

RRMC: Robust Revision-MI Control

Stopping Rule Evaluation on AR-Bench Detective Cases

RRMC Team

January 30, 2026

Active Reasoning Benchmark Experiments

Table of Contents

1. Background: Stopping Methods
2. Initial Method Runs
3. Parameter Grid Search
4. Validation Results & Analysis
5. Conclusions & Proposed Fixes

Background: Stopping Methods

The Active Reasoning Problem

Goal

When should an LLM **stop asking questions** and provide a final answer?

AR-Bench Detective Cases (DC):

- LLM plays detective, asks yes/no questions to an oracle
- Must identify the correct suspect from candidates
- Trade-off: More questions → more information but higher cost

Key Metrics:

- **Accuracy:** Correct suspect identification rate
- **Avg Turns:** Average questions asked before stopping

Method Overview: Simple Baselines

Fixed Turns

- Always ask exactly N questions
- No adaptive stopping
- Baseline comparison

Verbalized Confidence

- Ask LLM: “Rate confidence 1-10”
- Stop when confidence \geq threshold
- Simple but relies on self-assessment

Self-Consistency

- Sample k answers at temperature > 0
- Stop when majority agrees (\geq threshold)
- Measures answer stability

Semantic Entropy

- Cluster semantically similar answers
- Compute entropy over clusters
- Stop when entropy $<$ threshold

Method Overview: MI-Based Methods

MI-Only

- Estimate Mutual Information via self-revision
- $MI = H[\text{answers}] - H[\text{answers}|\text{revision}]$
- Stop when $MI < \text{threshold}$
- Low MI = new info won't change answer

Robust MI

- Multi-variant MI estimation
- Uses “base” and “skeptical” prompts
- More robust uncertainty quantification

Key Insight

MI Interpretation

High MI: Model uncertain, new questions likely to change answer

Low MI: Model confident, safe to stop

Threshold Effect

- Lower threshold → stricter stopping criterion
- Requires more certainty before stopping
- Results in more turns but higher accuracy

Method Overview: New Baselines

KnowNo

- Conformal prediction approach
- Build prediction sets with coverage guarantees
- Stop when set size \leq threshold
- From: Ren et al. (2023)

CIP-Lite

- Simplified conformal inference
- Track answer set across samples
- Stop when answers converge

UoT-Lite

- Inspired by Uncertainty of Thoughts
- Lightweight variant for stopping
- Uses answer diversity as signal

Implementation Status

These methods are newly ported from literature. Grid search reveals they may stop too early on DC tasks.

Initial Method Runs

Initial Baseline Results (5 puzzles)

Method	Threshold	Accuracy	Avg Turns
Fixed Turns (5)	–	40%	5.0
Self-Consistency	0.7	20%	1.0
Semantic Entropy	0.5	20%	1.0
Verbalized Confidence	8.0	20%	1.0
MI-Only	0.3	20%	1.0
Robust MI	0.3	20%	1.0

Problem Identified

Most methods stop at **Turn 1** with only 20% accuracy!

→ Thresholds too permissive, triggering immediate stops

Parameter Grid Search

Grid Search Methodology

Objective

Find optimal thresholds for each stopping method

Setup:

- 5 puzzles per threshold value (quick iteration)
- Parallel execution across 6 terminals
- Track accuracy and average turns

Threshold Ranges Tested:

Method	Values Tested
MI-Only	0.01, 0.02, 0.05, 0.1, 0.2, 0.3
Robust MI	0.01, 0.02, 0.05, 0.1, 0.2, 0.3
Verbalized Confidence	5.0, 6.0, 7.0, 8.0, 9.0
Self-Consistency	0.5, 0.6, 0.7, 0.8, 0.9, 1.0
Semantic Entropy	0.05, 0.1, 0.2, 0.3, 0.5

Grid Search Results

Method	Best Threshold	Accuracy	Avg Turns	Note
MI-Only	0.01	80%	6.0	★ Winner
Verbalized Conf.	9.0	80%	20.6	Expensive
Semantic Entropy	0.5	40%	1.0	Stops early
Self-Consistency	1.0	40%	1.6	Stops early
KnowNo	2–3	40%	1.0	Stops early
CIP-Lite	1	40%	1.6	Stops early

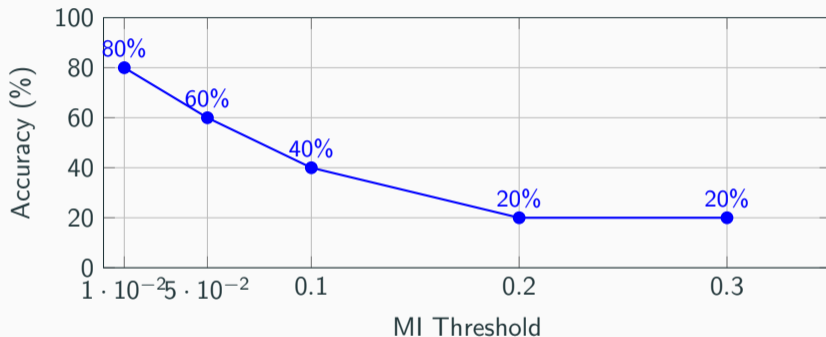
Key Finding

MI-Only with threshold=0.01 achieves 80% accuracy with only 6 turns average.

Issue

Methods stopping at Turn 1 are just guessing without gathering information.

MI-Only: Threshold vs Performance



Insight: Lower threshold \rightarrow stricter stopping \rightarrow more turns \rightarrow higher accuracy

Validation Results & Analysis

Validation Results: The Early-Stopping Problem

Grid Search Did NOT Generalize!

Results on 20 puzzles reveal a fundamental problem.

Method	Accuracy	Avg Turns	Turn 1 %	MI=0 %
CIP-Lite	45%	1.4	75%	—
KnowNo	40%	1.0	100%	—
Self-Consistency	35%	1.1	90%	75%
Semantic Entropy	35%	1.4	80%	75%
Fixed Turns (10)	30%	10.0	0%	—
Verbalized Conf.	30%	23.1	0%	—
MI-Only (0.01)	25%	4.1	70%	70%
Robust MI	25%	3.0	70%	45%

Grid search showed 80% for MI-Only, but validation shows only 25%!

Root Cause: $MI = 0$ at Turn 1

MI Scores for MI-Only (first 10 puzzles):

Puzzle	MI at Turn 1	Result
0-4	0.0	Stop → Wrong
5	0.45	4 turns → Wrong
6-7	0.0	Stop → Mixed
8	0.45	5 turns → Wrong
9	0.0	Stop → Correct

The Problem

At Turn 1, LLM generates **identical answers** across all $k=6$ samples.

Identical → zero entropy → **$MI = 0$**

$MI = 0 < 0.01$ → **Immediate stop**

Fundamental Issue

MI estimator conflates “model is confident” with “model has enough information.”

At Turn 1 with **zero clues gathered**, a confident model is just **guessing**.

Analysis: Why Grid Search Misled Us

Grid Search (5 puzzles):

- 80% accuracy, 7.4 avg turns
- Those 5 puzzles happened to have **diverse initial samples**
- Non-zero MI triggered more turns
- **Lucky variance!**

Validation (20 puzzles):

- 25% accuracy, 4.1 avg turns
- 70% of puzzles: $MI = 0$ at start
- 70% stopped at Turn 1
- True performance revealed

Key Lesson

5 puzzles is not enough for reliable threshold tuning.

Small sample variance can create misleading “winners.”

Turn Distribution Analysis

MI-Only Turn Distribution:

Turns	Count	Accuracy
1	14	28.6%
4–5	2	0%
8	1	100%
12–14	2	0%
25 (max)	1	0%

Verbalized Confidence:

Turns	Count	Accuracy
2–9	2	0%
25 (max)	18	33.3%

Opposite problem: never stops, always hits max turns!

Observation: Methods either stop too early (Turn 1) or too late (Turn 25). No method finds the sweet spot.

Concrete Example: Turn-1 Stop (No Questions Asked)

Puzzle 0 - WRONG

- Prediction: Suspect 0
- Ground Truth: Suspect 3
- MI at Turn 1: **0.0**
- Questions asked: **NONE**

Conversation: []

Model guessed without asking a single question!

Puzzle 1 - CORRECT (lucky)

- Prediction: Suspect 0
- Ground Truth: Suspect 0
- MI at Turn 1: **0.0**
- Questions asked: **NONE**

Conversation: []

Same behavior – just happened to guess correctly.

14/20 puzzles stopped at Turn 1 with empty conversations!

The model is “confidently guessing” – all $k=6$ samples agree on same answer before asking anything.

Concrete Example: Multi-Turn Success (Rare)

Puzzle 17 - CORRECT after 8 turns

MI Scores: [0.22, 0.64, 0.78, 0.41, 0.87, 0.64, 1.01, **0.0**]

Turn 1: Asked Eleanor Whitaker: "What were you doing near the studio?"

Turn 2: Asked Michael Turner: "Who was with you at the project site?"

Turn 3: Asked Clara Mitchell: "Did you have disagreements with Jonathan?"

...

Turn 8: MI drops to 0.0 → Stop → **Correct!**

What went right

- Model was uncertain from start (MI = 0.22 \leq 0)
- Asked diverse questions to different suspects
- Gradually gathered information until confident
- **Only 1 of 6 multi-turn puzzles succeeded (16.7%)**

Concrete Example: Multi-Turn Failure (Common)

Puzzle 18 - WRONG after 25 turns (max)

Prediction: Suspect 0 — Ground Truth: Suspect 2

MI never stabilized: [0.32, 0.26, 1.24, 0.87, ... , 0.87, ∞]

Turn 1: Asked Clara Bennett: "What were you doing 9-10pm?"

Turn 2: Asked Marcus Langley: "Did you have disputes with Jonathan?"

Turn 3: Asked Dr. Harper: "Were you in the office 9-10pm?"

...

Turn 22: Asked Marcus Langley: **SAME QUESTION**

Turn 23: Asked Dr. Harper: **SAME QUESTION**

Turn 25: Hit max turns → **Wrong!**

Problems identified

- Model asks **repetitive questions** (same suspect, same question)
- Doesn't synthesize information across turns
- MI never drops → runs to max turns → still wrong

Conclusions & Proposed Fixes

Key Takeaways

1. All current stopping methods fail on DC task

- Best accuracy: 45% (CIP-Lite) – barely better than random
- MI-Only dropped from 80% (5 puzzles) to 25% (20 puzzles)

2. The $MI = 0$ problem is fundamental

- At Turn 1, LLM samples are often identical $\rightarrow MI = 0$
- Zero MI triggers immediate stop before gathering information
- 70–100% of puzzles stop at Turn 1 for most methods

3. Grid search on small samples is unreliable

- 5 puzzles showed 80% accuracy (lucky variance)
- 20 puzzles revealed true 25% performance
- Need larger validation sets (50+) for reliable tuning

Proposed Fixes

1. Minimum Turns Guard

- Force at least N turns before checking MI
- E.g., `min_turns = 3`
- Ensures some information gathering

2. MI Floor with Context

- Only stop if $MI < \text{threshold}$ **AND** $\text{turns} > \text{min}$
- Combine uncertainty with effort

3. Diversity Detection

- Detect when samples are identical
- If all same \rightarrow don't trust $MI = 0$
- Force temperature increase or more sampling

4. Progressive Threshold

- Start with high threshold (keep asking)
- Gradually lower as turns increase
- Natural “explore then exploit”

Next Steps

- **Implement minimum turns guard:** Quick fix to prevent Turn-1 stops
- **Larger validation:** Run on 50+ puzzles for reliable metrics
- **Analyze sample diversity:** Understand when/why samples collapse
- **Cross-task evaluation:** Test on Situation Puzzles (SP), Guessing Numbers (GN)
- **Calibration pipeline:** Use risk-controlled threshold calibration with larger calibration set

Thank you!

Code: [github.com/\[repo\]/RRMC](https://github.com/[repo]/RRMC)