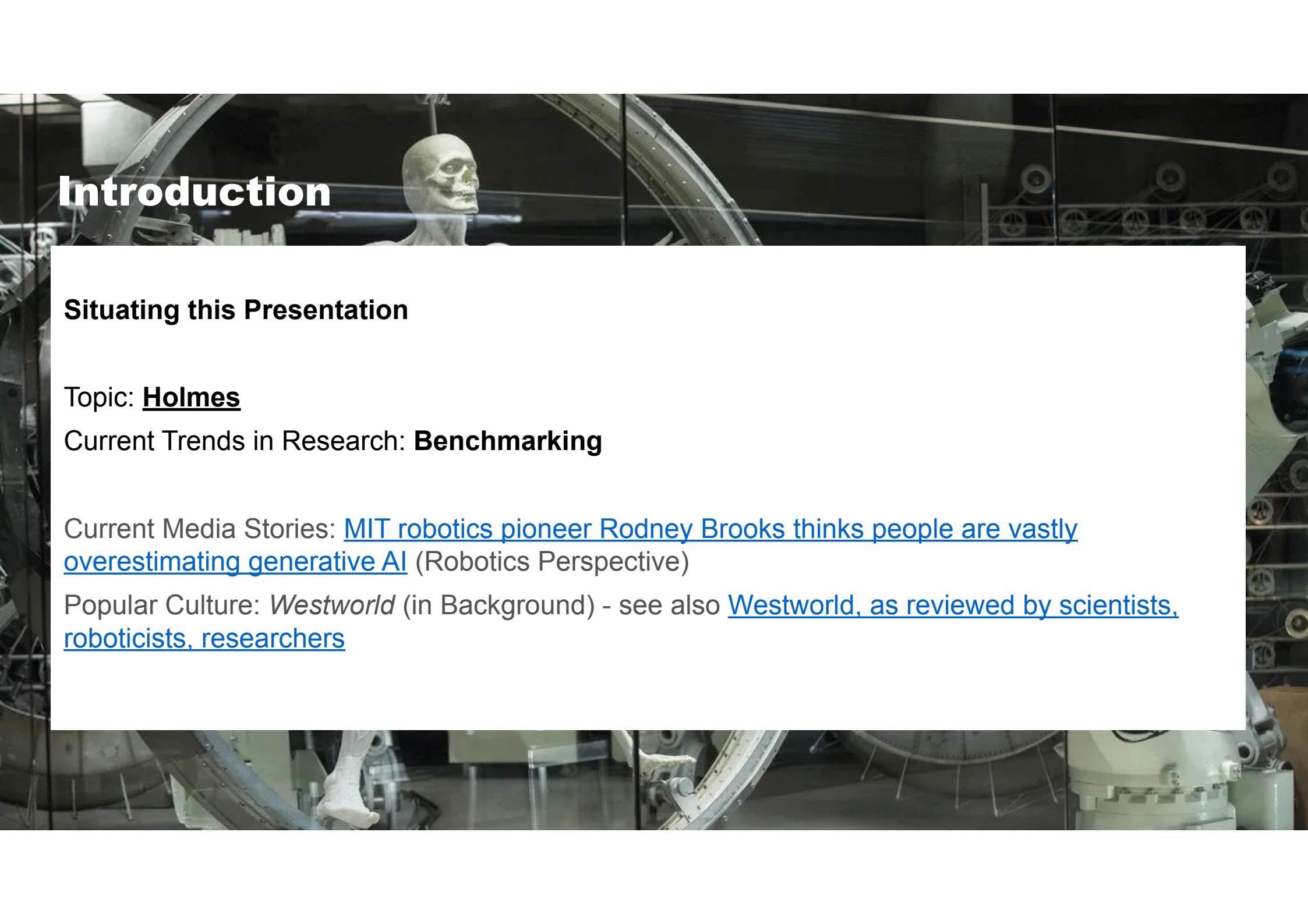


# LLM Linguistic Competence



**with a Focus on Benchmarking/HOLMES**

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



# Introduction

## Situating this Presentation

Topic: **Holmes**

Current Trends in Research: **Benchmarking**

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](#)

# Coming from Benchmarking: Linguistic Performance

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

**59 LLMs evaluated, 5 linguistic categories**

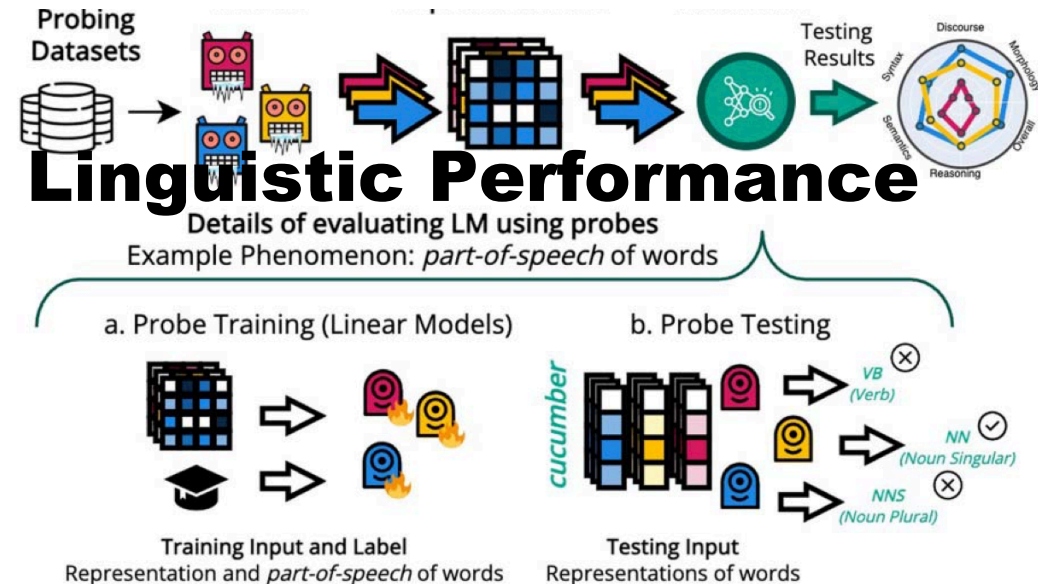
What is the evaluation?

(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 do my own *FlashHOLMES* test later - see next slide

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

**Goal: Training probes to predict linguistic phenomena:  
so that quality of prediction says something about the LM**



# Internal Representation: **DETAIL**

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

# 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:

# **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

# 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 ...

Activated sentiment neuron, this idea

# Technical Research: **FlashHOLMES** Test



# **Last-Layer Critique: Multi-Layer Probing?**

# Discussion I

# Discussion II

# Conclusion

# References

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

Third-party references throughout slides