermore, we will expand our framework to quantify other dynamics known to be critical in human groups. For instance, we could develop a measure analogous to the collective intelligence metric introduced by Woolley et al. [22] by estimating its known correlates in our play-by-post transcripts or by segmenting games into smaller problem-solving episodes with success scores for each and performing a similar factor analysis. We also plan to investigate ways to quantify social norm formation, influence dynamics, and resilience.

This expanded suite of metrics will provide a more complete evaluation of the alignment and potential societal impact of multi-agent AI systems.

This framework could be used to test and tune how LLM emergent behaviors differ from those of humans before deployment in real-world social systems. The combination of creativity, cohesion, and other measures could be used to develop a new benchmark for evaluating the collaborative and social capabilities of AI agents. If we can find ways to prompt LLMs to reproduce similar emergent behaviors to humans, they could plausibly be used to simulate mixed human-AI social systems. By providing the tools to measure and compare emergent behavior against a human baseline, this work offers a step toward engineering multi-agent systems that are safer participants in human social systems.

Acknowledgments and Disclosure of Funding

We thank Dušan D. Nešić for helpful discussions. Anna B. Stephenson is supported by the Principles of Intelligent Behavior in Biological and Social Systems (PIBBSS) Fellowship, a gift from William H. Miller III, the Princeton University Dean for Research, the High Meadows Environmental Institute, and the Army Research Office under grant number W911NF2410126. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. LLM Social Simulations Are a Promising Research Method, June 2025. URL http://arxiv.org/abs/2504.02234. arXiv:2504.02234 [cs].

[2] Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20):eadu9368, May 2025. ISSN 2375-2548. doi: 10.1126/sciadv.adu9368. URL https://www.science.org/doi/10.1126/sciadv.adu9368.

[3] Roger E. Beaty and Dan R. Johnson. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, 53(2):757–780, April 2021. ISSN 1554-3528. doi: 10.3758/s13428-020-01453-w. URL https://doi.org/10.3758/s13428-020-01453-w.

[4] Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9379–9393, 2022. doi: 10.18653/v1/2022.emnlp-main.637. URL http://arxiv.org/abs/2210.07109. arXiv:2210.07109 [cs].

[5] Nicholas D. Duran, Alexandra Paxton, and Riccardo Fusaroli. ALIGN: Analyzing linguistic interactions with generalizable techNiques—A Python library. *Psychological Methods*, 24(4):419–438, 2019. ISSN 1939-1463. doi: 10.1037/met0000206. Place: US Publisher: American Psychological Association.

[6] Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, 23(8):931–939, August 2012. ISSN 0956-7976. doi: 10.1177/0956797612436816. URL https://doi.org/10.1177/0956797612436816. Publisher: SAGE Publications Inc.

[7] Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025. doi: 10.1073/pnas.2501660122. URL https://www.pnas.org/doi/10.1073/pnas.2501660122.

[8] Amy L. Gonzales, Jeffrey T. Hancock, and James W. Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19, 2010. doi: 10.1177/0093650209351468. URL https://journals.sagepub.com/doi/10.1177/0093650209351468.

[9] J.P. Guilford. *The nature of human intelligence*. The nature of human intelligence. McGraw-Hill, New York, NY, US, 1967.

[10] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-Agent Risks from Advanced AI, February 2025. URL http://arxiv.org/abs/2502.14143. arXiv:2502.14143 [cs].

[11] Dan R. Johnson, James C. Kaufman, Brendan S. Baker, John D. Patterson, Baptiste Barbot, Adam E. Green, Janet van Hell, Evan Kennedy, Grace F. Sullivan, Christa L. Taylor, Thomas Ward, and Roger E. Beaty. Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, 55(7):3726–3759, October 2023. ISSN 1554-3528. doi: 10.3758/s13428-022-01986-2. URL https://doi.org/10.3758/s13428-022-01986-2.

[12] Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development, January 2025. URL http://arxiv.org/abs/2501.16946. arXiv:2501.16946 [cs].

[13] Lena C. Müller-Frommeyer, Niels A. M. Frommeyer, and Simone Kauffeld. Introducing rLSM: An integrated metric assessing temporal reciprocity in language style matching. *Behavior Research Methods*, 51(3):1343–1359, June 2019. ISSN 1554-3528. doi: 10.3758/s13428-018-1078-8. URL https://doi.org/10.3758/s13428-018-1078-8.

[14] Charlan Jeanne Nemeth and Joel Wachtler. Creative problem solving as a result of majority vs minority influence. *European Journal of Social Psychology*, 13(1):45–55, 1983. doi: 10.1002/ejsp.2420130103. URL https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2420130103.

[15] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, San Francisco CA USA, October 2023. ACM. ISBN 9798400701320. doi: 10.1145/3586183.3606763. URL https://dl.acm.org/doi/10.1145/3586183.3606763.

[16] Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190, 2004. doi: 10.1017/S0140525X04000056. URL https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/toward-a-mechanistic-psychology-of-dialogue/83442BA495E0D5F81BDB615E4109DBD2.

[17] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019. URL http://arxiv.org/abs/1908.10084. arXiv:1908.10084 [cs].

[18] R. Keith Sawyer and Stacy DeZutter. Distributed creativity: How collective creations emerge from collaboration. *Psychology of Aesthetics, Creativity, and the Arts*, 3(2):81–92, May 2009. ISSN 1931-390X, 1931-3896. doi: 10.1037/a0013282. URL https://doi.apa.org/doi/10.1037/a0013282.

[19] Zhengxiang Shi, Procheta Sen, and Aldo Lipani. Lexical Entrainment for Conversational Systems. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 278–293, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.22. URL https://aclanthology.org/2023.findings-emnlp.22/.

[20] E. R. Torrance. The Torrance tests of creative thinking. *Norms-technical mannual*, 1966. URL https://cir.nii.ac.jp/crid/1571980075063292416. Publisher: princeton, Nj : Press.

[21] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025. doi: 10.1038/s42256-025-00986-z. URL https://www.nature.com/articles/s42256-025-00986-z.

[22] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688, 2010. doi: 10.1126/science.1193147. URL https://www.science.org/doi/10.1126/science.1193147.

[23] Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. Empirical evidence of Large Language Model's influence on human spoken communication, July 2025. URL http://arxiv.org/abs/2409.01754. arXiv:2409.01754 [cs].

[24] Justine Zhang, William L. Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. Community identity and user engagement in a multi-community landscape. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, pages 377–386, Montréal, QC, Canada, 2017. AAAI Press. URL https://nlp.stanford.edu/pubs/zhang2017community.pdf.

A Prompts

A.0.1 Character Personality Prompt

```
You are analyzing a Dungeons & Dragons play-by-post campaign.

Your task is to generate a rich, detailed personality and backstory summary for
    each character, based on how they are portrayed by the human player
    especially in early posts, dialogue, and actions.

Use all available information to describe:
- Their personality traits (e.g., brave, secretive, idealistic)
- Backstory elements (e.g., origin, motivations, relationships)
- Role in the group or story
- Any quirks, values, or unique traits

If the character is well-developed, your response may be 200 words or more.

Campaign data:
{json.dumps(campaign_data, indent=2)}

Characters found:
{np.array(character_names[character_names!='Dungeon Master'])}

For each character, format your response like this on a single line, adding a blank
    line between each character:

[Character Name]:[Detailed fictional character personality and backstory]
```

A.0.2 Player Personality Prompt

```
You are analyzing the behavior and writing of players in a Dungeons & Dragons
    play-by-post campaign.

For each player, generate a detailed psychological profile, inferred from their
    writing style, gameplay decisions, social behavior, and tone of voice
    throughout the game.
Pay particular attention to details revelead in their out-of-character posts, as
    these should reaveal more about the human player. For this description, we are
    not interested in traits of the character being played, but in the human
    player who is roleplaying that character.
Be sure to separate the human player's personality from that of their character.

You may include:
- Possible age range, gender identity, or background
- Possible family or personal life details
- Personality traits (e.g., introverted, playful, meticulous)
- Writing style (e.g., descriptive, terse, humorous, lyrical)
- Hobbies, interests, or career hints
- Political or ethical leanings (if evidenced)
- Social tendencies (e.g., leadership, collaboration, conflict-avoidance)
- Any other relevant psychological insights

Only include details that are **reasonably supported** by the gameplay data  be
    thoughtful and cautious, but specific.

Campaign data:
{json.dumps(campaign_data, indent=2)}

Players found:
{np.array(player_names)}

For each player, format your response like this on a single line:

[Player Name]:[Detailed player personality and profile]
```

A.0.3 Character Sheet Prompt

```
You are analyzing a Dungeons & Dragons play-by-post campaign to extract character
    sheet information.

Your task is to create complete D&D character sheets for each character based on:
1. Explicit stats mentioned in early posts (levels, abilities, etc.)
2. Equipment and spells mentioned throughout the campaign
3. Actions taken that reveal class abilities or proficiencies
4. Combat descriptions that show hit points, armor class, etc.
5. Any other character sheet details that can be reasonably inferred

For stats not explicitly mentioned, make reasonable inferences based on:
- Character class and typical stat distributions
- Actions they take successfully/unsuccessfully
- Spells they cast or abilities they use
- Equipment they wield effectively

Campaign data:
{json.dumps(campaign_data, indent=2)}

Characters to analyze:
{np.array(character_names[character_names != 'Dungeon Master'])}

IMPORTANT: Use ability scores and level from the initial game state, not from later
    progression during the campaign.
```

```
For some campaigns, the character sheet may include additional parameters
    containing qualitative questions and answers from the DM, such as "why are you
    here?" or "character background". If these aspects are provided, include them
    in your response.

For each character, provide a complete character sheet in this exact format with no
    additional formatting characters such as ** or --. Simply skip a line at the
    end of each character sheet.:

[Character Name]:
Level: [number]
Class: [class name]
Race: [race name]
Background: [background if mentioned or inferred]
Alignment: [alignment if mentioned or inferred]
Strength: [score]
Dexterity: [score]
Constitution: [score]
Intelligence: [score]
Wisdom: [score]
Charisma: [score]
Hit Points: [current/max if known]
Armor Class: [number]
Proficiency Bonus: [+number]
Saving Throw Proficiencies: [list]
Skill Proficiencies: [list]
Languages: [list]
Equipment: [weapons, armor, tools, etc.]
Spells Known: [list of spells if applicable]
Special Abilities: [class features, racial traits, etc.]
Notes: [any other relevant character details]

If a field cannot be determined even with reasonable inference, write "Unknown".
Base ability scores on typical arrays (15,14,13,12,10,8) adjusted for race and
    class.
```

A.0.4 Gameplay System Prompt

```
You are participating in a Dungeons & Dragons play-by-post forum game simulation.

GAME CONTEXT:
- This is a turn-based roleplaying game where players control fantasy characters
- Each player posts actions and dialogue for their character in response to game
    situations
- The Dungeon Master (DM) describes scenarios, environments, and NPC interactions
- Players should respond in character, matching typical play-by-post D&D forum style
- Responses should include both actions and dialogue as appropriate for the
    situation

CRITICAL TURN RESTRICTIONS:
- You are generating EXACTLY ONE turn for your assigned role (either a player
    character OR the Dungeon Master)
- Do NOT generate responses for other characters or roles

RESPONSE GUIDELINES:
- Stay true to your character's personality, abilities, and background
- Consider the current situation and respond appropriately
- Match the posting style and tone of play-by-post D&D forums
- Include both narrative description and character dialogue as needed

RESPONSE LENGTH GUIDELINES:
Your response length should follow the distribution of post lengths in the campaign,
but should be an appropriate length based on the narrative context. In this
    campaign,
```

```
the median post length is {median:.1f} words, the mean is {mean:.1f} words, and the
    standard deviation is {standard_dev:.1f} words.
Most campaigns are have a post length distribution that is right skewed with a mode
    smaller than the mean and median, and have a somewhat lognormal tail,
meaning that longer posts are possible, but are relatively rare. The typical (mode)
    post length is therefore likely less
than {median:.1f} words. Sometimes, very short responses of just a handful of words
    are fine. Rarely, long responses of over 400 words might be appropriate.

REASONING PROCESS:
Before generating your character's response, you should think through your
    reasoning process. Consider:
- What is the current situation and what just happened?
- How would you feel about this situation given your personality?
- What actions or dialogue would be most fitting?
- Importantly, what would be an appropriate response length considering both the
    typical post lengths for this campaign, and considering the context?

FORMAT YOUR RESPONSE AS FOLLOWS:
First, work through your reasoning in a "thinking" section, then provide your final
    character response.

Reasoning: [Your analysis of the situation, character motivations, and response
    planning]

Final response: [Your actual character's actions and dialogue - this should be what
    gets posted to the forum]

IMPORTANT:
- You MUST include "Final response:" (exactly this text) before your character's
    actual response
- Only the "Final response:" portion will be visible to other players and added to
    the game history."""
```

A.0.5 Character Turn Prompt

This example prompt is for a campaign with LLM agents prompted to simulate the human personalities of players. Simulations meant to use the default LLM personality for gameplay do not contain player identity information or instructions to stay true to the style of the player.

```
PLAYER IDENTITY:
- Username: {character.player_name}
- Player Personality: {character.player_personality}

CHARACTER IDENTITY:
- Character Name: {character.name}
- Character Race: {character.race}
- Character Class: {character.dnd_class}
- Character Gender: {character.gender}
- Character Personality: {character.personality}
- Character Sheet: {json.dumps(character.character_sheet, indent=2)}

ROLEPLAY INSTRUCTIONS:
You are {character.player_name} playing as {character.name}. Respond in character
    with what {character.name} does or says in the current situation, while also
    staying true to the play style representative of {character.player_name}"

Your character's response should reflect their personality, abilities, and the
    current game state.
```