

# Can Language Models Understand Visual Information?

Petter Mæhlum and David Samuel, Language Technology Group (LTG), UiO



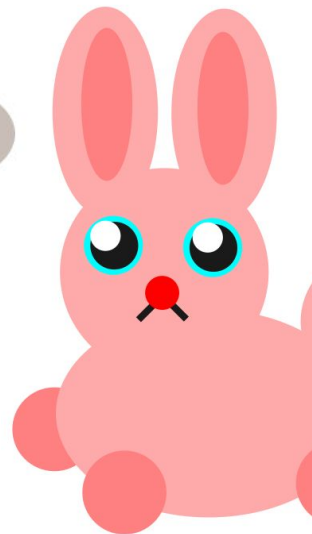
# Visual Data?

- We wanted to investigate SVGs
- We present our dataset and some investigations
- Slightly deceiving headline - Not really multimodality.
- Focus on interpretation of simplistic elements
- Secondary focus on aspects of SVG syntax



# The SVGs Ask themselves

“Can language models develop an understanding of visual properties – shapes, proportions, colors, and spatial relationships – from linguistic data?”



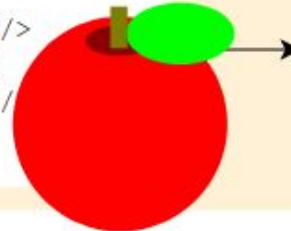
# SVG structure

- XML-based file format
- Opposed to rasterized images (pixels)
- Allows for infinite zooming (no quality loss)

Honey, I  
rasterized  
the kids!



```
<svg xmlns="http://www.w3.org/2000/svg" width="100" height="100">
  <circle cx="50" cy="50" r="50" fill="red"/>
  <ellipse cx="45" cy="15" rx="15" ry="6" fill="darkred"/>
  <ellipse cx="75" cy="15" rx="25" ry="15" fill="lime"/>
  <rect x="45" y="20" width="8" height="20" fill="olive"/>
</svg>
```

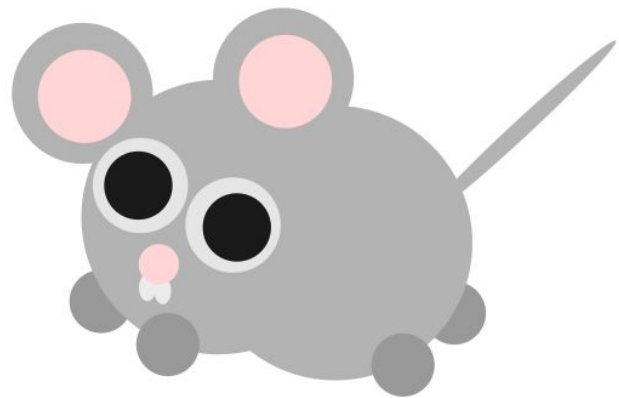


Which of the four captions best describes the following SVG image?

- A) strawberry
- B) apple**
- C) raspberry
- D) tomato

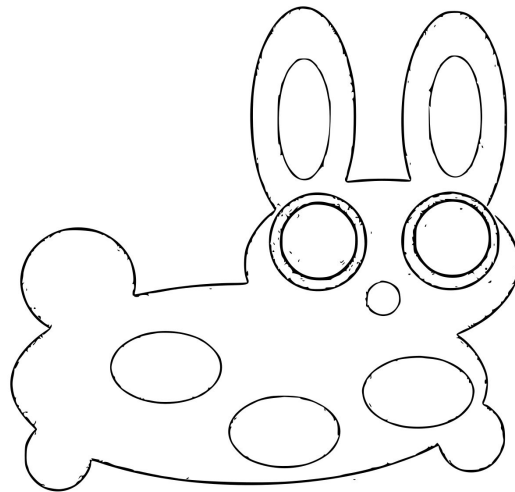
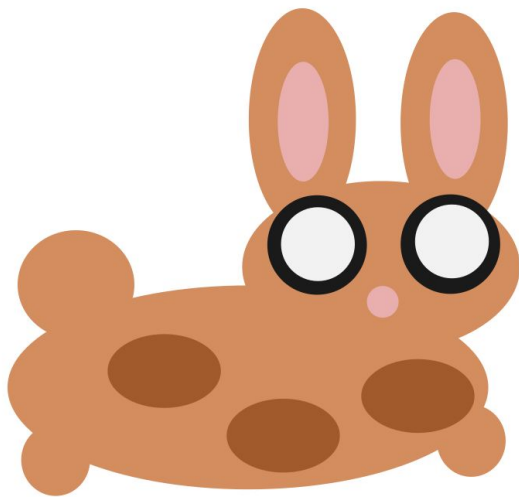
# Existing benchmarks

- There are many datasets out there
  - Sort-of-CLEVR
  - SVGEEditBench
  - SVG Taxonomy
  - VGBench
  - SGP-Bench
- But quality and usability varies
- Many datasets have dubious origins



# Vectorization of Rasterized Images is a Possibility

- “Easy” to make make bezier approximation - but not optimal



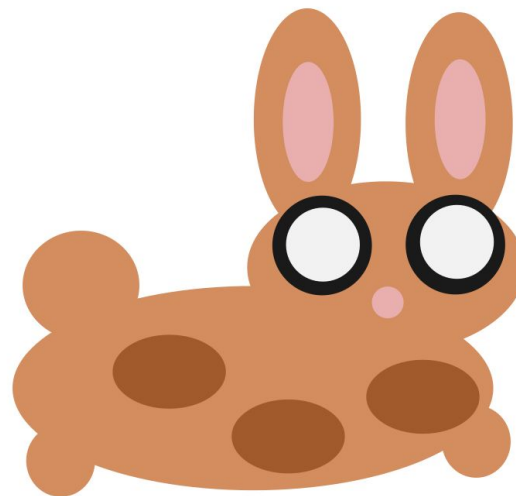
<https://picsvg.com/>

# Internal SVG Structure Can Vary

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <!-- Created with Inkscape (http://www.inkscape.org/) -->
3 <svg width="210mm" height="297mm" version="1.1" viewBox="0 0 210 297" xmlns="http://www.w3.org/2000/svg">
4   <g stroke-opacity="0.39772">
5     <g fill="#d38d5f">
6       <ellipse cx="57.96" cy="196.93" rx="6.1875" ry="12.817" stroke-width=".4"/>
7       <ellipse cx="75.484" cy="197.38" rx="6.1875" ry="12.817" stroke-width=".4"/>
8       <ellipse cx="51.678" cy="227.9" rx="27.749" ry="11.775" stroke-width=".45841"/>
9       <ellipse cx="67.054" cy="213.97" rx="16.039" ry="9.9126" stroke-width=".4"/>
10    </g>
11    <g stroke-width=".4">
12      <circle cx="59.665" cy="211.45" r="5.7455" fill="#1a1a1a"/>
13      <circle cx="75.054" cy="211.35" r="5.7455" fill="#1a1a1a"/>
14      <circle cx="59.76" cy="211.35" r="4.2618" fill="#f2f2f2"/>
15      <circle cx="75.23" cy="211.04" r="4.2618" fill="#f2f2f2"/>
16    </g>
17    <g stroke-width=".4">
18      <circle cx="67.242" cy="218.01" r="1.831" fill="#e9afaf"/>
19      <ellipse cx="29.485" cy="236.26" rx="3.9777" ry="4.1671" fill="#d38d5f"/>
20      <ellipse cx="77.456" cy="234.06" rx="3.9777" ry="4.1671" fill="#d38d5f"/>
21    </g>
22    <g fill="#a05a2c" stroke-width=".4">
23      <ellipse cx="42.018" cy="226" rx="6.5348" ry="4.2618"/>
24      <ellipse cx="55.756" cy="233.46" rx="6.5348" ry="4.2618"/>
25      <ellipse cx="71.3" cy="228.89" rx="6.5348" ry="4.2618"/>
26    </g>
27    <g stroke-width=".4">
28      <ellipse cx="58.055" cy="197.21" rx="2.9359" ry="6.9136" fill="#e9afaf"/>
29      <ellipse cx="75.597" cy="196.9" rx="2.9359" ry="6.9136" fill="#e9afaf"/>
30      <ellipse cx="31.832" cy="216.04" rx="6.7414" ry="6.295" fill="#d38d5f"/>
31    </g>
32  </g>
33 </svg>
34

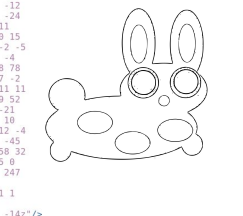
```



119 164 278 20 168 46 265 74 501 70 187 73 301 111 231

```
266 20 -10 21 -27 52 75
267 <path d="M4151 2924 c0 -11 3 -14 6 -6 3 7 2
268 <path d="M1617 2749 c7 -7 15 -10 18 -7 3 3 -
269 <path d="M4526 2508 c3 -5 10 -6 15 -3 13 9 1
270 <path d="M4245 2441 c-2 -5 -17 -11 -33 -14
```

```
16 -1 19 -3 4 -6 -2 -5 -13z"/>
2 9 -12 12 -14 6 -15 5 -6 -52z"/>
1 12 -6 12 -8 0 -12 -4 -9 -92z"/>
32 -7 -92 -60 -92 -83 0 -8 -5
```

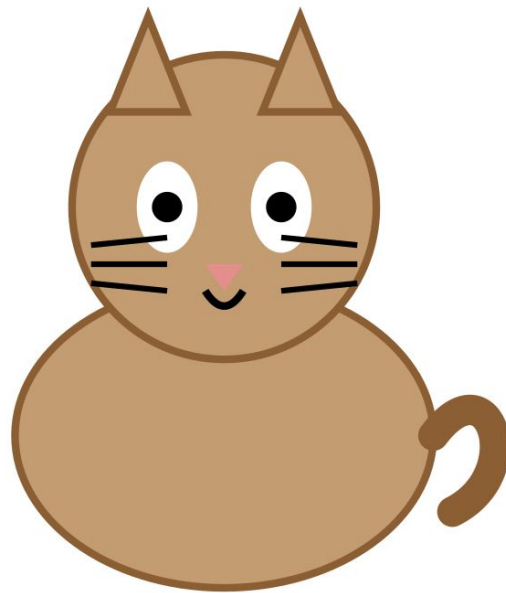


# Oops ... One more!



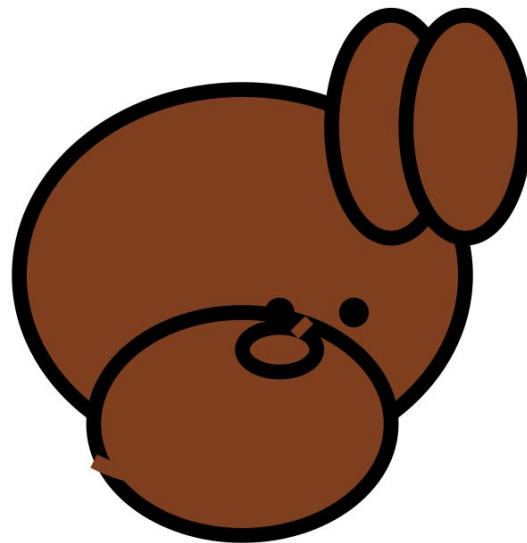
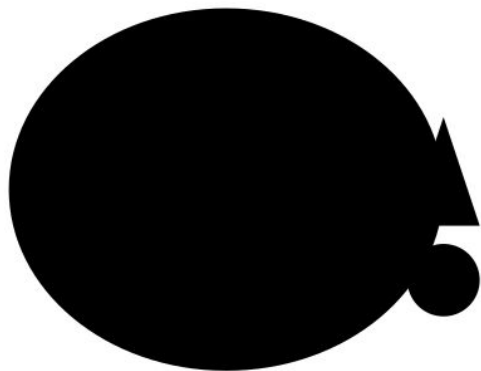
```
270<path d="M4245 2441 c-2 -5 -17 -11 -33 -14 -32 -7 -92 -60 -92 -83 0 -8 -5
271 -12 -11 -8 -8 4 -9 -3 -4 -24 3 -18 2 -33 -4 -37 -7 -4 -8 -12 -2 -24 6 -10 7
272 -27 4 -30 -3 -15 9 -24 11 -20 9 -3 -16 -11 16 -16 0 -34 90 -81 156 -75 62 6
273 96 27 133 82 31 47 33 54 27 113 -6 72 -25 99 -90 134 -42 22 -102 32 -111 192
274 m77 -21 c47 -13 93 -62 106 -111 37 -147 -125 -257 -248 -168 -23 17 -45 40
275 -48 51 -3 13 -7 16 -12 8 -7 -12 -3 75 6 117 6 31 52 82 88 98 38 17 61 18
276 108 52"/>
277<path d="M5336 2437 c-16 -12 -16 -13 2 -13 10 0 19 6 19 13 0 7 -1 13 -2 13
278 1 0 -9 -6 -19 -13"/>
279<path d="M701 2354 c0 -11 3 -14 6 -6 3 7 2 16 -1 19 -3 4 -6 -2 -5 -13"/>
280<path d="M2538 2363 c7 -3 16 -2 19 1 4 3 -2 6 -13 5 -11 0 -14 -3 -6 -62"/>
281<path d="M776 2172 c-3 -5 1 -9 9 -9 8 0 12 4 9 9 -3 4 -7 8 -9 8 -2 0 -6 -4
282 -9 -82"/>
283<path d="M827 2078 c9 -10 17 -18 17 -18 1 0 2 8 2 18 0 10 -8 18 -18 18 -16
284 0 -16 -2 -1 -182"/>
285<path d="M2015 1939 c-100 -10 -150 -23 -235 -62 -409 -188 -270 -598 223
286 -658 115 -14 245 -1 354 37 92 32 113 43 172 90 116 94 155 198 119 318 -18
287 60 -127 167 -209 206 -124 58 -281 83 -424 692 m210 -22 c63 -9 170 -40 200
288 58 16 -10 37 -22 45 -26 8 -3 38 -25 65 -48 128 -100 141 -256 33 -384 -15
289 -18 -32 -29 -38 -26 -6 4 -9 1 -8 -6 2 -7 -15 -23 -37 -35 -22 13 -53 -30
290 -68 -30 -16 -9 -35 -14 -43 -11 -8 3 -13 2 -9 -3 4 -8 -43 -22 -125 -38 -68
291 -12 -200 -14 -255 -3 -33 7 -70 13 -82 14 -13 0 -23 5 -23 10 0 4 -7 6 -15 2
292 0 -3 -15 -1 -15 5 0 6 -4 8 -8 5 -15 -9 -192 89 -192 106 0 4 -13 21 -30 38
293 -16 17 -30 40 -30 50 0 11 -4 20 -9 20 -5 9 -14 21 -20 46 -10 37 -9 56 5 183
294 17 61 87 156 106 145 6 -4 8 -2 5 4 -8 12 115 84 148 86 13 0 21 4 19 8 -4 6
295 49 19 155 38 43 8 159 6 226 -32"/>
296<path d="M5050 1836 c0 -2 8 -10 18 -17 15 -13 16 -12 3 4 -13 16 -21 21 -21
297 132"/>
298<path d="M5110 1790 c0 -5 5 -10 11 -10 5 0 7 5 4 10 -3 6 -8 10 -11 10 -2 0
299 4 -4 4 -102"/>
300<path d="M4535 1689 c-219 -20 -410 -138 -460 -284 -17 -50 -17 -63 -6 -118
301 -22 -106 93 -186 223 -249 110 -53 283 -73 338 -72 290 2 517 135 552 325 10
302 51 9 67 -7 112 -37 106 -138 195 -280 246 73 26 -109 33 -245 45 -19 1 -71
303 1 -115 -52 m212 -24 c35 -3 86 -13 115 -21 66 -19 190 -92 181 -107 -3 -6 -1
304 -7 6 -3 14 0 75 -50 66 -64 -4 -6 -1 -9 7 -8 8 2 12 -5 11 -20 -2 -13 1 -21 6
305 18 16 10 34 -53 29 107 -2 -28 -12 -68 -22 -88 -19 -41 -113 -139 -133 -139
306 -7 0 -13 -4 -13 -8 0 -5 -18 -15 -40 -22 -22 -7 -40 -16 -40 -20 0 -4 -39 -18
307 -87 -31 -111 -30 -291 -32 -393 -5 -110 29 -131 37 -124 48 4 7 2 8 -5 4 -15
308 10 -87 33 -138 82 -81 79 -115 194 -78 266 8 15 15 34 15 41 0 6 10 20 22 31
309 20 10 21 18 15 1 -4 -11 -3 -16 4 -12 5 4 -15 4 25 -4 13 5 26 32 47 62 48
310 200 116 216 106 5 -3 13 0 17 8 12 192 29 245 23 11 -2 48 -6 82 -92"/>
311<path d="M4140 1200 c6 -11 13 -20 16 -20 2 0 0 9 -6 20 -6 11 -13 20 -16 20
312 2 0 0 -9 6 -202"/>
313<path d="M3143 1289 c-140 -21 -297 -106 -363 -196 -98 -134 -70 -272 83 -480
314 84 -71 264 -123 423 -123 156 0 298 42 415 122 68 47 110 104 134 178 39 126
315 33 257 -188 340 -109 59 -190 81 -322 85 -66 2 -148 0 -182 -62 m274 -13 c17
316 0 29 -4 27 -7 -2 -4 -24 -12 57 -18 34 -6 78 -20 98 -31 20 -11 48 -26 64 -33
317 15 -6 25 -17 23 -24 -3 -7 0 -15 6 -19 7 -4 8 -3 4 4 -15 25 7 10 45 -29 21
318 -22 39 -50 39 -62 0 -16 3 -10 9 -8 14 21 33 -52 29 -109 -1 -26 0 -46 5 -43
319 11 6 9 -1 -12 -49 -21 -49 -108 -151 -122 -142 -5 3 -9 2 -9 -3 0 -14 -99 -64
320 168 -85 -96 -28 -283 -35 -384 -14 -95 20 -190 54 -201 72 -4 8 -16 14 -26
321 14 -20 0 -112 95 -136 139 -22 44 -30 113 -15 141 7 12 7 20 2 16 -6 -4 -8 2
322 5 14 3 11 9 20 14 20 5 0 9 9 10 9 11 5 23 12 27 7 4 8 3 4 -11 -18 0
323 -14 15 5 10 11 10 14 2 9 -30 -17 -9 12 36 54 44 40 72 57 171 105 35 17 113
324 33 195 40 38 4 71 7 72 8 1 1 25 0 55 -2 29 -2 67 -4 85 -52"/>
325<path d="M2310 470 c-9 -6 -10 -10 -3 -10 6 0 15 5 18 10 8 12 4 12 -15 02"/>
326<path d="M3978 453 c7 -3 16 -2 19 1 4 3 -2 6 -13 5 -11 0 -14 -3 -6 -62"/>
327</g>
328</svg>
```

# GPT5 and GPT OSS 120B



# NorwAI Magistral and Normistral

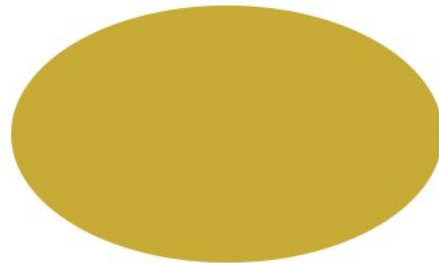
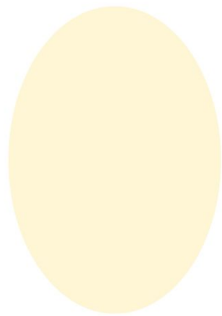
National



# Enter our Dataset

Initially planned to be part of a Kaggle competition

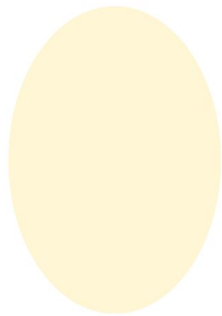
Simple shapes: Only ellipses, rectangles and polygons



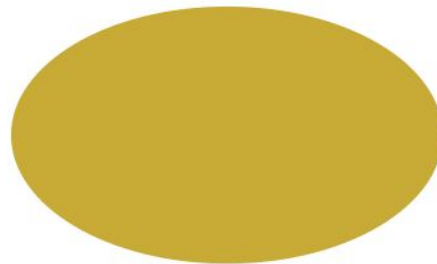
# Enter our Dataset

Initially planned to be part of a Kaggle competition

Simple shapes: Only ellipses, rectangles and polygons



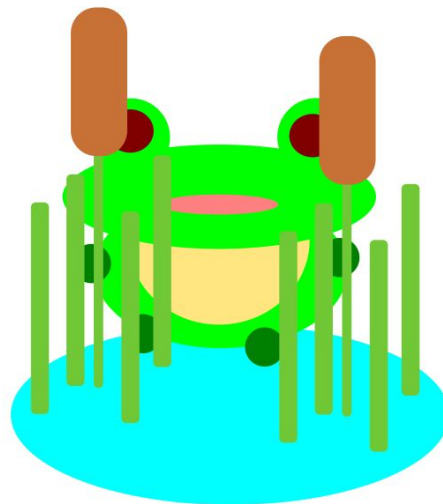
egg



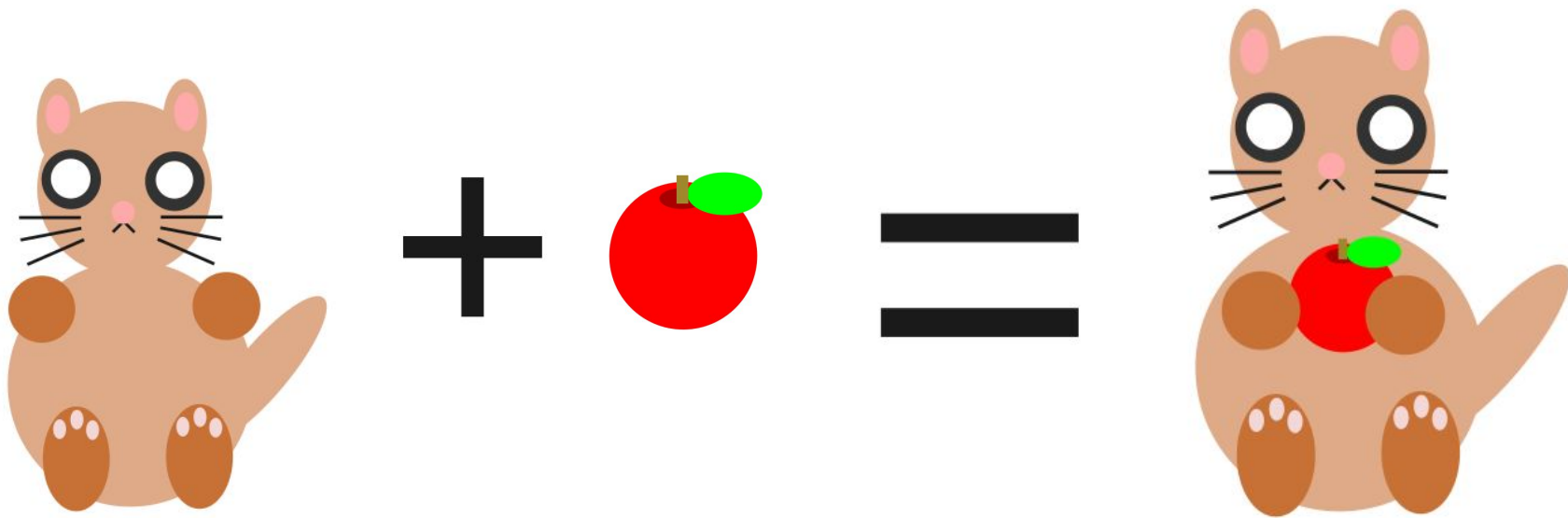
potato

# Artists in NLP

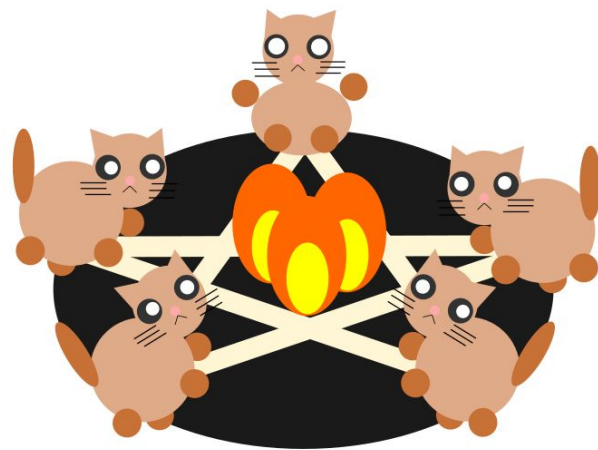
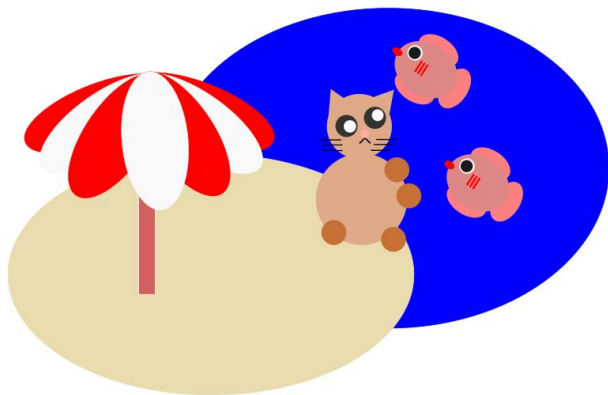
- No generated images
- 100% of content generators allow for model training and evaluation



# Compositionality - Somewhat Underutilized

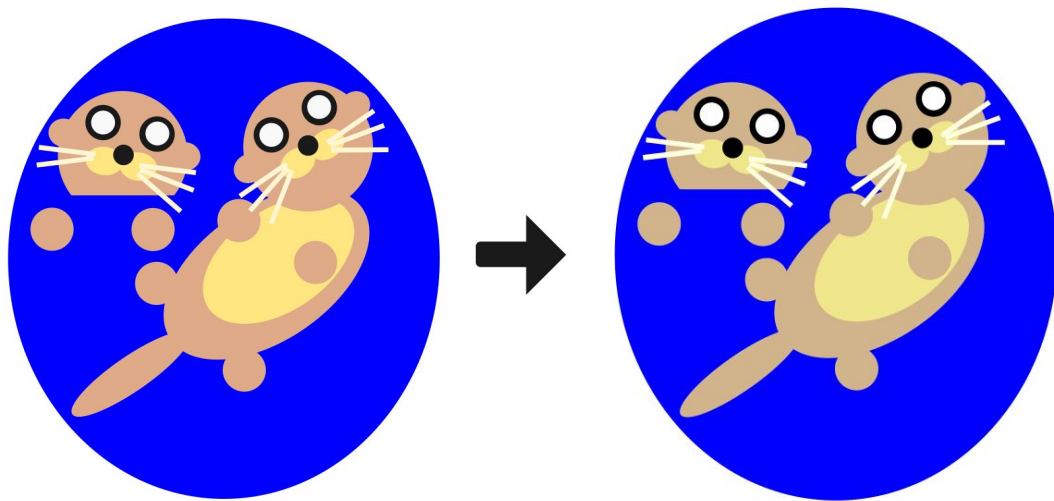
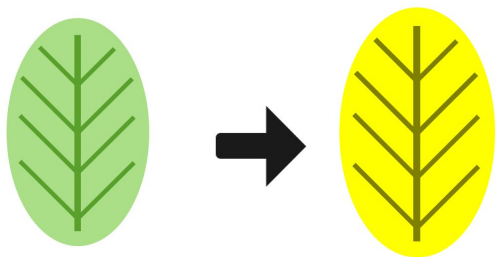


# A Myriad of Semantically Rich Scenes



# Normalization

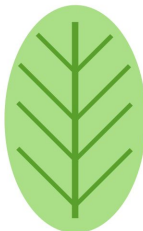
- For the test setup the dataset has undergone another normalization layer
- Quantization
- Color normalization
- Size standardization



# Non-normalized



```
1<?xml version="1.0" encoding="UTF-8" standalone="no"?>
2<!-- Created with Inkscape (http://www.inkscape.org/) -->
3
4<svg
5  width="210mm"
6  height="297mm"
7  viewBox="0 0 210 297"
8  version="1.1"
9  id="svg1"
10 inkscape:version="1.4.1 (1:1.4.1+202503302257+93de688d07)"
11 sodipodi:docname="leaf.svg"
12 xmlns:inkscape="http://www.inkscape.org/namespaces/inkscape"
13 xmlns:sodipodi="http://sodipodi.sourceforge.net/DTD/sodipodi-0.dtd"
14 xmlns:svg="http://www.w3.org/2000/svg">
15 <sodipodi:namedview
16   id="namedview1"
17   pagecolor="#ffffff"
18   bordercolor="#666666"
19   borderopacity="1.0"
20   inkscape:showpageshadow="2"
21   inkscape:pageopacity="0.0"
22   inkscape:pagecheckerboards="0"
23   inkscape:deskcolors="#d1d1d1"
24   inkscape:document-units="mm"
25   inkscape:zoom="0.52473017"
26   inkscape:cx="596.49705"
27   inkscape:cy="549.80639"
28   inkscape:window-width="1920"
29   inkscape:window-height="1016"
30   inkscape:window-x="0"
31   inkscape:window-y="0"
32   inkscape:window-maximized="1"
33   inkscape:current-layer="layer1" />
34 <defs
35   id="defs1" />
36 <g
37   inkscape:label="Layer 1"
38   inkscape:groupmode="layer"
39   id="layer1">
40 <ellipse
```

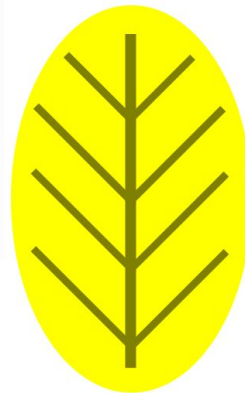


```
41 <ellipse
42   style="fill:#aade87;stroke-width:0.457186"
43   id="path1"
44   cx="105.69531"
45   cy="144.35956"
46   rx="74.429611"
47   ry="119.94855" />
48 <rect
49   style="fill:#5aa02c;stroke-width:0.457186"
50   id="rect1"
51   width="6.7663283"
52   height="206.37302"
53   x="102.00459"
54   y="42.249538" />
55 <rect
56   style="fill:#5aa02c;stroke-width:0.457186"
57   id="rect2"
58   width="4.6134062"
59   height="54.438187"
60   x="5.1508206"
61   y="85.440399"
62   transform="rotate(-45)" />
63 <rect
64   style="fill:#5aa02c;stroke-width:0.563614"
65   id="rect2-9-4"
66   width="4.6134062"
67   height="82.733742"
68   x="-30.411934"
69   y="94.656082"
70   transform="rotate(-45)" />
71 <rect
72   style="fill:#5aa02c;stroke-width:0.563614"
73   id="rect2-9-4-1"
74   width="4.6134062"
75   height="82.733742"
76   x="-60.187798"
77   y="121.86519"
78   transform="rotate(-45)" />
79 <rect
80   style="fill:#5aa02c;stroke-width:0.563614"
81   id="rect2-9-4-1-2"
```

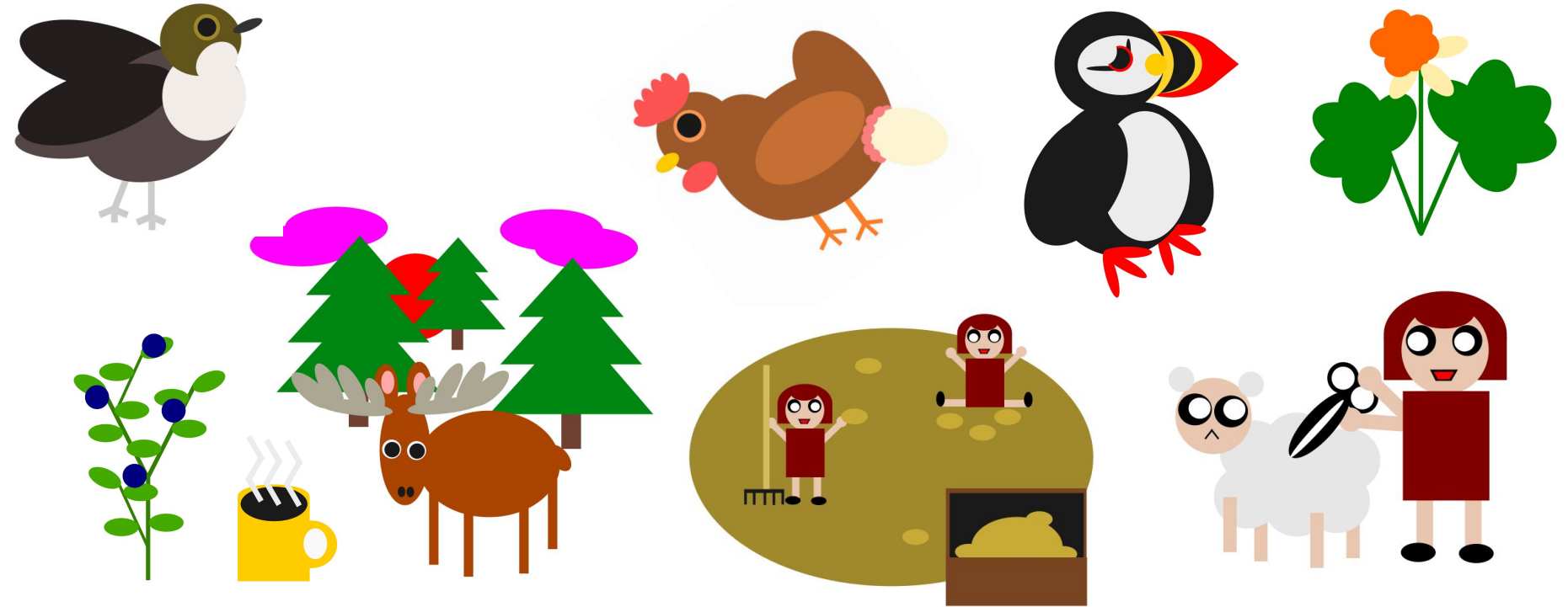
```
80   style="fill:#5aa02c;stroke-width:0.563614"
81   id="rect2-9-4-1-2"
82   width="4.6134062"
83   height="82.733742"
84   x="-94.113365"
85   y="155.79465"
86   transform="rotate(-45)" />
87 <rect
88   style="fill:#5aa02c;stroke-width:0.457186"
89   id="rect2-7"
90   width="4.6134062"
91   height="54.438187"
92   x="-144.03648"
93   y="-61.649498"
94   transform="matrix(-0.70710678,-0.70710678,0.70710678,0.70710678,0,0)" />
95 <rect
96   style="fill:#5aa02c;stroke-width:0.563614"
97   id="rect2-9-4-4"
98   width="4.6134062"
99   height="82.733742"
100   x="-179.59843"
101   y="-52.433815"
102   transform="matrix(-0.70710678,-0.70710678,0.70710678,0.70710678,0,0)" />
103 <rect
104   style="fill:#5aa02c;stroke-width:0.563614"
105   id="rect2-9-4-1-1"
106   width="4.6134062"
107   height="82.733742"
108   x="-209.37428"
109   y="-25.224728"
110   transform="matrix(-0.70710678,-0.70710678,0.70710678,0.70710678,0,0)" />
111 <rect
112   style="fill:#5aa02c;stroke-width:0.563614"
113   id="rect2-9-4-1-2-8"
114   width="4.6134062"
115   height="82.733742"
116   x="-243.29987"
117   y="8.7047386"
118   transform="matrix(-0.70710678,-0.70710678,0.70710678,0.70710678,0,0)" />
119 </g>
120 </svg>
```

# Normalized

```
1 <svg xmlns="http://www.w3.org/2000/svg" width="1000" height="1000">
2   <ellipse cx="500" cy="499" rx="310" ry="499" fill="yellow" />
3   <rect x="484" y="74" width="28" height="860" fill="olive" />
4   <rect x="403" y="94" width="20" height="227" fill="olive" transform="rotate(-45 413 208)" />
5   <rect x="367" y="209" width="20" height="345" fill="olive" transform="rotate(-45 377 382)" />
6   <rect x="360" y="377" width="19" height="345" fill="olive" transform="rotate(-45 369 549)" />
7   <rect x="360" y="577" width="19" height="345" fill="olive" transform="rotate(-45 369 749)" />
8   <rect x="569" y="101" width="19" height="226" fill="olive" transform="rotate(-135 578 214)" />
9   <rect x="604" y="215" width="20" height="345" fill="olive" transform="rotate(-135 614 388)" />
10  <rect x="612" y="383" width="19" height="345" fill="olive" transform="rotate(-135 622 556)" />
11  <rect x="612" y="583" width="19" height="345" fill="olive" transform="rotate(-135 622 756)" />
12 </svg>
```

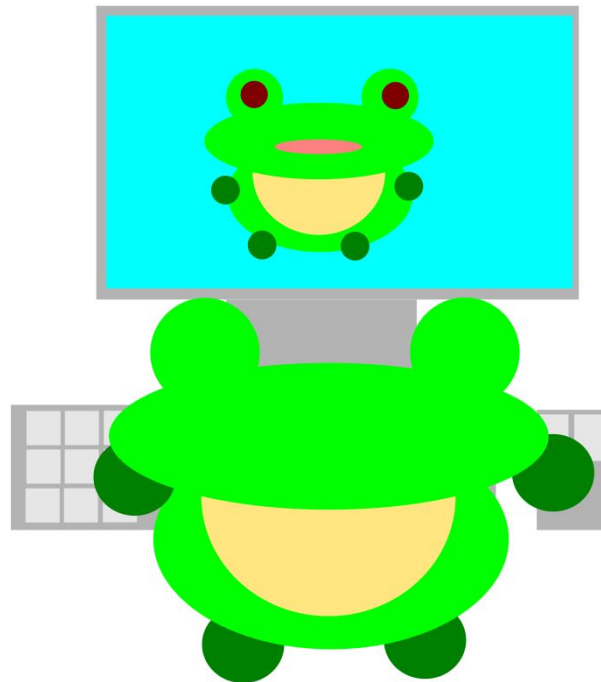


# Some Cultural Bias



# Initial investigations

- Closed problem: 4 options
- Some perturbations and challenge datasets



# Human baseline

- 7 annotators, no native speakers
- Average score  $98 \pm 1$

## SVG Caption Annotation

Log in

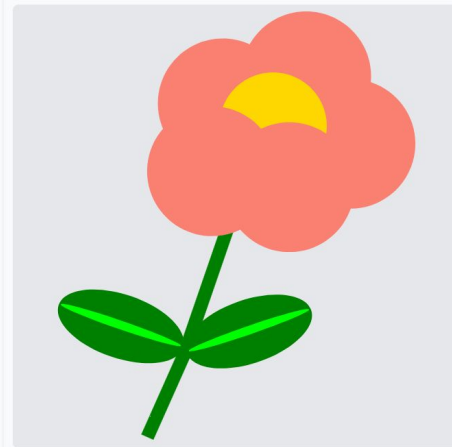
Enter your unique annotator ID to begin

Annotator ID

Login

Progress: 3/200

Which of the four captions best describes the following SVG image?



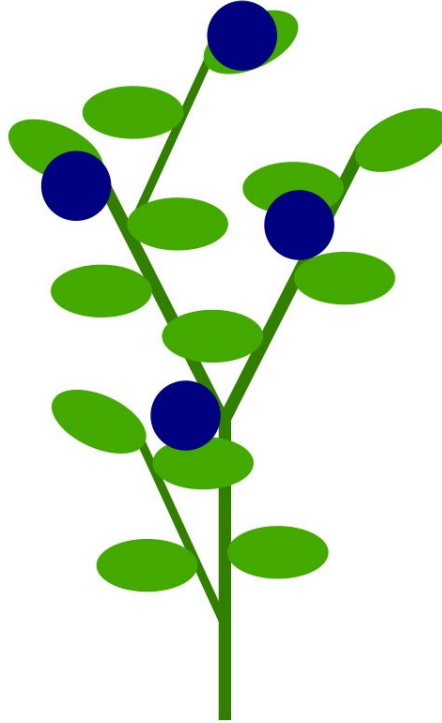
A: an apple tree

B: a flower

C: a blossoming tree

D: a peach tree

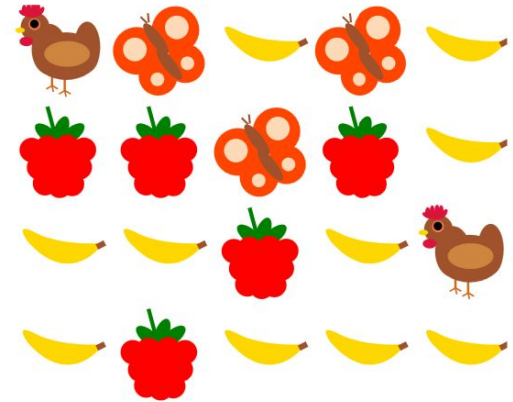
# The Humble Bilberry Causing some Issues



# Expansion 1 - Counting

- How many butterflies are there?
- Grid based - main object is entered first

(a) Counting objects



## Expansion 2 - Relative positions

- What is to the right of the dog
- Reference object is always unique
- The remainder is populated in a grid

**(b) Relative positions**



# Expansion 3 - Adversary colors

- A subset of the SVGs have a dedicated color mapping
- This mapping is easy to manipulate

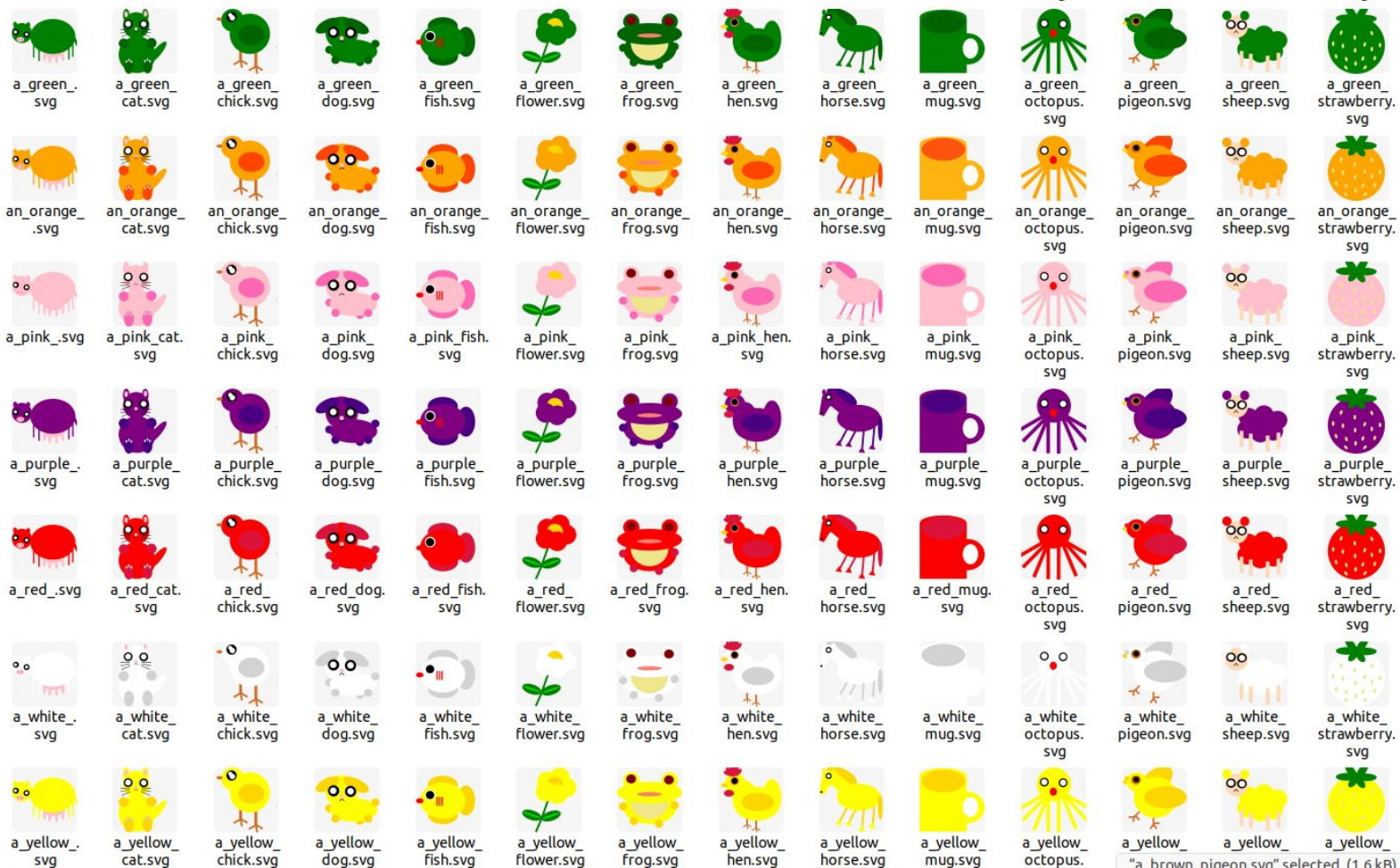
```

10 colorremap = {"black":{"mono":"black","light":"darkgray","dark":"black"},
11               "white":{"mono":"white","light":"white","dark":"lightgray"},
12               "gray":{"mono":"gray","light":"lightgray","dark":"gray"},
13               "gray":{"mono":"gray","light":"lightgray","dark":"gray"},
14               "brown":{"mono":"gray","light":"lightgray","dark":"gray"},
15               "red":{"mono":"red","light":"red","dark":"crimson"},
16               "orange":{"mono":"gray","light":"lightgray","dark":"gray"},
17               "yellow":{"mono":"gray","light":"lightgray","dark":"gray"},
18               "green":{"mono":"gray","light":"lightgray","dark":"gray"},
19               "blue":{"mono":"gray","light":"lightgray","dark":"gray"},
20               "pink":{"mono":"gray","light":"lightgray","dark":"gray"},
21               "purple":{"mono":"purple","light":"lightgray","dark":"gray"}}

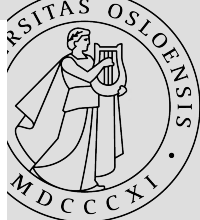
```

(c) Adversary colors





"a\_brown\_pigeon.svg" selected (1,6 kB)



# Main Task

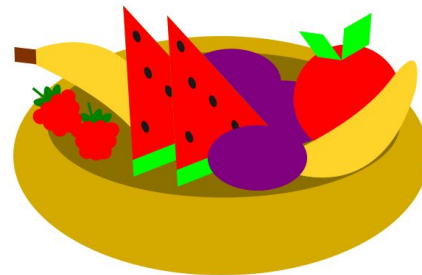
Which of the four captions best describes the following SVG image? Respond only with `[[A]]`, `[[B]]`, `[[C]]` or `[[D]]`.

**## Image**

```
```svg
{image}
```
```

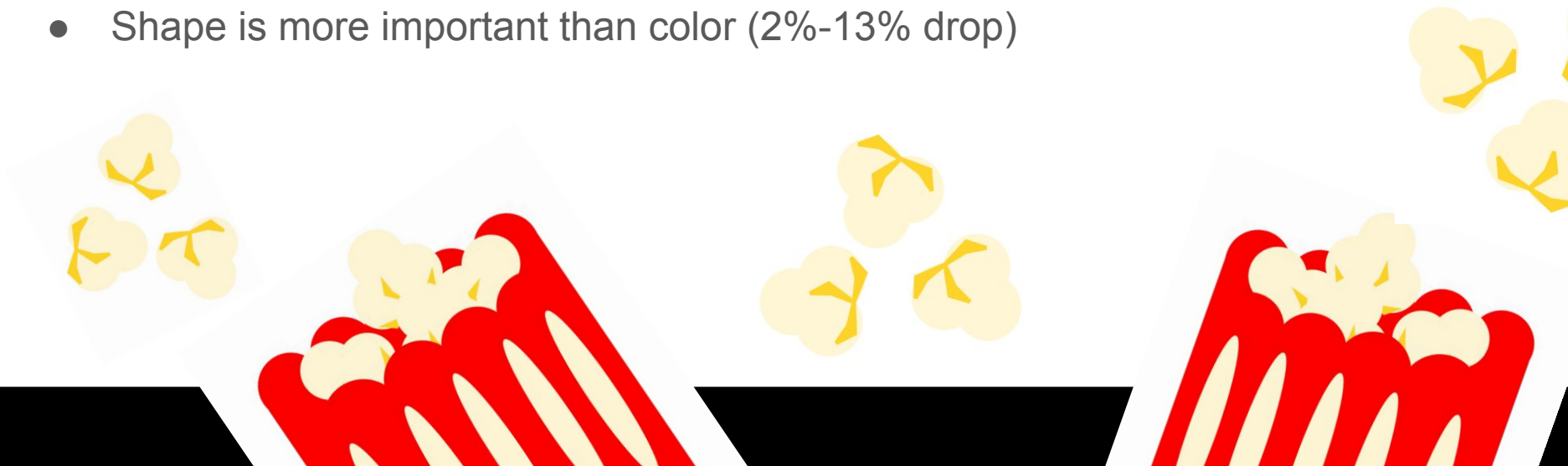
**## Captions**

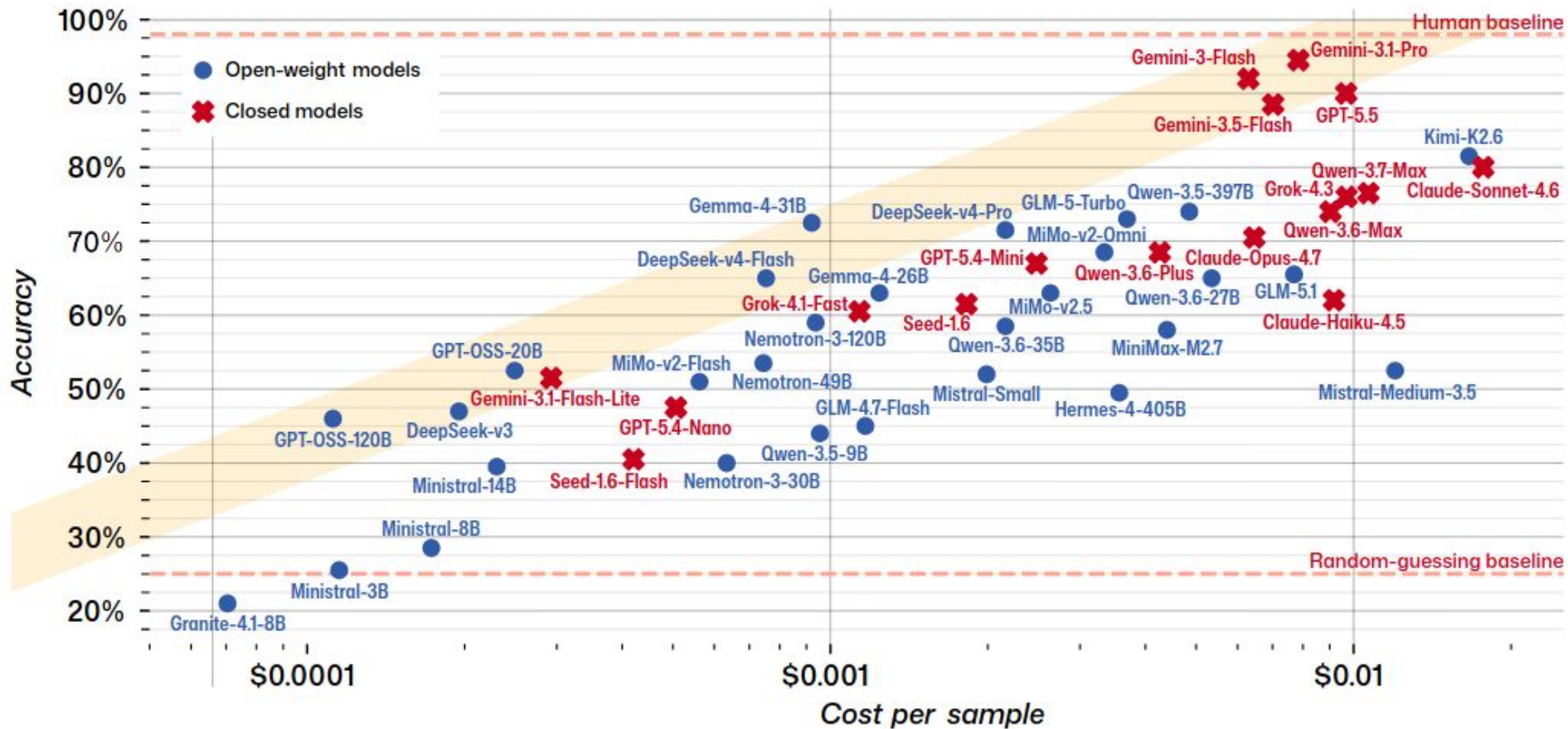
- A) {caption\_1}
- B) {caption\_2}
- C) {caption\_3}
- D) {caption\_4}



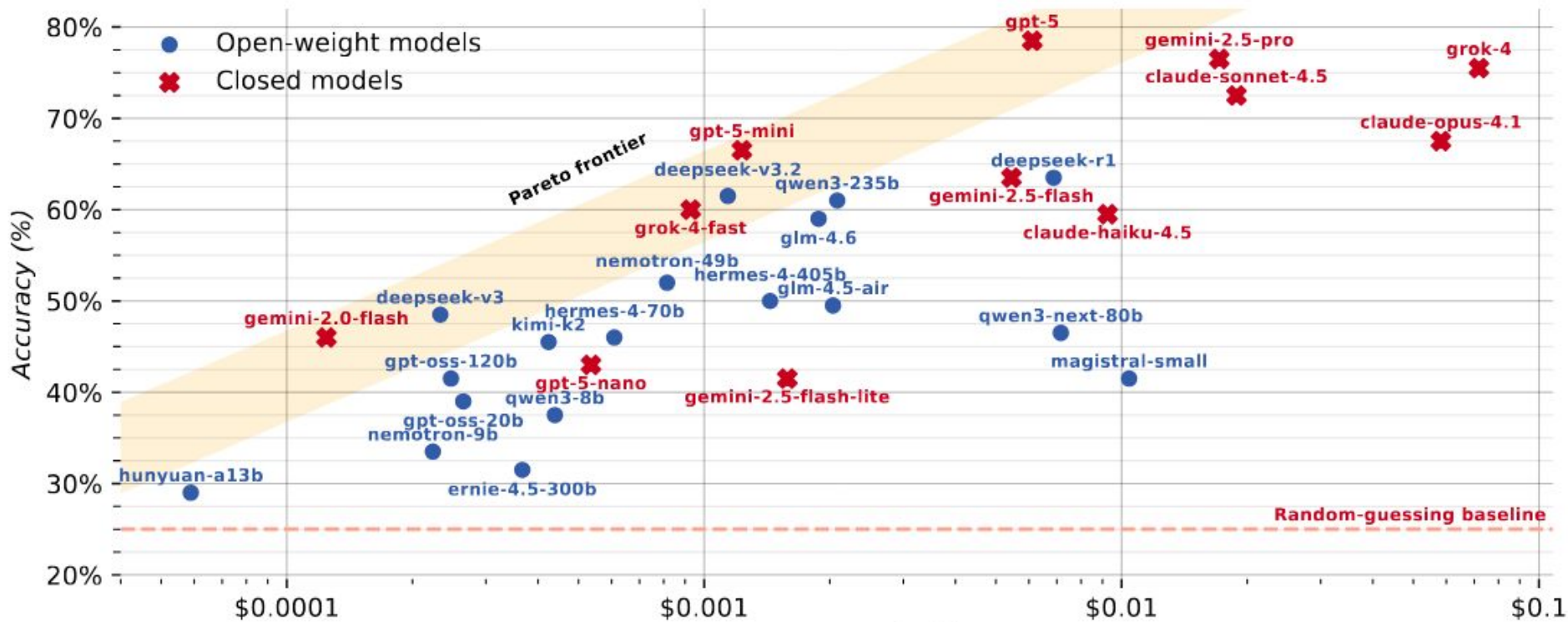
# Experiments

- Closed-weight models
- Open-weight models
- Reasoning seems efficient where it is comparable - not simple pattern matching
- Shape is more important than color (2%-13% drop)



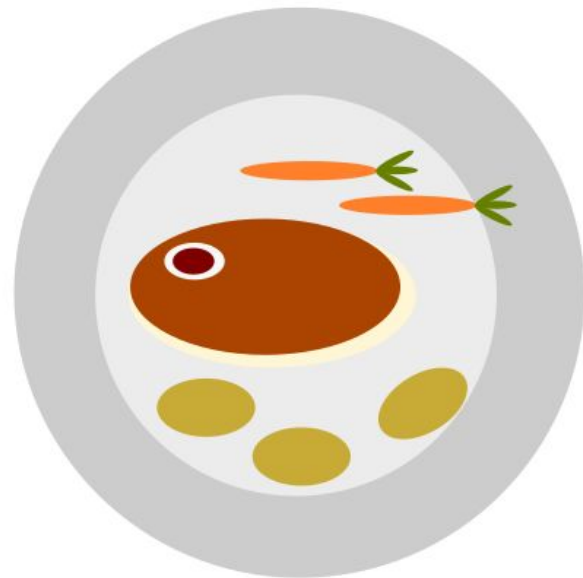


# But one year ago ...



# Results

- GPT5.5 Not far from Human baseline
- Many frontier models are below 80%
- Best open-weight model is Kimi K2.6 - 81-5



Most misclassified image

# Ablation 1

| Model             | full<br>scores | without<br>reasoning | without<br>colors   | randomized<br>coordinates |
|-------------------|----------------|----------------------|---------------------|---------------------------|
| Gemini 3 Flash    | 92.0           | 78.5 <i>(-13.5)</i>  | 80.5 <i>(-11.5)</i> | 38.5 <i>(-53.5)</i>       |
| GPT 5.5           | 90.0           | 79.5 <i>(-10.5)</i>  | 86.0 <i>(-4.0)</i>  | 36.5 <i>(-53.5)</i>       |
| Kimi K2.6         | 81.5           | 57.5 <i>(-24.0)</i>  | 68.5 <i>(-13.0)</i> | 39.5 <i>(-42.0)</i>       |
| Claude Sonnet 4.6 | 80.0           | 71.0 <i>(-9.0)</i>   | 70.5 <i>(-9.5)</i>  | 39.5 <i>(-40.5)</i>       |
| Gemma 4 31B       | 72.5           | 60.5 <i>(-12.0)</i>  | 66.5 <i>(-6.0)</i>  | 38.0 <i>(-34.5)</i>       |
| DeepSeek v3       | 47.0           | 44.0 <i>(-7.0)</i>   | 45.0 <i>(-2.0)</i>  | 33.0 <i>(-14.0)</i>       |

## Ablation 2

| <b>Model</b>      | <b>object<br/>counting</b> | <b>positional<br/>relations</b> | <b>adversary<br/>colors</b> |
|-------------------|----------------------------|---------------------------------|-----------------------------|
| Gemini 3 Flash    | 87.0                       | 94.4                            | 97.4                        |
| GPT 5.5           | 100.0                      | 100.0                           | 99.4                        |
| Kimi K2.6         | 39.5                       | 100.0                           | 96.7                        |
| Claude Sonnet 4.6 | 85.2                       | 88.8                            | 89.6                        |
| Gemma 4 31B       | 67.9                       | 50.0                            | 89.0                        |
| DeepSeek v3       | 29.0                       | 50.0                            | 57.8                        |

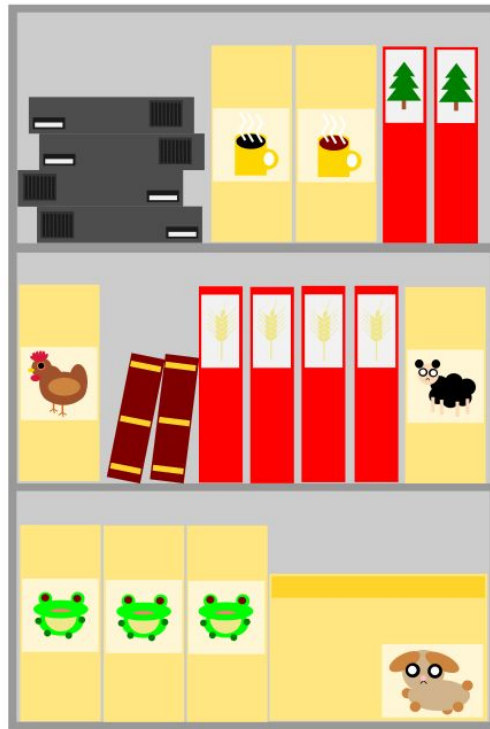
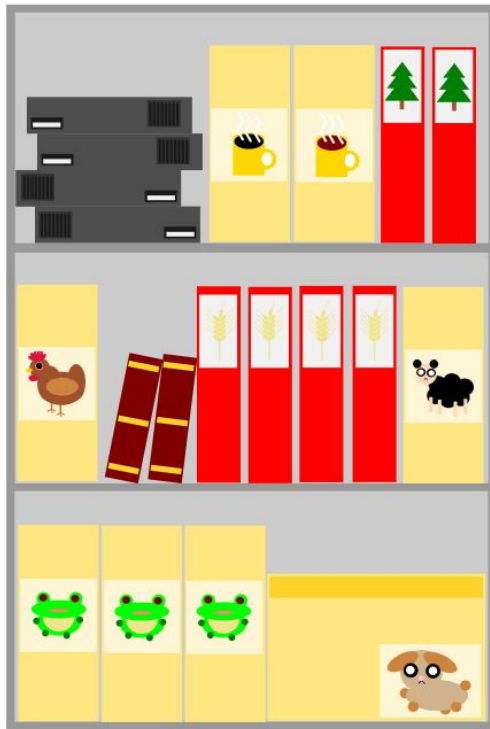
# Ablation



| <b>base model</b> | <b>text-only<br/>model</b> | <b>visual-language<br/>model</b> |
|-------------------|----------------------------|----------------------------------|
| GLM-5-Turbo       | 73.0                       | 71.0                             |
| Qwen3-235B-A22B   | 57.5                       | 60.5                             |
| Qwen3-30B-A3B     | 46.0                       | 47.0                             |
| Qwen3-8B          | 36.0                       | 37.5                             |

# Future additions

- Expand dataset
- Multilinguality
- More detailed error analysis
- Master's project



Thank you!

