HealthAlign-Agents: Reflective Self-Play for Culturally Safe Health Communication in Low-Resource Languages

Aura Arefeh Yavary

CLRLC-LLMs Workshop @ NeurIPS 2025

Why culturally aligned health AI matters

Large Language Models can solve many medical reasoning tasks, but still fail at something profoundly human: **understanding how people speak, feel, and make sense of illness.**

This failure disproportionately harms underrepresented language communities.

The main idea: a three-agent reflective system that improves itself through dialogue and critique — without any finetuning or dataset collection.

The gap nobody talks about

LLMs today:

- Strong medical reasoning
- Weak cultural grounding
- Poor metaphor + politeness understanding
- English-centric defaults
- Unsafe in many low-resource contexts

Medically correct $\neq$ culturally safe.

Our goal

We want health communication that:

- Speaks with **empathy**
- Respects **cultural norms**
- Uses **verified medical knowledge**
- Feels natural in phrasing, metaphor, and tone
- Stays **safe** in low-data settings

No personal data collection.

Why this is hard

Culture appears in:

- Idioms and metaphors
- Honorifics and politeness strategies
- Emotional expression
- Health beliefs and taboos
- Trust and risk perception

You can't just **finetune** culture into a model — but you can **simulate** it.

HealthAlign-Agents Architecture

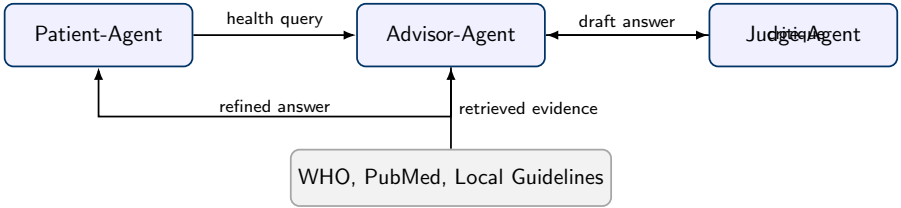

Figure: Architecture of HealthAlign-Agents. A culturally situated patient query is handled by an Advisor-Agent that drafts an answer using retrieved medical evidence. A Judge-Agent critiques the draft for factuality, empathy, and cultural alignment. The Advisor incorporates this critique to generate a refined, safer, and more culturally resonant answer.

Reflective self-play loop

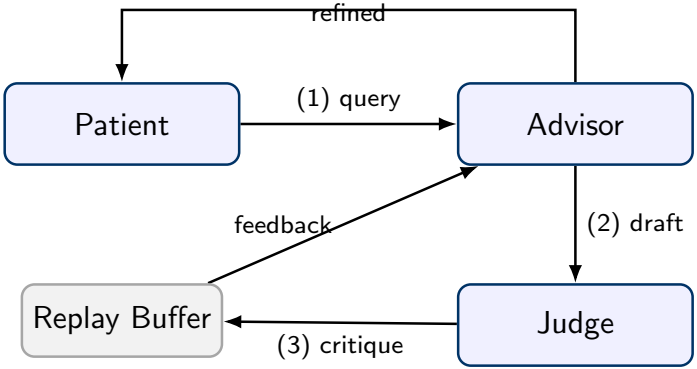

Figure: The reflective self-play loop. The Patient-Agent poses a culturally situated health question. The Advisor-Agent generates an initial draft. The Judge-Agent critiques it for factuality, empathy, and cultural alignment. A refined version is produced and fed back to the patient. This iterative process leads to emergent alignment without finetuning.

Optimization Objective

$$L_{\text{reflect}} = \mathbb{E}[\lambda_f R_f + \lambda_e R_e + \lambda_c R_c] + \beta \|\Delta P_t\|^2$$

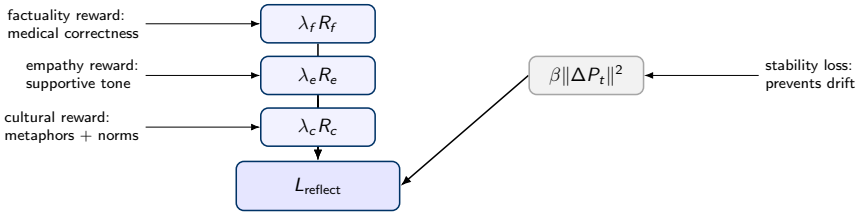

Figure: The optimization view of HealthAlign-Agents. Three reward terms guide the Advisor-Agent toward (1) medically accurate explanations (R_f), (2) empathetic and supportive tone (R_e), and (3) culturally grounded communication (R_c). A stability penalty $\beta \|\Delta P_t\|^2$ prevents prompt drift across self-play iterations, ensuring alignment emerges smoothly.

Emergent behaviors

HealthAlign-Agents learns:

- Pragmatic repair
- Tone shifting
- Cultural metaphor rewriting
- Moral and risk reasoning
- Stable empathy across turns

Qualitative example 1 — Diabetes advice

Context: Sweet tea is tied to hospitality in the user's culture.

Base LLM:

- "You must stop drinking sweet tea."
- Ignores hospitality norms and family expectations.
- Cold tone, no emotional support.

HealthAlign-Agents:

- "It's understandable to feel worried when hearing 'diabetes.'"
- Suggests culturally aligned adjustments: "Serve guests tea as usual — make your own cup less sweet."
- Uses local metaphor (e.g., "tired blood").
- Encourages involving family, not secrecy.

Qualitative example 2 — Postpartum care

Context: A new mother reports exhaustion, guilt, fear of "not being a good parent," and is balancing cultural expectations around rest, food, and family roles.

Base LLM:

- Gives clinically correct instructions ("Monitor bleeding... Watch for fever... Get enough rest.")
- Uses medicalized language unfamiliar to first-time mothers.
- No validation of emotional or social stress.
- Ignores cultural taboos around postpartum rest, household labor, or diet.

HealthAlign-Agents:

- Opens with reassurance: "Many new mothers feel overwhelmed — you're not doing anything wrong."
- Explains symptoms using **culturally familiar metaphors**: e.g., "Your body is recovering its strength after giving so much energy."
- Acknowledges social context: "It's okay to ask your sister or mother for help with meals or chores; this is part of the tradition in many families."
- Gives **clear medical guidance** while keeping tone supportive.
- Avoids blame or judgment; emphasizes shared responsibility with family.
- Frames rest and nutrition in culturally aligned ways (warm foods, herbal teas, reduced household work).

Qualitative Considerations

Base LLM:

- Factual but emotionally flat.
- Uses clinical terminology unfamiliar to many caregivers.
- Doesn't acknowledge fear, shame, or exhaustion.

HealthAlign-Agents:

- "Many new mothers feel this way; you're not alone."
- Adjusts metaphors to match cultural concepts of rest + recovery.
- Avoids stigmatizing phrasing.
- Still medically safe but emotionally human.

Quantitative improvements

- **+17%** factuality
- **+24%** empathy stability
- **+30%** cultural adequacy
- Zero finetuning / zero data collection

Impact

For low-resource communities

- Enables **safer health communication** in languages and dialects that have *no* labeled datasets.
- Reduces reliance on imported, English-centric content that ignores local beliefs and practices.
- Makes it easier to co-design tools with **community health workers** instead of external annotators.

For LLM alignment

- Shows that part of "alignment" can emerge from **structured self-play**, not only from RLHF on static data.
- Decomposes alignment signals into **factuality, empathy, and culture**, with an explicit objective L_{reflect}.
- Offers a reusable template for multi-agent alignment in other sensitive domains (mental health, crisis support, etc.).

For the broader ecosystem

- Encourages benchmarks and evaluations that center **underrepresented languages** instead of treating them as an afterthought.

Limitations & future work

Limitations

- Depends on retrieval quality and coverage.
- Judge may inherit LLM bias.
- No standard benchmarks for cultural adequacy.
- Multi-step self-play adds latency.

Future work

- Community-in-the-loop evaluation with clinicians + local health workers.
- Multi-judge setups for bias reduction.
- Cultural safety benchmarks across regions.
- Distillation of multi-agent behavior into efficient deployable models.

Culture-aware health AI should be universal.

Reflective self-play makes this scalable — without data.

Thank you!