RRMC — Week 5

AllSuspects, Ensemble, and DQS on DC & GN

2026-02-06

Qwen 2.5 7B Instruct via OpenRouter

Detective Cases — Stopping Rules

DC: All Methods Head-to-Head (20 puzzles)

Method	Accuracy	Avg Turns	Cost	Note
CIP-Lite	**45%**	1.4	1×	*best overall*
KnowNo	40%	1.0	1×	stops T1 always
Self-Consistency	35%	1.4	1×	
Semantic Entropy	35%	1.4	1×	
Fixed Turns (10)	30%	10.0	1×	
Verb. Confidence	30%	23.1	1×	never stops
MI-Only	25%	4.1	1×	MI=0 at T1
AllSuspects+CIP	15%	6.6	1×	↓30%
Ensemble+AS+CIP	25%	7.0	8×	↓20%

AllSuspects: 45% → 15%. Ensemble + AllSuspects: 45% → 25% at 8× cost.

The Turn-1 Paradox

CIP-Lite Stops at Turn 1 — And That's the Best Strategy

Subgroup	n	Acc
CIP-Lite stops T1	15	**47%**
CIP-Lite stops T2+	5	40%
AllSuspects (T5+)	20	15%

AllSuspects degradation ratio:

- 7 correct → wrong
- 1 wrong → correct
- **Net: −6 puzzles**

Why T1 works

CIP-Lite samples k=8 answers from the case background alone. When all 8 agree (set size = 1), the model **consistently** identifies the same suspect from the description.

This is genuine comprehension of the case text — the model's **strongest signal**.

Why more turns hurt

NPC responses are LLM-generated, generic alibis ("I was at home"). Every extra turn adds **noise** that overwrites the correct first impression.

Ensemble Failure Mode

Ensemble Consensus Is Anti-Correlated with Correctness

Consensus	n	Acc
2/6 (33%)	6	17%
3/6 (50%)	8	38%
4/6 (67%)	6	**17%**

No puzzle reached 5/6 or 6/6.

Higher confidence $\not\Rightarrow$ correctness.

Root cause: correlated errors

- Qwen 7B has **systematic biases** (suspects the “most vocal” character)
- 6 trajectories with temperature 0.5–1.0 make the **same mistake**
- Majority vote **locks in** the wrong answer

Example (Puzzle 6):

Baseline & AllSuspects: ✓ correct.

Ensemble: 4/5 vote for wrong suspect at 80% confidence.

→ Ensemble **killed** the correct answer.

Why Intuitions Failed

More Investigation $\neq$ Better Decisions

Intuition	Reality	Evidence
"Stopping at T1 is premature"	T1 accuracy (47%) is 3× AllSuspects (15%)	Case background is the strongest signal
"Question all suspects for fair coverage"	NPC responses are generic, undifferentiated	7:1 degradation ratio
"Ensemble reduces variance"	Errors are correlated, not independent	Consensus anti-correlates with accuracy

The bottleneck

NPC response quality, not stopping rules.

Case background	High signal
NPC turns 1–5	Low signal
NPC turns 5+	Noise

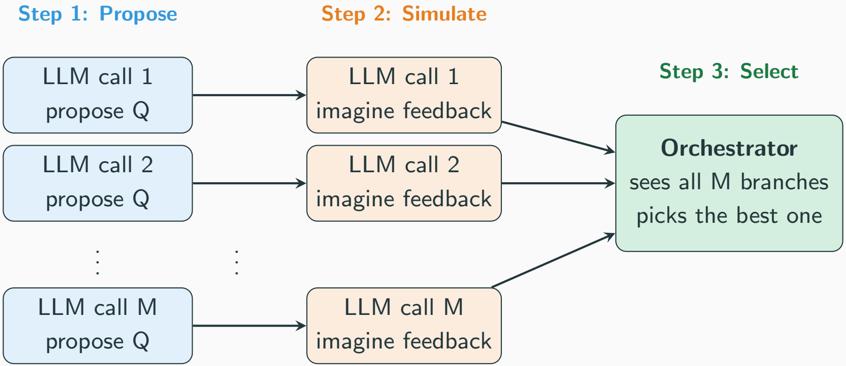
Implication

For Qwen 7B on AR-Bench DC, CIP-Lite at Turn 1 is **near-optimal**. Improvement requires better models or better NPCs.

DQS — Deliberative Question Selection

DQS: Can Better Questions Help?

Instead of improving *when* to stop, improve *what* to ask.



Cost: $2M+1$ LLM calls per turn. **LLM decides everything** — no algorithmic scoring.

Works for both DC (suspects/questions) and GN (guesses/feedback).

Guessing Numbers (GN)

GN Baseline: LLM Cannot Play Bulls & Cows

GN Baseline — 0% accuracy

Metric	Value
Accuracy	**0%**
Avg turns	25.0 (max)
Tokens	632K

Example (Puzzle 0, secret = 8362):

T14: Guess 8367 — **3 bulls!**
T15: Guess 2591 — 0 bulls
T16: Guess 9183 — 0 bulls
(throws away T14 info)
T24: Final: 7359 — wrong

Diagnosis

The LLM **ignores feedback**.

- Gets 3/4 digits right at T14
- Immediately guesses unrelated numbers
- Never returns to the near-miss
- Cannot track constraints across turns

Information-theoretic baseline

Optimal play solves Bulls & Cows in **5–6 guesses**.

5040 valid numbers $\xrightarrow{\text{entropy-optimal}}$ 1 in $\sim$5.5 turns.

LLM can't do this — it doesn't reason about elimination.

GN + DQS: Results Pending

What DQS does for GN:

1. M parallel LLM calls: each proposes a guess
2. M parallel LLM calls: each imagines the feedback
3. 1 orchestrator call: picks the most informative guess

Key question:

Can the LLM select better guesses when it *deliberates* about possible outcomes, even though it can't track constraints?

Method	Acc	Turns
GN Baseline	0%	25.0
GN + DQS	*running*	*running*
(Optimal)	*100%*	*5.5*

Expectation

DQS may improve over 0% — but unlikely to match optimal play. The LLM's constraint-tracking limitation is fundamental.

Key Findings

Week 5 Findings

1. **DC: CIP-Lite at Turn 1 is near-optimal for Qwen 7B**
   Model's case-background comprehension is its best signal. NPC responses degrade accuracy monotonically.

2. **"Improvements" consistently hurt**
   AllSuspects: −30pp. Ensemble: −20pp at 8× cost. 7:1 degradation ratio. Consensus anti-correlated with correctness.

3. **GN: LLM fundamentally cannot track constraints**
   0% accuracy on Bulls & Cows. Ignores feedback, guesses randomly. DQS may help question selection but can't fix reasoning.

4. **The bottleneck is model capability, not the framework**
   Stopping rules, ensembles, AllSuspects are sound in principle. They need a model that can synthesize multi-turn evidence.

Next Steps

- **DQS results** — GN and DC runs in progress
- **Stronger model** — test GPT-4 / Claude to see if AllSuspects / Ensemble / DQS become beneficial
- **GN: structured prompting** — step-by-step elimination strategy in system prompt (from AR-Bench templates)
- **Larger DC evaluation** — 50+ puzzles for reliable metrics

End