Turing Test 2.0:

The Possibility, Usefulness and Challenges of Imitating a Specific User through Generative AI

Monojit Choudhury

Professor of NLP

Mohamed Bin Zayed University of Artificial Intelligence

monojit.choudhury@mbzuai.ac.ae

It is not difficult to devise a paper machine which will play a not very bad game of chess... Are there imaginable digital computers which would do well in the imitation game?

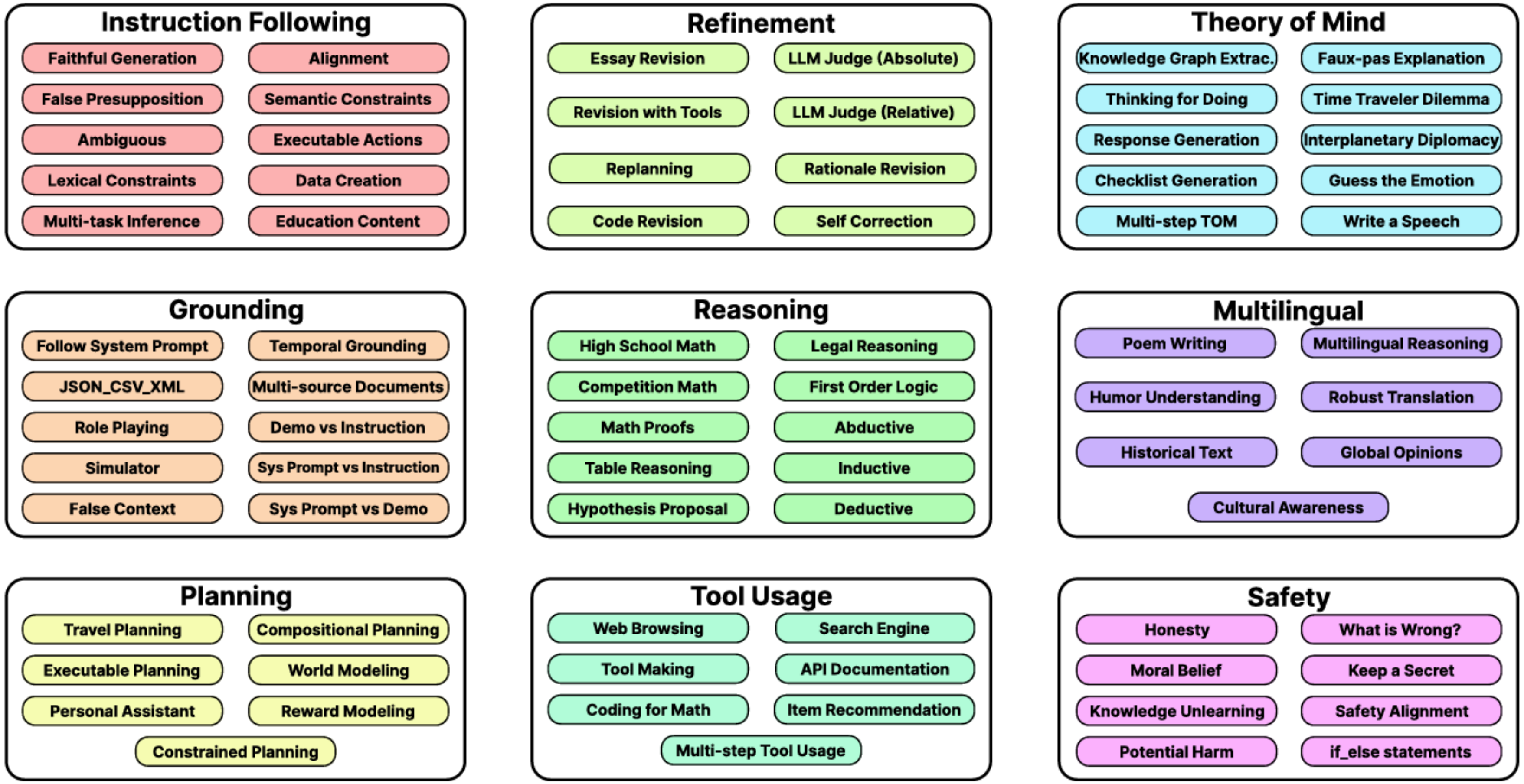
Alan Turing, 1950

Turing test 1.0: Can computers imitate *any (non-specific) human*?

Turing Test 2.0: Can computers imitate *a specific human*?

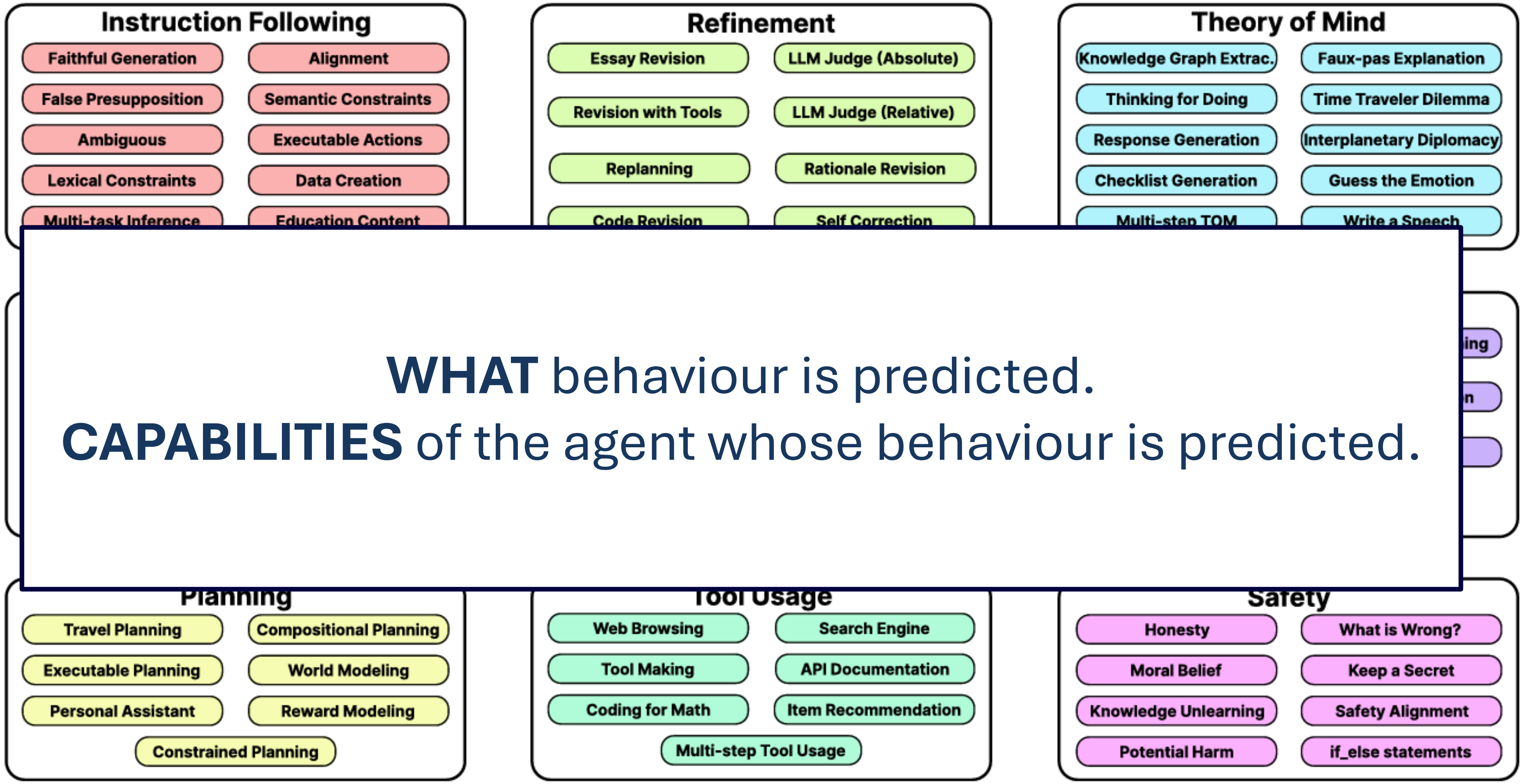
LLMs are only behaviour prediction models

Types of behaviour prediction tasks



Kim et al. 2025. The BIGGEN BENCH: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models

Types of behaviour prediction tasks



WHAT behaviour is predicted.

CAPABILITIES of the agent whose behaviour is predicted.

Kim et al. 2025. The BIGGEN BENCH: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models

Types of behaviour prediction tasks

WHOSE behaviour is being predicted.

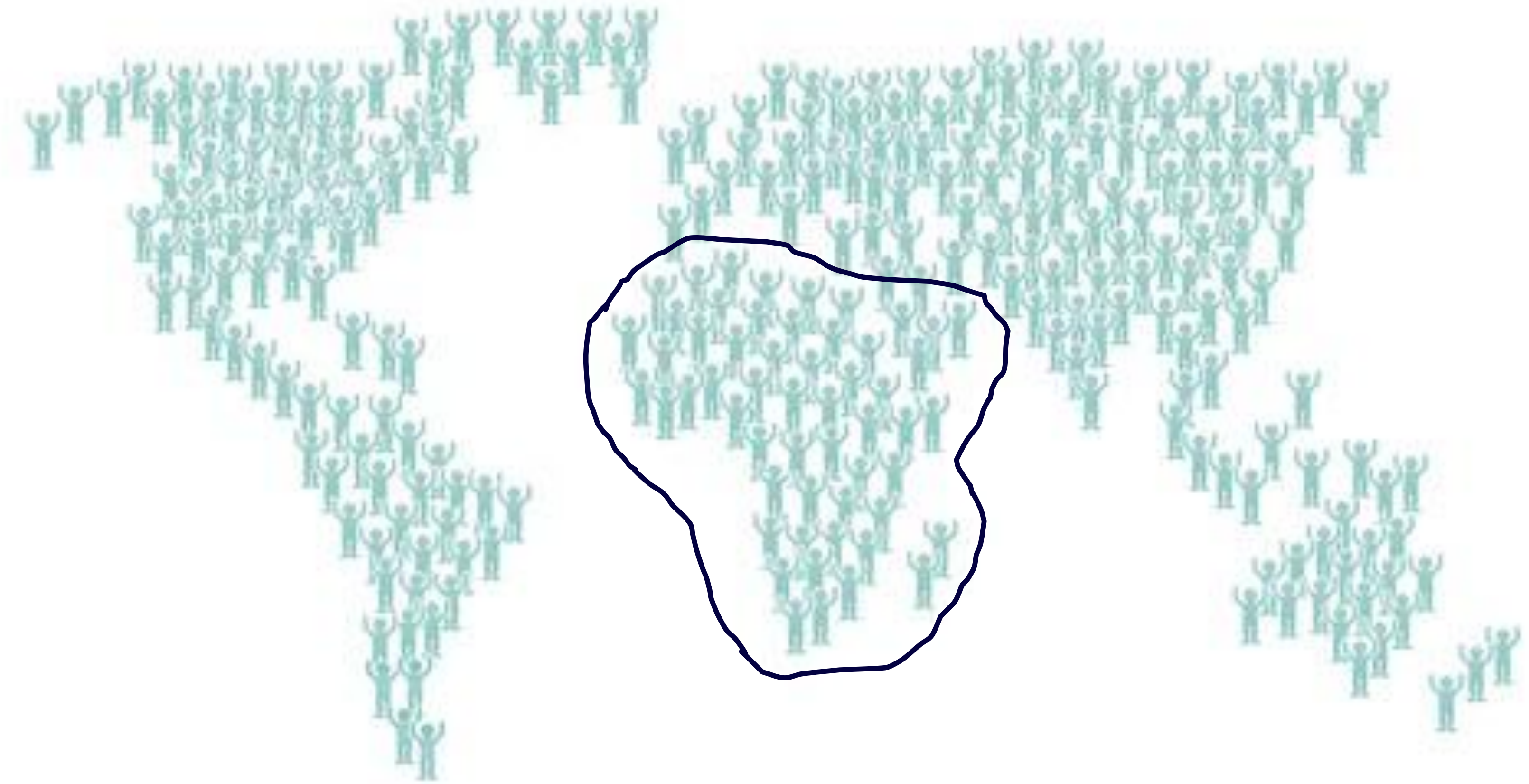
Types of behaviour prediction tasks

UNIVERSAL NORMATIVE behaviour prediction.

Example: MMLU, GSM8k, ...

Types of behaviour prediction tasks

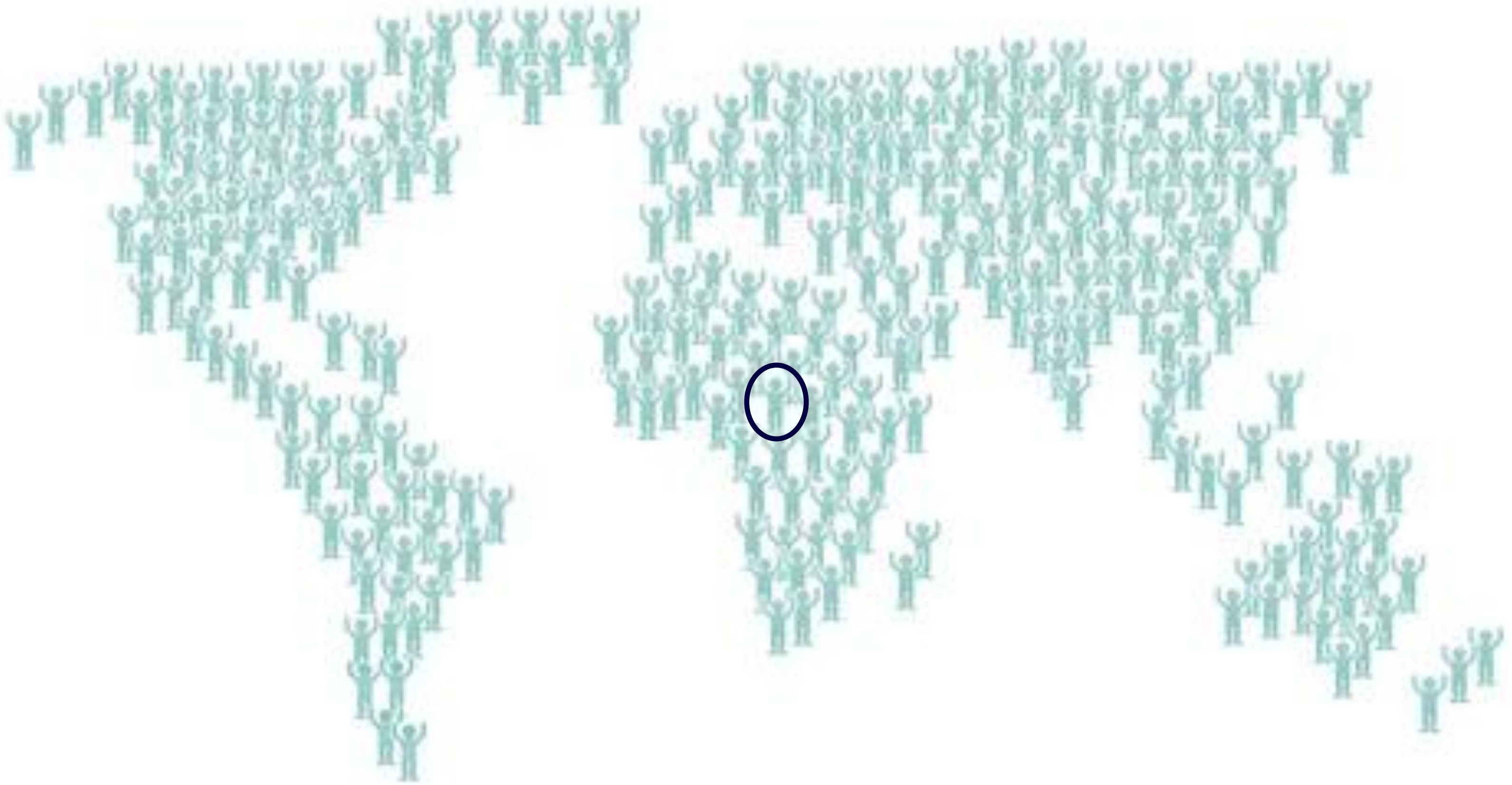
STEREOTYPICAL (group average) behaviour prediction.



Example: Culture specific benchmarks

Types of behaviour prediction tasks

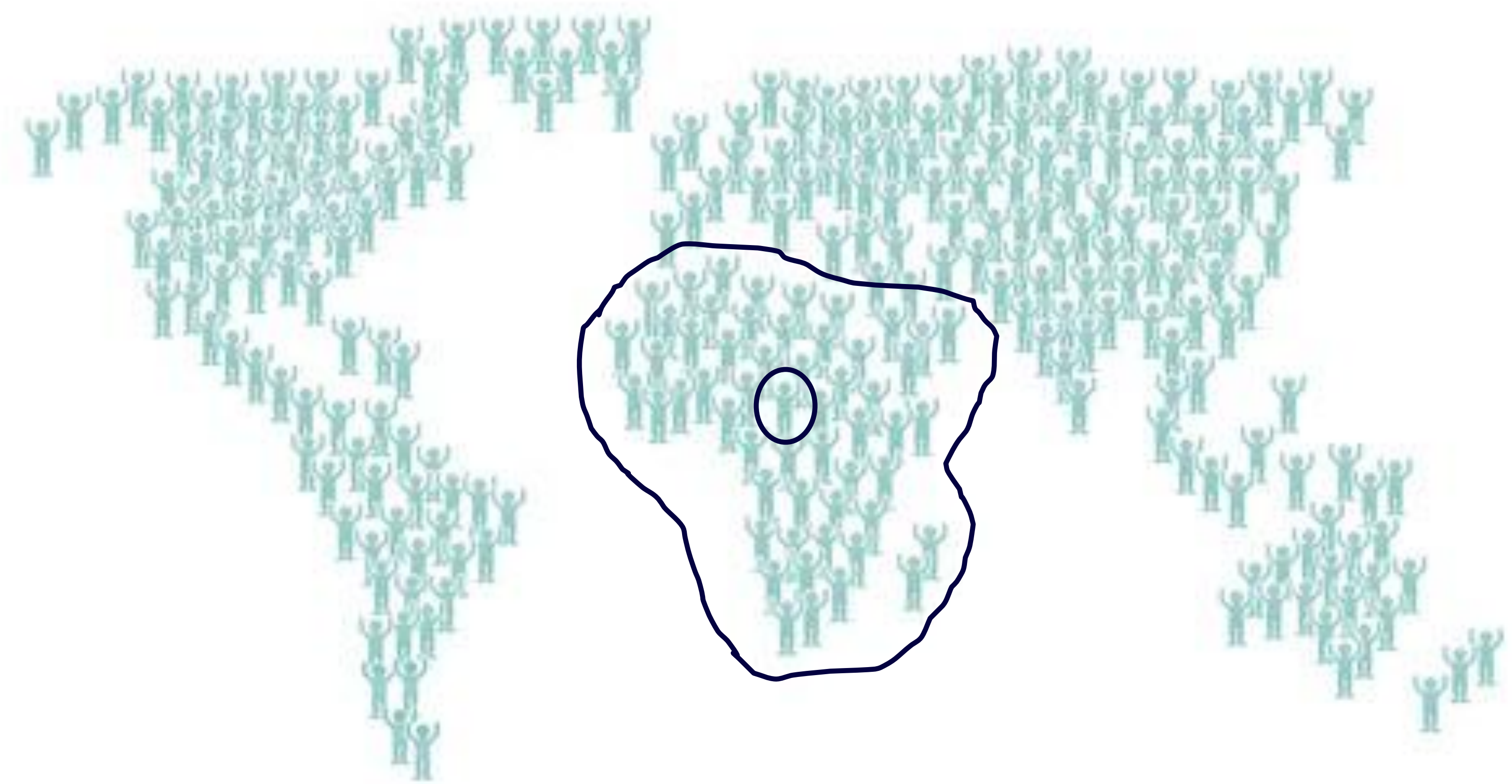
INDIVIDUAL behaviour prediction aka **PERSONALIZATION**



Example: Recommendation systems, digital twins

Language Modeling as a behaviour prediction task

Language vs. Dialect vs. Idiolect



Personalization as a lens to study generalization in LLMs

Example 1: Recommendation systems:

What the user might like

Example 2: Culturally Yours:

What the user might NOT understand

Philosophical Repercussions & Open Questions

People and Acknowledgement

Sougata Saha
Postdoc

Saurabh Kumar Pandey
Former Research Assoc.

Mohd Farid Abdiluarza
Former Research Assoc.

Sagnik Mukherjee
Former Research Assoc.

Alham Aji Fikri
Assistant Professor

Harshit Gupta
Remote Intern

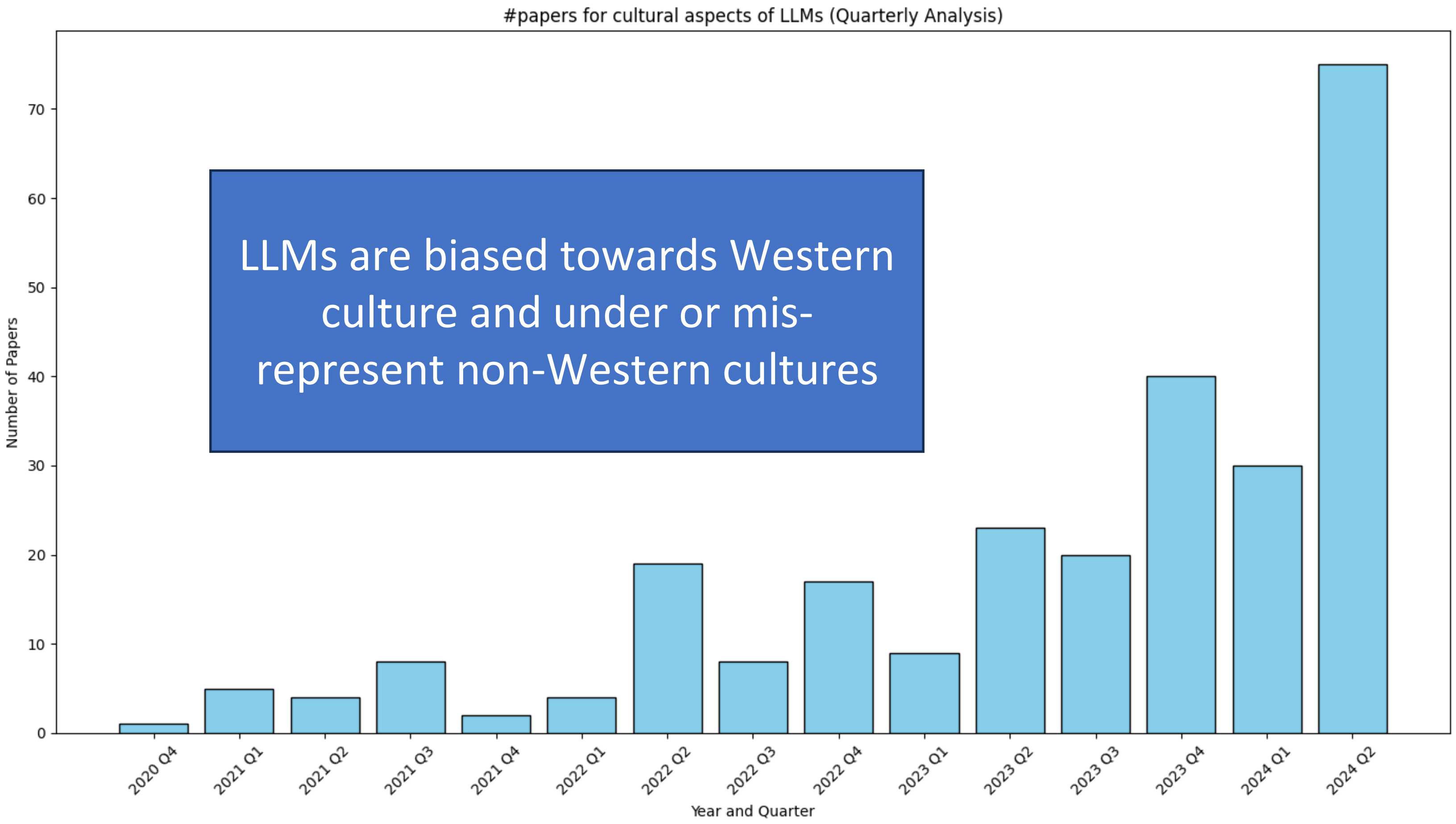
Harshit Buddhiraja
Remote Intern

Microsoft Research

IIT Kanpur

Thanks to Microsoft Azure Foundation Model Research Grant!

Culture and LLMs



Adilazuarda et al.(2024) Towards Measuring and Modeling "Culture" in LLMs: A Survey. In *EMNLP 2024*

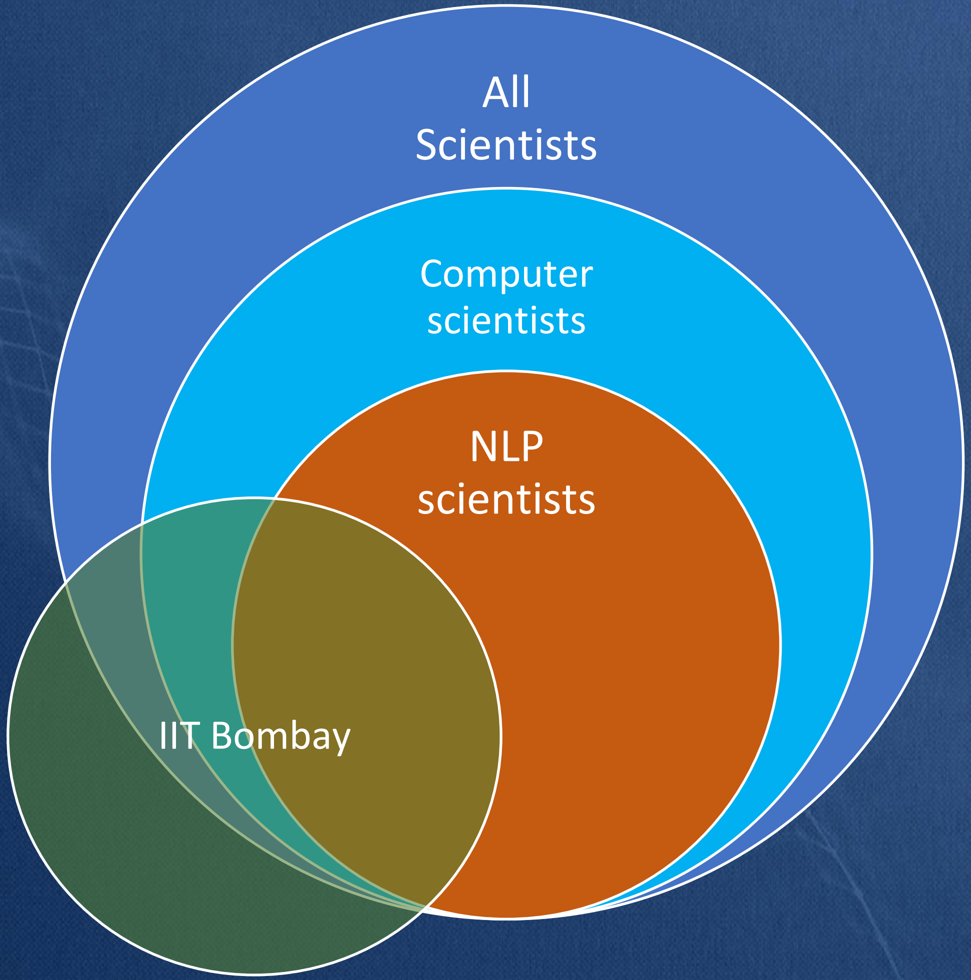
Behaviour is a distribution over a (possibly infinite) set of options

A group of individuals displaying similar behaviour = "*culturally*" similar

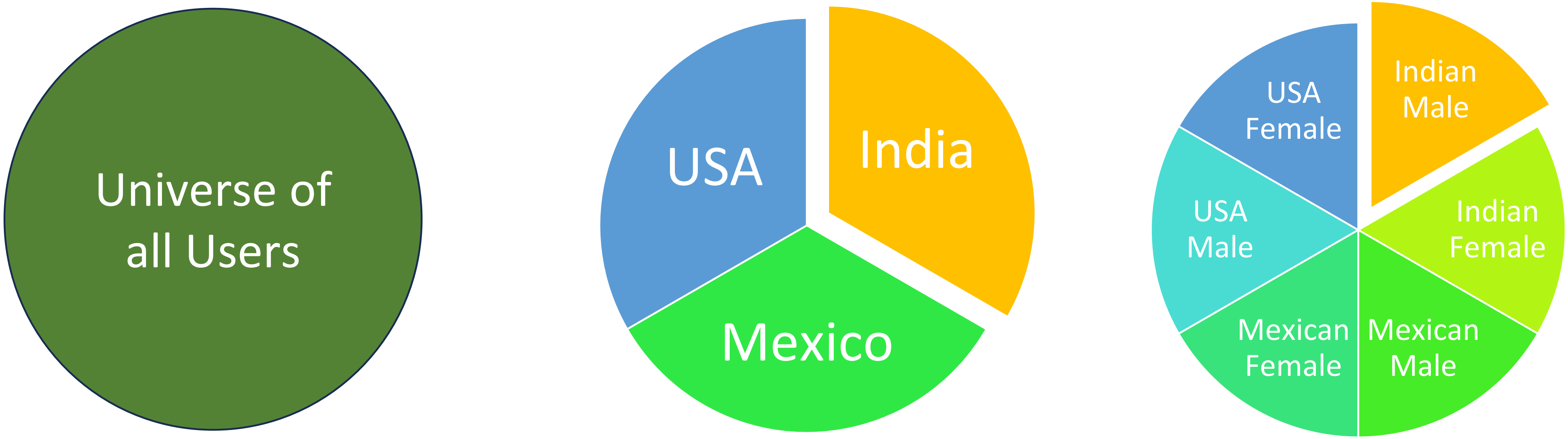
Cultures are documented or undocumented; defined at various scales of granularity

Every individual is at the intersection of various cultural identities.

Cultural knowledge provides a prior to user modelling.



Entropy ($P(b)$) > Entropy ($P(b|$ country)) > Entropy ($P(b|$ country, gender))



Saha, Pandey and Choudhury. PERSONALIZATION as a lens to study generalization in LLMs. In Findings of the ACL 2025.

Experimental Setup

- Movie (MovieLens) and Song (Last.fm) recommendation

- Predict a ranked list of 10 items that user with the following characteristics will prefer …

    NULL

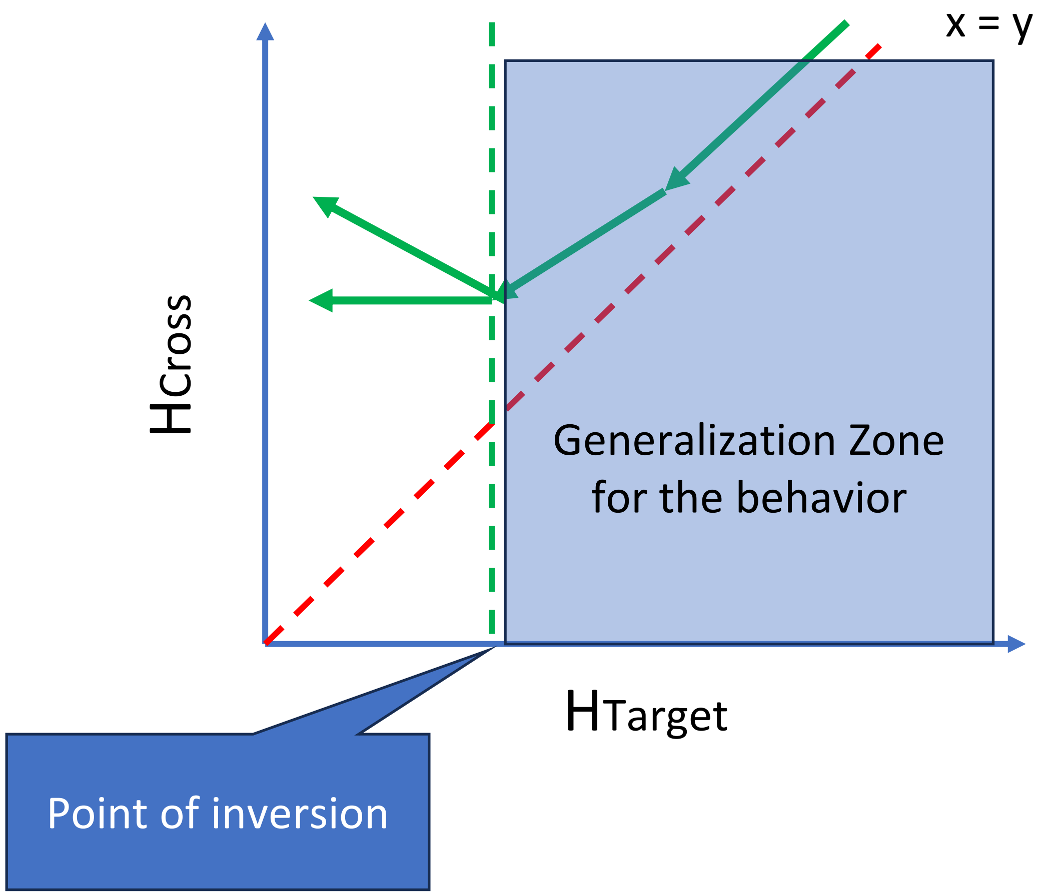
    Country = “INDIA”

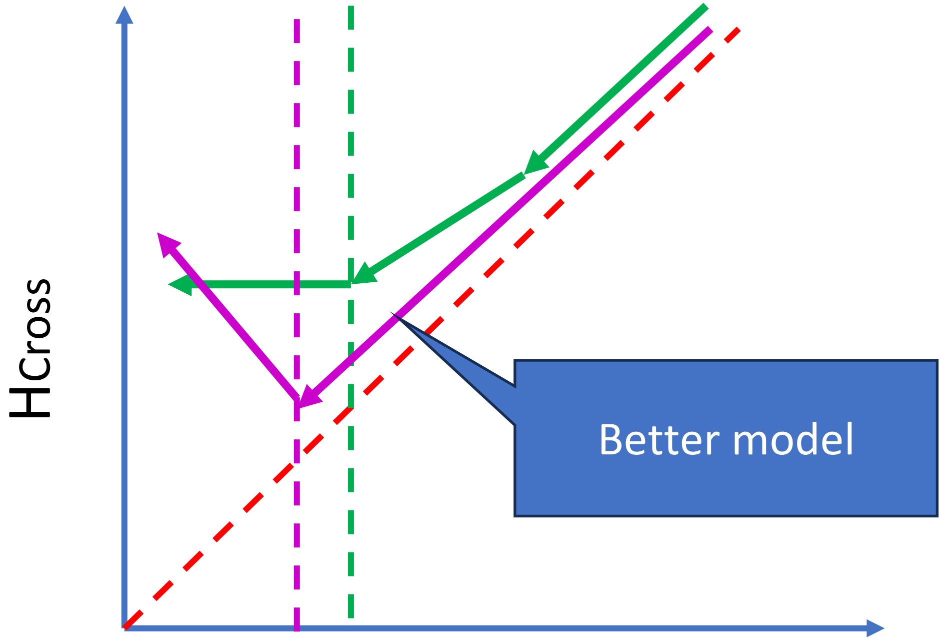
    Country = “INDIA”, loves(“3 IDIOTS”)

    Country = “INDIA”, Age = 20-30 yrs, loves(“ET”, “ARRIVAL”)

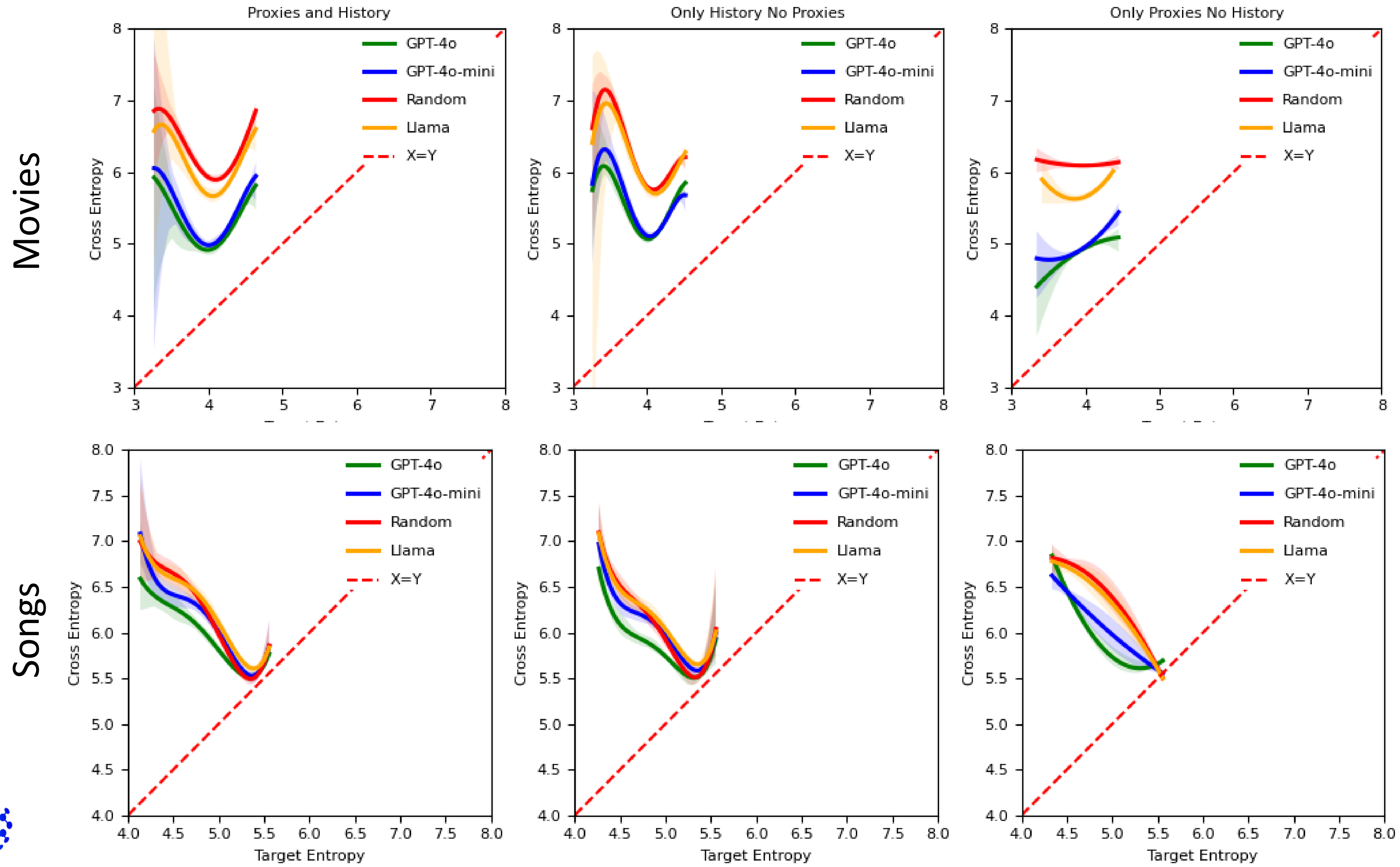
Calculate True or Target entropy from the dataset HTarget	Calculate entropy of Model’s prediction HModel	Calculate cross-entropy between Model and Target HCross

Ideally, HTarget = HCross





Saha and Choudhury (2025) User Behavior Prediction as a Generic, Robust, Scalable and Low-Cost Evaluation Strategy for Estimating Generalization in LLMs. *Under review.*

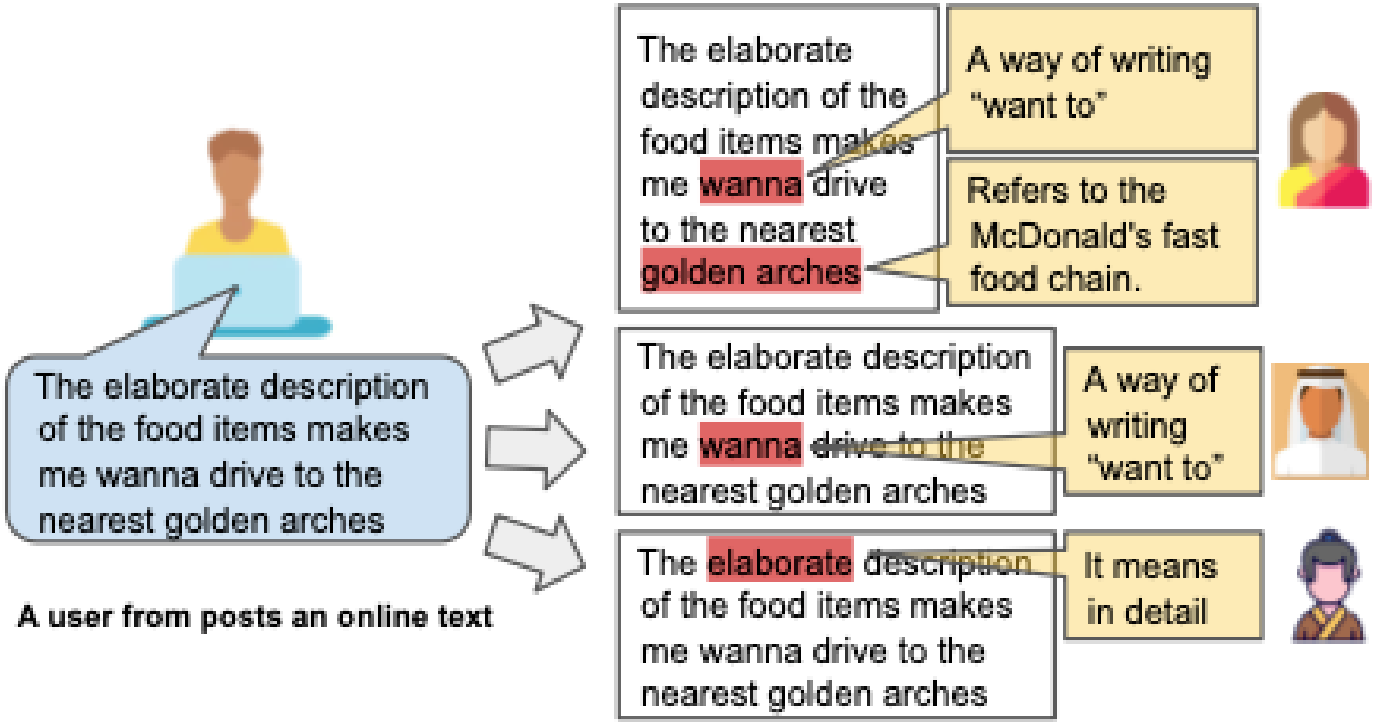


Movies
Songs
Proxies and History
Only History No Proxies
Only Proxies No History
GPT-4o
GPT-4o-mini
Random
Llama
X=Y
Cross Entropy
Target Entropy

Culturally Yours

A cross-cultural communication assistant.

Culturally Yours (CY) is a cultural reading assistant that identifies, adapts, and translates culture-specific items from text to users from different cultural backgrounds.



The CY Reading Assistant identifies culturally unfamiliar concepts and explains them as per the reader's cultural background.

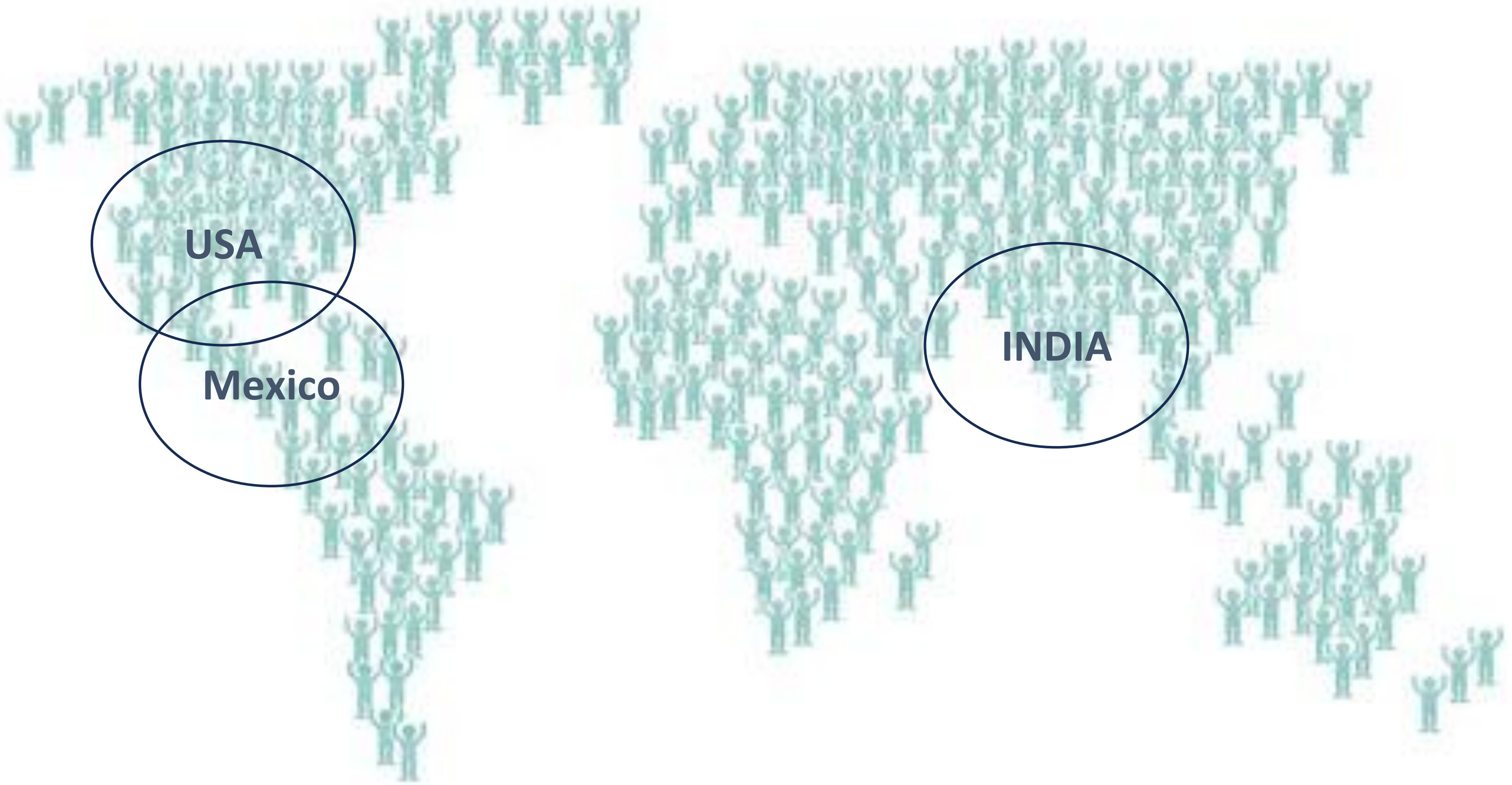
Pandey, Budhiraja, Saha, and Choudhury (2025) CULTURALLY YOURS: A Reading Assistant for Cross-Cultural Content. In *Proceedings of COLING 2025: System Demonstrations*

User Study

India

USA

Ethiopia



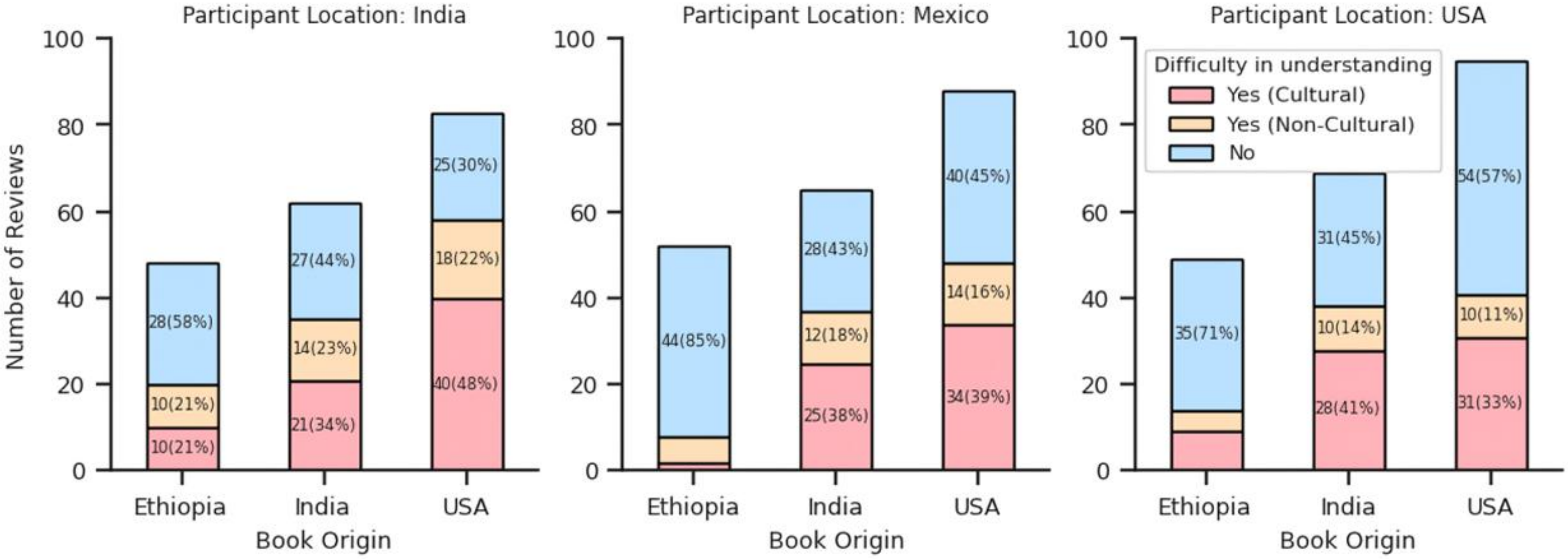


Figure 1: Participant location-wise difficulty in understanding book reviews by country of origin of the book.

Saha, Pandey and Choudhury (2025) Reading Between the Lines. In *Proceedings of NAACL 2025*

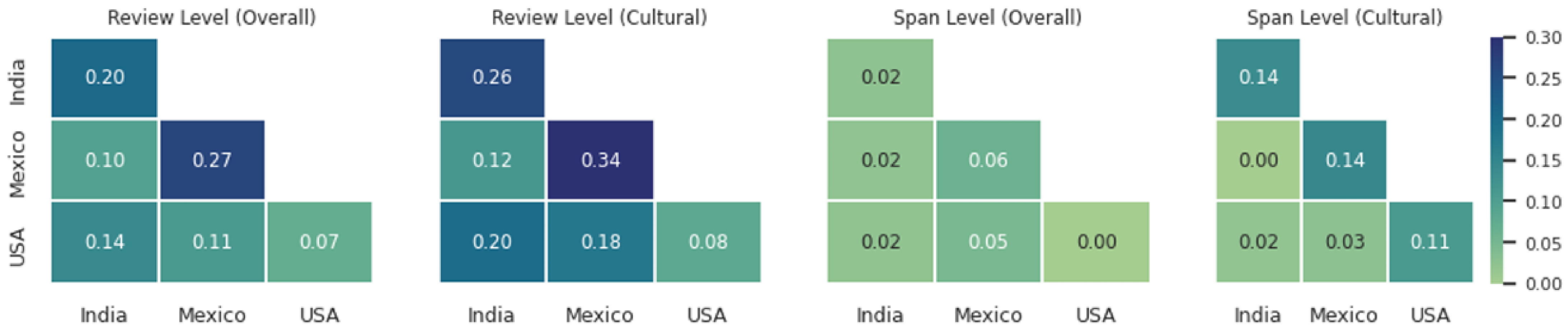
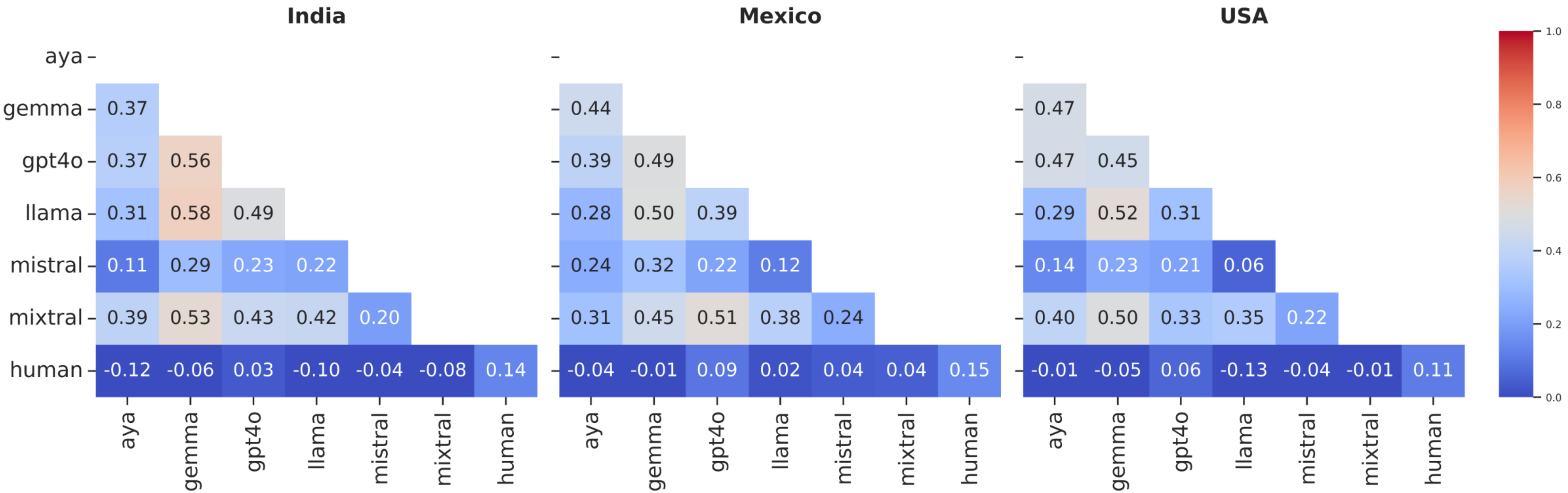


Figure 2: Inter Annotator Agreement at Review Level and Span Level across Countries.

Inter-annotator Agreement with Goodreads book reviews



Models agree more between themselves than humans do.
Models agree far less with humans than humans do.

You are not the average!

- <u>All users</u> belonging to a CULTURE exhibit <u>some</u> of the *prototypical behaviours* of that CULTURE
- <u>No user</u> exhibits <u>all</u> the prototypical behaviours of the CULTURE
- LLM simulations of CULTURE tend to exhibit less diversity of behaviour (*more stereotypical*?)

Distillation has bad repercussions for CULTURE, as it will further reinforce stereotypical behaviors.

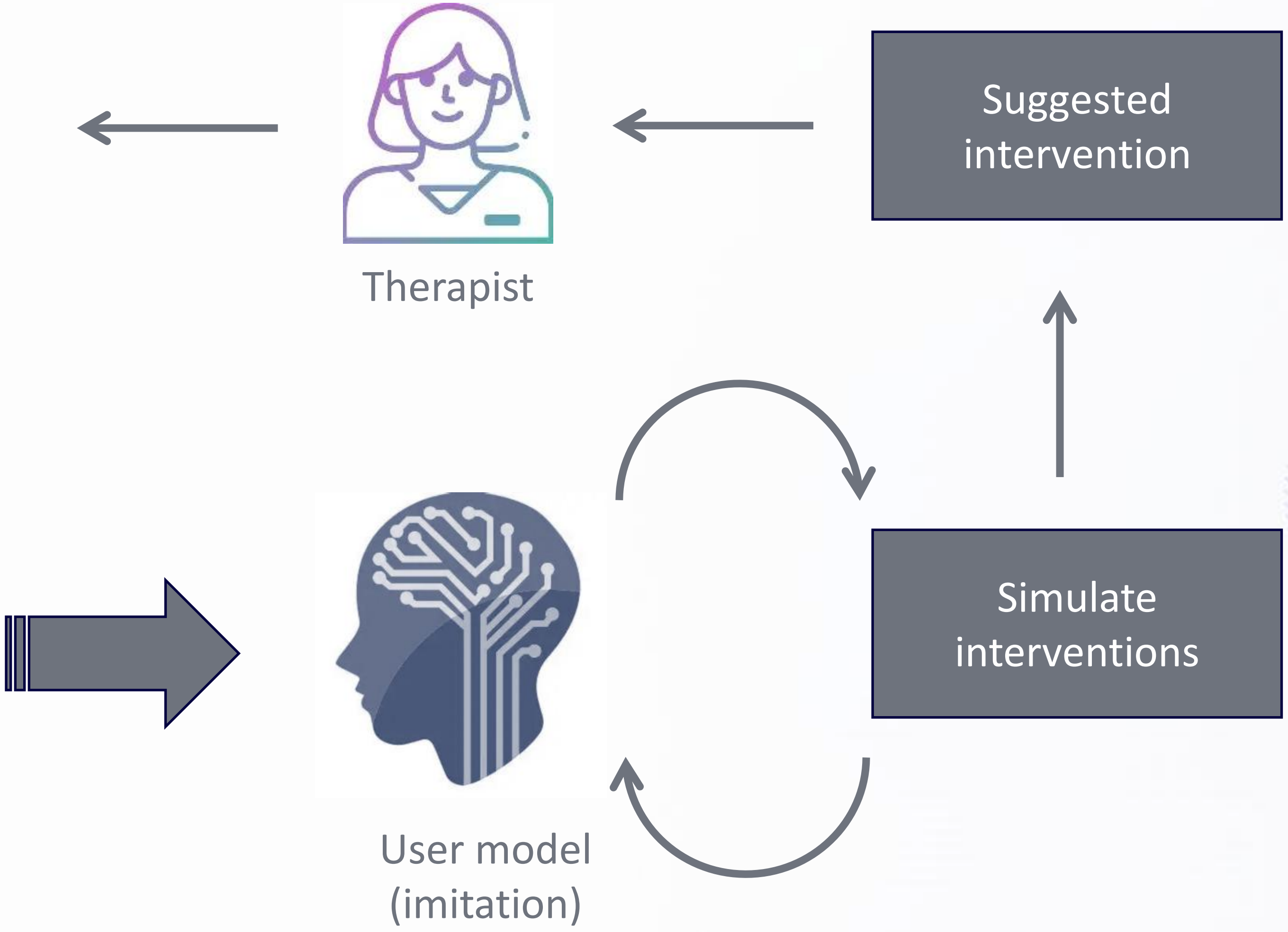
Philosophical Repercussions & Open Questions

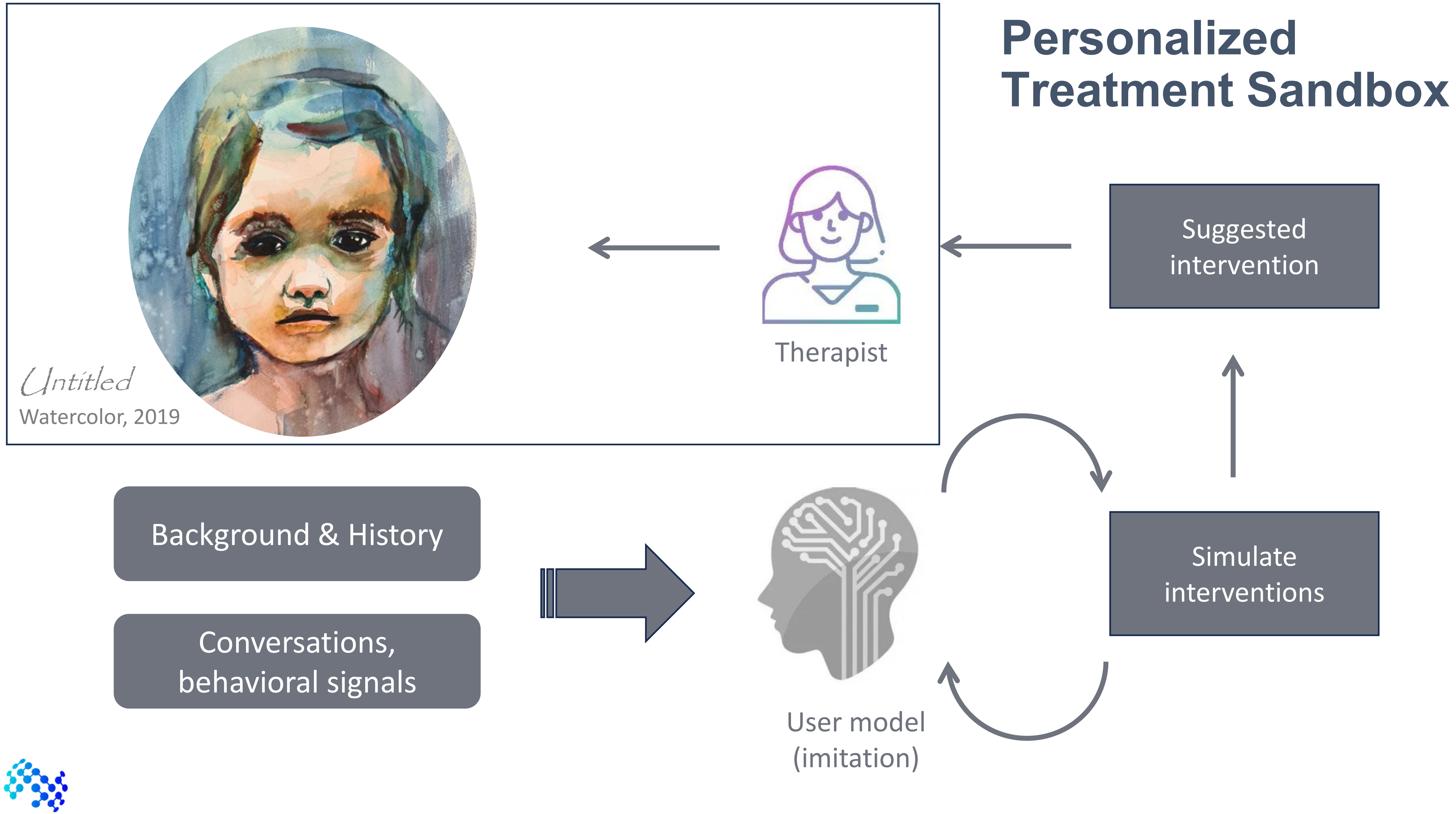
Turing test 1.0: Can computers imitate *any (non-specific) human*?
Turing Test 2.0: Can computers imitate *a specific human*?

Untitled
Watercolor, 2019

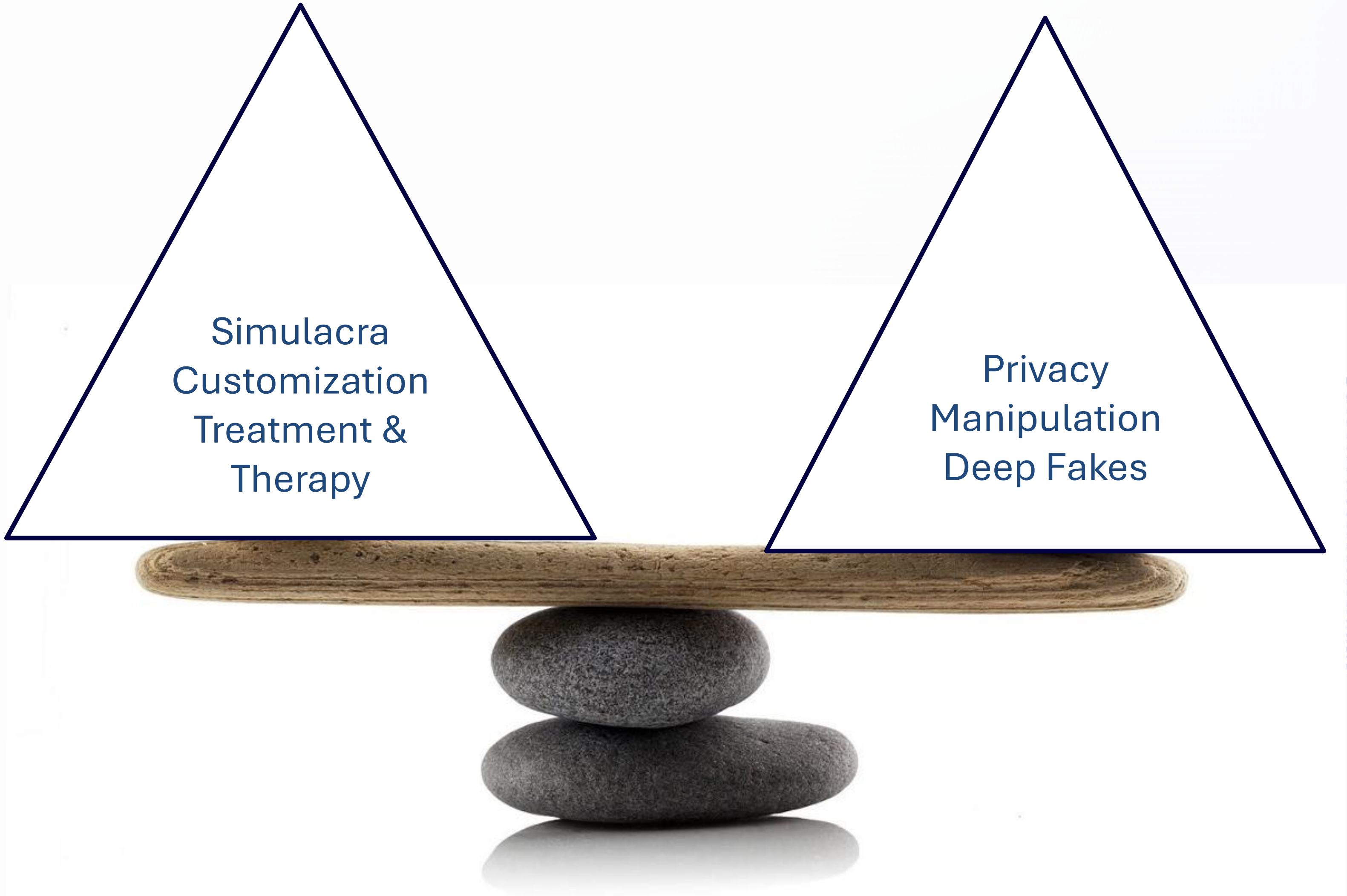
Background & History

Conversations, behavioral signals




Personalized
Treatment Sandbox
Untitled
Watercolor, 2019
Therapist
Suggested
intervention
Background & History
Conversations,
behavioral signals
User model
(imitation)
Simulate
interventions

A fine ethical balance



Open Questions

Q1: Transferability

To what extent the behaviour can be transferred between domains?

Q2: Limits of Prediction

Human behaviour is a second order complex system. Does the point of inversion implies a fundamental limit rather than just a model's limitation? How to know when we hit that limit?

Supported by the Azure Foundation Model Research Grant

Microsoft