UCLA Samueli
Computer Science

# Measuring the Psychological Depth of Language Models

**Fabrice**
Harel-Canada

**Hanyu**
Zhou

**Sreya**
Muppalla

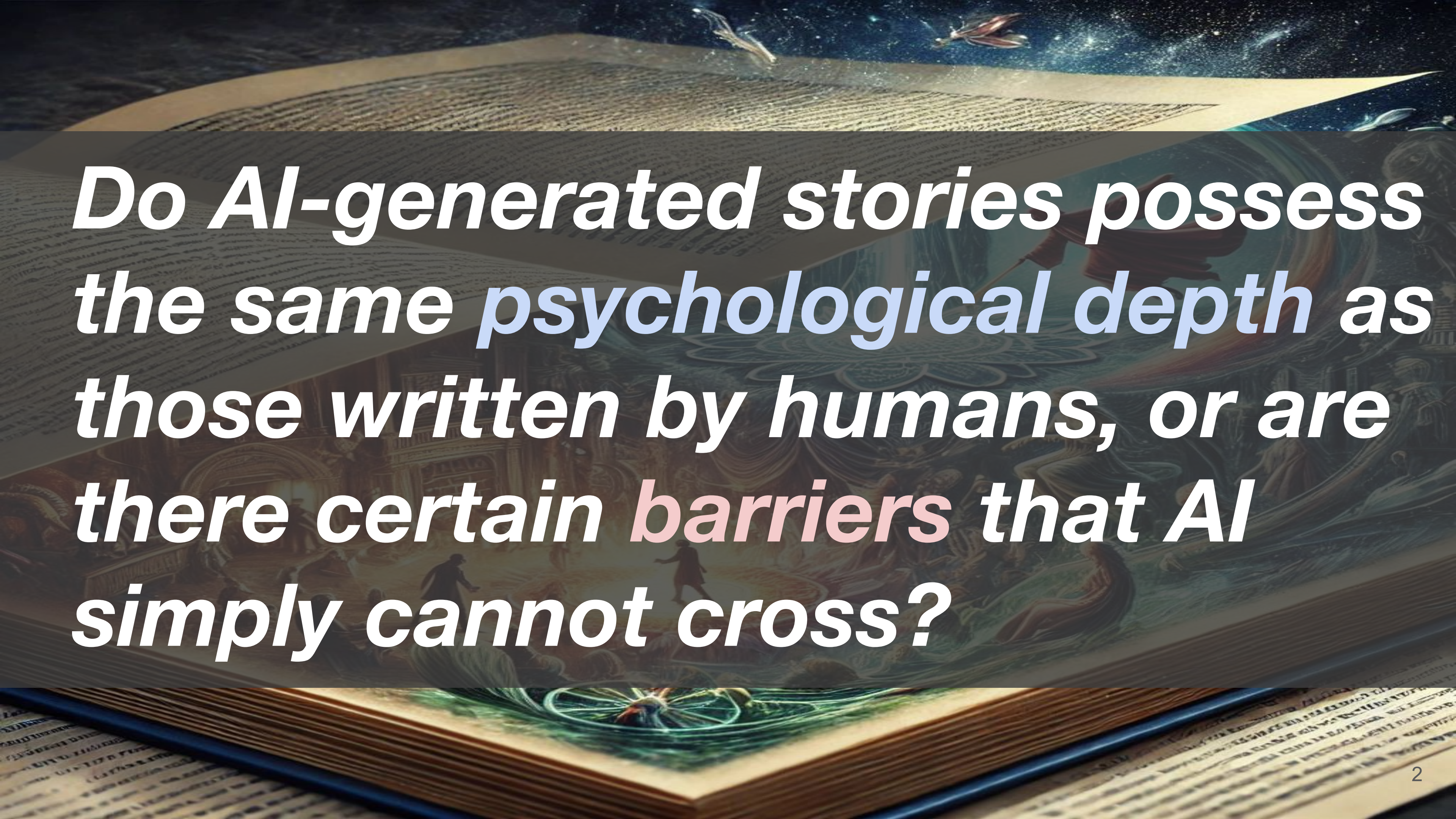
**Zeynep**
Yildiz

**Miryung**
Kim

**Amit***
Sahai

**Nanyun***
Peng

* equal co-advisors



# *Do AI-generated stories possess the same psychological depth as those written by humans, or are there certain barriers that AI simply cannot cross?*

# Why focus on stories?

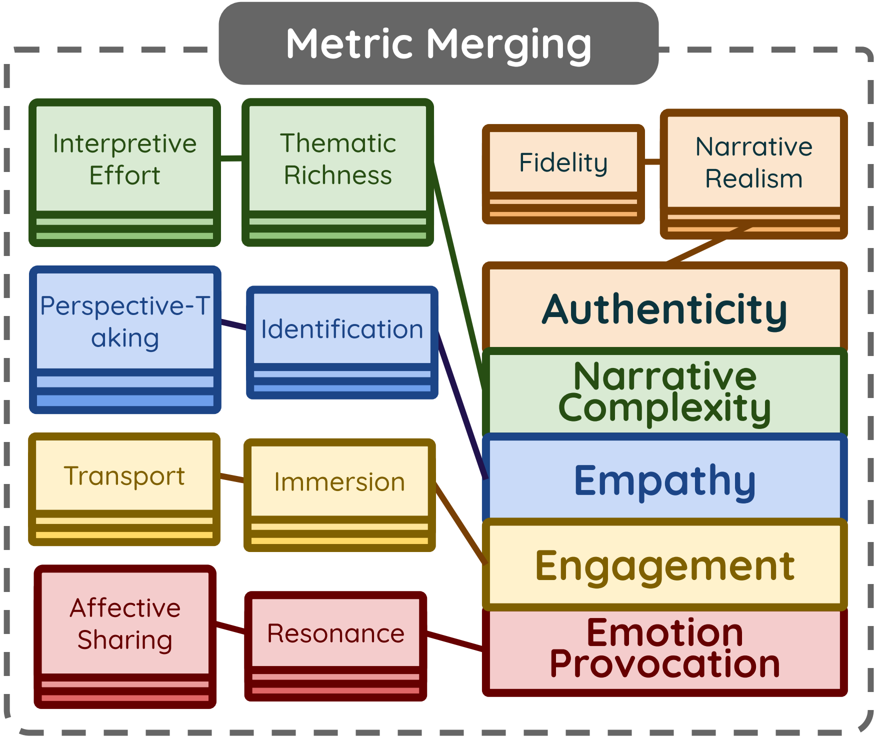
- Stories are **crucial for understanding** of ourselves and the world around us
- Writers of all skill levels are **using LLMs to augment** their stories
- Most of common metrics cannot answer this question because they focus on **objective** properties of the text: fluency, grammaticality, coherence, toxicity, bias, etc…
- Stories must be understood on a more **subjective** level

# Characterizing the Reading Experience

Reader-Response

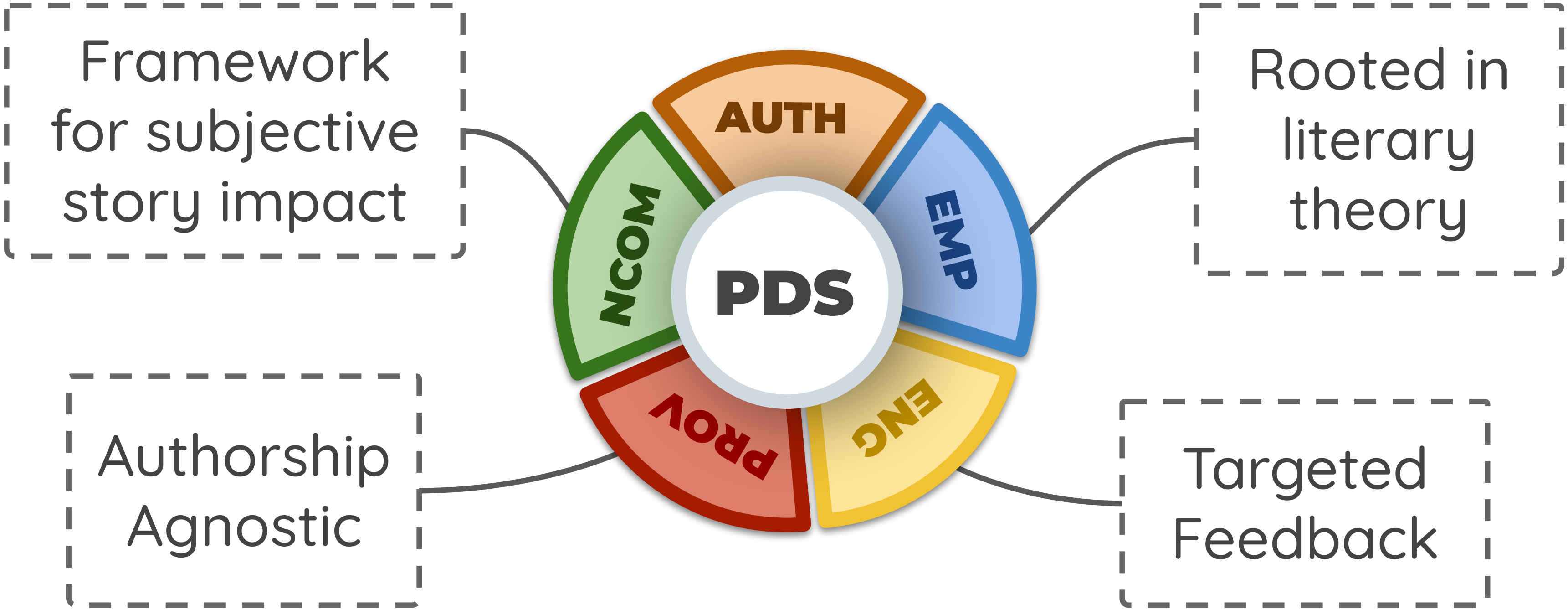
Text World

- Literature Survey
  - 95 scholarly articles & books
  - 143 candidates merged into 5 families via thematic analysis



Metric Merging

# The Psychological Depth Scale



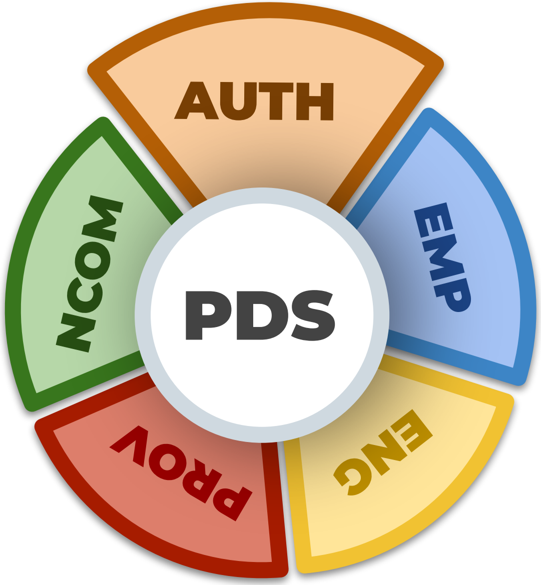
Framework for subjective story impact

Rooted in literary theory

Authorship Agnostic

Targeted Feedback

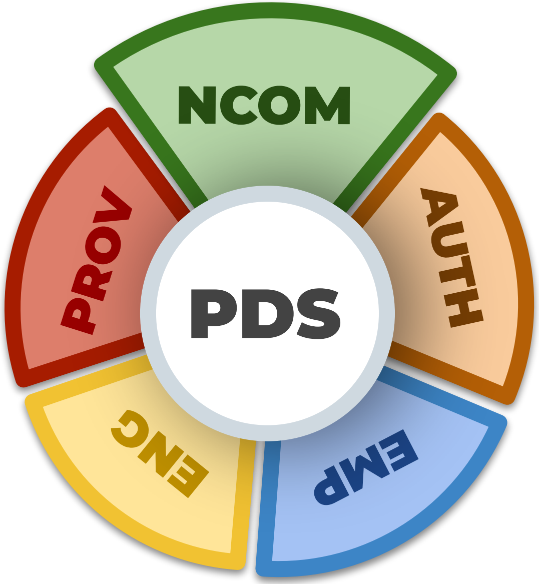
# What is Authenticity?



Expression of genuine human experiences and emotions that resonate with readers, reflecting real-life complexities and evoking a sense of connection

- Does the writing feel true to genuine human experiences?
- Does it represent psychological processes in a way that feels believable?

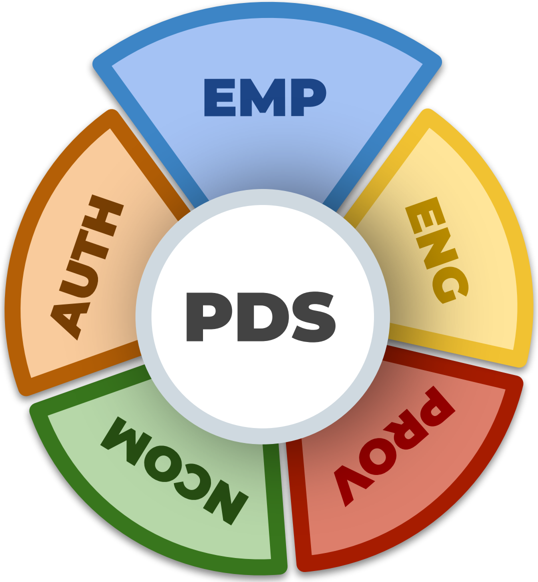
# And what about **Narrative Complexity?**



Presence of intricate storylines, character development, and layered perspectives, prompting cognitive mental restructuring

- Do characters exhibit multifaceted personalities and internal conflicts?
- Does the writing move beyond stereotypes or clichéd tropes?
- How deeply does the narrative explore relationships between characters?
- Does the story delve into intricate conflicts and their resolutions?

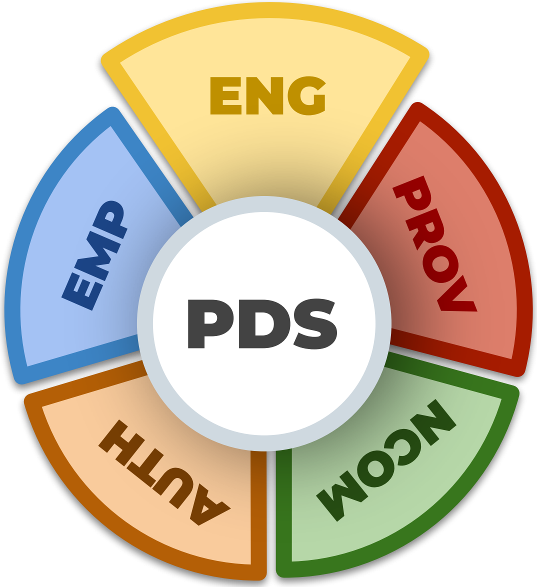
# And Empathy?



Ability to inspire readers to emotionally connect with characters, take on their perspectives, and engage in reflective insight

- Do you feel empathy for the characters and situations?
- Does the text lead you to introspection or new insights about yourself or the world?

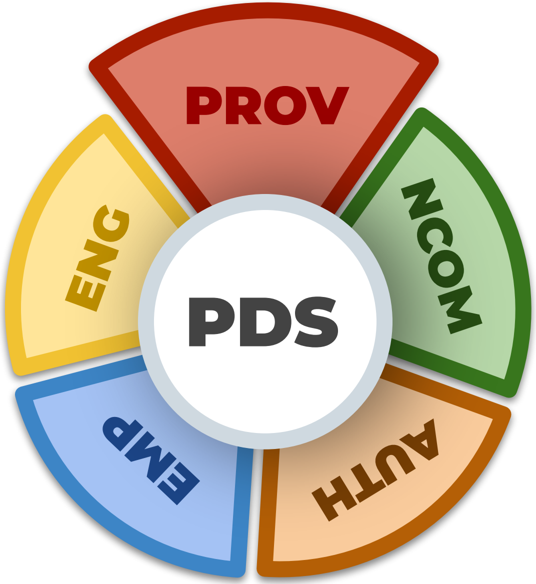
# There's Engagement, Too?



Ability to captivate and sustain the reader's attention, immersing them into the story

- Does the text engage you emotionally and psychologically?
- Do you feel compelled to continue reading?

# And Finally **Emotion Provocation**?

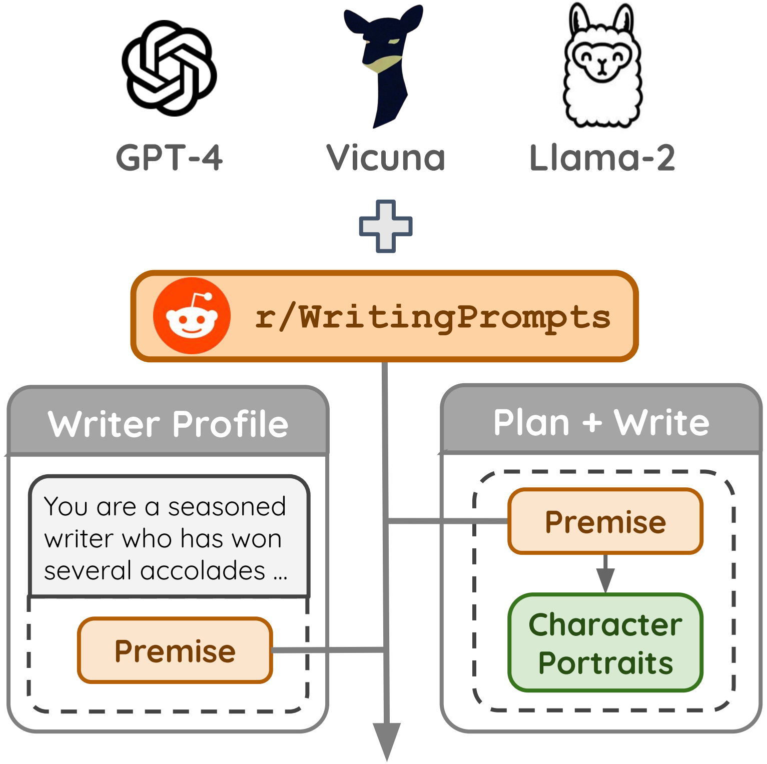


Capacity to elicit strong emotional responses, both **positive and negative**, by aligning emotional intensity and valence to maximize cognitive impact.

- Does the story explore the nuances of the characters' emotional states?
- Can the writing "show" a variety of emotions, rather than just "tell"?
- Do the portrayed emotions make sense within the story's context?

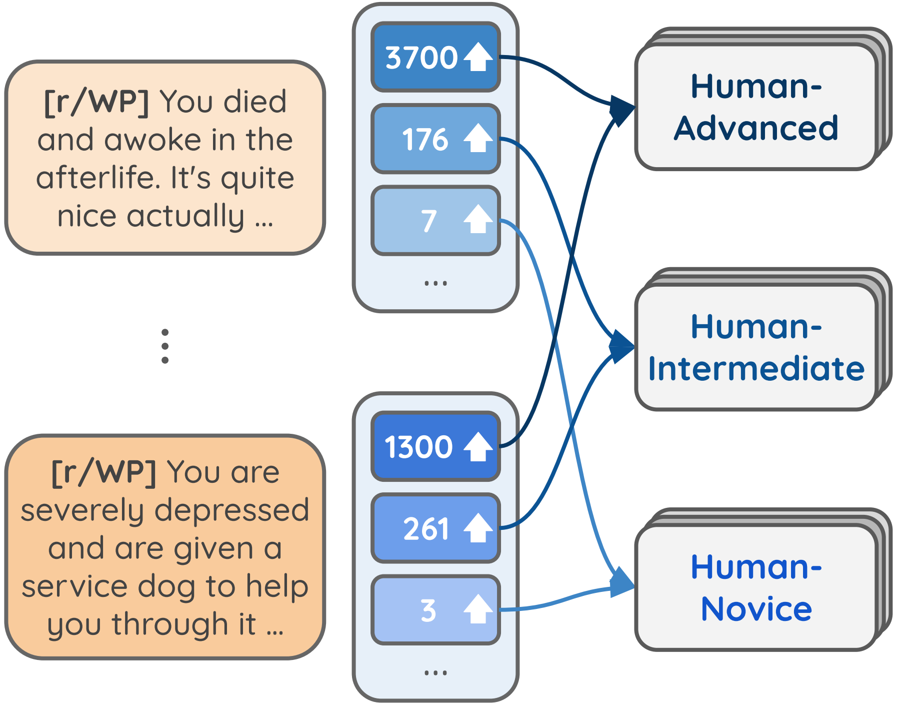
# Generating LLM Stories

- 5 models
  - `Llama-2-7B`
  - `Llama-2-13B`
  - `Llama-2-70B`
  - `Vicuna-33B`
  - `GPT-4`
- 2 prompting strategies
  - `Writer Profile`
  - `Plan + Write`
- 450 LLM-generated stories



# Collecting Human Stories

- **15 premises** collected from **`r/WritingPrompts`**
  - Posted after training data cutoff dates!
- **45 human-authored stories**
  - **3 stories per premise**
  - categorized to **three groups** based on their **upvote ranking**



**[r/WP]** You died and awoke in the afterlife. It's quite nice actually …
**[r/WP]** You are severely depressed and are given a service dog to help you through it …
3700
176
7
…
1300
261
3
…
Human-Advanced
Human-Intermediate
Human-Novice

# Empirical Study Setup

## Build Story Datasets

- 45 human + 450 LLM = 495 stories
- Sample 97 for human annotation

## Recruit Participants

- 47 applicants → selected 5 undergrads from English and Psychology departments @ UCLA
- Paid $100 each to read + rate

## Study Protocol

- Rate PDS on an 1-to-5 Likert scale
- Assess the likelihood of authorship from 1 (LLM) to 5 (human)
- Optionally justify ratings

## Human Judgements

- 2,425 psychological depth ratings
- 485 authorship likelihood ratings
- 1,128 free-form justifications

## RQ1

# PDS Reliability

How consistently can well-informed humans judge psychological depth?

0.72

Krippendorff’s **α**

**Takeaway**

The PDS enjoys a substantial degree of **consensus** and is a **coherent** framework for evaluating the depth of short stories

# RQ2

# Automated Measure

To what extent can psychological depth be measured automatically?

# Can LLMs mimic human depth judgements?

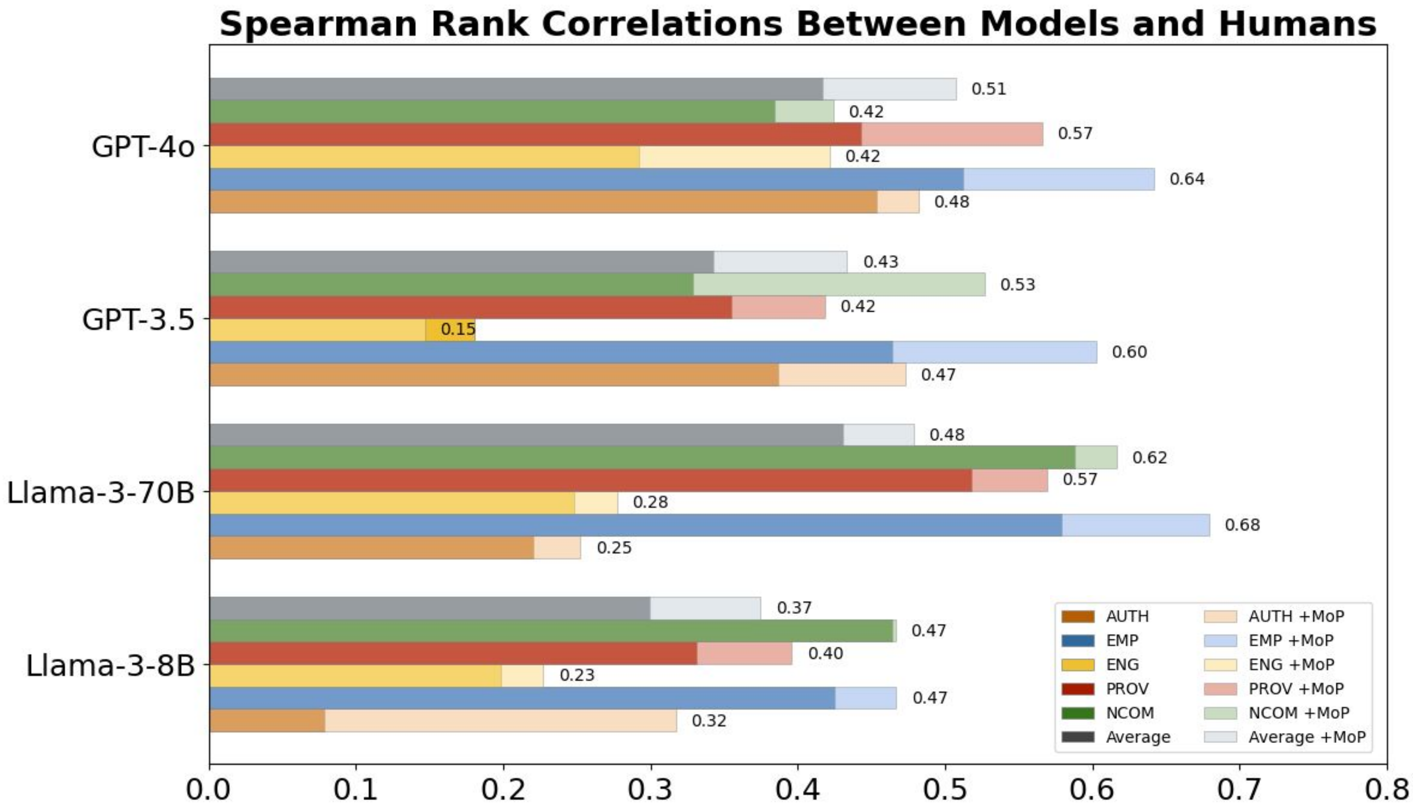
- Human labor is expensive
- Annotation is often boring

**We compared:**

- Vanilla zero-shot baseline
- Mixture-of-Personas (MoP)
  - Repeated the zero-shot rating with N = 5 different personas
  - Targeted roleplay of generic persons with specific textual analysis skills

**Metric:** Spearman Rank Correlation with human judgements

| | |
|---|---|
| Authenticity | ↑ **33.81%** |
| Empathy | ↑ **20.74%** |
| Engagement | ↑ **16.93%** |
| Emotional Provocation | ↑ **18.34%** |
| Narrative Complexity | ↑ **15.22%** |
| Average | ↑ **20.43%** |



## Takeaway

**GPT-4o** best avg corr **@ 0.51**

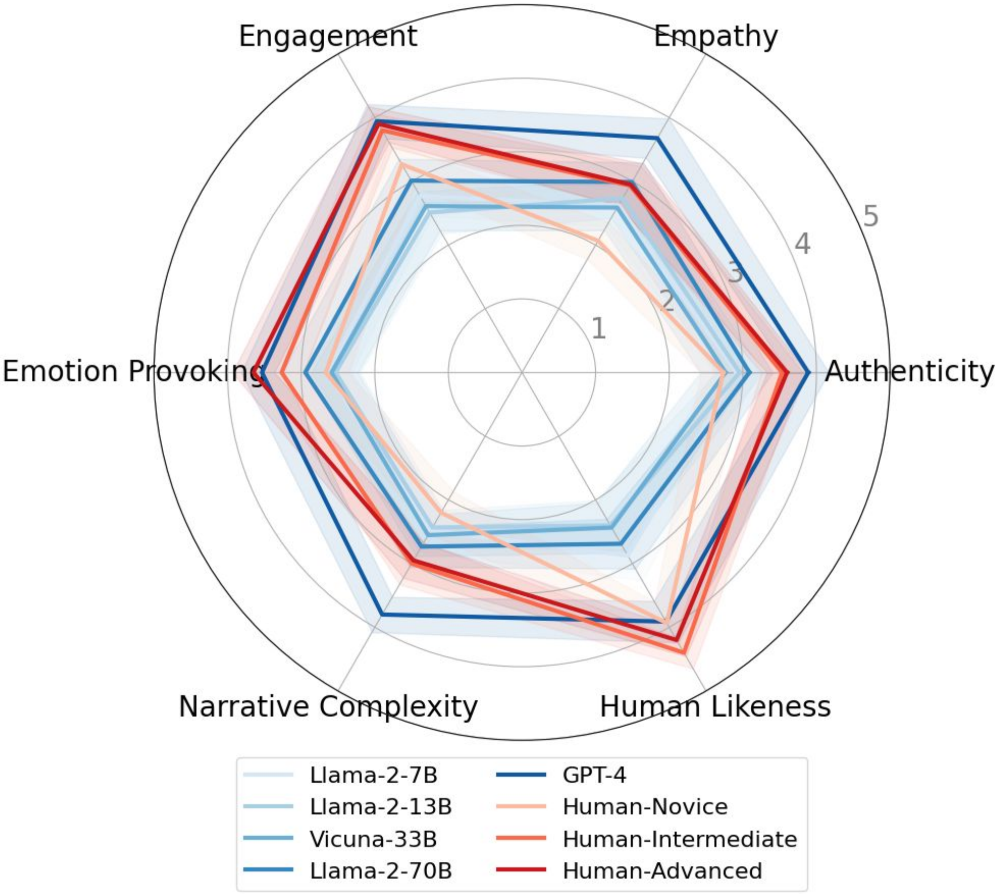
**Llama-3-70B** best overall corr **empathy @ 0.68**

Results are comparable to previous works like G-Eval (0.51 corr)

## RQ3

# LLM Capabilities

How do stories written by amateur humans and LLMs manifest psychological depth?



Takeaway

GPT-4 **significantly surpassed** popular human stories on **narrative complexity** and **empathy,** while being indistinguishable on all other PDS components.

# LLMs Stories are Hard to Detect

informality

"He gave me an emote" sounds like a very human thing to write.

grammaticality

...there are a lot of (usually incorrectly used) semi-colons, which is an error I see human authors make, so I'm more inclined to think this was written by a human...

creativity and nuance

The story exhibits a high level of creativity, emotional depth, and nuanced exploration of philosophical concepts, suggesting it was likely written by a human.

humor

I think this joke is only something that humans would get or would find funny

**Takeaway**

On average, participants identified **human vs. LLM authorship** with only **56% accuracy**, which dropped to **27%** for **GPT-4** stories.

# Takeaways and Next Steps

**PDS** is **validated**, **automated**, and **systematic** means of measuring the capacity of **LLMs to connect with humans** through the stories they tell

**Short stories can be effectively written by LLMs**

**What about other types of creative works?**

- Longer stories - novelettes, books
- Screenplays + comedy scripts
- Magazine articles
- Poems and more!

**We hope you'll help explore these questions in the future!**

https://github.com/PlusLabNLP/psychdepth

SCAN ME

website / paper link

# Thank You!

## Questions?