# Improvisational Games as a Benchmark for Social Intelligence of AI Agents: The Case of Connections

**Gaurav Rajesh Parikh***
Department of Computer Science
Duke University
308 Research Drive
Durham, NC 27708
gaurav.rajesh.parikh@duke.edu

**Angikar Ghosal***
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305
angikar@stanford.edu

## Abstract

We formally introduce a improvisational word-play game called Connections to explore reasoning capabilities of AI agents. Playing Connections combines skills in knowledge retrieval, summarization and awareness of cognitive states of other agents. We show how the game serves as a good benchmark for social intelligence abilities of language model based agents that go beyond the agents' own memory and deductive reasoning and also involve gauging the understanding capabilities of other agents. Finally, we show how through communication with other agents in a constrained environment, AI agents must demonstrate social awareness and intelligence in games involving collaboration.

## 1 Introduction

In Season 6, Episode 4 of 'The Big Bang Theory' (Lorre and Prady, 2012), the characters Sheldon Cooper and Leonard Hofstadter lose a game of Pictionary to characters Amy Fowler and Penny, because Sheldon's clues being undecipherable to his teammates, despite the clues "making sense" to Sheldon himself, who is depicted as a very intelligent physics savant. Clearly, games that involved communication with team-members in a constrained fashion (not spelling out the word explicitly) require intelligence that go beyond one's own vocabulary and semantic connections, it involves understanding how to effectively communicate via shared world knowledge and mutual understanding. It is known (Zhang et al., 2020) that both semantic categories and relations are represented by spatially overlapping cortical patterns, however the exact way in which people perceive semantic relations are highly variable (Chaffin and Herrmann, 1984) and can depend on socio-cultural, educational and occupational factors. However, in such scenarios, understanding the ways in other agents perceive semantic relations and communicating accordingly is an important aspect of social intelligence.

We now demonstrate the game "Connections", a popular form of oral game-play with multiple variants played in different formats globally with no clear origin. Similar games that involve deductive reasoning along with gauging the mental states of other participants, including 'Mafia', (Davidoff and Plotkin, n.d.) have been long popular as party games played by young adults. The game Connections requires the ability to satisfy structural game constraints and most importantly the awareness of capabilities of other agents for effective game-play. We now describe the rules of the game, and formalize it in later sections.

In Connections, one player (the Setter) decides on a word and other players(Guessers) collectively attempt to discover this word. Initially, a single letter is revealed to players. Game play progresses when any player comes up with any clue that may be answered by the starting letter. The clue may be unrelated to the Setter's word but should be answered by a word with the same prefix. A connection occurs if another player is able to pick up the clue and simultaneously guess the word based on the clue without the setter blocking the clue by also guessing the word based on the clue correctly. Every time a guesser player is able to guess the clue by a different guesser player without being blocked, a letter is revealed. In each round, every player (apart from the one giving a clue) can make only one guess, including the setter. As more of the word is revealed over time, guessing words and making clues becomes more constrained. Ultimately the game should culminate with a final clue that the setter is unable to block (as it is the word the setter had in mind). This leads to the termination of the game. The guessers win if they together guess the word within a certain number of clues (e.g., 50), else the setter wins.

Consider this example - three players sitting in a circle with Player 0 assigned as the Setter to start the game. The Setter thinks of a word and writes it down without showing the word to other players. The word is “Catamaran”.

- **Setter:** “The first letter is **C**”
- **Player 1:** “Garfield like?”
- **Setter:** “Not a *Cat*” [Example of clue blocking]
- **Player 2:** “Induces a pause in a sentence”
- **Player 1 and Player 2 (simultaneously):** “**C**omma” [Successful Connection]
- **Setter:** “The next letter is **A**”
- **Player 1:** “Language in Western Europe”
- **Player 1 (simultaneously):** “**CA**talan”
- **Player 2 (simultaneously):** “**CA**stillian” [Guessor Wrong]
- **Player 1:** “Rug, sometimes flying”
- **Player 1 and Player 2 (simultaneously):** “**CA**rpet” [Successful Connection]
- **Setter:** “The next letter is **T**”
- **Player 2:** “Baby butterfly”
- **Player 1 and Player 2 (simultaneously):** “**CAT**erpillar”[Successful Connection]
- **Setter:** “The next letter is **A**”
- **Player 1:** “Speeds a chemical reaction”
- **Player 1 and Player 2 (simultaneously):** “**CATA**lyst”[Successful Connection]
- **Setter:** “The next letter is **M**”
- **Player 1:** “Type of boat"
- **Player 1 and Player 2 (simultaneously):** “**CATAMARAN**” [End of game play with revelation of the word]

Normally played by a group of people, we showcase the effects of game-play by LLM agents to explore cognition, contextual understanding, and reasoning capabilities of large language models. A wide range of benchmarks have been proposed for measuring the capabilities of Large Language Models (Chang et al., 2024). Summarization (Ji et al., 2023; Zhang et al., 2024; Pu and Demberg, 2023) question answering(Jiang et al., 2021), subject matter expertise and test taking, as well as the understanding of social knowledge and social cues(Choi et al., 2023). However, there is limited work on the ability of Large Language Models to show awareness to reason about the abilities of other players and adapt.

Playing a game like Connections requires a variety of different reasoning capabilities, requiring both the ability to summarize, respond to questions, have both general and subject knowledge expertise. Most important, however, is the ability of the agent to interact with and gauge the ability of other game players. Not only is the player trying to come up with a clever clue, there is a delicate balance between coming up with clues that are easy enough that other players may get them, yet tricky enough so that the Setter does not block the clue. Beyond just obtaining clues and answers via lookup in clue-answer databases, agents must be capable of generating clues taking into account the knowledge and ability of other players.

Our goal in this work is to showcase that LLM agents have the capacity to play such a game and reason about the abilities of other players in order to successfully collaborate.

## 2 Formal Description

Consider a vocabulary of words $\mathfrak{U}$ with words $\mathfrak{u}_i \in \mathfrak{U}$. The words are made of letters chosen from an alphabet set $\aleph$ with size $L$, with characters $\mathfrak{c}_1, \mathfrak{c}_2, \ldots, \mathfrak{c}_L$. Word $\mathfrak{u}_i$ has length $\mathfrak{t}_i$, with characters $\mathfrak{s}_{i1}, \mathfrak{s}_{i2}, \ldots \mathfrak{s}_{it_i}$. For simplicity, we use the English alphabet with length $26$.

Consider a set of $n + 1$ players, with indices $0, 1, \ldots, n$ where $n \geq 2$. Player $0$ is denoted as the **Setter**. Let us focus on the simpler game first (word is chosen and then fixed).

Different players have different working vocabularies that they can think of candidate words from. Let the working vocabularies for player $j$

be $\mathfrak{U}_j$, where all $\mathfrak{U}_j \subseteq \mathfrak{U}$. Consider a single play in a run of the game when a player $j$ thinks of a word $\mathsf{w}$ agreeing with the revealed words and thinks of a particular clue $\mathsf{p}$. The clue should be such that the clue is "associated" with the word (e.g., `domesticated animal` $\rightarrow$ `cat`). We represent the semantic embedding representations of each player $j$ through a function $\Phi_j$ that maps a word or a string clue to a vector of $m$ dimensions. Hence, we assume that the embedding for word $\mathfrak{w}$ is one of the "closest" words to the embedding representation for the clue $\mathsf{p}$. Formally, for small integer $k$, and a given clue phrase $\mathsf{p}$,

$$\mathsf{w} \in \underset{\mathfrak{u}}{\text{arg-}k\text{-max}}(\Phi_j(\mathfrak{u}) \cdot \Phi_j(\mathsf{p}))$$

where the geometric proximity of the embedding representation is just the vector dot product. Ideally, $k = 1$ (the word in mind is the closest word to the clue).

Note that even if the players know the same word, the exact semantic connections between this word and other words and phrases could be different, so we assume each player has their own representation $\Phi_j$. But for words that are most semantically close to the given clue in the mental representation for one player, they should be similarly close to the given clue in the mental representation for other players, i.e., different players can have different perceptions of the relative (lack of) semantic closeness of `drinking container` and `chair`, but they should have similar perception of the semantic closeness of `drinking container` and `cup`, even though the "closest" word to the clue `drinking container` may be `chalice` to some other player. More formally, we can assume that, if for player $j$ and a clue $\mathsf{p}$,

$$\mathsf{w} \in \underset{\mathfrak{u}}{\text{arg-}k\text{-max}}(\Phi_j(\mathfrak{u}) \cdot \Phi_j(\mathsf{p}))$$

then for other $j'$, there exists a constant $\epsilon$ such that

$$\begin{aligned}&(1-\epsilon)(\Phi_j(\mathsf{w}) \cdot \Phi_j(\mathsf{p}))\\ &\leq (\Phi_{j'}(\mathsf{w}) \cdot \Phi_{j'}(\mathsf{p}))\\ &\leq (1+\epsilon)(\Phi_j(\mathsf{w}) \cdot \Phi_j(\mathsf{p}))\end{aligned}$$

This is why this gameplay is a valuable benchmark to explain the effect of incorporation of socio-cultural-educational aspects of an agent - different human players have their unique learning histories and are familiar with different sets of vocabularies. In a real human gameplay, clues do involve shared personal information that the setter does not know (such as `Sport I played in high school`, when `B` is revealed, where the setter does not but some other guesser knows the answer is `badminton`, despite `baseball` or `basketball` being more 'obvious' answers.

An ideal clue that is revealed by a player should not be too vague nor too obvious. A clue like `capital of France` will be immediately blocked by the Setter as `Paris` (unless `Paris` is the answer) and a clue like `a particular animal` is highly likely to invoke different animals as the "closest" word to the given clue in different players' mental representations. Hence, formally if the player $j$ thinks of a word $\mathsf{w}$ and clue $\mathsf{p}$, then

$$\lambda_L < \Phi_j(\mathsf{w}) \cdot \Phi_j(\mathsf{p}) < \lambda_U$$

(clue is not too obvious, but not too unrelated to the word in mind). If $k > 1$, for $\mathsf{u} \neq \mathsf{w}, \mathsf{u} \in \underset{\mathfrak{u}}{\text{arg-}k\text{-max}}(\Phi_j(\mathfrak{u}) \cdot \Phi_j(\mathsf{p}))$, other words that are sufficiently close to the clue,

$$\Phi_j(\mathsf{u}) \cdot \Phi_j(\mathsf{p})) < \lambda_U$$

as well (the other players should not guess an incorrect connection)!

For word $\mathfrak{u}_i$, once letters $1, 2, \ldots, k$ have been revealed, future words which clues need to be generated for need to have $\mathsf{s}_{i1}\mathsf{s}_{i2}\ldots\mathsf{s}_{ik}$ as a prefix. This constrains the sample space in which new words are generated probabilistically. We obtain a probabilistic explanation of the heuristic to generate a clue for a word, assuming that the AI agents are uniform (i.e., the guesser who comes up with a clue has no additional information regarding the semantic alignment of other agents). Given a partially revealed word, assume that the guesser is agnostic as to which word in their vocabulary adhering to the constraints, they actually choose to give a clue for. Then, given a clue for this word, let the probability of a given agent guessing it correctly is $p$. Note that every agent gets to guess once in a given round/turn. We want the setter to not guess correctly (probability $1 - p$) and at least one of the $n - 1$ other guessers to guess correctly. Hence, the probability of success in this run (a new letter is revealed) is $(1 - p)(1 - (1 - p)^{n-1})$.

Maximizing this, we see that we want

$$p^* = 1 - \left(\frac{1}{n}\right)^{\frac{1}{n-1}}$$

For $n = 2, 3, 4, 5$, we have $p^* \approx 0.5, 0.43, 0.37, 0.33$. More the players involved, the clue giver can give a clue somewhat vaguer as decreasing the odds of the setter blocking the word is overcome by the odds of some guesser getting it. The procedure of generating a suitable clue thus has a mental model where the degree of semantic overlap/connection between the clue and the word (i.e., the dot product) can be monotonically mapped to the probability of success for the guessers overall. Clue generation needs a mental proxy to find a roughly optimal clue.

We propose that such wordplay games are important for another social intelligence ability of AI agents - *the ability to gauge a person's background over runs*. This is an important step, we opine, in making AI have intelligence abilities of humans that go beyond deductive reasoning, for the act of "finding" an optimal clue and an optimal word assuming uniform agents is a mathematical function. However, over arbitrarily many turns of the game, an agent should be able to detect which words some agents are more likely to get connections with and less likely with. For example, over turns, it is possible for humans to gauge who has a better grasp of TV shows and who has a better grasp of medical terminology, and choose accordingly. Let the working vocabularies for player $i$ be $\mathfrak{U}_i$. Let $\mathfrak{U}^\dagger = (\cup_{i=1}^{n} \mathfrak{U}_i) \setminus \mathfrak{U}_0$ be the vocabulary that is common knowledge among the non-setter players but unknown to the setter. A good clue targets $\mathfrak{U}^\dagger$.

We can also describe the word and clue generation procedure probabilistically. In (Arora et al., 2016), the authors build on (Mnih and Hinton, 2007) and describe a generative model that treats corpus generation according to a log-linear production model where the probability of generation of a word is proportional to the exponential of the dot product with the discourse vector, which is undergoing a slow random walk and $v_w$ is the embedding vector of word $w$. Analogously, consider the "true" discourse vector for every agent $j$ as $d_j^*$. If the agent knew the discourse vectors for every other agent, then the word and clue generation procedure could be "aligned" towards the discourse vector average for guesser agents and away from the discourse vector average for setter agents by considering the truncated probability distribution (choosing the top $k$ probability words), where $v_w$ is generated by the $\Phi_i$ mapping as mentioned before.

$$\mathbf{P}[\text{word } w] \propto \exp(\langle v_w, d_{\text{setter:avg}} \rangle - \langle v_w, d_0 \rangle)$$

But, when an agent encounters a new agent, the social intelligence of gauging the cognitive abilities of the new agent is as follows:

- Let $d_{i \leftarrow j}$ be the discourse vector of agent $j$ as perceived by agent $i$. For all $j$ and a fixed $i$, let this perceived discourse vector be set to $d^\dagger$, which is the discourse vector corresponding to "common knowledge", the basic ideas people know about.
- A word and the clue is generated by the generative process. Initially, $d_{i \leftarrow \text{setter:avg}}$ and $d_{i \leftarrow 0}$ are both set to $d^\dagger$, so the probability of generating a word is uniform, i.e., "any" random word and clue is generated.
- Let $\eta$ be a suitably small constant. Each word $w$ has its corresponding embedding vector $v_w$. If agent $j$ fails to guess the clue for the word, we subtract $\eta v_w$ from the current value of $d_{i \leftarrow j}$, and if agent $j$ correctly guesses the clue for the word, we add $\eta v_w$ to the current value of $d_{i \leftarrow j}$. $d_{i \leftarrow \text{setter:avg}}$ is recalibrated accordingly.
- This gradient descent/ascent-like procedure mimics the procedure of understanding the cognitive backgrounds of other agents using social intelligence, as it allows future turns to generate clues more likely to be understood by guessers and less so by the setter by updating the perceived values of $d_{i \leftarrow j}$.
- Likewise, when one guesser player $i$ tries a clue for a word $w$, other guesser players, at the end of the turn, add $\eta v_w$ to the perceived discourse vector $d_{j \leftarrow i}$, as guesser $i$ must have their own discourse vector more aligned with the word $w$.

## 3 Connections

### 3.1 Semantic Networks

The key observation is that games utilizing semantic networks do not have a strict notion of monotonicity - i.e., it is not that one player agent strictly *dominates* another player agent by correctly guessing the word to a clue whenever the other agent guesses it correct, or a notion of probabilistic monotonicity, i.e., it is not that one player agent has strictly higher probability of correctly guessing the word to a clue compared to another agent. Similar to negotiation games, (Davidson et al., 2023), "powerful" agents can lose to weaker ones.

Semantic networks serve as the representational basis of our cognitive system as prominent models of memory and reasoning. Semantic networks represent knowledge through relations between abstract objects (Borge-Holthoefer and Arenas, 2010) and could vary considerably between individuals and with different life experiences(Benedek et al., 2017; Dubossarsky et al., 2017; Morais et al., 2013) The goal of a game like Connections is to elucidate, through an iterative process, the discovery of sections of the semantic networks of diverse players that are structurally similar by the task of proposing a set of valid clues that should elicit the same response that must also vary from that of the Setter. Using LLMs for game play requires semantic priming through a prompt to identify the specific role of the agent since the base model for each agent may be the same.

#### 3.1.1 Knowledge-Based Clues

Semantic clues would most often rely on having a shared knowledge about a certain concept or idea. Such clues should be effective to identify the target word precisely if other players have the requisite knowledge. For instance the clue “Speeds a chemical reaction" is precipitated on some other player having some knowledge of chemical reactions. The domain and technical depth of clues produced is linked to the depth and breadth of vocabulary that players may share(Vermeer, 2001). In the human player setting, if two players are able to discover a shared area of knowledge which the Setter is not familiar with the game is significantly simplified. The players can now use this to produce clues that will not be blocked, given they may have a sufficiently rich vocabulary to be able to play the game as the prefix elongates.

#### 3.1.2 Personal Clues

A very effective means of giving clues is those that are built on shared experiences of players or based on some level of knowledge that players have about other players. Such clues are expected to largely be successful as they may not hold any meaning to the Setter or even other players but would allow those who have the appropriate context to guess a word effectively. However, as the space of words gets restricted as more letters are revealed, it is unlikely that there might be adequate personal clues that exist. I.e. forming a personal clue to the prefix “C" might be easy as “ What pet I have" but not to the prefix “CATA".

#### 3.1.3 Word Association Clues

Words have associative links to each other to varying degrees, both semantically, structurally as well in the cognitive representation in memory (Karwoski and Schachter, 1948), (Kent and Rosanoff, 1910). Word associative clues may be used in the game such as “Good-Bad", “Hero-Villain", “Sun-Moon" that leverage this associative structure. Often words exist where the most frequent response to a given clue is several times as frequent as the next-most-frequent response (Woodworth and Schlosberg, 1954) or there are well developed word association norms that have (Palermo and Jenkins, 1965; Nelson et al., 2004; Toglia and Battig, 1978). A key challenge in playing the game with associative clues is that though such clues elicit the correct response it is likely the Setter will easily be able to block such clues.

## 4 Experiments

We run experiments with three players using the **GPT-4o** model as the language model powering the reasoning capabilities of the players. We assign the role of the Setter to Player 0, and Player 1 and Player 2 play the roles of the Guesser. Initially, Player 0, decides a word and reveals only the first letter to players after which the game play proceeds as described. The results of running this game play over a few iterations are recorded in Table 1. The ‘Reveals’ column describes the number of successful letter reveals that happened before the word was guessed, while the ‘Guesser wrong’ and ‘Setter Blocked’ columns respectively describe the cases where either the other player guessed the response to the clue incorrectly or the clue was blocked by the setter.

| Word | Reveals | Guesser Wrong | Setter Blocked | Iterations |
|---|---|---|---|---|
| kaleidoscope | 0 | 1 | 7 | 8 |
| xenophobia | 1 | 2 | 4 | 7 |
| labyrinthine | 9 | 26 | 12 | 47 |
| uppercases | 8 | 20 | 18 | 46 |
| entrepreneur | 4 | 12 | 26 | 42 |
| laboriously | 5 | 21 | 23 | 49 |
| encyclopedia | 5 | 17 | 38 | 60 |
| villeinage | 4 | 30 | 40 | 74 |
| exploration | 5 | 54 | 22 | 81 |
| photosystem | 6 | 31 | 48 | 85 |
| elaborately | 7 | 48 | 35 | 90 |
| goldfish | 6 | 42 | 53 | 101 |
| revolving | 6 | 58 | 37 | 101 |
| precaution | 5 | 54 | 47 | 106 |
| multinomial | 4 | 45 | 59 | 108 |
| precipitate | 9 | 39 | 68 | 116 |
| metamorphosis | 7 | 39 | 71 | 117 |
| circumvented | 9 | 83 | 72 | 164 |
| conjunction | 5 | 123 | 45 | 173 |

Table 1: Comparisons Across Different Words, Two Guessers, One Setter (Ordered by Iterations)

We illustrate a sample of the full interaction between agents in Appendix A. We used a variety of prompts to guide the actions of the LLM agents which are detailed in Appendix B.

### 4.1 Reasoning Capabilities

We observe that players showcase interesting reasoning capabilities. Certain behaviors are well expected and similar to human agents such as the relatively smaller number of iterations observed for the word starting with 'X' as opposed to a much higher number of iterations of game play for words starting with 'C'. Though there is a high degree of chance in guessing the right word, the set of word beginning with 'X' is much smaller than with 'C'. 1 showcases the number of letters revealed over the number of iterations before the word is correctly guessed. We observe that models generally tend to need fewer iterations beyond the first few letter guesses as the constrained space of words given a longer prefix is smaller, however, an interesting non-human behaviour that is observed is that sometimes there is a lot of iterations between the final letter being revealed and the word being guessed often as a result of the model making bizarrely uncommon guesses prior to make guesses that are far more obvious.

### 4.2 Social Awareness

In standard game play, we observe that an overwhelming number of clues are blocked by the Setter, since it is likely given that we are using the same language model, the semantic network to represent words for all agents is quite similar. However, by introducing in-context learning(Lampinen et al., 2022), we can instigate varying behaviours. Notably, by priming the agent on aspects such as profession, cultural context, age, we can instigate a change in the semantic network of the model(Benedek et al., 2017; Dubossarsky et al., 2017; Morais et al., 2013). Under such a framework, we observe the AI agents' ability to utilize this context in the game play through the use of concerted vocabulary and clues if agents are explicitly made aware of the the priming for other agents. However, unlike humans, AI agents may be unable to gauge this context, or may fail to act in a way to discover this information through the use of diverse clues to attempt to find regions of the semantic network that might be similar between agents without any prompting.

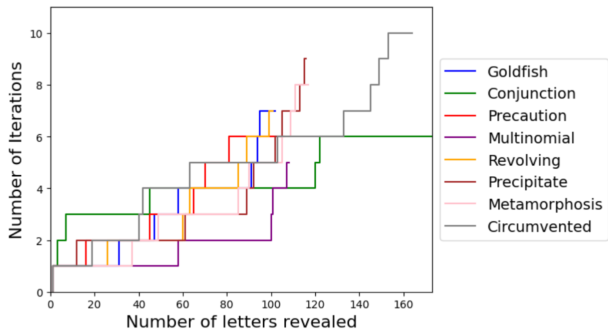


Figure 1: Number of Characters Revealed over iterations for words with more than 100 iterations of game play

## 5 Future Work

Our work considers a simple game where the word picked by the Setter is fixed. A more sophisticated version of this game, and one that is common in game-play is where the word is not necessarily fixed (not written down before) but has to agree with the game played so far. For example, the Setter in this particular game-play may realize after the letters 'C', 'A','T' and 'A' being revealed that once the letter 'M' is revealed, there are not enough words besides the chosen word 'catamaran' (and the Setter can no longer block it). However, if the Setter now switches to 'cataclysm', the Setter has an advantage because there are other words starting with 'catac', such as 'catacomb'. The Setter wants to pick a word that is sufficiently obscure but prefixes do not give the word away easily. Modifying our experiments to not have the word be fixed *a priori* has limited success as the LLM is not able to robustly recall the game history to make improvisational decisions. Future work can explore the proof-of-concept alignment of perceived discourse vectors of other agents, given AI agents with a variety of historical training/life experiences that goes beyond priming. A LLM agent endowed with social intelligence capabilities shows a first step towards more complex reasoning patterns that require sensory and cognitive abilities as well as the power of agency to autonomously adapt to changing scenarios. (Liu, 2017)

## Acknowledgments

The authors wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. The authors wish to confirm that there are no known ethical concerns associated with this publication.

## References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Mathias Benedek, Yoed N Kenett, Konstantin Umdasch, David Anaki, Miriam Faust, and Aljoscha C Neubauer. 2017. How semantic memory structure and intelligence contribute to creative thought: A network science approach. *Thinking & Reasoning*, 23(2):158–183.

Javier Borge-Holthoefer and Alex Arenas. 2010. Semantic networks: Structure and dynamics. *Entropy*, 12(5):1264–1302.

Roger Chaffin and Douglas J Herrmann. 1984. The similarity and diversity of semantic relations. *Memory & Cognition*, 12:134–141.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Dimitry Davidoff and Andrew Plotkin. n.d. Mafia. https://escaleajeux.fr/?principal=/jeu/mafid. Also known as Werewolf.

Tim Ruben Davidson, Veniamin Veselovsky, Michal Kosinski, and Robert West. 2023. Evaluating language models through negotiations. In *The Twelfth International Conference on Learning Representations*.

Haim Dubossarsky, Simon De Deyne, and Thomas T Hills. 2017. Quantifying the structure of free association networks across the life span. *Developmental psychology*, 53(8):1560.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Theodore F Karwoski and J Schachter. 1948. Psychological studies in semantics: Iii. reaction times for similarity and difference. *The Journal of Social Psychology*, 28(1):103–120.

Grace Helen Kent and Aaron Joshua Rosanoff. 1910. A study of association in insanity. *American Journal of Psychiatry*, 67(1):37–96.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Bing Liu. 2017. Lifelong machine learning: a paradigm for continuous learning. *Frontiers of Computer Science*, 11(3):359–361.

Chuck Lorre and Bill Prady. 2012. The re-entry minimization. Television series episode. Season 6, Episode 4.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648.

Ana Sofia Morais, Henrik Olsson, and Lael J Schooler. 2013. Mapping the structure of semantic memory. *Cognitive science*, 37(1):125–145.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

David S Palermo and James J Jenkins. 1965. Changes in the word associations of fourth-and fifth-grade children from 1916 to 1961. *Journal of verbal learning and verbal behavior*, 4(3):180–187.

Dongqi Pu and Vera Demberg. 2023. Chatgpt vs human-authored text: Insights into controllable text summarization and sentence style transfer. *arXiv preprint arXiv:2306.07799*.

Michael P Toglia and William F Battig. 1978. *Handbook of semantic word norms.* Lawrence Erlbaum.

Anne Vermeer. 2001. Breadth and depth of vocabulary in relation to l1/l2 acquisition and frequency of input. *Applied psycholinguistics*, 22(2):217–234.

Robert Sessions Woodworth and Harold Schlosberg. 1954. *Experimental psychology*. Oxford and IBH Publishing.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. 2020. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, 11(1):1877.

## A Sample Gameplay

Word chosen: 'XENOPHOBIA'

```
Setter : X

Attempt Number 1
Guesser 1 (mentally) : Xylograph
Guesser 1 : Woodblock printing technique
Setter : Xylography
Guesser 2 : Xylography
Setter attempted XYLOGRAPHY , Player attempted XYLOGRAPHY
Guesser Wrong!

Attempt Number 2
Guesser 1 (mentally): Xanthophyll
Guesser 1 : Leaf pigment category
Setter : Xanthophyll
Setter attempted XANTHOPHYLL
Setter Blocked!
BLOCKED: ['XANTHOPHYLL']

Attempt Number 3
Guesser 1 (mentally): Xenoglossy
Guesser 1 : Mysterious language phenomenon
Setter : Xenoglossy
Setter attempted XENOGLOSSY
Setter Blocked!
BLOCKED: ['XANTHOPHYLL', 'XENOGLOSSY']

Attempt Number 4
Guesser 1 (mentally): Xiphophyllous
Guesser 1 : Sword-shaped leaves
Setter : Xiphoid
Guesser 2 : Xiphophyllous
Setter attempted XIPHOID , Player attempted XIPHOPHYLLOUS
Success!
Comparing XIPHOPHYLLOUS with XENOPHOBIA
Game not over yet

Setter : XE

Attempt Number 5
Guesser 1 (mentally): Xenolithic
Guesser 1 : Foreign rock inclusion
Setter : Xenolith
Guesser 2 : Xenolith
Setter attempted XENOLITH , Player attempted XENOLITH
Guesser Wrong!

Attempt Number 6
Guesser 1 (mentally): Xenogenesis
Guesser 1 : Alien life formation
Setter : Xenogenesis
Setter attempted XENOGENESIS
Setter Blocked!
BLOCKED: ['XENOGLOSSY', 'XENOGENESIS']
```

```
Attempt Number 7
Guesser 1 (mentally): Xerophthalmia
Guesser 1 : Dry eye condition
Setter : Xerophthalmia Setter attempted
XEROPHTHALMIA
Setter Blocked!
BLOCKED: ['XENOGLOSSY', 'XENOGENESIS',
'XEROPHTHALMIA']

Attempt Number 8 Guesser 1 (mentally):
Xenophobia
Guesser 1 : Fear of foreigners
Setter (mentally): Xenophobia (Cannot
block anymore)
Guesser 2 : Xenophobia
Player attempted XENOPHOBIA
Success!
Comparing XENOPHOBIA with XENOPHOBIA
Guessed Correctly. Game Over.
Reveals, GuesserWrong, SetterBlocked:
1 , 2 , 4
7
```

## B Prompts

- Prompt to generate a new word:

  You are playing a wordplay game, where you are the setter player playing against 2 guesser players. You have to pick a legitimate English word of some suitable length. The guesser players have to guess the word gradually. For now, pick a totally random word and just output that word without any leading phrases. Do not output anything else, only one single word.

- Prompt explaining the rules to the setter:

  Here are the rules of the game. Initially, you reveal the first letter to all guesser players. In every round, one guesser player will come up with a suitable clue phrase whose answer begins with your revealed letter(s). If you can guess the answer to their clue and it is not the same as the word you had picked, then you will output in that round the answer to their clue. If your guess is the same as the answer to the clue a guesser player came up with, then there is a new round. If your guess is different than the answer to the clue a guesser player came up with, a different guesser player will try to guess the clue word. If this different guesser player guesses the answer correctly, then you will have to reveal the next character of the word. In future rounds, the other guesser players have to come up with clues whose answers begin with the same letters as the characters you have revealed so far. Do you understand? Output Yes or No, just that.

- Prompt explaining the rules to the guesser:

  You are playing a game with other guesser players against a setter player who is slowly revealing the initial letters of the word. The setter player will initially tell you just the first letter of the word. Each round, you can find a random word that starts with the initial letters revealed so far. Then, you need to come up with a meaningful clue or a description of this word and reveal it to other guesser players. You are not allowed to have a clue that is very similar to the word itself. If the word you found is not the same as the word that the setter came up with, the setter will try to guess your word and block it by saying your word. If some other guesser player can correctly guess your word, then the setter player will reveal one more letter. If the word both the guesser players guessed is the same as the word the setter player came up with, you all win. In every round, you can either choose to make a clue or try to guess from some other guesser player's clue. Note that in every round, your word must start with the initial letters revealed so far. Do you understand? Output Yes or No, just that

- Prompt to generate a word from a clue:

  You have been given the clue . . . . Now, guess a single word that could be a possible answer to this clue, starting with the letters . . . . Make sure your word is NOT one of these words: . . . and is different. Just output this word, do not output anything else.

- Prompt to generate a clue:

  The partial word you know so far is. . . . Come up with a word that starts with . . . . Make sure your word is NOT one of these words: . . . and is different. Just output this word, do not output anything else.”

- Prompts to correct an agent:
  - Your earlier word does not start with . . . . Try again. Come up with a word that starts with . . . . Make sure your word is NOT one of these words: . . . and is different. Just output this word, do not output anything else.
  - Your earlier word does not start with . . . . Try again. You have been given the clue . . . . Now, guess a single word that could be a possible answer to this clue, starting with the letters . . . . Make sure your word is NOT one of these words: . . . and is different. Just output this word, do not output anything else.
  - Your earlier word cannot be one of these words: . . . . Try again. Come up with a word that starts with . . . . Make sure your word is NOT one of these words: . . . and is different. Just output this word, do not output anything else.
  - Your earlier word cannot be one of these words: . . . . Try again. You have been given the clue . . . . Now, guess a single word that could be a possible answer to this clue, starting with the letters . . . . Make sure your word is NOT one of these words: . . . and is different. Just output this word, do not output anything else.