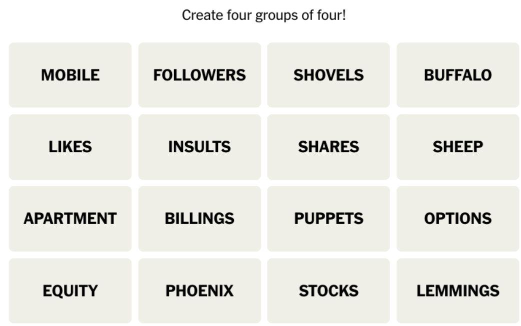
Connecting the Dots: Evaluating Abstract Reasoning Capabilities of LLMs Using the New York Times Connections Word Game

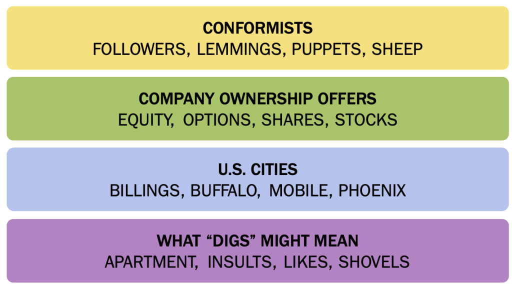
Anonymous ACL submission

Abstract

The *New York Times* Connections game has emerged as a popular and challenging pursuit for word puzzle enthusiasts. We collect 200 Connections games to evaluate the performance of state-of-the-art large language models (LLMs) against expert and novice human players. Our results show that even the best-performing LLM, GPT-4o, which has otherwise shown impressive reasoning abilities on a wide variety of benchmarks, can only fully solve 8% of the games. Compared to GPT-4o, novice and expert players perform better, with expert human players significantly outperforming GPT-4o. We create a taxonomy of the knowledge types required to successfully cluster and categorize words in the Connections game, revealing that LLMs struggle with associative, encyclopedic, and linguistic knowledge. Our findings establish the *New York Times* Connections game as a challenging benchmark for evaluating abstract reasoning capabilities in humans and AI systems.



(a) The unsolved connections game presented to a player



(b) The solved connections game with distinct categories sorted according to levels of difficulty—straightforward (yellow) to tricky (purple)

Figure 1: Example from a *NYT* Connections game

1 Introduction

Word puzzle enthusiasts have become captivated by Connections, an engaging game launched by the *New York Times* (*NYT*) in June 2023. This daily game presents players with a 4x4 grid containing 16 words and tasks them to identifying four distinct clusters that link the corresponding four words in each cluster through some shared characteristics (Figure 1 [a] vs [b]). Despite its seemingly straightforward premise, Connections delivers a stimulating linguistic workout that keeps players returning daily to test their mental acuity and semantic savvy. The categories 1 (yellow), 2 (green), 3 (blue), and 4 (purple) are arranged according to ascending levels of difficulty. Category 1 is the most intuitive, while Category 4 is the hardest. For instance, in Figure 1 (b), the most straightforward category is "Conformists" *{Followers, Lemmings, Puppets, Sheep}*, while the most challenging category includes *{Apartment, Insults, Likes, Shovels}* and requires the understanding that a single word (in this case, "digs") can have multiple meanings that differ in etymology or sense, depending on the context.

While the task may appear easy, many words clump easily into categories, acting as red herrings. For instance *Likes, Followers, Shares, Insult* might be categorized as “Social Media Interactions" at first glance. However, the game is designed to promote orthogonal thinking and pushes players to find unusual ways to group things. To group words across proper categories, as shown in Figure 1 (b), a player must reason with various forms of knowledge spanning from *Semantic Knowledge* (Conformists) to *Encyclopedic Knowledge* (U.S. cities).

Abstract reasoning represents a person's ability to solve problems, identify patterns, and work with logical systems (Barrett et al., 2018; Johnson et al., 2021). While the performance of large language models (LLMs) on arithmetic and language-based commonsense reasoning benchmarks has been the subject of recent analyses, it is unclear whether these LLMs possess abstract reasoning capabilities that are often challenging even for humans (Xu et al., 2023). Given its nature, we choose the *NYT* Connections Game as a test bed for investigating the abstract reasoning capabilities of both humans and large language models (LLMs). We collect 200 distinct Connection games and test the capabilities of four state-of-the-art large language models, namely Google's Gemini 1.5 Pro (Team et al., 2023), Anthropic's Claude 3 Opus (Anthropic, 2024), OpenAI's GPT4-Omni (OpenAI, 2023), and Meta's Llama 3 70B (AI@Meta, 2024) and compare them with human performance.

While all LLMs can partially solve some of the games, their performance is far from ideal. Experimental evidence shows that with few-shot and chain-of-thought prompting, even the best-performing LLM, GPT-4o, can only solve 8% of the games completely. We recruit human players at novice and expert levels of proficiency and compare their performance to GPT-4o. Our results show that the Connections game serves as a challenging benchmark for reasoning, with novice players performing only marginally better than GPT-4o. On the contrary, expert players perform significantly better than GPT-4o in solving games perfectly (Section 5). To better understand the challenging nature of this benchmark, we create a taxonomy of knowledge required to group words into their respective categories (Section 3.2). Our analysis shows that while LLMs are good at reasoning that involves some types of semantic knowledge, they struggle with other types of knowledge such as associative, encyclopedic, or muti-word expressions. Our code and data will be made available upon publication.

2 Related Work

Advancements in LLMs have led to a growing interest in exploring their potential to take on intricate and conceptual challenges by generating linguistic sequences. These models have shown potential in gaming by serving as players (Wang et al., 2023a; Tsai et al., 2023; Ciolino et al., 2020; Bakhtin et al., 2022; Noever et al., 2020), non-player characters (NPCs) (Park et al., 2023; Urbanek et al., 2019), and generating game content (Todd et al., 2023; Wang et al., 2023b; Sudhakaran et al., 2024; Ammanabrolu and Riedl, 2021).

Recent research has explored applying large language models (LLMs) and other natural language processing (NLP) techniques to solve and generate text-based puzzles. Zhao and Anderson (2023) use LLMs to tackle and create the weekly Sunday Puzzles featured on *National Public Radio (NPR)*. When presented with multiple-choice questions, GPT-3.5 attained an accuracy of up to 50%. However, when asked to generate novel and engaging puzzles, the model encounters challenges. Compared to the Connections game, *NPR*'s weekly puzzles tend to emphasize character-level word transformations and relatively common references, relying less on encyclopedic, associative, or semantic knowledge. Rozner et al. (2021) examined the potential of using "cryptic crossword" clues as an NLP benchmark. Wallace et al. (2022) propose automatic ways of solving crossword puzzles by generating answer candidates for each crossword clue using neural question answering models and combining loopy belief propagation with local search to find full puzzle solutions. Our work builds upon these efforts by utilizing the *NYT* Connections puzzle as a means to investigate the abstract reasoning capabilities of state-of-the-art LLMs.

The word association task (Galton, 1879) has been used extensively in psychological and linguistic research as a way of measuring connections between words in the mental lexicon. Responses in word association tasks have informed what we know about the structure and organization of semantic memory and the mental lexicon (De Deyne and Storms, 2008). In this work, we similarly show how one must utilize semantic and associative memories to solve the Connections game.

Chollet (2019) proposed the Abstraction and Reasoning Corpus (ARC), built upon an explicit set of priors designed to be as close as possible to innate human priors and argued that it can be used to measure a human-like form of general fluid intelligence, enabling fair general intelligence comparisons between AI systems and humans. Recently Xu et al. (2023) show that GPT-4 solves only 13/50 of the most straightforward ARC tasks, demonstrat-

Knowledge		Category	Words
Encyclopedic		TV SHOWS WITH HAPPY-SOUNDING NAMES	['CHEERS', 'EUPHORIA', 'FELICITY', 'GLEE']
		JACKS	['BLACK', 'FROST', 'MA', 'SPARROW']
Semantic	Synonym	COLLEAGUES	['ASSOCIATE', 'FELLOW', 'PARTNER', 'PEER']
	Polysemy	WHAT A MOLE CAN BE	['ANIMAL', 'BIRTHMARK', 'SPY', 'UNIT']
	Hypernym	PERIOD	['AGE', 'DAY', 'ERA', 'TIME']
Associative		ORIGIN	['CRADLE', 'FONT', 'ROOT', 'SOURCE']
		THINGS THAT ARE ORANGE	['BASKETBALL', 'CARROT', 'GOLDFISH', 'PUMPKIN']
Linguistic		NOUN SUFFIXES	['DOM', 'ION', 'NESS', 'SHIP']
		SILENT "W"	['ANSWER', 'TWO', 'WRIST', 'WRONG']
Multiword Expression		___WOOD	['DOG', 'DRIFT', 'HOLLY', 'SANDAL']
Combined		CITY HOMOPHONES	['DELI', 'NIECE', 'ROAM', 'SOUL']
		SOCIAL MEDIA APP ENDINGS	['BOOK', 'GRAM', 'IN', 'TUBE']

Table 1: Different types of knowledge required to group words into their respective categories

ing a significant gap in the abstract reasoning capabilities of LLMs. Prior work has also studied abstract reasoning in Neural Networks (Barrett et al., 2018) even in the presence of distracting features (Zheng et al., 2019). Our work builds upon these and presents the Connections game as a compelling benchmark for abstract reasoning capabilities for LLMs in the presence of distractors.

3 Data

3.1 Collection

To gather the necessary data, we found an archival site consisting of all possible answer choices and their corresponding categorizations. As the *NYT* does not maintain an archive of Connection puzzles, we resorted to an external, third-party site for data collection.[1] Our data spans daily problems from the conception of Connections, June 2023, to March 2024. In total, we gather 203 distinct games, out of which 3 are used for few-shot prompting, while the remaining 200 comprise the dedicated test set.

3.2 Types of Reasoning

Investigating the relationship between words offers insights into both the structure of language and the influence of cognition on linguistic tasks (Stella et al., 2018). To solve Connections games, players must draw on certain aspects of word knowledge, such as a word's meaning. To deepen our understanding, we bucket each <category, grouping> into the types of knowledge that are primarily required to solve them. Two experts annotate a total of 800 samples coming from 200 games into 6 broader categories. On 8.6% of the 800 samples where they disagree (See examples of disagreement in Appendix B), the experts engaged in discussion (Schaekermann et al., 2018; Chen and Zhang, 2023; Chen et al., 2019) to arrive at an individual category.

3.2.1 Semantic Knowledge

The majority of instances in the Connections game require possessing knowledge of *lexical semantics* (Cruse, 1986), particularly semantic relations such as synonymy (words with the same meaning), hypernymy/hyponymy (relation between a generic terms and its specific instance), and polysemy (many possible meanings for a word). Table 1 shows three examples of groups that use such Semantic Knowledge.

3.2.2 Associative Knowledge

To group words into their respective categories one often needs to think beyond the lexical semantic relations mentioned above. Associative learning (Shanks, 1995) occurs when an element is taught through association with a separate, seemingly unrelated pre-occurring element. To cluster words using Associative Knowledge, one either needs to focus on the connotative meaning of a word or the shared property that connects several words. For instance, as shown in Table 1, the words *Cradle, Root,* or *Font* in their literal sense do not refer to *Origin*; instead, one needs to rely on their connota-

[1] https://tryhardguides.com/nyt-connections-answers/

Semantic Knowledge	Associative Knowledge	Encyclopedic Knowledge	Mutiword Expressions	Linguistic Knowledge	Combined Knowledge
337	171	153	77	49	13

Table 2: Breakdown of instances of different types of reasoning across 800 categories from 200 Connections games

tive meaning for such a categorization. Similarly, on the surface level, *Basketball, Carrot, Goldfish,* and *Pumpkin* are unrelated. However, a shared property that connects them is their orange color.

3.2.3 Encyclopedic Knowledge

We notice that to group certain sets of words into their proper categories, one needs knowledge that spans beyond concepts and relies on entities in the real world found in knowledge bases such as Wikipedia (Mihalcea and Csomai, 2007). This can be seen in Table 1, where, to bucket the words *Black, Frost, Ma,* and *Sparrow* into the category of *Jacks*, one needs to possess knowledge across various domains: 'Jack Black' an American actor, 'Jack Frost' a character from English folklore who personifies winter, 'Jack Ma' the founder of Alibaba, and 'Jack Sparrow' the protagonist of the *Pirates of the Caribbean* film series. We label this type of knowledge Encyclopedic Knowledge.

3.2.4 Multiword Expressions

Multiword Expressions are complex constructs that interact in various, often untidy ways and represent a broad continuum between non-compositional (or idiomatic) and compositional groups of words (Moon, 1998). Higher difficulty levels (blue and purple) in the Connections game often require players to recognize that the four words can form a Multiword Expression if combined with an external word. Table 1 shows examples of Multiword Expressions where half of the expressions are given in the form of individual words and the player needs to find the other half to categorize words into the correct group.

3.2.5 Linguistic Knowledge

Linguistic competence (Coseriu, 1985) is the system of unconscious knowledge that one has when one knows a language. Such competence is often required to classify words into their appropriate categories. Several instances from the Connections game require knowledge of morphology, phonology, or orthography for correct categorization. For example, as shown in Table 1, one needs knowledge about morphology to group *Dom, Ion, Ness,* and *Ship* as *Noun Suffixes*. Similarly, one needs to rely on phonological knowledge about the sound patterns of *Answer, Two, Wrist,* and *Wrong* to categorize them as *Silent "W"*.

3.2.6 Combined

Some of the hardest examples in the *NYT* Connections game require reasoning with multiple types of knowledge. For instance, the example in Table 1 shows that to group *Deli, Niece, Roam,* and *Soul*, one requires the knowledge that these words have the same phonological form with the cities Delhi, Nice, Rome, and Seoul. This categorization requires the simultaneous use of Encyclopedic and Linguistic Knowledge. Similarly, to group the words *Book, Gram, In,* and *Tube* together one needs to identify that they are essentially parts of closed compounds (Face+Book, Insta+Gram, Linked+In, You+Tube) that also represent popular social media apps. This categorization requires one to use Encyclopedic and Linguistic Knowledge together.

4 Experimental Settings

4.1 LLMs as Game Players

To test the capabilities of large language models in solving the Connections game, we rely on recent advancements in in-context learning and chain-of-thought prompting (Wei et al., 2022). We provide 3 complete examples in our few-shot prompt along with rules and common strategies that players must use to solve the game. We also elicit chain-of-thought reasoning (Wei et al., 2022) requiring models to explain their groupings and categories chosen. Formulation of the prompt involved trial and error; the first iteration of the prompt included the Connections game instructions provided by the *New York Times* (Liu, 2023a), and included three demonstrations with gold labels asking the LLM to explain its reasoning in a step-by-step manner (Wei et al., 2022). We ran this first prompt with a few games on a development set of 30 games (different from our test set), using the 4 LLMs. After identifying commonalities in the types of errors made by the LLMs while playing the game, we added additional instructions about the game, specified the response format, and included some tips from a *NYT* article about playing Connections (Aronow and Levine, 2023). The entire prompt is in Appendix A. To ensure consistency and fairness in performance, we prompt 4 LLMs — Gemini 1.5 Pro, Claude 3 Opus, GPT-4o, and Llama 3 70B —with the same input and use the default sampling parameters (temperature and top_p). We use the scoring schema outlined in Section 4.3 to evaluate how all models perform in solving 200 Connections games spanning from June 15, 2023 to January 1, 2024.

4.2 Humans as Game Players

Alongside LLMs, we recruited 17 human evaluators in two subgroups: 12 novice players with little to no prior experience playing Connections and 5 expert or regular Connections players. The novice and the expert evaluators were peers of the authors of the paper who volunteered to participate without any payment.

We designed a human evaluation interface and randomly sampled 100 games from our set. Appendix E has more information about the interface. The first screen displays a shortened version of the instructions from the LLM's prompt so as to not overwhelm the human players. To ensure that the humans solve the game in a manner comparable to the LLMs setup, they were given one try to fully solve the game (i.e., make all 4 categorizations).

Playing these games is a significant cognitive burden. As such, each novice human evaluator played around 8-12 distinct games for a total of 100 randomly sampled games out of the 200 games in the test set, and expert participants each played 10 games for a total of 50 games from the subset of 100 games played by novices.

4.3 Evaluation Criteria

Our scoring schema was developed as a means to numerically interpret the outcome of each Connections game and standardize comparison across LLMs and human players. We outline two processes to obtain *clustering* and *categorical reasoning* scores of a game of Connections.

4.3.1 Clustering Score

The clustering score evaluates the ability to correctly group together all the words in the game. We consider two clustering scores. The first, or the *unweighted clustering score*, is calculated independently of the categories' supposed difficulty. In this simple scoring mechanism, we allocate one point for each correct cluster (when all 4 words in the group classified by the LLM or human correspond to the 4 words in the gold category/grouping). Ideally, a player's score should be close to the maximum of 4, signifying that all 4 groups were correctly identified. The equation is as follows:

$$score = n_0 + \ldots + n_3 \qquad (1)$$

where $n_x = 1$ for each correct grouping x and $n_x = 0$ for each incorrect grouping.

The second score, referred to as the *weighted clustering score*, takes into account the difficulty of each grouping. The worst weighted clustering score a player can obtain is 0, meaning that no words were grouped correctly. Ideally, a player's score should be close to the maximum of 10, signifying that all 4 categories were correctly classified. The equation for this score is as follows:

$$score = n_0 \cdot w_0 + \ldots + n_3 \cdot w_3 \qquad (2)$$

where n_x represents one of the 4 categories and is always equal to 1 for each category x. The reward procedures are as follows: $w_0 = 1$ for a Yellow (most straightforward) correct grouping, $w_1 = 2$ for a Green correct grouping, $w_2 = 3$ for a Blue correct grouping, $w_3 = 4$ for a Purple (trickiest) correct grouping. Our schema for the clustering scores does not incorporate the number of tries as a variable, since in our setup the LLMs are prompted once and take one try to solve the game.

4.3.2 Categorical Reasoning

While the weighted and unweighted clustering scores are calculated for LLMs and humans, the *categorical reasoning score* is used only for the LLMs' responses. If all 4 words in a category are correctly identified by an LLM, we conduct further analysis to evaluate whether the LLM reasoned correctly *why* the words in the groups belong together. We make this distinction in our evaluation so that — in conjunction with the taxonomy knowledge for Connections categories (Section 3.2) — we can assess the types of reasoning that the LLMs are most or least adept in. Since our prompt asks the LLM to include the category name and share the reasons why it grouped words together, we can evaluate whether the LLM's reasoning in its response is semantically analogous to the gold *NYT* Connections-provided category name. The decision of semantic equivalence between LLMs output and gold is done manually by a human judge to ensure accuracy.

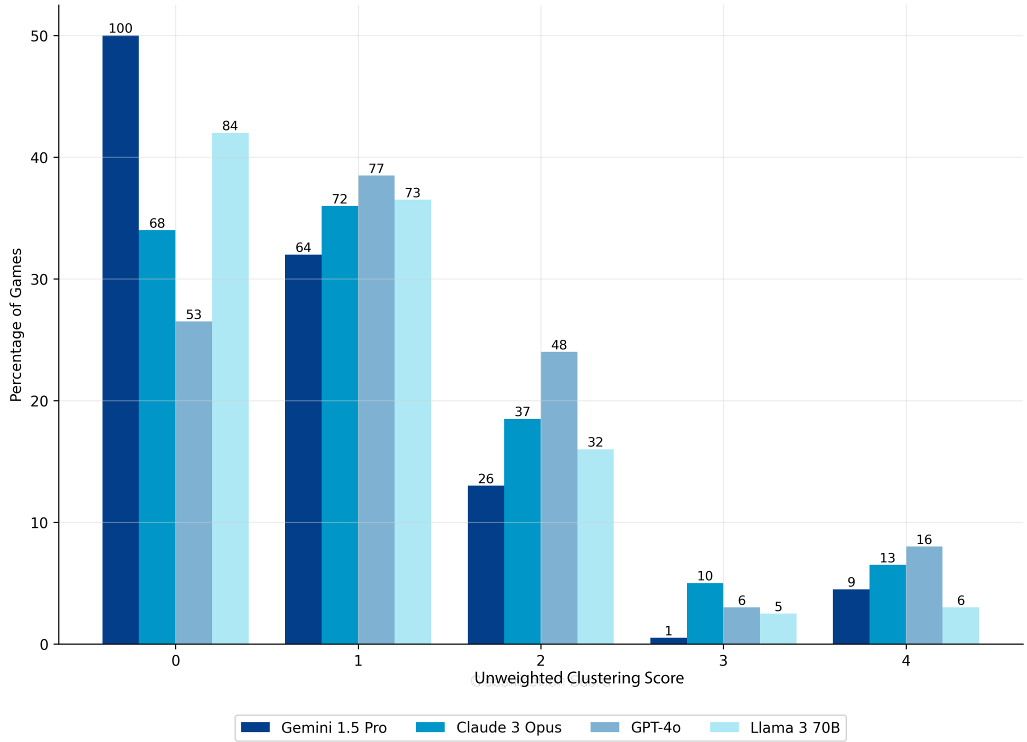


Figure 2: Frequency of unweighted clustering scores for 4 LLMs across 200 games. The number of games in which the respective unweighted clustering score was achieved is atop each bar.

5 Results

5.1 LLM performance

Overall, we find that GPT-4o performs best across all 200 games. Figure 2 shows the unweighted clustering scores for all 4 LLMs. GPT-4o has the lowest percentage of games in which it made no correct clustering (53 out of 200) and the most games solved perfectly (16 out of 200). Claude 3 Opus is a close second for perfectly solved games at 13 out of 200. Gemini 1.5 Pro performs the worst overall. While it could not make any correct clusters for half of the games, it was able to solve 9 games perfectly, outperforming Llama 3 70B's 6 games solved perfectly. In terms of weighted clustering scores for each model (Figure 3), Gemini 1.5 Pro and Llama 3 70B show similar results. Most of their scores are concentrated before 2, showing that these models showed a higher ability to correctly classify the easiest or second easiest categories. GPT-4o and Claude 3 Opus also showed similar results, with most of their weighted clustering scores concentrated before 4, meaning they were better at classifying more and harder categories. Weighted clustering scores ≥ 8 are very rarely represented in all the models. Appendix D.2 contains a more detailed breakdown.

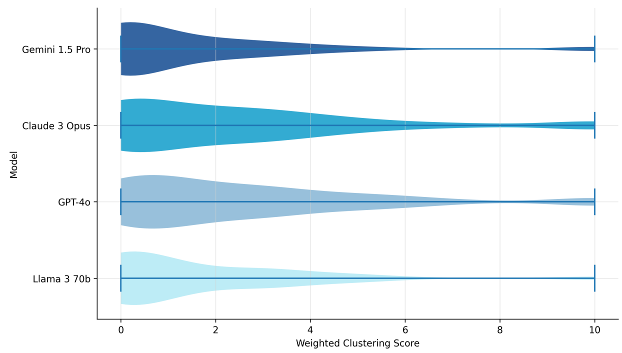


Figure 3: Spread of weighted clustering scores for each model across 200 Connections games

5.2 Human Performance

In human performance, we measure both novice and expert players against the best overall performing GPT-4o. For the 100 games played by novices and 50 games played by experts, we compare the same 100 and 50 games played by GPT-4o.

5.2.1 Novice Players

In the 100 games that the novice players completed, their average unweighted clustering score was 1.38, marginally better than GPT-4o's average of 1.17 on the same 100 games. GPT-4o and novice humans also had similar weighted clustering score distribu-

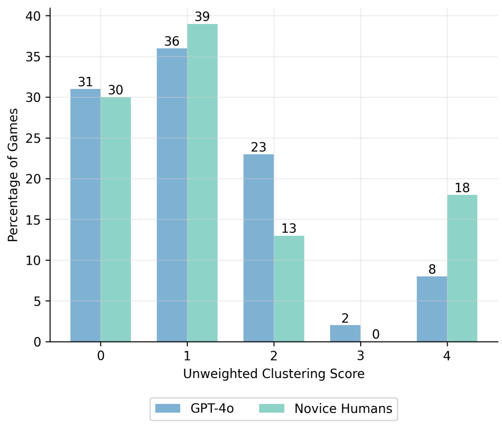


Figure 4: Frequency of clustering scores of GPT-4o and 12 novice humans across 100 games

tions. More details are in Appendix D.1. Due to the setup of the human interface, humans could not receive a clustering score of 3 (if humans correctly solve 3 groupings, the 4th is also correct). Because of GPT-4o's imperfect instruction-following abilities (repeating or omitting a word), it was still able to obtain a clustering score of 3, as shown in Figure 4.

5.2.2 Expert Players

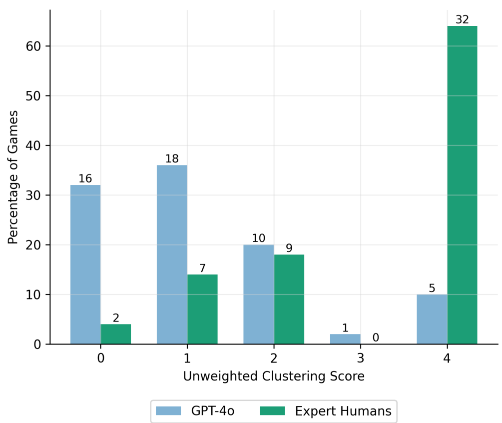


Figure 5: Frequency of clustering scores of GPT-4o and 5 expert humans across 50 games

Expert human players performed significantly better than novices and GPT-4o, with an average clustering score of 3 compared to GPT-4o's 1.22 (on the same 50 games) and an average weighted clustering score of 7.4 compared to GPT-4o's 2.32. The distribution of weighted clustering scores is also far more right-skewed (see Appendix D.1 for more). Figure 5 shows that experts perfectly solve over 60% of the 50 games, while GPT-4o only fully solved 5% of the games.

6 Discussion

6.1 What type of reasoning is hardest for LLMs and humans?

To answer this question we rely on our taxonomy of reasoning types we introduced in Section 3.2. The breakdowns of the reasoning types for the 800 categories in our 200-game dataset are shown in Table 2. The patterns in performance across the types of reasoning parallel the LLMs' overall performance for the most part, with Claude 3 Opus's performance in Multiword Expressions defying this pattern. The performance in reasoning categories across models is ranked from best to worst as follows: $Semantic > Associative > Encyclopedic > Linguistic > Multiword > Combined$. The model's performance in these categories corresponds to the frequency with which each category appears, except the category of Multiword Expressions. Only Claude 3 Opus has a greater than zero success rate (4 out of 77) in Multiword Expressions. In Connections games with Multiword Expressions, they are usually the purple category (most difficult), while semantic and associative reasoning appear most often as yellow or green (easier) categories. For Combined categories, no models output correct clusters.

We find that both novice and expert human players are better at all types of reasoning compared to GPT-4o, although neither cluster any Combined Knowledge categories successfully. Novice performance in reasoning categories across models is ranked from best to worst as follows: $Semantic \approx Linguistic > Associative > Encyclopedic > Multiword > Combined$. The greatest difference between GPT-4o and novices is in Multiword Expressions (difference of around 20%) and Linguistic Knowledge (difference of 25%). Expert performance ranked from best to worst: $Semantic > Encyclopedic > Linguistic > Associative > Multiword > Combined$.

For both LLMs and humans, the Combined Knowledge category seems hardest to grasp. Though experts rank low in Multiword Expressions compared with other types of reasoning, they still perform quite highly, with over 60% of Multiword Expression categories grouped correctly. However, perhaps because of a lack of familiarity with the game and types of categories, novice players, like LLMs, struggle with Multiword Expression, achieving an accuracy of just over 20%. Further breakdowns are in Appendix D.

6.2 How much do distractors prevent both LLMs and humans from correct categorization?

The Connections game is often formulated with item overlap in mind, according to the Connections puzzle creator (Liu, 2023b). These distractors, or red herrings, make the game far more challenging. Red herrings can appear in two ways — as a *red herring category* or *red herring word*. In the former instance, 3 ultimately unconnected words seem to form a category of their own with 1 word missing. In the latter, a category seems applicable to more than 4 words, but the extras belong to a separate grouping. Examples of each of these types of red herrings are in Appendix C.

Mistakes resulting from red herrings often occur in categories related to Associative Knowledge. Though the words may be associated in one dimension, the LLMs fail to conduct step-by-step reasoning to find another, perhaps more obscure, grouping (in the case of red herring categories) or the outlier (in the case of red herring words).

6.3 How often do LLMs group the words correctly but present incorrect reasons?

To measure the disparity between LLMs making correct clustering and providing the correct reasoning or category name for their choice, we use a measure calculated from the clustering and categorical reasoning scores. Since the categorical reasoning score is the number of categories reasoned correctly and the clustering score considers whether the grouping was correct independent of the reason behind it, $\frac{\text{unweighted clustering}}{\text{categorical reasoning}}$ tells us how common it is for LLMs to cluster categories correctly by chance. The average ratios in Table 3 are

Model	Average Ratio
Gemini 1.5 Pro	0.78
Claude 3 Opus	0.87
GPT-4o	0.86
Llama	0.76

Table 3: Average categorical reasoning to unweighted clustering score ratio by model

fairly high, close to or above 80%. The highest overall performing models GPT-4o and Claude 3 Opus have the highest ratios as well. Though it is fairly uncommon that a model will correctly group without correctly reasoning, there are very few instances where models received both a clustering score of 4 (fully solved game) and a categorical reasoning score of 4.

6.4 How can future work improve on such a benchmark?

Certain strategies grounded in reflective thinking could improve performance on such a benchmark. Instead of greedily choosing the first grouping, identifying red herring words or categories first could prevent the possibility of misclassification. Allowing LLMs to solve the game one category at a time and incorporating the feedback present to humans in the *NYT* Connections game, including whether a grouping is correct (and what difficulty level it is by color), incorrect, or one word away from a correct grouping, may improve performance as well. Retrieval Augmentation from WordNet or dictionaries for lexical connotations (Allaway and McKeown, 2020) could further improve such categorization. Finally, creating synthetic training data and training an LLM on this task could further close the gap between expert human and LLM performance. We leave such exploration for future work.

7 Conclusions

We introduced NYT Connection games as a benchmark to test abstract reasoning in state-of-the-art LLMs and evaluate their performance against expert and novice human players. We find that GPT-4o performs best, although it is still no match for expert human players. By examining the performance through our knowledge taxonomy, we obtain a more solid understanding of areas in which LLMs can improve to solve classification tasks. They are fairly deficient in certain types of reasoning required to be a skilled Connections player. Although most possess adequate semantic and associative reasoning capabilities, they struggle with Multiple Expressions and Combined Knowledge categories. Additional struggles arise because they cannot identify red herrings and use step-by-step logic to work around them. Ultimately, we find that excelling in Connections means having a berth of different knowledge types, and LLMs are unfortunately not yet suited for the task.

8 Limitations

Many of the limitations in this section stem from the lack of data available for Connections games and disparities in the comparison between LLMs and humans. Because it is a fairly recent invention and only one puzzle is released per day, there are only a few hundred games available. Since there are some category patterns learned through frequent play, ideally, a model trained on past Connections games might bridge the gap between LLM are expert human evaluators performance.

We acknowledge that human evaluators were not required to add justifications for the groupings they made. This could have made performance comparisons between humans and LLMs for the types of reasoning more equal. Additionally, because a score of 3 was impossible in the human evaluation interface, we cannot be certain that humans were adept in the type of knowledge of their last category grouped, as this could simply been a matter of grouping all options left. Other limitations of human evaluators include that because they were all peers or acquaintances of the paper's authors, sampling bias could exist. Though the age range of the humans recruited was 14-60, other demographic factors that may not have been accounted for in this sample.

9 Ethical Considerations

We collect the names of users in the human evaluation game's database simply for logistical purposes. Other than this, no personal data is collected. The data collected and its purpose were verbally conveyed to each evaluator before asking for their consent. We remove the names of evaluators in the data release. Besides this, we ensure that now and in the future, any data collection is transparent with users and is used in an ethical and responsible manner. Since our research primarily evaluates reasoning in a game environment, there are fewer potential real-world risks of its applications. However, biases in LLMs may be reproduced.

References

AI@Meta. 2024. Llama 3 model card.

Emily Allaway and Kathleen McKeown. 2020. A unified feature representation for lexical connotations. *arXiv preprint arXiv:2006.00635*.

Prithviraj Ammanabrolu and Mark O Riedl. 2021. Modeling worlds in text. *arXiv preprint arXiv:2106.09578*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. `https://www.anthropic.com/news/claude-3-family`.

Isaac Aronow and Elie Levine. 2023. How to Line Up a Great Connections Solve. *The New York Times*.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. 2018. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pages 511–520. PMLR.

Quan Ze Chen and Amy X. Zhang. 2023. Judgment sieve: Reducing uncertainty in group judgments through interventions targeting ambiguity versus disagreement. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Quanze Chen, Jonathan Bragg, Lydia B. Chilton, and Dan S. Weld. 2019. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Matthew Ciolino, Josh Kalin, and David Noever. 2020. The go transformer: natural language modeling for game play. In *2020 Third International Conference on Artificial Intelligence for Industries (AI4I)*, pages 23–26. IEEE.

Eugenio Coseriu. 1985. Linguistic competence: What is it really? *The Modern Language Review*, 80(4):xxv–xxxv.

D Alan Cruse. 1986. *Lexical semantics*. Cambridge university press.

Simon De Deyne and Gert Storms. 2008. Word associations: Network and semantic properties. *Behavior research methods*, 40(1):213–231.

Francis Galton. 1879. PSYCHOMETRIC EXPERIMENTS. *Brain*, 2(2):149–162.

Aysja Johnson, Wai Keen Vong, Brenden M Lake, and Todd M Gureckis. 2021. Fast and flexible: Human program induction in abstract reasoning tasks. *arXiv preprint arXiv:2103.05823*.

Wyna Liu. 2023a. Connections - How to Play.

Wyna Liu. 2023b. How Our New Game, Connections, Is Put Together. *The New York Times*.

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242.

Rosamund Moon. 1998. *Fixed Expressions and Idion1s in English: A Corpus-Based Approach*. Oxford University Press.

David Noever, Matt Ciolino, and Josh Kalin. 2020. The chess transformer: Mastering play using generative language models. *arXiv preprint arXiv:2008.04057*.

OpenAI. 2023. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Josh Rozner, Christopher Potts, and Kyle Mahowald. 2021. Decrypting cryptic crosswords: Semantically complex wordplay puzzles as a target for nlp. *Advances in Neural Information Processing Systems*, 34:11409–11421.

Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

David R Shanks. 1995. *The psychology of associative learning*. Cambridge University Press.

Massimo Stella, Nicole M Beckage, Markus Brede, and Manlio De Domenico. 2018. Multiplex model of mental lexicon reveals explosive learning in humans. *Scientific reports*, 8(1):2259.

Shyam Sudhakaran, Miguel González-Duque, Matthias Freiberger, Claire Glanois, Elias Najarro, and Sebastian Risi. 2024. Mariogpt: Open-ended text2level generation through large language models. *Advances in Neural Information Processing Systems*, 36.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo

Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Graham Todd, Sam Earle, Muhammad Umair Nasir, Michael Cerny Green, and Julian Togelius. 2023. Level generation through large language models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, pages 1–8.

Chen Feng Tsai, Xiaochen Zhou, Sierra S Liu, Jing Li, Mo Yu, and Hongyuan Mei. 2023. Can large language models play text games well? current state-of-the-art and open questions. *arXiv preprint arXiv:2304.02868*.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. *arXiv preprint arXiv:1903.03094*.

Eric Wallace, Nicholas Tomlin, Albert Xu, Kevin Yang, Eshaan Pathak, Matthew Ginsberg, and Dan Klein. 2022. Automated crossword solving. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3073–3085, Dublin, Ireland. Association for Computational Linguistics.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Ruoyao Wang, Graham Todd, Xingdi Yuan, Ziang Xiao, Marc-Alexandre Côté, and Peter Jansen. 2023b. ByteSized32: A corpus and challenge task for generating task-specific world models expressed as text games. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13455–13471, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias Boutros Khalil. 2023. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *Transactions on Machine Learning Research*.

Jingmiao Zhao and Carolyn Jane Anderson. 2023. Solving and generating npr sunday puzzles with large language models. *arXiv preprint arXiv:2306.12255*.

Kecheng Zheng, Zheng-Jun Zha, and Wei Wei. 2019. Abstract reasoning with distracting features. *Advances in Neural Information Processing Systems*, 32.

A Prompt

Solve today's NYT Connections game. Here are the instructions for how to play this game:
Find groups of four items that share something in common.
Category Examples:
FISH: Bass, Flounder, Salmon, Trout
FIRE ___: Ant, Drill, Island, Opal
Categories will always be more specific than '5-LETTER-WORDS', 'NAMES', or 'VERBS.'

Example 1:
Words: ['DART', 'HEM', 'PLEAT', 'SEAM', 'CAN', 'CURE', 'DRY', 'FREEZE', 'BITE', 'EDGE', 'PUNCH', 'SPICE', 'CONDO', 'HAW', 'HERO', 'LOO']
Groupings:

1. Things to sew: ['DART', 'HEM', 'PLEAT', 'SEAM']
2. Ways to preserve food: [CAN', 'CURE', 'DRY', 'FREEZE']
3. Sharp quality: ['BITE', 'EDGE', 'PUNCH', 'SPICE']
4. Birds minus last letter: ['CONDO', 'HAW', 'HERO', 'LOO']

Example 2:
Words: ['COLLECTIVE', 'COMMON', 'JOINT', 'MUTUAL', 'CLEAR', 'DRAIN', 'EMPTY', 'FLUSH', 'CIGARETTE', 'PENCIL', 'TICKET', 'TOE', 'AMERICAN', 'FEVER', 'LUCID', 'PIPE']
Groupings:

1. Shared: ['COLLECTIVE', 'COMMON', 'JOINT', 'MUTUAL']
2. Rid of contents: ['CLEAR', 'DRAIN', 'EMPTY', 'FLUSH']
3. Associated with "stub": ['CIGARETTE', 'PENCIL', 'TICKET', 'TOE']
4. __ Dream: ['AMERICAN', 'FEVER', 'LUCID', 'PIPE'])

Example 3:
Words: ['HANGAR', 'RUNWAY', 'TARMAC', 'TERMINAL', 'ACTION', 'CLAIM', 'COMPLAINT', 'LAWSUIT', 'BEANBAG', 'CLUB', 'RING', 'TORCH', 'FOXGLOVE', 'GUMSHOE', 'TURNCOAT', 'WINDSOCK']
Groupings:

1. Parts of an airport: ['HANGAR', 'RUNWAY', 'TARMAC', 'TERMINAL']
2. Legal terms: ['ACTION', 'CLAIM', 'COMPLAINT', 'LAWSUIT']
3. Things a juggler juggles: ['BEANBAG', 'CLUB', 'RING', 'TORCH']
4. Words ending in clothing: ['FOXGLOVE', 'GUMSHOE', 'TURNCOAT', 'WINDSOCK']

Categories share commonalities:

- There are 4 categories of 4 words each
- Every word will be in only 1 category
- One word will never be in two categories
- As the category number increases, the connections between the words and their category become more obscure. Category 1 is the most easy and intuitive and Category 4 is the hardest
- There may be a red herrings (words that seems to belong together but actually are in separate categories)
- Category 4 often contains compound words with a common prefix or suffix word
- A few other common categories include word and letter patterns, pop culture clues (such as music and movie titles) and fill-in-the-blank phrases

You will be given a new example (Example 4) with today's list of words. First explain your reason for each category and then give your final answer following the structure below (Replace Category 1, 2, 3, 4 with their names instead)

Groupings:
Category1: [word1, word2, word3, word4]
Category2: [word5, word6, word7, word8]
Category3: [word9, word10, word11, word12]
Category4: [word13, word14, word15, word16]

Remember that the same word cannot be repeated across multiple categories, and you need to output 4 categories with 4 distinct words each. Also do not make up words not in the list. This is the most important rule. Please obey

Example 4:
Words : [InsertGame]
Groupings

B Disagreements in Annotations

There was certain extent of disagreements between Semantic and Encyclopedic Knowledge. For instance ['PIKE', 'SPLIT', 'STRADDLE', 'TUCK'] are GYMNASTICS POSITIONS and requires domain specific knowledge so could be thought of as Encyclopedic knowledge but it could be classified under Semantic Knowledge (*Type Of* relation) as many of these words appear in Wordnet. Some disagreements also occurred between between Associative and Encyclopedic Knowledge. For instance here the shared property for ['BASE', 'BOND', 'ELEMENT', 'SOLUTION'] being CHEMISTRY TERMS requires using Associative Knowledge but this could still require Encyclopedic Knowledge about Chemistry. However since we consider Encyclopedic knowledge only as ones related to a Knowledge Base and entities instead of domain concepts we treat this as Associative Knowledge. Finally due to their colloquial use in the English language sometimes there can be confusion amongst what Semantic and Associative Knowledge. For instance ['BOMB', 'DUD', 'FLOP', 'LEMON'] can be thought of synonyms of FAILURE and hence fall under category of Semantic Knowledge , however Lemon is rarely used for Failure and requires using connotative knowledge (shared property) and hence falls more appropriately under Associative Knowledge.

Figure 6: Example of red herring category where the 3 words outlined in red might seem as though they belong together.

C Red Herrings

In the puzzle in Figure 6, a red herring category is present. In the highest performing models Claude 3 Opus and GPT-4o created a category called "Milk" with *Whole, Skim,* and *Soy* and included a random fourth word that did not fit. Each of these three words, however, belongs to a different category: *Whole* to *Kinds of Numbers*, *Skim* to *Touch Lightly*, and *Soy* to *Sauces in Chinese Cuisine*. In other puzzles including a red herring category like this one, all models make similar rationalizations.

Figure 7: Example of red herring word where the 5 words outlined in red may seem like they belong together.

The game in Figure 7 is an example of a game with a red herring word. The five words that appear as though they belong together are outlined in red. However, *Mistletoe, Reindeer, Snowman,* and *Stocking* form the *Christmas Related* category, while *Candy Cane* belongs to the category *Things with Stripes*. In this game, GPT-4o, Gemini 1.5 Pro, and Llama 3 70B all made the mistake of grouping *Candy Cane* with some combination of three of the other Christmas-related words.

D Performance

D.1 Humans

The frequency of clustering scores for novice human players in 100 games and scores for GPT-4o in the same games are shown in Table 4. The frequency of clustering scores for expert human players in 50 games and scores for GPT-4o in the same games are in Table 5.

Figures 8 and 9 show the distribution for the weighted clustering scores of GPT-4o against novice humans and expert humans, respectively. Finally, Figures 10 and 11 depict the performance of novice and expert humans, respectively, by reasoning type from our taxonomy of knowledge. Because the total counts of types of reasoning required across the 400 (novice) and 200 (expert) categories are unbalanced, the count of the categories reasoned correctly is shown above each bar.

Unweighted Clustering Score	GPT-4o	Novice Humans
0	31	30
1	35	39
2	29	13
3	2	0
4	8	18

Table 4: Frequency of clustering scores 0-4 for GPT-4o and novice human players across 100 Connections games

Unweighted Clustering Score	GPT-4o	Expert Humans
0	16	2
1	18	7
2	10	9
3	1	0
4	5	32

Table 5: Frequency of clustering scores 0-4 for GPT-4o and expert human players across 100 Connections games

D.2 LLMs

Table 6 shows the frequency of the unweighted clustering scores (number of categories correctly grouped) for each LLM. The total number of games played by each model is 200. Table 7 is slightly different and shows the frequency of categorical reasoning scores (the categories correctly grouped and reasoned) for each model. Because a caveat for receiving a categorical reasoning score greater than 0 is matching gold categories, a score of 0 is more common than in the unweighted clustering scores.

Figure 12 shows the performance of each LLM by reasoning type from our taxonomy of knowledge. Because the total counts of types of reasoning required across the 800 categories are unbalanced, the count of the categories reasoned correctly is shown above each bar.

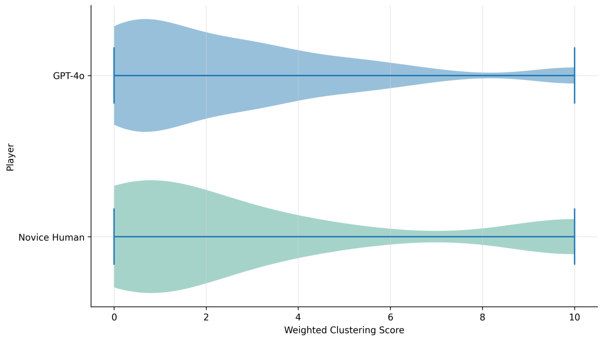


Figure 8: Spread of weighted clustering score for GPT-4o and novice human players across 100 Connections games

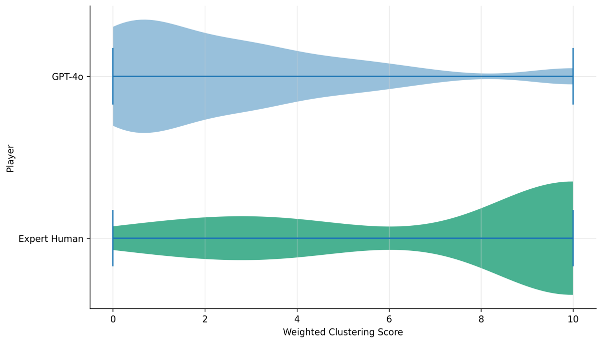


Figure 9: Spread of weighted clustering score for GPT-4o and expert human players across 50 Connections games

E Human Evaluation Interface

Figure 13 shows the two main screens of the evaluation interface provided to both novice and expert human evaluators. (a) is the instruction screen, while (b) is an example of a game screen after the user hits the "Play" button. To solve the game in one shot, all 16 words from a game are displayed on the screen in separate boxes, with one drop-down per box. The drop-down consists of four labels:

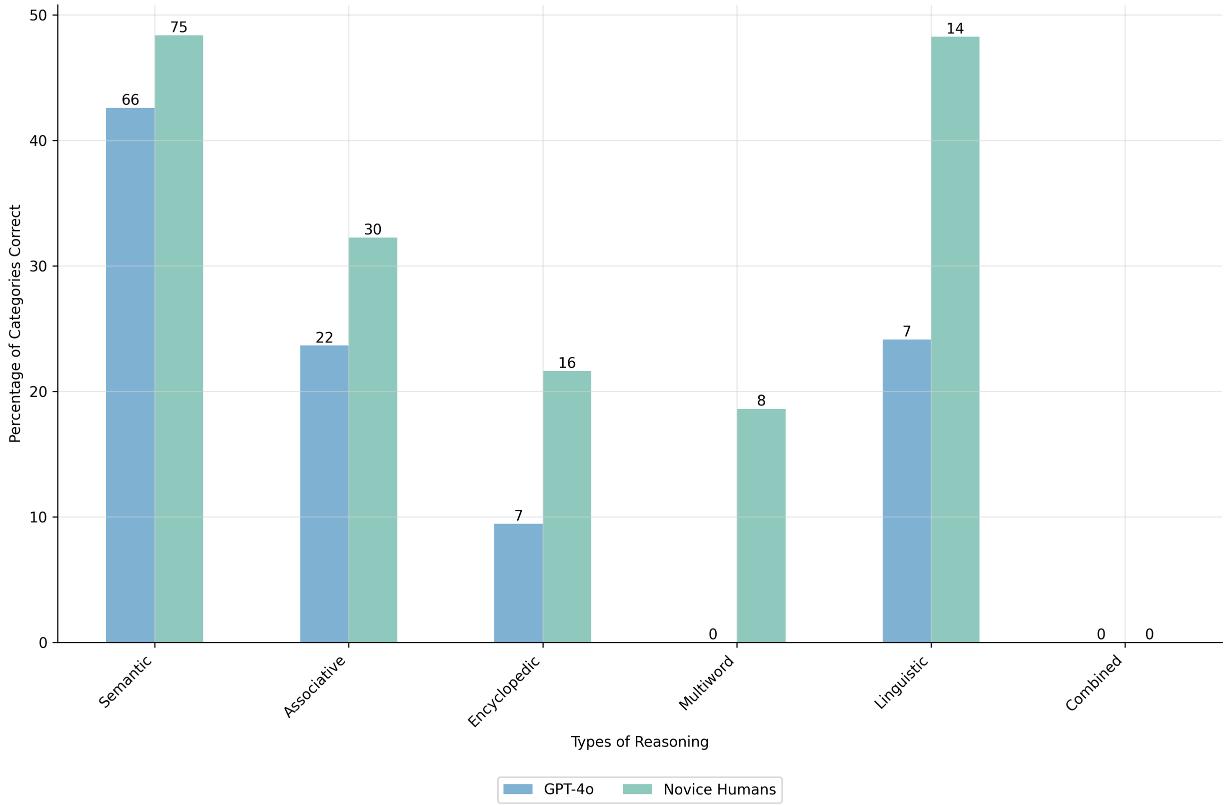


Figure 10: Percentage of categories from each knowledge type correctly classified and reasoned by GPT-4o and novice human players across 100 games. The counts of categories correctly reasoned are displayed above each bar.

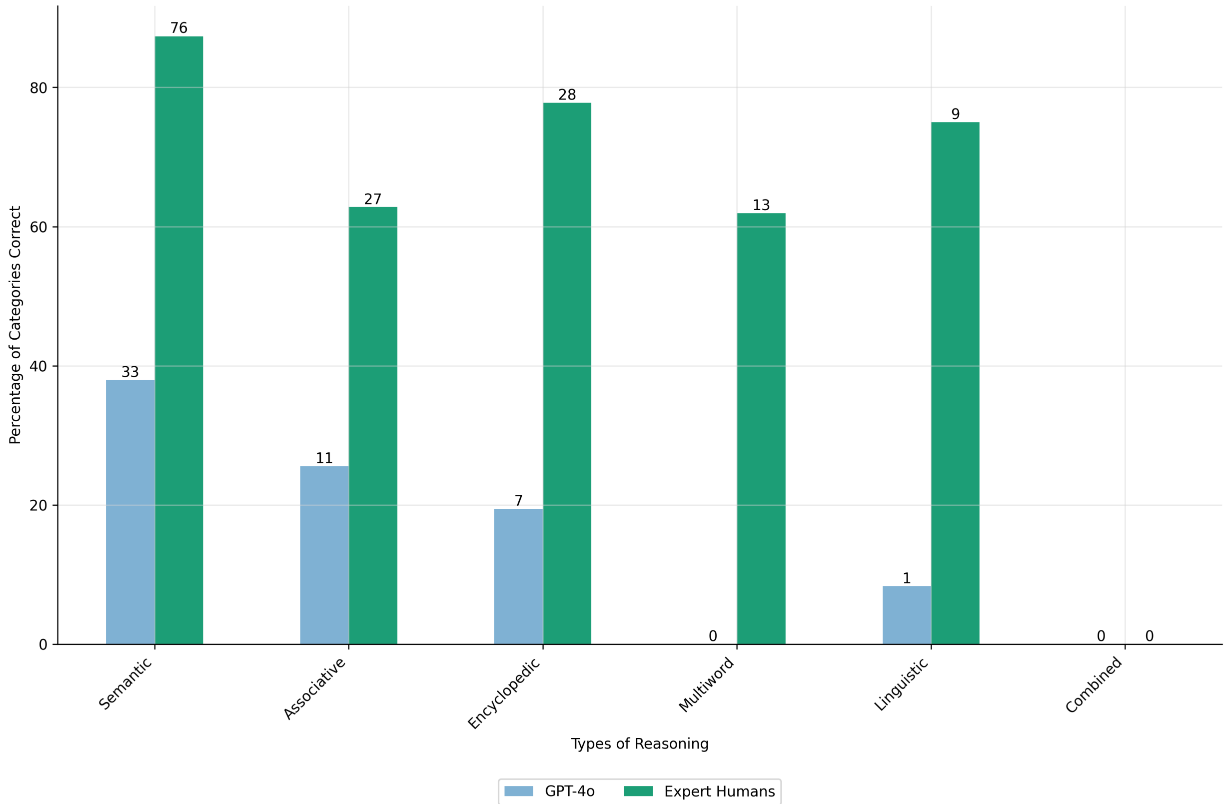


Figure 11: Percentage of categories from each knowledge type correctly classified and reasoned by GPT-4o and expert human players across 50 games. The counts of categories correctly reasoned are displayed above each bar.

Unweighted Clustering Score	Gemini 1.5 Pro	Claude 3 Opus	GPT-4o	Llama 3 70B
0	100	68	53	84
1	64	72	77	73
2	26	37	48	32
3	1	10	6	5
4	9	13	16	6

Table 6: Frequency of unweighted clustering scores 0-4 for 4 LLMs across 200 Connections games

Categorical Reasoning Score	Gemini 1.5 Pro	Claude 3 Opus	GPT-4o	Llama 3 70B
0	116	80	59	98
1	59	66	87	73
2	18	36	41	23
3	7	13	11	4
4	0	5	2	2

Table 7: Frequency of categorical reasoning scores 0-4 for 4 LLMs across 200 Connections games

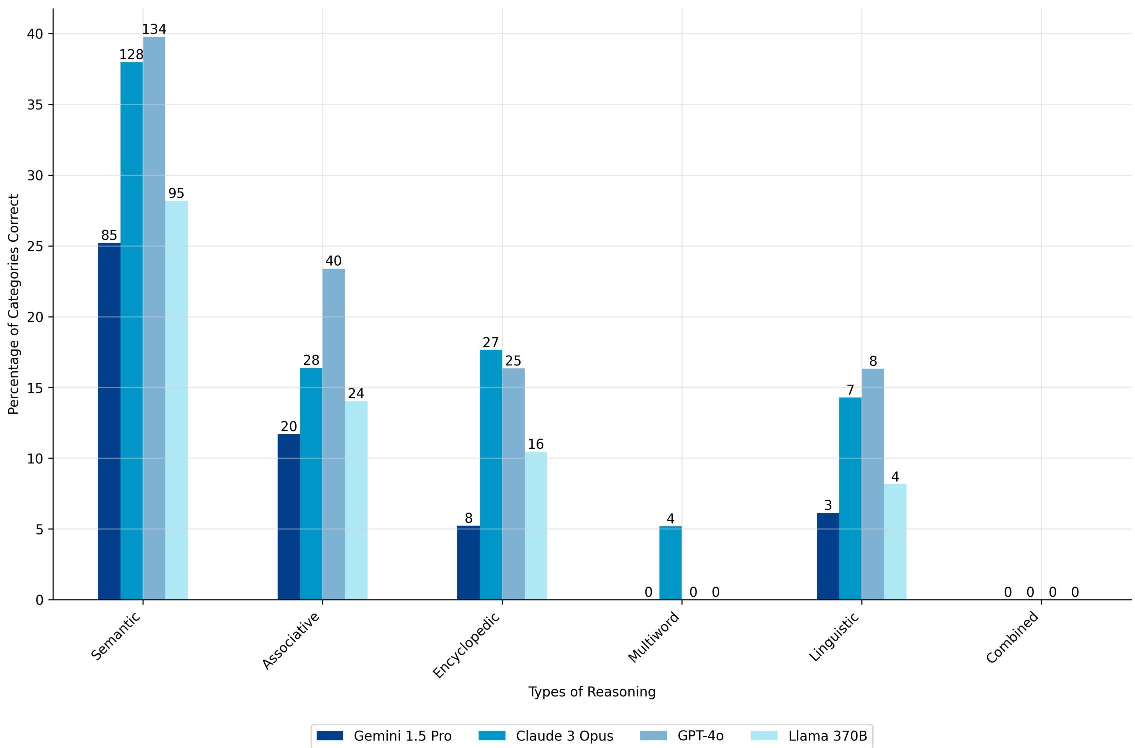
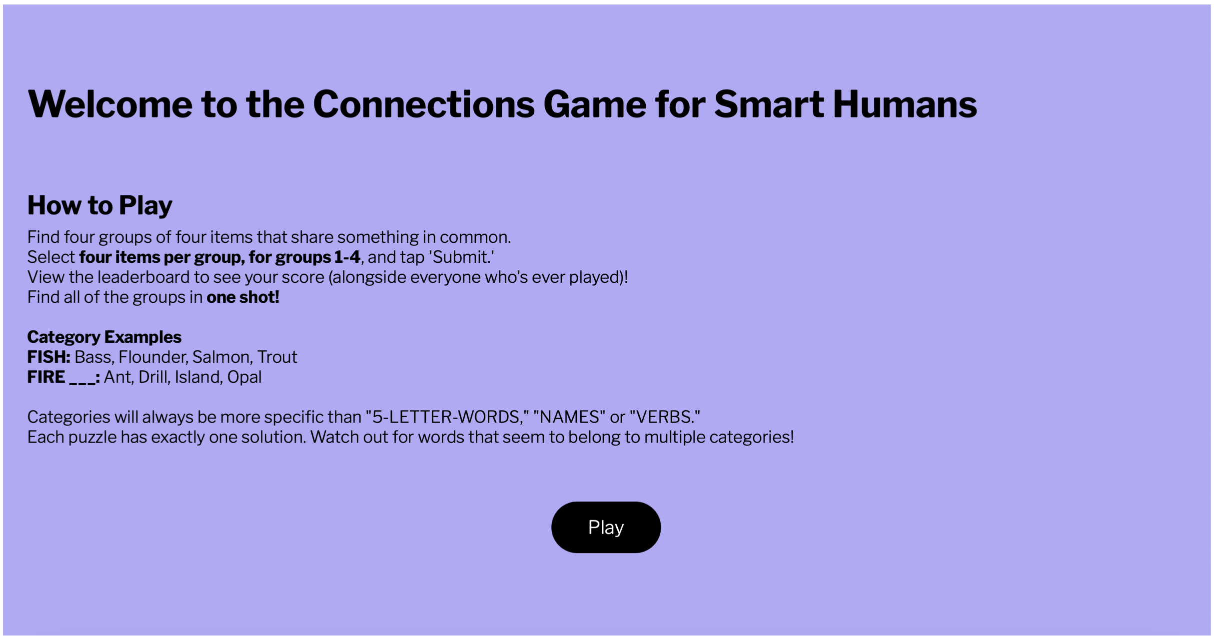


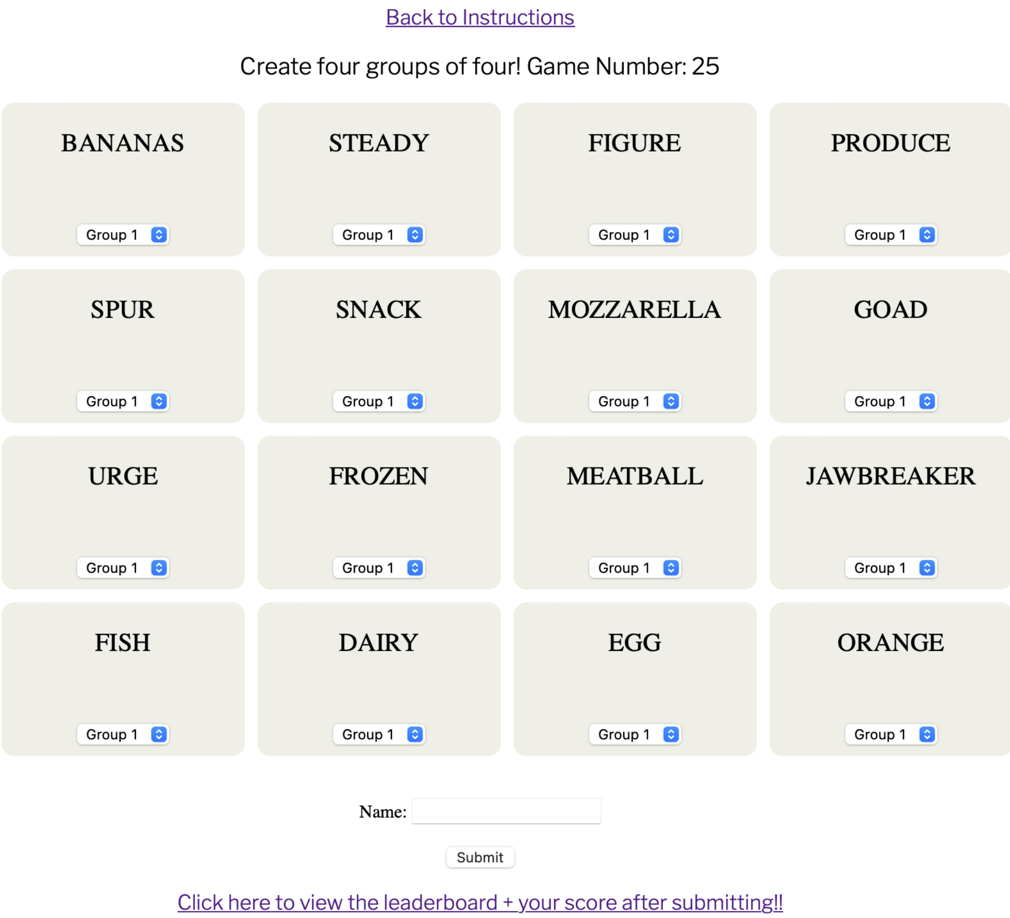
Figure 12: Percentage of categories from each knowledge type correctly classified and reasoned by the models across 200 games. The counts of categories correctly reasoned are displayed above each bar.

Group 1, Group 2, Group 3, and Group 4. The user's job is to create 4 groups of 4 words using the given labels. Because the groups are chosen from a drop-down menu where the default option is Group 1, a clustering score of 3 is impossible.

We stored the data collected in a SQLite database. Other than any name of choice users were prompted to enter in the "Name" text entry box, no personal data was collected. Each evaluator was then assigned initials in the final dataset collected for evaluation. These initials are not included in the data release. The data that would be collected and its purpose were verbally conveyed to each evaluator before asking for their consent.



(a) Instruction screen



(b) Example of game play

Figure 13: Human evaluation interface