UCLA
Computer Science

Measuring Psychological Depth in Language Models

Fabrice Harel-Canada, **Hanyu** Zhou, **Sreya** Muppalla, **Zeynep** Yildiz
Miryung Kim, **Amit** Sahai*, **Nanyun** Peng*
* equal co-advisors

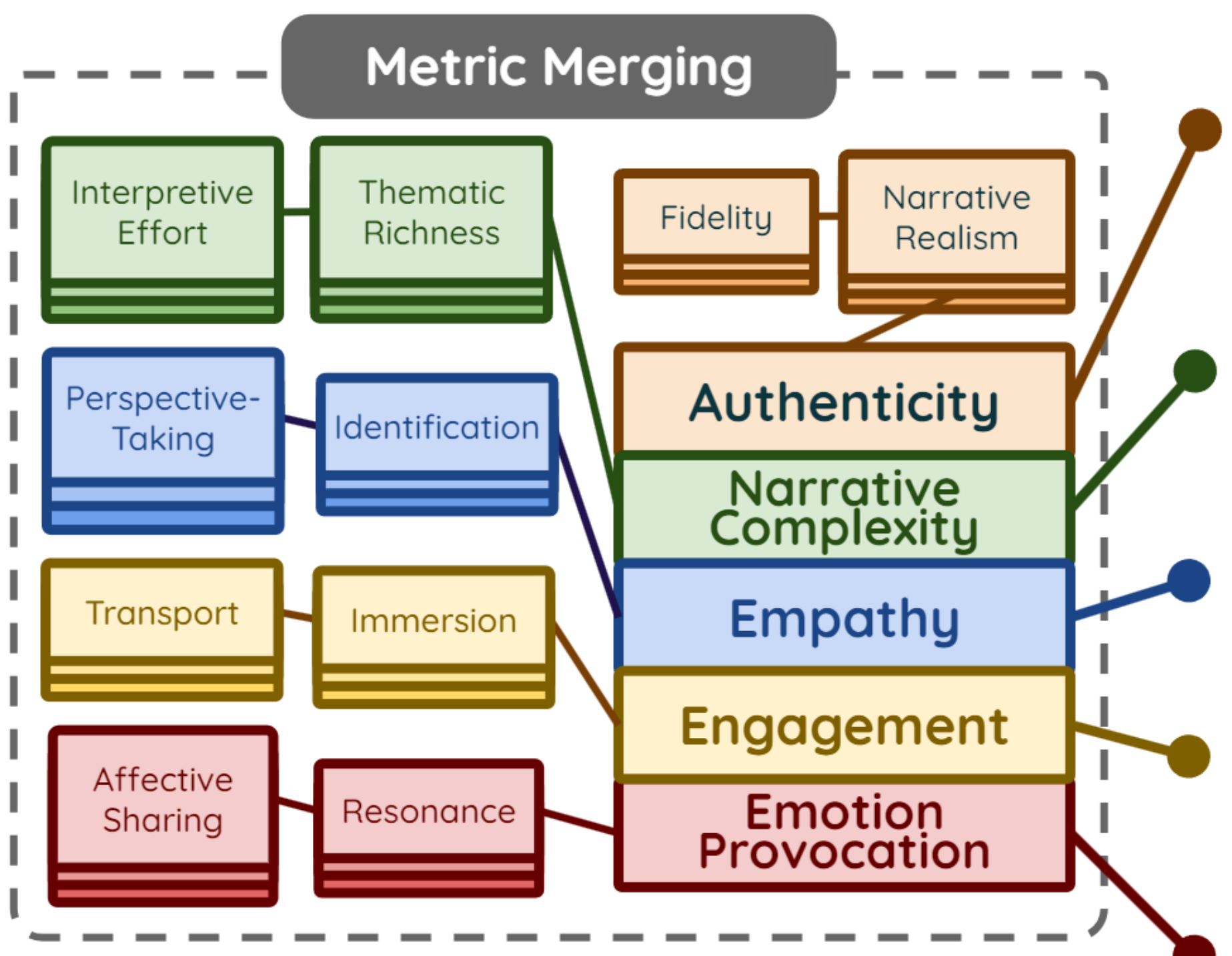
SCAN ME
paper link

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

Reader-Response

Text World

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



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

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

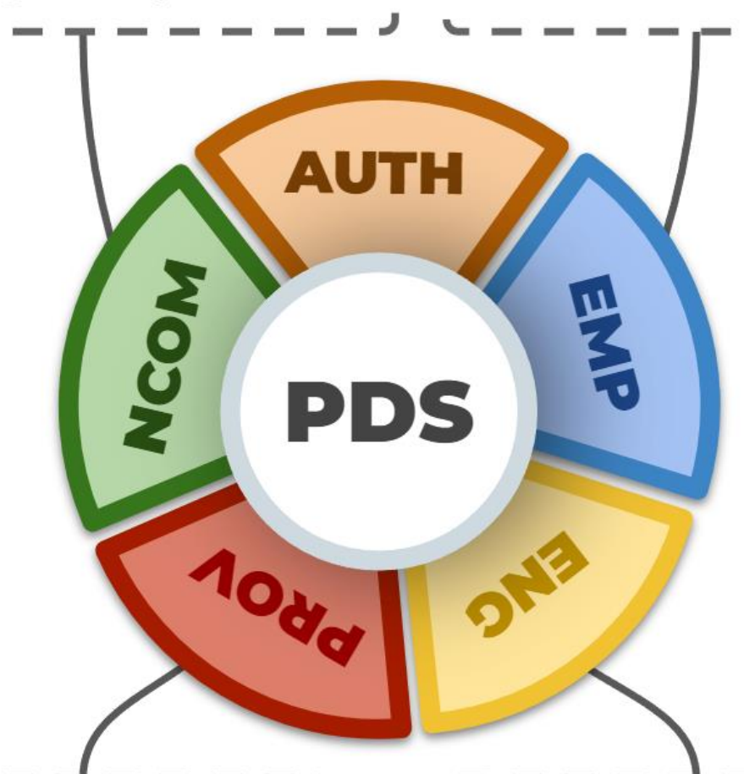
Ability to inspire readers to connect emotionally with characters, take on their perspectives, and engage in reflective insight

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

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

Framework for subjective story impact

Rooted in literary theory



Authorship Agnostic

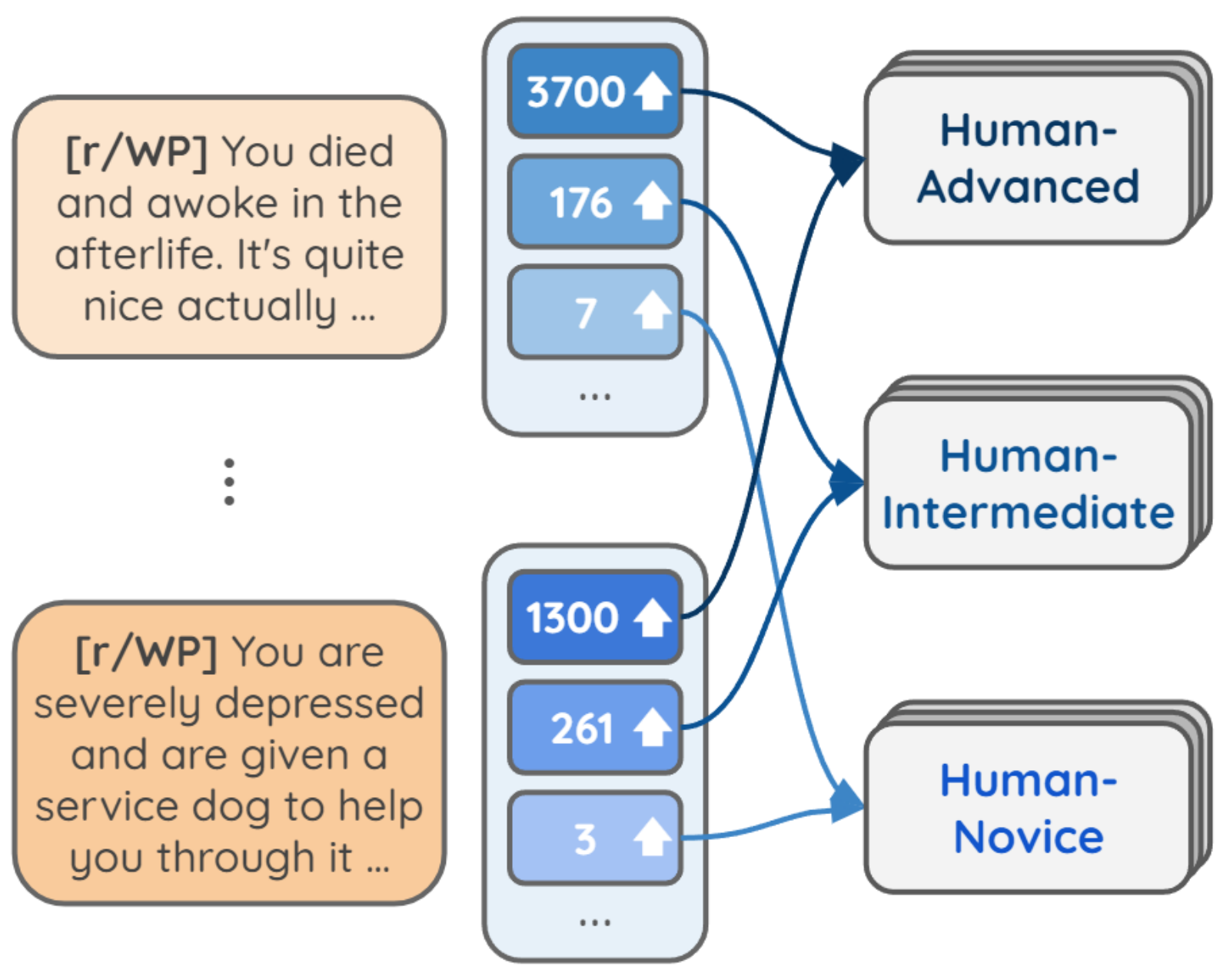
Targeted Feedback

Collecting Human Stories

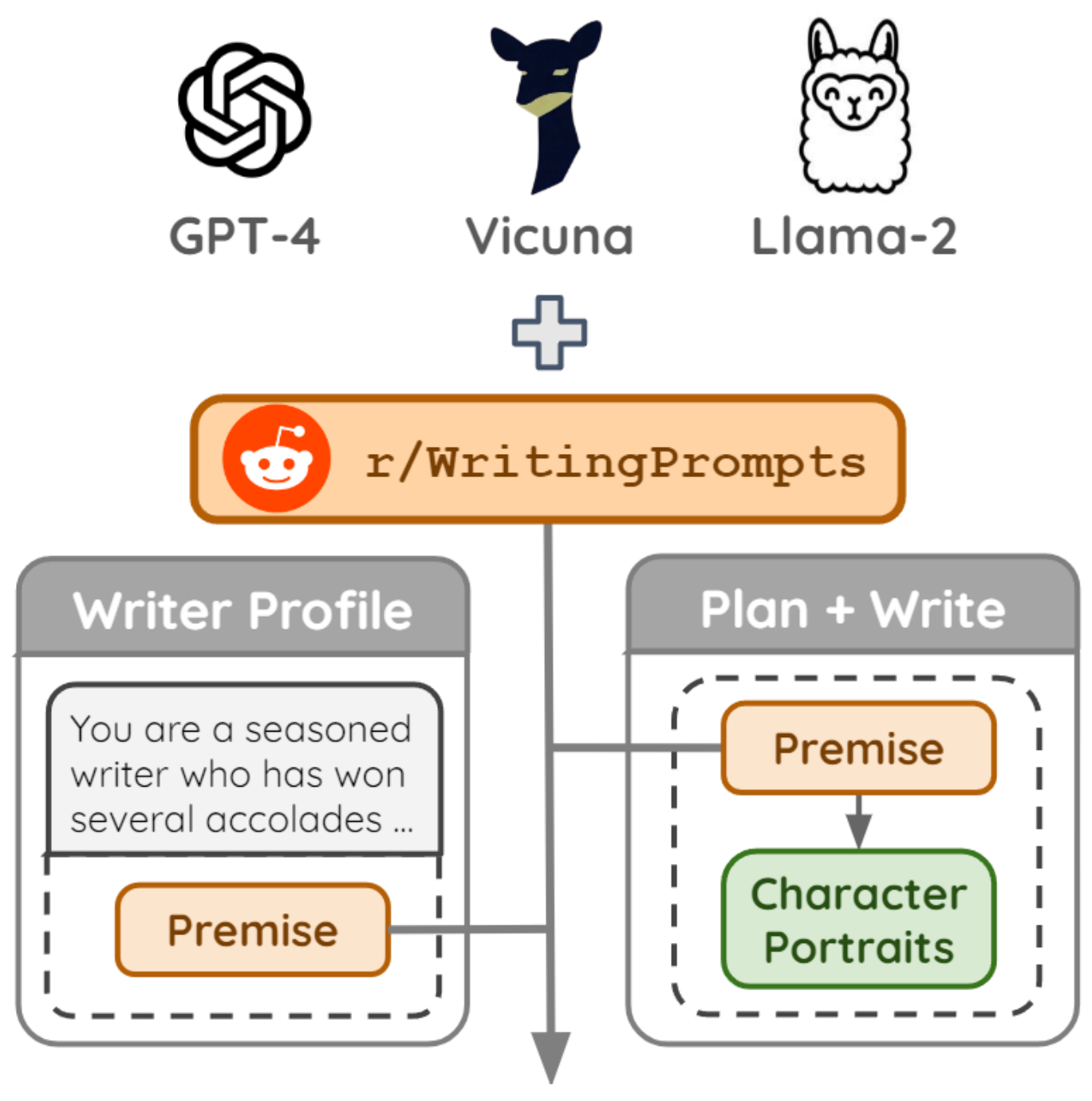
- **15 premises** collected from `r/WritingPrompts`
  - Posted after training data cutoff dates!
- **45 human-authored stories**
  - **3 stories per premise**
  - categorized to **three groups** based on their **upvote ranking**

Build Story Datasets

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


[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

Generating LLM Stories



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

Recruit Participants

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

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

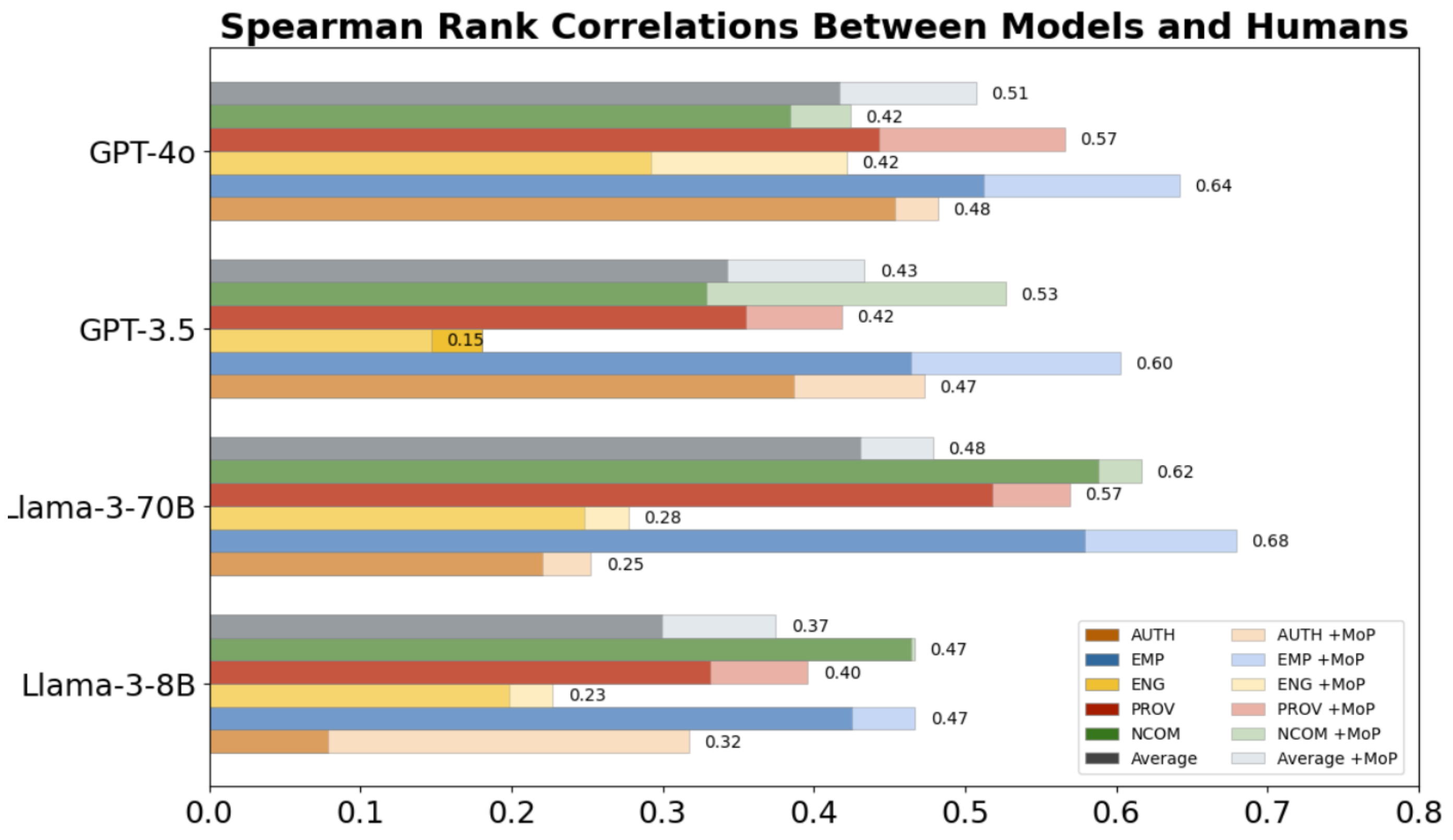
RQ2
Automated Measure

To what extent can psychological depth be measured automatically?

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

Spearman Rank Correlations Between Models and Humans



Takeaway

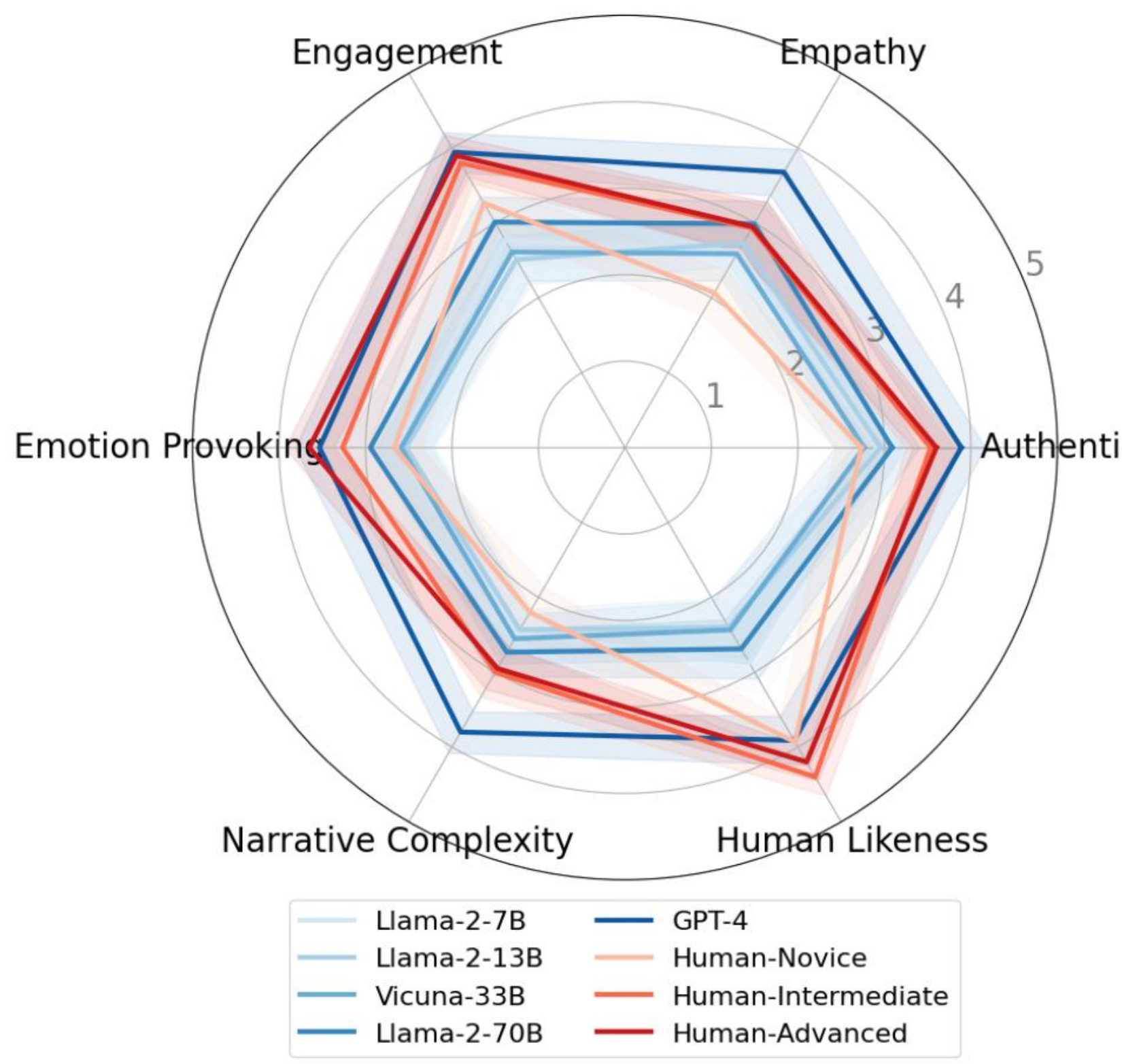
GPT-4o best avg corr **@ 0.51**

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.

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.