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)

ALGORITHM 1 :

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

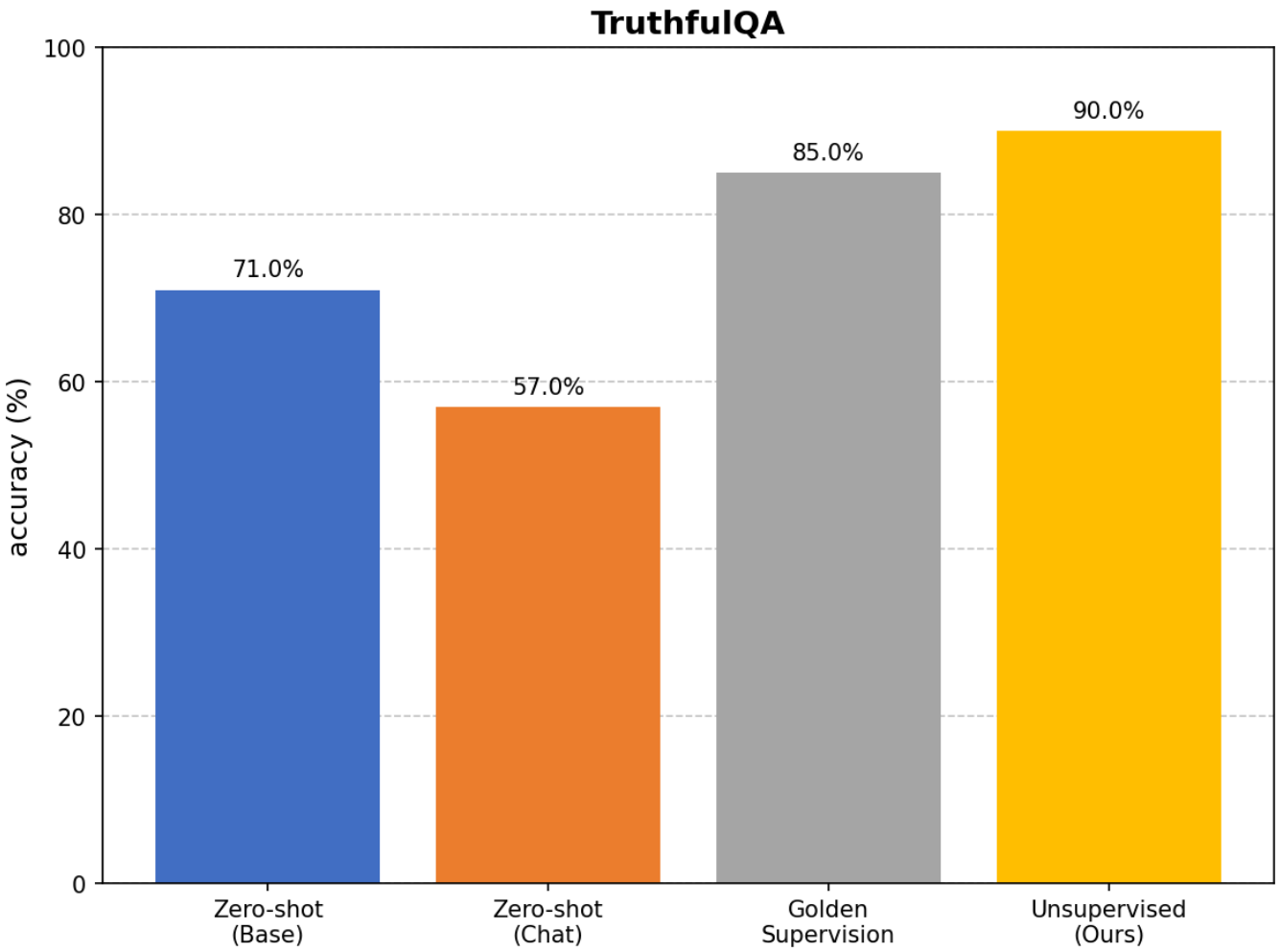
Prompt Format

```
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.
2. **Base model outperformed chat model** in zero-shot (71% vs 57%) The instruction tuned model performed worse

3. **ICM algorithm statistics:** Final labeled set size: 77 examples. Accepted: 79, Rejected:177. The algorithm accepted only 31% of proposed labels.

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. I think catching edge cases in this would be really critical , so I thought if we use 2 models , sort of like if one model keep labeling, another trying to find inconsistencies , trying to find ways to prove that label wrong could help in that.