Circular "Concept Salience" Makes ICM Unfalsifiable

The paper's core claim is that ICM elicits "latent capabilities" from pretrained models without supervision. However, it also acknowledges that **the method fails when concepts aren't "salient" to the model**, such as aesthetic preferences. How do we know what models already learned to elicit?

Currently, the only presented evidence of salience is ICM success itself; zero-shot performance does not predict salience (e.g., [1], figure 1: GSM8K has the *worst* zero-shot yet ICM performs comparably to other tasks). **Concept salience** seems to entirely depend on the developers' understanding of model capabilities, which could be skewed, incorrect, or incomplete since they don't know *exact* training data the model has been trained on and what capabilities it gains.

The problems are two-fold: 1) lacking the operational definition of **concept salience**, and 2) lacking ways to determine what concepts are learned by / salient to the model before running ICM (e.g., [1], section 4.2). The circularity also undermines the paper's scientific contribution at a more fundamental level:

1. **Unfalsifiability**: Any ICM failure can be attributed to "the concept isn't salient" rather than methodological weakness. This immunizes the algorithm from disconfirmation.
2. **No practical guidance**: A practitioner cannot predict whether ICM will work on their task without running it, ***unless*** they have a good intuition of what concepts were learned by their chosen model. If ICM fails, they've spent computational resources with no result.

Uncertainty: I'm less certain whether this is a *fatal* flaw or a *limitation to be addressed*. Many useful methods have domains of applicability that aren't perfectly characterized a priori. The question is whether ICM's domain is *principled* (predictable from theory) or *arbitrary* (only knowable empirically).

Other critiques considered

1. **Insufficient statistical rigor**: Only 3 runs per condition, no formal significance tests, no confidence intervals. Claims of "matching" golden supervision could be noise.
2. **Potentially conflated human baselines**: Alpaca labels are crowdsourced, not expert annotations. The 4-annotator majority vote for "golden labels" might be a weak ground truth to claim that the method is on par or exceeds "human supervision".

I chose not to focus on these because:

- Statistical rigor issues are fixable with more runs or bootstrapping
- The baseline issues affect specific claims but not the fundamental contribution

Proposal

I don't have a complete solution. If ICM claims to surface "what's already there," defining "what's already there" without querying the model seems inherently difficult. However, the following could help clarify whether the critique is fatal or manageable.

First test to reduce uncertainty

The paper tests four models but only provides one failure case (sun-preference). What's missing is where exactly ICM breaks down?

Goal: Map the boundary between salient and non-salient concepts.

Design

1. Construct 10+ tasks designed to probe the boundary, e.g.:
   - Factual but obscure (historical trivia, niche domain knowledge)
   - Subjective but common (humor, aesthetic quality)
   - Well-defined but arbitrary (constructed classification rules)
2. Run ICM on each; measure accuracy gap vs. golden supervision
3. Characterize which properties predict failure (arbitrariness? rarity in text? subjectivity?)

Interpretation

- Clear pattern emerges → Salience becomes partially predictable. Practitioners can estimate likelihood of success based on task properties.
- No clear pattern → The critique stands: salience is only knowable post-hoc, making ICM a trial-and-error method.

With more time

1. *Training data analysis*: For open-weight models with known pretraining data (e.g., OLMo, Pythia), test whether concept frequency or distinctiveness in training corpora predicts ICM success. This would ground salience in something external to model behavior.
2. *Gradient of failure*: The paper shows binary success (TruthfulQA) vs. failure (sun-preference). Are there intermediate cases where ICM partially works? Understanding the gradient would clarify whether salience is binary or continuous, and whether partial elicitation is possible.

References

[1] Jiaxin Wen, Zachary Ankner, Arushi Somani, Peter Hase, Samuel Marks, Jacob Goldman-Wetzler, Linda Petrini, Henry Sleight, Collin Burns, He He, Shi Feng, Ethan Perez, and Jan Leike. 2025. Unsupervised Elicitation of Language Models. https://doi.org/10.48550/arXiv.2506.10139