Appendices

A Additional Experiments

Task 1 – Grouping In addition to grouping clue words using token embeddings (discussed in the main paper §4), we also ran grouping the words by clustering on 'contextual' embeddings. We experimentally induce 'context' by joining the sixteen (16) word tokens (in a random order) into a single pseudo-sentence. The embeddings for each token were different based on the ordering of the tokens. We repeat the random ordering sixteen times and report the mean and variance of the results obtained in Table 6.

	WD ↓	FMS ↑	ARI ↑	AMI ↑	# Solved Walls	# Correct Groups
ELMo$_{\text{LARGE}}$	90.0 ± .3	23.6 ± .4	4.5 ± .5	5.6 ± .7	0 ± 0	19 ± 3
DistilBERT$_{\text{BASE}}$	88.4 ± .7	26.7 ± .3	8.3 ± .4	10.4 ± .5	0 ± 0	30 ± 4
BERT$_{\text{LARGE}}$	**87.2 ± .6**	**28.3 ± .5**	**10.4 ± .6**	**12.8 ± .7**	0 ± 0	**46 ± 5**
BERT$_{\text{BASE}}$	87.7 ± .5	28.0 ± .2	10.0 ± .3	12.4 ± .4	0 ± 0	39 ± 2
RoBERTa$_{\text{LARGE}}$	88.4 ± .5	25.9 ± .2	7.4 ± .3	9.3 ± .4	0 ± 0	30 ± 4
all-mpnet$_{\text{BASE}}$	87.6 ± .5	28.0 ± .3	10.0 ± .4	12.4 ± .5	0 ± 0	38 ± 3
E5$_{\text{LARGE}}$	87.7 ± .5	28.1 ± .3	10.2 ± .4	12.7 ± .5	0 ± 0	37 ± 4
E5$_{\text{BASE}}$	**87.2 ± .3**	28.2 ± .2	10.2 ± .3	12.5 ± .4	0 ± 0	**46 ± 5**
Human Performace	–	–	–	–	285 / 494	1405 / 1976

Table 6: Results of selected models on Task 1 (Grouping) using contextual embeddings. WD: Wasserstein Distance. FMS: Fowlkes Mallows Score. ARI: Adjusted Rand Index. NMI: Normalized Mutual Information. Mean ± standard deviation over 16 random seeds is shown. **Bold**: best scores.

Task 2 – Connections In addition to prompting based results on GPT-4 (discussed in §4), we ran experiments on additional LLMs like LLaMa [67] (7B, 13B) using pre-trained configuration weights obtained by permission from Meta AI. However, without additional fine-tuning on the specific task, these LLMs were unable to solve the task in a meaningful manner. To elucidate, LLaMa generated a bunch of hallucinated words with unequal group sizes. We omit these unintelligible results for brevity.

B Additional Figures

In this section, we provide additional t-SNE projections of embeddings from various methods used.

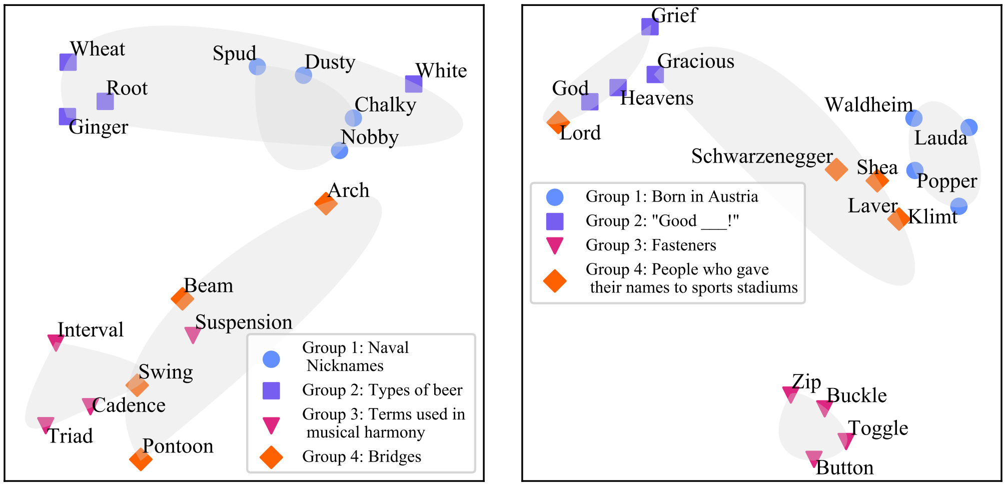

Figure 7: Solved wall for Task 1 (Grouping) using GloVe. **Left**: (`wall_id="7ed3"`), the embedding model erroneously associated the clue "*Suspension*" with the connection "*Bridges*"; however, this association is an example of a red herring. "*Suspension*" is "*a term used in musical harmony*" in this context. **Right**: (`wall_id="5e3c"`), shows that clue "*Lord*" is close to "*God, Heavens, and Grief*" in the embedding space, which matches the "*Good ___!*" connection. However, this is another example of a red herring as, in this context, "*Lord*" refers to "*Lord's cricket Ground*", a cricket stadium named after "*Thomas Lord*".

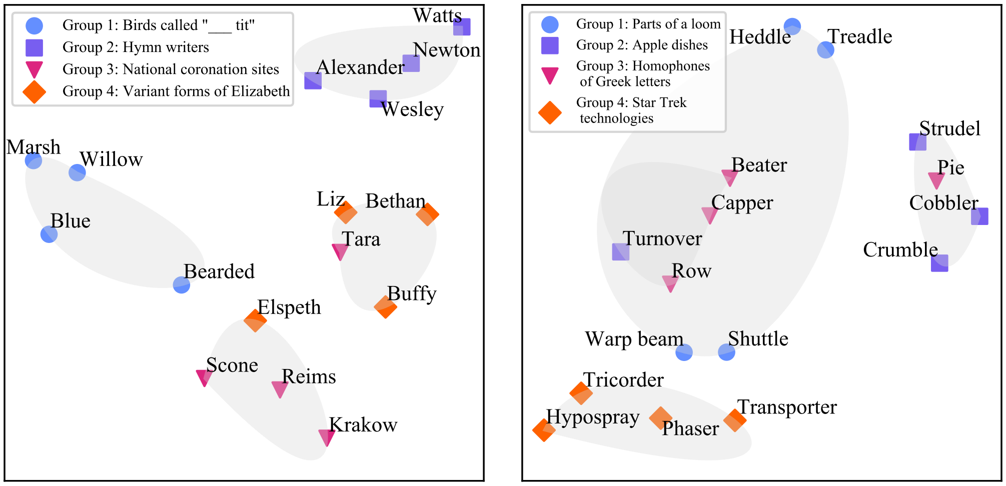

Figure 8: Solved wall for Task 1 (Grouping) using FastText (Crawl). **Left**: (`wall_id="d5e6"`), the embedding model erroneously associated the clue "*Tara*" other girls' names; but here, "*Tara*" is short for "*Hill of Tara*" and belongs to the "*national coronation sites*" group. **Right**: (`wall_id="4c22"`), shows that clue "*Pie*" associated with the connection "*Apple*". Even though it is acceptable in general context, here it represents a homophone for the Greek letter "π".

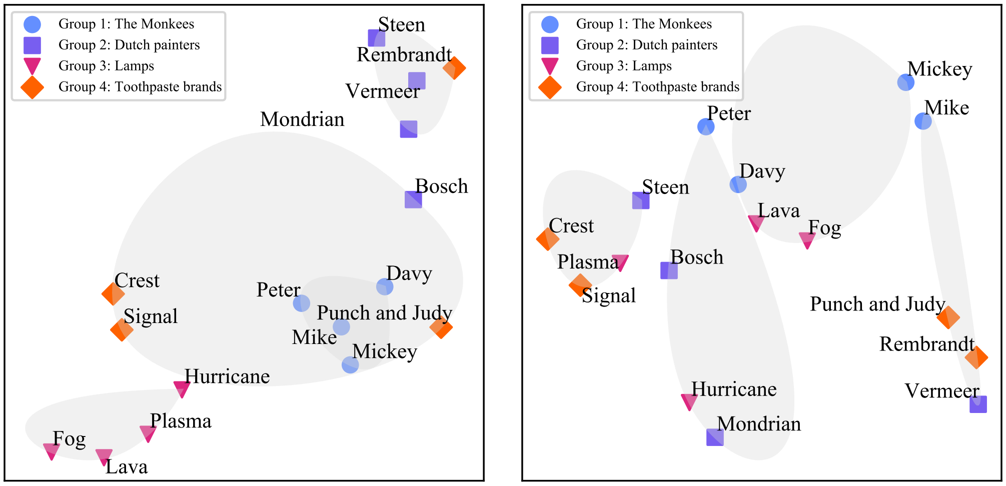

Figure 9: Solved wall (`wall_id="2d8f"`) for Task 1 (Grouping) using BERT$_{\text{LARGE}}$ with both static and contextual embeddings. **Left**: contextual embedding solved 3/4 groups. Here the clue “*Rambrandt*” is placed near other Dutch painters. The correct grouping for this clue in this wall is “*Toothpaste Brands*”. **Right**: static embedding solved 0/4 groups.

C Datasheet

The following section provides answers to questions listed in datasheets for datasets.

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The OCW dataset is created to be an analogical proxy for the Remote Associates Test (RAT) [45] from cognitive neuroscience in evaluating LLMs for human-imitative *creative problem-solving*. The presented clues have heterogeneous connections with open-domain knowledge retrieval and contain red herrings or misleading stimuli by design. The two tasks entails *grouping* sixteen (16) jumbled up clue words into associated groups, and naming the right *connection* for each group. To the best of our knowledge, there are no existing tasks for evaluating LLMs for human-like creative problem solving in existing, and concurrent benchmarks including the BIG-Bench, HELM, Global-Bench. Thus, this dataset and tasks are valuable additions for overall LLM evaluation and measuring progress towards human-imitative AI.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset has been collectively curated by the authors of this paper.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
This work was supported by Natural Sciences and Engineering Research Council of Canada (NSERC).

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Each instance contains a connecting wall puzzle and its solution from the popular quiz show Only Connect.

How many instances are there in total (of each type, if appropriate)?
618 wall puzzles (instances of the dataset), for a total of 2,472 groups, and 9,888 clues.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?
The dataset has been curated from the first fifteen seasons of the "Only Connect" show, which accounts for approximately 81% of the total seasons. The latest season, Season 18, was concluded in March 2023.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance contains a connecting wall puzzle and its clues and solution. All instances are in English and provided as text strings in JSON format.

Is there a label or target associated with each instance? If so, please provide a description.
Yes. The labels for Task 1 are the solved walls, and for Task 2 the ground-truth connections.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
N/A

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
Each wall is given a unique ID. Clues and solutions associated with each wall belong to the same JSON object as that wall.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
We randomly split the dataset into the training/dev/test set according to a proportion of 1:1:8. The primary goal of our dataset is to evaluate the zero- and few-shot creative problem-solving abilities of Large Language Models; as such, we elect to set the size of the test set to be much greater than train or validation sets.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
The dataset has undergone a thorough review and is subjected to both automated and manual checks as part of a strict quality control protocol.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?
The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
N/A.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
N/A.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
The dataset does not have individual-specific information.

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The wall puzzles were scraped from the fan website ocdb.cc as well as manually watching the episodes. Human performance results were manually curated from the episodes. all data verified

through manual watching of episodes.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
We utilized python's BeautifulSoup library to scrape only connect fan websites. all episodes were watched manually for human performance collection, and the same procedure validated the data collection.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs
Experiments were run using NVIDIA GeForce RTX 2080 Ti GPU system.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
The authors of this paper.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
N/A.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
The dataset does not have individual-specific information.

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
N/A.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
N/A.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
N/A.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for?
Evaluation of Large Language Models for creative problem-solving as well as Artificial General Intelligence tasks.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
N/A.

Are there tasks for which the dataset should not be used? If so, please provide a description.
We caution regarding unethical reuse of the dataset, specifically for the purpose of training future reasoning engines for unethical use cases.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The code and link to the dataset is available at `https://github.com/TaatiTeam/OCW`

When will the dataset be distributed?
Now.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset is released under MIT License.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The dataset is hosted on University of Toronto Computer Science Department servers and will be maintained by the authors of this paper.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The maintainers can be contacted via email: saeid.alavi@mail.utoronto.ca, raeidsaqur@cs.toronto.edu, john.giorgi@mail.utoronto.ca, mozhgans@stanford.edu, babak.taati@uhn.ca.

Is there an erratum? If so, please provide a link or other access point.
No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
The authors plan to continue updating the dataset, including but not limited to scaling the dataset to include more seasons, providing new test/dev sets, and organizing shared tasks with the dataset. The updates will be yearly and communicated to users through public shared tasks.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
Yes, the authors are committed to maintaining and updating the older versions of the dataset.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Any potential contributors are welcome to expand the dataset to larger size through contacting the authors of the paper.

D Effects of Red-Herrings: Additional Experiments, Analysis and Results

D.1 Additional Datasets

Both of the additional datasets described in this section for ablation experiments have been made available via our code repository.

D.1.1 OCW-Randomized Dataset

This test dataset generates a version of the test set where red herrings are removed or largely reduced in frequency. This is achieved by rebuilding every wall using a randomly selected group from different walls. We only applied the process to the (original OCW) test set, the train and validation sets are left untouched.

Method For each wall in the existing test set, we leave the first group untouched, and sample three new groups, each from a different wall, such that none of the groups share a word in common. The connections for each group are unmodified. The result is a new version of the test set where every wall is composed of 4 random groups from 4 different walls.

D.1.2 OCW-WordNet Dataset

WordNet [46, 20] is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. We use the hypernym/hyponym (or superlative/subordinative) hierarchical lexical structure aggregated in WordNet to generate an easy test set to further analyze the effects of red-herring in OCW.

Method We use the existing words in a wall to select synonyms from the word's synsets. We only consider synsets that have at least five synonymous lexical names, then randomly sample four words. The original test set word and its definition (`ss.definition()`) subsequently becomes the connection phrase for the group. Four groups were generated for each wall, and the easy wall generation process is repeated for all the total number of walls (494) in the original test data set.

For the group connections, we concatenate the superlative parent word with a synset definition giving a description of the word. This allows for an ideal semantic similarity score to be calculated using BERTScore. For a few cases (approx. 70/494 walls in test set), number of generated groups per wall is less than four, due to the unavailability of direct synonyms from word synsets. In those edge cases, we generate and append groups using common hypernym words like animal, mammal, furniture etc. to ensure a wall is valid with four groups.

A sample generated easy group is shown below, where we prefix the group_id from original OCW dataset with 'easy' to aid with mapping or identification.

```
{
    ...
    "group_3": {
    "group_id": "easy_691a_3",
    "gt_words": ["gibe","shaft","jibe","barb"],
    "gt_connection": "Shaft: an aggressive remark directed at a person
    like a missile and intended to have a telling effect"
    ...
}
```

Further, we generate easy train and validation sets mimicking the original dataset, package and release these three additional easy sets, as **OCW-WordNet** as added contributions.

D.2 Results of Ablation Experiments

D.2.1 PLMs: Performance on Task 1 (Grouping)

We perform and present the results using 'static' embeddings due to the noted superior results and the word order related deficiency already shown with using contextual embeddings pertinent to our task setup.

	WD ↓	FMS ↑	ARI ↑	AMI ↑	# Solved Walls	# Correct Groups
			Classic Word Embeddings			
GloVe	76.8 ± .7	39.2 ± .3	24.0 ± .4	27.7 ± .4	7 ± 1	213 ± 8
FastText (Crawl)	76.1 ± .5	40.5 ± .3	25.0 ± .6	28.6 ± .7	**13 ± 1**	236 ± 7
FastText (News)	79.3 ± .5	36.8 ± .3	21.0 ± .3	24.5 ± .4	5 ± 1	176 ± 6
			Pre-trained Language Models (PLMs)			
$\text{ELMo}_{\text{LARGE}}$	80.9 ± .4	35.2 ± .3	18.9 ± .3	22.2 ± .4	3 ± 1	154 ± 6
$\text{DistilBERT}_{\text{BASE}}$	82.3 ± .6	34.2 ± .4	17.7 ± .5	21.1 ± .5	1 ± 1	124 ± 8
$\text{BERT}_{\text{LARGE}}$	86.2 ± .4	29.2 ± .3	11.5 ± .3	14.2 ± .4	0 ± 0	66 ± 4
$\text{BERT}_{\text{BASE}}$	87.5 ± .4	27.7 ± .3	9.6 ± .6	11.8 ± .5	0 ± 0	48 ± 4
$\text{RoBERTa}_{\text{LARGE}}$	86.7 ± .5	28.6 ± .2	10.8 ± .3	13.4 ± .3	1 ± 0	56 ± 4
			Sentence Transformers			
$\text{all-mpnet}_{\text{BASE}}$	81.4 ± .4	35.1 ± .4	18.9 ± .5	22.0 ± .6	8 ± 1	154 ± 7
E5_{LARGE}	76.0 ± .5	40.7 ± .3	25.9 ± .4	29.7 ± .4	8 ± 1	230 ± 5
E5_{BASE}	**75.1 ± .8**	**41.8 ± .3**	**27.2 ± .3**	**31.1 ± .3**	8 ± 1	**249 ± 8**
Human Performance	–	–	–	–	–	–

Table 7: Results of **OCW-Randomized** using static embeddings. WD: Wasserstein Distance. FMS: Fowlkes Mallows Score. ARI: Adjusted Rand Index. NMI: Normalized Mutual Information. Mean ± standard deviation over 16 random seeds is shown. **Bold**: best scores.

	WD ↓	FMS ↑	ARI ↑	AMI ↑	# Solved Walls	# Correct Groups
			Classic Word Embeddings			
GloVe	43.0 ± 1.0	66.1 ± .4	57.4 ± .5	60.9 ± .5	118 ± 3	886 ± 1
FastText (Crawl)	30.6 ± 1.0	75.8 ± .6	69.6 ± .7	72.4 ± .7	195 ± 6	1173 ± 18
FastText (News)	44.9 ± 1.2	64.9 ± .5	55.9 ± .6	59.5 ± .6	105 ± 3	844 ± 12
			Pre-trained Language Models (PLMs)			
$\text{ELMo}_{\text{LARGE}}$	52.5 ± 1.1	58.9 ± .3	48.2 ± .4	52.5 ± .4	67 ± 3	682 ± 9
$\text{DistilBERT}_{\text{BASE}}$	45.5 ± 1.0	64.1 ± .4	55.0 ± .5	58.7 ± .5	105 ± 3	835 ± 13
$\text{BERT}_{\text{LARGE}}$	76.9 ± 1.0	38.9 ± .2	23.4 ± .3	27.5 ± .3	7 ± 0	197 ± 6
$\text{BERT}_{\text{BASE}}$	73.0 ± 1.3	42.5 ± .5	27.9 ± .6	32.5 ± .6	8 ± 2	268 ± 12
$\text{RoBERTa}_{\text{LARGE}}$	57.4 ± 1.3	54.8 ± .3	43.3 ± .3	47.5 ± .3	48 ± 2	573 ± 8
			Sentence Transformers			
$\text{all-mpnet}_{\text{BASE}}$	**22.6 ± .7**	**81.9 ± .4**	**77.1 ± .5**	**79.4 ± .4**	**256 ± 4**	**1365 ± 12**
E5_{LARGE}	23.6 ± .8	80.9 ± .4	75.9 ± .5	78.3 ± .4	250 ± 4	1347 ± 12
E5_{BASE}	26.9 ± .9	78.0 ± .4	72.3 ± .5	75.0 ± .5	224 ± 4	1259 ± 10
Human Performance	–	–	–	–	–	–

Table 8: Results of **OCW-WordNet** using static embeddings. WD: Wasserstein Distance. FMS: Fowlkes Mallows Score. ARI: Adjusted Rand Index. NMI: Normalized Mutual Information. Mean ± standard deviation over 16 random seeds is shown. **Bold**: best scores.

D.2.2 LLMs: Performance on Task 1 (Grouping) using GPT3.5/4

Here we present the results of repeating Task 1 (grouping) on the ablation datasets OCW-Randomized (D.1.1) and OCW-Wordnet (D.1.2) to analyze the effects of red-herrings in walls on LLM performance.

	# In-context Examples	WD ↓	FMS ↑	ARI ↑	AMI ↑	# Solved Walls	# Correct Groups
GPT-3.5-turbo	0-shot	74.3	40.4	26.4	29.8	5	274
	1-shot	72.0	43.1	29.0	32.3	12	315
	3-shot	72.7	43.4	29.4	32.9	10	306
	5-shot	70.7	44.6	30.9	34.4	16	337
	10-shot	70.5	43.8	30.0	33.5	17	333
GPT-4	0-shot	58.2	56.2	45.4	48.8	59	595
	1-shot	55.1	**58.0**	**47.5**	**51.0**	57	644
	3-shot	55.0	57.5	46.9	50.3	62	649
	5-shot	**54.1**	**58.0**	**47.5**	50.9	**68**	**655**
	10-shot	56.6	56.1	45.1	48.5	55	614
Human Performance		–	–	–	–	–	–

Table 9: Results of **OCW-Randomized** using Large Language Models. WD: Wasserstein Distance. FMS: Fowlkes Mallows Score. ARI: Adjusted Rand Index. NMI: Normalized Mutual Information. **Bold**: best scores.

The results adhere to the expected results of superior performance with the dilution/removal of red-herrings from the walls.

	# In-context Examples	WD ↓	FMS ↑	ARI ↑	AMI ↑	# Solved Walls	# Correct Groups
GPT-3.5-turbo	0-shot	15.9	86.3	83.4	84.9	337	1522
	1-shot	24.8	76.4	74.4	75.4	320	1400
	3-shot	8.65	92.7	91.2	91.8	415	1748
	5-shot	8.09	94.0	92.4	93.1	415	1759
	10-shot	6.55	95.3	94.0	94.7	428	1800
GPT-4	0-shot	**1.51**	**98.5**	**98.0**	**98.2**	**471**	**1926**
	1-shot	19.2	87.9	84.3	83.7	304	1581
	3-shot	21.5	86.6	82.5	81.8	279	1537
	5-shot	19.1	88.1	84.5	83.8	298	1584
	10-shot	11.2	92.9	90.7	90.4	378	1742
Human Performance		–	–	–	–	–	–

Table 10: Results of **OCW-WordNet** using Large Language Models. WD: Wasserstein Distance. FMS: Fowlkes Mallows Score. ARI: Adjusted Rand Index. NMI: Normalized Mutual Information. **Bold**: best scores.