
MAPPING LANGUAGE MODELS TO GROUNDED CONCEPTUAL SPACES

Roma Patel & Ellie Pavlick (ICLR 2022)

Presenter: Abdulrahman Alabdulakreem - 16 April 2024

Motivation

- LLMs have achieved amazing results in text-based benchmarks

Motivation

- LLMs have achieved amazing results in text-based benchmarks
- A main criticism of LLMs: lack of grounding
 - The ability to tie word's representation from textual domain to referent in non-linguistic world
 - Some say “Text alone is not enough”
- Want to analyze to what extent this is true

Motivation

- It is indisputable that text-only models do not learn representations of concepts that are grounded in the non-text world
- BUT, is it possible for the structure of relations between concepts in text form to be similar to what a grounded model would learn?


Relevant Works

Relevant Works

- Experience Grounds Language (Bisk, et al. 2020)
 - NLU is held back by lack of physical world grounding
 - Need to prioritize grounding and agency

We define five levels of **World Scope**:

WS1. Corpus (*our past*)

 WS2. Internet (*most of current NLP*)

WS3. Perception (*multimodal NLP*)

WS4. Embodiment

WS5. Social

Relevant Works

- Experience Grounds Language (Bisk, et al. 2020)
 - NLU is held back by lack of physical world grounding
 - Need to prioritize grounding and agency
- Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data (Bender & Koller, 2020)
 - LLMs do not “understand” or “comprehend” natural language
 - “the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning”

Relevant Works

- Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color (Abdou, et al. 2021)
 - Previous paper we just covered

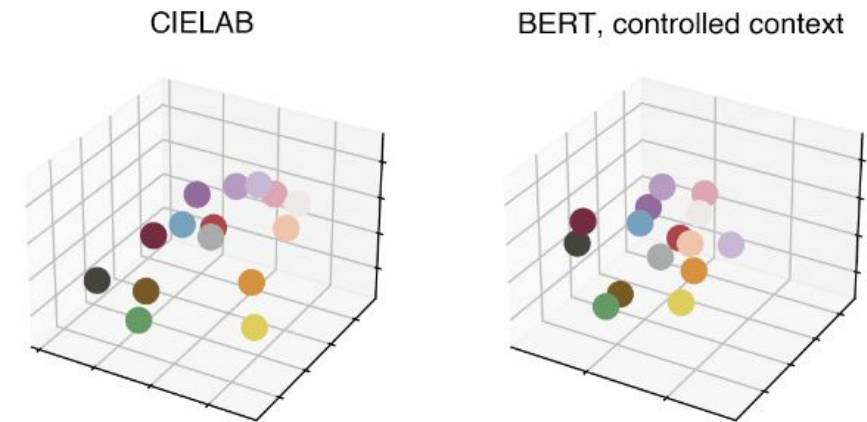


Figure 1: Right: Color orientation in 3d CIELAB space. Left: linear mapping from BERT (CC, see §2) color term embeddings to the CIELAB space.

Relevant Works

- Implicit Representations of Meaning in Neural Language Models (Li, et al. 2021)
 - Word embeddings model entities and situations in stories
 - Contains each entity's current properties and relations
 - can be manipulated with predictable effects on language generation
- Models build a representation of the input beyond basic linguistic relationships

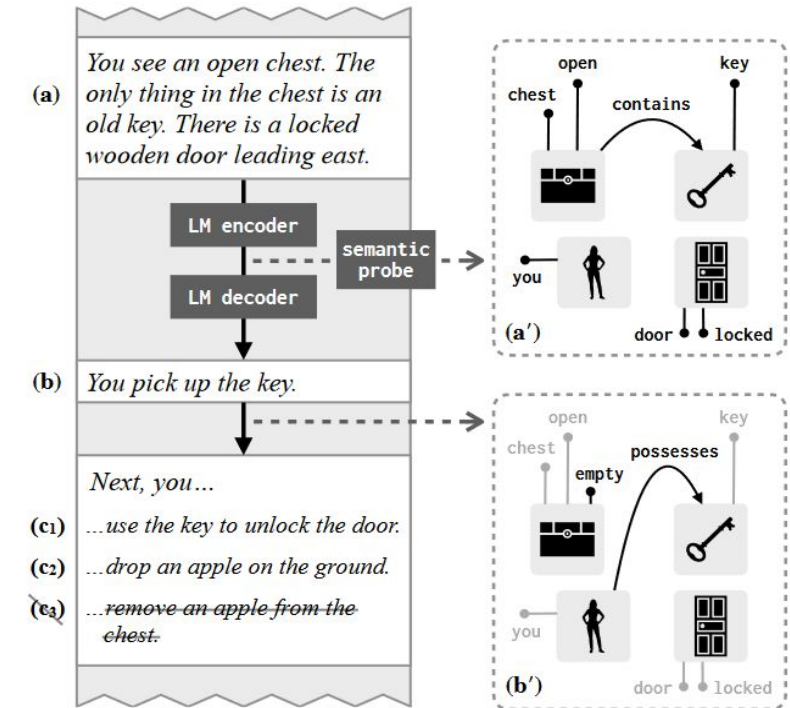


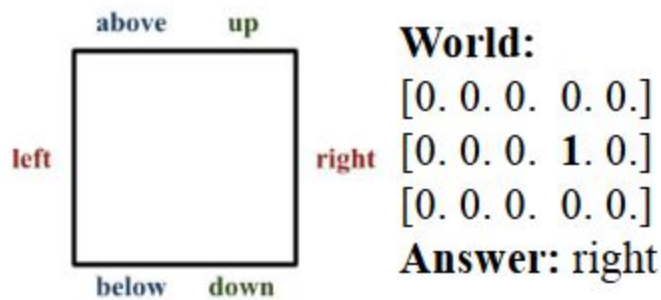
Figure 1: Neural language models trained on text alone (a–c) produce semantic representations that encode properties and relations of entities mentioned in a discourse (a'). Representations are updated when the discourse describes changes to entities' state (b').

Methodology

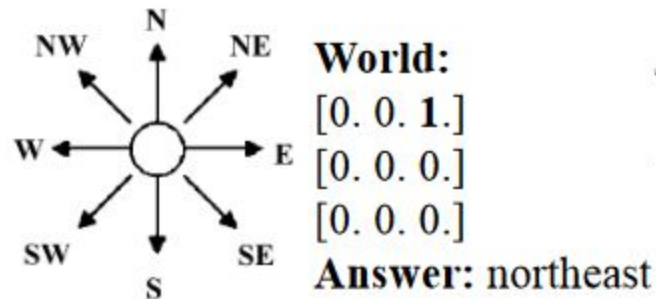
Methodology

- 3 Worlds
 - Spatial Terms: 6 concepts (left, right, up, down, top, bottom)
 - Cardinal Directions: 8 concepts (north, south, east, west, northeast, ...)
 - Colour Terms: 367 colors mapped to 3d space.

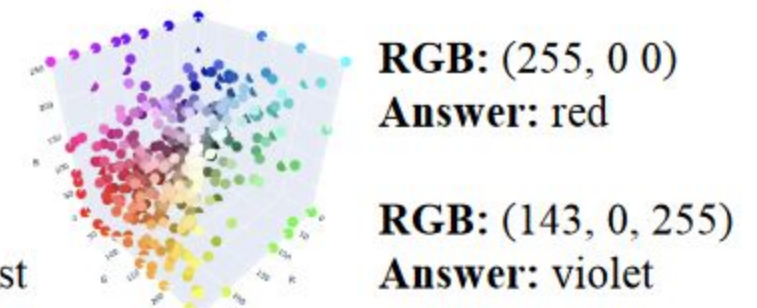
Spatial Terms



Cardinal Directions



RGB Colours



Methodology

- Spatial Terms example prompt:

Example Input (20 in-context-learning examples followed by prompt)

World: [0. 0. 0.] [0. 0. 0.] [0. 0. 1.] Answer: right	World: [0. 0. 0.] [0. 0. 0.] [0. 0. 1.] Answer: right	World: [0. 0. 0.] [0. 0. 0.] [0. 0. 1.] [0. 0. 0.] [0. 0. 0.] Answer: right	World: [1. 0.] [0. 0.] [0. 0.] [0. 0.] [0. 0.] Answer: left
World: [1. 0. 0. 0.] Answer: left ...13 more...	World: [0. 0.] [1. 0.] [0. 0.] Answer: left	World: [0. 1. 0. 0.] Answer: left	World: [1. 0. 0. 0.] [0. 0. 0. 0.] Answer:

Example Model Outputs

GPT-2 (124M)	
world	P=0.09
0. 0.]]	P=0.08
[0 [0	P=0.01

GPT-3 (175B)	
left	P=0.20
right	P=0.11
leftmost	P=0.01

Prevent Memorization

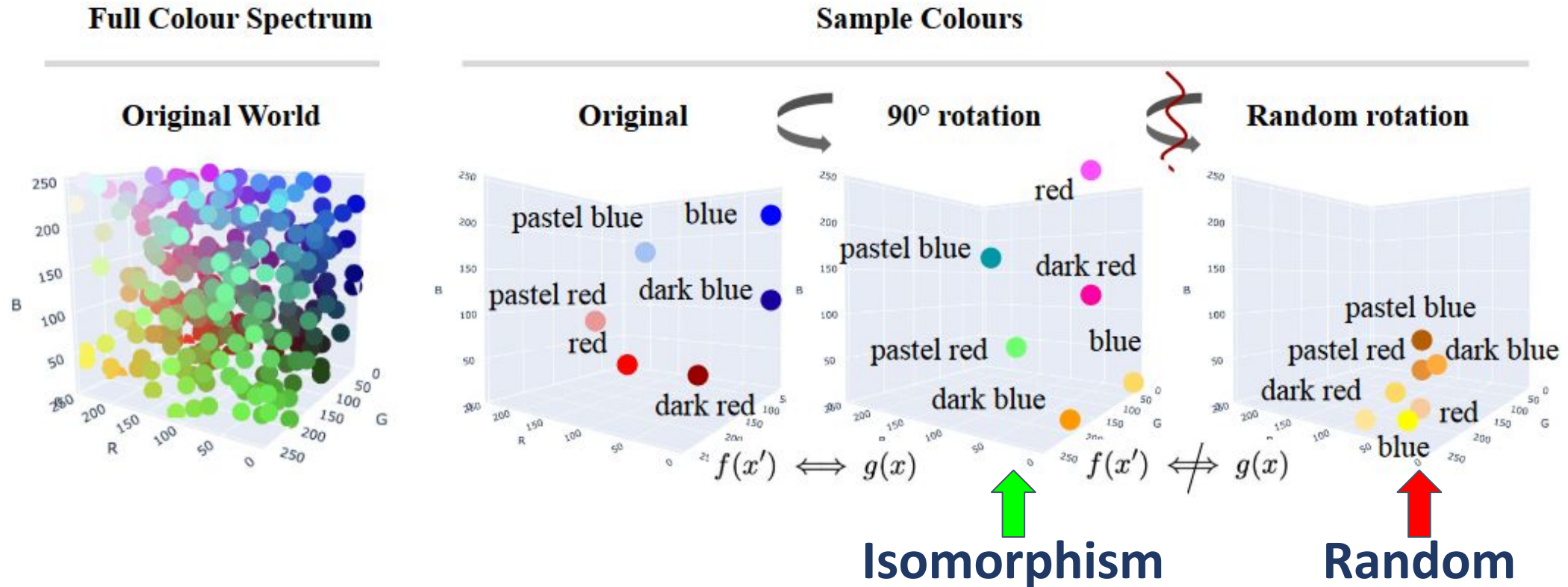
- Need to ensure that models don't simply answer based on training data
 - Plausible that world encountered in training data

Prevent Memorization

- Need to ensure that models don't simply answer based on training data
 - Plausible that world encountered in training data
- **Control #1:** Use isomorphic transformation on the space while preserving structural relations between concepts.
 - Example: Swap the pairs (left-right, above-below, up-down)
 - Example: Rotate compass by {90, 180, 270} degrees.
 - Example: Rotate RGB space around axis by {90, 180, 270} degrees.
 - LLM should maintain accuracy under isomorphic transformation
- **Control #2:** Use random perturbations to disturb structural relations
 - Example: make left "north" and right "east"
 - LLM should NOT be able to achieve good results here

Prevent Memorization

- Example control on RGB color world



Methodology

- 5 Models
 - GPT-2 (124M, 355M, 774M, 1.5B) and GPT-3 (175B)
 - pre-trained on OpenAI Web-Text dataset
- Only In-Context learning (no gradient updates)
 - 20 grounded concepts for spatial terms, 60 for colours
 - No significant increase in performance past that
- 5 tokens per prompt, average of 3 samples per prompt
- Baselines
 - R-IV: random token from vocabulary
 - R-ID: random label from possible answer
- Metrics: Top-1 & Top-3 Accuracy
 - Substring of ground truth is correct (e.g. “red” is correct if ground is “deep red”)
 - Grounding distance for RGB world

Results

- 3 Sections:
- #1: Generalisation to unseen worlds
 - Cardinal Directions: Have seen north in several sized grids, identify north in new unseen grid size
- #2: Generalisation to unseen but related concepts
 - Cardinal Directions: Have seen north and east in several sized grids, identify south and west in new grid
- #3: Analysis of predictions and errors made by models
 - Evaluate in-domain/out-domain incorrect answers
 - Color Space: Map predicted color into space and report distance as error

Results - Experiment #1

- #1: Generalisation to unseen worlds
 - Spatial and Cardinal worlds

	Top-1 Accuracy						Top-3 Accuracy					
	Spatial			Cardinal			Spatial			Cardinal		
	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.
R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.16	0.16	0.16	0.13	0.13	0.13
124 M	0.11	0.10	0.10	0.13	0.12	0.11	0.23	0.21	0.10	0.25	0.24	0.10
355 M	0.12	0.12	0.10	0.11	0.14	0.10	0.24	0.25	0.15	0.23	0.14	0.12
774 M	0.08	0.09	0.10	0.11	0.12	0.11	0.12	0.19	0.14	0.18	0.17	0.11
1.7 B	0.10	0.11	0.11	0.10	0.11	0.10	0.11	0.18	0.15	0.12	0.12	0.13
175 B	0.45	0.44	0.16	0.43	0.46	0.18	0.76	0.75	0.19	0.88	0.76	0.21

Results - Experiment #1

- #1: Generalisation to unseen worlds
 - Spatial and Cardinal worlds

	Top-1 Accuracy						Top-3 Accuracy					
	Spatial			Cardinal			Spatial			Cardinal		
	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.
R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.16	0.16	0.16	0.13	0.13	0.13
124 M	0.11	0.10	0.10	0.13	0.12	0.11	0.23	0.21	0.10	0.25	0.24	0.10
355 M	0.12	0.12	0.10	0.11	0.14	0.10	0.24	0.25	0.15	0.23	0.14	0.12
774 M	0.08	0.09	0.10	0.11	0.12	0.11	0.12	0.19	0.14	0.18	0.17	0.11
1.7 B	0.10	0.11	0.11	0.10	0.11	0.10	0.11	0.18	0.15	0.12	0.12	0.13
175 B	0.45	0.44	0.16	0.43	0.46	0.18	0.76	0.75	0.19	0.88	0.76	0.21

Results - Experiment #1

- #1: Generalisation to unseen worlds
 - Spatial and Cardinal worlds

	Top-1 Accuracy						Top-3 Accuracy					
	Spatial			Cardinal			Spatial			Cardinal		
	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.
R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.16	0.16	0.16	0.13	0.13	0.13
124 M	0.11	0.10	0.10	0.13	0.12	0.11	0.23	0.21	0.10	0.25	0.24	0.10
355 M	0.12	0.12	0.10	0.11	0.14	0.10	0.24	0.25	0.15	0.23	0.14	0.12
774 M	0.08	0.09	0.10	0.11	0.12	0.11	0.12	0.19	0.14	0.18	0.17	0.11
1.7 B	0.10	0.11	0.11	0.10	0.11	0.10	0.11	0.18	0.15	0.12	0.12	0.13
175 B	0.45	0.44	0.16	0.43	0.46	0.18	0.76	0.75	0.19	0.88	0.76	0.21

Results - Experiment #1




	Top-1 Accuracy						Top-3 Accuracy					
	Spatial			Cardinal			Spatial			Cardinal		
	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.	Orig.	Rot.	Rand.
R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.16	0.16	0.16	0.13	0.13	0.13
124 M	0.11	0.10	0.10	0.13	0.12	0.11	0.23	0.21	0.10	0.25	0.24	0.10
355 M	0.12	0.12	0.10	0.11	0.14	0.10	0.24	0.25	0.15	0.23	0.14	0.12
774 M	0.08	0.09	0.10	0.11	0.12	0.11	0.12	0.19	0.14	0.18	0.17	0.11
1.7 B	0.10	0.11	0.11	0.10	0.11	0.10	0.11	0.18	0.15	0.12	0.12	0.13
175 B	0.45	0.44	0.16	0.43	0.46	0.18	0.76	0.75	0.19	0.88	0.76	0.21

- Original/Isomorphism world scores higher than random world
- No significant degradation in isomorphisms
- GPT-2 models are not better than random guessing

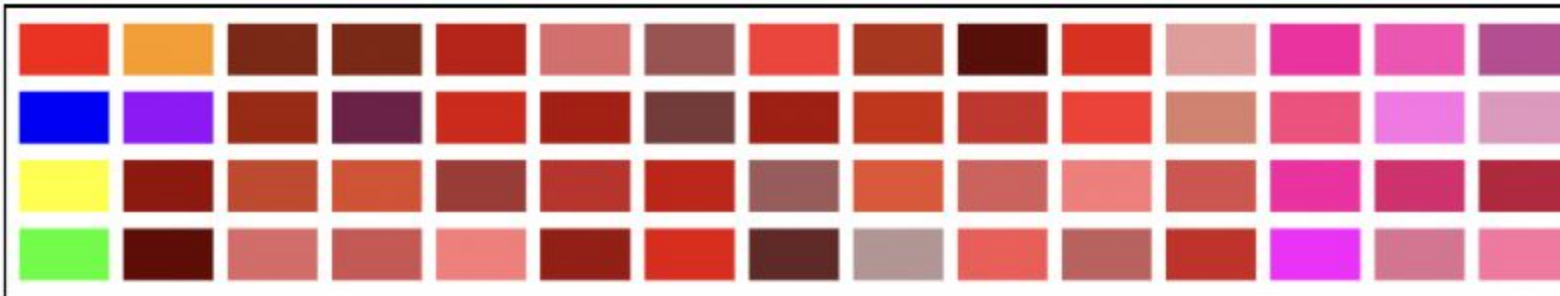
Results - Experiment #2

- #2: Generalisation to unseen but related concepts
 - Example: ICL shades of red -> test shades of blue

Example Input (60 in-context-learning examples followed by prompt)

 RGB: (48, 213, 200) Answer: orange	 RGB: (220, 20, 60) Answer: crimson	 RGB: (0, 0, 128) Answer:
---------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------

All 60 Training Examples



6 Primary and Secondary Colours

red, blue, yellow, green, orange, violet

57 Colours Within a Sub-space

dark red, maroon, crimson, fuchsia, rust, bright red..

Example Model Outputs

GPT-2 (124M)

color	P=0.14
black	P=0.08
rgb	P=0.02

GPT-3 (175B)

navy	P=0.24
dark blue	P=0.13
blue	P=0.08

Results - Experiment #2

		Spatial			Cardinal			Colours		
		Original	Rotated	Random	Original	Rotated	Random	Original	Rotated	Random
Top-1 Accuracy	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.10	0.11	0.04	0.11	0.10	0.05	0.08	0.09	0.03
	355 M	0.10	0.10	0.04	0.10	0.11	0.06	0.06	0.07	0.04
	774 M	0.09	0.11	0.03	0.13	0.12	0.08	0.11	0.09	0.01
	1.5 B	0.14	0.14	0.12	0.13	0.14	0.10	0.10	0.09	0.06
	175 B	0.28	0.27	0.13	0.30	0.29	0.08	0.23	0.21	0.11
Top-3 Accuracy	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.13	0.12	0.07	0.09	0.09	0.08	0.06	0.05	0.04
	355 M	0.24	0.17	0.15	0.19	0.17	0.10	0.14	0.11	0.12
	774 M	0.19	0.24	0.12	0.17	0.15	0.11	0.15	0.16	0.14
	1.5 B	0.32	0.29	0.20	0.21	0.20	0.14	0.19	0.18	0.16
	175 B	0.64	0.65	0.21	0.60	0.61	0.09	0.34	0.36	0.13

Results - Experiment #2

		Spatial			Cardinal			Colours		
		Original	Rotated	Random	Original	Rotated	Random	Original	Rotated	Random
Top-1 Accuracy	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.10	0.11	0.04	0.11	0.10	0.05	0.08	0.09	0.03
	355 M	0.10	0.10	0.04	0.10	0.11	0.06	0.06	0.07	0.04
	774 M	0.09	0.11	0.03	0.13	0.12	0.08	0.11	0.09	0.01
	1.5 B	0.14	0.14	0.12	0.13	0.14	0.10	0.10	0.09	0.06
	175 B	0.28	0.27	0.13	0.30	0.29	0.08	0.23	0.21	0.11
	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.13	0.12	0.07	0.09	0.09	0.08	0.06	0.05	0.04
Top-3 Accuracy	355 M	0.24	0.17	0.15	0.19	0.17	0.10	0.14	0.11	0.12
	774 M	0.19	0.24	0.12	0.17	0.15	0.11	0.15	0.16	0.14
	1.5 B	0.32	0.29	0.20	0.21	0.20	0.14	0.19	0.18	0.16
	175 B	0.64	0.65	0.21	0.60	0.61	0.09	0.34	0.36	0.13

Results - Experiment #2

		Spatial			Cardinal			Colours		
		Original	Rotated	Random	Original	Rotated	Random	Original	Rotated	Random
Top-1 Accuracy	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.10	0.11	0.04	0.11	0.10	0.05	0.08	0.09	0.03
	355 M	0.10	0.10	0.04	0.10	0.11	0.06	0.06	0.07	0.04
	774 M	0.09	0.11	0.03	0.13	0.12	0.08	0.11	0.09	0.01
	1.5 B	0.14	0.14	0.12	0.13	0.14	0.10	0.10	0.09	0.06
	175 B	0.28	0.27	0.13	0.30	0.29	0.08	0.23	0.21	0.11
Top-3 Accuracy	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.13	0.12	0.07	0.09	0.09	0.08	0.06	0.05	0.04
	355 M	0.24	0.17	0.15	0.19	0.17	0.10	0.14	0.11	0.12
	774 M	0.19	0.24	0.12	0.17	0.15	0.11	0.15	0.16	0.14
	1.5 B	0.32	0.29	0.20	0.21	0.20	0.14	0.19	0.18	0.16
	175 B	0.64	0.65	0.21	0.60	0.61	0.09	0.34	0.36	0.13

Results - Experiment #2

		Spatial			Cardinal			Colours		
		Original	Rotated	Random	Original	Rotated	Random	Original	Rotated	Random
Top-1 Accuracy	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.10	0.11	0.04	0.11	0.10	0.05	0.08	0.09	0.03
	355 M	0.10	0.10	0.04	0.10	0.11	0.06	0.06	0.07	0.04
	774 M	0.09	0.11	0.03	0.13	0.12	0.08	0.11	0.09	0.01
	1.5 B	0.14	0.14	0.12	0.13	0.14	0.10	0.10	0.09	0.06
	175 B	0.28	0.27	0.13	0.30	0.29	0.08	0.23	0.21	0.11
	175 B	0.64	0.65	0.21	0.60	0.61	0.09	0.34	0.36	0.13

- GPT-3 achieves high performance despite unseen concepts & in isomorphic world
 - Thus not using memorization

Results - Error Analysis

- Are incorrect labels in or out-of-domain
 - If ground is “right” then “left” is wrong and “[0 [0” or “hello” are also wrong
- GPT-3 almost always in-domain unlike GPT-2 models
 - If only in-domain: GPT-3 gets 98%, smallest model gets 53% (RGB world)

Results - Using Exact Match

		Spatial			Cardinal			Colours		
		Original	Rotated	Random	Original	Rotated	Random	Original	Rotated	Random
Top-1 Accuracy	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.00	0.01	0.00	0.01	0.00	0.00	0.02	0.01	1.00
	355 M	0.02	0.01	0.00	0.03	0.02	0.01	0.03	0.03	0.01
	774 M	0.02	0.02	0.01	0.03	0.02	0.00	0.07	0.06	0.03
	1.5 B	0.06	0.07	0.03	0.07	0.07	0.05	0.09	0.09	0.05
	175 B	0.09	0.09	0.04	0.10	0.09	0.05	0.19	0.20	0.08
Top-3 Accuracy	R-IV	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	R-ID	0.16	0.16	0.16	0.13	0.13	0.13	0.00	0.00	0.00
	124 M	0.01	0.01	0.00	0.02	0.03	0.01	0.05	0.05	0.02
	355 M	0.04	0.04	0.02	0.03	0.04	0.02	0.06	0.05	0.04
	774 M	0.04	0.05	0.04	0.03	0.04	0.00	0.10	0.09	0.03
	1.5 B	0.09	0.09	0.04	0.09	0.10	0.05	0.12	0.11	0.06
	175 B	0.30	0.31	0.10	0.35	0.32	0.13	0.26	0.25	0.10

Results - Error Analysis

- RGB world Grounding distance instead of Top-1/Top-3 Accuracy
 - Map predicted color into space and report distance as error

$$d(c_1, c_2) = \begin{cases} \sqrt{(c_{1x} - c_{2x})^2 + (c_{1y} - c_{2y})^2 + (c_{1z} - c_{2z})^2} & \text{for } c_1 \in C \\ 500 & \text{for } c_1 \notin C \end{cases}$$

True G	Predicted G & Distance
dark red	wine (76.5), light crimson (208.1) dark slate gray (144.7)
light green	beige (126.6), light sea green (129.7) cerulean (185.7), violet (262.6)

Results - Error Analysis

- RGB world Grounding distance instead of Top-1/Top-3 Accuracy
 - Map predicted color into space and report distance as error

	124M	355M	774M	1.5B	175B
C	328.3	309.5	209.6	190.7	96.3
R-IV	334.9	334.9	334.9	334.9	334.9
R-ID	174.9	174.9	174.9	174.9	174.9

Table 3: Table shows average distance (lower is better) between model-predicted groundings and true groundings in the world, averaged over all instances in the test set. We see that the largest model has an average distance of predictions significantly lower than random

Discussion

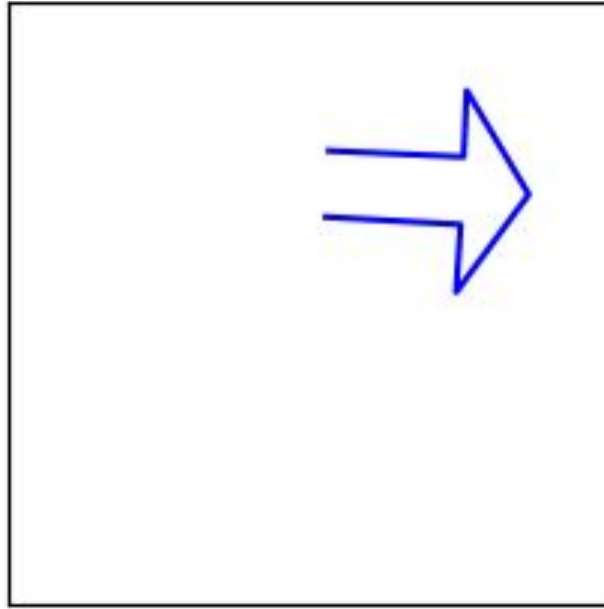
- GPT-3 able to learn a grounded conceptual space with few examples
 - Success in isomorphic and unseen concepts mean they aren't memorizing or simply recalling training data
 - Possibly exploiting conceptual structure of textual space (English dataset) to map onto novel spaces at test time
- Limitations:
 - GPT-2/GPT-3 only accept textual input
 - Most visual/sensory grounded concepts don't translate to text format/questions
 - Restricted to simple visual concepts that have textual representation (color/direction)
 - Naively converting to text loses complex structure

Assessment

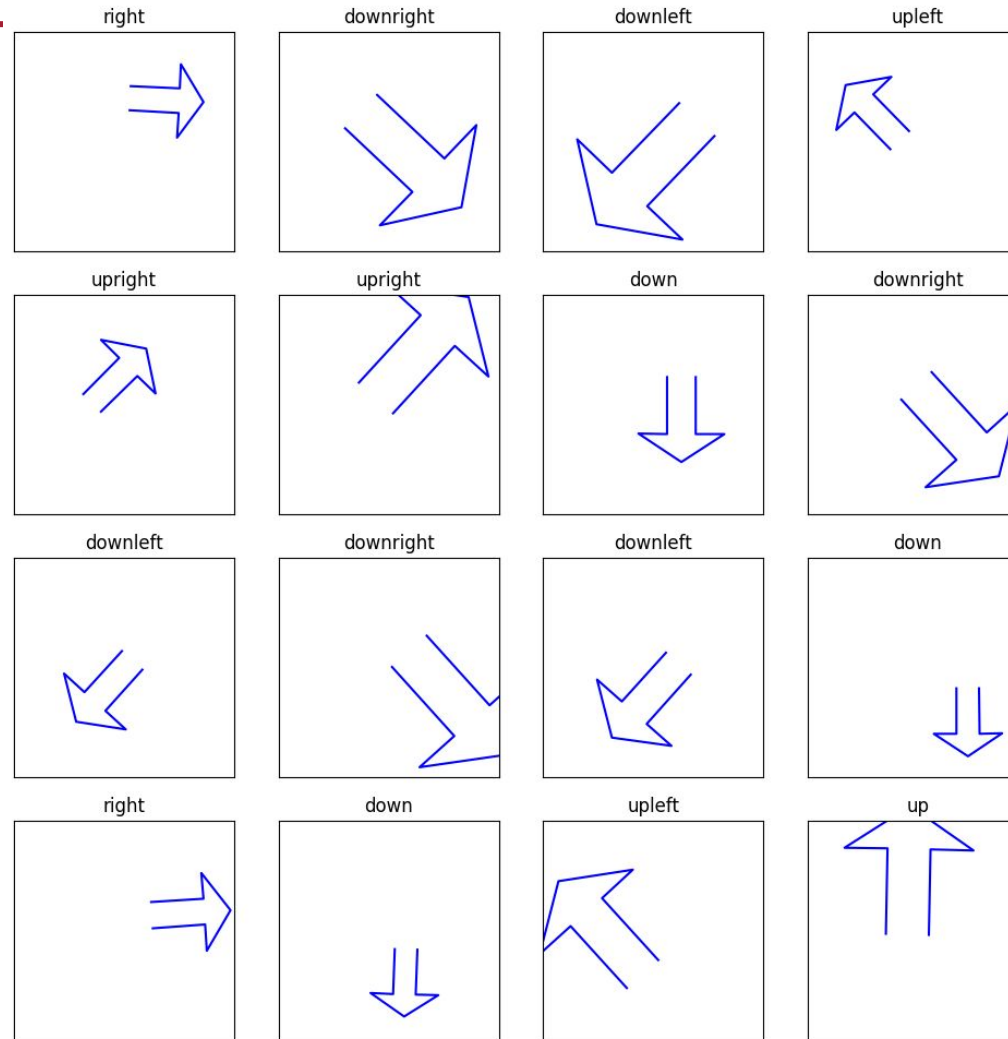
- Performance of models other than GPT-2/GPT-3?
 - Models not trained on OpenAI's dataset
- Why does GPT-2 get slightly worse with size?
 - Is it a consistent trend or random error?
- Are the tests sufficient to support the claim?
 - Many of the accuracies are small, can simple function $f(x)$ achieve similar accuracy from correlations?
 - Are isomorphisms sufficiently different from online datasets
- Future work using LVLMs?

- Created a new visual grounding task based on points on a plot forming an arrow
 - Example Input: [(1.2, 10.1), (10.1, 9.7), (10.3, 14.1), (14.4, 7.2), (9.6, 0.7), (9.9, 5.2), (1.0, 5.7)]
 - output: “right”

- Created a new visual grounding task based on points on a plot forming an arrow
 - Example Input: $[(1.2, 10.1), (10.1, 9.7), (10.3, 14.1), (14.4, 7.2), (9.6, 0.7), (9.9, 5.2), (1.0, 5.7)]$
 - output: “right”
 - Visual:



Extra



Extra

AR You

Solve the following visual grounding task which requires knowing where the arrow that the points form is pointed, only use single word answers:

Input: [(1.2, 10.1), (10.1, 9.7), (10.3, 14.1), (14.4, 7.2), (9.6, 0.7), (9.9, 5.2), (1.0, 5.7)]

Answer: right

Input: [(-2.1, 8.5), (10.1, -3.1), (15.9, 3.0), (13.2, -12.0), (-1.6, -15.3), (4.3, -9.2), (-7.9, 2.4)]

Answer: downright

Input: [(11.1, 1.0), (-1.1, -11.8), (5.3, -17.9), (-10.3, -15.1), (-13.8, 0.4), (-7.5, -5.7), (4.7, 7.0)]

Answer: downleft

Input: [(-5.0, -1.4), (-11.5, 5.3), (-14.8, 2.1), (-13.1, 10.3), (-4.8, 11.8), (-8.1, 8.6), (-1.6, 1.9)]

Answer: upleft

Input: [(-7.4, 1.9), (-0.9, 8.5), (-4.2, 11.8), (4.0, 10.2), (5.7, 2.0), (2.4, 5.2), (-4.2, -1.3)]

Answer: upright

Input: [(-5.4, 4.0), (5.8, 16.3), (-0.3, 21.9), (14.5, 19.6), (18.1, 5.1), (12.0, 10.7), (0.8, -1.6)]

Answer: upright

Input: [(7.7, 5.0), (7.7, -5.4), (12.9, -5.4), (5.1, -10.5), (-2.7, -5.3), (2.5, -5.4), (2.5, 5.0)]

Answer: down

Input: [(2.5, 5.9), (12.5, -5.1), (18.0, -0.1), (14.7, -13.1), (1.4, -15.1), (7.0, -10.1), (-3.0, 0.9)]

Answer: downright

Input: [(3.3, -0.4), (-3.4, -7.9), (0.3, -11.3), (-8.7, -9.9), (-10.9, -1.1), (-7.2, -4.5), (-0.4, 3.0)]

Answer: downleft

Input: [(6.9, 5.8), (18.3, -6.9), (24.6, -1.2), (20.8, -16.0), (5.6, -18.2), (11.9, -12.5), (0.6, 0.1)]

Answer: downright

Input: [(6.8, -1.3), (-1.2, -10.3), (3.3, -14.3), (-7.5, -12.8), (-10.2, -2.2), (-5.7, -6.2), (2.3, 2.7)]

Answer: downleft

Input: [(11.1, -3.8), (11.2, -12.1), (15.3, -12.0), (9.1, -16.2), (2.9, -12.1), (7.0, -12.1), (7.0, -3.8)]

Answer: down

Input: [(4.9, 5.2), (14.4, 5.8), (14.0, 10.5), (19.3, 3.7), (15.0, -3.7), (14.7, 1.1), (5.2, 0.4)]

Answer: right

Input: [(5.2, -3.5), (4.9, -11.7), (9.0, -11.8), (2.8, -15.7), (-3.3, -11.4), (0.8, -11.6), (1.1, -3.4)]

Answer: down

Input: [(-4.8, -10.5), (-15.0, 0.8), (-20.6, -4.3), (-17.2, 9.0), (-3.7, 11.1), (-9.3, 5.9), (0.9, -5.4)]

Answer: upleft

Input: [(-5.8, -0.6), (-5.5, 15.0), (-13.3, 15.1), (-1.4, 22.7), (10.1, 14.6), (2.3, 14.8), (2.0, -0.8)]

Answer:

< 10 / 10 >

ChatGPT

upright



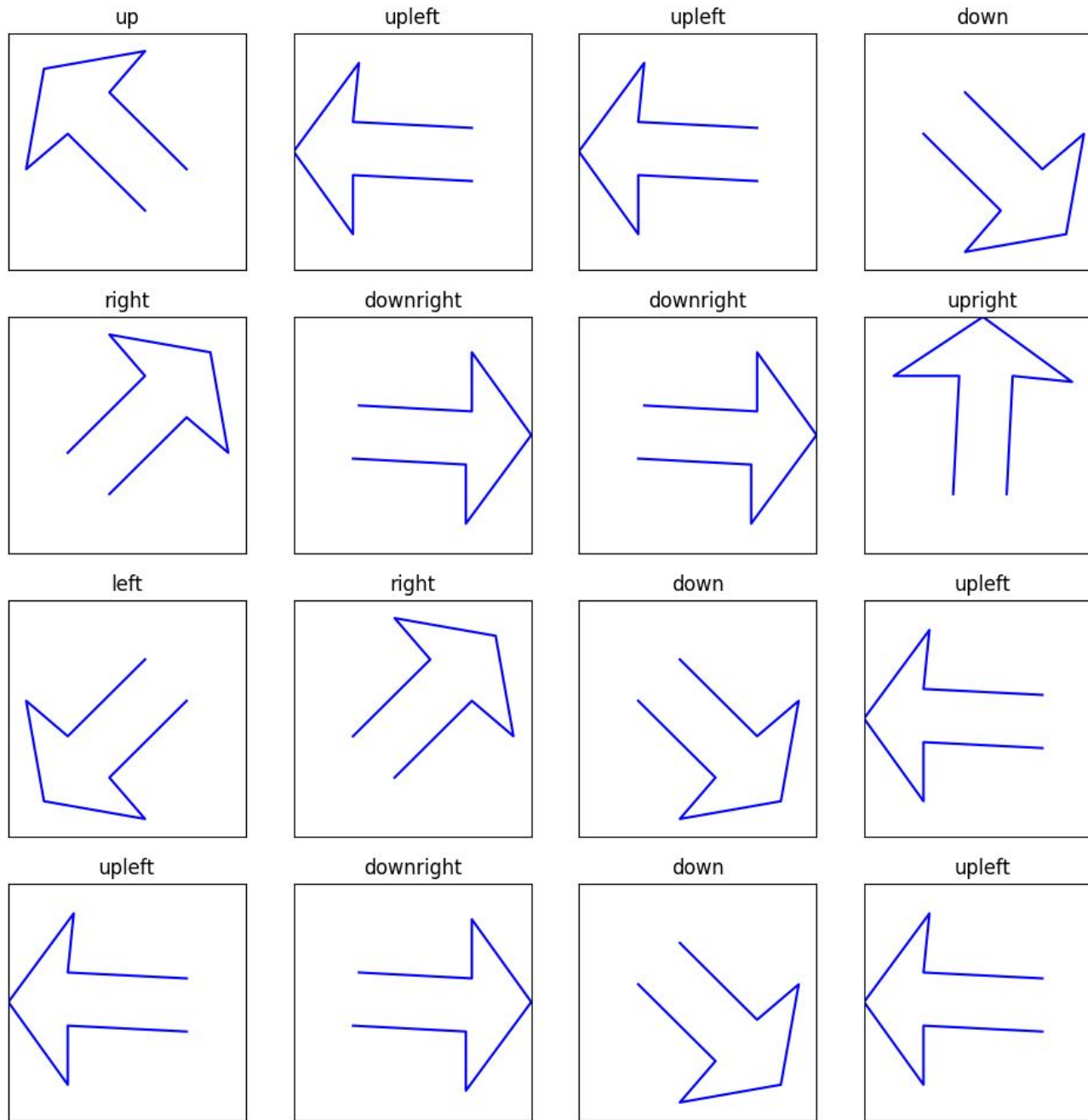
Incorrect, It is pointing up|



- Results GPT 3.5:
 - Without any scaling or translation: ~100% (5/5)
 - Answer is contained in the context
 - Without any scaling or translation (excluding label from examples): ~100% (5/5)
 - Without any scaling or translation (excluding subset from examples): ~100% (5/5)
 - Input only contains [“down”, “left”, “downleft”] target is “upright”
 - With random scale/translate and perturbations: ~10% (1/10)

Extra

- 45 degree Isomorphism



- Results GPT 3.5 - Isomorphism (45 degrees)
 - Without any scaling or translation: ~100% (5/5)
 - Without any scaling or translation (excluding label from examples): ~100% (5/5)
 - Without any scaling or translation (excluding subset from examples): ~100% (5/5)
 - Input only contains [“down”, “left”, “downleft”] target is “upright”
 - With random scale/translate and perturbations: ~0% (0/10)

Discussion

- GPT-3 able to learn a grounded conceptual space with few examples
 - Success in isomorphic and unseen concepts mean they aren't memorizing or simply recalling training data
 - Possibly exploiting conceptual structure of textual space (English dataset) to map onto novel spaces at test time
- Limitations:
 - GPT-2/GPT-3 only accept textual input
 - Most visual/sensory grounded concepts don't translate to text format/questions
 - Restricted to simple visual concepts that have textual representation (color/direction)
 - Naively converting to text loses complex structure