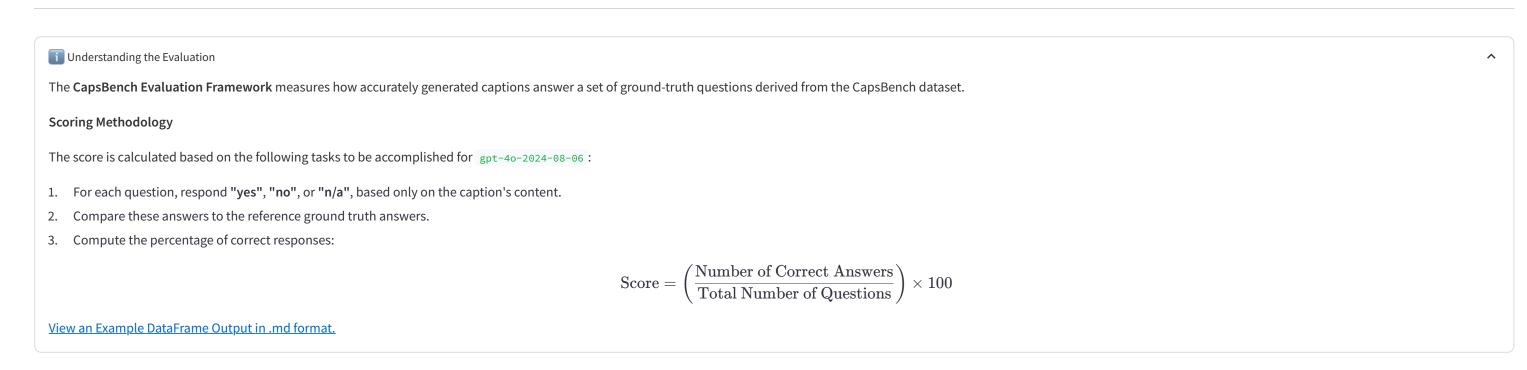




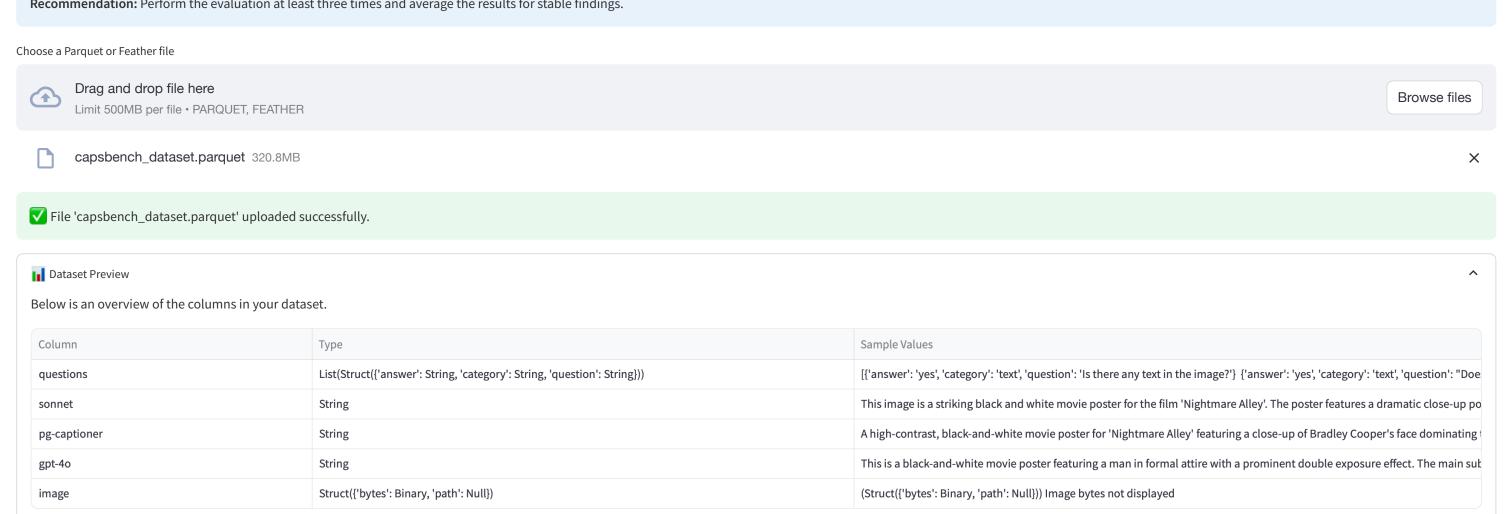
A Benchmark for evaluating the accuracy of synthetic captions against the CapsBench dataset.



Upload the CapsBench Dataset

You can download .parquet or .feather files from https://example.com/huggingFace/playgroundai/CapsBench.

Recommendation: Perform the evaluation at least three times and average the results for stable findings.



Processing Parameters

Advanced Configuration		^
Image Bytes Column	Questions Column	②
image	questions	
Caption Column	③ Start Index	②
pg-captioner	0	- +
Output File Path	② End Index (not inclusive)	②
results_gpt-4o-2024-08-06_pg-captioner_iter_1parquet	200	- +
	Step Size	③
	1	- +

Processing: 100%

▼ Processing complete! Results saved to 'results_gpt-4o-2024-08-06_pg-captioner_iter_1_.parquet'.

□ → **II** CapsBench Accuracy

CapsBench Accuracy (%)

76.59 **↑** 76.59%

Snapshot of Evaluation Results

Below are the first 10 evaluated entries:

candidate_caption	captioned_by	ground_truth	openai_answers	score	token_count
A surreal 3D digital illustration features a woman's face in an avant-garde compositio	pg-captioner	[{'answer': 'yes', 'category': 'color', 'question': 'Is the color of the butterflies primarily l	1. Is the color of the butterflies primarily blue? - yes 2. Is the face mask on the right	66.67	357
A vintage-style coffee label features an ornate frame with goldenrod (#d4a964) text or	pg-captioner	[{'answer': 'yes', 'category': 'text', 'question': 'Is there any text in the image?'} {'answe	1. Is there any text in the image? - yes 2. Does the text include the word 'COFFEE'?	90.91	390
A surreal digital illustration features a central female figure with long, straight teal ha	pg-captioner	[{'answer': 'yes', 'category': 'color', 'question': "Is the hair's color light green?"} {'answ	1. Is the hair's color light green? - no 2. Is the character located in the center of the	72.73	457
Three men are seated at a black table covered with a dark blue cloth, engaged in wha	pg-captioner	[{'answer': 'yes', 'category': 'text', 'question': 'Is there any text in the image?'} {'answe	1. Is there any text in the image? - yes 2. Does the image contain the text 'AKSHAY'	68.18	370
A close-up photograph captures an intimate moment between two individuals seated	pg-captioner	[{'answer': 'Yes', 'category': 'color', 'question': 'Is the background color off-white?'} {'a	1. Is the background color off-white? - yes 2. Is the person with the white bow tie o	72.73	425
A high-contrast, black-and-white movie poster for 'Nightmare Alley' featuring a close-	pg-captioner	[{'answer': 'yes', 'category': 'text', 'question': 'Is there any text in the image?'} {'answe	1. Is there any text in the image? - yes 2. Does the image contain the text 'NIGHTMA	86.67	334
A high-angle aerial photograph captures an intense nighttime scene of a large bonfire	pg-captioner	[{'answer': 'yes', 'category': 'color', 'question': 'ls the fire orange?'} {'answer': 'yes', 'ca	1. Is the fire orange? - yes 2. Is the fire in the center of the image? - n/a 3. Are the	84.62	386
A promotional graphic featuring a vibrant contrast between a bright red background	pg-captioner	[{'answer': 'yes', 'category': 'text', 'question': 'Is there any text in the image?'} {'answe	1. Is there any text in the image? - yes 2. Does the image contain the text 'MOTIVATI	72.22	399
A cozy kitchen scene with an emphasis on vintage and rustic elements, featuring a wh	pg-captioner	[{'answer': 'yes', 'category': 'text', 'question': 'Is there any text in the image?'} {'answe	1. Is there any text in the image? - yes 2. Does the image contain the text 'LINEN'?	75	280
A movie poster for 'The Secret Life of Pets' features a close-up view of a brown and w	pg-captioner	[{'answer': 'yes', 'category': 'text', 'question': 'Is there any text in the image?'} {'answe	1. Is there any text in the image? - yes 2. Does the image contain the text 'JULY 8'?	82.35	421

Token Metrics

Average Token Length Vocabulary Size 6456.00 357.04 **↑** 357.04 **↑** 6456.00

Total Number of entries successfully processed for CapsBench Accuracy: 200

▼ Evaluation completed successfully.

Developed by <u>alexferdg</u> | Enhanced by the community