Think Globally, Group Locally: Evaluating LLMs Using Multi-Lingual Word Grouping Games

César Guerra-Solano **Zhuochun Li** **Xiang Lorraine Li**
School of Computing and Information
University of Pittsburgh
{cguerrasol, zhl163, xianglli}@pitt.edu

Abstract

Large language models (LLMs) can exhibit biases in reasoning capabilities due to linguistic modality, performing better on tasks in one language versus another, even with similar content. Most previous works evaluate this through reasoning tasks where reliance on strategies or knowledge can ensure success, such as in commonsense or math tasks. However, abstract reasoning is vital to reasoning for everyday life, where people apply "out-of-the-box thinking" to identify and use patterns for solutions, without a reliance on formulaic approaches. Comparatively, little work has evaluated linguistic biases in this task type. In this paper, we propose a task inspired by the New York Times *Connections*: GLOBALGROUP, that evaluates models in an abstract reasoning task across several languages. We constructed a game benchmark with five linguistic backgrounds – English, Spanish, Chinese, Hindi, and Arabic – in both the native language and an English translation for comparison. We also proposed game difficulty measurements to evaluate models on games with similar difficulty, enabling a more controlled comparison, which is particularly important in reasoning evaluations. Through experimentation, we find English modalities largely lead to better performance in this abstract reasoning task, and performance disparities between open- and closed-source models.[1]

1 Introduction

Abstract reasoning is driven by dynamic thinking, encouraging identifying and building patterns to solve problems, rather than largely relying on previous knowledge or structured solutions. Large Language Models (LLMs) have demonstrated strong performance in reasoning-centered tasks, where models must reason given a set of information (DeepSeek-AI et al., 2025). This performance can be language-dependent, manifesting as differential task performance depending on the language, even given similar content (Etxaniz et al., 2024; Ahuja et al., 2023; Vayani et al., 2024). However, these investigations have left abstract reasoning tasks relatively unexplored, with a prevalence of English tasks here, presenting a gap for further research (Xu et al., 2024b; Huang et al., 2024).

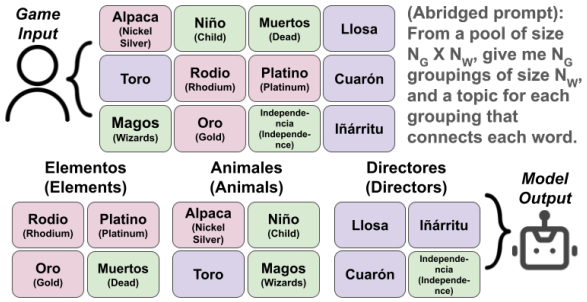

Figure 1: An example Spanish (ES) GLOBALGROUP game. The model is given 12 words and is instructed to output 3 groups of 4 words, along with the topics it identifies connects each group of words. In the figure, same-colored words form a group. The model made mistakes in its grouping.

An important aspect of abstract reasoning is the ability to understand and utilize knowledge to reason in complex situations, such as reasoning under constraints. This has been explored through work investigating the usage of the New York Times (NYT) *Connections*, a constrained word grouping game, as an abstract reasoning evaluation. The game tasks players to, with a pool of 16 words, find 4 groups of 4 words, each linked under a topic.

While the game-based format affords a unique and customizable test-bed for assessing abstract reasoning capabilities, as the game's format could include different groupings to assess reasoning capabilities relative to different content, the game is only available in English and no work has investigated multilingual derivatives. To bridge these gaps, we introduce a game-based reasoning benchmark to evaluate an LLM's ability when solving

[1] **Code and data**: https://github.com/cgsol/globalgroup

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20156–20177
November 4-9, 2025

abstract reasoning tasks in different languages.

We propose GLOBALGROUP, a novel word grouping game inspired by NYT *Connections*, that evaluates a facet of a model's abstract reasoning capabilities in different languages. In GLOBALGROUP, the LLM is given a pool of words, and with them, must create equal groups of words and provide a topic (oftentimes more abstract than just pertaining to semantic relatedness) that connects each group's words. The model must reason about the given set of words to succeed because of the game's constraints, i.e., limiting word group size. This involves optimizing the word groupings by identifying commonalities among words without interfering with other grouping choices. Due to this game-based format enforcing lateral thinking, such as inductive reasoning via recognizing group patterns, rather than knowledge proficiency or other forms of structured problem-solving, GLOBALGROUP assesses abstract reasoning capabilities across languages, addressing a gap in current reasoning evaluation research. Figure 1 presents a Spanish GLOBALGROUP example. The dataset contains groupings from 5 languages – Spanish, Chinese, Hindi, Arabic, and English – and 511 *Connections* and 600 *Connections*-derived games.

We evaluate four open- and two closed-source models and find:

- Across all models, we see a **bias towards English representations**, represented by increases in performance when translating non-English groupings into English.
- By comparing open- and closed-source model results, the **immense value of a multilingual-focused training paradigm** is shown, as this allows a small open-source model to perform on-par with far larger closed- and open-source LLMs. We additionally note a **significant contribution of model size** to both overall performance and translation-related performance differences.
- Through a game difficulty-based analysis, we identify **three game features ("difficulty metrics") correlating with model performance**, hinting at possible game factors impacting model reasoning ability.

2 Related Work

Reasoning capabilities of LLMs have become significantly more advanced over time. Recent work has explored evaluating these capabilities, spanning across a variety of reasoning types. For example, mathematical and code reasoning evaluations test skills relating to recognizing and applying formulas or algorithms to achieve a correct answer (Hendrycks et al., 2021; Austin et al., 2021; Li et al., 2025; Wang et al., 2024a). Commonsense reasoning integrates a variety of reasoning types, with a dependence on prior "commonsense knowledge". Many of this commonsense knowledge relies on social customs and norms, which is also the focus of many social reasoning evaluation benchmarks (Talmor et al., 2019; Sap et al., 2019).

An important reasoning type is abstract reasoning, where people employ skills like abductive or inductive reasoning to identify and build patterns for solutions, rather than relying on existing content, like formulas or specific knowledge. With this, evaluations encourage out-of-the-box thinking and less reliance on pre-set strategies or formulaic approaches (Huang et al., 2024; Xu et al., 2024b). While multilingual studies are prevalent in assessing other forms of reasoning (Etxaniz et al., 2024; Ahuja et al., 2023; Vayani et al., 2024), little work has done so for abstract reasoning, with primarily English evaluations here. With the importance of abstract reasoning to everyday life and human intelligence, it is vital that we evaluate the extent of a potential linguistic disparity, informing future developments in creating equitable and diverse models.

NYT *Connections*, a game released in June 2023, has recently been used as an abstract reasoning evaluation for LLMs. In the game, the player, with a pool of 16 words, must find 4 groups of 4 words, with each group of words connected by a topic, varying in abstractness. These topics relate to domains like alternate word usages, pop culture, morphology, and more. *Connections* games additionally oftentimes have word overlap between groupings (words that could reasonably be put in several prospective groupings in a game), or red herrings (words that could be within a topic/group, but aren't actually, such as "Sponge", "Bob", "Square", and "Pants"). Its use for this purpose is due to its emphasis on abstract reasoning skill, rather than just knowledge, for success in-game – the game has users identify and optimize potential groupings, in light of potential red herrings, through a lateral thinking paradigm where pre-set strategies may not lead to success (Merino et al., 2024).

Prior work converted the Connection game to

a standard English abstract reasoning evaluation task (Samadarshi et al., 2024; Todd et al., 2024). In these experiments, evaluation was performed using *Connections* games, with no new groupings or games created. These works also do not require the model to produce possible group names, making it difficult to understand rationale behind model predictions. Additionally, little work has used games (or *Connections* derivatives) as multilingual reasoning evaluations for LLMs. Our work strives to fill this gap, using a multilingual game to assess abstract reasoning.

3 Dataset

GLOBALGROUP has a pool of x words, with the game goal being to form m groups of n words where $m \times n = x$. Words within the same group share a common topic or category, distinguishing them from words in other groups. Finding a solution to the game requires a reasoning process, which we aim to evaluate in different languages. Models are prompted with the word pool and instructed to return groupings as previously specified and topics connecting each grouping's words.

We create GLOBALGROUP games by sampling m groupings from a language's word grouping dataset, and sampling n words from each grouping to create a game with a pool of x words.

3.1 Word Groupings

We create 6 word groupings datasets – English (EN), Spanish (ES), Chinese (ZH), Hindi (HI), Arabic (AR), and compiled groupings from 511 *Connections* games (NYT). To create the grouping dataset for a language, native speaker annotators are asked to create groups of 4 words connected by a topic. Due to the prevalence in *Connections* of culturally-related groupings (with topics pertaining to aspects of culture in English-speaking regions, such as slang, cultural dishes, or pop culture) and non-culturally-related groupings (with topics less pertaining to aspects of culture, such as animals), annotators are asked to generate groups with culturally-related and non-culturally-related topics relative to their specific language to mirror this quality, labeling groups as such. The authors then agree on these labels. Annotator background information can be found in Appendix F.

Following a quality check of annotator groupings[2], this resulted in 48 EN, 48 ES, 80 ZH groupings, 49 HI, and 40 AR groupings. We additionally translate non-English groupings into English, creating four additional groupings datasets: ES-EN (ES in English), ZH-EN (ZH in English), HI-EN (HI in English), and AR-EN (AR in English). These translations were performed by the annotators, with verification assistance from online translation tools Google Translate[3] and DeepL[4]. Further details on the translation protocol are in Appendix F.

3.2 The New York Times Connection Game

Additionally, a dataset of groupings from *Connections* was collected from an online guide[5], featuring all groupings across 511 *Connections* games (2,044 groupings total). These groupings were used as sources for GLOBALGROUP games via random sampling. These were included in experimentation to compare performance between author-created GLOBALGROUP groups and games (especially in EN) to existing *Connections* groups and games.

3.3 GLOBALGROUP

The default GLOBALGROUP game consists of 16 words forming four groups of four. To evaluate model performance across various game settings, we also created games with 2, 3, and 4 groups, each containing 2, 3, or 4 words per group. These variations were generated through random sampling without replacement, resulting in 9 experimental settings for EN, ES, ES-EN, ZH, ZH-EN, HI, HI-EN, AR, and AR-EN. For each setting, 600 games were created via sampling and then evenly split into development and test sets with 300 games for each set. The authors further ensured that no repeating words or group topics appeared within a game, guaranteeing that each game had a single correct solution. This is further discussed in Appendix C.

Additionally, to compare performance between author-created GLOBALGROUP groups and games and *Connections* groups and games, *Connections* groupings from 511 games were also used. These games all have 4 groupings per game. To create diverse experiments, words per grouping in *Connections* games were randomly sampled to create NYT-SEQ, resulting in 3 experimental settings (2-, 3-, and 4-word 4-group games) with 211 and 300 games in the development and test sets respectively.

[2]The quality check included ensuring that the groupings consist of words connected by a topic, and that topics in a groupings dataset do not repeat.

[3]https://translate.google.com/

[4]https://www.deepl.com/en/translator

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

Dataset Background Subsets	Available Languages	Non-Culturally-Related Group Topic (in en)	Culturally-Related (C-R) Group Topic (in en)	C-R Group Amount	Total Group Amount	Game Settings ($N_{words} \times N_{groups}$)
en	en	Professions	American Holidays	22	48	$\{2,3,4\} \times \{2,3,4\}$
es	es, en	Aquatic Animals	Characters From Don Quixote	27	48	$\{2,3,4\} \times \{2,3,4\}$
zh	zh, en	Tree Composition	The Five Classics	35	80	$\{2,3,4\} \times \{2,3,4\}$
hi	hi, en	Parts of a Train	Fairy Tales	32	49	$\{2,3,4\} \times \{2,3,4\}$
ar	ar, en	Things That Spin	Traditional Food	19	40	$\{2,3,4\} \times \{2,3,4\}$
nyt-seq	en	Basic Directions	Starts of U.S. Presidents	N/A	2044	$4 \times 2, 4 \times 3, 4 \times 4$
nyt-shuf	en	Basic Directions	Starts of U.S. Presidents	N/A	2044	$\{2,3,4\} \times \{2,3,4\}$

Table 1: Dataset Overview: Our dataset includes games from five languages, with 11 subsets, including the English translations of ES, ZH, HI, and AR. The NYT-SEQ and NYT-SHUF dataset is sourced from the New York Times Connections, separated due to different game creation procedures. We provide examples of both non-culturally-related and culturally-related groups to show the dataset diversity. Except for NYT-SEQ, which has three game settings, all other subsets have nine game settings. Additional dataset examples can be found in Appendix G.

Intuitively, words in the *Connections* game have the potential to count for multiple groupings (*overlap* between groupings), which can make the task more difficult as players need to consider multiple group possibilities at once (Merino et al., 2024). This notion of word overlap is explicitly used to create challenging *Connections* games (Liu, 2023). Thus, we hypothesize that purposeful word overlap between groupings from *Connections* games increases difficulty. To evaluate this, we also randomly sampled groups and words from the original games to create NYT-SHUF, resulting in 9 experimental settings (2-, 3-, and 4-word 2-, 3-, and 4-group games) with 600 games evenly split into development and test sets, providing an English subset similar to other language subsets in the benchmark.

4 Experiments

For experimentation, we evaluate closed- and open-source models on 4-group, 4-word games for each subset in GLOBALGROUP. Section 4.1 explains the evaluation metrics for all games. Section 4.2 describes how we measure game difficulty and Section 4.3 provides details on the baseline models.

4.1 Evaluation

For evaluation, a model is provided with the GLOBALGROUP game premise, the shuffled word pool for a given game, and instructions regarding the output format. The model is instructed to output several groups of words, along with an associated topic for each group. The exact prompts used can be seen in Appendix A. We evaluate the model's performance based on both the predicted groups and topics.

Matching Attempts to Ground Truth Groups As the models are not required to output their responses in a particular order associated with the true answer, matching the model's predicted groups to the ground truth groups can be challenging. We use an intuitive strategy: true groupings are assigned to the predicted groupings with the greatest set intersection, referred to as "attempt groups." If the number of predicted groups is fewer than the ground-truth groups, the unmatched true groupings are mapped to NULL and scored as complete misses. If the model predicts too many groups, the remaining predicted groups after group mapping are ignored from evaluation. We do so, rather than penalizing for additional group responses, to accommodate the potential for extra groupings outside of lapses in reasoning, such as answer explanations. This is further discussed in Appendix D.

Grouping Evaluation We use **group-level F1** to compare each matched pair of predicted and true groups. For each word in a ground-truth group,

a positive point is defined as the word appearing in the attempted group; otherwise, it is negative. The F1 score for each game is then averaged across groups for a game. Unlike metrics used in works that evaluate with *Connections*, such as the unweighted clustering score (Samadarshi et al., 2024), which considers a predicted group correct only if all words are correct (further referred to as "CTD"), group-level F1 allows a granular analysis of model performance, with model predictions evaluated on a scale based on their degree of correctness. We test and support this hypothesis in Appendix H.

Topic Evaluation We also evaluate the predicted group topics using a **"Topic Achieved" (TA)** score, a boolean indicating if a topic was successfully predicted. The TA score is calculated by comparing predicted and ground-truth topics using FastText embeddings (Bojanowski et al., 2017). We compute the cosine similarity between the predicted topic and its matched true topic, as well as the other true topics in the same game. If the predicted topic and its matched true topic have a similarity ≥ 0.3, and is more similar than the predicted topic and the other true topics, a TA score of 1 is given. We report the success rate of topic prediction.

We experimented with both BERTScore (Zhang* et al., 2020) and FastText embeddings with cosine similarity as the basis for the TA score, testing a range of similarity thresholds from 0.1 to 0.7 in 0.1 increments. To determine the best setting, two human annotators annotated the topic achievement of 200 Chinese grouping attempts. We calculated Randolph's Kappa between the annotators, treating automatic predictions (either BERTScore or FastText similarity) as one of the annotators (Randolph, 2010). From this, we found that a FastText embeddings-based approach with a similarity threshold of 0.3 performed best, attaining a Randolph's Kappa score of 0.55. Further details are in Appendix E.

4.2 Game Difficulty Measurement

Linguistic modality is not the only factor influencing model performance. Due to the nature of any game, varying difficulty levels (like differing game factors) can impact performance. Therefore, we experimented with different difficulty metrics for GLOBALGROUP games. By classifying game difficulties, we can compare model performance across languages while controlling for these difficulty metrics, allowing finer-grained reasoning analysis which is often coveted when evaluating reasoning in LLMs. Possible sources of game difficulty include handling larger amounts of information, such as having more words to group, or unusual groupings, like dissimilar words in the same group. Other challenges arise from the game's structure – perceived word overlap between groups (where a word could belong to several topics) has been identified as a significant contributor to difficulty in *Connections* (Merino et al., 2024; Liu, 2023). We explore all of these factors as explained below. We experimented with the difficulty measurements in Section 6.

Group Size We hypothesize that game size is a strong contributor to game difficulty. We change game size by varying the number of word groupings in a game, and words per grouping in a game.

Clusters by Word Embeddings Another source of difficulty could arise when dissimilar words belong to the same group, meaning group word connectedness is less tied to semantic meaning. To verify this, we represent word similarity within a topic by mapping true word groupings to groups formed by clustering words using semantic embeddings (creating groups connected by semantic similarity) and evaluate inter-group similarity using the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985). Words within each grouping are represented by their embedding, calculated by tokenizing each word or phrase and averaging the embeddings across the tokens. Groupings with poor ARI scores relative to their matched similarity-based grouping indicate low semantic similarity between words potentially making the game more difficult.

Word Overlap between Groups The final difficulty measurement we explored extends beyond individual groupings in the game. We hypothesize that "word overlap" between groupings, represented by words that could fall under several prospective group topics, is a reasonable predictor of difficulty. This has been confirmed as a significant contributor to difficulty in *Connections*, where word overlap is considered a desirable feature for creating a challenging game (Liu, 2023) and is used in modeling custom game creation (Merino et al., 2024). To measure word overlap in the game, we ask GPT-4o-mini to propose grouping candidates without the game group size constraints, and then compute the average group intersection (via set intersection) across these proposed groupings for a

Models	F1 Score										
Datasets	es	es-en	zh	zh-en	hi	hi-en	ar	ar-en	en	nyt shuf.	nyt seq.
GPT-3.5-Turbo	0.828	0.855	0.927	0.836	0.886	0.948	0.756	0.916	0.874	0.594	0.531
GPT-4	**0.892**	**0.920**	**0.971**	**0.935**	**0.963**	**0.964**	**0.877**	**0.957**	**0.943**	**0.813**	**0.707**
Llama3-8B	0.676	0.743	0.631	0.849	0.724	0.878	**0.829**	0.815	0.776	0.606	0.610
Llama3.1-70B	**0.830**	**0.860**	**0.954**	**0.944**	**0.942**	**0.957**	0.823	**0.916**	**0.886**	**0.708**	0.642
Mistral-7B	0.565	0.677	0.771	0.825	0.497	0.927	0.612	0.789	0.790	0.584	0.577
Aya-8B	0.745	0.794	0.884	0.843	0.873	0.940	0.734	0.814	0.822	0.656	**0.672**
	% FastText Topic Achieved (threshold = 0.3)										
GPT-3.5-Turbo	0.597	0.693	0.661	0.706	0.626	0.755	0.495	0.648	0.700	0.312	0.288
GPT-4	**0.663**	**0.725**	**0.721**	**0.712**	**0.775**	**0.790**	**0.622**	**0.792**	**0.758**	**0.463**	**0.407**
Llama3-8B	0.418	0.543	0.434	0.661	0.346	0.676	**0.548**	0.563	0.554	0.293	0.280
Llama3.1-70B	**0.622**	**0.697**	**0.691**	**0.719**	**0.673**	**0.785**	0.541	**0.658**	**0.705**	**0.387**	**0.338**
Mistral-7B	0.257	0.550	0.448	0.650	0.296	0.744	0.429	0.562	0.572	0.288	0.296
Aya-8B	0.445	0.507	0.596	0.578	0.572	0.685	0.437	0.481	0.522	0.244	0.228

Table 2: The averaged results by model across each 4-group, 4-word game used during experimentation. Results are divided by each dataset and between closed- and open-source models. A bar plot is in Appendix I.4.

game, giving each game a word overlap score. This is further discussed in Section 6.

4.3 Baselines

We experimented with zero-shot prompting for multiple models. The prompts include the rules of the game and the pool of words for the given game. Additionally, within the prompt, the output format is specified with an example to support the parsing of LLM answers. We experimented with 11 subsets of the datasets for models using each game in each respective dev dataset and reported the results on the test subset. Model responses were parsed using regular expressions and then organized, mapping guessed groupings to guessed topics. During this process, if any game caused issues with answer parsing due to bad response formatting, the model response was reformatted via prompting of GPT-4o-mini. Please find the exact prompts used across experimentation in Appendix A.

Models Six models, including both closed-source and open-source models, are experimented with using the benchmark. For closed-source LLMs, we utilized OpenAI's GPT-3.5-Turbo (Brown et al., 2020) and GPT-4 (OpenAI et al., 2024). For open-source LLMs, we utilized Meta's Llama-3-8B-Instruct and Llama-3.1-70B-Instruct (Dubey et al., 2024) and Mistral AI's Mistral-7B-Instruct-v0.2 (Jiang et al., 2023). These three models are three of the highest-performing open-source LLMs at the time of writing, with good performance in various NLP tasks such as reasoning, mathematics, and code generation (Xu et al., 2024a). This selection of models additionally allows us to observe potential scaling effects, with two different sizes of the same model. An additional open-source multilingual LLM, Cohere Labs' Aya-23-8B (Aryabumi et al., 2024), was also used in the experimentation due to its multilingual focus[6]. More details of the model parameters are shown in Appendix B. These models were chosen not only due to their high performance across a variety of benchmarks, but also to offer a comparison between privately-owned LLMs of larger size, versus smaller, open-source models.

5 Results

We report the model performance across models and subsets of 4-group, 4-word (4×4) games in the benchmark in Table 2. Additional analyses, omitted due to spatial constraints, can be seen in Appendix I.

Performance with Language of Origin Considering both the F1 and TA scores, we can see in Table 2 that ES, HI, and AR groupings performed significantly better when translated to English across all models. ZH groupings had a differing relationship, with better performance in Chinese rather than English translations in closed-source models, but English better than Chinese in open-source models, except for Aya-8b. With the multilingual

[6]We have tested other open-source and multilingual LLMs: Llama2-7B-chat (Touvron et al., 2023), Apollo-7B (Wang et al., 2024b), PolyLM-13b (Wei et al., 2023), and Bloomz-7b1 (Muennighoff et al., 2023). However, these LLMs can't understand our tasks and give meaningful output. Therefore, their results are not included.

Mean F1 Score Difference (Non-Culturally-Related - Culturally-Related) Across All Dataset Subsets				
Models	GPT-3.5-Turbo	GPT-4		
Mean	**0.050**	0.032		
Models	Llama3-8B	Llama3.1-70B	Mistral-7B	Aya-8B
Mean	0.044	0.045	**0.086**	0.065

Table 3: A table depicting the average difference in performance in groups labeled as non-culturally-related versus culturally-related in 4×4 games averaged across all dataset subsets (excluding NYT-SEQ and NYT-SHUF) for open- and closed-source models.

focus of Aya-8b's training (Aryabumi et al., 2024), we assume this differing pattern in performance is due to larger amounts of Chinese text training data. This highlights a strong bias towards English language representations in these models, reflecting limited multilingual reasoning capabilities even for the same knowledge, and stresses the need for diverse language choices when developing training paradigms and benchmarks. An additional figure depicting this trend can be seen in Figure 7.

Performance with Culture of Origin Utilizing the culture-relatedness judgment for each grouping from its respective annotator, we investigate the average F1 performance on culturally-related and non-culturally-related groupings, across all datasets. A seen in Table 3, we find that, when compared to culturally-related groupings, non-culturally-related groupings lead to average F1 performance improvements across all models. This substantiates a potential cultural basis to model performance and reasoning capabilities as well. The full results, including the performance difference by dataset, can be seen in Table 10.

Crowd-created Games vs *Connections* Games In Table 2, the first 9 columns represent crowd-created GLOBALGROUP games, while the last 2 columns show games derived from the NYT *Connections* games. We can see that performance on the NYT games is significantly lower across all models, particularly for closed-source models when evaluated using the F1 score, and all models when evaluated with the TA score, suggesting that the NYT *Connections* game may be more challenging, especially when compared to the in-house EN game. By stratifying *Connections* results by the *Connections*-set group difficulty, we find our EN games are similar in performance to "Yellow" *Connections* groupings, the easiest connection grouping type. Further analysis is in Appendix I.2.

When comparing the two variations of the Connections game, we also observed a performance drop for NYT-SEQ, potentially indicating intentionally difficult game design. This is possibly due to intentional group overlap, as that is the only difference between NYT-SEQ and NYT-SHUF, due to the random sampling approach used for NYT-SHUF.

Closed-Source vs Open-Source Model Results Throughout the experiments, one can note a clear grouping of models with respect to performance – the smaller, closed-source LLMs Llama3-8B and Mistral-7B perform similarly badly compared to all other models, even considering Aya-8B, a open-source model of similar size. This demonstrates the value of a multilingual-focused training paradigm, as Aya-8B is able to attain comparable results as the far larger LLMs, despite having a substantially smaller size. This can be seen across all scored metrics, and across all dataset backgrounds, pointing to a clear difference in these models. We hope this informs future model development efforts that seek to efficiently adapt to a variety of linguistic and cultural settings.

Additionally, we can note that model size has a large impact on performance. When comparing results from Llama3-8B to Llama3.1-70B, we can note a general increase in performance, with many subsets now reaching $> 90\%$ F1 score. We can additionally see how model size impacts performance differences relative to translation, as there are considerably smaller performance deficits between English/non-English subsets, pointing to model size directly impacting the extent of linguistic modality-related performance biases.

While the performance gap between open- and closed-source models (aside from Llama3.1-70B) is large in English-based games, this performance gap is far smaller for non-English-based games, such as Arabic games, contrary to initial expectations. This can inform future work relating to LLM training, and offers a perspective on what tasks can smaller, open LLMs, perform on par with larger, oftentimes cost-intensive, closed LLMs. In this case, we see that these smaller, open LLMs can perform similarly to the larger, closed LLMs on Arabic tasks.

We additionally perform a principal component analysis, where each model is represented by a vector of their mean 4×4 game results by dataset subset, plotting the first two principal components for each model. As seen in Figure 2, the closed-

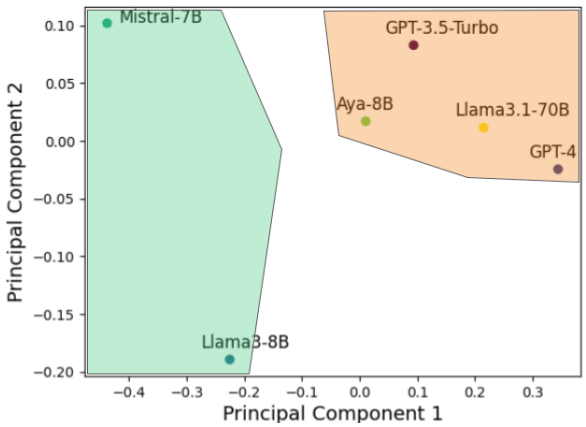

Figure 2: PCA plot of model representations based on their performance. Each point corresponds to a model, projected onto the first two principal components.

source models are clustered together while the open-source models form another cluster, additionally seen in K-means clustering with $k = 2$. Investigating the contributions of each dataset to the first two principal components, we find that HI, ZH, and AR have the largest weights, with all of these being the non-Latin script dataset subsets.

6 Difficulty-Aware Benchmark Results

Model performance on each subset of the game depends on several factors, such as linguistic/cultural differences and game difficulty, which involves both the complexity within the groups and the difficulty across groups within the game. In order to provide granularity in model evaluation, through using games of a specific "difficulty level", and control for potential confounding factors to better identify language-related differences in performance, we propose several measures to assess game difficulty. We evaluate the effectiveness of these measures in this section. Additional analyses, such as correlation calculations and categorized difficulty results, can be seen in Appendix K.

A good difficulty measure should correlate well with model performance, i.e., the harder the game, the worse the model's performance when game difficulty (agnostic of language or cultural setting) is the only difference between games. Following this intuition, we calculate model performance results for different difficulty metric slices of the benchmark, such as varying group counts or group sizes in the game. For ARI score and word overlap, we bin them into 3-4 categories and calculate the average performance by bin. The results for GPT-4 and GPT-3.5-Turbo on all game sizes are in Table 4.

Models	**Average F1 Score**			
Group Size	2	3	4	
GPT-3.5-Turbo	0.881	**0.890**	0.884	
GPT-4	0.925	0.940	**0.949**	
Group Count	2	3	4	
GPT-3.5-Turbo	**0.930**	0.883	0.842	
GPT-4	**0.962**	0.940	0.912	
Adjusted Rand Index Range	[-0.5, 0.0]	(0.0, 0.5]	(0.5, 1.0]	
GPT-3.5-Turbo	0.924	0.927	**0.951**	
GPT-4	0.969	0.974	**0.990**	
Candidate Proposal Overlap Range	[0.0, 0.75]	(0.75, 1.5]	(1.5, 2.25]	(2.25, 3.0]
GPT-3.5-Turbo	**0.932**	0.873	0.813	0.831
GPT-4	**0.977**	0.923	0.863	0.859

Table 4: The averaged F1 score for closed-source models across each 4-word game (4-group for group size analysis) whose group size, group count, ARI, and Word Overlap score fell within the specified ranges. Results are divided by 3-4 ranges, and averaged across all dataset backgrounds, excluding NYT-SEQ.

We additionally use these difficulty metrics for controlled comparisons across models and languages, although due to spatial constraints, that has been moved to Appendix K.

Group Size & Group Count We report results from 4-group games for analyzing the impact of group size, and 4-word groups for analyzing the impact of group count. We hypothesize that the number of words in each group is associated with game difficulty. From a player's perspective, increasing the group size or context may offer additional opportunities to identify group topics, making the game easier. However, as shown in Table 4, the observed patterns are inconsistent. While performance clearly increases in games as group size increases for GPT-4, there is an inconsistent pattern for GPT-3.5-Turbo, with performance unusually peaking with a group size of 3, indicating that grouping size may not be a good predictor of game difficulty. In contrast, varying the number of groups reveals clear patterns in model performance. This is reflected with the average performance relative to the group count of games considerably declining with an increase in group count for both models. Thus, we adopt game group count as one of the main difficulty measurements.

Adjusted Rand Index We represent words in each ground-truth group using FastText embeddings and apply K-means clustering, setting k to the number of groups in a game. We use 2, 3, and 4-group games with a group size of 4, giving us

30 subsets across the 10 benchmark subsets. For each subset, we evaluate the clustering results with the Adjusted Rand Index (ARI), binning the results into three categories. We hypothesize that a higher ARI indicates greater semantic similarity within the groups, making the game easier. To test this hypothesis, we compare model performance in games within 3 different bins of ARI values. As shown in Table 4, both GPT-4 and GPT-3.5-Turbo exhibit very strong trends, with model performance consistently improving with higher ARI scores. We conclude that increased group inter-word similarity is associated with better performance.

Word Overlap Assessment To calculate word overlap, we conducted candidate proposal experiments. Here, we prompted GPT-4o-mini with the word pool of a game and the final group topics, asking it to assign words to topics, with replacement. By allowing words to be assigned to multiple topics, we capture the word overlap between groups in-game. We calculate the overlap by parsing each list of words per topic and computing the set intersection between the lists. The average number of overlapped words in a given game is then calculated by averaging this value across all topics. To validate GPT-4o-mini's performance in this task, a subset of candidate proposal assignments for EN were cross-checked by the authors. We observe only valid responses and no illogically overlapped words, therefore we proceed with this experimentation. We then follow the same experimental setup to calculate the average F1 score in games within 4 bins of overlap values, as done in the clustering experiment. As shown in Table 4, word overlap proves to be associated with performance, with performance trending as expected, decreasing as word overlap increases. Consequently, we conclude that word overlap is also associated with game difficulty.

7 Conclusion

We have presented a new multilingual dataset, GLOBALGROUP, to assess LLM abstract reasoning across languages through a novel, game-based perspective. By having game mechanics which require high performance in abstract reasoning/lateral thinking (encouraging inductive reasoning and pattern recognition), we have created an evaluation method that can assess this trait across multiple languages. As evidenced through experimentation, GLOBALGROUP can be used to identify language-related bias, seen through performance increases with English translations, and note consistent performance differences between the closed and open-source models that were tested, observing the value of a multilingual-focused training paradigm for smaller models. Additionally, we provide three difficulty metrics for predicting game difficulty for a model, reflecting previous judgments of aspects contributing to game difficulty for humans. With this benchmark, we provide an avenue for evaluating LLM reasoning abilities across languages, providing a valuable means for assessing linguistic bias within LLMs – a key component of creating diverse and equitable language systems.

8 Limitations and Ethical Considerations

In this paper, we only cover five languages, while recognizing that there are thousands of languages in the world that could be studied, along with a multitude of cultural backgrounds that are not represented. We encourage community members to add to our dataset using our explicit annotator request and dataset construction procedures.

For the language we covered in the paper, we recognize the vast and diverse nature of language and culture. Since our annotators do not reflect the complete demographics of speakers of a given language, some form of bias may have been introduced. We also acknowledge that culture is a complex term that involves different levels of communities and that our approximation of cultural-relatedness may not fully encompass the vastness of what culture represents. Another limitation is that while the authors filtered for unethical or harmful content in the word groups, while very unlikely, there is a possibility that we might have overlooked some.

Acknowledgments

We would like to thank the following entities for their gracious and continual support of our work through resources, feedback, and discussion:

- This research was supported in part by the University of Pittsburgh Center for Research Computing and Data, RRID SCR_022735, through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483.
- The PittNLP community, for consistent help in data annotation and thoughtful feedback throughout the development of this project.

- All annotators, including members of the PittNLP community, the Durrant Lab, the Dietrich School of Arts and Sciences, and community members outside of the University of Pittsburgh, for incredible and diligent work with annotation and translation.

- The Stamps Foundation, for providing scholarship support and additional funding throughout the project for the first author.

- We also thank Allyson Ettinger and Vered Shwartz for the fruitful discussion and feedback for the project.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Haitao Zheng. 2024. LatEval: An interactive LLMs evaluation benchmark with incomplete information from lateral thinking puzzles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10186–10197, Torino, Italia. ELRA and ICCL.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *J. Classif.*, 2(1):193–218.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhuochun Li, Yuelyu Ji, Rui Meng, and Daqing He. 2025. Learning from committee: Reasoning distillation from a mixture of teachers with peer-review. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4190–4205, Vienna, Austria. Association for Computational Linguistics.

Wyna Liu. 2023. 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. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, pages 87–96.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-

der, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Justus Randolph. 2010. Free-marginal multirater kappa (multirater free): An alternative to fleiss fixed-marginal multirater kappa. volume 4.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Prisha Samadarshi, Mariam Mustafa, Anushka Kulkarni, Raven Rothkopf, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Connecting the dots: Evaluating abstract reasoning capabilities of llms using the new york times connections word game. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Graham Todd, Timothy Merino, Sam Earle, and Julian Togelius. 2024. Missed connections: Lateral thinking puzzles for large language models. *2024 IEEE Conference on Games (CoG)*, pages 1–8.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, Sanjay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Hafizi Amirudin, Muhammad Ridzuan, Daniya Najiha Abdul Kareem, Ketan More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Fabian Farestam, Muztoba Rabbani, Sanoojan Baliah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Xavier, Amit Bhatkal, Hawau Olamide Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman H. Khan, and Fahad Shahbaz Khan. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19565–19575.

Lun Wang, Chuanqi Shi, Shaoshui Du, Yiyi Tao, Yixian Shen, Hang Zheng, Yanxin Shen, and Xinyu Qiu. 2024a. Performance review on llm for solving leetcode problems. *2024 4th International Symposium on Artificial Intelligence and Intelligent Manufacturing (AIIM)*, pages 1050–1054.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024b. Apollo: Lightweight multilingual medical llms towards de-

mocratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Han Xu, Xingyuan Wang, and Haipeng Chen. 2024a. Towards real-time and personalized code generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5568–5569.

Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias Boutros Khalil. 2024b. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *Trans. Mach. Learn. Res.*

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Prompts

A.1 Model Response Prompting

From these prompts, "#*#" is replaced with the total number of words in the pool, $*$ is replaced with the total number of groups in the game, and is replaced with the provided pool of words for a given game. A new prompt is generated for each new game, with each game's metrics being used to fill these spaces in the prompts.

English prompt for 2-group games:
I am going to give you a pool of #*# words. These #*# words can be separated into $*$ equal groups of 4 words linked under some category. I want you to tell me the four groups and what category you think connects them. Here is an example: Given the pool ['Mile', 'League', 'Jazz', 'Heat', 'Yard', 'Cabaret', 'Carousel', 'Nets', 'Gobble', 'Scarf', 'Foot', 'Bucks', 'Chow', 'Wolf', 'Cats', 'Chicago'], you would output: <NBA TEAMS>: ['Bucks', 'Heat', 'Jazz', 'Nets'], <UNITS OF LENGTH>: ['Foot', 'League', 'Mile', 'Yard'], <Synonyms For Eat>: ['Chow', 'Gobble', 'Scarf', 'Wolf'], <Musicals Beginning With 'C'>: ['Cabaret', 'Carousel', 'Cats', 'Chicago']. Now, given the pool: . The answer must be $*$ groups, each of them containing 4 words and defined by one category, and the output format must be the same as the example. Give me the answer immediately.

English prompt for 3-group games:
I am going to give you a pool of #*# words. These #*# words can be separated into $*$ equal groups of 4 words linked under some category. I want you to tell me the four groups and what category you think connects them. Here is an example: Given the pool ['Water', 'Happiness', 'Fire', 'Earth', 'Mercury', 'Surprise', 'Wind', 'Sadness', 'Venus', 'Pluto', 'Angry', 'Mars'], you would output: <Natural Elements>: ['Water', 'Fire', 'Earth', 'Wind'], <Emotions>: ['Happiness', 'Sadness', 'Angry', 'Surprise'], <Planets>: ['Mercury', 'Venus', 'Pluto', 'Mars']. Now, given the pool: . The answer must be $*$ groups, each of them containing 4 words and defined by one category, and the output format must be the same as the example. Give me the answer immediately.

English prompt for 4-group games:
I am going to give you a pool of #*# words. These #*# words can be separated into $*$ equal groups of 4 words linked under some category. I want you to tell me the four groups and what category you think connects them. Here is an example: Given the pool ['Water', 'Fire', 'Sad', 'Wind', 'Happy', 'Earth', 'Angry', 'Surprised'], you would output: <Natural Elements>: ['Water', 'Fire', 'Earth', 'Wind'], <Emotions>: ['Happy','Sad','Angry','Surprised']. Now, given the pool: . The answer must be $*$ groups, each of them containing 4 words and defined by one category, and the output format must be the same as the example. Give me the answer immediately.

A.2 Model Response Reformatting Prompt

English prompt for model response reformatting:
Please reformat the following response into a format mapping topic names, encased by < and > with no single quotes, to a list of Python single-quote strings containing the words in a topic (removing the single quotes within each word, like the ' in 'Newton's'). An example of the format is: <Water>: ['Lake', 'River', 'Pond', 'Ocean'], <Family>: ['Uncle', 'Niece', 'Mom', 'Sister'], <American Food>: ['Burger', 'Pizza', 'Wings', 'Steak'], <Precious Metals>: ['Gold', 'Rhodium', 'Platinum', 'Nickel Silver']. Please reformat the following response, outputting only the final reformatted version:

A.3 Candidate Groups for Word Overlap Prompt

English prompt for word overlap grouping proposal for 4-group 4-word games (as other game sizes follow a similar format):

I am going to give you a list of words, as well as a list of topics each word could fall under. Please group the words under each topic, with replacement, allowing words to be grouped under several topics. Each word must be used at least once. Output only these groups in Python dictionary format, mapping a string representing the topic to a list of strings representing the selected candidate words for that topics. Here is an example – given 4 topics ["NBA Teams", "Units of Length", "Synonyms For Eat", "Musicals Beginning With 'C'"] and 16 words ["Bucks", "Foot", "Chow", "Gobble", "Scarf", "Cabaret", "Carousel", "Cats", "Chicago", "Wolf", "League", "Mile", "Yard", "Heat", "Jazz", "Nets"], you could output in Python dictionary format: "NBA TEAMS": ["Bucks", "Heat", "Jazz", "Nets", "Mile", "Chicago"], "UNITS OF LENGTH": ["Foot", "League", "Mile", "Yard", "Carousel", "Bucks"], "Synonyms For Eat": ["Chow", "Gobble", "Scarf", "Wolf", "Cats", "Bucks"], "Musicals Beginning With 'C'": ["Cabaret", "Carousel", "Cats", "Chicago", "Chow"]. Here is the list of topics: LIST_TOPICS . Here is the pool of words from which you can select from: LIST_WORDS. Output solely the groups in the previously specified Python dictionary format.

B LLM Parameter Settings

For the efficiency and accuracy of our evaluation process, we fixed the parameters for the same model during different game evaluations. As our task requires LLMs to have flexible thoughts to explore the internal connections between words and group them, we maintain the default settings for open-source models, as official documents suggest. In addition, we found that greedy sampling can generate nonsense responses with occasional repetitions of one single word. Therefore, we didn't use it in our open-source tasks.

- Parameter settings for Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.2:
  - do_sample: *True*
  - top_k: *50*
  - top_p: *0.9*
  - temperature: *0.6*
  - num_return_sequences: *1*
  - max_length: *512*
- Parameter settings for Aya-23-8B:
  - do_sample: *True*
  - temperature: *0.3*
  - num_return_sequences: *1*
  - max_length: *512*

C Handling the Risk of Multiple Solutions Per Game

When generating a given game, we ensure that there are no repeating words or repeating group topics within a game. This ensures that no more than one solution to the game, utilizing the groupings from which the games are constructed, exists per game.

While there may be additional possible groupings that could be created from a game's pool (like answering with groupings that do not natively exist in a games' language's groupings dataset, such as an additional word counting under a grouping topic, or swapping N words between groups), this possibility is accounted for as an aspect of difficulty in the game through our word overlap analysis. Additional grouping variabilities can also occur in answers, such as more/less words, or repeated words in a grouping; however, these variabilities are due to lapses in reasoning from the LLMs, such as misunderstanding the task or the output format.

Additionally, as we consider the "correct answer" for a game to be all correct groupings submitted (while still scoring individual groupings for the sake of evaluation and analysis), each game would require N complete groupings as the correct solution. While there can be individual grouping member variations (such as word overlap or the potential to "force" words into groupings with new topics by stretching word meaning or usage), for this to constitute a new solution for a game, this would need to occur across all groupings in the game, which is highly unlikely. While there may be *plausible* solutions to each GLOBALGROUP game (as there exist plausible solutions as one plays *Connections*, the inspiration behind GLOBALGROUP), these do not constitute complete solutions.

D Handling Additional Groups in Model Responses

We considered penalizing the grouping evaluation score if a model provided additional groupings, but however opted instead to only consider N answered groupings per game (where N is the group count for a game) for a few reasons. While there is the potential for a model to output extra groupings

for a "game-breaking" reason (such as outputting all permutations of groupings for a game to ensure attaining the correct answer), this phenomenon has not been observed, with models outputting extra groupings for non-game-breaking reasons being far more common.

For example, we can observe that models can partially repeat their answer, or following their answer, output smaller straggler groups that do not match the game group size requirements. Here is an example response from Llama-3.1-8B in the ES-EN 4-group/4-word dataset:

- "Here is the solution: The four groups are: <VEGETABLES>: ['Celery', 'Carrot', 'Broccoli', 'Leaf'], <PLANT PARTS>: ['Branch', 'Trunk', 'Root', 'Leaf'], <WRITING TOOLS>: ['Pen', 'Pencil', 'Eraser', 'Envelope'], <VERBS>: ['Sang', 'Leaped', 'Yelled', 'Cried']. The category that connects them is that they are all related to everyday life, but they belong to different domains. The first group is related to food, the second is related to plants, the third is related to writing and communication, and the fourth is related to human actions and emotions. Here is the answer in the required format: <VEGETABLES>: ['Celery', 'Carrot', 'Broccoli', 'Leaf'], <PLANT PARTS>: ['Branch', 'Trunk', 'Root', 'Leaf'], <WRITING TOOLS>: ['Pen',"

Additionally, models sometimes restate their groupings in subsequent discussion in their answer, sometimes with small variations (like omitting a group, or omitting a word in a group), or in a "reasoning process", as we see models correct themselves in their own answer. Here is an example response from Llama-3.1-8B in the ES-EN 4-group/4-word dataset, where the model corrects itself (with this correction being used for scoring):

- "Solution: The 16 words can be separated into four groups linked under the category of ""Titles of Positions in the Catholic Church"". Here is the output:<TITLES OF POSITIONS IN THE CATHOLIC CHURCH>: ['Archbishop', 'Cardinal', 'Father', 'Priest'], <UNITS OF LENGTH>: ['Cool', 'Ho', 'Powder', 'Shell'], <COOKIES>: ['Shortbread', 'Solid', 'Fantasy', 'Cc'], <RELIGIOUS TITLES>: ['Man', 'Monk', 'Pas', 'Roll']. Note: The word 'Fantasy' does not belong to the category, it should be 'Fantastic'. Also, the word 'Pas' does not belong to the category, it should be 'Coc'. The correct answer is:<TITLES OF POSITIONS IN THE CATHOLIC CHURCH>: ['Archbishop', 'Cardinal', 'Father', 'Priest'], <UNITS OF","

In these situations, penalizing models for their responses having extra groups will then unfairly penalize models that offer additional discussion aside from solely their response (where they restate their groupings and inconsistencies can creep in), or when they provide reasoning steps toward their response (with intermediate groupings then being penalized). Because of the frequency of these cases observed in preliminary testing, and as observed in our evaluations using GLOBALGROUP, we opt instead not to penalize models for providing extra groupings, instead only considering N groupings per model response.

E Determining Topic Achieved

The threshold value of 0.3 was selected through experimentation. The basis of the Topic Achieved score was evaluated using both BERTScore and FastText embeddings-based cosine similarity as measures of similarity between topics. 7 Different threshold values were used for score generation (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, and 0.7). For 100 games of the 2-group, 2-word ZH dataset, we had human annotations of topic achieved scores. Human annotators were provided the same rules for calculating the Topic Achieved score but were asked to base it on their own perception of similarity. As a means of selecting the metric most similar to human judgment, we calculated the inner-annotator agreement utilizing Randolph's Kappa between the human annotations and each of the methods used for calculation. From this, we found that a FastText embeddings-based approach with a similarity threshold of 0.3 performed best, attaining a Randolph's Kappa score of 0.55. A BERTScore-based approach with a similarity threshold of 0.5 had the second-best score of 0.53.

Additionally, due to the lack of context provided in *Connections* and GLOBALGROUP to inform word meaning and group membership, semantic embeddings were preferred.

F Annotator Information

F.1 Annotator Request

Annotators were initially provided with the following message, choosing to submit groupings for usage on a volunteer basis:

"Hello [ANNOTATOR NAME]! My name is [NAME], and I am an working with [NAME] on a project using word grouping games (similar to "Connections" from the New York Times) to assess multilingual reasoning. We are creating word grouping datasets in source languages, composed of groupings of four words or very small phrases connected by an associated topic. These topics don't necessarily pertain just to semantic meaning of the words, and can relate to other potential uses/meanings of the words. Here's a sample from our English dataset:

[Teacher, Scientist, Engineer, Doctor] – Professions [Pool, Jacuzzi, Ocean, Lake] – Places to Swim [Rocky, Appalachian, Adirondack, Sierra Nevada] – Mountain Ranges [Hansel, Witch, Bear, Wolf] – Fairy Tale Figures

The groupings can be connected by non-culture-related topics (i.e., topics whose contents can be recognized across cultures, such as "Professions" or "Places to Swim") or culture-related topics (i.e., topics whose contents are better recognized in cultures associated with that language, such as Western/North American culture being familiar with the specific mountain range names in "Mountain Ranges"). With your knowledge of [LANGUAGE], I wanted to ask if you would be available to write custom word groupings in [LANGUAGE]. We want to collect 25-50 groupings each in additional languages, with an even split between non-culture-related group topics and culture-related group topics per language. No worries at all if that is a bit too much – however many groupings/languages you could commit to would be fine. I have attached the dataset of our English groupings for your reference. If this would be possible, please let me know! I am available via Slack and via email [EMAIL] if you have any questions."

Annotators, upon responding favorably to this request, were then provided with further information on the task, including how the groups would be used (random sampling to create games) and providing examples when needed. Annotators were then informed of the translation protocol for translating groupings in their source language into English. This is further detailed in Appendix F.3.

F.2 Annotator Demographics

The 2 annotators for the EN groupings had the following background statistics:

- 2 were university-educated
- 2 were born and raised in the United States
- 1 was male, 1 was female

The 7 annotators for the ES groupings had the following background statistics:

- 5 were university-educated, 2 were not
- 3 were born and raised in Mexico, 4 were born and raised in the United States within Mexican families
- 3 were male, 4 were female

The 4 annotators for the ZH groupings had the following background statistics:

- 2 were university-educated, 2 were not
- 4 were all born and raised in China.
- 1 were male, 3 were female

In addition, some Chinese groupings were created by searching and checking online resources, as well as Chinese LLMs such as the Baidu ERNIE Bot (https://yiyan.baidu.com/). We used Baidu ERNIE for several simple Chinese groups creation. For example, we used the prompt: "Please list several animals that are common to see in our daily lives". We then manually checked and picked 4 animals in Chinese from their answers to form one simple Chinese group, for example: cat, dog, bird, and fish.

The 3 annotators for the HI groupings had the following background statistics:

- 3 were university-educated
- 2 were born and raised in India, 1 was born in the United Kingdom and raised in the United States
- 3 were male

The 2 annotators for the AR groupings had the following background statistics:

- 2 were university-educated
- 2 were born and raised in Egypt
- 1 was male, 1 was female

F.3 Protocol for Translation into English

Annotators were instructed to translate each word within a group, along with the group topic, from the initial language into English. In situations where a direct translation did not exist, annotators were instructed to translate literally, word for word. This process was assisted with the online translation tools Google Translate[7] and DeepL[8]. The authors verified translations using these tools as well. The same translation process was applied to every grouping within our dataset subsets in order to ensure parity.

G Example WGG Games

G.1 en-es

EN-ES 4-group, 4-word				
Shuffled Game Word Pool				Topic
Neruda	History	Ladybug	Mosquito	Bugs
Mystery	Poems	Mistral	Chilindrina	Latin-American Poets
Locust	Chavo	Butterfly	Martí	Genres
Ramón	Acevedo	Quico	Romance	Characters from El Chavo del Ocho

Table 5: An example 4-group, 4-word game from the EN-ES dataset background. Here, you can see a mix of both colloquial topics, such as "Latin-American Poets", and non-colloquial topics, such as "Bugs".

G.2 en-zh

EN-ZH 4-group, 4-word				
Shuffled Game Word Pool				Topic
Positivism	Idealism	Great Wall	Terracotta Warriors and Horses	Ruminants
Ox	Forbidden City	Existentialism	Water	Ancient Building
Qin Shi Huang Mausoleum	Fire	Earth	Sheep	Natural Elements
Reindeer	Pragmatism	Wind	Sika Deer	Philosophical Concept

Table 6: An example 4-group, 4-word game from the EN-ZH dataset background. Here, you can see a mix of both colloquial topics, such as "Ancient Building", and non-colloquial topics, such as "Natural Elements".

H Group Evaluation Metric Comparison

As can be seen in Figure 3, CTD invokes a great penalization on model responses, with peaks in lower-valued score buckets and a sharp decrease in median-valued buckets suggesting a lack of granularity in model response scoring. Contrastingly, F1 score offers a gradual increase in frequency in increasing-valued score buckets, and no peaks at opposing ends of the score value range, providing a more complete and fine-toothed scoring of model performance in this task.

[7] https://translate.google.com/

[8] https://www.deepl.com/en/translator

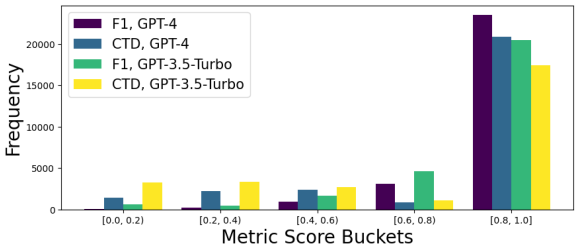

Figure 3: Depicting the distribution of game average F1 and CTD scores across 5 buckets of score ranges, encapsulating the total score range for both metrics. This is evaluated for closed-source models.

Models	**Mean F1 Score in Groups of Tag**		
Group Tag	General Knowledge	Cultural and Pop Culture Knowledge	Linguistic
GPT-3.5-Turbo	**0.902**	0.871	0.826
GPT-4	**0.954**	0.937	0.916
Llama3-8B	**0.804**	0.767	0.724
Llama3.1-70B	**0.935**	0.900	0.859
Mistral-7B	**0.779**	0.694	0.669
Aya-8B	**0.866**	0.818	0.789

Table 7: The averaged performance on groups, separated by tag, by model across each non-NYT 4-group, 4-word game used during experimentation. Results are divided by each group tag and between closed- and open-source models.

I Additional Model Performance Analyses

I.1 Performance Relative to Topical Tagging of GLOBALGROUP Groupings

We additionally perform tagging/typing of each grouping in our grouping datasets, categorizing groupings under three labels:

- **General Knowledge**, where groupings pertain to encyclopaedic knowledge or knowledge gained from everyday life, largely agnostic of cultural origin
- **Cultural and Pop Culture Knowledge**, where groupings pertain to knowledge strongly grounded in culture (such as food, religion, important sites), or pop culture (such as movies, music, and sports)
- **Linguistic**, where groupings pertain to linguistic concepts, such as synonyms, morphology, phonetics, idioms, and slang

Models	Mean F1 Score in Groups of Tag, By Dataset														
Dataset		es			es-en			zh			zh-en				
Group Tag	GK	CPCK	L	GK	CPCK	L	GK	CPCK	L	GK	CPCK	L			
GPT-3.5-Turbo	**0.909**	0.810	0.774	**0.949**	0.841	0.781	**0.959**	0.907	0.904	0.841	**0.866**	0.761			
GPT-4	**0.950**	0.890	0.840	**0.977**	0.924	0.856	**0.981**	0.965	0.960	0.932	**0.954**	0.900			
Llama3-8B	**0.786**	0.656	0.600	**0.833**	0.723	0.680	0.635	**0.649**	0.585	0.845	**0.882**	0.788			
Llama3.1-70B	**0.894**	0.831	0.766	**0.941**	0.857	0.783	**0.982**	0.948	0.906	**0.964**	0.946	0.899			
Mistral-7B	**0.705**	0.536	0.470	**0.809**	0.644	0.589	**0.820**	0.746	0.719	**0.857**	0.846	0.718			
Aya-8B	**0.816**	0.722	0.707	**0.881**	0.754	0.759	**0.930**	0.881	0.783	0.859	**0.863**	0.770			
Dataset		hi			hi-en			ar			ar-en			en	
Group Tag	GK	CPCK	L	GK	CPCK	L	GK	CPCK	L	GK	CPCK	L	GK	CPCK	L
GPT-3.5-Turbo	**0.898**	0.882	0.851	**0.957**	0.940	0.952	**0.779**	0.765	0.702	**0.933**	0.911	0.895	0.889	**0.916**	0.811
GPT-4	0.963	0.960	**0.979**	0.978	0.951	**0.984**	0.882	**0.888**	0.852	**0.968**	0.944	0.957	**0.959**	0.956	0.911
Llama3-8B	**0.766**	0.703	0.665	0.879	**0.880**	0.865	0.830	**0.838**	0.814	**0.859**	0.775	0.799	0.800	**0.803**	0.719
Llama3.1-70B	**0.944**	0.943	0.927	**0.968**	0.947	**0.968**	**0.860**	0.823	0.759	**0.945**	0.900	0.888	**0.915**	0.909	0.832
Mistral-7B	**0.545**	0.464	0.474	**0.941**	0.915	0.939	**0.652**	0.577	0.588	**0.850**	0.728	0.771	**0.833**	0.789	0.749
Aya-8B	**0.914**	0.846	0.858	**0.958**	0.928	0.931	0.713	**0.760**	0.736	**0.864**	0.774	0.787	**0.856**	0.837	0.769

Table 8: The averaged performance on groups, separated by tag, by model across each non-NYT 4-group, 4-word game used during experimentation. Results are divided by each dataset, by each group tag, and between closed- and open-source models. GK = General Knowledge, CPCK = Cultural and Pop Culture Knowledge, L = Linguistic

We do not further type groupings within each category, such as "Linguistic", due to a small number of examples per category.

This annotation was performed by two annotators (university-educated, male, Indian and American backgrounds) until full agreement was reached. From this, we perform an analysis, looking at F1 performance in groups within these tags, across datasets, across models. We report results averaged across all 4-group/4-word datasets, per model, in Table 7, and separated by language dataset in Table 8.

From this, we can observe a clear trend in performance relative to the type of content present within a grouping – across all models, average performance in groupings tagged as a category averaged across all datasets decreases by 20% from "General Knowledge" to "Cultural and Pop Culture Knowledge" to "Linguistic". We can observe similar trends with the results split by the respective language dataset, as in Table 8.

I.2 Performance with *Connections* Games

One can note in Table 2 how performance greatly differs between the custom groupings-sourced games and the *Connections* groupings-sourced games. However, it is important to note the difficulty paradigm present within *Connections* – each game features groupings of four different difficulties, represented by the assigned color of a grouping, with difficulty said to be scaled from easiest to hardest in the order of yellow, green, blue, and purple. The NYT SEQ. and NYT SHUF. games utilized in Table 2 utilize groupings of all difficulties.

Models	F1 Score			
Color Difficulty Backgrounds	Yellow nyt shuf.	Green nyt shuf.	Blue nyt shuf.	Purple nyt shuf.
GPT-3.5-Turbo	**0.671**	0.638	0.575	0.523
GPT-4	**0.904**	0.876	0.805	0.702
	% FastText Topic Achieved (threshold = 0.3)			
GPT-3.5-Turbo	**0.365**	0.303	0.337	0.259
GPT-4	0.508	**0.525**	0.507	0.365

Table 9: The averaged results on *Connections* groupings for closed-source models across each 4-group, 4-word nyt shuf. game used during experimentation. Results are divided by the *Connections* color difficulty associated with a given grouping. The color difficulties are ordered yellow, green, blue, purple, from easiest to hardest.

When stratifying performance in NYT SHUF. (chosen to act as a better comparison to the randomly-generated custom groupings-sourced games) by the difficulty color, as seen in Table 9, we can see clear trends in F1 score relative to the color difficulty. This aligns with the listed difficulty scaling of these colors, with F1 score decreasing from yellow, to green, to blue, to purple. This reaffirms the listed difficulties of these colors in the *Connections* games, with difficulty (represented by a decrease in performance) following the prescribed pattern.

We can utilize performance on our EN games to compare our efforts in grouping creation to the groupings seen in *Connections*. By comparing performance in EN from Table 2 to performance stratified by difficulty color in Table 9, we can see the closest similarity to yellow groupings from *Connections*. This aligns with our intention to create an easier, more discriminative benchmark, as discussed in Section 2.

I.3 Evaluating True and Predicted Groups as Clusters with Silhouette Score

Additional insight can be gained by evaluating the true and predicted groupings of a game as if they were clusters. By using unsupervised cluster metrics, such as silhouette score, we can assess the inter-word similarity of a true/guessed grouping relative to the other groupings of the same classification in the same game. This allows us to then see comparisons in trends of silhouette scores for true and predicted groupings, offering a perspective into a possible basis for model group formation, and how that compares to actual groupings in a game. A silhouette score of 0 would mean that word groupings are just as dissimilar within-group as they are between-group, meaning that groupings are not solely composed by semantically similar words, as represented by their word embeddings.

For each scored game, we evaluate the silhouette score of the set of true groupings and the set of guessed groupings. We compile this by ranges of silhouette scores, reporting the percentage of total games observed to have a silhouette score within a given value bucket for each classification of grouping. By comparing these distributions, we can see if models tend to reason along the intended abstract line of reasoning for a given game (i.e. grouping words based on abstract connection, represented by 0-centered silhouette scores), rather than purely by word similarity (represented by silhouette scores at the ends of the metric range). In order to investigate potential differences in this aspect relative to performance, pointing towards different group formation paradigms leading to increased or decreased performance, we stratify these results by the F1 score achieved in a game.

Due to the overlapping distributions for the varying classifications of groupings across the evaluated models, as seen in Figures 4 and 5, we can infer that models employ reasoning patterns reminiscent of the intended abstract line of reasoning for WGG games. This is further emphasized by these distributions trending away from the extremes of the silhouette score ranges, being 0-centered, showing that a lack of inter-word similarity is common for both true and predicted groups, as predicted.

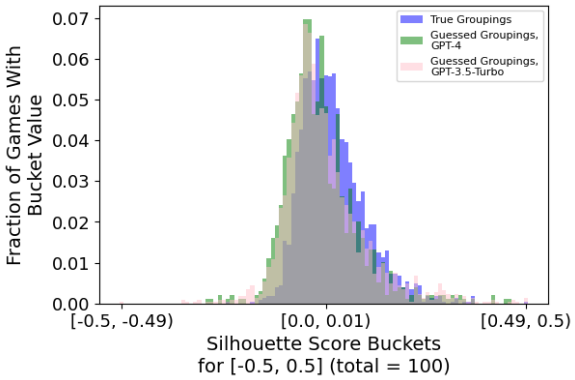

Figure 4: Depicting the distribution of total game composition of games with silhouette score values across 100 buckets of silhouette score ranges, for games where a model scored an F1 ≤ 0.5. This is evaluated for closed-source models, with game silhouette scores calculated using both the true groupings, and guessed groupings for a given game.

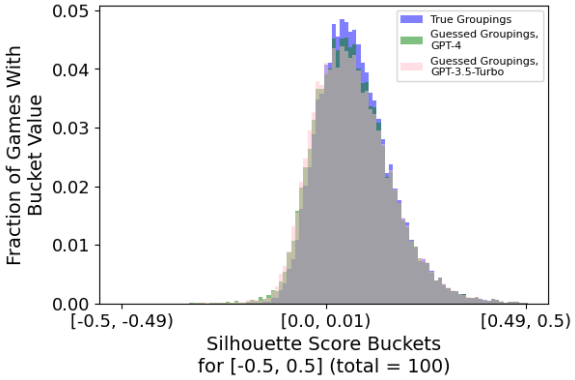

Figure 5: Depicting the distribution of total game composition of games with silhouette score values across 100 buckets of silhouette score ranges, for games where a model scored an F1 > 0.5. This is evaluated for closed-source models, with game silhouette scores calculated using both the true groupings, and guessed groupings for a given game.

I.4 Additional Raw Result Analysis

J Licenses and Intended Use

We utilized the OpenAI API to prompt the GPT-3.5-Turbo, GPT-4, and GPT-4o-mini models dur-

Models	Mean F1 Score Difference (Non-Culturally-Related - Culturally-Related)									
Datasets	es	es-en	zh	zh-en	hi	hi-en	ar	ar-en	en	Mean
GPT-3.5-Turbo	**0.128**	**0.130**	**0.054**	**0.041**	**0.026**	0.024	**0.054**	**0.036**	-0.047	**0.050**
GPT-4	0.098	0.092	0.013	0.018	0.001	**0.026**	0.036	0.022	**-0.017**	0.032
Llama3-8B	0.137	0.128	-0.005	0.025	0.048	-0.003	0.018	0.082	-0.033	0.044
Llama3.1-70B	0.100	0.118	0.031	0.028	0.017	0.019	0.078	0.044	-0.027	0.045
Mistral-7B	**0.182**	**0.198**	0.068	**0.061**	0.042	**0.036**	**0.080**	**0.109**	-0.005	**0.086**
Aya-8B	0.130	0.144	**0.070**	0.052	**0.065**	0.032	-0.011	0.095	**0.008**	0.065

Table 10: A table depicting the average difference in performance in groups labeled as non-culturally-related versus culturally-related in 4×4 games per dataset subset for open- and closed-source models. Note the strong pattern of performance improvement with non-culturally-related groupings, across nearly all languages.

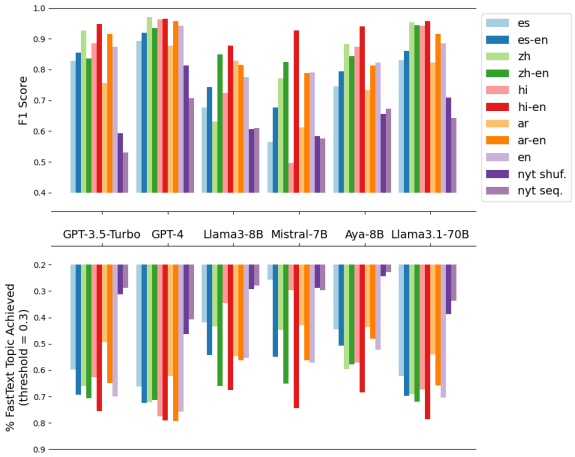

Figure 6: A figure depicting the averaged results by model across each 4-group, 4-word game used during experimentation.

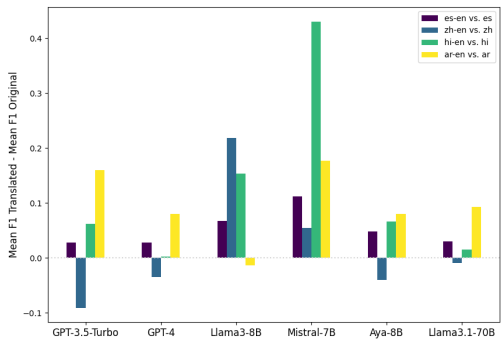

Figure 7: A figure depicting the difference in the average 4×4 game results for the translated and non-translated version of each dataset. Note a strong trend of improvement after translation into English, present across nearly all models in nearly all dataset translation pairs.

ing evaluation, and for help with reformatting other model responses. This usage complied with OpenAI's terms of use.

We utilized the Llama3-8B and Llama3.1-70B models during evaluation. Our usage complied with Meta's LLAMA 3.1 Community License Agreement.

We utilized the Mistral-7B model during evaluation. Our usage complied with the Apache 2.0 License under which this model was released.

We utilized the Aya-8B model during evaluation. Our usage complied with the Apache 2.0 License under which this model was released.

We utilized pretrained word embeddings from FastText, an embedding library released by Meta, for our various clustering analysis. Our usage complied with the MIT License under which these embeddings were released.

We additionally use packages gensim (Řehůřek and Sojka, 2010), NLTK (Bird et al., 2009), and jieba[9] for text tokenization. We use scikit-learn (Pedregosa et al., 2011) for the implementation of KMeans clustering, silhouette score calculation, and adjusted Rand index calculation. We use statsmodels (Seabold and Perktold, 2010) for the Randolph's Kappa calculation.

K Additional Difficulty Metric Analysis

K.1 Ranking Correlation for Difficulty Metric Results

We additionally calculate Spearman's rank correlation coefficient between the difficulty measurements and model performance results for different slices of the benchmark. The categorization of the dataset depends on the chosen difficulty measure, such as varying group sizes or numbers of groups in the game. For continuous difficulty measures, such as clustering metrics based on the Adjusted Rand Index or word overlap, we bin them into three

[9]https://github.com/fxsjy/jieba

Models	Average F1 Score				Average Correlation
Group Size	2	3	4		
GPT-3.5-Turbo	0.881	**0.890**	0.884		0.250
GPT-4	0.925	0.940	**0.949**		0.750
Group Count	2	3	4		
GPT-3.5-Turbo	**0.930**	0.883	0.842		1.000
GPT-4	**0.962**	0.940	0.912		1.000
Adjusted Rand Index Range	[-0.5, 0.0]	(0.0, 0.5]	(0.5, 1.0]		
GPT-3.5-Turbo	0.924	0.927	**0.951**		0.900
GPT-4	0.969	0.974	**0.990**		0.900
Candidate Proposal Overlap Range	[0.0, 0.75]	(0.75, 1.5]	(1.5, 2.25]	(2.25, 3.0]	
GPT-3.5-Turbo	**0.932**	0.873	0.813	0.831	-0.800
GPT-4	**0.977**	0.923	0.863	0.859	-0.800

Table 11: Average F1 results in games within buckets of specific difficulty metrics. Additionally, the ranking correlation of ranking average experimental setting performance by F1, and by the difficulty metric ranking, is provided.

or four categories and calculate the average model performance for each bin. Note that Spearman correlation focuses on ranking order rather than the correlation between the values, so although the scales of these metrics may differ, we argue that variations in their rankings should correspond with variations in model performance rankings. For 10 subsets of the benchmark, i.e., EN, ES, ES-EN, ZH, ZH-EN, HI, HI-EN, AR, AR-EN, and NYT-SHUF, we calculate the correlation between ranking average performance across experimental settings by the F1 score and by value of the difficulty metric (or expected ranking for group count/size). This was then averaged over all of the subsets to get their average correlation. The results of GPT-4 and GPT-3.5-Turbo are shown in Table 11.

Group Size & Group Count Contrary to our hypothesis for group size, as shown in Table 11, the observed patterns are inconsistent. This is reflected by the large average correlation score of 0.750 for GPT-4, contrasted with a very low score of 0.250 for GPT-3.5-Turbo, indicating that game performance does not consistently correlate with variations in word grouping size. In contrast, varying the number of groups reveals clear patterns in model performance. This is reflected in the correlation value, with both models boasting perfect correlations of 1.000, validating our hypothesis for group count.

Adjusted Rand Index We compare ranking experimental settings by model performance to ranking by the ARI scores. As shown in Table 11, both GPT-4 and GPT-3.5-Turbo exhibit near-perfect correlation coefficients of 0.900. We conclude that increased inter-word similarity within a group increases performance, confirming our initial hypothesis.

Word Overlap Assessment We follow the same experimental setup to calculate the average correlation, as done in the clustering experiment. As shown in Table 11, word overlap proves to be a significant predictor of game difficulty, with both models exhibiting a near-perfect correlation of -0.800, with performance trending as expected, decreasing as word overlap increases. Consequently, we conclude that word overlap is another indicator of game difficulty.

K.2 Difficulty-aware Results

After establishing effective difficulty measures, we report performance trends across different dataset backgrounds while controlling for game difficulty, as shown in Table 12. The datasets are categorized into four difficulty combinations based on two metrics: ARI score (where higher values indicate easier games) and word overlap score (where higher values indicate harder games). All experiments were conducted on 4-word, 4-group games across 10 benchmark subsets.

We can note that in general, more difficult groups lead to lower performance for both models. We see this clearly when comparing C3 and C4 (with an increase in word overlap), with a stark decline in performance across all models and all backgrounds, except for GPT-3.5-Turbo with zh and GPT-4 with ar. We additionally see the same general pattern when comparing C1 and C3 (with an increase in ARI), with an increase in performance across all models and nearly all backgrounds. It is important to note that the backgrounds where this pattern is strong in the opposite direction are HI and AR, and that the expected pattern is partially restored in the English-translated versions of these datasets. These backgrounds utilize non-Latin script text, where the pretrained embeddings may not be adequately capturing word meaning, and therefore, may impact the ARI calculation.

We also note that most games fall within the easiest difficulty groups, which is likely due to the random sampling approach. This method is less likely to generate NYT-SEQ-like games with higher word overlap between games, although it does not completely eliminate the possibility, as

Models	Average F1 Score									
Backgrounds	en	es	es-en	zh	zh-en	hi	hi-en	ar	ar-en	nyt-shuf
Metric/Range	C1: ARI [-1.0, 0.0], Word Overlap [0.0, 1.5]									
Game Count	32	31	10	1	13	164	7	27	38	37
GPT-3.5-Turbo	0.883	0.790	0.838	0.875	0.755	0.901	**0.982**	0.791	0.911	0.581
GPT-4	0.947	0.868	0.975	0.875	0.889	**0.976**	0.946	0.914	0.964	0.794
Metric/Range	C2: ARI [-1.0, 0.0], Word Overlap (1.5, 3.0]									
Game Count	0	0	5	0	2	4	3	0	4	5
GPT-3.5-Turbo	N/A	N/A	0.788	N/A	**0.938**	0.766	0.875	N/A	0.828	0.612
GPT-4	N/A	N/A	0.850	N/A	**1.000**	0.906	0.917	N/A	0.844	0.750
Metric/Range	C3: ARI (0.0, 1.0], Word Overlap [0.0, 1.5]									
Game Count	217	231	215	295	231	120	259	270	226	181
GPT-3.5-Turbo	0.875	0.835	0.864	0.927	0.851	0.872	**0.956**	0.755	0.921	0.618
GPT-4	0.948	0.902	0.927	0.971	0.938	0.966	**0.974**	0.873	0.959	0.840
Metric/Range	C4: ARI (0.0, 1.0], Word Overlap (1.5, 3.0]									
Game Count	24	21	27	4	26	7	10	2	15	44
GPT-3.5-Turbo	0.849	0.824	0.829	**0.938**	0.846	0.848	0.831	0.562	0.871	0.578
GPT-4	0.898	0.856	0.910	**0.969**	0.907	0.827	0.853	0.875	0.917	0.777
Metric/Range	C5: ARI (0.5, 1.0], Word Overlap [0.0, 0.75]									
Game Count	12	4	14	50	14	0	56	12	10	4
GPT-3.5-Turbo	**0.979**	0.906	0.911	0.956	0.848	N/A	0.971	0.807	0.942	0.688
GPT-4	0.990	0.969	0.964	0.978	0.920	N/A	0.990	0.922	**1.000**	**1.000**

Table 12: The averaged F1 score for closed-source models across 4-word, 4-group game. Results are divided by 2 ranges for the ARI and word overlap scores, creating 4 combinations (C1-4 in the first 4 rows), and by each dataset background, excluding NYT-SEQ. The number of games for each dataset background for each combination is provided as Game Count. C5 is a more refined difficulty metric combination, representing the easiest cases.

shown by some games still appearing in higher overlap ranges.

K.3 Devising An Integrated Difficulty Metric

Utilizing the correlation results from Appendix K, we can create an integrated difficulty metric for a GLOBALGROUP game by considering group count, ARI, and word overlap together. For this, we:

- Project each metric's value, considering the metric's range, onto a range of $[0, 1]$. For group count we project from the range $[2, 4]$, for ARI we project from the range $[-1, 1]$, and for word overlap we project from the range $[0, 3]$.

- Utilizing the previously calculated correlation coefficients, we perform a weighted summation of the projected values, with each value weighted by its metric's respective correlation coefficient in the direction of difficulty. For group count this would be 1.000 (as a higher group count is associated with more difficulty), for ARI this would be -0.900 (as a higher ARI is associated with less difficulty), and for word overlap this would be 0.800 (as a higher word overlap is associated with more difficulty).

This results in an integrated difficulty metric within the range of $[-0.9, 1.8]$. This is finally projected into the range $[0, 1]$ as the final score.

Performing an identical correlation analysis with the integrated difficulty metric as done for ARI and word overlap yields a perfect correlation of -1.000, meaning that as the integrated difficulty metric value of a game increases, F1 performance on that game decreases.