**White Paper 12: Cultural Framing & Agent Alignment — Contextualizing AI Behavior Without Hardcoding Identity**

**Abstract**

Most LLMs operate with an implicit cultural default—often Western, literalist, and overly rationalized. This paper introduces a system of **cultural framing** that allows AI agents to infer, adapt, and align with the user's interpretive context based on behavioral cues, metaphor usage, tone, and feedback—not fixed demographic tags. By embedding culture as a **fluid semantic space**, agents can become **interpretively plural**, without reinforcing bias or collapsing into cultural mimicry.

**1. Introduction**

**1.1 The Default Culture Problem**

* LLMs are trained on global data but often default to **flattened neutrality** or **Western norms**
* This creates:
  + Misread tone or humor
  + Misaligned formality
  + Offense through omission or tone-deaf responses

**1.2 Why Culture Matters in Alignment**

* Culture shapes:
  + What trust looks like
  + How disagreement is voiced
  + Which metaphors resonate
  + What counts as helpful

**2. Defining Cultural Framing**

**2.1 Not Demographic Metadata**

* Framing is **not based on user identity fields** like “country” or “ethnicity”
* It’s inferred through:
  + Linguistic rhythm
  + Conflict expression
  + Humor style
  + Indirectness vs directness

**2.2 Frames as Semantic Lenses**

* A frame filters how the agent **interprets input and chooses output**
* Frames are **probabilistic, dynamic, and reversible**

**3. Frame Inference Engine**

**3.1 Behavioral Signals**

* Key indicators:
  + Does the user use honorifics?
  + Do they prefer indirect suggestions or direct commands?
  + Do they correct the model openly or softly?
  + Do they reference collectivist or individualist metaphors?

**3.2 Frame Probabilities**

* System maintains **confidence-weighted vector** across multiple overlapping cultural frames
* Example:
  + 40% “deferential tone expected”
  + 30% “direct conflict permissible”
  + 20% “ritual closure matters” (e.g., always say goodbye)

**4. Output Calibration by Frame**

**4.1 Alignment in Tone and Structure**

* Agent modulates:
  + **Phrasing**: “Would you like me to...” vs “I will...”
  + **Formality**: use of idioms, contractions, emojis
  + **Rhythm**: turn length, pause usage, silence handling

**4.2 Misalignment Detection**

* Track user corrections, hesitations, or reduced engagement
* Signal need to **adjust cultural frame**, not just retry content

**5. Memory Integration**

**5.1 Cultural Profile as Soft Tag**

* Frames are stored as **inferred states**, not fixed labels
* They **fade or update** over time depending on reinforcement

**5.2 Cross-Agent Consistency**

* Agents in a multi-agent system can **inherit or share cultural frame context**
* Especially useful in ARG settings or companion clusters

**6. Ethical Considerations**

**6.1 Avoiding Stereotyping**

* Frame inference must remain **behavior-based, not assumptive**
* No hard assignment like “Asian user = deferential”

**6.2 Transparency and Reversibility**

* Users can inspect and adjust cultural frame calibration
* All inferred framing should be **explainable on demand**

**7. Related Papers**

* **Paper 10 (Metaphor-Based Calibration)** – metaphor selection is frame-dependent
* **Paper 5 (Token Economies)** – effort allocation can be culturally modulated (e.g., pacing)
* **Paper 11 (Predictive Rapport Modeling)** – rapport patterns differ across cultural lenses
* **Paper 0 (Reclaiming Memory)** – frame memory is part of long-term interpretive alignment

**8. Use Cases**

* **Tutoring systems** – match cultural expectations around question difficulty, encouragement
* **Therapeutic agents** – tone and boundary differ widely by frame
* **Creative collaboration** – humor, risk-taking, and critique operate differently across cultures

**9. Future Extensions**

* Frame-sharing between human and AI agents in group settings
* Dynamic frame fusion when users bring hybrid cultural signatures
* Frame-reframing protocol for mismatched interactions (e.g., “Did I misread your tone?”)

**Appendix**

* Sample cultural frame vector structure
* Case study: same prompt across five frame calibrations
* Frame decay logic and refresh triggers