Evaluating Numerical Reasoning in Text-to-Image Models

Ivana Kajić
Google DeepMind

Olivia Wiles
Google DeepMind

Isabela Albuquerque
Google DeepMind

Matthias Bauer
Google DeepMind

Su Wang
Google DeepMind

Jordi Pont-Tuset
Google DeepMind

Aida Nematzadeh
Google DeepMind

Task 1: Exact Quantities

Generate images containing an **exact** quantity

Task 2: Approximate Quantities

Interpret **approximate** quantities expressed linguistically

Task 3: Complex Reasoning

Understand more **complex** numerical concepts

How do we evaluate numerical reasoning?

1. Design a set of **text prompts** for each of the 3 tasks
   - Task 1: Exact Number Generation
   - Task 2: Approximate Number Generation
   - Task 3: Complex Reasoning
2. Generate images using 12 different **text-to-image models**
3. **Annotate** images with counts/descriptions of objects
4. Use annotations to **evaluate model accuracy**

Step 1: Prompt Templates

Task 1

Simple Numeric	{2,3}-additive	Colors	Spatial Relationships
• 3 cats. • Two koalas. • 7 cinnamon sticks. • 1 okra. • 6 paper clips. • Ten flutes.	• 1 chair and 3 kangaroos. • 4 coconuts and five cats. • 4 corkscrews, 1 olive and 2 pistachios. • 4 spoons, 4 pistachios and five parsnips.	• **Two green** apples. • **1 red** koala and **two** black apples. • **One black** mushroom and **3 black** bottles.	• There are four pistachios **to the right** of 4 flies. • There are 2 mushrooms **above** 3 tables. • There are **two** dogs **below 1** tree.

Task 2

Sentence Numeric	Approximate Quantifiers
• **An image showing** 5 mushrooms. • **There are** 5 mushrooms. • **There are** 5 mushrooms **in this image.**	• An image with some ants and some flutes. There are **fewer** ants than flutes. • An image of a vase. There are **many** flowers in the vase. • An image of a vase. There are **no** flowers in the vase.

Task 3

Fractional (simple, complex)	Part-whole
• A pizza cut into **3 slices**. • A cake cut into **quarters.** • An image of a pencil where **one half** of it is **red** and the **other half** is **blue**.	• There are **2** forks on the table, but one fork is broken into **two pieces**. • There are **4** plates on the table, but one plate is broken into **two pieces**.

Step 1: Prompt Templates

Task 1

Simple Numeric

- 3 cats.
- Two koalas.
- 7 cinnamon sticks.
- 1 okra.
- 6 paper clips.
- Ten flutes.

Spatial Relationships

- There are four pistachios **to the right** of 4 flies.
- There are 2 mushrooms **above** 3 tables.
- There are **two** dogs **below 1** tree.

Sentence Numeric

- **An image showing** 5 mushrooms.
- **There are** 5 mushrooms.
- **There are** 5 mushroo **in this image.**

3

Part-whole

- There are **2** forks on the table, but one fork is broken into **two pieces**.
- There are **4** plates on the table, but one plate is broken into **two pieces**.

are **no** flowers in the vase.

1386 Prompts

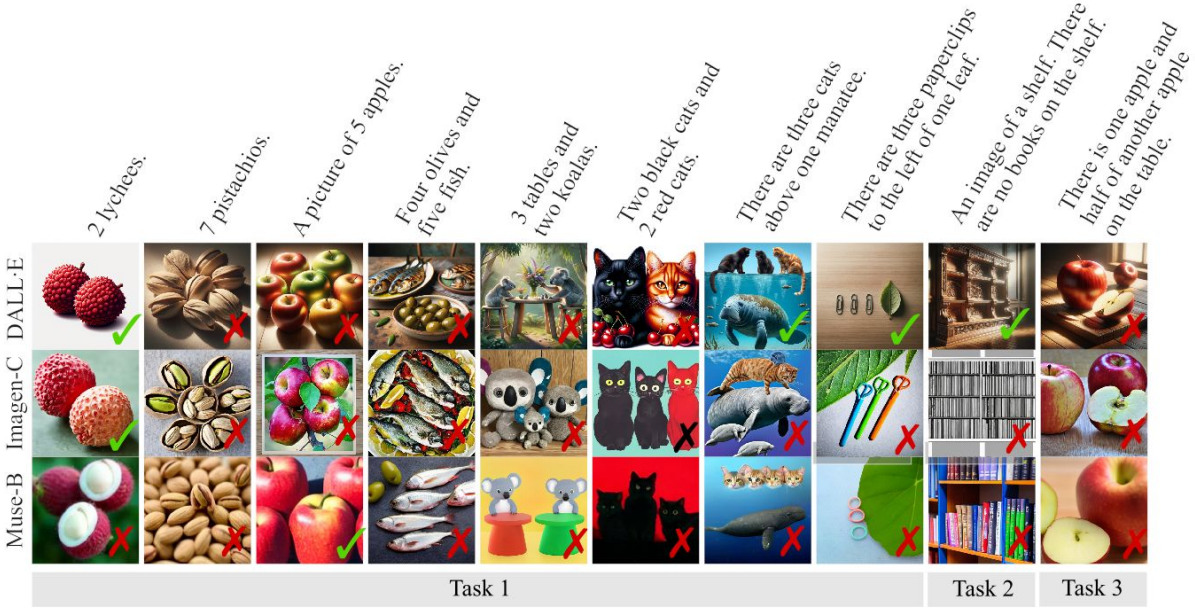
	Prompt Type	# of Prompts	Numbers
Task 1	numeric-simple	600	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
	attribute-color	160	1, 2, 3, 4
	numeric-sentence	100	1, 2, 3, 4, 5
	2-additive	100	1, 2, 3, 4, 5
	2-additive-color	100	1, 2, 3, 4, 5, 6, 7, 8
	3-additive	100	1, 2, 3, 4, 5
	attribute-spatial	100	1, 2, 3, 4, 5
Task 2	approx-1-entity	24	no, few, many
	approx-2-entity	45	fewer, as many as, more
Task 3	fractional-simple	36	1, 2, 3, 1/2, 1/3, 1/4, 1/5
	part-whole	15	1/2
	fractional-complex	6	1/3 + 2/3, 1/2

How do we evaluate numerical reasoning?

1. Design a set of **text prompts** for each of the 3 tasks
   - Task 1: Exact Number Generation
   - Task 2: Approximate Number Generation
   - Task 3: Complex Reasoning
2. Generate images using 12 different **text-to-image models**
3. **Annotate** images with counts/descriptions of objects
4. Use annotations to **evaluate model accuracy**

Step 2: Image generation

MODELS
1. Dalle-3
2. Midjourney v6
3. Imagen-A
4. Imagen-B
5. Imagen-C
6. Imagen-D
7. Muse-A
8. Muse-B
9. SD 1.5
10. SD 2.1
11. SDXL
12. SD 3



Total: ~83K Images

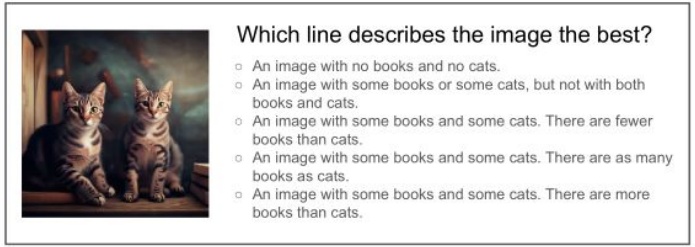
How do we evaluate numerical reasoning?

1. Design a set of **text prompts** for each of the 3 tasks
   - Task 1: Exact Number Generation
   - Task 2: Approximate Number Generation
   - Task 3: Complex Reasoning
2. Generate images using 12 different **text-to-image models**
3. **Annotate** images with counts/descriptions of objects
4. Use annotations to **evaluate model accuracy**

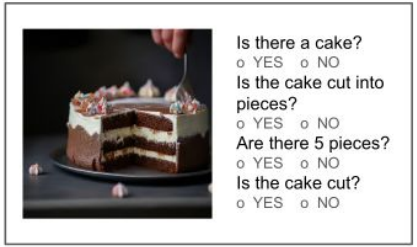
Step 3: Collecting Annotations

Annotation Task 1

Annotation Task 2



Annotation Task 3



- AT3: Questions are LLM-generated (Cho et al 2023, ICLR)

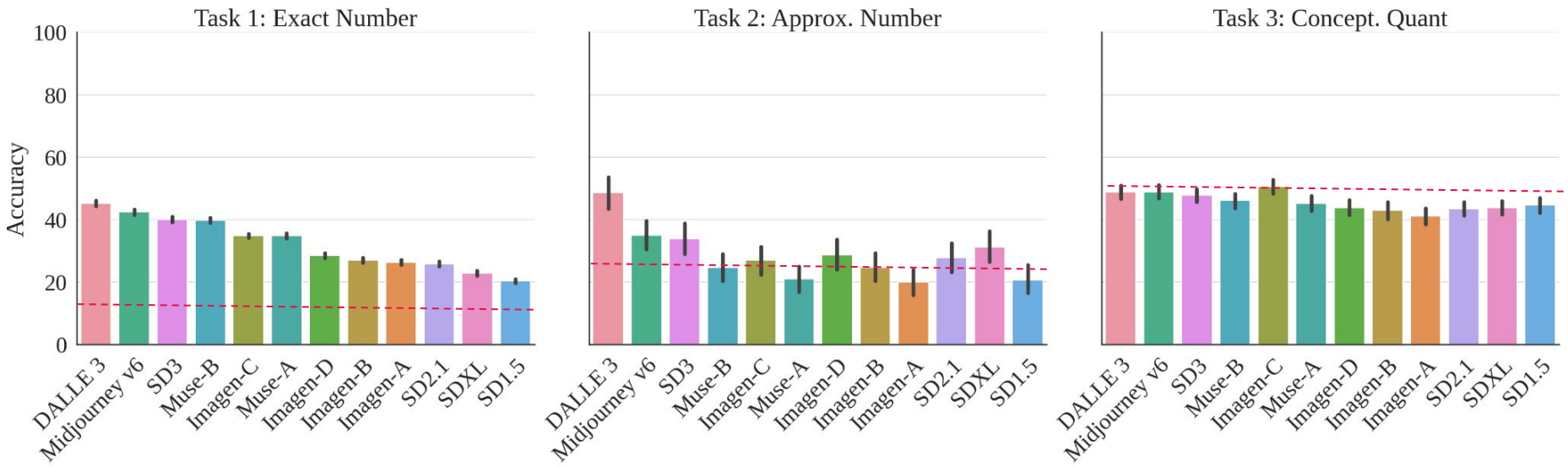
Total number of collected annotations (image labels)		
Task 1	**Task 2**	**Task 3**
718K (143K)	21K (4K)	62K (62K)

~801K annotations

How do we evaluate numerical reasoning?

1. Design a set of **text prompts** for each of the 3 tasks
   - Task 1: Exact Number Generation
   - Task 2: Approximate Number Generation
   - Task 3: Complex Reasoning
2. Generate images using 12 different **text-to-image models**
3. **Annotate** images with counts/descriptions of objects
4. Use annotations to **evaluate model accuracy**

Results



Random annotation baseline

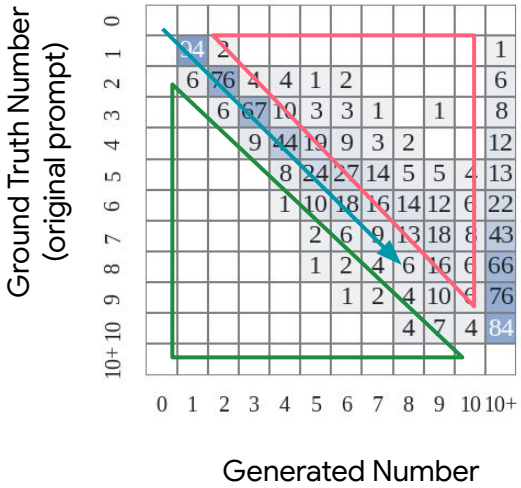
Accuracy

	DALLE 3	Midjourney v6	Imagen-A	Imagen-B	Imagen-C	Imagen-D	Muse-A	Muse-B	SD1.5	SD2.1	SDXL	SD3	Random	
numeric-simple	76	83	69	59	81	55	64	73	59	68	60	82	10	Task 1
attribute-color	79	84	69	62	76	54	72	69	52	64	52	75	25	Task 1
numeric-sentence	66	67	71	62	78	57	78	79	50	54	35	69	20	Task 1
2-additive	48	44	26	42	41	51	52	67	25	22	14	42	20	Task 1
2-additive-color	48	46	25	31	43	38	45	56	18	24	21	50	20	Task 1
3-additive	52	43	16	25	32	38	42	57	18	19	18	41	20	Task 1
attribute-spatial	40	33	10	19	25	20	23	42	15	12	16	28	20	Task 1
approx-1-entity	66	47	42	43	48	40	44	42	38	52	45	48	33	Task 2
approx-2-entity	40	29	8	15	16	23	9	16	12	15	24	26	20	Task 2
fractional-simple	50	53	52	53	58	54	54	56	54	52	52	52	50	Task 3
part-whole	46	45	25	30	42	28	34	33	29	30	28	44	50	Task 3
fractional-complex	48	32	19	16	29	22	21	21	28	25	31	32	50	Task 3

easier

harder

DALLE 3



- Diminishing accuracy
- Underestimation
- Overestimation

Conclusion

- **Dalle-3** has the **strongest** numerical reasoning capability
  - Difficulty: Task 1 < Task 2 < Task 3
- While helpful, **scaling alone does not seem enough** to develop an abstract, robust concept of numerosity
- More: auto-metrics, counting VQA benchmark, open-sourced benchmark

`github.com/google-deepmind/geckonum_benchmark_t2i`