

NEURIPS 2026 · TRACK ON EVALUATIONS & DATASETS

Evaluating Whether LLMs Can Reliably Connect the DOTs?

Eftekhar Hossain · John Salvador · Santu Karmaker

Bridge-AI Lab, University of Central Florida

Project Page

A HYPOTHETICAL SCENARIO





Severe flooding

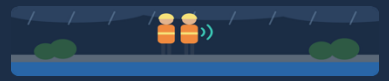
People waiting for help

Rescue team sent

The emergency coordinator

10:00 AM · Initial observation

Field team (radio)
10:00 AM
"On main road. Moving slowly."



Facebook post
9:58 AM
"Just drove out on Main Rd. Water near the edge."



Emergency call
9:57 AM
"Water coming up our street."



Weather service
9:55 AM
"River water may rise further."



THE UNKNOWN GAP

?

40 minutes, no contact.
What happened to the road?

10:40 AM · Final observation

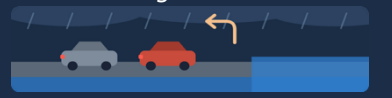
Fox News
10:35 AM
"Flood warning extended."



Field team (radio)
10:40 AM
"Heading to the school."



Twitter / X post
10:38 AM
"Cars turning back."



Anonymous call
10:36 AM
"No one can reach us."



THE AI'S ROLE

Before
After



LLM



Plausible event



OUR STUDY

How reliably can LLMs fill missing information while staying consistent with what comes before and after?

What could have happened in between?



A Gradual flooding

Plan: Wait for it to recede or send a boat.



B Bridge collapse

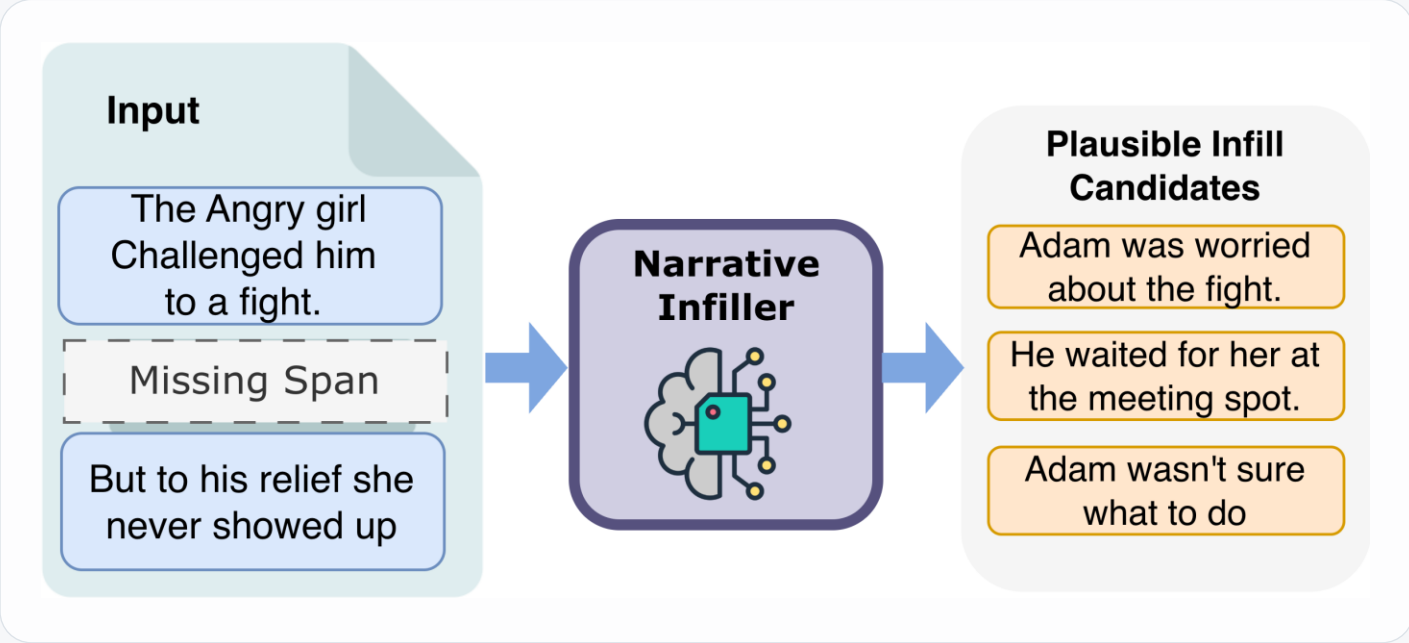
Plan: Reroute; this road may be gone for days.



C Debris blockage

Plan: Send a clearing crew with the rescue team.

Narrative infilling



The Benchmark

9,142
instances

4
narrative domains

1–3
sentences masked

20
open LLMs, 1.5B–70B

Instances per source



How we evaluate

Instruction detail

Level 0 · Level 1 · Level 2 · Level 3 · Level 4

From raw text (Level 0) to few-shot examples (Level 4)

Reasoning guidance

Abductive · Causal · Chain-of-Thought · Counterfactual

Each added on top of the Level 2 prompt

JUDGE · GPT-4.1-MINI

Infilling quality, scored 1–5

Fluency · Context Faithfulness · Bidirectional Coherence · Narrative Consistency · Informativeness

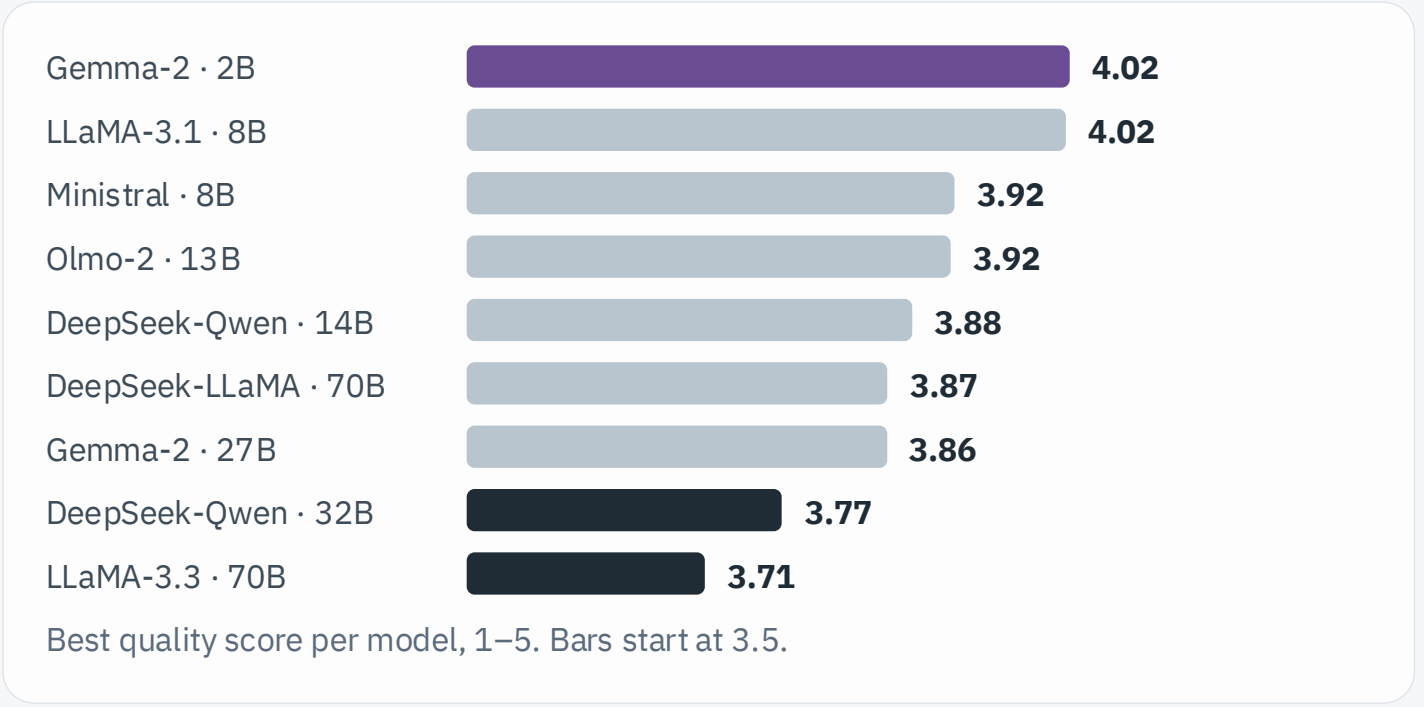
Validated on 300 human-annotated completions

0.65
human–human κ

0.58
GPT–human ρ

Finding 1

Scale doesn’t guarantee quality



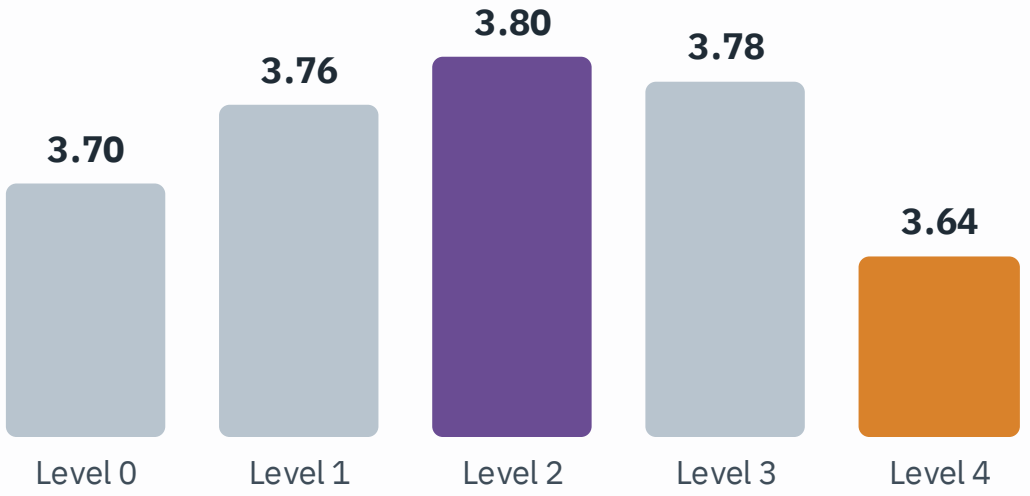
Finding 2

Level 2 is enough, and few-shot examples hurt

+3% quality from Level 0 to Level 2

+16.7% automatic scores from Level 0 to Level 2

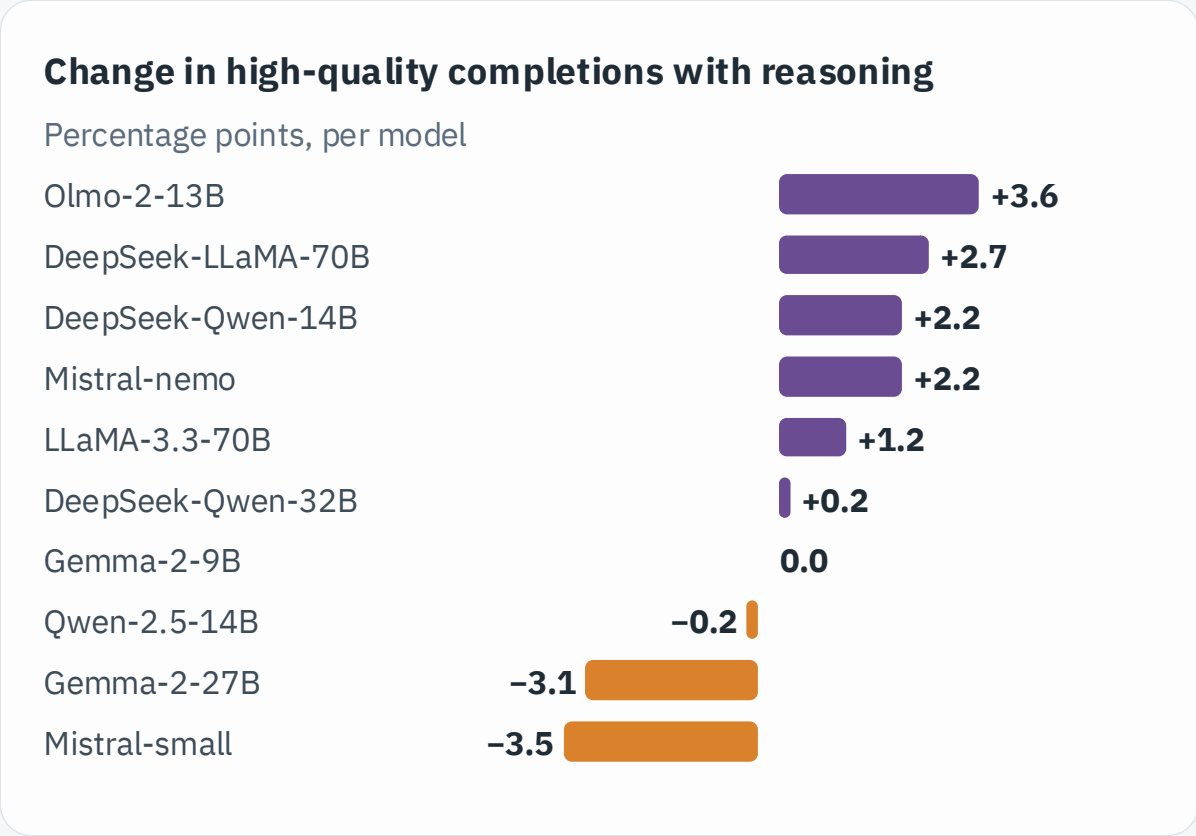
3.64 quality at Level 4, with few-shot examples



Average quality over all 20 models. Bars start at 3.5.

Finding 3

Reasoning adds little (+0.6% overall)



Finding 4

Short stories fail most

16.0%

low-score rate for 5-sentence stories, the worst case

<2%

low-score rate for stories of 20+ sentences

High-quality completions (%): standard → reasoning prompts

Domain	Opening	Middle	Closing
ROCStories	63.2 → 69.3	67.0 → 72.6	54.0 → 55.0
SIND	44.6 → 50.8	48.6 → 53.8	46.9 → 48.8
CNN/DailyMail	67.9 → 63.8	65.4 → 61.3	65.3 → 60.8
Wikipedia	57.1 → 52.4	60.5 → 57.3	60.2 → 56.3

Green rows improve with reasoning prompts. Orange rows get worse.

Reasoning helps story-like text but hurts factual text, where knowledge matters more than inference.

Takeaways

1. Infilling is far from solved: 3.80/5 on average, 4.02 at best.
2. Neither scale, detailed prompts, nor reasoning reliably help infilling.
3. Short stories and domain drive difficulty more than gap position.

Acknowledgements







UNIVERSITY OF CENTRAL FLORIDA

Thank you