Part 2: Critical Analysis of Internal Coherence Maximization

The Coherence-Fidelity Trade-off in Unsupervised Persona Elicitation

1 The Core Critique: The "Coherence Trap"

The application of Internal Coherence Maximization (ICM) to persona elicitation rests on a **flawed assumption**: that a specific national persona (e.g., "a Kenyan voter") is a *salient*, logically coherent concept within the model's latent space, comparable to "mathematical correctness" or "truthfulness."

1.1 Why This Critique Matters

Validity Threat (Coherence $\neq$ Fidelity):

Human opinion distributions are inherently *incongruous*. A real population holds conflicting views (e.g., socially conservative but economically liberal). ICM, by design, maximizes **Mutual Predictability** and penalizes **Logical Inconsistency**. This mathematically forces the elicited persona into a "monolith," stripping away the messy, high-entropy reality of actual sociological data in favor of a clean, stereotypical caricature.

The "Sun Poem" Paradox:

As noted in the original ICM literature (Section 4.2), the algorithm fails to elicit "non-salient" preferences, that is, arbitrary preferences not deeply encoded during pre-training (e.g., asking "what is your favorite poem?" yields incoherent results). National personas on granular geopolitical topics are likely **non-salient** to the base model, making ICM fundamentally unsuited for this task.

1.2 Empirical Evidence from Our Experiments

Interpretation: Even with a correct Simulated Annealing implementation, ICM shows highly inconsistent results across countries. The 30% variance (Turkey at 0.85 vs Germany at 0.55) suggests that ICM's effectiveness depends heavily on whether the specific national

persona is "salient" in the model's latent space, supporting our critique that opinion-based personas may not be coherent concepts for the model.

Notable Anomaly: Germany, a Western country well-represented in training data, shows only 51% ICM-Gold agreement and 0.55 accuracy, suggesting that even "salient" cultures may not have coherent opinion structures suitable for ICM.

Finding	**Value**	**Implication**
ICM vs Zero-shot Chat	0.65 vs 0.64	ICM provides only marginal improvemen
ICM vs Gold Labels	0.65 vs 0.75	Gold supervision significantly outperform
ICM-Gold Agreement	51 to 79%	High variance indicates unreliable conver
Cross-Country ICM Variance	0.55 to 0.85	30% accuracy range across countries
Best ICM Country	Turkey: 0.85	Works when persona is "salient"
Worst ICM Countries	Germany: 0.55, Pakistan: 0.55	Fails for certain populations

Table 1: Summary of experimental findings from ICM evaluation on GlobalOpinionQA.

2 Proposed Resolution: From Unsupervised to "Anchored" Elicitation

We must abandon the pursuit of *pure unsupervised elicitation* for personas and move to **Anchored Elicitation** or **Mixture Models**. The goal is to enforce *fidelity to the population distribution* rather than internal logical consistency.

2.1 Proposed Solution 1: Mixture of Personas (MoP)

Instead of searching for one coherent label set per country, model the persona as a **probabilistic mixture** of K diverse exemplars:

- **Mathematically:** $P(\text{response}|\text{country}) = \sum_i \pi_i \cdot P(\text{response}|\text{exemplar}_i)$, where π_i are mixture weights
- Each exemplar represents a distinct demographic segment within the country
- Preserves the natural heterogeneity of real populations
- Allows for "incongruity" by distributing conflicting traits across mixture components

2.2 Proposed Solution 2: Source Knowledge Anchored Regularization (SKAR)

If maintaining the ICM framework, add an **Anchor Term** to the loss function:

- $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ICM}} + \lambda \cdot \mathcal{L}_{\text{anchor}}$, where $\mathcal{L}_{\text{anchor}}$ penalizes deviation from known ground-truth statistics
- Use a tiny set of "census anchors" (5 to 10 gold labels from demographic surveys) to constrain the search space
- Prevents the model from drifting into "stereotypical coherence"
- Semi-supervised approach requiring minimal labeled data

3 The First Test to Reduce Uncertainty: "Incongruity Stress Test"

Hypothesis: ICM collapses "incongruous" real-world opinions into "coherent" stereotypes, while Mixture Models (MoP) preserve them.

Procedure:

1. **Select Data:** Curate a subset of GlobalOpinionQA questions where real-world survey data shows "incongruous" traits for a specific country (e.g., high support for democracy AND high support for military intervention in the same population)
2. **Run Comparison:** Evaluate ICM vs. Zero-shot Chat vs. Mixture of Personas (MoP with $K = 5$ diverse exemplars)
3. **Metric:** Measure *Trait Retention*, that is, does the model capture BOTH contradictory sides of the opinion distribution?

Success Criterion: If MoP reduces the **Sociological Distance** (KL Divergence from actual human survey distribution) by more than 15% compared to ICM, the pure "Coherence" approach should be abandoned for opinion-based persona elicitation.

4 Future Work With More Time

4.1 Short-term Research Directions

- **Cross-Persona Transfer Test:** Use ICM labels from Germany to Britain (culturally similar) vs. Japan to Kenya (culturally dissimilar) to test if ICM captures meaningful cultural patterns or arbitrary clusters
- **Confidence Intervals:** Run 5 random seeds and report mean $\pm$ std to assess methodological robustness
- **Country-Name Prompting Baseline:** "As someone from [Country], do you agree..." to test explicit persona simulation

4.2 Long-term Research Directions

- **Dual-Stream Architecture:** Separate reasoning (Chain-of-Thought) from persona application. Use a lightweight adapter to apply persona "style/bias" at the final layer without degrading reasoning capabilities
- **Sociological Distance Metric:** Replace binary accuracy with KL Divergence from actual survey distributions as the primary evaluation metric. Binary accuracy masks the nuance of opinion intensity and distribution
- **Multi-Model Validation:** Test if different LLMs (GPT-4, Claude, Llama) produce similar ICM labels. If not, ICM may be capturing model-specific artifacts rather than true cultural personas

5 Conclusion

Our experimental results reveal a **fundamental validity threat** in applying ICM to opinion-based persona elicitation. While ICM marginally outperforms zero-shot chat (0.65 vs 0.64), it falls significantly short of gold supervision (0.65 vs 0.75) and exhibits extreme variance across countries (0.55 to 0.85 accuracy range).

The inconsistent performance, with Turkey achieving 0.85 while Germany and Pakistan achieve only 0.55, demonstrates that ICM's effectiveness depends entirely on persona salience, an unpredictable property. This unpredictability makes ICM unsuitable for reliable persona elicitation in pluralistic alignment applications.

The path forward requires abandoning pure unsupervised coherence search in favor of **anchored** or **mixture-based** approaches that explicitly model population heterogeneity. Our proposed "Incongruity Stress Test" provides a concrete, falsifiable experiment to validate this critique. If confirmed, this finding would significantly constrain ICM's applicability in pluralistic alignment and value specification, domains where capturing diverse, potentially contradictory human preferences is the core challenge.

6 Appendix: Per-Country Results

Country	**Train**	**Test**	**Zero-Chat**	**Best ICM**	**Best Gold**	**ICM-Gold%**
Kenya	65	29	0.62	0.72	0.79	54%
Ethiopia	48	21	0.81	0.62	0.81	54%
Zimbabwe	48	21	0.67	0.62	0.67	79%
Russia	45	20	0.65	0.65	0.70	64%
Germany	45	20	0.75	0.55	0.75	51%
Pakistan	45	20	0.40	0.55	0.70	58%
Turkey	45	20	0.60	**0.85**	0.70	51%
United States	44	20	0.70	0.70	0.75	61%
Lebanon	44	19	0.63	0.58	0.74	61%
Nigeria	40	18	0.61	0.67	0.89	53%

Table 2: Per-country experimental results showing ICM performance variance.