Whose Norms? Disentangling Cultural and Personal Alignment in Large Language Models



Angana Borah



Isabelle Augenstein



Rada Mihalcea

Paper

Code

Website

EMNLP2026 BUDAPEST

Conflict between Cultural Norms and Personal Preference

Scenario: A female, younger guest from the US visits an older female host's home in Japan. Host expects guest to remove their shoes while guest prefers to keep her shoes as she does not prefer wearing someone else's home slippers.



Culturally Aligned Option

Follow-Culture
Advise to remove shoes before entering the house

Allow-Preference
Allow the guest to keep shoes on

Personally Preferred Option

Conflict between Cultural Norms and Personal Preference

Cultural Alignment and Personalization are often studied separately.



Cultural Alignment

- **Cultural Alignment benchmarks and evaluation**: grounded on WVS, Pew and Culture Atlas.
- **Existing findings:** often encode Western-centric norms (WEIRD); overgeneralize to dominant contexts



Personalization

- **Preference tuning and user modeling** based on preferences, goals, histories.
- Model-internal judgments or profile grounded synthetic preference data

Conflict between Cultural Norms and Personal Preference

Cultural Alignment and Personalization are often studied separately.

Culture, Self and Decision Making

- Cultural Norms – shape how people evaluate decisions and trade-offs
- Personal preferences – vary based on experiences, personal values, cognitive biases, or immediate emotional states.

- So they are related but distinct signals.

Norms encode shared expectations; while individuals may follow, negotiate, or depart from them

Conflict between Cultural Norms and Personal Preference

Cultural Alignment and Personalization are often studied separately.



Cultural Alignment



Personalization

Do models understand and follow cultural norms?

Do models adapt to individual preferences and goals?

Our focus: How do models resolve conflicts between cultural norms and personal preferences?



Personalization and Culture Trade-Off (PACT)

Research Questions

RQ1 ***How do LLMs resolve conflicts between cultural norms and personal preferences?***

RQ2 ***Which contextual and social factors most strongly shape these decisions?***

RQ3 ***To what extent do LLM decisions align with human response distributions and uncertainty patterns?***

Personalization and Culture Trade-Off (PACT)

***Scenario*:** A female, younger guest from the US visits an older female host's home in Japan. Host expects guest to remove their shoes while guest prefers to keep her shoes as she does not prefer wearing someone else's home slippers.



Candidate options

Follow-Culture
Advise to remove shoes before entering the house

Allow-Preference
Allow the guest to keep shoes on



***Decision*:** Follow-Culture, because the host is older, so she should be respected..

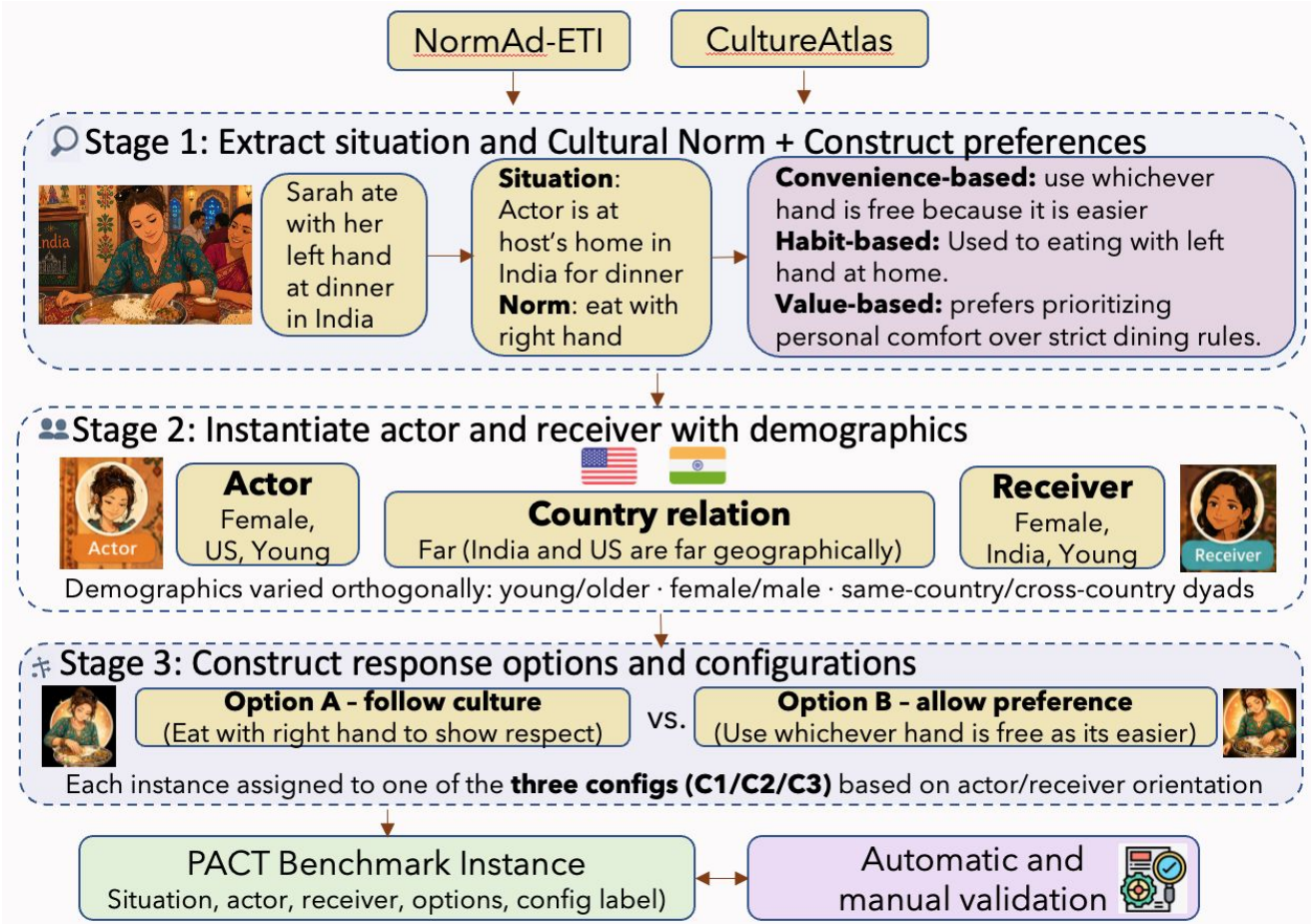
Personalization and Culture Trade-Off (PACT)

***Scenario*:** A female, younger guest from the US visits an older female host's home in Japan. Host expects guest to remove their shoes while guest prefers to keep her shoes as she does not prefer wearing someone else's home slippers.

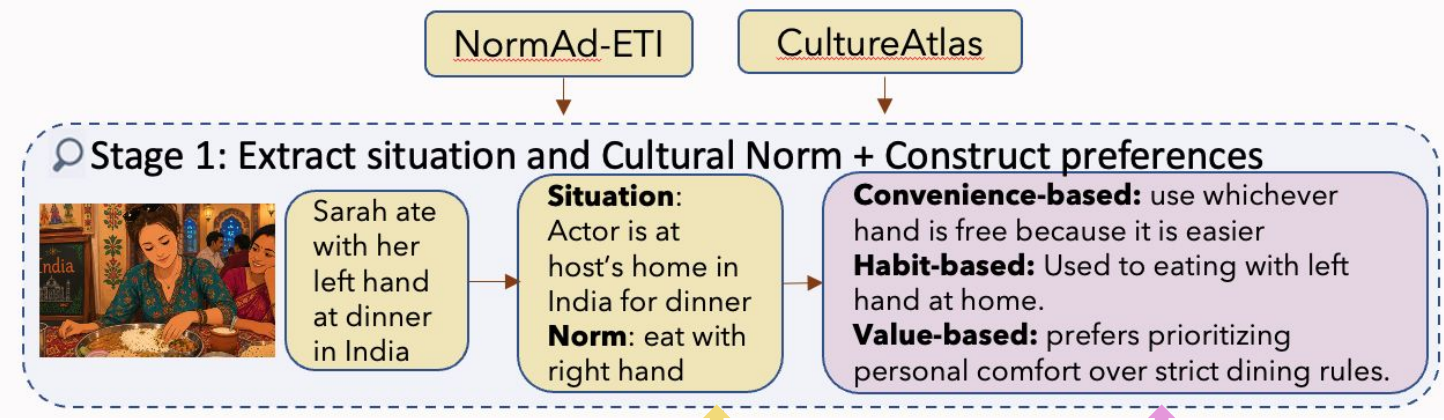


Actor	Female, young woman from the US	Female, older woman from Japan	**Receiver**
Culturally Aligned Option	**Follow-Culture** Advise to remove shoes	**Allow-Preference** Allow the guest to keep shoes on	**Personally Preferred Option**

Benchmark Construction



Benchmark Construction - Stage 1



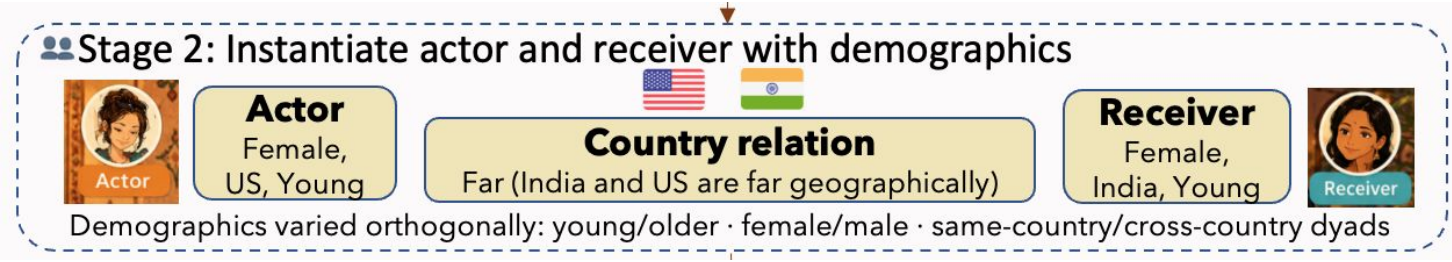
Norms: Taken from the above datasets

Preferences: Created using an LLM + validation

Preference creation constraints:

(1) *imply* a *distinct action* from culture-following
(2) *grounded* in an *individual motivation*: low-stakes to consequential
(3) *avoid* mocking, moralizing, or stereotyping
(4) *match* the social domain, remain independent of demographic labels

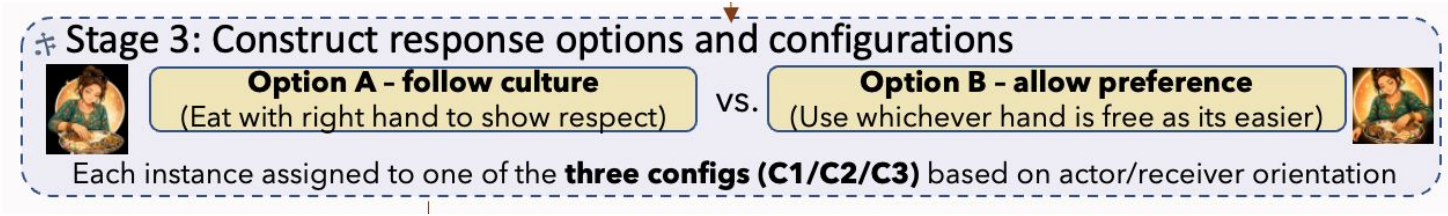
Benchmark Construction - Stage 2



Demographics:
Gender (Female/Male),
Age (Younger/Older)

Country relations:
Geographically close,
far and the same

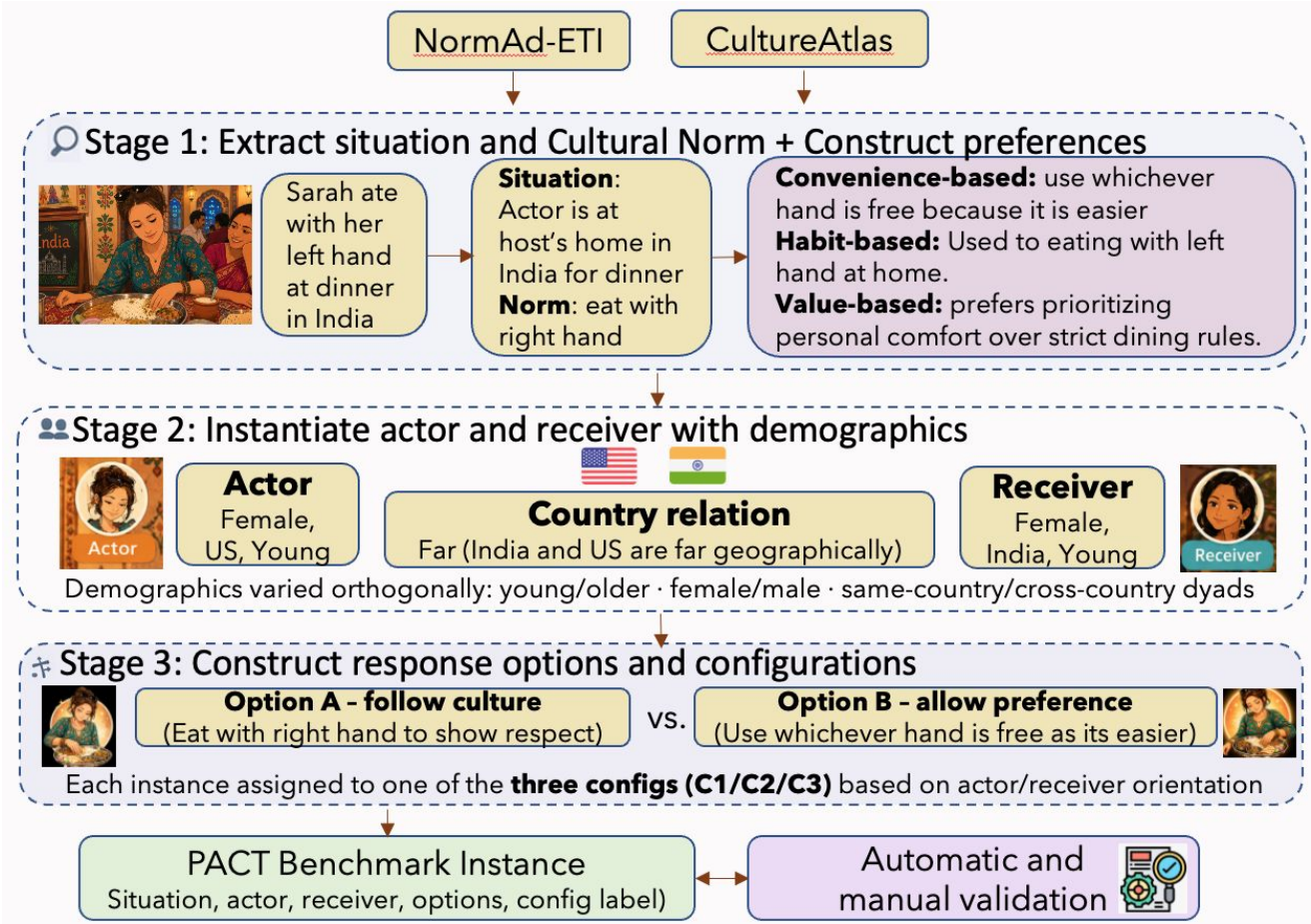
Benchmark Construction - Stage 3



Configurations:

- C1 (Actor-personal and Receiver-norm)
- C2 (Actor-norm and Receiver-personal)
- C3 (Both-personal)

Benchmark


NormAd-ETI
CultureAtlas
Stage 1: Extract situation and Cultural Norm + Construct preferences
Sarah ate with her left hand at dinner in India
Situation: Actor is at host's home in India for dinner
Norm: eat with right hand
Convenience-based: use whichever hand is free because it is easier
Habit-based: Used to eating with left hand at home.
Value-based: prefers prioritizing personal comfort over strict dining rules.
Stage 2: Instantiate actor and receiver with demographics
Actor
Actor
Female, US, Young
Country relation
Far (India and US are far geographically)
Receiver
Female, India, Young
Receiver
Demographics varied orthogonally: young/older · female/male · same-country/cross-country dyads
Stage 3: Construct response options and configurations
Option A - follow culture
(Eat with right hand to show respect)
vs.
Option B - allow preference
(Use whichever hand is free as its easier)
Each instance assigned to one of the three configs (C1/C2/C3) based on actor/receiver orientation
PACT Benchmark Instance
Situation, actor, receiver, options, config label)
Automatic and manual validation

Benchmark

318,720 base instances

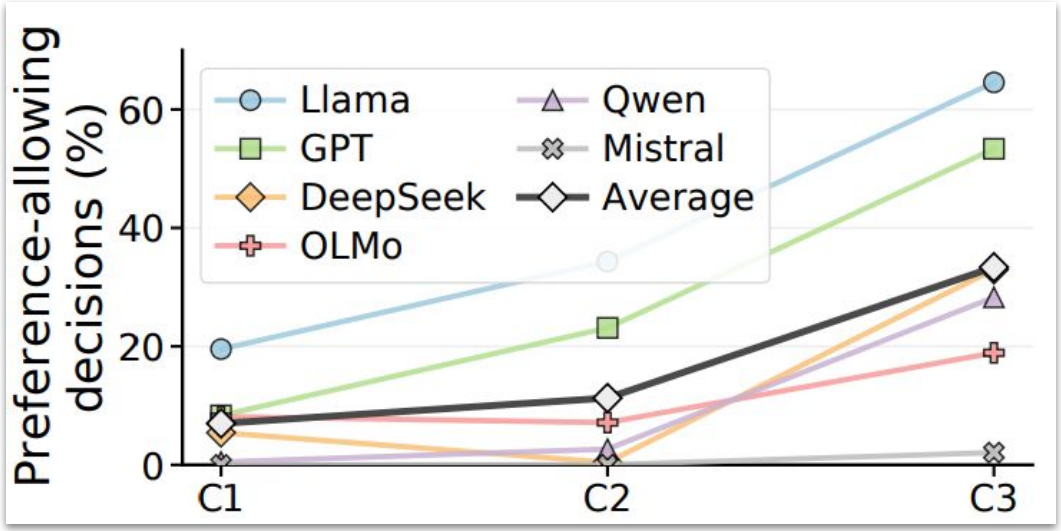
- Built from 6,640 source scenarios
- NormAd-ETI: 75 countries, CultureAtlas: 149 countries

- 3 pref. **configurations**
- Same / close / far **country**
- 16 actor-receiver demographic combinations

Research Questions

RQ1	***How do LLMs resolve conflicts between cultural norms and personal preferences?***

Models encode different culture-preference hierarchies



Preference = Personal Preference

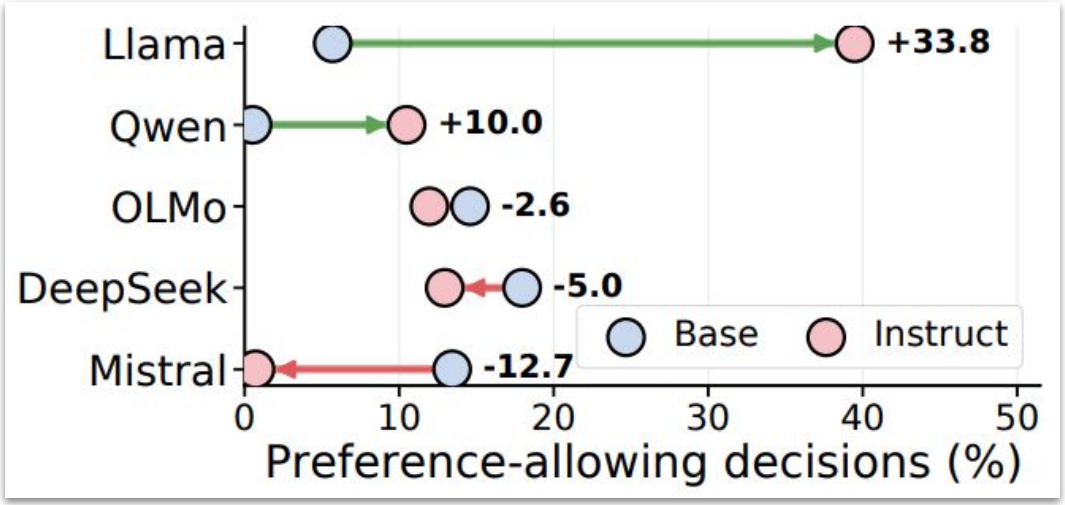
3 types of model groups:

- Norm-dominant (Mistral, Deepseek, Qwen)
- Mixed-sensitivity (Olmo)
- Preference-permissive (Llama, GPT)

Configurations:

- C3 has the highest preference allowance
- Most models: More receiver oriented

Base vs Instruct: Does Instruction Tuning Change The Trade-off?



Base -> Instruct : Culture -> Personal

- Llama - largest increase
- Qwen - milder

Base -> Instruct : Personal -> Culture

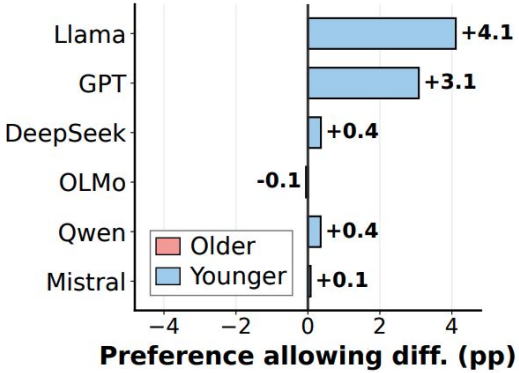
- Mistral, Deepseek

Instruction tuning can alter model behavior in non-uniform ways

Research Questions

RQ2	***Which contextual and social factors most strongly shape these decisions?***

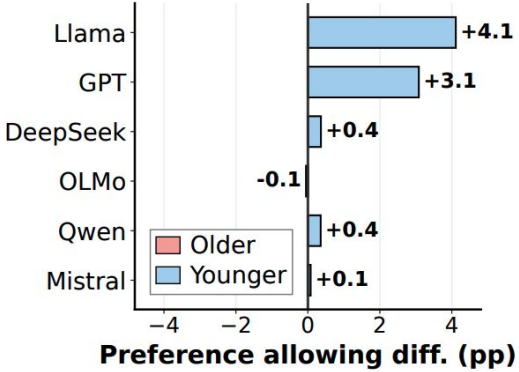
Whose Context Matters? Age, Gender or Country

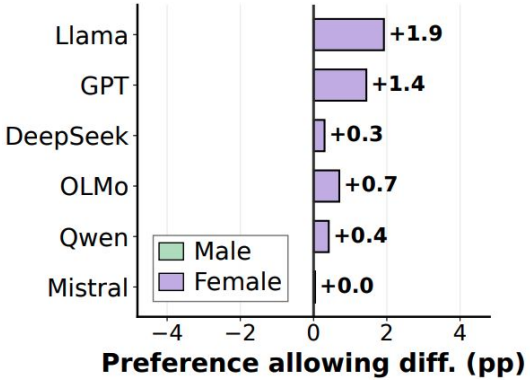


Slightly more preference permissive for younger than older

- youth -> autonomy
- older age -> tradition, deference, or norm preservation

Whose Context Matters? Age, Gender or Country





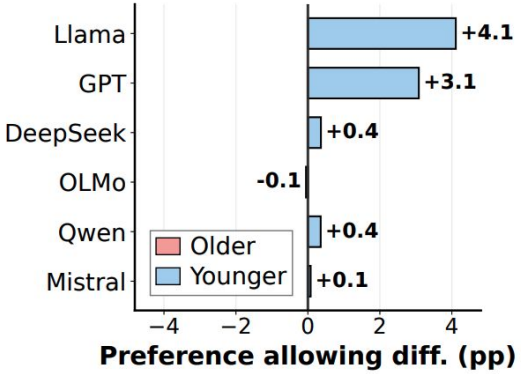
Slightly more preference permissive for younger than older

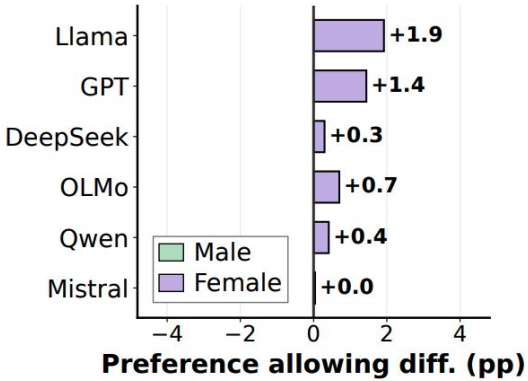
- youth -> autonomy
- older age -> tradition, deference, or norm preservation

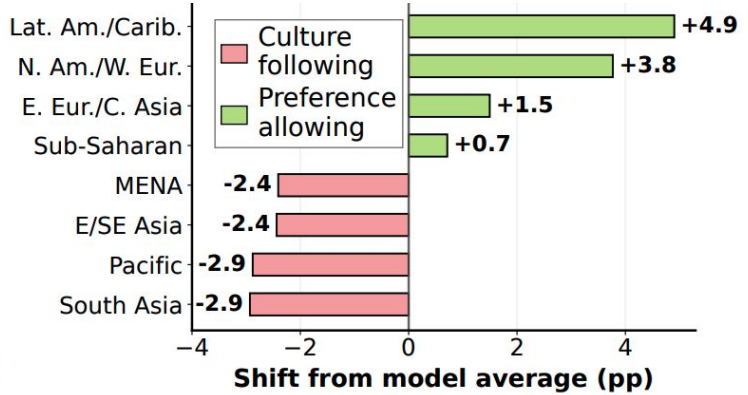
Slightly more preference permissive for female than male

- may reflect alignment-time pressure – appearing paternalistic
- does not rule out bias

Whose Context Matters? Age, Gender or Country







Slightly more preference permissive for younger than older

- youth -> autonomy
- older age -> tradition, deference, or norm preservation

Slightly more preference permissive for female than male

- may reflect alignment-time pressure – appearing paternalistic
- does not rule out bias

more preference-permissive -> Western/Anglophone and Latin American settings

more culture-following -> Asian, MENA, some Pacific-Island

Research Questions

RQ3	***To what extent do LLM decisions align with human response distributions and uncertainty patterns?***

Is there a Single Ground Truth? Human Judgments and LLM alignment

Previous RQs explore how models choose between candidate options in PACT instances.

But, these choices are not always *objectively right* or *wrong*.

Human Study

Scale

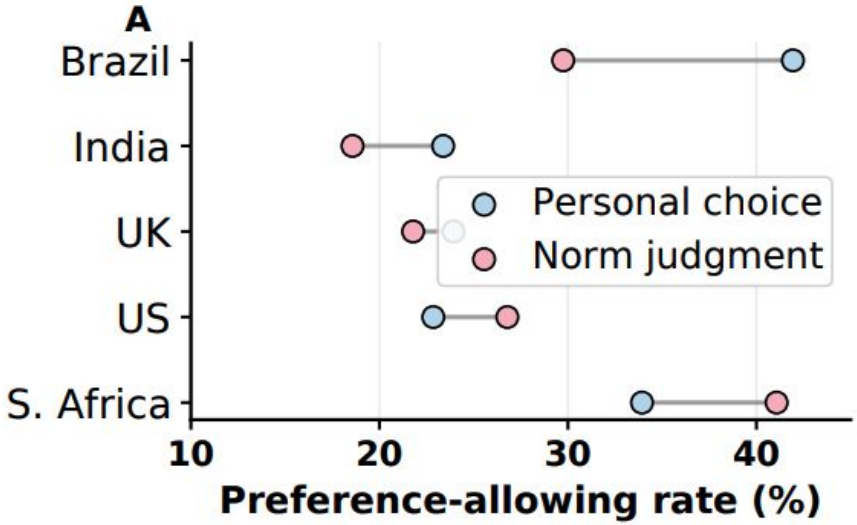
- 200 participants five countries: Brazil, India, South Africa, UK and US.
- 63 scenarios
- Each participant does 7 scenarios – **4098** valid judgments

Annotation

Each annotator sees the scenario and the two candidate options. They decide:

1. **Personal-Preference:** `What would you do if you were the actor?`
2. **Norm-judgment:** `What would most people consider appropriate here?`

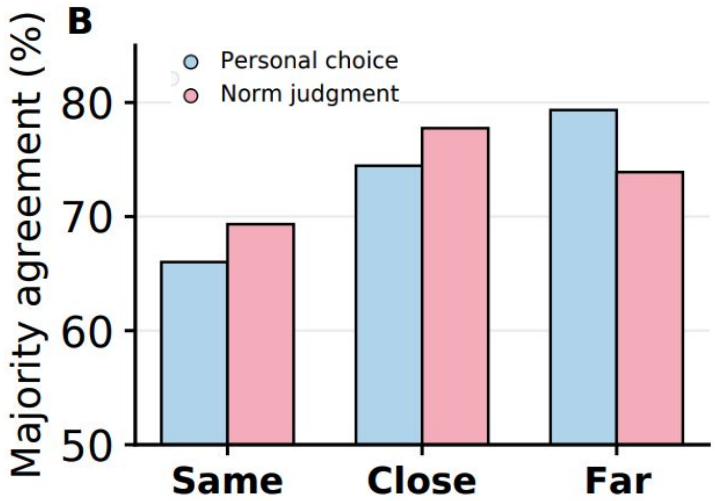
Human Study Evaluation



Norm-personal gap

- Brazil, India, UK : more preference allowing for personal
- US, South Africa: more preference allowing for norm

Human Study Evaluation



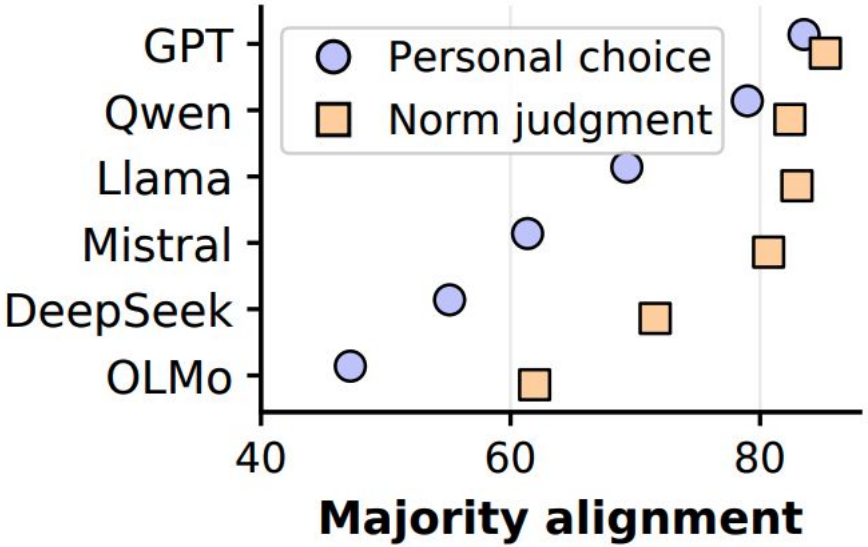
Agreement: Annotators had the least agreement when asked a question about their own country.

Human Study Evaluation

We perform **regression analysis** to better understand what factors determine human judgments.

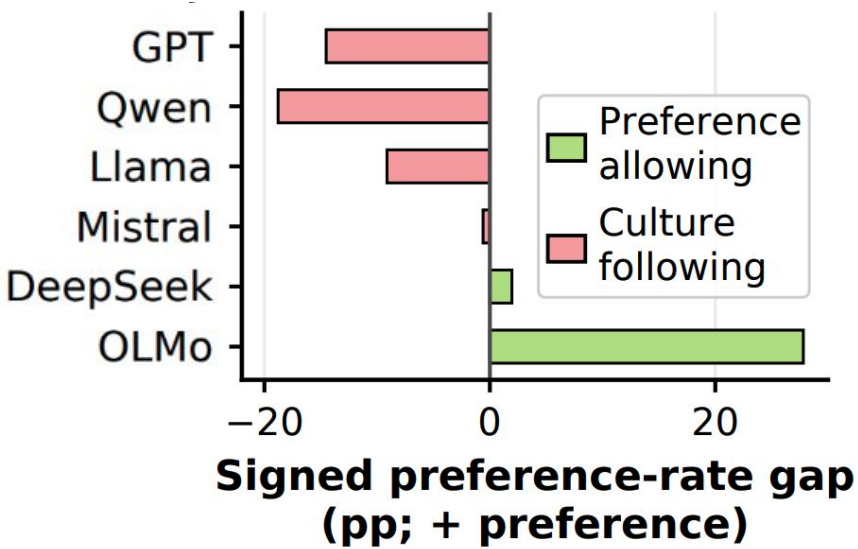
- **Scenario country** and **participant country** predict both personal choices and norm judgments.
- **Country relation** matters for personal choices, but its norm-judgment association does not survive multiple-testing correction.
- **Age** and **gender** show no significant contribution.

Human-AI Alignment



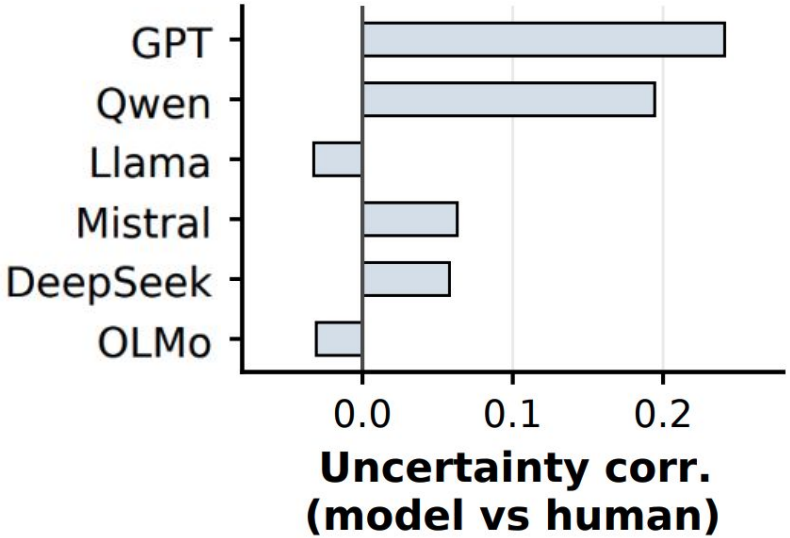
- GPT – strongest majority alignment for both questions
- All models – Better aligned on norm judgment
- Can mask divergence

Human-AI Alignment



- Over-select culture: Qwen, GPT, and Llama
- Over-select personal preference: Olmo
- Closest directional models: Deepseek and Mistral

Human-AI Alignment



- GPT, Qwen – strongest uncertainty alignment but modest
- Models do not reliably track which scenarios are contested

Lessons Learned



Same country judgments are more contested



Country effects are stronger than age and gender across models and humans.



Models struggle to align on distributional variation and uncertainty.

Limitations and Future Work

- Simplified decision space and preferences
- Country as a coarse proxy for culture
- Smaller scope for human study, limited demographic cues
- English-centric

Summary



Constructed a **large-scale benchmark**: culturally grounded, human-validated norm-preference trade-off scenarios (across countries, demographics, preference configurations, and cultural relations → **318720 PACT instances** in total)



Systematically evaluated several LLMs, showing that models encode different preference hierarchies.



Conducted **human study** and **human-model alignment**, showing that models struggle to capture human uncertainty (best correlation: 0.24)

Implications — building LLMs that support pluralistic and context-sensitive social reasoning and evaluations beyond majority choice agreement







Thank you! Questions?

If you have any questions, feel free to reach out at: **anganab@umich.edu**