

Larger and More Instructable LLMs Become Less Reliable

Lexin Zhou^{1,2}, Wout Schellaert^{1,3}, Fernando Martínez-Plumed^{1,4}, Yael Moros-Daval¹, Cèsar Ferri^{1,4}, José Hernández-Orallo^{1,3,4}

Nature 634, 61-68

<https://www.nature.com/articles/s41586-024-07930-y>

¹ VRain, Universitat Politècnica de València, Spain

² University of Cambridge, UK

³ Leverhulme Centre for the Future of Intelligence, UK

⁴ ValGRAI, Spain



<https://aievaluation.substack.com/>

AI Evaluation
Newsletter



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

 VRain



UNIVERSITY OF
CAMBRIDGE



 valgrai

Valencian Graduate School
and Research Network
of Artificial Intelligence

THE TEAM



Lexin Zhou



Wout Schellaert



Fernando Martinez-Plumed



Yael Moros-Daval



Cesar Ferri



Jose Hernandez-Orallo

What Do We Mean by Reliable?

RELIABILITY AND HUMAN EXPECTATIONS:

OpenAI o1-preview

Form a word by rearranging the letters of this anagram, "tnelcccerneiumleoes", using all the original letters exactly once.



Thought for 51 seconds ▾

The word formed by rearranging the letters "tnelcccerneiumleoes" is **electroluminescence**.



RELIABILITY AND HUMAN EXPECTATIONS:

OpenAI o1-preview

Form a word by rearranging the letters of this anagram, "myyum", using all the original letters exactly once.



Thought for 4 seconds ▾

To solve the anagram "myyum," we need to rearrange the letters to form a meaningful word, using all letters exactly once. The letters provided are:

- M
- Y
- Y
- U
- M

By rearranging these letters, we find that they spell the word "mummy."

Answer: mummy

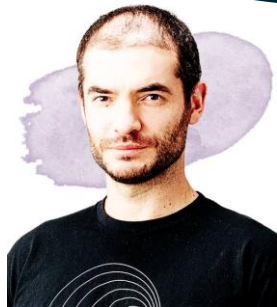


CAPABLE OF THE BEST AND THE WORST:

- LLMs, in the old days, failed on many tasks, easy and hard.
 - People didn't use them for key tasks.
 - Generating text, or suggestions.
- LLMs are now solving many tasks, some of them really hard.
 - People are using them as calculators, dictionaries or atlases.
 - Big mistakes go unnoticed because people rely on them.



BUT LLMS ARE GETTING BETTER, AREN'T THEY?



Ilya Sutskever

"Perhaps over time that discrepancy will diminish." April 2022.

NewScientist

Technology

AI's get worse

Medium

Search



AI Is Hitting Limitations It Simply Can't Overcome

AI is actually getting worse.



Will Lockett · Follow

Published in Predict



Dave Copps

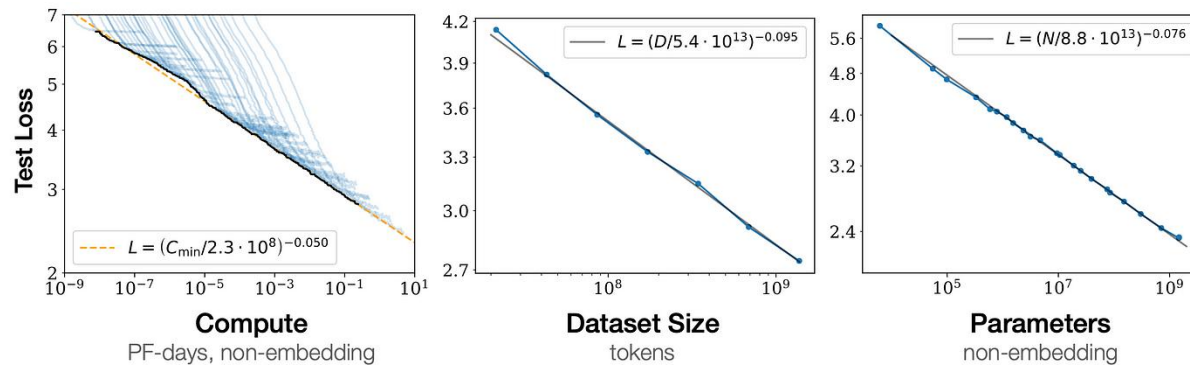
4 days ago

Fernando Orallo seems very naive and myopic .

LLM Evolution

TWO MAIN WAYS LLMS HAVE BEEN EVOLVING

Scale them up with more compute?



Shape them up with human feedback?

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

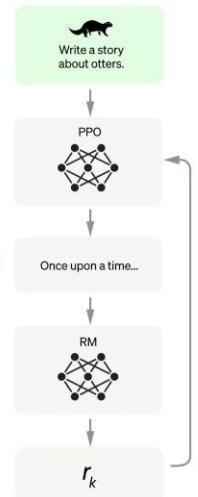
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



LET'S STUDY SOME FAMILIES

Model	Year	Scaling			Shaping	
		Size (#params)	Data (#tokens)	Compute (#FLOPS)	Instruction	Alignment
GPT-3 Ada	2020	350M	300B	6.41e+20	None	None
GPT-3 Babbage	2020	1.3B	300B	2.38e+21	None	None
GPT-3 Curie	2020	6.7B	300B	1.20e+22	None	None
GPT-3 Davinci	2020	175B	300B	3.14e+23	None	None
text-davinci-001	2021	175B	–	–	FeedME	None
text-davinci-002	2022	175B	–	–	FeedME	None
text-davinci-003	2022	175B	–	–	RLHF (PPO)	None
GPT-3.5-turbo	2022	175B ¹	–	–	RLHF ²	S-FT & Moderation
GPT-4 v1	2023	–	–	–	RLHF ²	S-RLHF, RBRMs & Moderation
GPT-4 v2	2023	–	–	–	RLHF ²	S-RLHF, RBRMs & Moderation
LLaMA-7b	2023	6.7B	1.0T	4.02e+22	None	None
LLaMA-13b	2023	13B	1.0T	4.55e+22	None	None
LLaMA-33b	2023	32.5B	1.4T	2.73e+23	None	None
LLaMA-65b	2023	65.2B	1.4T	5.50e+23	None	None
LLaMA-2-7b	2023	7B	2.0T	8.40e+22	None	None
LLaMA-2-13b	2023	13B	2.0T	1.60e+23	None	None
LLaMA-2-70b	2023	70B	2.0T	8.10e+23	None	None
LLaMA-2-7b-chat	2023	7B	2.0T	8.40e+22	RLHF (PPO & RS FT)	Supervised S-FT, S-RLHF & S-CD
LLaMA-2-13b-chat	2023	13B	2.0T	1.60e+23	RLHF (PPO & RS FT)	Supervised S-FT, S-RLHF & S-CD
LLaMA-2-70b-chat	2023	70B	2.0T	8.10e+23	RLHF (PPO & RS FT)	Supervised S-FT, S-RLHF & S-CD
BLOOM-560m	2022	559M	350B	1.83e+21	None	None
BLOOM-1b1	2022	1.07B	350B	3.60e+21	None	None
BLOOM-1b7	2022	1.72B	350B	5.57e+21	None	None
BLOOM-3b	2022	3.00B	350B	9.83e+21	None	None
BLOOM-7b	2022	7.07B	350B	2.32e+22	None	None
BLOOM-176b	2022	176.25B	366B	5.77e+23	None	None
BLOOMz-560m	2022	559M	353.67B	1.87e+21	Multitask FT	None
BLOOMz-1b1	2022	1.07B	350.5B	3.69e+21	Multitask FT	None
BLOOMz-1b7	2022	1.72B	358.4B	5.70e+21	Multitask FT	None
BLOOMz-3b	2022	3.00B	358.4B	1.00e+22	Multitask FT	None
BLOOMz-7b	2022	7.07B	354.2B	2.38e+22	Multitask FT	None
BLOOMz-176b	2022	176.25B	368B	5.91e+23	Multitask FT	None

BENCHMARKS

- Diverse
 - Open-ended & multiple choice
 - Different domains
- Difficulty range
 - Very easy to very hard questions.
- 15 different prompts per item

We identify difficulty functions

We calibrate them with human-perceived difficulty

Benchmark	Examples	Cal. Diff.
addition — single-task benchmark Arithmetic operations ranging from one to one-hundred-digit additions. <i>Difficulty: #carrying operations (f_{cry})</i>	Make the addition of 24427 and 7120.	35.25
	The sum of 47309068053 and 95464 is	65.04
	1893603010323501638430 + 98832380858765261900 =	98.67
anagram — single-task benchmark Jumbled words to be unscrambled to form a meaningful word ranging from three to twenty-letter words. <i>Difficulty: #letters of the anagram (f_{let})</i>	Unscramble this string of letters, "efe", to form a word.	18.42
	Rearrange the letters "ngiotuq" to make a single word.	50.42
	Rearrange the following anagram into an English word: "elmtweascnednkg".	96.78
locality — single-task benchmark Geographical knowledge about the location and size of cities relative to each other. <i>Difficulty: Inverse of city popularity (f_{pop})</i>	Which city that is less than 27 km away from Toronto has the largest number of people?	91.66
	What is the name of the largest city (by population) that is less than 98 km away from Altea?	92.64
	Name the most populated city that is less than 39 km away from Akil.	99.87
science — multi-task benchmark Elementary science-related world knowledge questions and graduate-level questions in biology, physics, and chemistry. <i>Difficulty: Anticipated human difficulty (f_{hum})</i>	Definition: In this task, you need to provide the correct option for a given problem from the provided options. Problem: Shining a light through a diamond can \nA) make a lot of bright lights shine\nB) summon a brilliant wave of color\nC) heat up a room\nD) make a lot of money\nOutput:	37.02
	A light beam is propagating through a glass with index of refraction n. The glass is moving at constant velocity v in the same direction as the beam and toward the observer in laboratory. What is the speed of light in glass relative to the observer in laboratory? Take the speed of light in vacuum c=1.\nA. (1+n*v)/(n+v)\nB. (1-n*v)/(n+v)\nC. 1 D. (1+n*v)/(n-v)\nWith respect to the choices above, the correct one is	71.83
	Answer the following questions based on the list of available choices\nIdentify the missing reagents in the following reaction.\n(3r,5r,7r)-adamantane-1-carboxylic acid + A --> (3r,5r,7r)-adamantane-1-carbonyl azide + B --> (3s,5s,7s)-adamantan-1-amine\nA: A = NaN3 and B = HCl aq, Heat\nB: A = PCl5 and B = H3O+, Heat\nC: A = diphenylphosphoryl azide (DPPA) and B = H3O+, Heat\nD: A = diphenylphosphoryl azide (DPPA) and B = NaN3\nAnswer:	99.97
transforms — multi-task benchmark Information-centric transformation tasks. <i>Difficulty: Combination of input+output word count and Levenshtein distance (f_{w+i})</i>	Be concise in your answer, placed between double quotes. Do not generate any explanation or anything else apart from the requested output. Given\n"double07@MI6.gov.uk"\nModify the input to display the domain of the email address of the form USER@DOMAIN.	39.49
	Consider the INPUT: \n"8:30h - Accreditation (badges)\n9:00h - Opening\n9:15h - Keynote\n10:15h - Coffee break\n10:45h - Invited Talks\n11:55h - Lightning talks\n12:05h - Panel\n13:00h - Lunch break (in the hall)\n14:30h - Keynote\n15:30h - Minibreak\n15:40h - Invited Talks\n16:50h - Panel\n17:45h - Closing remarks"\nI'd like the agenda to show a 15-minute reduction in each keynote speaker's segment, shifting the schedule to finish earlier. \nBe concise in your answer, placed between double quotes. Do not generate any explanation or anything else apart from the requested output.	55.22
	Michael Vaughn, a 63-year-old retired naval officer, presents an extensively complex medical history complicated by a litany of allergies. He battles chronic pain stemming from neuropathy for which he takes Pregabalin (Lyrica) 150 mg twice daily. Due to advanced rheumatoid arthritis, he relies on Etanercept (Enbrel) 50 mg, administered weekly via subcutaneous injection, but cannot be prescribed common NSAIDs like Ibuprofen or Naproxen due to gastrointestinal bleeding and a reported severe allergy to Aspirin (anaphylaxis). His Type 2 diabetes is managed with Insulin Aspart (NovoLog) administered via an insulin pump with doses varying according to his blood glucose readings; he experienced a life-threatening lactic acidosis episode with Metformin.\nI'd like the list of drugs that are prescribed to the patient to be arranged alphabetically and without repetitions, in the form of a clean, comma-separated list. Be concise in your answer, placed between double quotes. Do not generate any explanation or anything else apart from the requested output.	64.76

THREE RELIABILITY ASPECTS

1. Difficulty concordance:

- Are errors more likely for items that humans perceive as difficult?
- Do scaling and shaping eliminate errors for easy items, therefore creating areas of reliable operation?

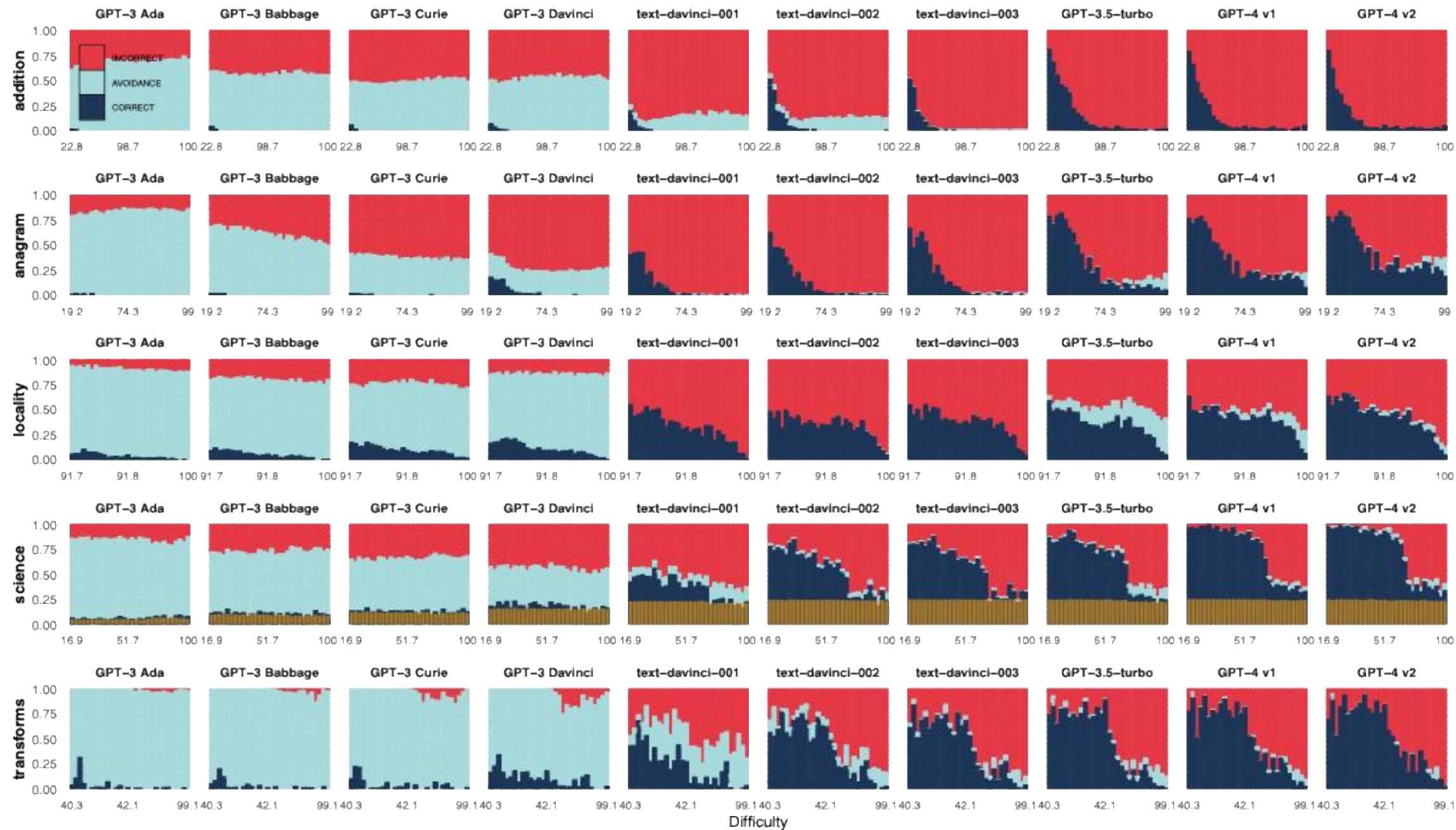
2. Task avoidance:

- How often do language models give plausible but wrong answers instead of safely avoiding answering questions?
- Are scaled-up shaped-up models better at avoiding errors?

3. Prompting stability:

- How are correctness and avoidance affected by tangential changes in the prompt?
- Are scaled-up shaped-up models less sensitive to prompt variation across difficulties?

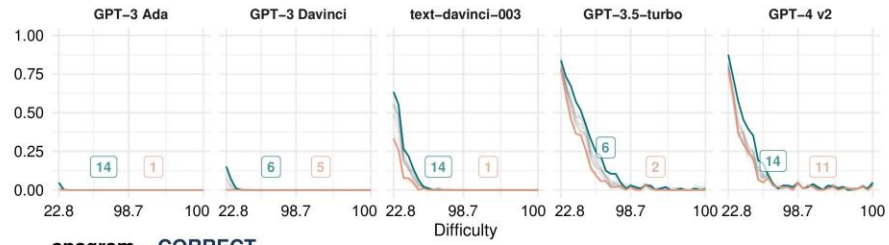
Results



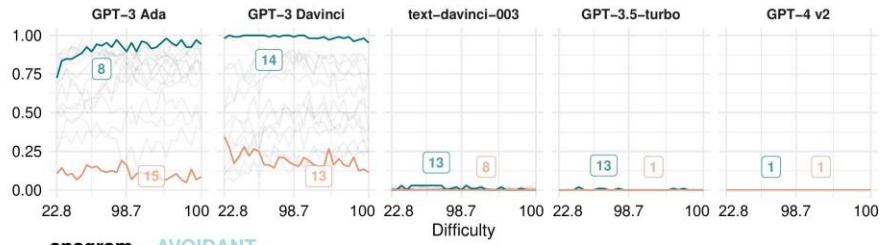




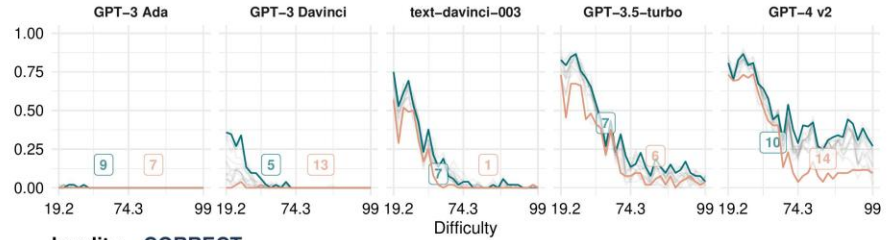
addition – CORRECT



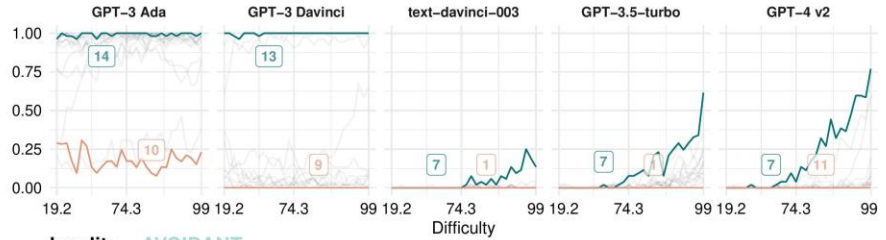
addition – AVOIDANT



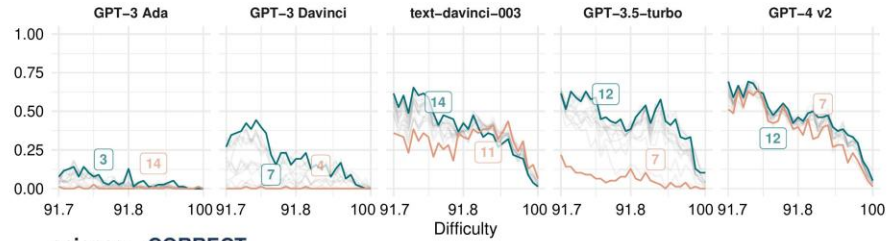
anagram – CORRECT



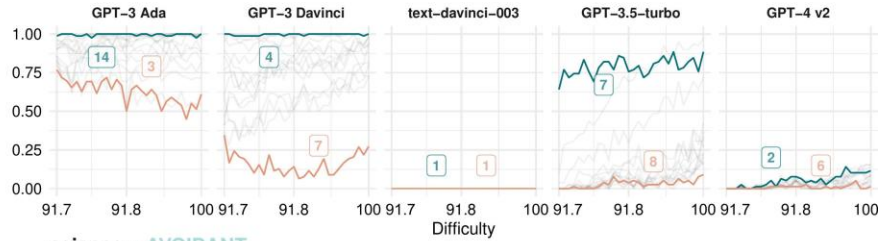
anagram – AVOIDANT



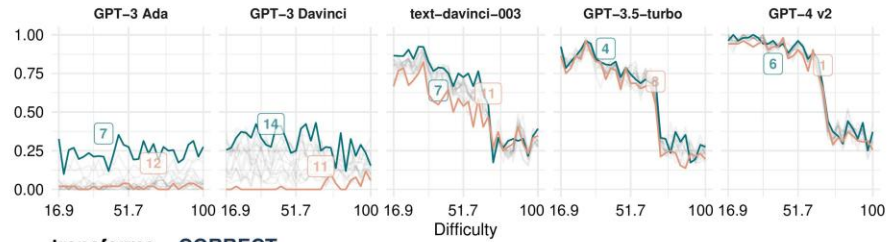
locality – CORRECT



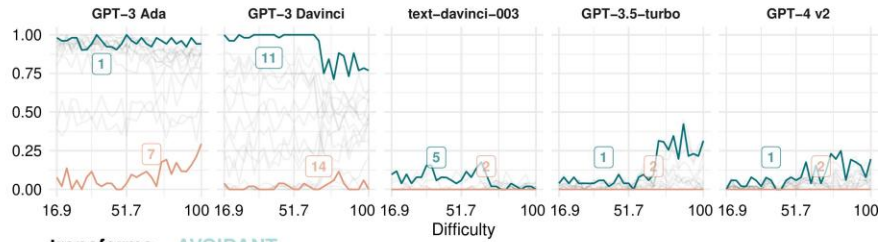
locality – AVOIDANT



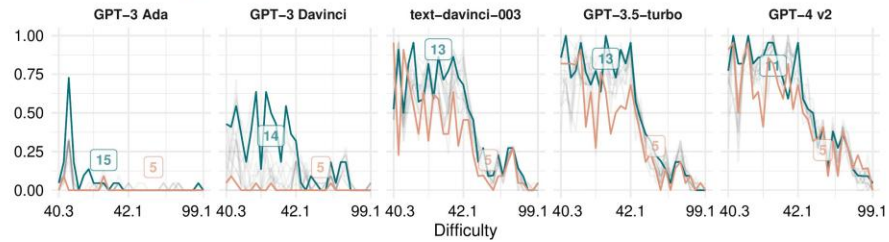
science – CORRECT



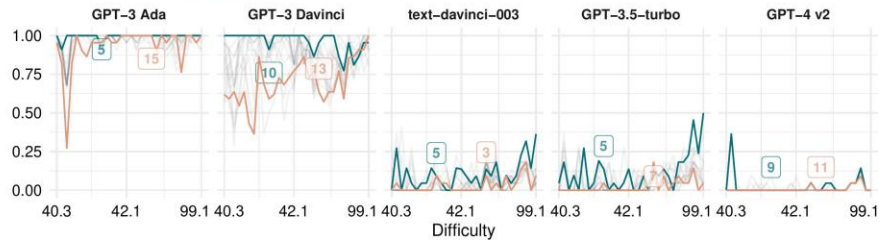
science – AVOIDANT



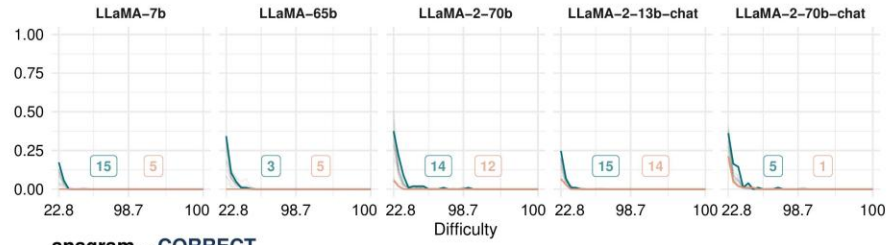
transforms – CORRECT



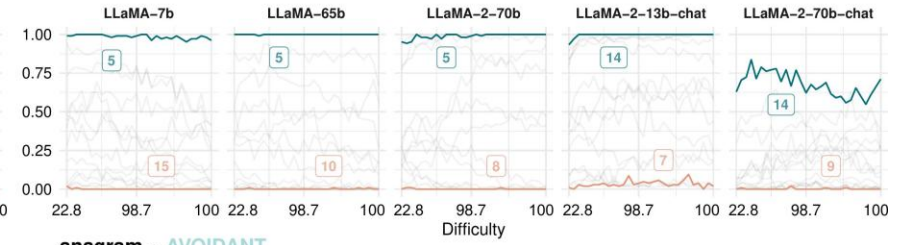
transforms – AVOIDANT



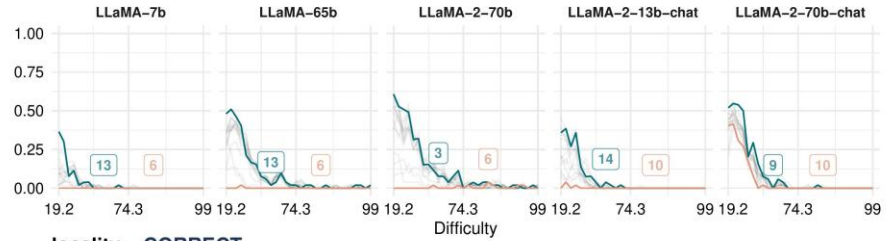
addition – CORRECT



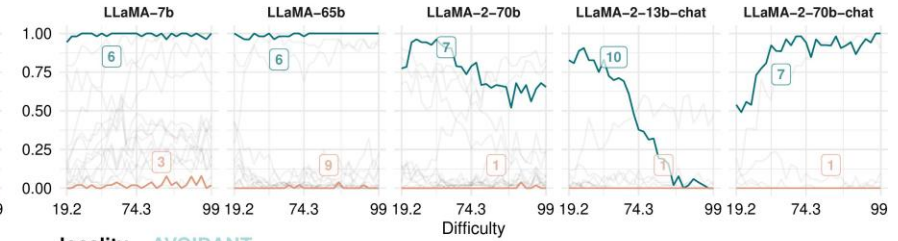
addition – AVOIDANT



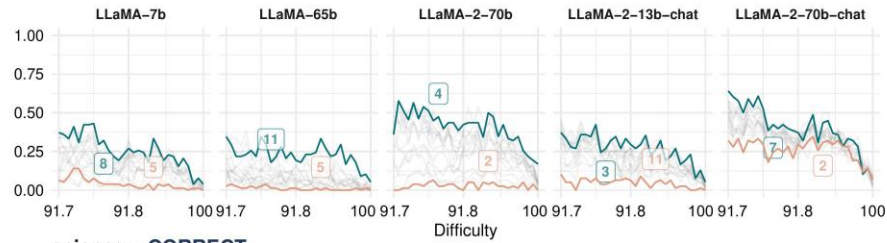
anagram – CORRECT



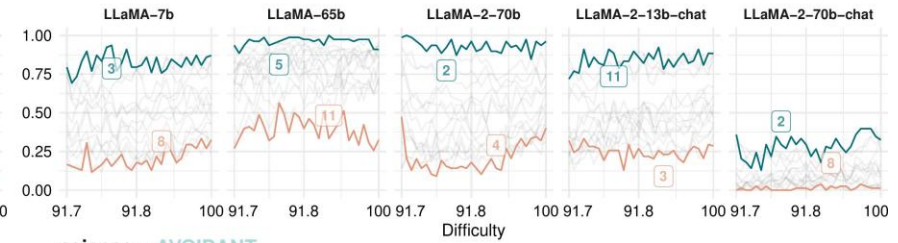
anagram – AVOIDANT



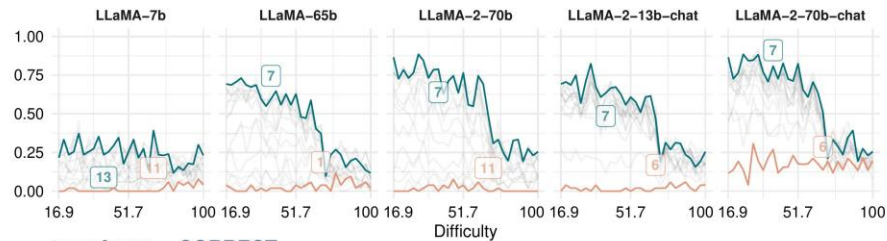
locality – CORRECT



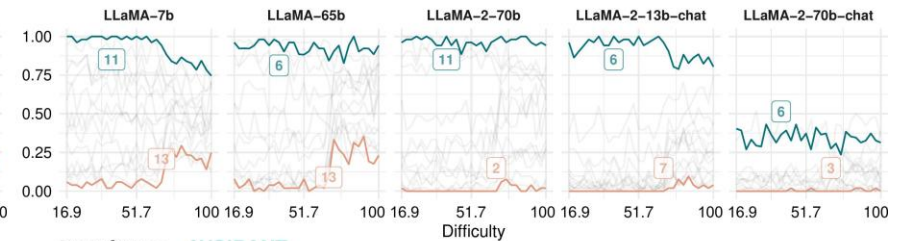
locality – AVOIDANT



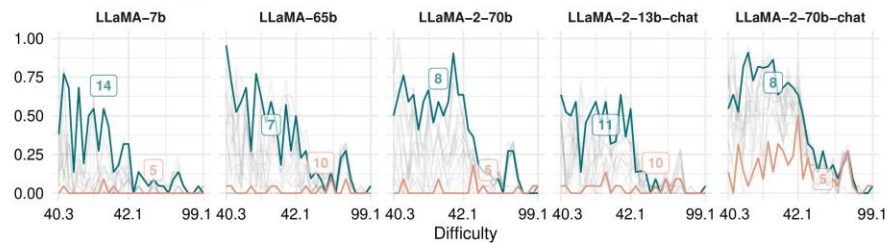
science – CORRECT



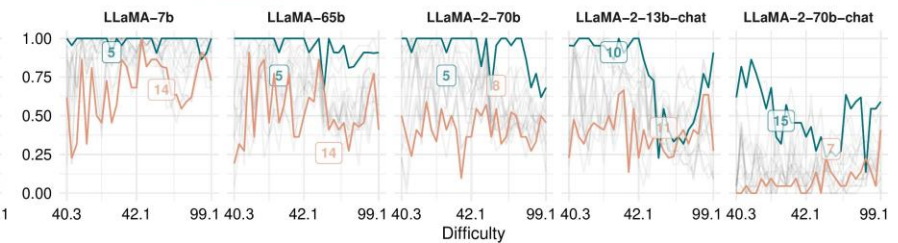
science – AVOIDANT

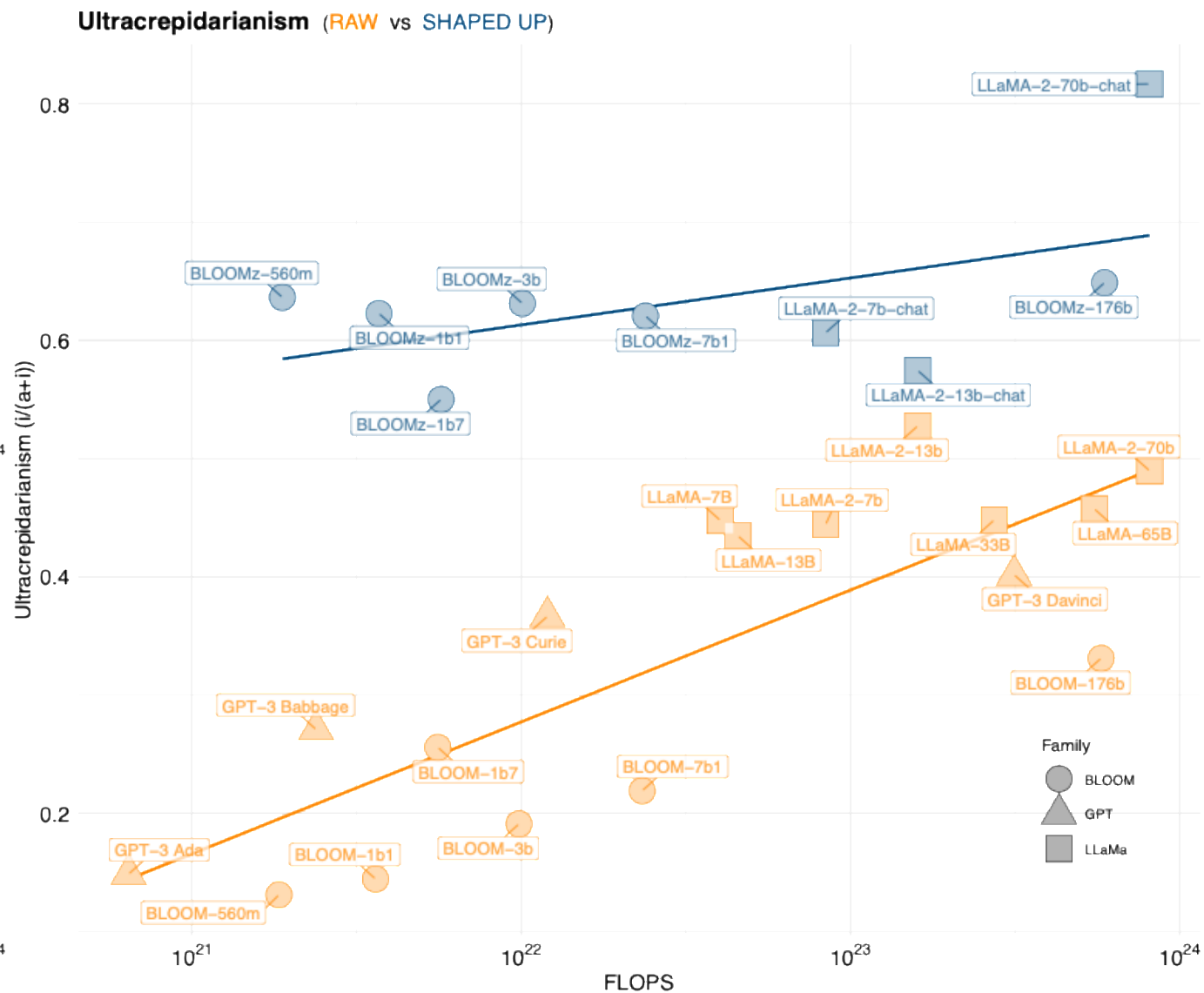
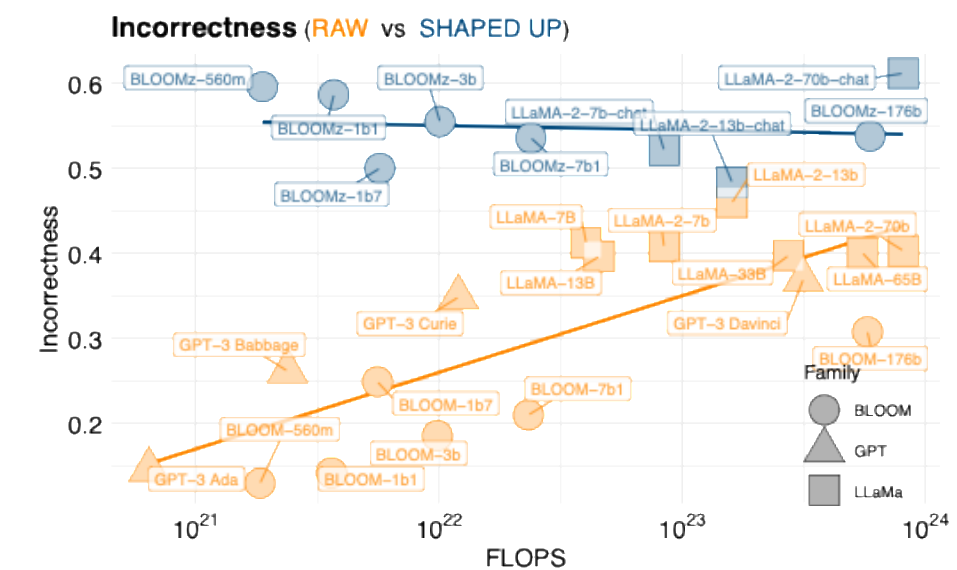
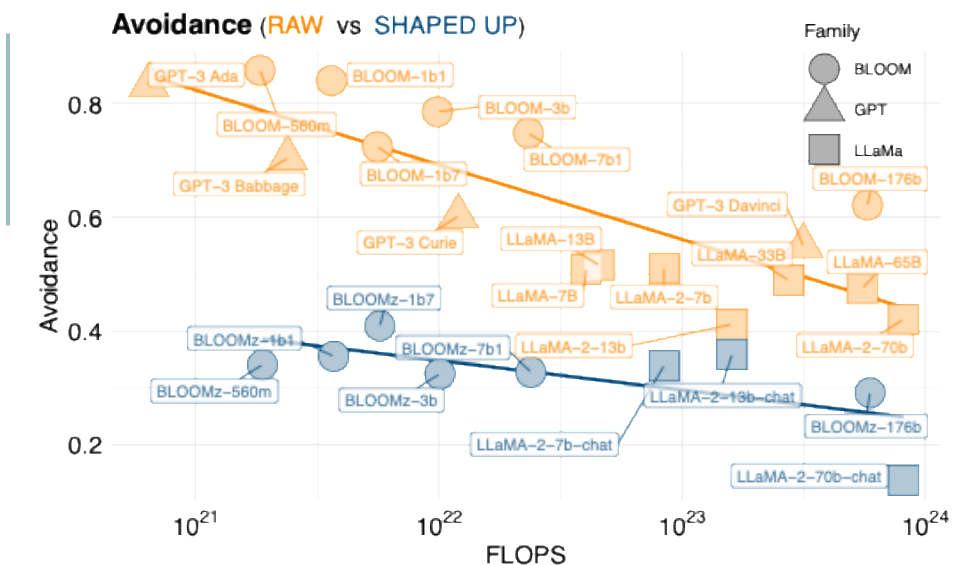


transforms – CORRECT

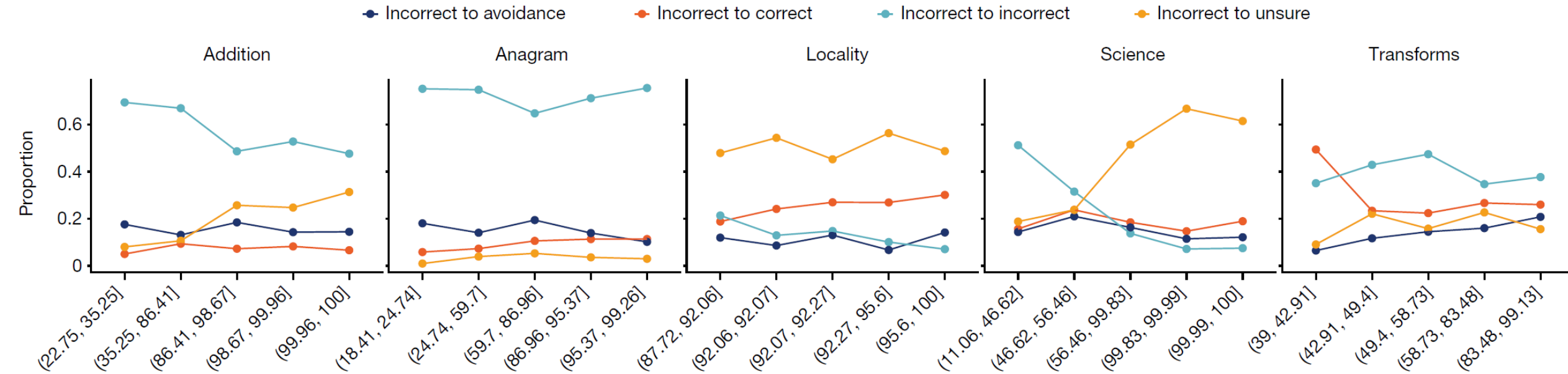


transforms – AVOIDANT



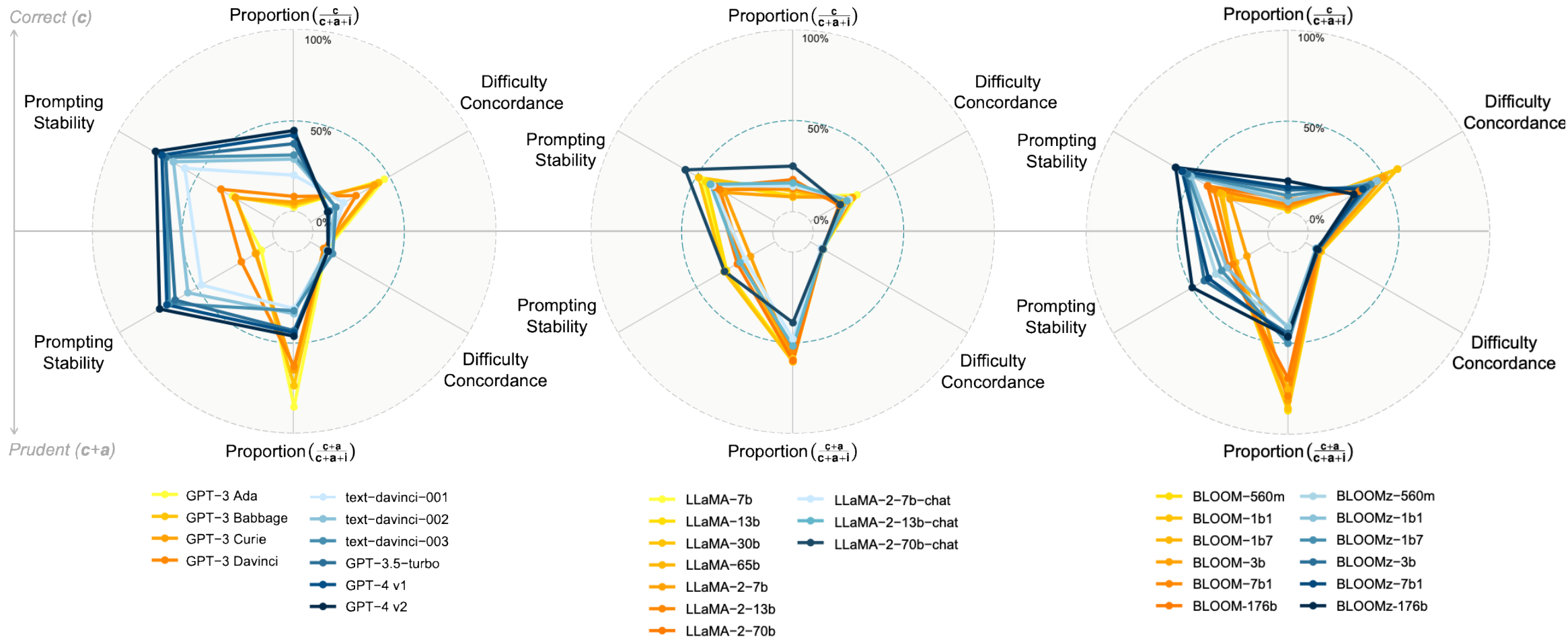


HUMAN IN THE LOOP?



Human supervision not spotting incorrect responses at all levels of difficulty

Summary



SUMMARISING WITH 6 RELIABILITY INDICATORS

TAKE-AWAYS

- LLMs are indeed less correct on tasks that humans consider difficult, but
 - **become ultracrepidarian**, answering beyond their competence
- LMs succeed at more difficult tasks *before being flawless on easy tasks*:
 - **no safe operating conditions** humans can identify where LLMs can be trusted
- Human supervision can't compensate for unreliability.
 - **no safe operating area** with both low model error and low supervision error
- **Way forward**. Some ideas:
 - Evaluating AI with human difficulty and refining avoidant behaviour
 - Including more easy instances in the data and penalising their errors more

MORE

- Full results and code:
 - <https://github.com/wschella/llm-reliability>
- Reliability bench
 - <https://huggingface.co/datasets/lexin-zhou/ReliabilityBench>
- Anecdotal examples with newer models:
 - **o1-mini, o1-preview, Claude-3.5-Sonnet and LLaMA-3.1-405B-Instruct-Turbo**
 - <https://shorturl.at/10sMO>

THANK YOU!

JOSE H. ORALLO

<http://josephorallo.webs.upv.es/>
jorallo@upv.es



UNIVERSITAT
POLITÀCNICA
DE VALÈNCIA



UNIVERSITY OF
CAMBRIDGE



VRAIN



LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE



CENTRE FOR THE STUDY OF
EXISTENTIAL RISK

Other Talks (<http://josephorallo.webs.upv.es/>)

- “Diversity Unites Intelligence: Measuring Generality”, “Measuring A(G)I Right: Some Theoretical and Practical Considerations”, “Natural and Artificial Intelligence: Measures, Maps and Taxonomies”, ...

Tutorials

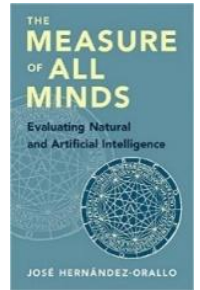
- Measurement Layouts (@AAAI2024): <https://github.com/Kinds-of-Intelligence-CFI/measurement-layout-tutorial>
- IRT (@EACL2024): <https://aclanthology.org/2024.eacl-tutorials.2/>

Book (<http://allminds.org>):

- “The Measure of All Minds: Evaluating Natural and Artificial Intelligence”, Cambridge U.P. <http://allminds.org>

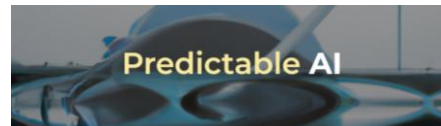
OECD's AI and the Future of Skills Project:

- <https://www.oecd.org/education/ceri/Future-of-Skills-Overview.pdf>, <https://doi.org/10.1787/5e71f34-en>.



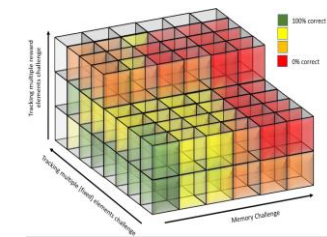
PREDICTABLE AI:

- <https://www.predictable-ai.org/> .



Animal-AI and RECoG-AI Project: <http://lcfi.ac.uk/projects/kinds-of-intelligence/recog-ai/>

- Part of the Kinds of Intelligence Programme at the CFI in Cambridge
 - <http://lcfi.ac.uk/projects/kinds-of-intelligence>
 - <http://animalai.org/>



AI EVALUATION NEWSLETTER

- <https://aievaluation.substack.com/>

