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

Prisha Samadarshi[1]   Mariam Mustafa[1]   Anushka Kulkarni[1]   Raven Rothkopf[1]
Tuhin Chakrabarty[2]*†   Smaranda Muresan[1]*

[1] Department of Computer Science; Barnard College, Columbia University
[2] Department of Computer Science, Stony Brook University

{ps3203, mm5970, ajk2256, rgr2124}@alum.barnard.edu, tuhin.chakrabarty@stonybrook.edu, smuresan@barnard.edu

Abstract

The *New York Times* Connections game has emerged as a popular and challenging pursuit for word puzzle enthusiasts. We collect 438 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, Claude 3.5 Sonnet, which has otherwise shown impressive reasoning abilities on a wide variety of benchmarks, can only fully solve 18% of the games. Novice and expert players perform better than Claude 3.5 Sonnet, with expert human players significantly outperforming it. We create a taxonomy of the knowledge types required to successfully cluster and categorize words in the Connections game. We find that while LLMs perform relatively well on categorizing words based on semantic relations they struggle with other types of knowledge such as Encyclopedic Knowledge, Multiword Expressions or knowledge that combines both Word Form and Meaning. Our results establish the *New York Times* Connections game as a challenging benchmark for evaluating abstract reasoning capabilities in AI systems.

1 Introduction

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; Ji et al., 2022). 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). We propose the *NYT* Connections Game as a test bed for investigating the abstract reasoning capabilities of both humans and large language models (LLMs).

Create four groups of four!

MOBILE	FOLLOWERS	SHOVELS	BUFFALO
LIKES	INSULTS	SHARES	SHEEP
APARTMENT	BILLINGS	PUPPETS	OPTIONS
EQUITY	PHOENIX	STOCKS	LEMMINGS

(a) The unsolved connections game presented to a player

CONFORMISTS
FOLLOWERS, LEMMINGS, PUPPETS, SHEEP

COMPANY OWNERSHIP OFFERS
EQUITY, OPTIONS, SHARES, STOCKS

U.S. CITIES
BILLINGS, BUFFALO, MOBILE, PHOENIX

WHAT "DIGS" MIGHT MEAN
APARTMENT, INSULTS, LIKES, SHOVELS

(b) The solved connections game with correct categories shown in ascending level of difficulty—straightforward (yellow) to tricky (purple)

Figure 1: Example from a *NYT* Connections game

Connections is 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 with identifying four distinct clusters that link the corresponding four words in each cluster through some shared characteristics (Figures 1 [a] and [b]). Categories 1 (yellow), 2 (green), 3 (blue), and 4 (purple) are arranged in ascending level 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

*Equal contribution. †denotes work done at Columbia

(in this case, "digs") can have multiple meanings that differ in etymology or sense, depending on the context.

While the task might seem easy, many words can be grouped easily into multiple categories, acting as red herrings. For instance, from the game in Figure 1, *Likes, Followers, Shares, Insult* might be categorized as "Social Media Interactions" at first glance. Unlike common categories (e.g., "Fruit," "Furniture"), the game is designed to promote ad hoc category formations that violate the correlational structure of the environment and are not well established in memory (Barsalou, 1983). 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).

We test the capabilities of five state-of-the-art large language models, namely Google's Gemini 1.5 Pro (Team et al., 2023), Anthropic's Claude 3.5 Sonnet (Anthropic, 2024), OpenAI's GPT-4Omni (OpenAI, 2023), Meta's Llama 3.1 405B, (AI@Meta, 2024) and Mistral Large 2 (Mistral-AI, 2024) on 438 distinct *NYT* Connections games and compare them with human performance on a subset of these games. Our experimental results show that while all LLMs can partially solve some games, their performance is far from ideal. Even the best-performing model Claude 3.5 Sonnet (with few-shot and chain-of-thought prompting), can only solve 18% of the games completely. In addition, we recruit human players at novice and expert levels of proficiency and compare their performance to Claude 3.5 Sonnet. Our results show that the *NYT* Connections game serves as a challenging benchmark for reasoning, with novice players performing marginally better than Claude 3.5 Sonnet and expert players performing significantly better than Claude 3.5 Sonnet in solving games perfectly (Section 5).

In addition, to better understand the LLMs abstract reasoning capabilities or the lack thereof, we propose a taxonomy of knowledge required to group words into their respective categories (Section 3.2). Our analysis at Section 6.1 shows that while LLMs are relatively better at reasoning that involve Semantic Relations, they struggle with other types of knowledge such as Multiword Expressions and combined knowledge about Word Form and Word Meaning (Section).

Our code and data will be made available to the public at `https://github.com/mustafamariam/LLM-Connections-Solver`.

2 Related Work

The growing popularity of Large Language Models has led to an exciting array of research using natural language processing techniques for text-based games. Recent work has studied whether these models can act as players in agentic environments (Huang et al., 2024; Wang et al., 2023; Wu et al., 2023; Noever et al., 2020) or conversational settings (Qiao et al., 2023). In addition to acting as players, researchers have also tested the abilities of transformer-based language models in generating games (Ammanabrolu and Riedl, 2021; Todd et al., 2023; Sudhakaran et al., 2024; Hu et al., 2024; Chen et al., 2023; Merino et al., 2024)

Recent research has explored applying large language models (LLMs) and other natural language processing (NLP) techniques to solve and generate text-based puzzles. 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. Zhao and Anderson (2023) release PUZZLEQA, a multiple-choice dataset comprising 15 years of on-air Sunday Puzzle word-games and show that while ChatGPT can solve these questions with an accuracy of around 50%, they still struggle with generating novel and engaging puzzles. Unlike the *NYT* Connections game, PUZZLEQA relies on character-level word transformations compared to encyclopedic, associative, or semantic knowledge. Rozner et al. (2021) examined the potential of using "cryptic crossword" clues as an NLP benchmark.

Most relevant to our paper is the contemporaneous work by Todd et al. (2024) who test the performance of various LLMs (from BERT and RoBERTA to GPT-4 and GPT-4 with Chain-of-Thought prompting) in solving the *NYT* Connections game. Our work is similar in that it utilizes the *NYT* Connections puzzles as a means to investigate the abstract reasoning capabilities of state-of-the-art LLMs. However, our contribution focuses not only on the ability of LLMs to solve the game, but also on studying the types of knowledge required to so. Moreover, we evaluate both state–of-the-art models (open and closed weights) and humans (novices and experts). We also benchmark on a

larger number of *NYT* Connections game. Additionally, Todd et al. (2024)'s experimental setup mirrors the way the original *NYT* Connections game is played (i.e., one category at a time, with an allotment of 4 incorrect guesses), while we require both humans and LLMs to provide all categories at once in only one attempt.

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 *NYT* 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, demonstrating a significant gap in the abstract reasoning capabilities of LLMs. Prior work has also studied abstract reasoning in Neural Networks (Barrett et al., 2018) in the presence of distracting features (Zheng et al., 2019). Our work builds upon these and presents the *NYT* Connections game as a compelling benchmark for abstract reasoning capabilities of 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 *NYT* Connections puzzles, we resorted to an external, third-party site for data collection.[1] Our data spans daily problems from the conception of *NYT* Connections in June 2023 to August 2024. In total, we gather 441 distinct games, out of which 3 are used for few-shot prompting, while the remaining 438 comprise the dedicated test set.

---

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

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 the *NYT* Connections game, players must draw on certain aspects of word knowledge, such as a word's meaning or form and sometimes both simultaneously. To deepen our understanding, we bucket each <category, grouping> into the types of knowledge that are primarily required to solve them. Three linguists annotate a total of 1,752 samples coming from 438 games into 3 broader categories which give rise to 8 subcategories (Figure 2). We take majority voting to arrive at a unique label for each <category, grouping>. We restrict annotations to sub-categories and not sub-sub-categories (e.g., Types of Semantic Relations). We obtain a Fleiss Kappa of 0.78 showing substantial agreement.

3.2.1 Word Form

Word Form refers to the specific shape or appearance a word takes in a given context. It encompasses various aspects of a word's structure and representation and is broadly used in the *NYT* Connections game, testing knowledge on *Phonology, Orthography, Morphology* and *Multiword Expressions*. Phonology in Word Form deals with the sound structure of words, including pronunciation, stress patterns, and phonetic variations. Orthography relates to the conventional spelling system of a language, which may not always directly correspond to pronunciation. Morphology examines the internal structure of words, including roots and affixes and how they combine to create meaning. For example, as shown in Figure 2, one needs morphological knowledge to group *Dom, Ion, Ness,* and *Ship* as "Noun Suffixes." Similarly, one needs to rely on phonological knowledge of the sound patterns of *Answer, Two, Wrist,* and *Wrong* to categorize them as "Silent 'W'." Multiword Expressions (MWE) have a fixed or semi-fixed form and are typically non-compositional (i.e., their meaning cannot be predicted from their individual components) (Moon, 1998).

3.2.2 Word Meaning

Semantic Relations: The majority of instances in the *NYT* Connections game require possessing knowledge of *semantic relations* (Murphy, 2003; Cruse, 1986), such as synonymy (words with the same meaning), hypernymy/hyponymy (relation
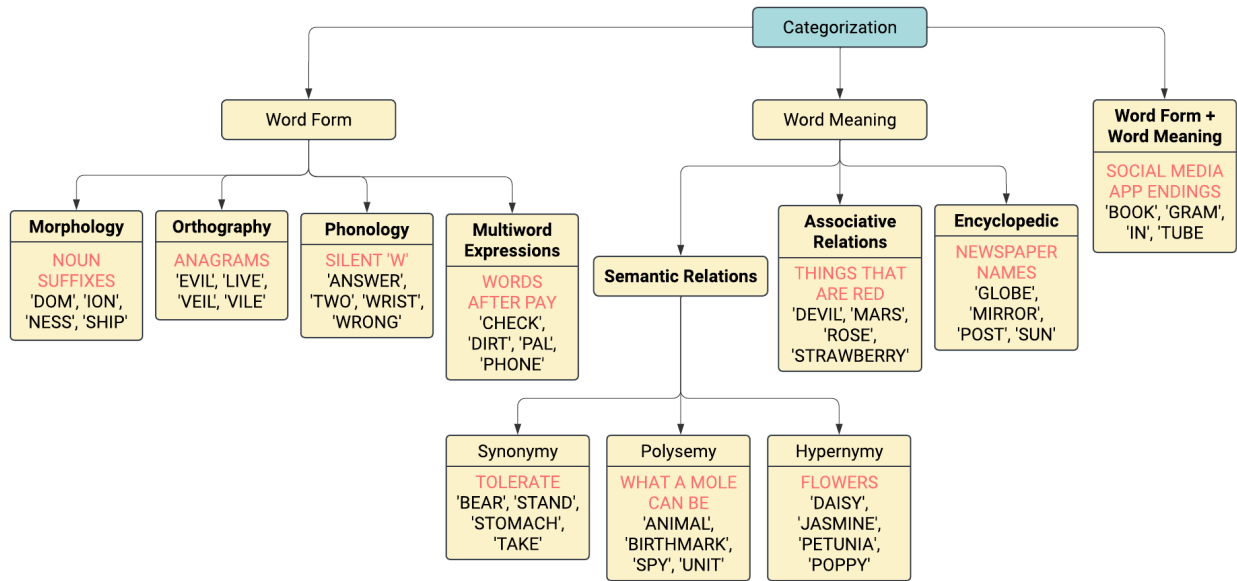
Figure 2: Proposed taxonomy of knowledge types required to solve the Connection games. In our evaluation we used the categories in bold

between a generic term and its specific instance), and homonymy and polysemy (many possible meanings of a word). Figure 2 shows three examples of groups that use such semantic relations.

Associative Relations Prior work has studied models of automatic priming for word identification that are typically divided into groups based on associative relations (e.g., spreading activation) and others based on semantic similarity (e.g., distributed models) (Thompson-Schill et al., 1998). In contrast to category members that share semantic features and category nodes, Associative Relations are elements of specific situations or thematic contexts, with little or no overlap between their semantic features (e.g. "Things that are red": *Mars* and *Strawberry*; Figure 2) (Rose and Rahman, 2016; Shanks, 1995; Barsalou, 1983).

Encyclopedic 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 Figure 2, where, to bucket the words *Globe, Mirror, Post,* and *Sun* into the category of "Newspaper Names," one needs to possess knowledge that Globe refers to the *Boston Globe*, Mirror to the *Daily Mirror*, a UK tabloid, Post to the *Washington Post* and Sun to *The Sun*, another UK tabloid. We label this type of knowledge

Encyclopedic.

3.2.3 Word Form + Word Meaning

Some of the hardest examples in the *NYT* Connections game require reasoning of both word form and meaning. For instance, the example in Figure 2 shows that to group the words *Book, Gram, In,* and *Tube*, 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 the use of knowledge on Word Meaning (Encyclopedic) as well as Word Form (Morphology).

4 Experimental Settings

4.1 LLMs as Game Players

To test the capabilities of large language models in solving the *NYT* 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 *NYT* Connections game instructions provided by the *New York Times* (Liu, 2023a), and included

Word Form		Word Meaning			Word Meaning + Word Form
Phonology/Orthography/ Morphology	Multiword Expressions	Semantic Relations	Associative Relations	Encyclopedic	92
44	168	1045	137	266	

Table 1: Distribution of different knowledge types required to categorize words across 438 games

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 on a development set of 30 games, using 5 LLMs. After identifying commonalities in the types of errors made by the LLMs while playing the game, we added additional instructions, specified the response format, and included some tips from a *NYT* article about playing the *NYT* Connections game (Aronow and Levine, 2023). The entire prompt is in Appendix A. To ensure consistency and fairness in performance, we prompt 5 LLMs — Gemini 1.5 Pro, Claude 3.5 Sonnet, GPT-4o, and Llama 3.1 405B and Mistral Large 2 — with the same input. We use the default sampling parameters (temperature and top_p) and the scoring schema outlined in Section 4.3 to evaluate how all models perform in solving 438 *NYT* Connections games spanning from June 2023 to August 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 *NYT* Connections and 5 expert or regular *NYT* Connections players. The novice and the expert evaluators were peers of the first four authors, who volunteered to participate without any payment. We designed a human evaluation interface and randomly sampled 100 games from our test set. Appendix E has more information about the interface. The first screen displays an abridged 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 solve the game (i.e., make all 4 categorizations at once). This aligns with (Todd et al., 2024)'s *challenge mode*.

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 438 in the test set, and expert participants each played 10 games for a total of 50 randomly sampled games.

4.3 Evaluation Criteria

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

4.3.1 Clustering Score

The clustering score evaluates the ability to correctly group 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 player match 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.
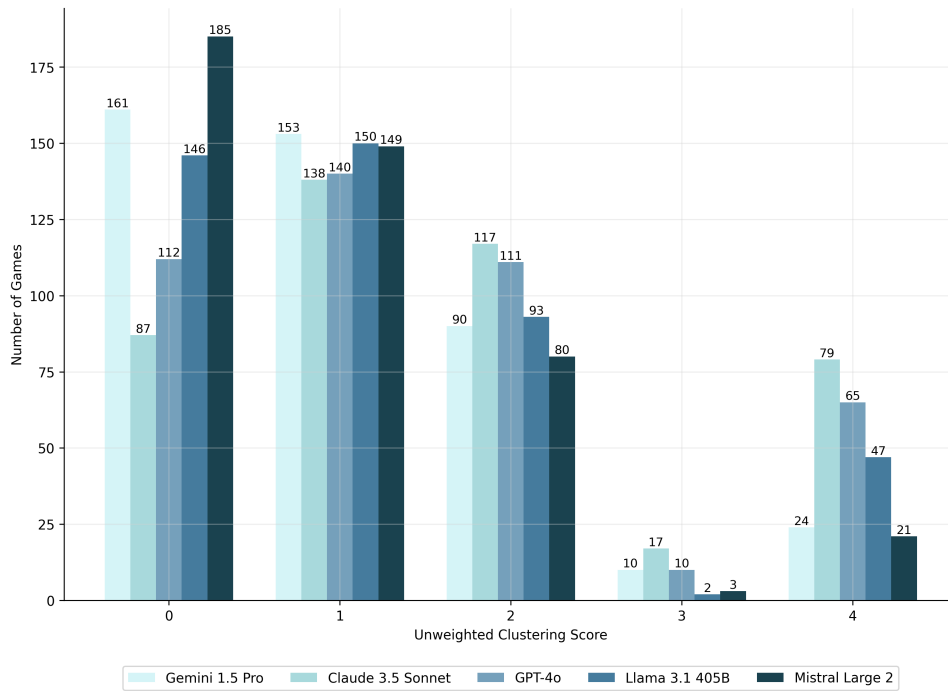
Figure 3: Frequency of unweighted clustering scores for 5 LLMs across 438 games. The number of games in which the respective unweighted clustering score was achieved is atop each bar. 0 means no correct clusters while 4 means all correct clusters

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 of knowledge for *NYT* 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, we can evaluate whether the 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.

5 Results

5.1 LLM performance

Overall, we find that Claude 3.5 Sonnet performs best across thw 438 games. Figure 3 shows the unweighted clustering scores for all 5 LLMs. Claude has the lowest percentage of games in which it

made no correct clustering, at about 20%, and most games solved perfectly at 18%. Mistral Large 2 performs the worst overall. It could not make any correct clusters for 42% of the games, and it perfectly solved the least amount of games. In terms of
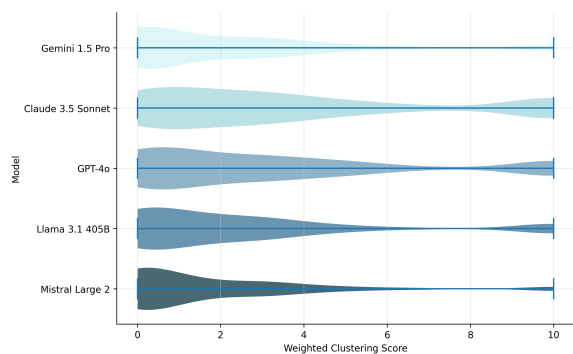

Figure 4: Spread of weighted clustering scores for each model across 438 *NYT* Connections games

weighted clustering scores for each model (Figure 4), Gemini 1.5 Pro and Mistral Large 2 show similar results. Most of their scores are concentrated before 2, showing that these models had a higher ability to correctly classify the easiest or second easiest categories. GPT-4o and Claude 3.5 Sonnet had most of their weighted clustering scores concentrated before 5, 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.

5.2 Human Performance

In human performance, we measure both novice and expert players against the best overall performing Claude 3.5 Sonnet. For the 100 games played by novices and 50 games played by experts, we compare the same games played by Claude 3.5 Sonnet.
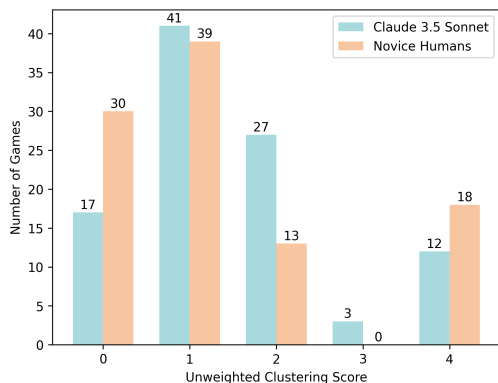
5.2.1 Novice Players

Figure 5: Frequency of clustering scores of Claude 3.5 Sonnet and 12 novice humans across 100 games

In the 100 games that the novice players completed, their average unweighted clustering score was 1.37, slightly worse than Claude's average of 1.52 in the same 100 games. Claude and novice humans also had similar weighted clustering score distributions. 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 fourth is also correct). Because of Claude's imperfect instruction-following abilities (repeating or omitting words), it was still able to obtain a clustering score of 3, as shown in Figure 5.

5.2.2 Expert Players

Expert human players performed significantly better than novices and Claude 3.5 Sonnet, with an average clustering score of 3.06 compared to Claude's 1.78 (on the same 50 games). The distribution of weighted clustering scores is also far more right-skewed (see Appendix D.1 for more). Figure 6 shows that experts perfectly solved over 60% of the 50 games, while Claude 3.5 Sonnet fully solved 20% of these.
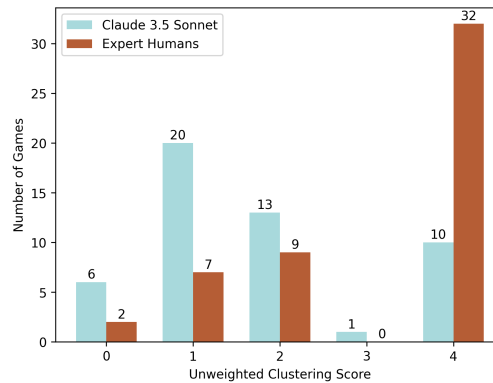

Figure 6: Frequency of clustering scores of Claude 3.5 Sonnet and 5 expert humans across 50 games

6 Discussion

6.1 What type of reasoning is hardest for LLMs?

To answer this question we rely on our taxonomy of reasoning types introduced in Section 3.2. The breakdowns of the reasoning types for the 1752 categories in our 438-game dataset are shown in Table 1. Figure 7 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 1,752 categories are unbalanced, the count of categories reasoned correctly is shown above each bar. The categories were only counted as correct if the model both clustered all the words correctly and justified its reasoning in a manner semantically analogous to the gold *NYT* Connections-provided category name.

The patterns in performance across the types of reasoning parallel the LLMs' overall performances for the most part, with Llama 3.1 405B's performance in Multiword Expressions and Word Meaning + Word Form (slightly better than GPT-4o) and Mistral Large 2's performance in Word Meaning + Word Form (slightly better than Gemini 1.5 Pro) defying this pattern. The performance in reasoning categories across models is ranked from best to worst as follows: *Semantic > Encyclopedic > Associative > Morphology/Orthography/Phonology > Multiword > Word Meaning + Word Form*. This aligns with the game makers' perceived level of category difficulty, as Word Meaning categories appear most often as yellow or green (easier) groupings, while Multiword Expressions and Word Meaning + Word Form categories are usually the purple groupings (most difficult). This is also consistent with the fact that the two types of reasoning
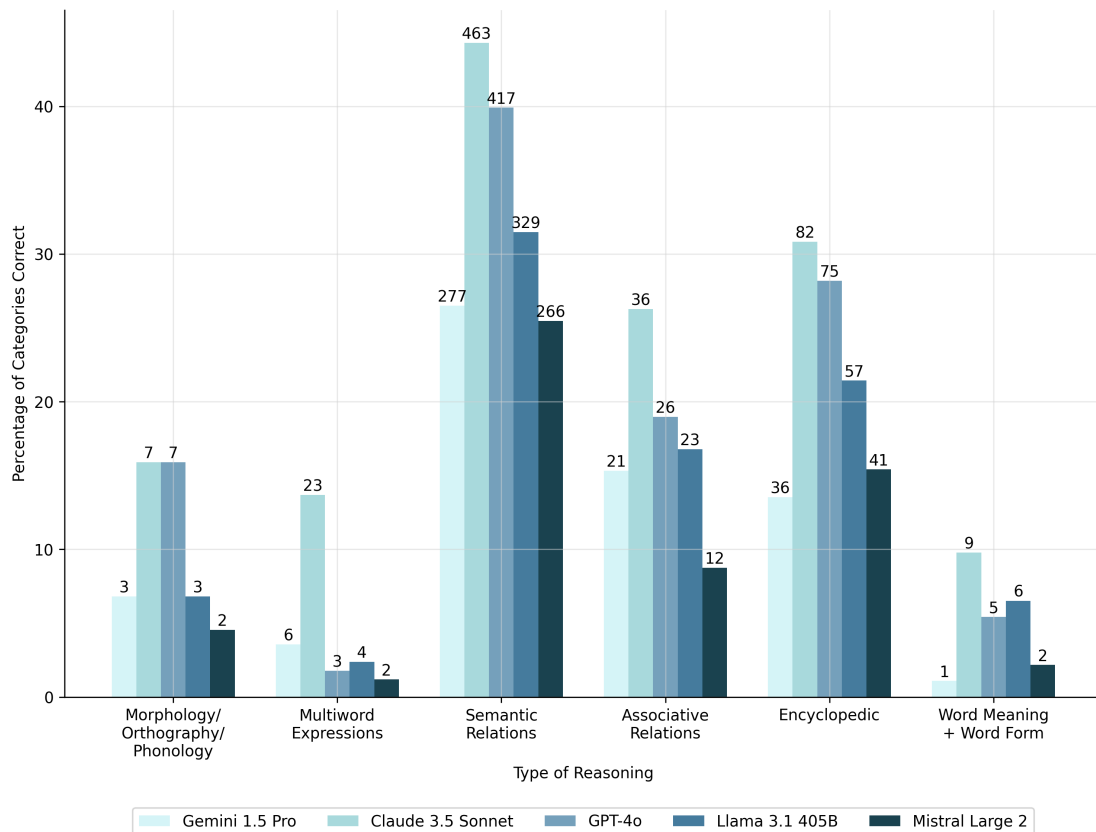
Figure 7: Percentage of categories from each knowledge type correctly classified and reasoned by the models across 438 games. The counts of categories correctly reasoned are displayed above each bar.

LLMs performed best in—Semantic Relations and Encyclopedic Knowledge—exist in web-scraped information included in pre-training data, while other types of knowledge are more obscure and require iterative and deductive reasoning within each game. In every type of reasoning, however, LLMs perform with less than 50% accuracy. Though they may be clustering words together accurately, they are not always justifying these clusters correctly, and this is visible in the differences between the frequencies of unweighted clustering scores of 4 and categorical reasoning scores of 4 (Appendix D.2 Tables 5 and 6).

Since humans were not asked to provide justifications for their clusters, we do not include reasoning comparisons between LLMs and humans.

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

The *NYT* Connections game is often formulated with item overlap in mind, according to the *NYT* 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 case, 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, $\frac{\text{categorical reasoning}}{\text{unweighted clustering}}$ tells us how common it is for LLMs to cluster categories correctly by chance. The average ratios in Table 2 are

Model	Average Ratio
Gemini 1.5 Pro	0.76
Claude 3.5 Sonnet	0.87
GPT-4o	0.84
Llama 3.1 405B	0.81
Mistral Large 2	0.82

Table 2: Average categorical reasoning to unweighted clustering score ratio by model

fairly high, close to or above 80%. The highest overall performing models Claude 3.5 Sonnet and GPT-4o have the highest ratios as well. Though it is fairly uncommon that a model will correctly group words without correctly naming the reasoning behind that grouping, 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?

Instead of choosing the first grouping, strategies grounded in System2 Thinking (Evans, 2003) could improve performance on such a benchmark. Generating multiple chains-of-thought reasoning and learning a model that assigns a higher reward to the correct reasoning chain prevents the model from arriving greedily to a suboptimal categorization. Allowing LLMs to solve the game one category at a time and incorporating the feedback present to humans in the original *NYT* Connections game, such as 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 the *NYT* Connections 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 Claude 3.5 Sonnet 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. Although most LLMs possess Word Meaning reasoning capabilities, they struggle with Multiword Expressions and combined knowledge categories. In addition, red herrings pose a challenge to current LLMs. Overall, we find that solving the *NYT* Connections requires a breadth of different knowledge types, which current LLMs do not seem to fully master.

8 Limitations

Many of the limitations in this section stem from the lack of data available for*NYT* 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 *NYT* Connections games might bridge the gap between LLM and expert human evaluators' performance.

We acknowledge that human evaluators were not required to add justifications for the groupings they made. This might 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 have 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 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*.

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.

Lawrence W. Barsalou. 1983. Ad hoc categories. *Memory & Cognition*, 11:211–227.

Dake Chen, Hanbin Wang, Yunhao Huo, Yuzhao Li, and Haoyang Zhang. 2023. Gamegpt: Multi-agent collaborative framework for game development. *arXiv preprint arXiv:2310.08067*.

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

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.

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10):454–459.

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

Chengpeng Hu, Yunlong Zhao, and Jialin Liu. 2024. Generating games via llms: An investigation with video game description language. *arXiv preprint arXiv:2404.08706*.

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. 2024. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*.

Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. 2022. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 582–601.

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*.

Tim Merino, Sam Earle, Ryan Sudhakaran, Shyam Sudhakaran, and Julian Togelius. 2024. Making new connections: Llms as puzzle generators for the new york times' connections word game. *arXiv preprint arXiv:2407.11240*.

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.

Mistral-AI. 2024. Mistral large 2 : Large enough. https://mistral.ai/news/mistral-large-2407/.

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

MLynne Murphy. 2003. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge 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/.

Dan Qiao, Chenfei Wu, Yaobo Liang, Juntao Li, and Nan Duan. 2023. Gameeval: Evaluating llms on conversational games. *arXiv preprint arXiv:2308.10032*.

Sebastian Benjamin Rose and Rasha Abdel Rahman. 2016. Cumulative semantic interference for associative relations in language production. *Cognition*, 152:20–31.

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.

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

21230

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 Peng, 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.

Sharon L Thompson-Schill, Kenneth J Kurtz, and John DE Gabrieli. 1998. Effects of semantic and associative relatedness on automatic priming. *Journal of memory and language*, 38(4):440–458.

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.

Graham Todd, Tim Merino, Sam Earle, and Julian Togelius. 2024. Missed connections: Lateral thinking puzzles for large language models. *arXiv preprint arXiv:2404.11730*.

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. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

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.

Yue Wu, Xuan Tang, Tom M Mitchell, and Yuanzhi Li. 2023. Smartplay: A benchmark for llms as intelligent agents. *arXiv preprint arXiv:2310.01557*.

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: [1COLLECTIVE', '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 some disagreement between Semantic and Encyclopedic Knowledge labeling. One example is the grouping *pike, split, straddle, tuck* which are "Gymnastics Positions" and requires domain-specific knowledge, meaning it could be thought of

as Encyclopedic. However, it could also be classified under Semantic Relations (*Type Of* relation) as many of these words appear in WordNet. There can be confusion about what is considered to be Semantic or Associative Knowledge. For example, *card, chocolate, heart, rose* might be thought of to be semantically related as gifts given to a loved one and hence labeled under semantic relations; however, they can be viewed as elements of specific situations or thematic context such as "Seen on Valentine's Day". Hence it falls more appropriately under Associative Relations.

C Red Herrings

Figure 8: Example of red herring category where the 3 words outlined in red might seem as though they belong together.

In the puzzle in Figure 8, a red herring category is present. Gemini 1.5 Pro 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, most models make similar rationalizations.

The game in Figure 9 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, all models except Claude 3.5 Sonnet made the mistake of grouping *Candy Cane* with some combination of three of the other Christmas-related words.

Figure 9: Example of red herring word where the 5 words outlined in red may seem like they belong together.

D Performance

D.1 Humans

The frequency of clustering scores for novice human players in 100 games and scores for Claude 3.5 Sonnet in the same games are shown in Table 3. The frequency of clustering scores for expert human players in 50 games and scores for Claude 3.5 Sonnet in the same games are in Table 4.

Figures 10 and 11 show the distribution of the weighted clustering scores for Claude 3.5 Sonnet against novice humans and expert humans, respectively.

Unweighted Clustering Score	Claude 3.5 Sonnet	Novice Humans
0	17	30
1	41	39
2	27	13
3	3	0
4	12	18

Table 3: Frequency of clustering scores 0-4 for GPT-4o and novice human players across 100 *NYT* Connections games

D.2 LLMs

Table 5 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 438. Table 6 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 category words and names, a

Unweighted Clustering Score	Claude 3.5 Sonnet	Expert Humans
0	6	2
1	20	7
2	13	9
3	1	0
4	10	32

Table 4: Frequency of clustering scores 0-4 for GPT-4o and expert human players across 50 *NYT* Connections games
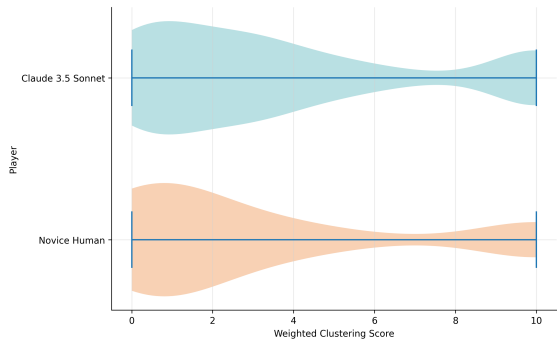

Figure 10: Spread of weighted clustering score for Claude 3.5 Sonnet and novice human players across 100 *NYT* Connections games
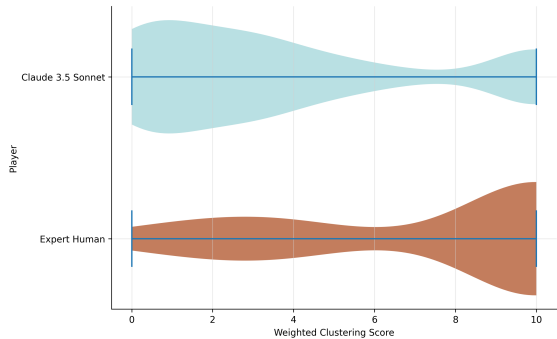

Figure 11: Spread of weighted clustering score for Claude 3.5 Sonnet and expert human players across 50 *NYT* Connections games

score of 0 is more common than in the unweighted clustering scores.

E Human Evaluation Interface

Figure 12 shows the two main screens of the evaluation interface provided to both novice and expert human players. (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: 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.

Unweighted Clustering Score	Gemini 1.5 Pro	Claude 3.5 Sonnet	GPT-4o	Llama 3.1 405B	Mistral Large 2
0	161	87	112	146	185
1	153	138	140	150	149
2	90	117	111	93	80
3	10	17	10	2	3
4	24	79	65	47	21

Table 5: Frequency of unweighted clustering scores 0-4 for 5 LLMs across 438 *NYT* Connections games

Categorical Reasoning Score	Gemini 1.5 Pro	Claude 3.5 Sonnet	GPT-4o	Llama 3.1 405B	Mistral Large 2
0	203	101	131	176	215
1	154	152	157	151	144
2	59	109	94	73	61
3	16	53	36	25	13
4	6	23	20	13	5

Table 6: Frequency of categorical reasoning scores 0-4 for 5 LLMs across 438 *NYT* Connections games

(a) Instruction screen

Create four groups of four! Game Number: 25

BANANAS	STEADY	FIGURE	PRODUCE
Group 1	Group 1	Group 1	Group 1
SPUR	SNACK	MOZZARELLA	GOAD
Group 1	Group 1	Group 1	Group 1
URGE	FROZEN	MEATBALL	JAWBREAKER
Group 1	Group 1	Group 1	Group 1
FISH	DAIRY	EGG	ORANGE
Group 1	Group 1	Group 1	Group 1

Name:

Submit

Click here to view the leaderboard + your score after submitting!!

(b) Example of game play

Figure 12: Human evaluation interface