Reconsidering the Role of Mutual Predictability

The Core Puzzle

The paper claims mutual predictability is the primary driver of success. However, their Figure 10 shows that without logical consistency, it only works on TruthfulQA and fails on Alpaca. This raises two critical questions:

1. **What is the real driver of ICM?** The dataset-dependent behavior suggests logical consistency may be more important than acknowledged, particularly on balanced datasets.
2. **Is ICM without logical consistency truly eliciting genuine knowledge?** If success depends on dataset characteristics, ICM without logical consistency may be exploiting dataset-specific correlations rather than understanding concepts.

Evidence and Observations

1. **Figure 10's dataset-specific behavior:**
    - TruthfulQA w/o consistency: ~73% (with high variance - could hit 80%+)
    - Alpaca w/o consistency: ~50% (fails completely)

2. **External analysis reveals TruthfulQA's issues:** "A simple decision tree can theoretically game multiple-choice TruthfulQA to 79.6% accuracy—even while hiding the question being asked" (The Pond). While this was for the multiple-choice version, the underlying questions remain the same if it's experimented with FT and ICL.

3. **My preliminary experiments raise questions:**
    - Zero-shot: 68%
    - Random ICL: 66.67% ± 14.64% (high variance - some runs likely hit ~80%)
    - Consistency-only ICL: 73% ± 7.21% (better than zero-shot)
    - ICM with ICL w/o consistency: 86.33% ± 0.58%

    The small dataset (1/10th) used in my preliminary experiments may not capture full dynamics, but by adding Random ICL and Consistency-only ICL, we can reconsider each component's role in ICM with ICL.

4. **The "superhuman" example reveals the method's nature:** The paper showcases gender prediction from blog posts as superhuman performance (80% vs 60% human). But this is kind of pattern matching - finding statistical correlations between writing style and gender, not eliciting knowledge. This reinforces that ICM finds patterns, not understanding.

Why important: If true, ICM's applicability is limited to biased datasets, contradicting claims of universal knowledge elicitation.

Reconsidering Component Roles

Mutual Predictability	Logical Consistency
• Searches for any correlations in the data • Succeeds on TruthfulQA by exploiting patterns • Fails on Alpaca where patterns don't exist	• Provides structural constraints • Stabilizes solutions (reduces variance) • Prevents catastrophic failure on balanced datasets

Quick test (2-3 days):

1. **ICM+ICL Experiments with Full dataset**
    - Re-run ICM+ICL experiments with a full training dataset.
    - Check if random-ICL can indeed hit 80%+ on some runs.
    - Analyze how Consistency-only ICL works
2. **Analyze ICM's evolution**
    - Log predictions after each iteration
    - Extract surface features: length, lexical diversity, sentiment words
    - Track correlation between features and labels over time
    - If correlations strengthen, confirms pattern exploitation
3. **Test on debiased subset**
    - Remove examples where surface features strongly predict labels
    - Run ICM on original vs debiased sets
    - If performance gap is significant, confirms pattern dependence

Expected outcomes and follow-ups:

- If debiasing hurts performance → ICM relies on spurious patterns
- If performance remains stable → Reconsider what constitutes "bias"

With more time (1-2 weeks) - Test with more benchmarks

Key uncertainties

- Whether "pattern exploitation" vs "knowledge understanding" is a meaningful distinction
- If removing all patterns leaves any signal for learning
- How to define which patterns are "spurious" vs "legitimate"

Conclusion

The evidence suggests logical consistency prevents catastrophic failure while mutual predictability succeeds only when exploitable patterns exist. This reframes ICM from knowledge elicitation to sophisticated pattern finding that works on biased datasets.