CRUGS: A Language Dataset for Compositional and Relational Understanding Using Geometric Shapes

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Kevin Ivey

Spring, 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Yanjun Qi, Department of Computer Science and Affiliated Faculty of Center for Public Health Genomics

CRUGS: A Language Dataset for Compositional and Relational Understanding Using Geometric Shapes

Kevin Ivey
Department of Computer Science
University of Virginia
Charlottesville, Virginia, USA
ki4km@virginia.edu

Abstract

New advances in large language models are rapid, with a new model touting better performance being released seemingly every week. As such, the performance of models needs to be quantitatively compared across standardized datasets. We present CRUGS, a language dataset for Compositional and Relational Understanding using Geometric Shapes. Our dataset is simple in nature, relying on synthetically generated canvases with ground-truth attributes and answers. CRUGS utilizes a series of increasingly challenging tasks that require no human experts and are flexible to the method of prompting and the number of shapes in the canvas. We evaluated six large language models on CRUGS and found that despite the simplicity of the tasks, gaps exist in the performance of the models.

1 Introduction

Recent large language models (LLMs) have shown remarkable capabilities. While these models are trained on simple next token prediction for a given piece of text, this skill appears to generalize across a number of different tasks, such as solving logic puzzles and writing code. In fact, any task that can be framed as completion of a sequence of text can be fed into a LLM. As LLMs grow in size and in training data, their capabilities appear to become better. However, such *qualitative* observations of performance need to be verified *quantitatively* in order to justify their use for critical real-world tasks.

Beyond the problem of evaluating the performance of a LLM is trying to evaluate the methods that a LLM uses to complete a particular task. Research has suggested that despite being trained on next token prediction, LLMs go beyond memorizing surface level statistics and instead exhibit world models that internally represent the task (Li et al., 2023). However, such research relies on adhoc data, not a standardized benchmark.

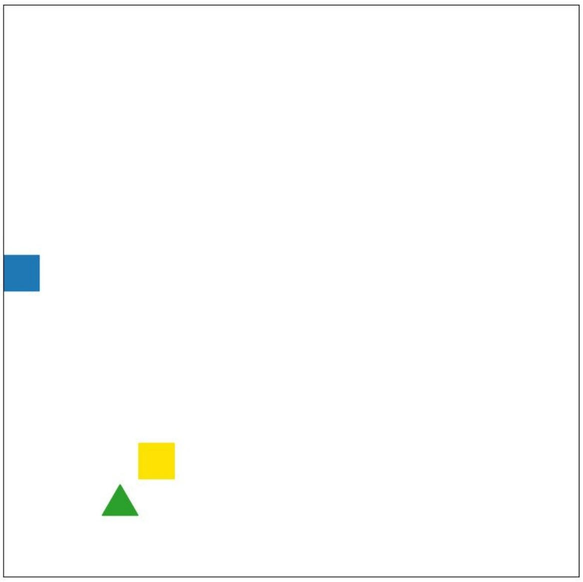
In this paper, we propose CRUGS, a dataset to evaluate LLMs on their Compositional and Relational Understanding using Geometric Shapes. CRUGS is a simple dataset consisting of geometric shapes in a 2D canvas. We developed a series of increasingly difficult compositional tasks that asked questions about the state of the canvas along with natural language descriptions of the canvas that encode the attributes of the geometric shapes and their positions.

While each canvas must be represented in natural language for the model, they can be visualized on the 2D plane. Hence, interpretability methods that seek to reconstruct internal representations of the model can be matched against the ground-truth state.

We focused on a synthetic self-contained dataset. This allows for a large amount of flexibility during data generation. Specifically, our dataset is configured to be extensible to new tasks without the need for human experts, supports flexible prompting methods, and contains ground-truth answers in both a textual and visual format. In summary, this paper makes the following contributions:

- The development of CRUGS, a self-containted geometric dataset of shapes in a 2-D canvas. Each canvas has descriptions of the canvas in natural language and a series of increasingly challenging compositional tasks that ask about the state of the canvas. This dataset is easily extensible for additional tasks and additional methods of prompting.

- An evaluation of past and current state-of-the-art LLMs on CRUGS. Six different LLMs are evaluated via our dataset. Of particular note are models that are finetuned from another model on instruction-following data.



Transitive Description: There are 3 shapes in a canvas. There is a large yellow square in the canvas. Below the large yellow square is a large green triangle. Above the large yellow square is a large blue square.
Transitive Task Question: Where is the large blue square relative to the large green triangle?
Transitive Ground-Truth Answer: Above

Partial Description: There are 3 shapes in a canvas. There is a large green triangle in the canvas. A large yellow square is to the above right of this large green triangle. A large blue square is to the above left of this large green triangle. There is a large yellow square in the canvas. A large blue square is to the above left of this large yellow square. There is a large blue square in the canvas.
Count Task Question: Is there a blue square?
Count Ground-Truth Answer: Yes

Figure 1: An example of a canvas with three shapes. Both the transitive task and the count task are shown with their respective descriptions of the canvas.

2 Related Works

Interpretability In recent years, research has been conducted to investigate whether LLMs memorize surface level statistics or if they utilize more complex internal representations of the world state.

In a previous study, GPT was used to predict legal moves in the game Othello (Li et al., 2023). Through the use of latent saliency maps and probing techniques, the authors concluded that the model encodes a nonlinear internal representation of the board state despite having no knowledge of the game rules. While small in scope, this points towards a more complex form of representation present in LLMs.

Evaluation Benchmarks A number of datasets have been developed to evaluate the capabilities of LLMs.

Google's Beyond the Imitation Game benchmark (BIG Bench) presents a total of 204 tasks across a diverse set of domains and reasoning types (Srivastava et al., 2022). Each task comes with a human-evaluator and a human expert baseline. The dataset is designed to be challenging for current state-of-the-art LLMs in order to measure future capabilities.

The Holistic Evaluation of Language Models (HELM) provides a framework for evaluating LLMs through the taxonomy of scenarios and metrics (Liang et al., 2022). A scenario is an instance where a LLM can be applied (e.g. question answering or sentiment analysis) while metrics describe what we want the LLM to do (e.g. accuracy or fairness). Additionally, HELM uses the idea of adaptations which allows a LLM to run on a new instances of a scenario (e.g. prompting or finetuning).

While BIG Bench provides a large number of tasks across domains and HELM provides a framework for benchmarks, neither benchmark meets all three aspects of (1) extensibility for new tasks, (2) scalability without the need for human experts, and (3) flexibility for new methods of prompting.

Synthetic Diagnostic Datasets In computer vision, the Compositional Language and Elementary Visual Reasoning diagnostics dataset (CLEVR) is a synthetic dataset used to measure the capabilities of visual question answering models (Johnson et al., 2017). CLEVR generates a 3D scene consisting of simple shapes along with questions about the scene to measure the capabilities of a model. Due to the synthetic nature of the dataset, the ground truth is known for every shape in a scene. Our work is inspired by the development of CLEVR.

3 The CRUGS Dataset

3.1 Overview

CRUGS provides a synthetically generated dataset with ground truth answers. The data consists of n canvases containing s shapes. Each shape is a square, circle, or triangle, can be small or large in size, and can be blue, red, orange, or green in color. In total, this allows for 24 unique shapes. By construction, each canvas defines the ground truth, as the position, shape type, size, and color is recorded for every shape in the canvas. From the canvas, multiple natural language descriptions are generated that encodes relevant information about the shapes and their positions. For every canvas, question-answer pairs are generated according to a set of tasks. See Figure 1 for an example of a canvas and tasks.

Note that simple, common geometric shapes along with common colors, and often used words for sizes are used so the dataset focuses on reasoning skills, not on the prior likelihoods of each shape.

3.2 Canvas Generation

To generate a canvas, the shape type, color, and size are randomly chosen with the constraint that no two shapes are the exact same. Additionally, the coordinates of the shape are chosen to ensure that the boundary of the shape is within the 64 by 64 canvas and that no two shapes overlap or intersect.

3.3 Canvas Descriptions

To provide the LLM with the state of the canvas, a natural language representation must be provided. We identified two main methods of encoding the attributes and positions of the shapes in the canvas in natural language: relative and coordinate descriptions.

Relative Descriptions In a relative description, the position of each shape in the canvas is provided relative other shapes. To avoid ambiguity, we provide either the position of a shape relative to every other shape in the canvas or the position of a shape relative to only one other shape. In the latter case, we reduce the description to one-dimension (i.e. horizontal or vertical) and only give the relative position of two shapes if there is no other shape in between them.

Coordinate Descriptions In a coordinate description, the x-y position of the center of the shape is given along with its numerical size (i.e. radius or sidelength).

3.4 Tasks

There are six compositional tasks that range from simple tasks, such as questions of existence, to more complex tasks, such as modelling a sequence of swaps. Each task is self contained and generates question-answer pairs given a canvas. Additionally, some tasks may define an additional canvas description or additional information to add to a canvas description. Note also that the answer distribution of the tasks is fixed to create a balanced dataset.

Existence Task The existence tasks asks whether an object with one or more attributes exists in the canvas. Note that descriptions do not contain distracting phrases, so this task could be completely solved by an exact string match. The expected answers are ”Yes” or ”No.”

Count Task The count task asks how many objects with one or more attributes exist in the canvas (see Figure 1). Counting is a precursor to more complicated arithmetic tasks and measures a basic capability. Additionally, some reasoning is required as the same shape may be referred to multiple times in the description.

Transitivity Task The transitivity task presents the relative positions of the shapes in a transitive manner (see Figure 1) and then asks for the relative position of two shapes that is not directly described. This task measures the ability for a model to understand the transitive property.

Coordinate Task The coordinate task presents the coordinate positions of the objects and their numerical sizes (i.e. radius or side length) and asks for the relative position of two shapes. In general, this task requires the understanding of integer inequalities, but certain cases require the understanding of basic geometric formulas to determine the boundary points of shapes.

Existence Tracking Task The existence tracking task presents the initial set of shapes in the canvas and then presents a sequence of shapes that are added to or removed from the canvas. This task requires understanding of a dynamic situation involving mental bookkeeping or internal model.

Shuffle Tracking Task The shuffle tracking tasks presents the initial positions of the shapes in a single line. A series of swaps of shapes is then given

Task Name	GPT-J	Dolly	LLaMA	Alpaca	Flan-T5	ChatGPT
Existence	0.4730	0.5560	0.5570	0.5840	**1.0000**	0.9980
	0.6700	0.6770	0.8560	0.6120	**1.0000**	0.9910
Count	0.0620	0.0170	0.0490	0.0120	**0.4060**	0.2710
	0.4990	0.5420	0.4960	0.2570	**0.5840**	0.4360
Transitivity	0.6730	0.6450	0.3190	0.4595	**0.9830**	0.8755
	0.5115	0.4840	0.4410	0.2360	**0.9210**	0.3790
Coordinate	0.0020	0.0035	0.0045	0.1655	0.0010	**0.4470**
	0.5935	0.5275	0.2215	0.2150	0.3245	0.4610
Existence Tracking	0.5010	0.5000	0.6520	0.8160	0.7310	**0.9740**
	0.4770	0.5210	0.7100	0.6660	0.8260	**0.9820**
Shuffle Tracking	0.3270	0.3250	0.3100	0.3100	**0.3640**	0.3400
	0.9390	0.9130	0.9790	**0.9810**	0.7390	0.8950

Table 1: Task level performance for 3 shapes in a canvas. For each task, the top row reports the 0-shot performance and the bottom row reports the 3-shot performance. Results in bold are the best in the respective task.

and asks for the shape present at a certain position. This task requires modelling a sequence of modifications to an initial state.

3.5 Task Scoring

Each task defines its own method to score a textual answer given the ground truth answer. However, each score lies in the range $[0, 1]$ where a score of 0 corresponds to a completely incorrect answer and a score of 1 corresponds to a correct answer. The scoring methods award points for a correct answer and subtract points for incorrect or hallucinated information.

3.6 Few-Shot Prompting

Each task generates three question-answer pairs for the given canvas. This allows for easy implementation of few-shot prompting. Note that few-shot prompting relies on one canvas description and that the few-shot question-answer pairs are fixed for a given canvas. Both zero and few-shot prompting ask the same question for the model to answer.

4 Experiments

Data Summary The experiments were conducted with the number of shapes s in the canvas as 3 and 5. For each value of s, $1,000$ examples were used. Additionally, for each value of s we tested both the zero-shot setting and a few-shot setting with the number of shots fixed at 3. Note that the dataset was generated prior to evaluation, so all canvases and shots are fixed across every iteration.

Models Six different models were evaluated: GPT-J, Dolly-1.0, Flan-T5, LLaMA, Alpaca, and ChatGPT (Wang and Komatsuzaki, 2021; Conover et al., 2023; Chung et al., 2022; Touvron et al., 2023; Taori et al., 2023; OpenAI, 2022). Notably, Dolly-1.0 and Alpaca are finetuned with instruction-following data from GPT-J and LLaMA respectively. All models were configured to return a maximum of 50 tokens, which is significantly longer than any expected answer, the `top_p` parameter was set to 0.9, and early stopping and sampling were disabled. Note that for ChatGPT, the `gpt-3.5-turbo` model was used.

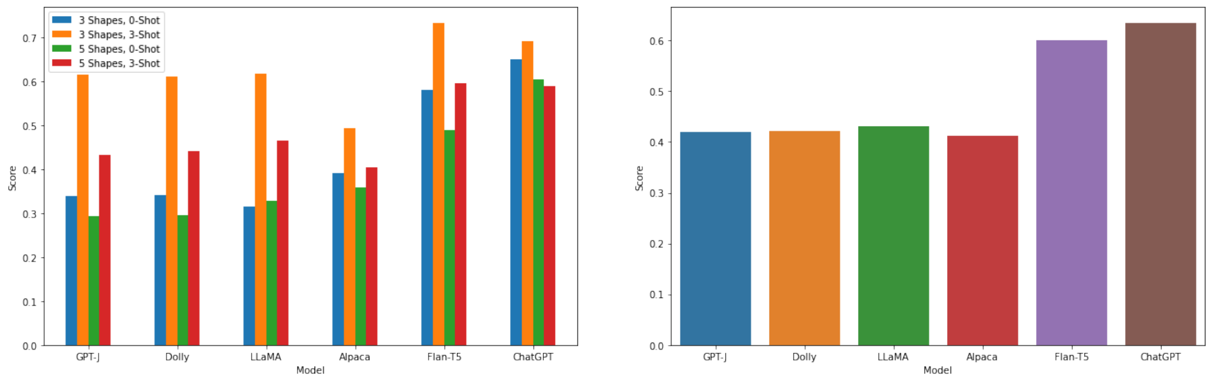
Metrics Models were evaluated by their average score on a task. For each task, scores lie in the range $[0, 1]$, where 1 corresponds to correctly answering the task and 0 corresponds to a completely incorrect answer.

5 Evaluation

Task Level Performance In Table 1 we present the results of each model on each task with three shapes in the canvas.

For the existence task, only Flan-T5 and ChatGPT performed significantly above the 0.5 baseline of random chance when given zero-shot prompts. From empirical observations, the models typically gave an incorrect answer of ”Yes” when the ground truth answer is ”No.” This may be a result of the word ”Yes” having a higher prior likelihood than the word ”No.” With three-shot prompts, the performance of the majority of models increased to at least outperform random chance.

Performance on the count task was poor. With zero-shot prompting, only Flan-T5 outperformed



(a) Performance of each model on the tasks aggregated via the number of shapes in the canvas and the number of shots.

(b) The overall average performance of each model. The average is the standard arithmetic mean.

Figure 2

the random chance of 0.33. With three-shot prompting, performance improved, but not to a much higher standard. When answering incorrectly, the models often counted the number of times the shape in question appeared in the prompt, not the number of times it appeared in the canvas. While the scores could then easily be influenced by the method of prompting, this task can easily be solved by a human with the current method of prompting.

All models performed above random chance on the transitivity task. Flan-T5 scored almost perfectly on this task and ChatGPT also had a high score. The other models performed above random chance, with the exception of Alpaca on three-shot prompts.

For the coordinate task, performance was quite low with zero-shot prompting, with the exception of ChatGPT. However, performance increased significantly with three-shot prompting. Given the difficulty of this task, which requires knowledge of 2-D spatial relationships, inequalities, and basic geometric formulas, the performance of the models is encouraging. In the case of three-shot prompts, the models outperformed random chance.

For the two dynamic tasks, existence and shuffle tracking, the models performed surprisingly well, especially with three-shot prompting. It's interesting that LLaMA and Alpaca had higher performance for the dynamic task of existence tracking than the static existence task when presented with zero-shot prompts. On the shuffle tracking task, three-shot prompting made the models perform exceptionally well, with five out of six models having a score of 0.8950 or higher. Given the difficulty of this task, which requires modelling shuffles to an initial state and the understanding of relational directions, these scores are impressive.

Shots In Table 1 the performance for zero and three-shot prompts is given. Note that for the majority of tasks, three-shot prompting increased the performance of a model. This is most notable in the coordinate task; for example, GPT-J goes from a score of 0.0020 to 0.5935, an improvement of almost 300x.

However, the inclusion of multiple shots did not universally increase performance. For example, the transitivity task did not benefit from three-shot prompting as the performance of all models decreased. In the case of ChatGPT, performance was more than halved. Since the addition of multiple shots increases the length of the prompt, this decrease in performance may be due to the phenomena of catastrophic forgetting, where the model forgets old information as new information is presented. The decrease in performance may also be attributed to a poor method of prompting.

Data Size As seen in Figure 2a, the performance of the models generally decreased as the number of shapes in the canvas increases. This is expected as the prompts become longer in length as the number of shapes increases, which increases the likelihood of catastrophic forgetfulness and decreases the performance of pure chance. Additionally, having more shapes in the canvas increases the general complexity of a task. Note that the answer distribution of a task was fixed so having more shapes does not increase the chance of an answer (e.g. for the existence task, having more shapes increases the number of questions with an answer of ”Yes”, but there is an even distribution of ”Yes” and ”No” answers).

The one exception where a model performed better with more shapes in the canvas is LLaMA. LLaMA had *slightly* better performance with five shapes than three shapes when presented with zero-shot prompts. LLaMA's high overall performance on five shapes with zero-shot prompts is attributed to LLaMA having a score of 0.3560 on the task count with five shapes, as compared to 0.0490 with three shapes. On the other tasks, LLaMA performed worse with five shapes than three shapes.

Overall Performance In Figure 2b, the aggregated performance of each model is shown. Despite being finetuned with instruction-following data, Alpaca and Dolly did not perform significantly better than their base models of LLaMA and GPT-J. In fact, Alpaca actually performed worse than LLaMA. Flan-T5 and ChatGPT both significantly outperformed the other models, with their average score almost 0.2 points higher. Overall, ChatGPT displayed the best performance.

6 Conclusion

We presented CRUGS, which is a simple geometric dataset to aid in the evaluation of LLMs. The goal of CRUGS is to have a dataset centered around a singular reference with tasks of increasing difficulty. The usage of simple shapes and properties allows for the reasoning skills of the model to be tested. Additionally, our dataset is easily extensible for new tasks without the need of human experts and flexible to new methods of prompting.

While the dataset is simple in nature, we showed that current state-of-the-art models perform poorly on certain tasks, most notably on the count task. Additionally, we showed that the models typically perform better with multiple shots and fewer shapes in the canvas.

Future work could include interpretability studies to evaluate how a model internally represents the canvas, prompting studies to evaluate how the model performance is affected by the prompt, or studies to determine how the complexity of the task affects the model.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Ali Ghodsi, Patrick Wendell, and Matei Zaharia. 2023. Hello dolly: Democratizing the magic of chatgpt with open models.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Los Alamitos, CA, USA. IEEE Computer Society.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

OpenAI. 2022. Chatgpt: Large language model.

Aarohi Srivastava et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.