Cultural Learning-Based Culture Adaptation of Language Models

Chen Cecilia Liu[1] and **Anna Korhonen**[2] and **Iryna Gurevych**[1]
[1] Ubiquitous Knowledge Processing Lab,
Department of Computer Science and Hessian Center for AI (hessian.AI),
Technical University of Darmstadt
[2] Language Technology Lab, University of Cambridge
www.ukp.tu-darmstadt.de

HUMANE Lab 석사과정 고경빈

2025 ACL

2025.09.11

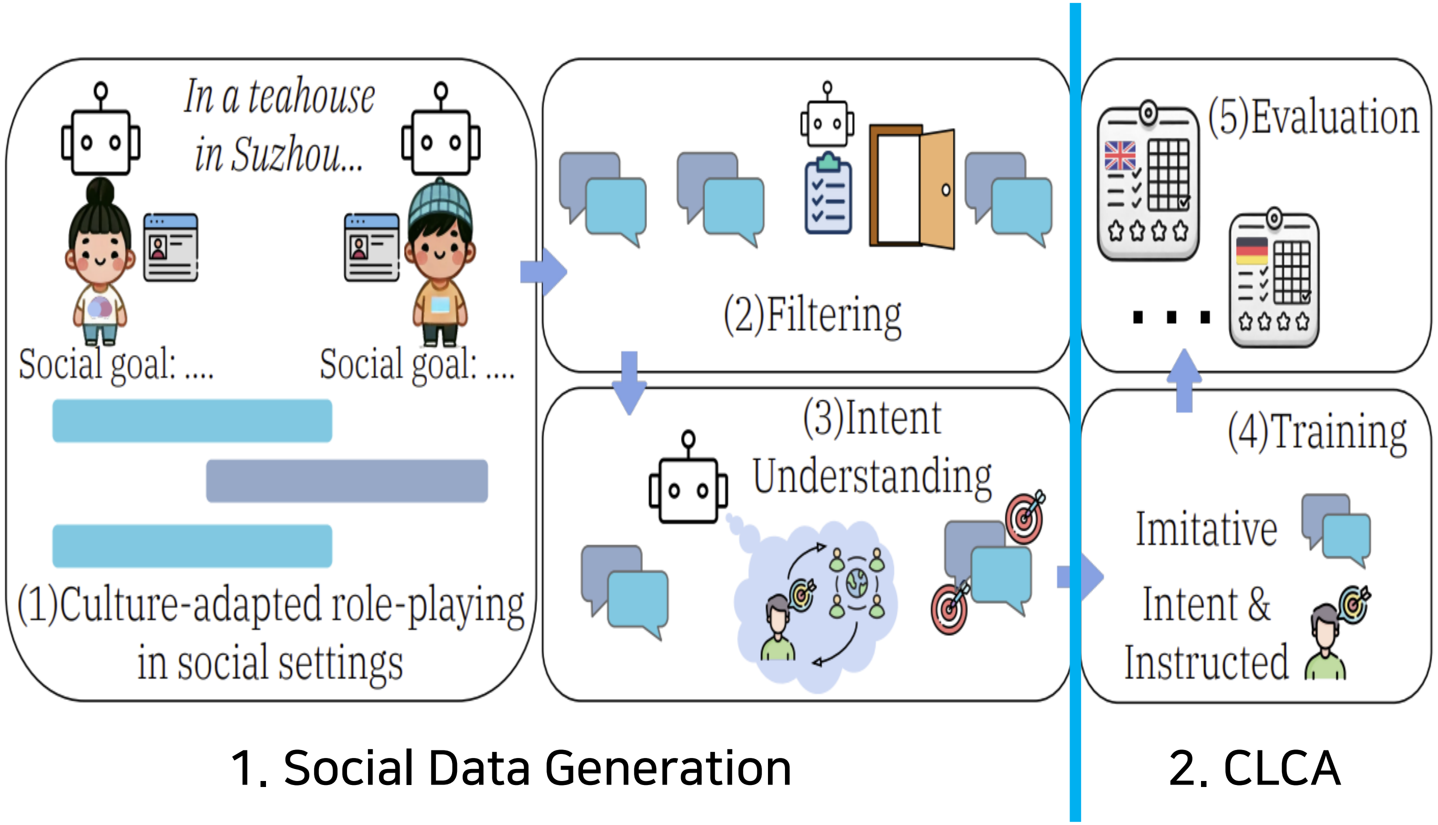
Background

- LLMs aligns with WEIRD values, showing limited cultural competence and global applicability.
- Existing methods for adapting LLMs to diverse cultural values rely on prompt engineering
  - LLMs must contain enough cultural values from pre-training…
- Recently, cultural learning has become important in AI training

Cultural Learning

- Process of learning behaviors, knowledge, and culture from environment
- How can we learn?
  - Imitative learning: observing and replicating the actions of others
  - Instructed learning: being explicitly conveyed or demonstrated
  - Collaborative learning
- Why ?
  - basic ways through which individuals first learn culture
- Key: Ability to understand the intentions of others during interactions

Method


In a teahouse in Suzhou...
Social goal: ....
Social goal: ....
(1)Culture-adapted role-playing in social settings
(2)Filtering
(3)Intent Understanding
(5)Evaluation
(4)Training
Imitative
Intent &
Instructed
1. Social Data Generation
2. CLCA

Social Data Generation

1. Culture-Adapted Social Scenarios
    - Setup: descriptions of social scenarios, 2 participant profiles, their respective private social goals for the interaction
    - Perform automatic culture adaptations of social settings using GPT-4
        - Names are localized: Anthony → Kenji
        - Settings are adapted: a bar in London → a teahouse in Suzhou
    - Generating new scenarios based on social and cultural norms from Social Chemistry and Culture Atlas

Social Data Generation

2. Interaction Data Generation
   - Two LLMs are role-playing the participants
   - Data generation incorporates cultural context from Hofstede's cultural dimensions and Inglehart-Welzel cultural map

3. Filtering(LLM-as-a-Judge)
   - Generate data twice for each scenario and apply the filtering process
   - 1st step: general generation quality & cultural adherentness
   - 2nd step: Meta-Evaluation(quality + confidence)
   - Discard: high-confidence bad meta-evaluation or general generation quality

Social Data Generation

4. Intent Generation

- Identifies the intent of each conversational turn
- Evaluates its alignment with social and cultural expectations

Example

Setting: At a bustling shopping mall a tourist is trying to find the nearest restroom.
Li Wei: 35 / Male, Mall Security Guard
Goal: To assist Zhang in finding the restroom while maintaining the flow of traffic in the mall.
Zhang Qi: 45 / Male, Tourist
Goal: To find the nearest restroom as quickly as possible.

Li Wei: Excuse me, sir. Are you looking for something?
Intent: Offering help.
Zhang Qi: Ah, restroom. I'm looking for the restroom. Could you tell me where it is?
Intent: To get directions to the restroom. Generic

Li Wei: The restroom is just down that hallway, sir. You can't miss it.
Intent: Li Wei's intent is to politely and efficiently provide directions to the tourist.
Zhang Qi: Thank you, I'll just go take a look.
Intent: Zhang Qi intends to politely thank Li Wei and follow his directions to find the restroom.

Li Wei: Sir, would you like me to escort you to the restroom?
Intent: Li Wei is showing respect and courtesy, as is customary in Chinese culture, especially when interacting with an older person (Zhang Qi is years older than Li Wei).
Zhang Qi: Thank you for your help, I'll just go now.
Intent: Politeness and appreciation. Cultural

Cultural Learning-Based Culture Adaptation (CLCA)

- Using a multi-task training approach leveraging the generated data
  - Multi-Turn Conversation Training → Imitative learning
    - Each conversation is trained from both participants' perspectives
  - Intent Understanding → Instructed learning
    - Generating the underlying intention of the conversation turn
    - Learning its relevance to social and cultural expectations.

Experimental Setup

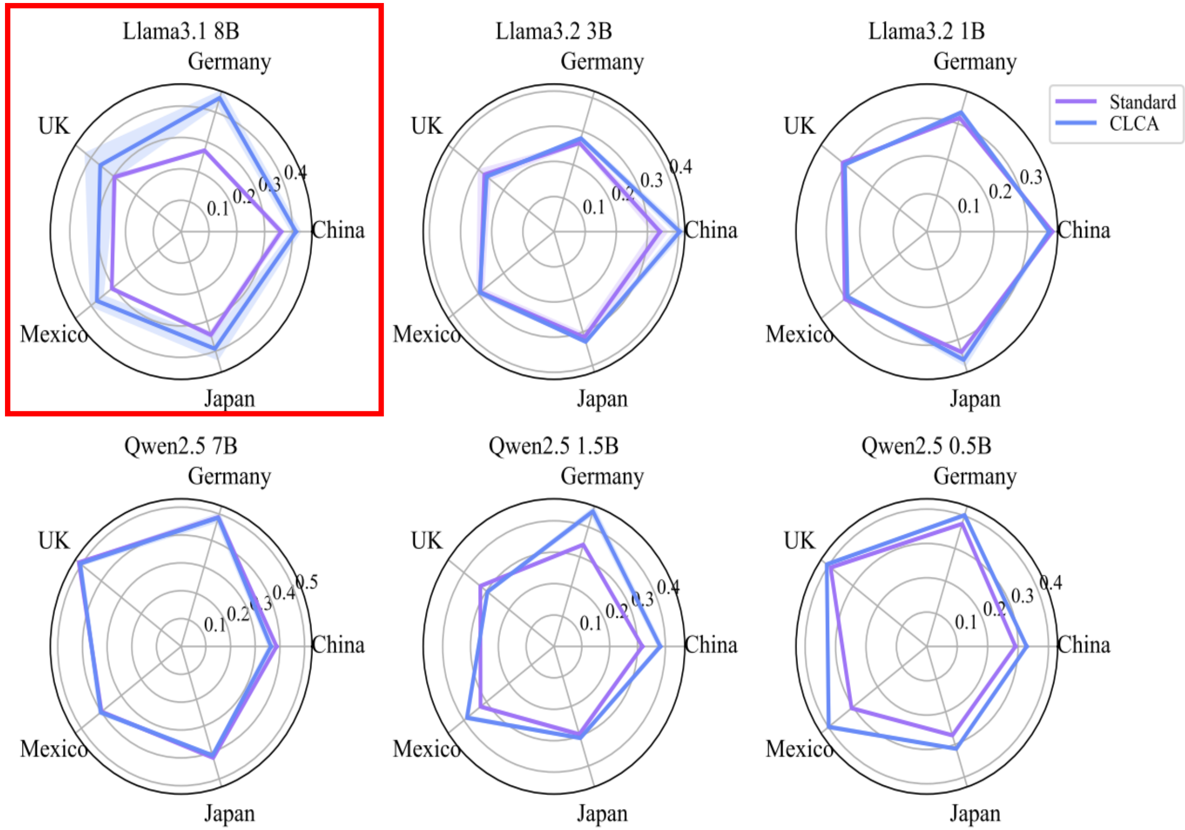
- Evaluating using the World Values Survey(WVS)
  - 5 different cultures: UK, China, Germany, Mexico, and Japan
  - Topics: Social Values, Norms, Stereotypes (44 questions per culture)
  - Using participant profiles, generating 1,000 personas for each culture
- Model (Instruction-tuned)
  - Llama - 3.2 1B/3B, 3.1 8B // Mistral - v0.3 7B // Qwen - 2.5 0.5B/1.5B/7B
- Methods
  - Persona(zero-shot) // Cultural(no demographics) // CLCA
- Metrics
  - Kullback-Leibler Divergence
  - Individual-level Accuracy

Cultural Learning Aligns Models to Surveys

	China	Germany	UK	Mexico	Japan	Avg. KL-D ↓
Llama3.1 8B	0.5958	0.6717	0.6268	0.5391	0.5721	0.6011
Llama3.1 8B$_{\texttt{cultural}}$	0.5881	0.6690	0.6431	0.5437	0.5660	0.6020
Llama3.1 8B$_{\textbf{CLCA}}$	0.5462	0.4935	0.5510	0.4630	0.5024	**0.5112** △0.0899
Llama3.2 3B	0.6174	0.6903	0.6631	0.5667	0.6221	0.6319
Llama3.2 3B$_{\texttt{cultural}}$	0.5996	0.6729	0.6375	0.5569	0.6042	0.6142
Llama3.2 3B$_{\textbf{CLCA}}$	0.5337	0.6732	0.6695	0.5525	0.6100	**0.6078** △0.0241
Llama3.2 1B	0.5936	0.6479	0.6384	0.5584	0.6024	0.6081
Llama3.2 1B$_{\texttt{cultural}}$	0.5905	0.6840	0.6675	0.5209	0.6664	0.6259
Llama3.2 1B$_{\textbf{CLCA}}$	0.5671	0.6208	0.6348	0.5683	0.5743	**0.5931** △0.0150
Qwen2.5 7B	0.5692	0.4610	0.4221	0.4509	0.5053	**0.4817**
Qwen2.5 7B$_{\texttt{cultural}}$	0.5984	0.5051	0.5355	0.4961	0.5467	0.5364
Qwen2.5 7B$_{\textbf{CLCA}}$	0.5917	0.4605	0.4439	0.4390	0.5047	0.4880 −△0.0063
Qwen2.5 1.5B	0.6315	0.6069	0.6040	0.5134	0.6225	0.5956
Qwen2.5 1.5B$_{\texttt{cultural}}$	0.6271	0.6406	0.6540	0.5476	0.6343	0.6207
Qwen2.5 1.5B$_{\textbf{CLCA}}$	0.5614	0.4895	0.6414	0.4559	0.6129	**0.5522** △0.0434
Qwen2.5 0.5B	0.6381	0.5589	0.5205	0.5192	0.6373	0.5748
Qwen2.5 0.5B$_{\texttt{cultural}}$	0.5661	0.6382	0.6093	0.5305	0.5818	0.5852
Qwen2.5 0.5B$_{\textbf{CLCA}}$	0.6130	0.5173	0.5061	0.4428	0.5794	**0.5317** △0.0431
Mistral-v0.3 7B	0.6216	0.6414	0.6249	0.5069	0.6458	0.6081
Mistral-v0.3 7B$_{\texttt{cultural}}$	0.6155	0.6733	0.6553	0.5219	0.6475	0.6227
Mistral-v0.3 7B$_{\textbf{CLCA}}$	0.6171	0.6407	0.6178	0.5074	0.6341	**0.6034** △0.0047

- CLCA > Persona > Cultural
- In LLaMA models, larger models align better, but this scaling trend isn't seen in Qwen models.

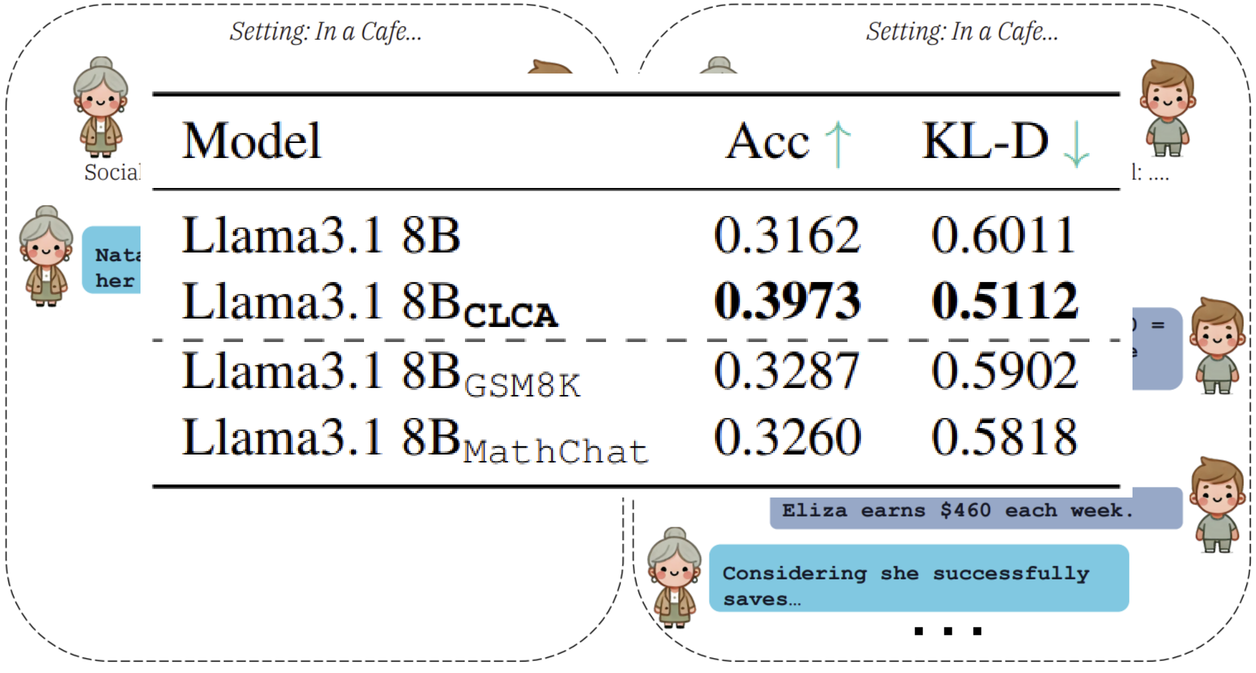
Cultural Learning Aligns Models to Surveys



- Llama 3.1 8B is the best

Social Interaction Plays a Significant Role

- Is social interaction data important for improving culture alignment?



GSM8K MathChat

Model	Acc ↑	KL-D ↓
Llama3.1 8B	0.3162	0.6011
Llama3.1 8B$_{\text{CLCA}}$	**0.3973**	**0.5112**
Llama3.1 8B$_{\text{GSM8K}}$	0.3287	0.5902
Llama3.1 8B$_{\text{MathChat}}$	0.3260	0.5818

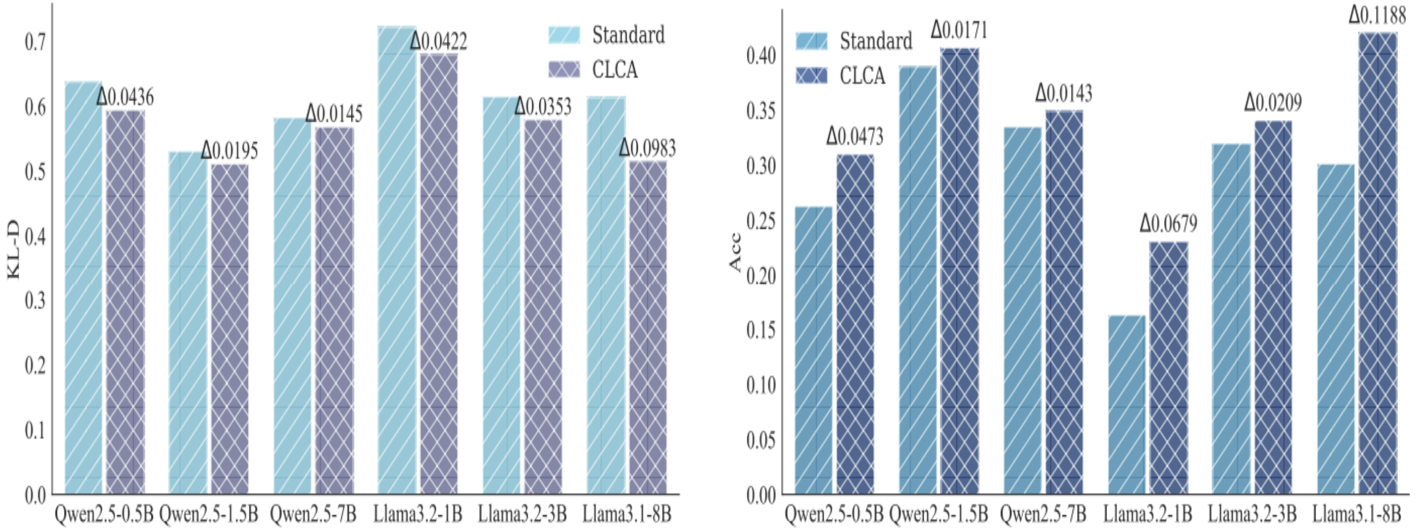
Its effect on cultural alignment is minimal compared to social interaction data

Intent Understanding is Important in CLCA

Model	Acc ↑	KL-D ↓
Llama3.1 8B	0.3162	0.6011
Llama3.1 8B CLCA	**0.3973**	**0.5112**
Llama3.1 8B CLCA `intent_only`	0.3117	0.6037
Llama3.1 8B CLCA `dialogue_only`	0.3453	0.5704

- dialogue_only: slightly improve
- intent_only: barely improve
- CLCA(dialogue + intent): greatly improve

Zero-shot Value Transfer to Other Languages



(a) Kullback-Leibler Divergence (KL-D, lower is better) between the model prediction and WVS data.

(b) Individual-level accuracy (higher is better) between the model prediction and WVS data.

- Overall, models show consistent improvements in both KL-D and accuracy.
- LLaMA models improve more than Qwen models
- Qwen2.5 7B improves in multilingual but not English evaluations

Data Generation Model

- Does the adaptation work only with the Llama3.1 70B as a teacher?
    - Collect simulation data from the Qwen2.5 32B and train the Llama3.1 8B

  → KL-D: 0.5617 // Accuracy: 0.3487

    - These results exceed the baselines but lag behind LLaMA3.1 70B

✓ Teacher model and data quality matter, but cultural learning proves effective

Conclusion

- Proposed CLCA, using culturally adapted scenarios, interactions, and norms
- CLCA effectively aligns LLMs with diverse cultural values across model architectures and sizes

Open Question

- Llama와 Qwen에서의 결과가 차이가 나는 이유