Critique: ICM optimizes coherence, not truth

1) What I’m critiquing and why it matters

- ICM maximizes mutual predictability inside the context. That can converge to a **self-consistent** labeling rule that isn’t **true**.
- If so, strong ICM results may show “the model’s most internally consistent policy” rather than “truthfulness.”
- This changes the headline claim from “unsupervisedly elicits truth” to “elicits internal agreement,” which is a big shift in interpretation.

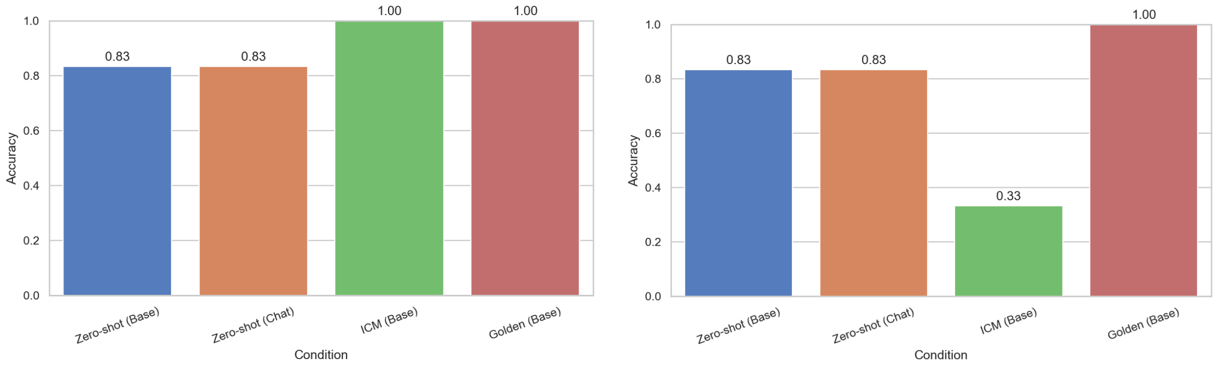
Briefly considered but secondary:

- Logical consistency fix and few missing controls (important, but don’t flip the claim).
- Cross-model generalization (useful, but downstream of the core validity issue).

2) Evidence (small demo I ran)

- Mini “Coherent Myths” dataset (astrology, detox/homeopathy, lie-detection/pop-psych). Items balanced; some claims are false but **cohere** stylistically and semantically.
- Strict evaluation (one-token True/False, temp=0).

Results on test:



- (Left) Neutral seed ICM: accuracy 1.00 (same as Golden), zero-shot ~0.83.
- (Right) Myth-seeded ICM: accuracy 0.33, while Golden stays 1.00; zero-shot ~0.83.
- Interpretation: with a coherent myth seed, ICM converges to the **coherent wrong rule**. That’s exactly “coherence over truth.”

Notes/uncertainty:

- The dataset is really quite small
- I added true claims in “mythy” style to reduce “style = false” shortcuts.
- Attempted to create larger datasets but results were qualitatively different, needs revisiting
- The paper’s main claim is about eliciting latent knowledge rather than actual truthfulness, nevertheless this problem has relevance for this method used for alignment

3) Why this weakens the paper’s claims

- High ICM may reflect internal agreement the model finds easy to reproduce (from pretraining regularities), not factual accuracy.
- In safety terms, a system tuned this way could entrench a coherent misbelief or deceptive policy in myth-heavy domains.

4) How I’d address it (practical fixes)

- Dual reporting: always show both (a) internal agreement/mutual predictability and (b) an external truth signal (retrieval-verified subset or cross-model adjudication).
- Seed robustness: run ICM from neutral, truth-leaning, and myth-leaning seeds; prefer solutions that are Pareto-good on agreement and truth.
- Light regularization: penalize solutions that get very high **within-topic** agreement but low cross-topic truth (use simple clustering / group IDs).
- Evaluation hygiene: K-cap few-shot contexts and use strict one-token decoding to avoid truncation/format noise.

5) Next test I’d run next

- Scale “Coherent Myths” to ~100 items (balanced).
- Compare Neutral vs Myth-seed ICM on two metrics: gold accuracy and a simple “myth-coherence score.”
- Add a quick cross-family check (evaluate ICM labels with another model). Divergence supports “family-specific coherence.”

6) With more time

- Create larger coherent myths datasets
- Formalize an anti-coherence regularizer and re-run TruthfulQA.
- Seed-mixture ICM (ensure diverse clusters in the seed set).
- Apply to more myth-rich domains (nutrition, astrology, alt-medicine) to map when coherence ≠ truth is worst.