CogniLoad

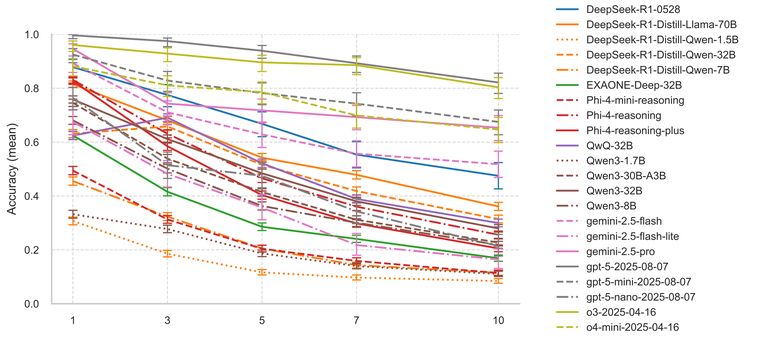
Long-context reasoning is not one challenge - it is three.

A Synthetic Natural Language Reasoning Benchmark With Tunable Length, Intrinsic Difficulty, and Distractor Density

Daniel Kaiser · Arnoldo Frigessi · Ali Ramezani-Kebrya · Benjamin Ricaud

Integreat – Norwegian Centre for knowledge-driven machine learning · UiT The Arctic University of Norway · University of Oslo

Intrinsic difficulty (d)



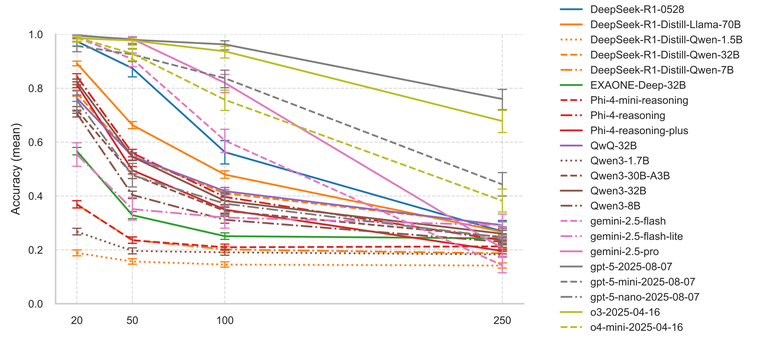
More interacting people, attributes, and clauses

CLT: intrinsic load

12 / 22 models drop below 50% by d=5

Difficulty hurts steadily for almost all models

Task length (N)



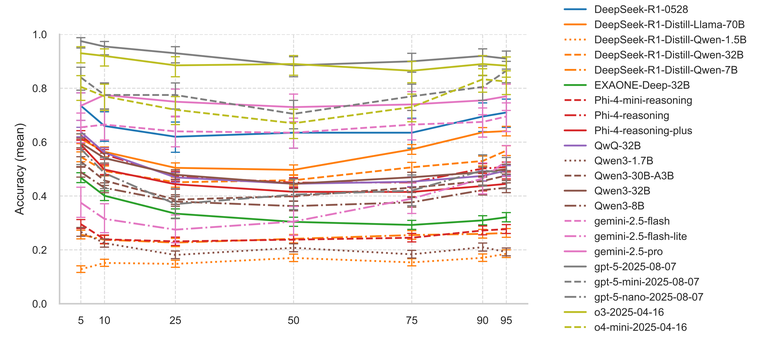
Longer puzzle with a longer state to track over time

CLT: Germane Load

Only gpt-5 and o3 stay above 50% at N = 250

Sequential state tracking is the sharpest bottleneck.

Needle-to-hay ratio (ρ)



More or fewer distractors relative to relevant statements

CLT: Extraneous Load

Many models are worst near ρ≈25%–50%.

Distractor sensitivity is U-shaped rather than monotonic.

Why CogniLoad ?

Benchmarks for long-context reasoning blur mix intrinsic reasoning difficulty, distractor interference, and sequential state tracking. When a model fails, the real cause is therefore ambiguous.

- Long-context suites stress context length

- Logic benchmarks stress intrinsic difficulty

- Needle-in-a-haystack tasks typically stress distractor filtering

CogniLoad controls **all three** in **one synthetic puzzle family** to make the evaluation of reasoning failures more **diagnostic, reproducible, and scalable**. The generation algorithm provides clear **gold states** and detailed **metadata** for each puzzle while allowing the creation of an **infinite amount** of randomized puzzles to **avoid test set poisoning.**

What does a puzzle look like?

(i) Puzzle Instruction: Solve this logic puzzle. You MUST finalize your response with a single sentence about the asked property (e.g., "Peter is in the livingroom.", "Peter is wearing blue socks",..). Solve the puzzle by reasoning through the statements in a strictly sequential order.

(ii) Initial State:

- Brent is wearing green socks and is wearing purple gloves and last listened to classical music.
- Anthony is wearing purple socks and is wearing yellow gloves and last listened to disco music.
- . . .

(iii) Update Statements:

1. The people wearing green socks listen to electronic music.
2. The people who last listened to classical music and wearing purple gloves put on yellow gloves.
3. . . .

(iv) Query: What color of socks is Brent wearing?

Cognitive Load Theory (CLT)

The three load dimensions are inspired by and grounded in Cognitive Load Theory (Sweller, 1988), an established theory from cognitive psychology on the factors influencing the capacity of working memory .

Cognitive Fingerprints

We evaluate **22 models** on **140 parameter configurations**, and up to **14,000 puzzles** per model. To compare models we fit a GLM on the beta-coefficients of the parameters and derive 50% capacity thresholds.

Model	β_0	β_d	β_S	β_ρ	β_{ρ^2}	ECL_{50}	NT_{50}	ID_{50}
gemini-2.5-pro	22.51^{***}	-0.41^{***}	-9.15^{***}	-1.67	1.76	153.3	−−	12.74
gemini-2.5-flash	18.14^{***}	-0.44^{***}	-7.56^{***}	-1.79	2.16	111.5	−−	8.56
gemini-2.5-flash-lite	3.19^{***}	-0.30^{***}	-1.22^{***}	-2.88^{**}	3.82^{***}	8.8	0.93	1.53
gpt-5-2025-08-07	17.34^{***}	-0.39^{***}	-5.11^{***}	-7.04^{***}	5.62^{***}	382.8	−−	14.78
gpt-5-mini-2025-08-07	11.09^{***}	-0.22^{***}	-3.96^{***}	-4.87^{***}	5.10^{***}	164.1	−−	11.72
gpt-5-nano-2025-08-07	6.50^{***}	-0.31^{***}	-2.43^{***}	-4.30^{***}	4.12^{***}	35.7	0.94	2.87
o3-2025-04-16	12.83^{***}	-0.22^{***}	-4.26^{***}	-2.72	2.07	356.9	−−	19.07
o4-mini-2025-04-16	13.00^{***}	-0.23^{***}	-4.89^{***}	-5.99^{***}	6.37^{***}	132.1	−−	10.86
DS-R1-0528	13.70^{***}	-0.39^{***}	-5.28^{***}	-4.13^{***}	4.19^{***}	104.6	−−	7.51
DS-Llama-70B	8.36^{***}	-0.30^{***}	-3.28^{***}	-3.50^{***}	3.92^{***}	69.8	0.53	5.14
DS-Qwen-32B	5.15^{***}	-0.19^{***}	-2.12^{***}	-2.07^{***}	2.29^{***}	54.3	0.78	3.95
DS-Qwen-7B	1.74^{***}	-0.23^{***}	-0.96^{***}	-0.45	0.58^{*}	2.9	−−	−0.53
DS-Qwen-1.5B	-0.35^{**}	-0.20^{***}	-0.33^{***}	0.47	-0.14	0.0	−−	−3.95
Phi-4-reasoning-plus	9.52^{***}	-0.45^{***}	-3.58^{***}	-4.21^{***}	3.41^{***}	45.7	0.16	3.68
Phi-4-reasoning	9.11^{***}	-0.39^{***}	-3.35^{***}	-4.62^{***}	3.99^{***}	52.3	0.92	4.08
Phi-4-mini-reasoning	1.70^{***}	-0.24^{***}	-0.81^{***}	-1.40^{***}	1.45^{***}	1.3	−−	−0.55
EXAONE-Deep-32B	4.09^{***}	-0.26^{***}	-1.55^{***}	-3.16^{***}	2.45^{***}	14.1	−−	1.0
QwQ-32B	5.68^{***}	-0.21^{***}	-2.08^{***}	-3.70^{***}	3.07^{***}	48.0	0.95	3.56
Qwen3-32B	7.22^{***}	-0.29^{***}	-2.75^{***}	-3.21^{***}	2.75^{***}	53.9	0.94	4.1
Qwen3-30B-A3B	5.83^{***}	-0.30^{***}	-2.25^{***}	-2.96^{***}	2.87^{***}	36.7	0.99	3.05
Qwen3-8B	5.40^{***}	-0.26^{***}	-2.19^{***}	-2.87^{***}	2.66^{***}	30.8	−−	2.2
Qwen3-1.7B	0.62^{***}	-0.17^{***}	-0.46^{***}	-1.53^{***}	1.24^{***}	0.0	−−	−4.07

Why is the U-shape interesting?

Increasing ρ creates two competing effects:

Different models balance these pressures differently, so distractor sensitivity is not monotonic.

Load dimensions compound

Difficulty and length are super-additive in 17/22 models: long, complex tasks degrade accuracy beyond what either factor alone predicts. Difficulty also amplifies distractor harm. Frontier models (gpt-5, o3) show no detectable interactions, consistent with near-separable processing.

What goes wrong?

State-tracking mistakes dominate and grow with d and N. Context-budget overflows are model-specific: Gemini-2.5 exhausts 32K tokens at N=250 due to verbosity, while gpt-5/o3 stay well within limits. Smaller models increasingly produce formatting drift or non-answers (e.g., "unsolvable") as load grows.

Supported by the Research Council of Norway through Integreat (332645), FRIPRO (356103), and computational resources provided through Sigma2 / LUMI (nn12027k)