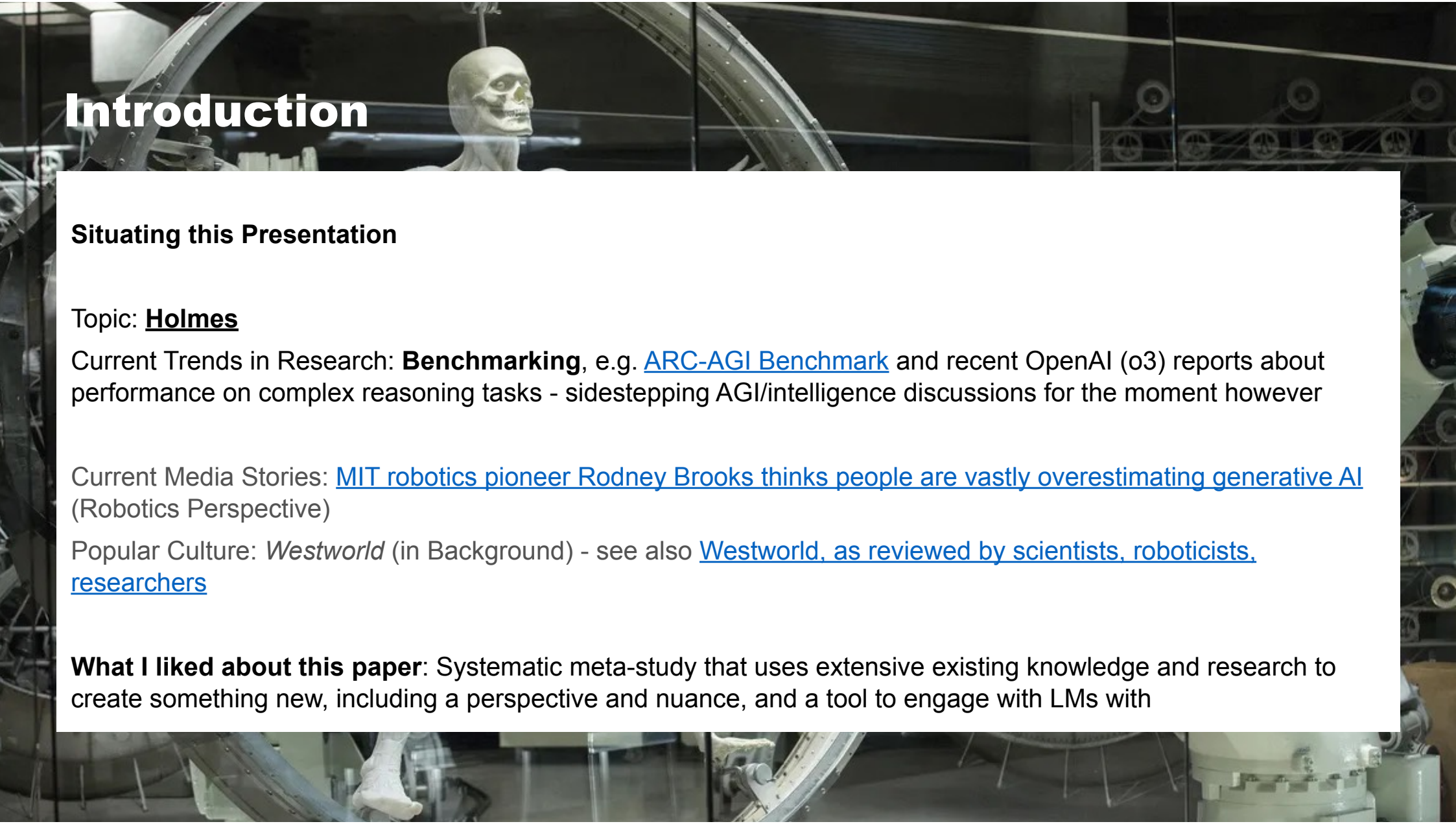
LLM Linguistic Competence

with a Focus on Benchmarking/HOLMES

Jack Heseltine **for IT:U NLP Group - Pres. 2 | 12 mins + 8 mins Q&A**

JOHANNES KEPLER UNIVERSITY LINZ



Introduction

Situating this Presentation

Topic: **Holmes**

Current Trends in Research: **Benchmarking**, e.g. ARC-AGI Benchmark and recent OpenAI (o3) reports about performance on complex reasoning tasks - sidestepping AGI/intelligence discussions for the moment however

Current Media Stories: MIT robotics pioneer Rodney Brooks thinks people are vastly overestimating generative AI (Robotics Perspective)

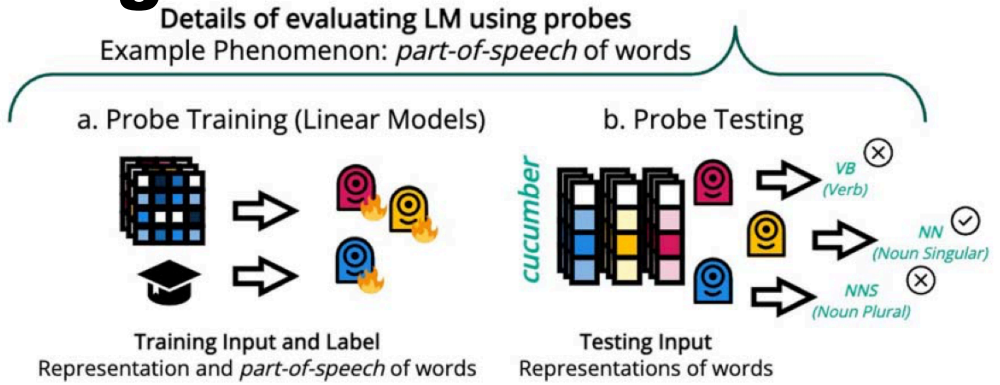
Popular Culture: *Westworld* (in Background) - see also Westworld, as reviewed by scientists, roboticists, researchers

What I liked about this paper: Systematic meta-study that uses extensive existing knowledge and research to create something new, including a perspective and nuance, and a tool to engage with LMs with

Coming from Benchmarking: Linguistic Performance

Skipping to results section (5) of the paper,
before working backward:

59 LLMs evaluated, 5 linguistic categories



What is the evaluation?

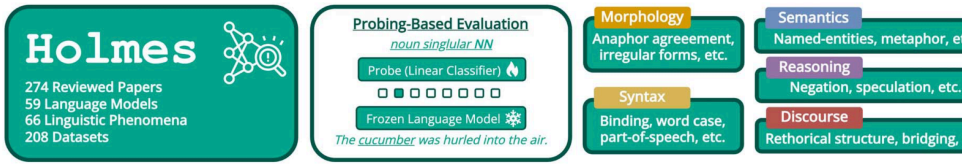
(From 4.3) internal representation of the last layer of LMs. This was my first hurdle of understanding, just exactly what we are talking about, and I referred to the Appendix as suggested - and would like to **suggest my own *FlashHOLMES* take-home exercise if helpful**

Re: this PhD application!

Absolute prediction performance of the probes (see XAI cross-reference later)

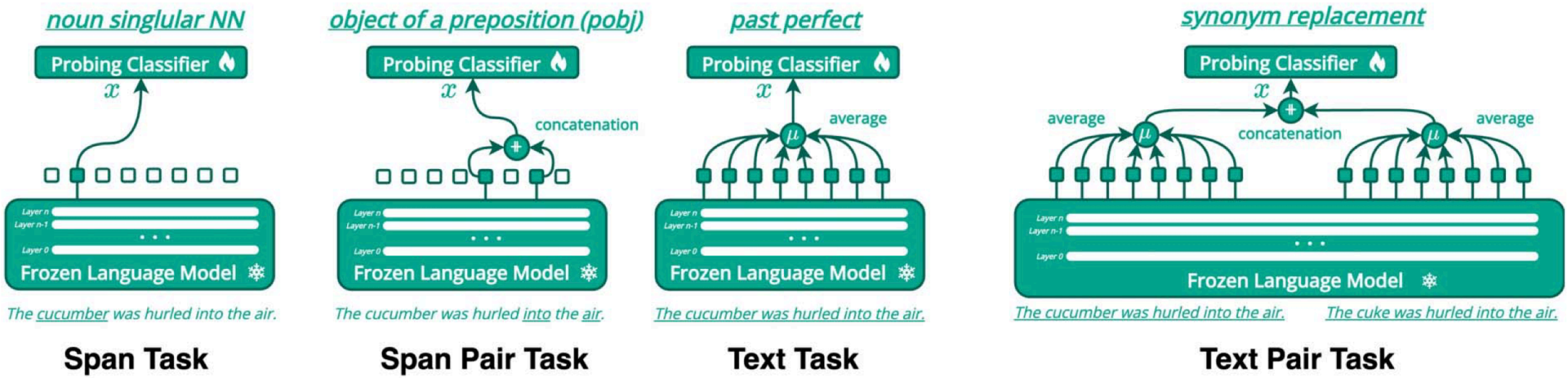
Reliability evaluation using control tasks and from information theory perspective - will also go into this

<u>Goal: Training probes to predict linguistic phenomena; so that quality of prediction says something about the LM</u>



Internal Representation: DETAIL

What can we say about an LM based on a probe/what will the downstream metrics tell us? (This was my main question going in.)



Linguistic Performance: DETAIL

What are the morphology, syntax, semantics, reasoning, and **discourse** phenomena/how do we check for these (what is the input)?

Phenomena	Illustrative Example	Text	Text-Pair	Span	Span-Pair	Weischedel et al. (2013)	Pandit and Hou (2021)	Nie et al. (2019)	Narayan et al. (2018)	Webber et al. (2019)	Carlson et al. (2001)	Zeldes (2017)
bridging	The disease and symptoms of advanced infection. ⇒ Valid Bridge	1			1		✓					
co-reference resolution	National Taiwan University opened the doors of five of its graduate schools. ⇒ Valid Co-Reference				1	✓						
discourse connective	Leaning against his hip. He reclined with his feet up on the table. ⇒ when		1					✓				
discourse representation theory	This is an old story. We're talking about years ago. ⇒ Implicit Relation			8							✓	
next-sentence prediction	Sentence A, Sentence B ⇒ Valid Next Sentence		1						✓			
rethorical structure theory	The statistics quoted by the '' new '' Census Bureau report ⇒ Elaboration			6	8					✓		✓
sentence order	Given Sentence B, C, and D ⇒ C is at position 2	1							✓			

Table 7: Overview of resources and linguistic phenomena mapping for *discourse*. We give an illustrative example for each phenomenon (*indicates the right option, if options are given) and the number of datasets for the phenomenon by dataset type.

Connecting to the Few-Shot Learner Idea

Own Presentation for IML JKU: Language Models are Few-Shot Learners (GPT-3 paper)

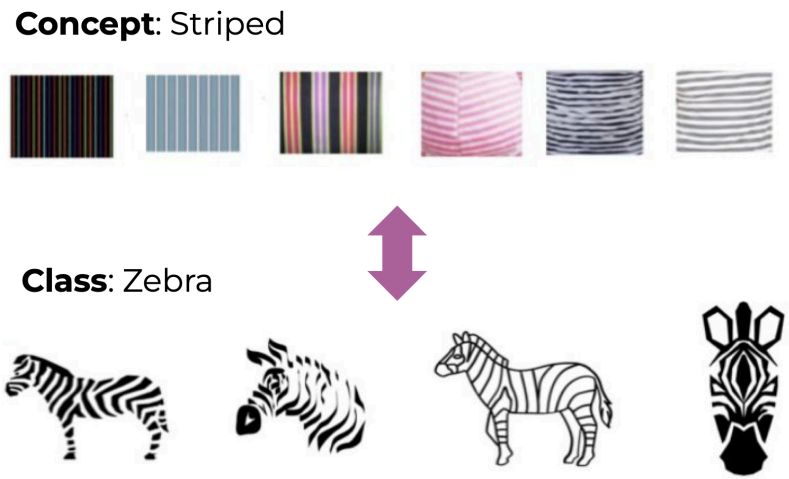
> "language models begin to learn […] tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText" (core translation example, developed from GPT-2 Unsupervised Multitask Learners paper)

We looked at some of this in the context of Making PDFs Accessible and Barrierfree already

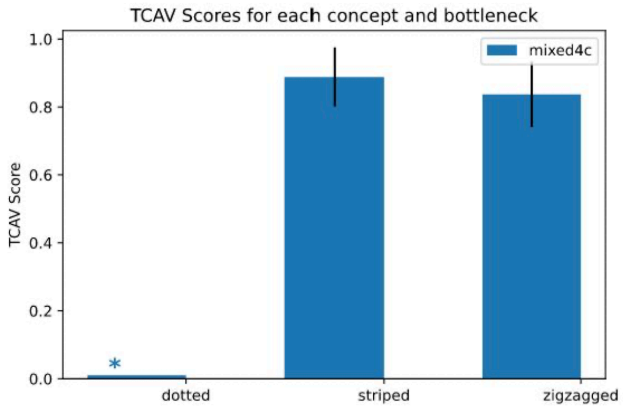
From the Linguistic Performance view: I see a connection to Lu et al. and the idea that "emergent abilities are not truly emergent, but result from a combination of in-context learning, model memory, and **linguistic knowledge**" especially functional linguistic abilities (the term in this paper) where the question here is if and to which extent emergent abilities are present in the linguistic performance/functional sense absent examples (they fall on no, not present) - opening this autonomy (safety issue!) problem space. I noted the appearance of the paper in the setup and evaluation section, noting on efforts to avoid few-shot examples

XAI Angle/Probing Idea (Another Connection)

Coming from other JKU work in Explainable AI (XAI)/I am seeing similar ideas picked up and explored in different areas of AI: Testing with **Concept Activation Vectors** (TCAV)



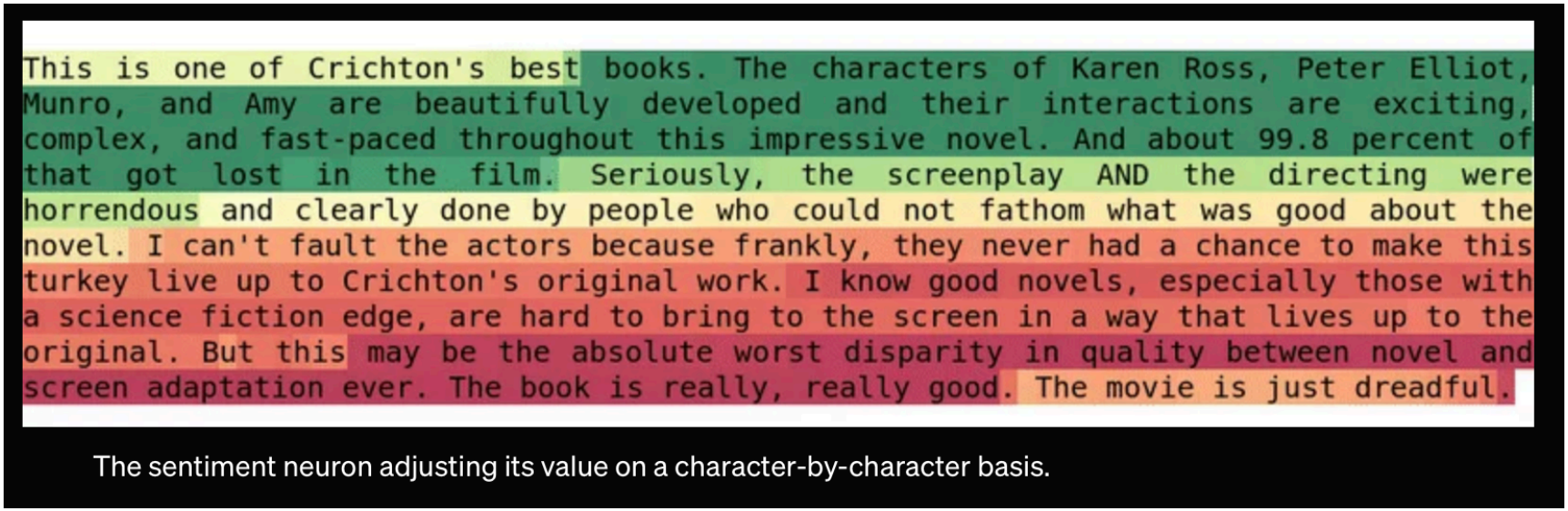
http://proceedings.mlr.press/v80/kim18d/kim18d.pdf



Molnar, Fig. 10.11

LSTM: (One More) Connection

Circling back to this idea of an isolated part of a model telling us a whole lot with regards to a pattern we are interested in: the single "**sentiment neuron**" that's highly predictive of the sentiment value - Radford et al. (2017) trained a multiplicative LSTM (Krause et al., 2016) for next character prediction on a corpus of Amazon reviews. They used this unsupervised model as an encoder by representing a review by the memory cell vector after the LSTM had processed it



The sentiment neuron adjusting its value on a character-by-character basis.

Last-Layer Questioning: Multi-Layer Probing?

So we are looking at the final layer of a representation, and, yes, while relevant pattern capturing can be super localized as in the single sentiment neuron example, I wonder how potentially complex linguistic phenomena might actually be represented at various levels (in the same model), and would be interested how a classifier on such a set of representations would fare

- in the given situation of testing for these five phenomena

- but also potentially other use cases that might more obviously require multiple (semantic) levels, say jokes, where jokes are often funny because they work on multiple levels, as a off-the-top-of-my-head example

I would be curious about research in this **potential research tangent** - turns out this gets picked up as a limitation at the end of the paper, see final slides

Evaluation

I was curious about the mean winning rate *mwr*, that is, a relative (to other LMs) measure: this was chosen in addition to standard deviation of the prob across seeds (robustness to noise) and performance measure macro F1/Pearson - but this makes sense in a cohort of LMs

Since also compression score/ratio and selectivity is considered

Possible limitations

Complexity of the benchmark result: but is **one score**/rate possible?

Selectivity*: "unambiguous" labels were chosen so as to allow for performance evaluation on randomized training signals. I wonder if a generalized, category-encompassing linguistic performance needs to be considered as well (lowering selectivity scores): would this be a different benchmark, i.e. multi-label linguistic phenomena? I am posing the question naively, I wonder if there is relevant background in linguistics, about interlace of syntax and semantics, say*

Limitations/Review

I thought the paper was convincing especially in the Results. Potential Limitations raised directly in the paper

- (English) language, non-multilingual: obvious one, easy to see why
- Last Layer Internal Representation: again obviously pragmatic, I already noted on this but will say this thought came pretty fast, as in why limit/positive, it would be super intriguing to explore multiple levels
- **Coverage**: this one I considered the least initially, coming up on the Limitations sections, but on a fundamental level “one fundamental aspect in composing a benchmark [is] acknowledging its incompleteness” (and Data Availability, Bias and Dataset Contamination)

I would close on this last point, that this is also a pretty good reason why one would need to be distrustful of an “AGI benchmark” as noted in the intro, and that this question of linguistic competence explored throughout actually spins the AGI question to begin with.

Food for thought and maybe comedic: an excellent soft-skills/well-spoken LM that has no as little knowledge as possible of the world, or, limited formal knowledge.

Next reading: Mahowald et al., Dissociating language and thought in large language models - this seems to go in this nice direction, and yes, likely requires “evidence from cognitive science and neuroscience,” remaining as interdisciplinary as ever!

References

Full presentation and context: https://heseltime.github.io/rDai#it-u

Third-party references throughout slides

Reminder: ***FlashHOLMES* take-home exercise suggestion**, as needed, formally or not, happy to take a look at this a bit further to get a real working knowledge