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×	↓30pp
Ensemble+AS+CIP	25%	7.0	8×	↓20pp
DQS + Fixed(10)	**60%**	**10.0**	**11×**	**↑30pp**

DQS is the **only method that improves** over baseline: 30% → 60%. AllSuspects/Ensemble both degrade.

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.

Why AllSuspects & Ensemble Fail

AllSuspects Failure: Puzzle 1 — Correct → Wrong

Victim: Dr. Jonathan Reed — Guilty: **Evelyn Carter (#0)**

Baseline: ✓ Evelyn Carter (T1)

Reads case background → immediately identifies Carter.

AllSuspects: ✗ Samantha Greene (T6)

Forced to question all 5 suspects:

T1: Carter — "I was at home, reviewing data."
T2: Whitmore — "I was at home, relaxing."
T3: Collins — "I was at home, making sense."
T4: Donovan — "I was at home, relaxing."
T5: Greene — "I was at home watching a movie, but I heard a **sighting near the park**."

What went wrong

Every suspect says "I was at home."

Greene adds one irrelevant detail ("sighting near the park") → model latches onto it as evidence → **changes answer from Carter to Greene**.

The NPC responses carry **no signal** — all generic alibis. The model over-interprets minor phrasing differences.

Pattern

7 of 20 puzzles: baseline correct, AllSuspects flipped it wrong. Only 1 puzzle went the other direction. **Ratio: 7:1 against.**

Ensemble Failure: Puzzle 6 — 80% Confident and Wrong

Guilty: **Michael Turner (#1)**

Method	Pred	Correct?
Baseline	#1 Turner	✓
AllSuspects	#1 Turner	✓
Ensemble	#0 Carter	✗

Ensemble consensus: **4/5 trajectories** (80%) voted for the **wrong** suspect.

Both baseline and AllSuspects got it right independently — the ensemble **killed** the correct answer.

What happened

All 6 trajectories run at the **same temperature**[1]. The model has a systematic bias toward Carter (most narrative connection to victim).

4 of 5 valid trajectories follow the same reasoning path → same wrong answer.

Majority vote **amplifies** the bias instead of correcting it.

Broader pattern

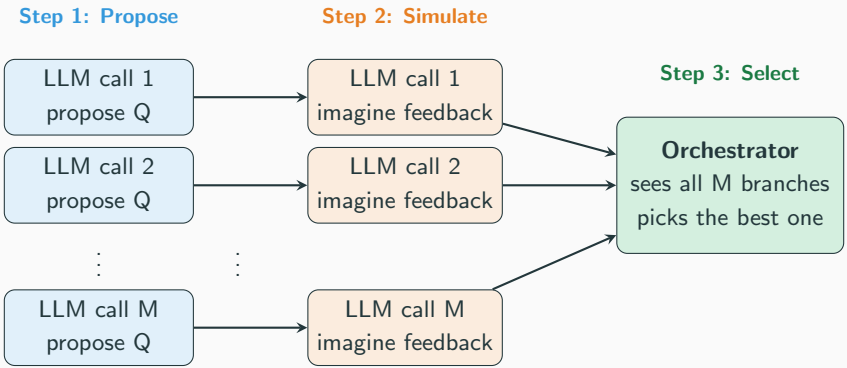
Ensemble consensus anti-correlates with correctness. 4/6 consensus: only 17% accurate. No puzzle reached 5/6 or 6/6.

[1] Temperature diversity was configured but not applied due to a wiring bug. All trajectories ran at temp=0.7.

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).

DC + DQS: The First Method That Actually Helps

Method	Accuracy	Avg Turns	API Calls	Cost
Fixed Turns (baseline)	30%	10.0	∼200	1×
CIP-Lite (best stopping)	45%	1.4	∼500	2.5×
AllSuspects + CIP-Lite	15%	6.6	∼500	2.5×
Ensemble + AS + CIP-Lite	25%	7.0	∼4200	21×
DQS + Fixed Turns	**60%**	**10.0**	**2179**	**11×**

What DQS proves

- Same 10 turns, **2× accuracy**
- The bottleneck was **question quality**, not stopping rules or turn count
- Deliberation over 5 candidates per turn makes NPC responses **actually informative**

Why it works

AllSuspects/Ensemble: same bad questions, more of them.

DQS: **better questions** → NPCs reveal discriminative information → model can identify the guilty suspect.

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: Deliberation Cannot Fix Reasoning

Method	Acc	Turns	Calls
GN Baseline	0%	25.0	500
GN + DQS	5%	24.7	7640
(Optimal)	*100%*	*5.5*	–

1 out of 20 correct. 15× the cost. 73 minutes.

DQS helps DC but not GN

DC: DQS picks better *questions* → NPCs give more useful answers → model can identify the suspect.

GN: DQS picks better *guesses* → feedback is already precise (bulls/cows) → but model **still can't use it**.

GN Deep Dive: DQS Doesn't Get Closer Either

Metric	Base	DQS
Avg best-bulls in play	**2.0**	1.9
Reached 3+ bulls	**5**	2
Reached 2+ bulls	16	17
Avg final-pred bulls	1.1	1.1

Per-puzzle comparison:

Who got closer?	n
Baseline closer	**5**
DQS closer	3
Tied	12

Example — Puzzle 0 (secret = 8362):

Baseline: Guessed 8367 at T14 (**3 bulls!**)
→ then guessed 2591, 9183...
→ threw away the near-miss.

DQS: Best guess was 0 bulls all game.
→ deliberation didn't help at all.

Conclusion
DQS improves the **input side** (which question/guess to make).

It cannot fix the **output side** (reasoning about feedback to narrow possibilities).

For GN, the output side is the bottleneck.

Key Findings

Week 5 Findings

1. **DQS doubles DC accuracy: 30% → 60%**
   The **only** method that improved over baseline. Same turns, better questions. "Ask smarter, not more."

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

3. **The bottleneck is question quality, not stopping rules**
   CIP-Lite at T1 (45%) beats all stopping-rule variants. But DQS at T10 (60%) beats CIP-Lite by asking better questions.

4. **GN: DQS can't help when model can't reason (0% → 5%)**
   DQS improves inputs (questions), not outputs (constraint tracking). 15× cost for +5pp. Fundamental model limitation.

Next Steps

- **DQS + CIP-Lite** — combine better questions with adaptive stopping
- **GN DQS** — can deliberation help constraint tracking? (running)
- **Stronger model** — does DQS benefit scale with model capability?
- **Larger DC evaluation** — 50+ puzzles; 60% CI is [40%, 80%]
- **DQS ablations** — vary M (candidates), measure cost-accuracy tradeoff

End