RRMC — Week 6

Scaling to 100 Puzzles & Cross-Model Validation

2026-02-13

Qwen 2.5-7B + Gemma 3 27B via OpenRouter

What Changed This Week

Scale & Models

- 20 → **100 puzzles** (all 3 tasks)
- Added **Gemma-3-27B** (cross-model)
- Tested Qwen3-Coder-30B-MoE on GN

New Methods

- **Zero-shot** baseline (0 turns)
- **MI-Only + min_turns** guard
- min_turns ablation (3, 5, 10)

Key Bug Fix

`_clone_stopping_rule()` in the parallel evaluator dropped `min_turns` when cloning MI stopping rules for threaded execution.

All MI experiments silently ran with `min_turns=0`. Detected via GN showing avg turns = 1.1.

Total: 19 experiment runs, 3 tasks, 2 models, 8 methods

Detective Cases (DC)

DC Results: Qwen 7B (100 puzzles)

Method	Accuracy	Avg Turns	Note
KnowNo	**41%**	1.0	always stops T1
Zero-shot (T=0)	38%	1.0	no questions asked
MI-Only (min3)	38%	9.1	= zero-shot at 9× turns
CIP-Lite	37%	1.3	stops early
Self-Consistency	34%	1.5	
DQS + Fixed(10)	29%	10.0	collapsed from 60% at n=20
Fixed Turns (10)	25%	10.0	questioning hurts
DQS + MI (min3)	24%	8.9	worst overall

DQS Collapse
n=20: 60% → n=100: 29%.

The +30pp story was **lucky variance**.

Zero-shot is the baseline
Zero-shot (38%) > Fixed(10) (25%).

Reading the case is better than investigating.

DC: Why Questioning Hurts — Puzzle 5

Ground truth: **Suspect D** — Zero-shot: **D** ✓ Fixed(10): **E** ✗

What happened during 10 turns of questioning:

T1: Clara Whitfield: "I was at my usual cafe..."
T4: Victoria Blackwood: "I was in my hotel room..."
T5: Eleanor Grey: "I was in the library, locking up."
T9: Blackwood: "Our disagreements were mostly academic."

Every suspect gave **plausible alibis** and **denied conflicts**. The model was confused by uniform denials → **switched** from correct D to wrong E.

The Pattern (24 puzzles)

Model starts with correct intuition from case description.

Asks uninformative questions → generic alibis → model **second-guesses itself** → switches to wrong answer.

Disagreement Counts

Zero-shot ✓, Fixed(10) ✗	**24**
Zero-shot ✗, Fixed(10) ✓	11
Both correct	14
Both wrong	51

Net: −13 puzzles from questioning.

DC: A Directly Answerable Case — Puzzle 17

Victim: **Jonathan Blake** (architect) Weapon: **blood-stained sculpture piece** Zero-shot: **A** ✓
Fixed(10): **E** ✗

Case description already answers it:

- **A:** **Eleanor Whitaker** — artist known for **avant-garde sculptures**, "fiercely competitive, public disagreements"
- B: Sophia Carter — art curator
- C: Michael Turner — real estate developer
- D: Clara Mitchell — landscape architect
- E: Oliver Bennett — park worker (found body)

Murder weapon = **sculpture**. Only one suspect makes sculptures. Zero-shot sees this immediately.

What 10 turns of questioning did:

- T1: Whitaker: "interactions were always tense"
- T4: Whitaker: "sculpture was central to my installation"
- T8: Bennett: "someone had been fiddling with it"
- T9: Whitaker: "blood-stained when Jonathan accidentally..."

Questioning Obscured the Signal
Whitaker deflected ("accidentally"). Bennett's proximity to body → looked suspicious. Model switched A → E.

Solvable from the description alone.

DC: When Questioning Helps — Puzzle 3

Ground truth: **E / Alex Turner** — Zero-shot: **D ✗** Fixed(10): **E ✓**

Key turns:

T5: Alex Turner: "I was setting up for a freelance project meeting, but it didn't go as planned."

T9: Turner: "Yes, we had some disagreements about the direction of the project..."

Meanwhile, all other suspects denied conflicts:

- Reynolds: "No, I had no disputes"
- Davis: "No, I don't think so"
- Mitchell: "No, I did not have any disputes"

Turner's admission stood out → model corrected to E.

Success Pattern (11 puzzles)

The guilty suspect **admits to conflicts** or suspicious behavior while innocents uniformly deny.

This contrast gives the model a clear signal — but it only works **11 out of 100 times**.

Takeaway

Questioning helps only when NPCs **differentially leak guilt**.

Most of the time, all suspects give the same generic alibis → noise.

DC Cross-Model: Gemma 27B (100 puzzles)

Method	Gemma-3-27B	Qwen-2.5-7B	Δ
Zero-shot	**57%**	38%	+19pp
Fixed(10)	40%	25%	+15pp
DQS+Fixed(10)	30%	29%	+1pp

Model Scale Dominates

Gemma-27B zero-shot (**57%**) gains more from scale (+19pp) than any method ever achieved.

No stopping rule or DQS comes close.

Questioning Hurts More with 27B

27B: 57% → 40% (−17pp)

7B: 38% → 25% (−13pp)

DQS hurts 27B further: 40% → 30% (−10pp).

Task-inherent, not model-specific.

DC: Why DQS Hurts the Stronger Model — Puzzle 40

Ground truth: **E** — Zero-shot: **E ✓** Fixed(10): **E ✓** DQS: **D ✗**

Fixed(10) — simple, focused questions:
"What were you doing between 9pm and 10pm?"

Systematically covered each suspect. Answers naturally revealed useful patterns.

DQS — elaborate, leading questions:
"Given your reputation for dealing in valuable art and your close ties to Eleanor, can you provide any insight into why someone would want to harm her. . . ?"

Suspects gave longer, more **evasive** non-answers.

DQS Failure Modes

1. **Leading questions** telegraph suspicion → evasive answers
2. **Fixation**: DQS asked one suspect 4/9 turns, creating an echo chamber
3. **Complexity**: compound questions generate compound non-answers

10 "easy" puzzles broken

Both zero-shot AND Fixed(10) correct, but DQS still got wrong.

DQS introduces noise even on easy puzzles.

Situation Puzzles (SP)

SP Results: Qwen 7B (100 puzzles)

Method	Bool Acc	Avg F1 (char)	Avg Turns	Note
MI-Only (min3)	99%	**0.799**	3.0	best F1, $p < 0.001$
Zero-shot	100%	0.719	1.0	baseline
CIP-Lite	100%	0.718	1.5	$\approx$ zero-shot
Fixed Turns (10)	100%	0.716	10.0	10 turns, no F1 gain

MI's One Clean Win

F1: 0.72 $\rightarrow$ **0.80** ($+0.08$, paired t-test $p < 0.001$).

MI-Only(min3) outperforms on **69/100** puzzles. The only task where MI-based stopping genuinely helps.

Boolean Acc is Uninformative

All methods score 99–100%.

Character-level F1 is the **only** metric that differentiates methods.

Generic questions don't help: Fixed(10) $\approx$ zero-shot.

SP: Why MI Works — Puzzle 53

Puzzle: "Sarah is seen in two places at once. How?"

Ground truth: *Sarah has an identical twin who covers for her.*

Zero-shot F1: 0.168 — MI-Only F1: **0.333** (+0.165)

MI-Only Q&A (stopped at turn 4):

T1: Q: Technology to appear in two places?
A: **No** (MI=0.45 — high uncertainty)

T2: Q: Remote working setup?
A: **No** (MI=0.0)

T3: Q: Does she have a **double**?
A: **Yes** (MI=0.45 — still uncertain)

T4: *(stopped — MI=0.0, converged)*

Prediction: "Sarah has a **twin sister**..."

Zero-shot (no questions):

"Sarah might be using **virtual reality** or advanced telepresence technology..."

Why It Works

Yes/no feedback is **ground-truth** — not LLM-generated noise.

T1–T2 **eliminated** technology hypotheses.

T3 **confirmed** the "double" direction.

Systematic elimination → correct answer.

SP: When MI Hurts — Puzzle 48

Puzzle: "John has never left his hometown, yet seen in two countries. How?" Ground truth: *Identical twin + video calls.*

Zero-shot F1: **0.377** — MI-Only F1: 0.216 (−0.161)

MI-Only Q&A:

T1: Q: Does John use technology to appear in both?
A: **No** (misleading — video calls ARE technology)

T2: Q: Did John travel at different times?
A: **No**

Model **ruled out technology** based on misleading "No" → guessed "conspiracy or hoax."

Zero-shot: guessed "hologram / projection" — closer to video calls → higher F1.

Failure Mode

SP uses an LLM referee for yes/no answers.

Referee answered "No" to a technology question when the ground truth **involves** technology (video calls).

Misleading feedback **poisoned** the reasoning — model ruled out the correct direction.

Net Result

MI better on **69** puzzles, worse on **29**.
Referee quality limits the ceiling.

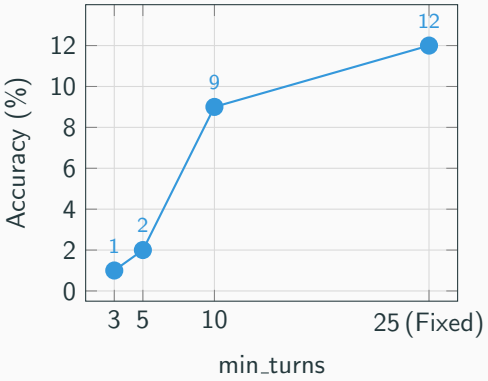
Guessing Numbers (GN)

GN Results: Gemma 27B (100 puzzles)

Method	Accuracy	Avg Turns	Note
Algorithmic DQS	*100%*	*5.4*	*not LLM*
Fixed(25)	**12%**	23.8	best LLM method
MI-Only (min10)	9%	17.5	
DQS + Fixed(25)	4%	24.5	DQS hurts again
MI-Only (min5)	2%	8.8	stops too early
MI-Only (min3)	1%	6.1	way too early
Self-Consistency	0%	2.8	immediate wrong stop
Qwen 7B (any method)	*0%*	—	*model too weak*
Qwen3-Coder-30B-MoE	*0%*	—	*3B active = too weak*

First non-zero LLM result on GN. Dense 27B breaks the 0% barrier. Model needs **maximum turns** — any early stopping hurts.

GN: min_turns Ablation



Clear Monotonic Trend
More investigation turns = higher accuracy.

min_turns	Acc	Avg Turns
3	1%	6.1
5	2%	8.8
10	9%	17.5
25 (Fixed)	**12%**	23.8

MI Stopping is Wrong for GN
MI detects "convergence" but the model converges on **wrong answers**.

GN needs all available evidence before answering.
Early stopping = premature commitment.

GN: What Gemma Does Right — Puzzle 51 (Solved in 4 turns)

Secret: **5670**

Turn	Guess	Feedback	Reasoning
1	1234	0B 0C	eliminates 1,2,3,4
2	5678	3B 0C	3 of {5,6,7,8} in position
3	5679	3B 0C	swapped 8→9: same 3B ⇒ 8 *and* 9 not at pos 4 ⇒ pos 4 must be 0
4	**5670**	**4B**	**Correct!**

Perfect deduction: T1 eliminates half the digits. T2 finds 3 bulls. T3 tests the remaining position. T4 fills in the only possibility.

Solve Distribution (12/100)

Category	Count
Early (<15 turns)	6
Late (15–25 turns)	6
Fastest	4 turns (idx=51)
Slowest	24 turns (idx=93)

Even split between fast solvers (clean deduction) and slow solvers (stumble into the answer).

GN: The "Last Mile" Problem — Puzzle 56 (3 Bulls, 5 Times, Never Solved)

Secret: **6728** — Result: **Wrong** (predicted 6718)

Turn	Guess	Result	
14	6723	**3B** 0C	pos 1,2,3 correct!
15	6745	2B 0C	*abandoned pos 3!*
18	6738	3B 0C	found 67_8
22	6729	3B 0C	tried 9 at pos 4
23	6708	3B 0C	tried 0 at pos 3
24	6718	3B 0C	tried 1 at pos 3

Turn 14 had 672_ — only needed to try digits at position 4.

But the model **abandoned** the correct '2' at pos 3 and **never returned to it**.

No Constraint Memory

The model doesn't maintain a record of "which positions are proven correct."

After finding 67__, it tried 6738, 6729, 6708, 6718 — systematically **avoiding** the correct 672_ prefix.

Scale of the Problem

31 of 88 failed puzzles reached 3 bulls at some point but couldn't close.

Avg first 3-bull hit: turn 14.6.

The model **can** narrow to 3/4 digits but cannot **systematically** test the last one.

Cross-Task Summary

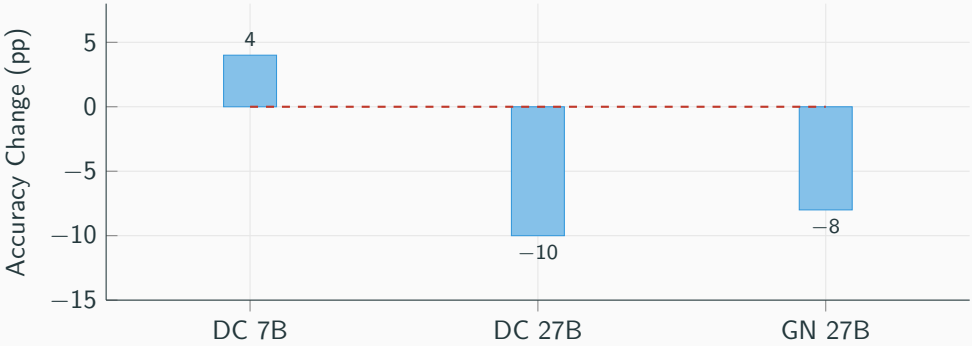
Cross-Task Summary

	DC (7B)	DC (27B)	SP (7B)	GN (27B)
Best LLM method	KnowNo 41%	Zero-shot **57%**	MI-Only **F1=0.80**	Fixed(25) 12%
Questioning helps?	No (−13pp)	No (−17pp)	Yes (+0.08)	Yes (+12pp)
DQS helps?	No	−10pp	—	−8pp
MI stopping helps?	No	—	Yes	No
Best strategy	Stop T1	Stop T1	Stop T3	Use all 25

The Critical Factor: Feedback Quality

DC: LLM-generated NPC responses = noise. **SP:** Ground-truth yes/no = signal. **GN:** Precise constraints = signal (but needs strong model to exploit).

DQS: A Universal Negative Result



Why DQS Fails

1. Leading questions → evasive answers
2. Fixation on one suspect/digit
3. Compound questions → compound noise

Implication
The bottleneck is **not** question quality.

DC: NPC responses are noise regardless of question.

GN: Model can't use feedback regardless of

Key Findings & Next Steps

Week 6 Findings

1. **DQS collapsed: 60% $\rightarrow$ 29% at** n=100 — Week 5 headline was a statistical artifact. DQS hurts across all tasks.
2. **Questioning hurts DC — confirmed across 7B and 27B** — Zero-shot $>$ Fixed(10) for both models. NPC responses are noise.
3. **MI-based stopping works on SP (F1: 0.72 $\rightarrow$ 0.80,** $p < 0.001$**)** — The one clean positive result. Ground-truth feedback enables systematic elimination.
4. **Gemma-27B breaks GN 0% barrier: 12% accuracy** — Dense 27B $>$ 3B-active MoE $>$ 7B. Model scale is the dominant factor.
5. **GN needs max turns — MI stopping is counterproductive** — min_turns ablation: 3$\rightarrow$5$\rightarrow$10$\rightarrow$25 maps to 1%$\rightarrow$2%$\rightarrow$9%$\rightarrow$12%.

Next Steps

- **DC with Gemma-27B + MI/KnowNo** — does MI add value when the base model is stronger (57% zero-shot)?
- **SP with Gemma-27B** — can the stronger model push F1 beyond 0.80?
- **GN with larger dense model** — Qwen3-Coder full (80B) or GPT-4o-mini. If 27B = 12%, 80B might reach 30%+.
- **GN error analysis** — 31/88 failures hit 3 bulls. Can we add a constraint-tracking scaffold to close the last digit?
- **Variance estimation** — run each method 3× to quantify run-to-run instability.

End