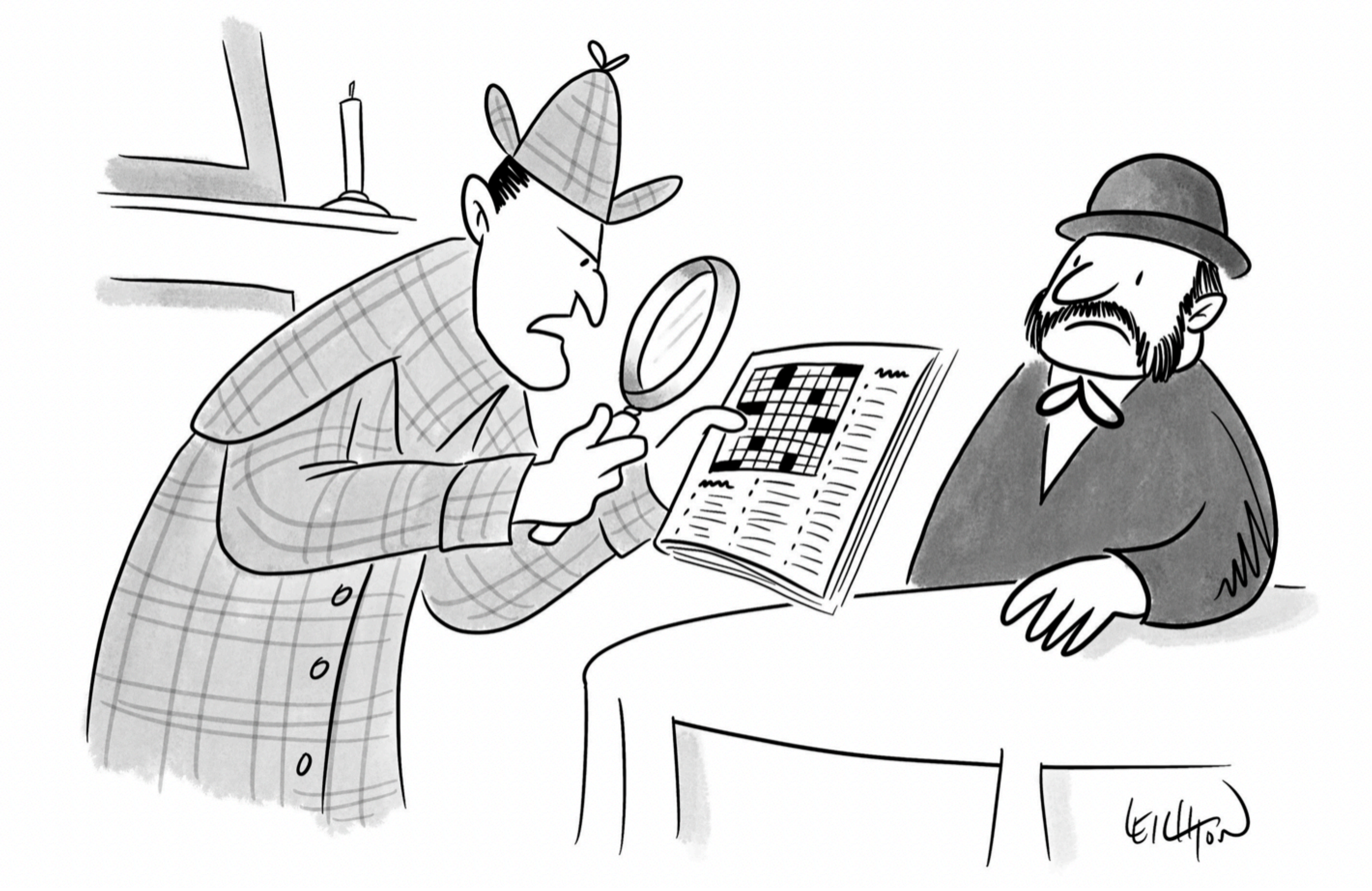
Humor in AI: Massive Scale Crowd-Sourced Preferences and Benchmarks for Cartoon Captioning

Jifan Zhang*, Lalit Jain*, Yang Guo*, Jiayi Chen†, Kuan Lok Zhou†,
Siddharth Suresh, Andrew Wagenmaker, Scott Sievert,
Timothy Rogers, Kevin Jamieson, Robert Mankoff, Robert Nowak

THE NEW YORKER

CARTOON CAPTION CONTEST

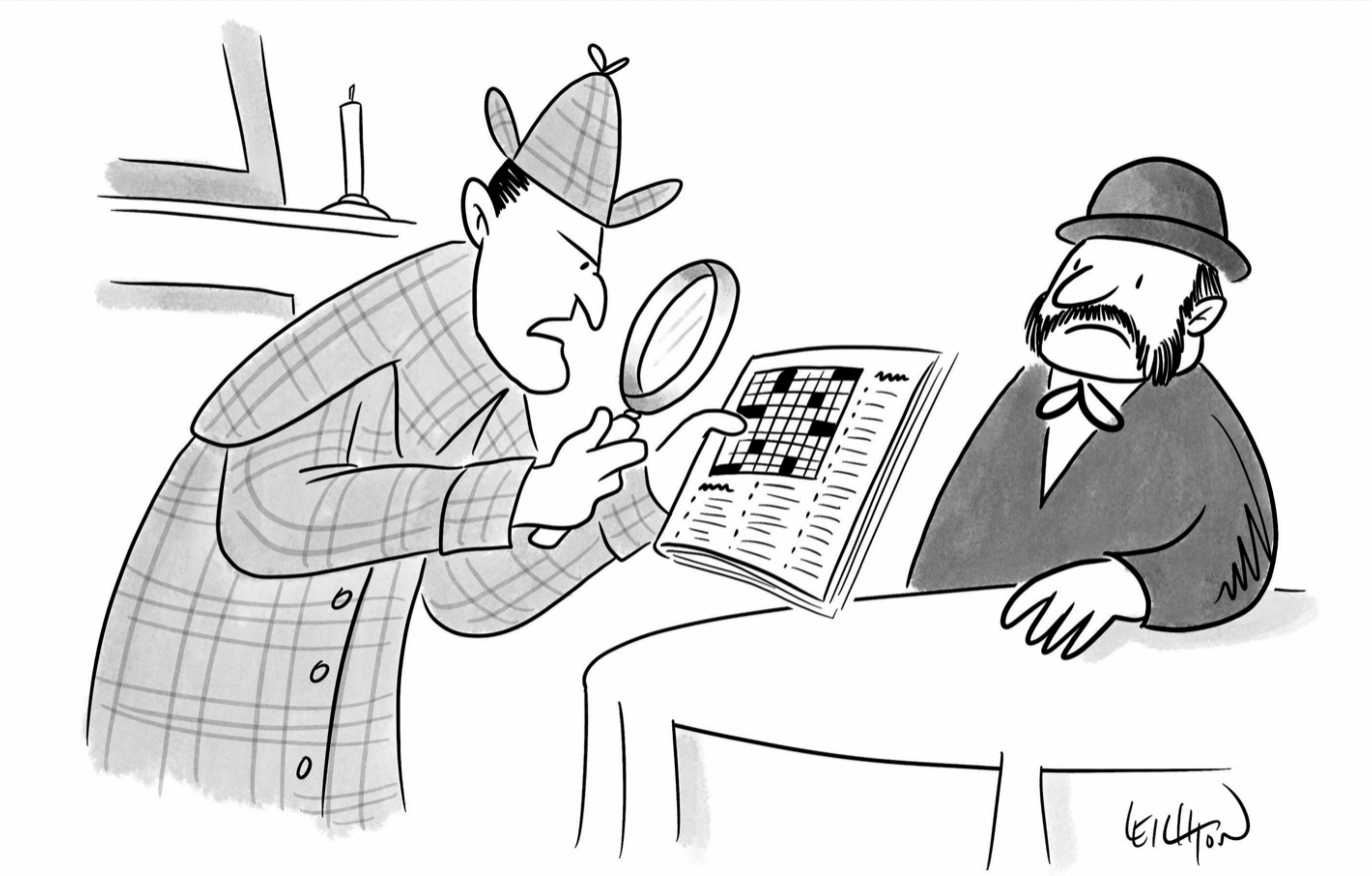
THIS WEEK'S CONTEST | THE FINALISTS | WINNING CAPTION



THE NEW YORKER

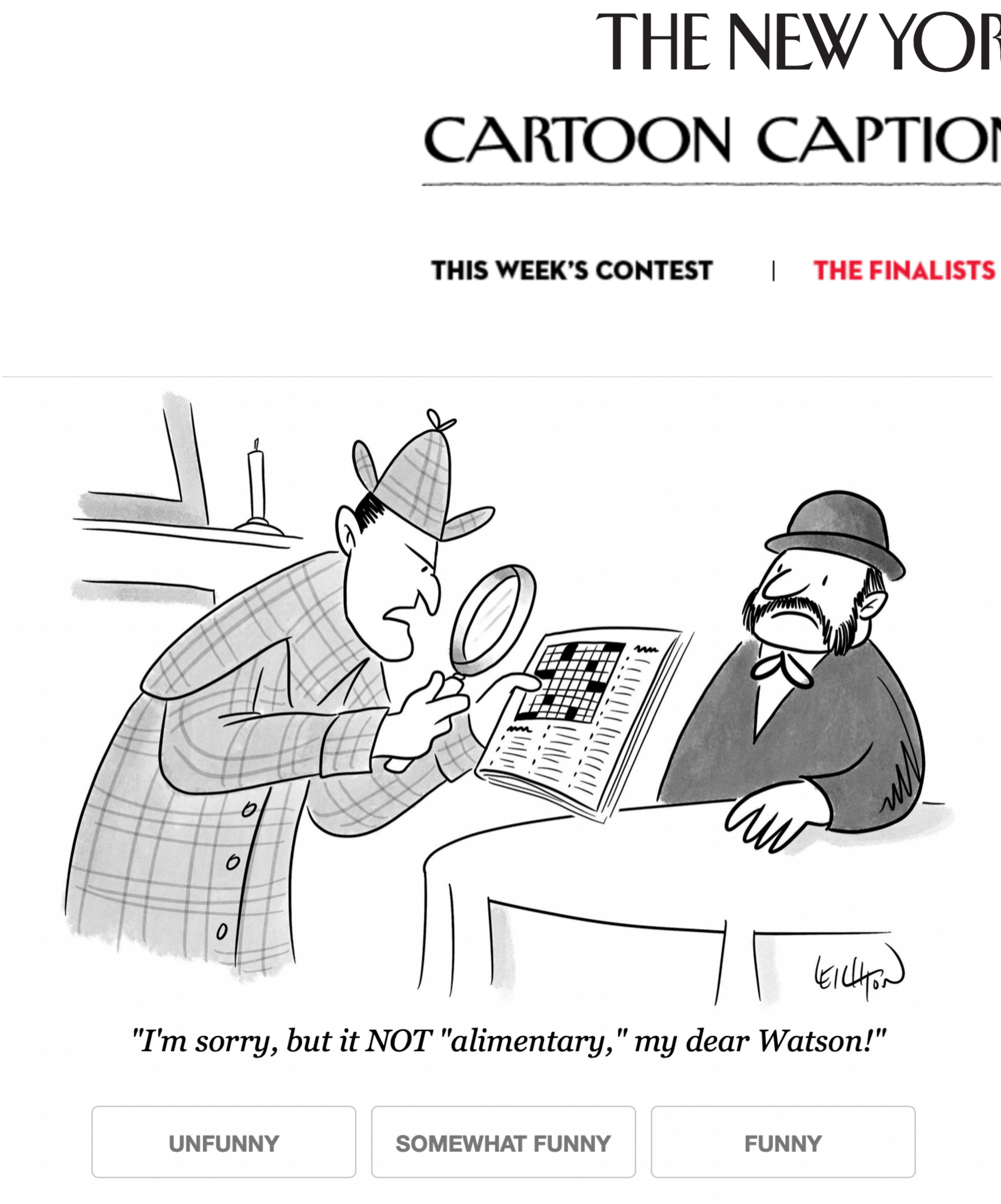
CARTOON CAPTION CONTEST

THIS WEEK'S CONTEST | THE FINALISTS | WINNING CAPTION



"I'm sorry, but it NOT "alimentary," my dear Watson!"

- ~7000 captions per week



THE NEW YORKER

CARTOON CAPTION CONTEST

THIS WEEK'S CONTEST | THE FINALISTS | WINNING CAPTION

"I'm sorry, but it NOT "alimentary," my dear Watson!"

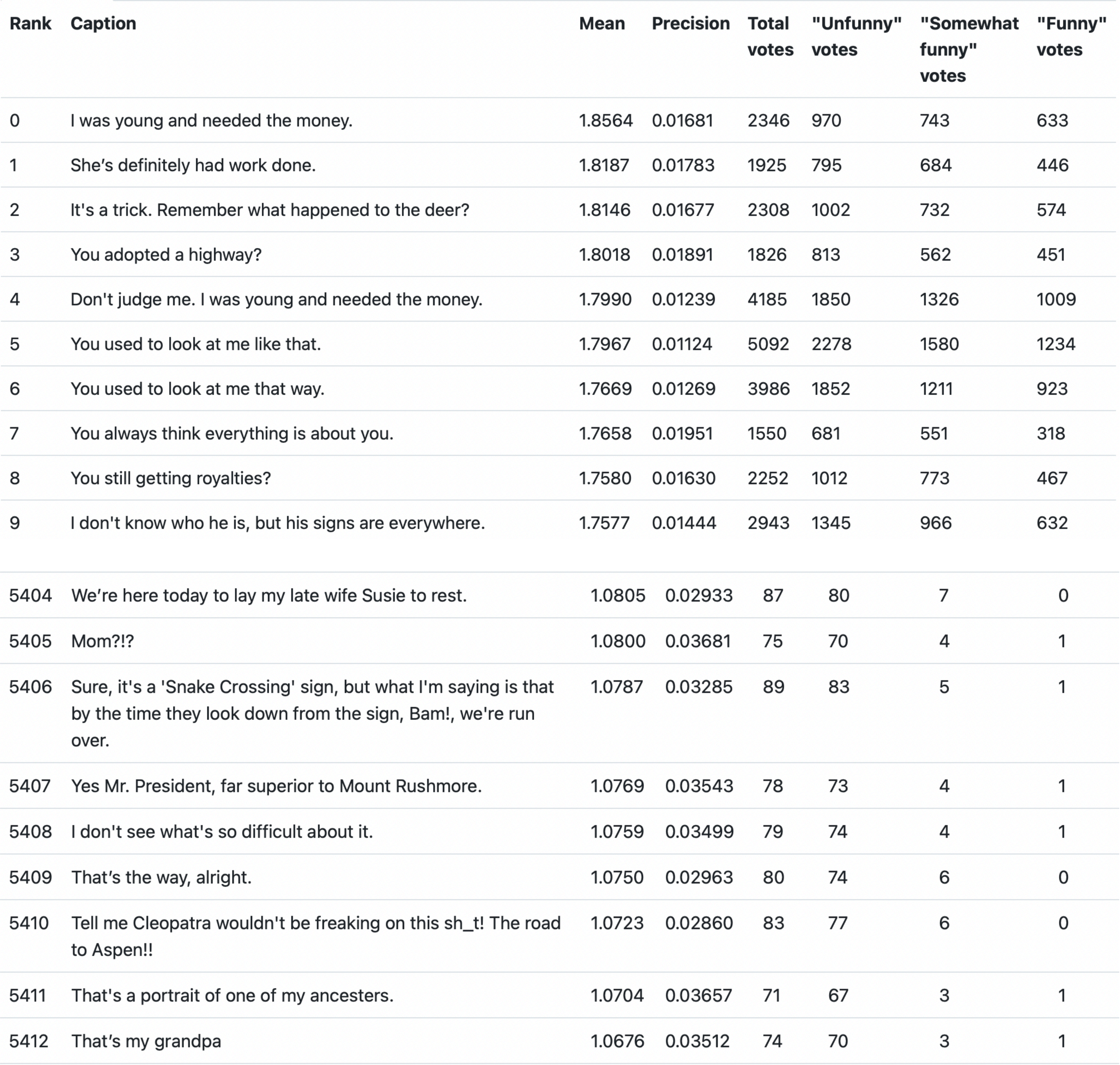
- ~7000 captions per week
- 500K - 1M ratings per week
- 300+ contests, 2.2M+ captions, 250M+ ratings (8 years)
- License: cc-by-nc 4.0 for research only

Ranking Collected by Multi-Armed Bandit Algorithm

Identify the funniest caption

Statistically significant captions receive 1-5K ratings.

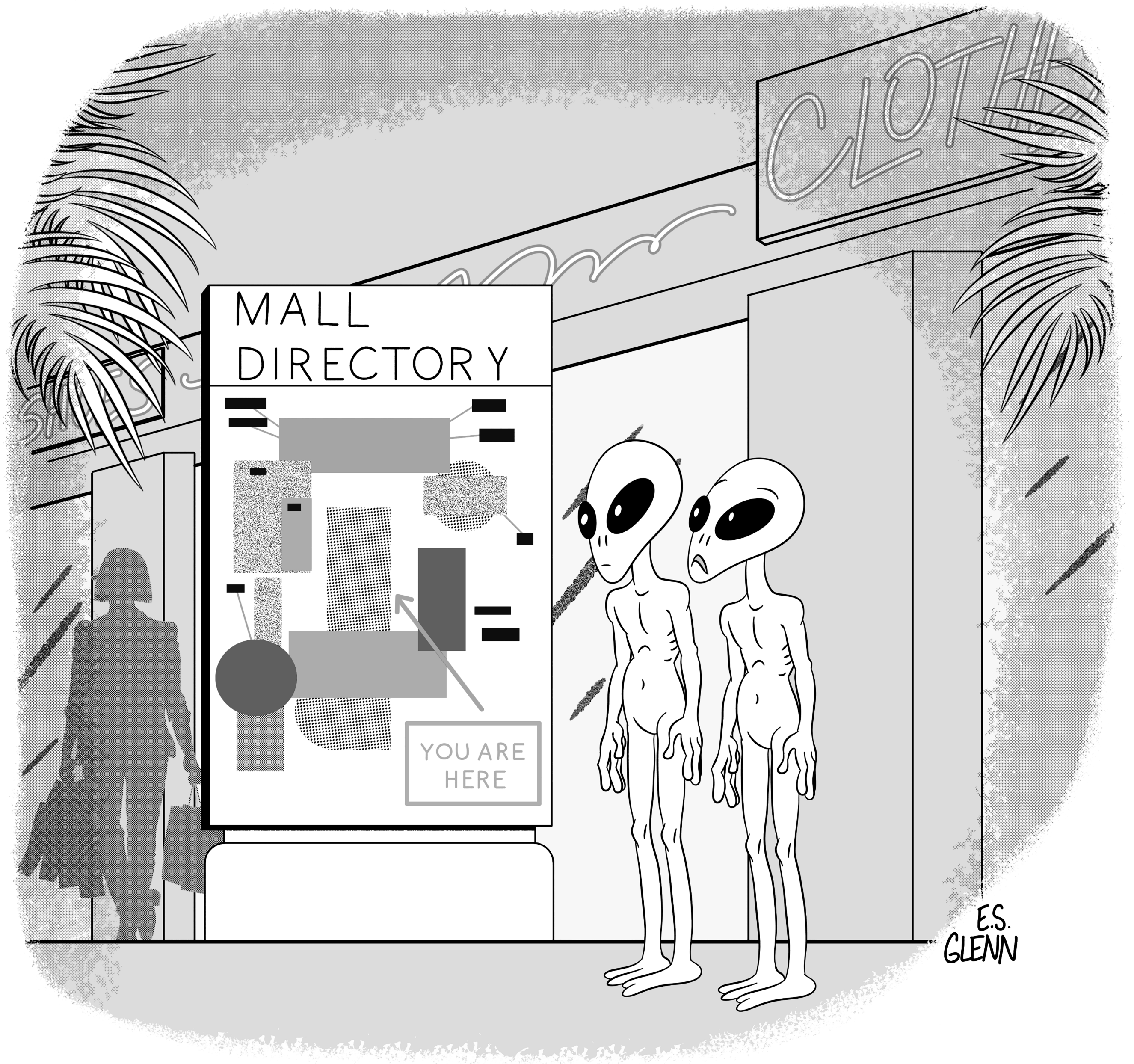
Captions unlikely to be the best receive <100 ratings.



Rank	Caption	Mean	Precision	Total votes	"Unfunny" votes	"Somewhat funny" votes	"Funny" votes
0	I was young and needed the money.	1.8564	0.01681	2346	970	743	633
1	She’s definitely had work done.	1.8187	0.01783	1925	795	684	446
2	It's a trick. Remember what happened to the deer?	1.8146	0.01677	2308	1002	732	574
3	You adopted a highway?	1.8018	0.01891	1826	813	562	451
4	Don't judge me. I was young and needed the money.	1.7990	0.01239	4185	1850	1326	1009
5	You used to look at me like that.	1.7967	0.01124	5092	2278	1580	1234
6	You used to look at me that way.	1.7669	0.01269	3986	1852	1211	923
7	You always think everything is about you.	1.7658	0.01951	1550	681	551	318
8	You still getting royalties?	1.7580	0.01630	2252	1012	773	467
9	I don't know who he is, but his signs are everywhere.	1.7577	0.01444	2943	1345	966	632
5404	We’re here today to lay my late wife Susie to rest.	1.0805	0.02933	87	80	7	0
5405	Mom?!?	1.0800	0.03681	75	70	4	1
5406	Sure, it's a 'Snake Crossing' sign, but what I'm saying is that by the time they look down from the sign, Bam!, we're run over.	1.0787	0.03285	89	83	5	1
5407	Yes Mr. President, far superior to Mount Rushmore.	1.0769	0.03543	78	73	4	1
5408	I don't see what's so difficult about it.	1.0759	0.03499	79	74	4	1
5409	That’s the way, alright.	1.0750	0.02963	80	74	6	0
5410	Tell me Cleopatra wouldn't be freaking on this sh_t! The road to Aspen!!	1.0723	0.02860	83	77	6	0
5411	That's a portrait of one of my ancesters.	1.0704	0.03657	71	67	3	1
5412	That’s my grandpa	1.0676	0.03512	74	70	3	1

https://nextml.github.io/caption-contest-data/
https://huggingface.co/datasets/yguooo/newyorker_caption_ranking

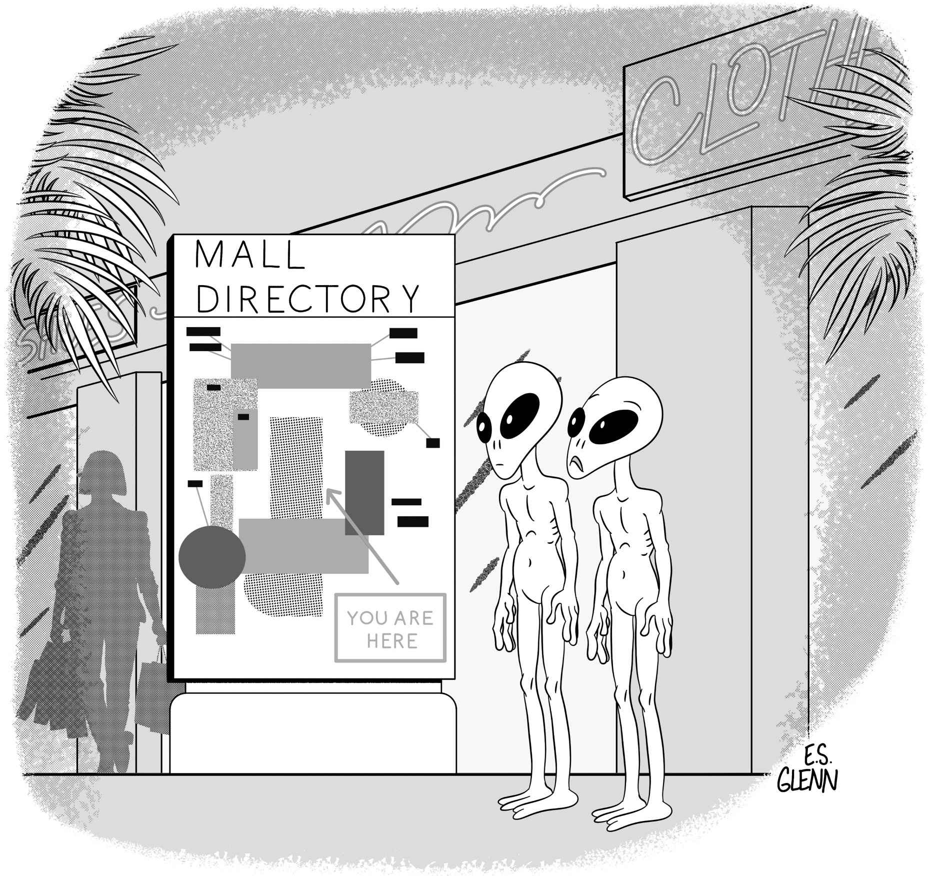
How well do LLMs generate funny captions?



GPT-4o Vision

“We travel light-years, and we still need directions”

Automated Evaluation using LLMs



Human Submission

1. Do you think death rays would be considered electronics or sporting goods?
2. Oh sure, now you look at a map.
3. I've never heard of planet fitness.
4. Wait, was it "conquer the mall" or "conquer them all"?
5. To Bed, Bath, and Beyond!
6. What do they mean "Space for Rent"?
7. You idiot! I said Area 51, not Forever 21!
8. They deny our existence but know where we are?
9. We're good; they're all looking at their phones.
10. Dairy Queen, that must be their leader

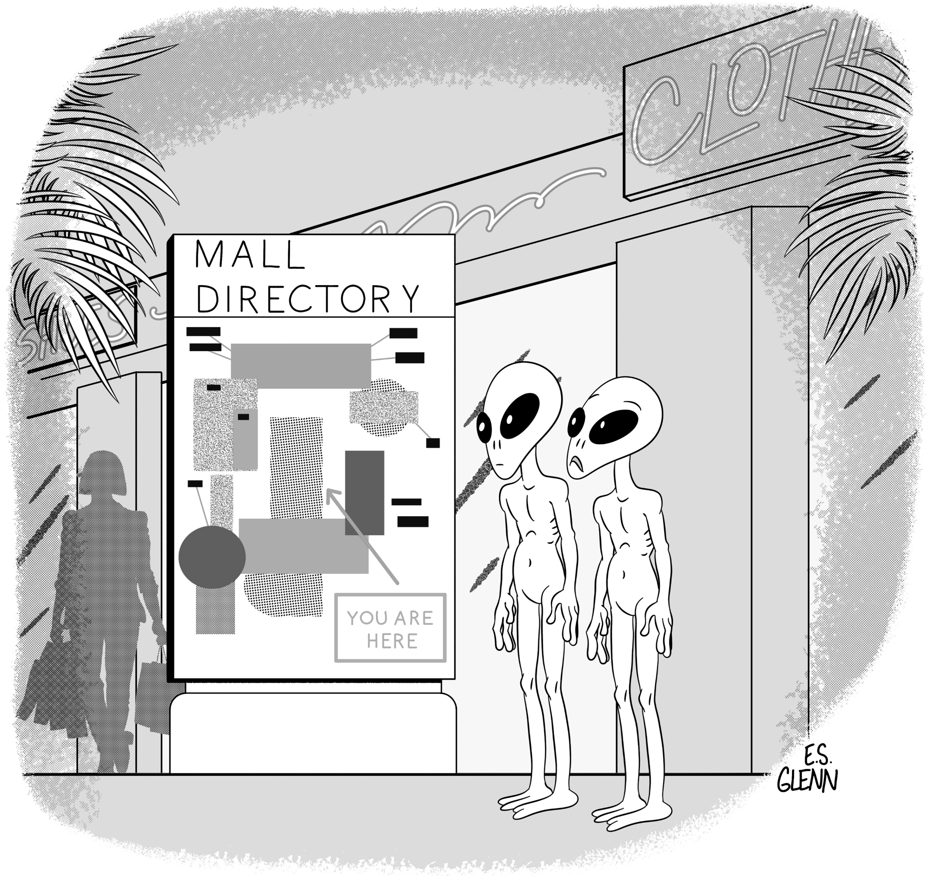
V.S.

LLM Generated

1. We've got to stop letting them treat us like surfboards.
2. He calls it 'extreme paddleboarding.' I call it rude.
3. At least it's not another barrel roll request.
4. He calls it surfing, I call it stubborn resistance.
5. Don't worry, this always turns into a trust exercise.
6. Apparently, we're the new ride-sharing service.
7. If this doesn't go viral, I'm quitting social media.
8. Just be glad he's not a juggler.
9. At least he's not asking for a joyride.
10. He calls it surfing, I call it acquiescing.

5-Shot Prompting

Human Judgement



Human Submission

1. Do you think death rays would be considered electronics or sporting goods?
2. Oh sure, now you look at a map.
3. I've never heard of planet fitness.
4. Wait, was it "conquer the mall" or "conquer them all"?
5. To Bed, Bath, and Beyond!
6. What do they mean "Space for Rent"?
7. You idiot! I said Area 51, not Forever 21!
8. They deny our existence but know where we are?
9. We're good; they're all looking at their phones.
10. Dairy Queen, that must be their leader

V.S.

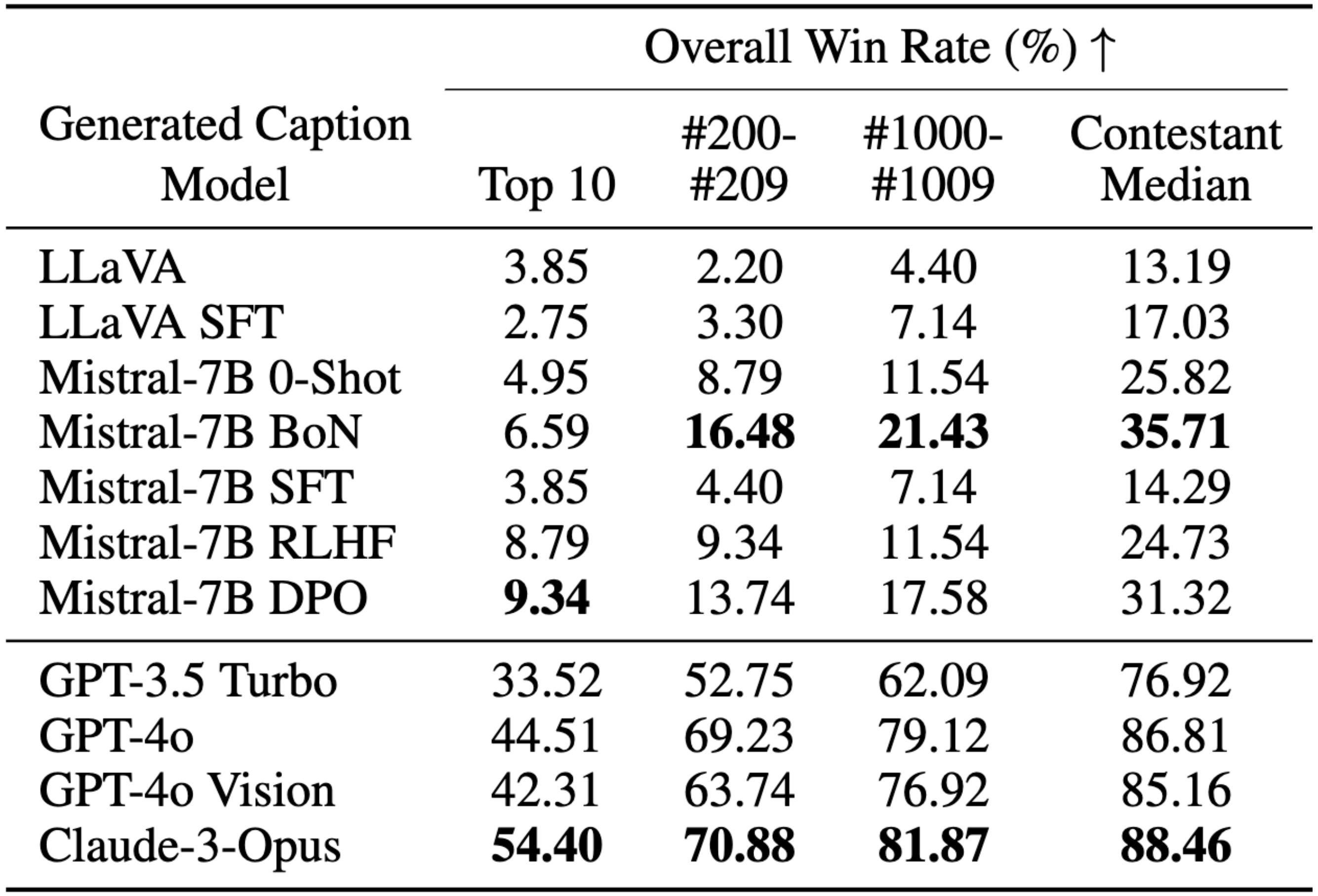
LLM Generated

1. We've got to stop letting them treat us like surfboards.
2. He calls it 'extreme paddleboarding.' I call it rude.
3. At least it's not another barrel roll request.
4. He calls it surfing, I call it stubborn resistance.
5. Don't worry, this always turns into a trust exercise.
6. Apparently, we're the new ride-sharing service.
7. If this doesn't go viral, I'm quitting social media.
8. Just be glad he's not a juggler.
9. At least he's not asking for a joyride.
10. He calls it surfing, I call it acquiescing.

Expert: former cartoon editor of New Yorker
Worker: crowdsource worker from Prolific

How Well Do LLMs Generate Funny Captions?

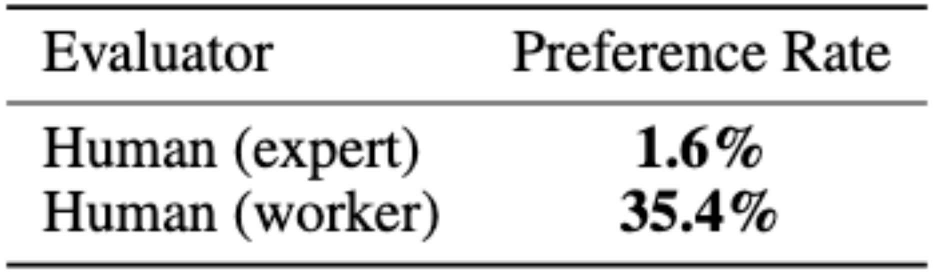
GPT4-Turbo as Judge



Generated Caption Model	Overall Win Rate (%) ↑			
	Top 10	#200-#209	#1000-#1009	Contestant Median
LLaVA	3.85	2.20	4.40	13.19
LLaVA SFT	2.75	3.30	7.14	17.03
Mistral-7B 0-Shot	4.95	8.79	11.54	25.82
Mistral-7B BoN	6.59	**16.48**	**21.43**	**35.71**
Mistral-7B SFT	3.85	4.40	7.14	14.29
Mistral-7B RLHF	8.79	9.34	11.54	24.73
Mistral-7B DPO	**9.34**	13.74	17.58	31.32
GPT-3.5 Turbo	33.52	52.75	62.09	76.92
GPT-4o	44.51	69.23	79.12	86.81
GPT-4o Vision	42.31	63.74	76.92	85.16
Claude-3-Opus	**54.40**	**70.88**	**81.87**	**88.46**

Human Judgement

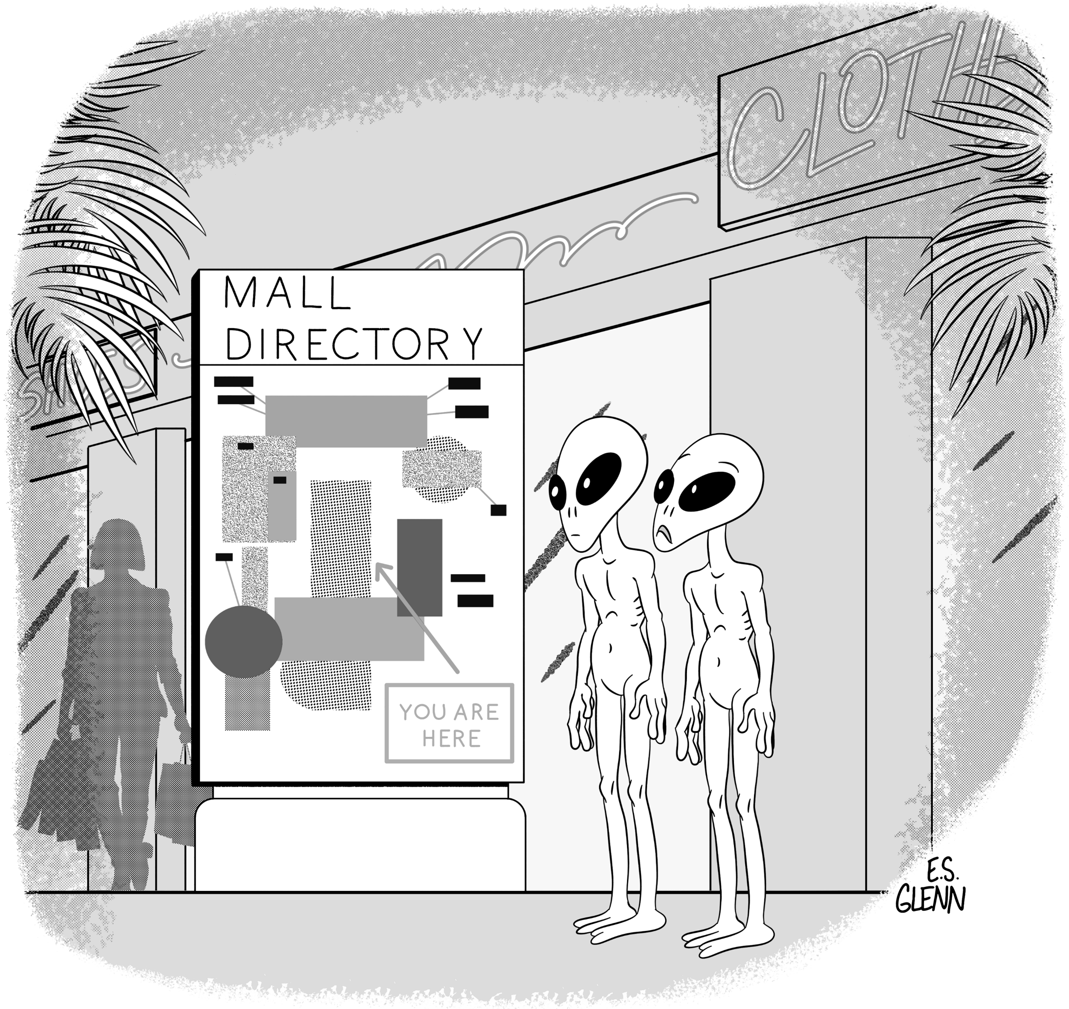
Claude-3-Opus
vs Human Top 10



Evaluator	Preference Rate
Human (expert)	**1.6%**
Human (worker)	**35.4%**

From expert's perspective,
LLM generated captions are far from great

Challenge 1: New Yorker humor has multiple layers of funny interpretations



Top 10 Human Caption

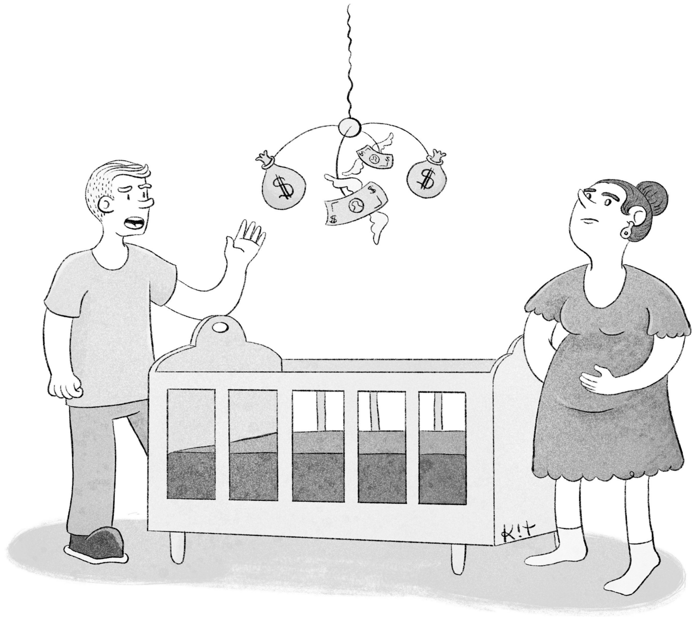
“Oh sure, now you look at a map”

v.s.

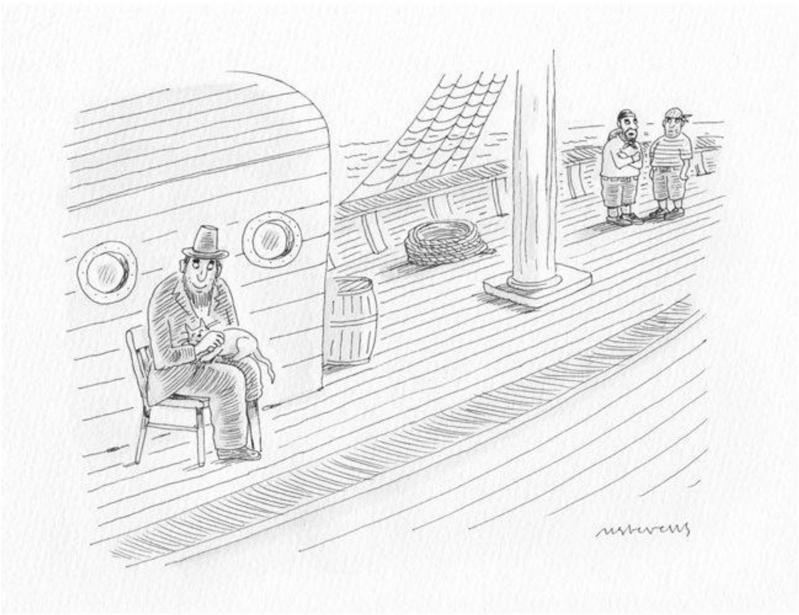
GPT-4o Vision

“We travel light-years, and we still need directions”

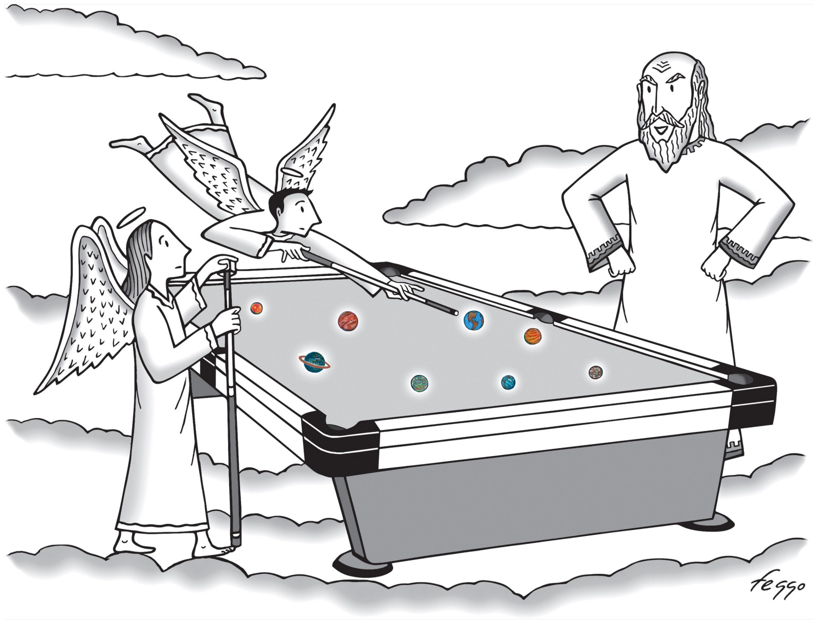
Challenge 1: New Yorker humor has multiple layers of funny interpretations (cultural references, tropes, puns, wordplays)



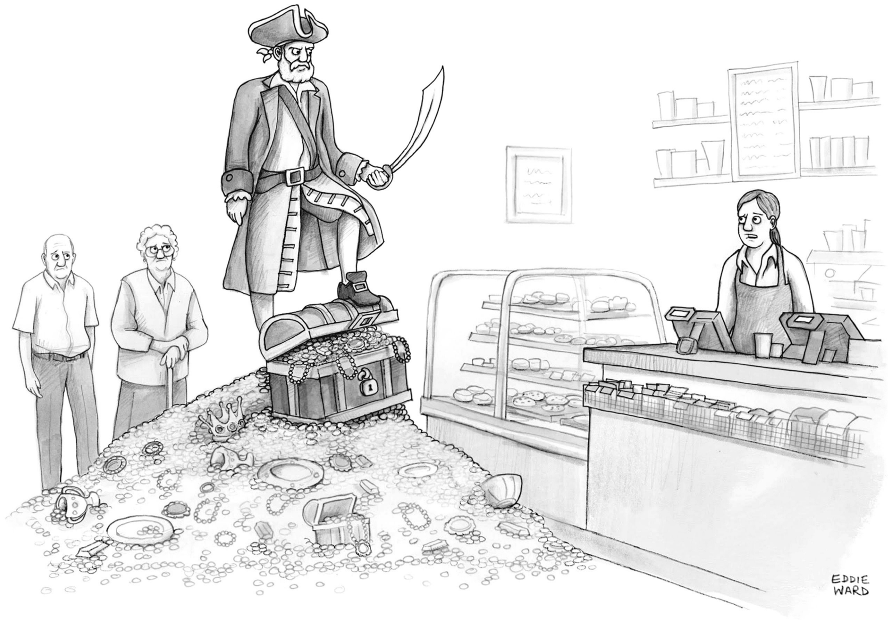
“It’s attached to the debt ceiling.”



"He calls it Ishmeow."

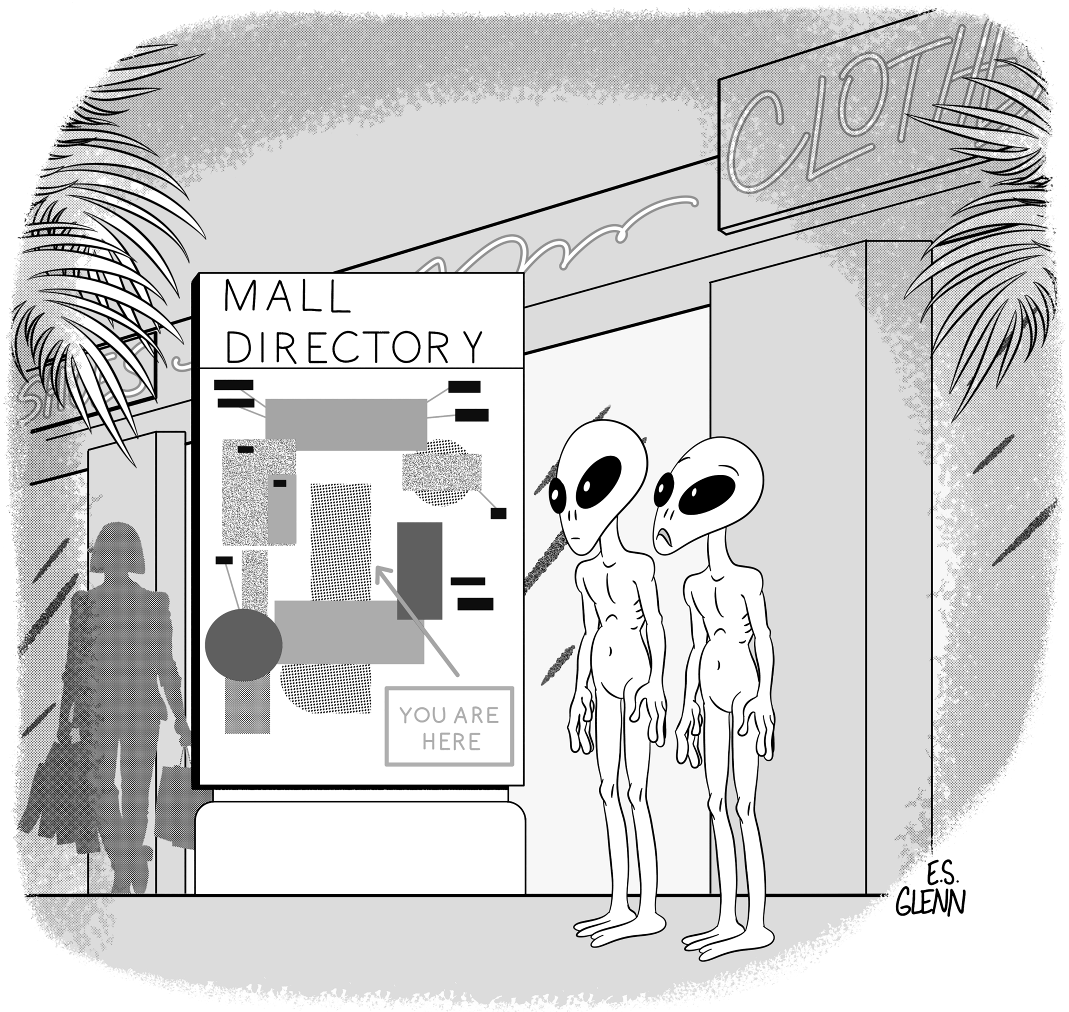


“I don't care what Satan lets his kids do.”



“Sir, this is a Whole Foods. The only payment we accept is an arm and a leg.”

Challenge 2: Current LLMs generate new words and phrases



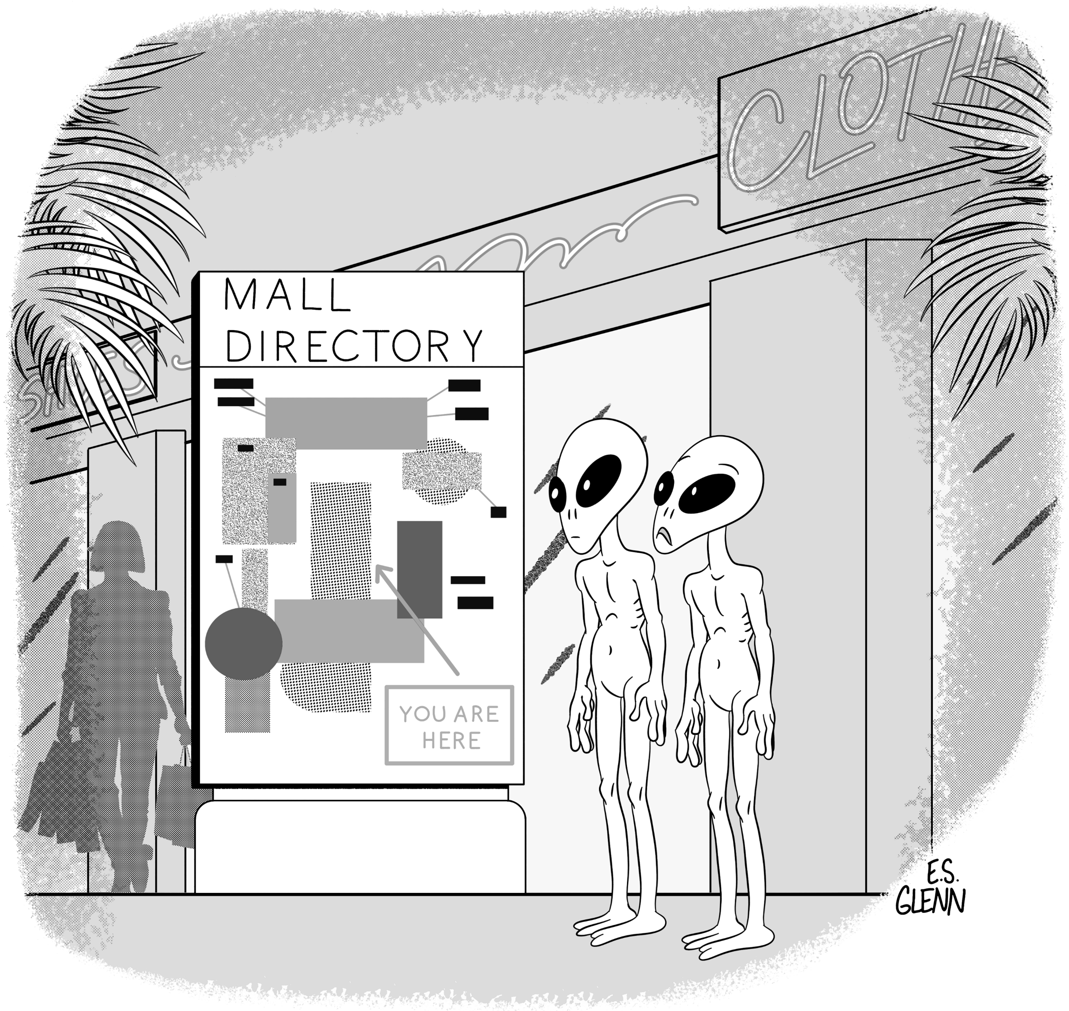
GPT-4o Vision

“Do they have a Black Hole Friday sale?”

Claude 3 Opus

“I don’t see ’Invasion Supplies’ listed anywhere...”

Challenge 3: Multimodal LLMs currently miss visual details



GPT-4o Vision

“So the alien abduction statistics were right. Malls are the prime hunting grounds”

More in our paper

- Open research problems for preference-based alignment methods (RLHF and DPO)
- More results on pairwise and group comparison performance of LLMs
- More dataset/codebase details
- Diversity comparison of AI vs human captions

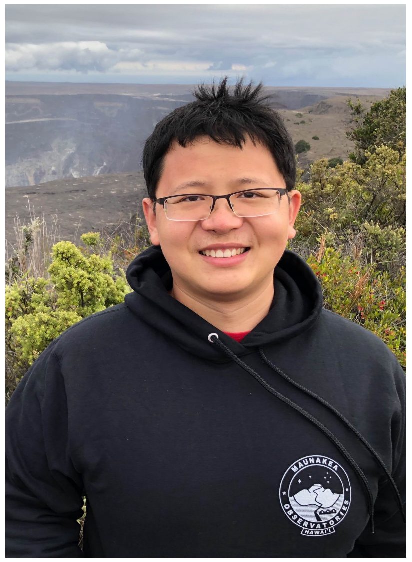
Poster: Wed 11 Dec
11am — 2pm

Lalit Jain

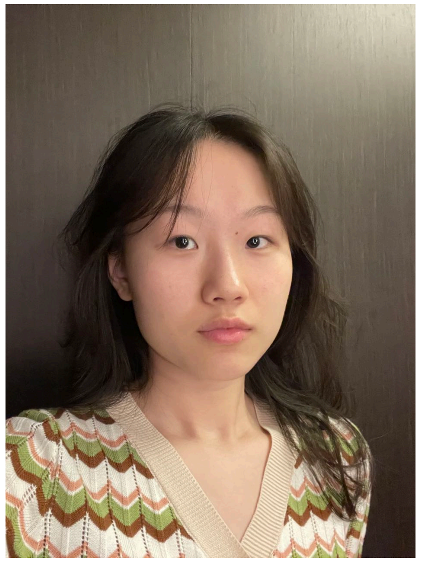
Robert
Nowak

Paper Link

Lalit Jain



Yang Guo



Jiayi Chen

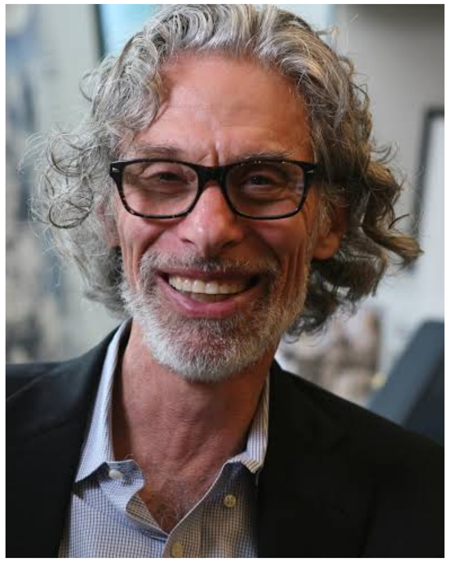
David Zhou

Siddharth Suresh

Andrew Wagenmaker

Timothy Rogers

Kevin Jamieson



Bob Mankoff

Robert Nowak

Give This Cartoon A Funny Caption