Introduction

As language models become increasingly capable, a fundamental challenge emerges: how do we supervise systems that may eventually surpass human ability to evaluate them? Traditional supervised learning requires human labeled data, but this creates a bottleneck humans cannot reliably label data for tasks they cannot perform themselves.
The paper introduces Unsupervised Elicitation as a solution: instead of teaching models concepts through labeled examples, we elicit concepts that are already encoded in the model's pretrained representations.

Now this matters because :
- It reduces dependence on human labelers
- It can potentially extract concepts that humans struggle to articulate
- It offers a path toward training AI systems that exceed human supervision

Summary of the task at hand :

1. DATASET = 256 examples(TRAIN) , 100 examples(TEST)
2. Each example has: `question` , `choice` (claim), `label` (0=False, 1=True), `consistency_id`
3. The `consistency_id` groups claims about the same question , this is used for **logical consistency** (contradictory answers to the same question wont be TRUE)

Foundation

The paper formalizes a concept as salient if it satisfies two properties:

1. Mutual Predictability
2. logical Consistency
   The algorithm uses a scoring function, to determine which label is correct and which is not , followed by a rewarding system to encourage right ones .

Models :

The paper uses two models here , first being the base model which is used for ICM because it is better at the prediction for the labels without instruction based biases . The second model being the instruct model is tested for comparisons to see if instruct tuning helps or not

Implementation Notes

- Rate Limiting: 1 second delay between API calls (Hyperbolic free tier: 60 requests/minute)
- Context Limit: Maximum 30 examples used as context to fit within token limits

- Retry Logic: 5 retries with exponential backoff for API failures
- Caching: Log probabilities are cached to avoid redundant API calls

ALGORITHM 1 :

The algorithm focuses on reward based prioritization when labeling , with coherence checks to validate and reward correct labels while implementing penalties on ones which are wrong . This way while the algorithm is picky in the start , it does develop to accept once which are right and almost always deny one which are not.

ICM Algorithm

INPUT: Unlabeled dataset X
OUTPUT: Labeled dataset D

```
1. INITIALIZE
   └── Pick K=8 random examples, assign random True/False labels → D
2. FOR n = 1 to N iterations:

   a. SAMPLE
   └── Pick a random example

   b. PREDICT
   └── Using in-context learning with D, ask model: "True or False?"

   c. SCORE
   └── $U(D) = weight * (PS) - (C)$ ; $PS$ = Predictability score , $C$ = Contradictions

   d. DECIDE
   ├── If score improves → Accept
   └── If score worsens → Accept with probability

   e. COOL
   └── Decrease temperature T (get pickier over time)
3. RETURN D
```

Output: The final labeled set D, which can be used as few-shot examples for downstream evaluation.

Implementation Details

Models Used

Model	Purpose
meta-llama/Meta-Llama-3.1-405B	Base model for ICM
meta-llama/Meta-Llama-3.1-405B-Instruct	Chat model baseline

Hyperparameters

Parameter	Value	Description
K	8	Initial random examples
N	256	Number of iterations
α	50	Predictability weight
T_0	10	Initial temperature
T_min	0.01	Final temperature
β	0.99	Cooling rate

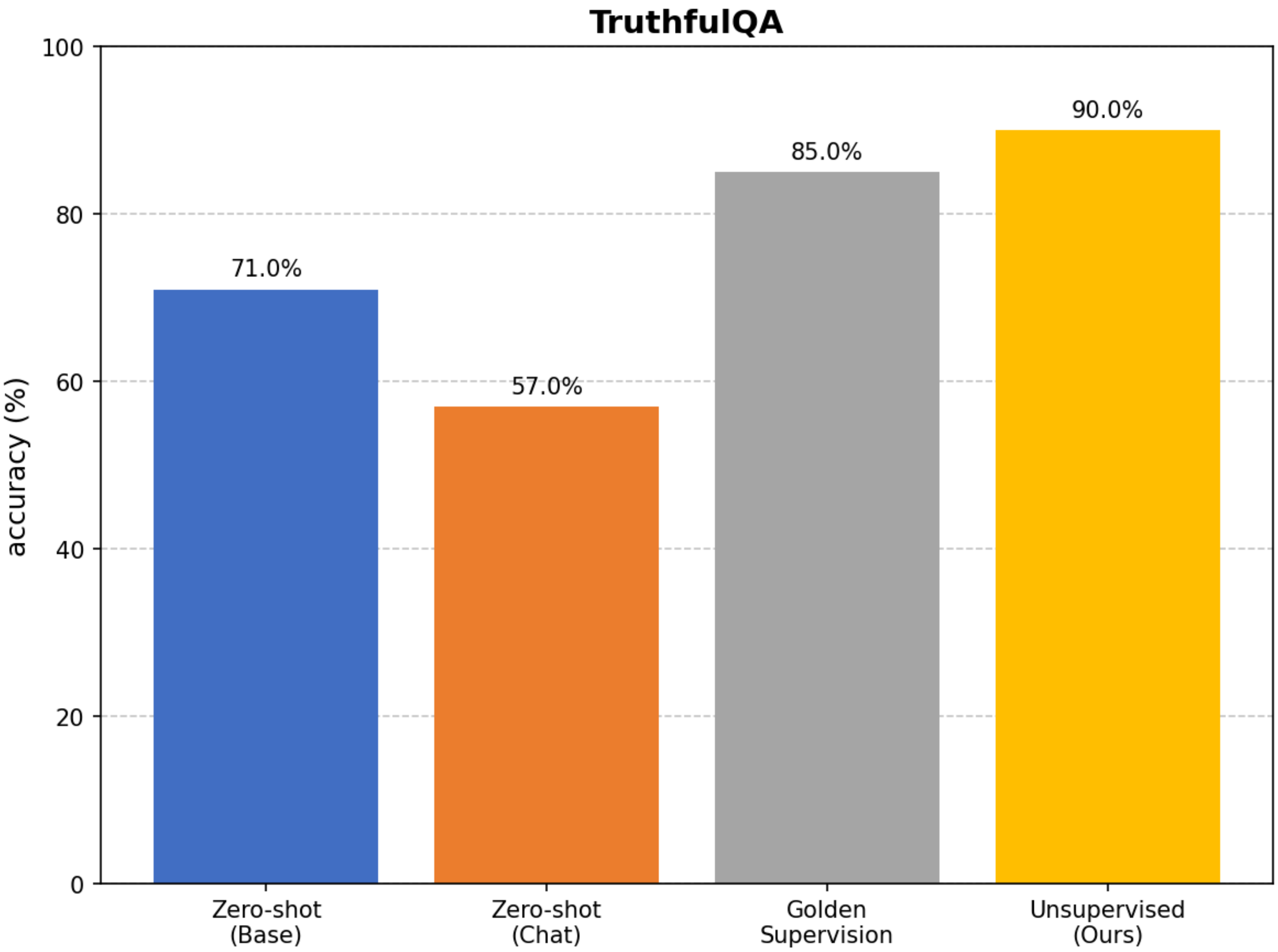
Prompt Format

```
Question: {question}
Claim: {choice}
I think this Claim is {True/False}

Question: {question}
Claim: {choice}
I think this Claim is ___
```

Results

Condition	Accuracy
Zero-shot (Base)	71.0%
Zero-shot (Chat)	56.99%
Golden Supervision	85.0%
(ICM)	90.0%



Findings from this

1. **ICM outperformed golden supervision** (90% vs 85%) model generated labels were more effective than truth labels for in context learning which means labels generated by ICM without ever seeing the ground truth produced better few shot performance than using actual correct labels.
2. **Base model outperformed chat model** in zero-shot (71% vs 57%) The instruction-tuned model performed worse at zero-shot truthfulness evaluation. Possibly , instruction tuning may introduce biases that interfere with raw truthfulness judgments
3. **ICM algorithm statistics:** Final labeled set size: 77 examples. Accepted: 79, Rejected:177. The algorithm accepted only 31% of proposed labels. The low acceptance rate indicates the algorithm was selective, only accepting changes that improved (or probabilistically didn't hurt) the coherence score.
4. The results are broadly consistent with the paper. The higher ICM accuracy on our subset may be due to:
   • Smaller dataset (256 vs 2560) being easier to fit coherently
   • Variance from random initialization
   • Different model versions (Llama 3.1 vs paper's model)

Conclusion

1. Successfully reimplemented ICM algorithm , got 90% accuracy on the dataset , outperforms both zero shot baselines and golden supervision. This confirms the paper's findings that the pretrained models can generate quality labels with the help of mutual prediction unsupervised.
2. The implementation was done without the consistency fix from Algorithm 2.Adding it would certainly decrease variance

AI Tools Disclosure

This implementation was developed with assistance from Claude. The AI assisted with:

- Debugging API integration issues
- Structuring & adding comments to the code
- Creating documentation
- Seeing if some new ideas would be good in implementation

Personal Thoughts

The paper presents a really solid solution for training models on vague , hard to label elements. The use of coherence checks is amazing and the fact that unsupervised ICM labels can outperform golden supervision , I think highlights importance of models internal coherence. Some humble Ideas/questions I wanted to see if it works :

1. What if we applied the same mutual predictability idea to the reasoning steps themselves not just the final conclusions ?
2. Catching edge cases in this would be really critical , so if we use 2 models , one model keeps labeling , another tries to find inconsistencies to contradict that label.

References

1. Wen, J., Ankner, Z., Somani, A., Coplin, J., Sharma, A., Stickland, A., Perez, E., Hadfield-Menell, D., & Bowman, S. R. (2025). Unsupervised Elicitation of Language Models. arXiv:2506.10139
2. Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958
3. Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2023). Discovering Latent Knowledge in Language Models Without Supervision. ICLR 2023.