COGNILOAD: A SYNTHETIC NATURAL LANGUAGE REASONING BENCHMARK WITH TUNABLE LENGTH, INTRINSIC DIFFICULTY, AND DISTRACTOR DENSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Current benchmarks for long-context reasoning in Large Language Models (LLMs) often blur critical factors like intrinsic task complexity, distractor interference, and task length. To enable more precise failure analysis, we introduce **CogniLoad**, a novel synthetic benchmark grounded in Cognitive Load Theory (CLT). CogniLoad generates natural-language logic puzzles with independently tunable parameters that reflect CLT's core dimensions: intrinsic difficulty (d) controls intrinsic load; distractor-to-signal ratio (ρ) regulates extraneous load; and task length (N) serves as an operational proxy for conditions demanding germane load. Evaluating 22 SotA reasoning LLMs, CogniLoad reveals distinct performance sensitivities, identifying task length as a dominant constraint and uncovering varied tolerances to intrinsic complexity and U-shaped responses to distractor ratios. By offering systematic, factorial control over these cognitive load dimensions, CogniLoad provides a reproducible, scalable, and diagnostically rich tool for dissecting LLM reasoning limitations and guiding future model development.

1 INTRODUCTION

Cognitive Load Theory (CLT) (Sweller, 1988) characterizes three types of cognitive load on human working memory when solving problems (Sweller, 1988; Paas et al., 2003; Lieder and Griffiths, 2020): intrinsic (ICL), extraneous (ECL), and germane (GCL). ICL stems from the inherent complexity and element interactivity of the task (Halford et al., 1998). ECL is induced by suboptimal task presentation requiring the processing of elements that are not task-relevant (Chandler and Sweller, 1991). GCL concerns effective remaining resources allocated to engaging with the intrinsic task demands for mental schema construction (Ericsson and Kintsch, 1995; Sweller, 2010).

Large language models (LLMs) demand analogous computational resources when solving reasoning tasks. The essential element interactivity of a reasoning chain mirrors ICL; distractor elements reflect ECL; and sustained engagement with intrinsically relevant information over a long reasoning process acts as an operational proxy for germane-like processing - the constructive effort to maintain a coherent problem representation.

To the best of our knowledge, no study has based the evaluation of problem-solving capacities of LLMs in CLT by distinguishing these three load types, and existing benchmarks often confound them: LongBench (Bai et al., 2024a) and L-Eval (An et al., 2024) vary context length but not necessarily the intrinsic reasoning depth; LogicBench (Parmar et al., 2024) probes ICL with minimal demands on ECL or context-induced load; BABILong (Kuratov et al., 2024) mixes multi-step reasoning with fixed distractor ratios, obscuring precise failure attribution.

We introduce **CogniLoad**, a controllable synthetic benchmark for long-context reasoning, inspired by CLT, that operationalizes these load types through tunable parameters in randomized natural-language logic puzzles: **(i) Intrinsic Load** via Intrinsic Difficulty d controls the number of interacting entities, attributes, and logical clauses, directly manipulating ICL by varying essential element interactivity and reasoning depth. **(ii) Extraneous Load** via Distractor Density ρl dictates distractor density; lower ρ increases irrelevant elements, manipulating ECL. **(iii) Germane Load Proxy** via Task Length N serves as an operational proxy for demanding germane-like cognitive work.

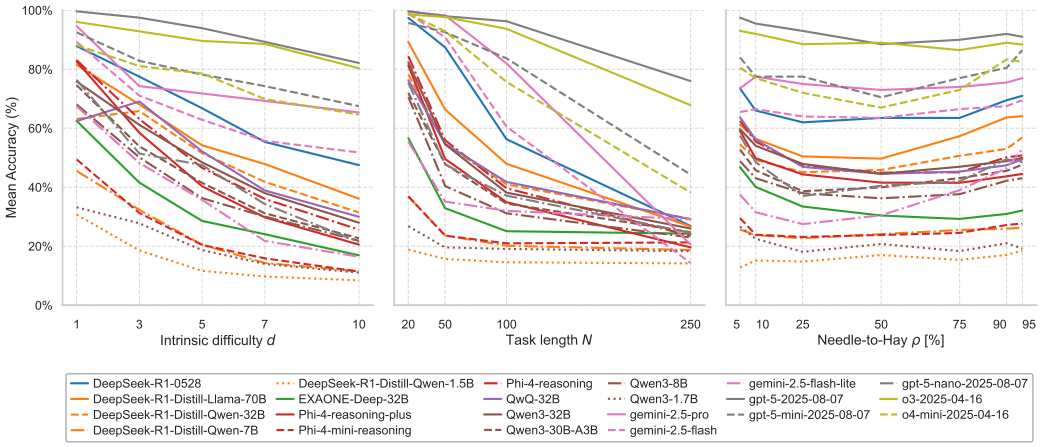

Figure 1: The average accuracy of models across the evaluated parameter space for $d \in \{1, 3, 5, 7, 10\}$ (left panel), $N \in \{20, 50, 100, 250\}$ (center panel), and $\rho \in \{5, ..., 95\}$ (right panel). Each plot selects one dimension for the X-axis and averages the accuracy of all evaluated puzzles for the other two dimensions relative to it.

Our key contributions are summarized as follows:

1. We ground the evaluation of LLMs in CLT, precisely defining benchmark parameters that control ICL, ECL, and an operational proxy for the conditions conducive to GCL.
2. We introduce *CogniLoad*, the first benchmark designed to independently control these three dimensions of cognitive load, while scaling to arbitrarily long contexts.
3. We provide an algorithm for the automatic randomized generation and evaluation of puzzle instances, enabling large-scale and reproducible comparison of LLM capabilities.
4. We report empirical results on 22 state-of-the-art (SotA) reasoning LLMs (see Figure 1), revealing distinct failure regimes across the (d, N, ρ) dimensions and highlighting specific targets for improving LLM design.

Together, these contributions translate CLT into a precise diagnostic framework for understanding and advancing long-context reasoning in LLMs.

1.1 Related Work

Long-context Benchmarks (Working Memory Capacity). A line of work starting with Long-Range Arena (LRA) (Tay et al., 2020) and followed by several recent benchmarks probe LLM performance on long sequences, often framed as testing "memory load" or context utilization. Earlier studies such as SCROLLS (Shaham et al., 2022), BookSum (Kryściński et al., 2021), and QMSum (Zhong et al., 2021) scale document length without manipulating intrinsic difficulty. LongBench (Bai et al., 2024a;b) and L-Eval (An et al., 2024) aggregate multi-task corpora up to 200k tokens, while BABILong (Kuratov et al., 2024), LongReason (Ling et al., 2025), RULER (Hsieh et al., 2024), ZeroSCROLLS (Shaham et al., 2023), and Michelangelo (Vodrahalli et al., 2024) increase context while the inherent difficulty of individual sub-tasks (ICL) may vary unsystematically and distractor density (ECL) is often not a controlled variable. Consequently, performance degradation could be due to sheer length overwhelming processing capacity, or an inability to sustain germane-like cognitive work over extended relevant information, but the precise cause of failure is not clear.

Logical-reasoning Benchmarks (Intrinsic Load). A complementary line of benchmarks focuses on ICL by presenting tasks with high inherent complexity but often within minimal context lengths or distractors. Notable classical suites include ReClor (Yu et al., 2020), LogiQA (Liu et al., 2020), and BIG-Bench-Hard (BBH) (Suzgun et al., 2022). AutoLogic (Zhu et al., 2025) is a benchmark that explicitly focuses on scaling ICL through controllable complexity. LogicBench (Parmar et al., 2024),

CLUTRR (Sinha et al., 2019), and ZebraLogic (Lin et al., 2025) also exemplify this by formulating symbolic logic puzzles that demand processing many interacting elements (e.g., multi-step deductions, handling negation, and constraint satisfaction). Similarly, mathematical reasoning datasets, e.g., GSM8K (Cobbe et al., 2021) and abstract rule induction tasks, e.g., ARC-AGI (Chollet et al., 2024) primarily escalate ICL by increasing the complexity of essential rules and their interdependencies.

Needle in a Haystack Benchmarks (Extraneous Load). Needle in a haystack (NIAH) designs (Gkamradt, 2023) specifically target ECL by embedding relevant facts ("needles") within large volumes of distractor text ("hay"). Variants such as Sequential NIAH (Yu et al., 2025) and Nolima (Modarressi et al., 2025) investigate the impact of such distractors, which constitute non-essential elements requiring processing for filtering, thereby imposing ECL. While these benchmarks effectively isolate the impact of distractors on information retrieval, the "needle" tasks themselves typically involve low ICL (e.g., simple fact lookup).

Need for Multi-dimensional Evaluation. CLT highlights the interplay of ICL, ECL, and germane processing under finite working memory (Paas et al., 2003). Existing LLM reasoning benchmarks, however, typically manipulate only one dimension without systematic, independent control over the others. Even benchmarks like MIR-Bench (Yan et al., 2025) which combine high ICL with extensive input, do not offer the factorial control needed to disentangle these loads, hindering precise diagnostics. Similar to our work, GSM-∞ (Zhou et al., 2025) allows manipulating noise and difficulty. However, these parameters are not adjusted independently of task length.

Contribution of CogniLoad. CogniLoad addresses this critical gap by providing a framework for independently controlling parameters that influence: (i) ICL via intrinsic puzzle difficulty (d), (ii) ECL via distractor density (ρ), and (iii) the demands for sustained, germane-like processing via task length (N), all within a single synthetically generated natural language puzzle. This factorial design enables a precise diagnosis of LLM failure modes, specifically the inability to handle increased intrinsic complexity, susceptibility to extraneous distractors, or incapacity to maintain coherent reasoning over an extended number of sequences. By explicitly grounding these dimensions in CLT, CogniLoad offers the first benchmark to diagnostically map LLM capability surfaces across these distinct cognitive demands, thereby complementing and extending the insights from evaluations that focus on a single factor.

2 Benchmark Design: CogniLoad Logic Puzzles

2.1 Puzzle Definition and Construction

CogniLoad is a family of natural-language logic-grid puzzles explicitly crafted to probe sequential reasoning capabilities of LLMs. The design goals are threefold: each puzzle (i) necessitates sequential multi-step deduction where order fundamentally matters; (ii) embeds a controllable number of relevant "needle" facts within the context of a controllable number of "hay" distractor statements; and (iii) provides parameters that control distinct dimensions of cognitive load. This section formalizes the task and describes the puzzle generation process, the control parameters, and key design choices.

Each puzzle in CogniLoad (see Figure 2) consists of a set of people with independent and mutable attributes. A series of statements, applied in strictly sequential order, updates these attributes according to conditions specified in each statement. The puzzle generation is parameterized by three key parameters: intrinsic difficulty d, total number of statements N, and needle-to-hay ratio ρ.

2.1.1 Basic Puzzle Construction

A puzzle is formally characterized by the following components:

- **People**: A set $P = \{p_1, p_2, \ldots, p_n\}$ of persons in the puzzle, and $n = \max(d, 2)$.
- **Person of Interest (PoI)**: A randomly selected person $p^* \in P$ about whom the final question is asked.
- **Attribute Categories**: A set $A = \{c_1, c_2, \ldots, c_d\}$ of attributes randomly selected from a predefined taxonomy of 12 categories. Each category takes values in a Value Domain with a given finite cardinality, smaller or equal to 10.

(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?

Figure 2: Example CogniLoad puzzle with intrinsic difficulty $d = 3$, statements $N = 20$, and needle-to-hay ratio $\rho = 50\%$. Only a subset of the initial state and update statements is shown.

- **Value Domains**: For each category $c \in A$, a value domain $V_c = \{v_{c,1}, v_{c,2}, \ldots, v_{c,\ell_c}\}$ where $\ell_c = d + 1$ for $d > 1$ or $\ell_c = 3$ when $d = 1$. See Appendix F for the complete ontology.
- **State Function**: $S_t(p, c)$ represents the value of attribute c for person p at step t. Each person has values for the d attribute of the selected attribute categories A, thus the state value represents a vector of dimension d.

2.1.2 Puzzle Initialization

A puzzle starts with initialization statements ($t = 0$) that assign unique attribute values to each person: $\forall p \in P, \forall c \in A : S_0(p, c) \in V_c$ such that $\forall p_i, p_j \in P, i \neq j, \exists c \in A : S_0(p_i, c) \neq S_0(p_j, c)$.

2.1.3 Statement Generation Process

For each step t from 1 to N, a statement is generated that changes the state of a person. If it updates the PoI, the statement is called a *needle*. An update for a non-PoI is called a *hay*.

1. **Statement Type Selection**: Given N and ρ, let n^t_{needle} and n^t_{hay} be the remaining numbers of needles and hays to satisfy the desired proportion ρ in the complete puzzle. The probability of selecting a needle statement is then $\mathbb{P}(T_t = \text{needle}) = n^t_{\text{needle}}/(N - t)$. The total number of needle statements in the puzzle is calculated as $n^0_{\text{needle}} = \max(1, \min(N, \text{round}(N \cdot \rho/100)))$.

2. **Reference Person Selection**: Given the selected statement type T_t, the algorithm selects the reference person r_t: if $T_t = \text{needle} \implies r_t = p^*$ and if $T_t = \text{hay} \implies r_t \sim \text{Uniform}(P \setminus \{p^*\})$.

3. **Statement Structure**: For each statement, CogniLoad samples a number of conditions $k_t \sim \text{Uniform}\{1, \ldots, d\}$, and a number of state updates $m_t \sim \text{Uniform}\{1, \ldots, d\}$ and uniformly sample attribute categories $C_t \subseteq A,\ |C_t| = k_t$ and state updates $U_t \subseteq A,\ |U_t| = m_t$.

4. **Condition and Update Value Specification**: For each category $c \in C_t$, the condition value is set by the reference person's current state: $v_{c,t} = S_{t-1}(r_t, c)$. For needles, these conditions target the PoI, while for hays the conditions can match multiple people. For update values, if $T_t = \text{needle} \implies u_{c,t} \sim \text{Uniform}(V_c)$ and if $T_t = \text{hay} \implies u_{c,t} \sim \text{Uniform}(V_c \setminus \{S_{t-1}(p^*, c)\})$.

5. **Logical Form**: The statement at step t has the logical form:

$$\forall p \in P : \Big(\bigwedge_{c \in C_t} S_{t-1}(p, c) = v_{c,t} \Big) \Rightarrow \Big(\bigwedge_{c \in U_t} S_t(p, c) = u_{c,t} \Big).$$

Attributes not mentioned in the update set remain unchanged $\forall p \in P, \forall c \in A \setminus U_t : S_t(p, c) = S_{t-1}(p, c)$. This is not specified in the prompt but implicitly assumed by the LLMs.

2.1.4 Validation Constraints

A sequence of validations verifies that the generated statement does not result in a state that prevents the generation of further needles and hays. If all validations pass, the statement is appended to the puzzle; otherwise a new statement is generated.

For hay statements ($r_t \neq p^*$): After the update, the state of affected non-PoIs must not become identical to PoI $\forall p \in P \setminus \{p^*\}$ such that $\forall c \in C_t : S_{t-1}(p, c) = v_{c,t}$, $\exists c \in A : S_t(p, c) \neq S_t(p^*, c)$ and the update must not affect the PoI $\exists c \in C_t : S_{t-1}(p^*, c) \neq v_{c,t}$.

For needle statements ($r_t = p^*$): The update must not affect all non-PoI people $\exists p \in P \setminus \{p^*\} : \exists c \in C_t : S_{t-1}(p, c) \neq v_{c,t}$ and after the update not all non-PoIs have identical states as the PoI $\exists p \in P \setminus \{p^*\} : \exists c \in A : S_t(p, c) \neq S_t(p^*, c)$.

To prevent the distractors from becoming too trivial to track at lower difficulties, we require that a hay statement does not result in all non-PoIs become identical so the set $P \setminus \{p^*\}$ must contain at least two persons with distinct attribute values. CogniLoad construction ensures that each hay statement $T_t =$ hay affects at least one non-PoI $\exists p \in P \setminus \{p^*\} : \forall c \in C_t : S_{t-1}(p, c) = v_{c,t}$.

2.1.5 Final Question Generation

After all N statements have been generated, the puzzle concludes with a question about a random attribute of the PoI, sampled as a random category $c_q \sim \text{Uniform}(A)$. The correct answer to the puzzle is $S_N(p^*, c_q)$ obtained from the final state of the PoI.

2.1.6 Evaluation metrics

We evaluate each puzzle by exact-matching of the queried attribute value in the output of model M with accommodating minor phrasing and common lexical variants. See Appendices C and D for the specifics of the evaluation pipeline and an overview of the granular failure types that contextualize model specific error modes. The accuracy of model M across the evaluation set is calculated as $\text{acc}(M) = \frac{1}{|Z|} \sum_{z \in Z} \mathbf{1}\big[\text{answer}_M(z) = S_N(p^*, c_q)\big]$ where $S_N(p^*, c_q)$ represents the final state value of the queried attribute c_q for the PoI p^* after all N statements have been processed.

2.2 Tunable Parameters

To systematically probe long-context reasoning, CogniLoad employs three independent parameters. These parameters are designed to operationalize distinct cognitive load dimensions as defined by CLT (Paas et al., 2003), allowing the creation of puzzles with varying characteristics. Together, they define the load profile of a puzzle instance.

Intrinsic Difficulty (d) for $d \in \{1, 3, 5, 7, 10\}$ controls multiple facets of puzzle complexity (see Table 1), directly manipulating ICL which according to CLT hinges on element interactivity (Halford et al., 1998). Increasing d increases ICL via: (i) combinatorial growth in state space ($\approx (d+1)^d$), (ii) increased interactivity between persons, attributes, and values, and (iii) increased rule complexity (up to d conditions/updates per statement).

Task Length (N) for $N \in \{20, 50, 100, 250\}$ sets the total number of sequential state-update statements. While directly determining sequence length, N serves as an operational proxy for conditions demanding GCL. Increasing N, particularly with a large d and ρ, compels deeper reasoning through more essential interacting elements (Sweller, 2010). Additionally, increasing N necessitates the maintenance of a coherent (stateful) problem representation over a longer term with the construction of an efficient schema for it (Ericsson and Kintsch, 1995).

Needle-to-hay Ratio (ρ) for $\rho \in \{5, ..., 95\}$ sets the percentage of PoI-relevant ("needle") versus distractor ("hay") statements, directly manipulating ECL. ECL arises from processing non-essential elements (Chandler and Sweller, 1991). Decreasing ρ increases ECL via increased distractor density which challenges filtering. Increasing ρ controls ECL by focusing resources on relevant information. Critically, CogniLoad's "hay" statements are syntactically similar to "needles" and involve valid state updates for non-PoIs, imposing a more challenging ECL than easy to distinguish distractor text.

3 Results

We have evaluated the performance of 13 open weights LLMs on 100 random CogniLoad puzzles per (d, N, ρ) configuration resulting in 14'000 puzzle instances per LLM in total. In addition, we have

Table 1: Key parameters controlling the puzzle generation.

Symbol	Name	Definition	Cognitive Load Affected
d	Intrinsic Difficulty	Controls cardinality of people set $\|P\| = \max(d, 2)$, attribute categories $\|A\| = d$, for each category $c \in A$ the cardinality of value domains $\|V_c\| = \max(d + 1, 3)$, and the distribution of conditions and updates per statement: $k, m \sim$ Uniform$\{1, ..., d\}$.	ICL: Element interactivity, state space/rule complexity.
N	Task Length	Total number of sequential state transitions in the puzzle.	GCL Proxy / Task Length: Demands sustained engagement with core elements.
ρ	Needle-to-hay Ratio	Percentage of statements directly influencing the PoI (needles) versus distractor statements (hay)	ECL: Distractor density challenges filtering, selective attention, and imposing load from processing non-essential elements.

evaluated the proprietary Gemini-2.5 and gpt-5[1] models and DeepSeek-R1-0528 on 10 CogniLoad puzzles per configuration (i.e., 1'400 puzzle instances). The maximum context length (input + output) was set to 32K tokens and LLMs run with their preset default decoding settings and system prompts.

Figure 1 shows mean accuracy across models as each load dimension varies with trends corroborated by our regression analysis (Section 3.1).

Intrinsic Difficulty (d) Performance declines monotonically with d for most models, although a few mid-tier models show small bumps at $d = 3$ (e.g., QwQ-32B: 0.62→0.69; DS-Qwen-32B: 0.63→0.66). Top models degrade only slightly from $d = 1$ to $d = 3$ (o3: 0.96→0.93; gpt-5: 1.00→0.97), while smaller or distilled models drop by 0.10-0.25 in the same range. By $d = 5$, 12 of 22 models fall below 50% accuracy. Beyond $d \geq 7$ the marginal decline flattens for the majority of models: the strongest models maintain their performance even at $d = 10$ (gpt-5: 0.82; o3: 0.80), and the weakest ones approach 0.10-0.15.

Task Length (N) This parameter remains the dominant stressor. Most models exhibit their steepest decline between $N = 20$ and $N = 50$ (e.g., DS-Llama-70B: 0.89→0.66; Qwen3-8B: 0.71→0.40), while the best performing ones show relative resilience (gpt-5: 1.00→0.98; o3: 0.99→0.98). Accuracy declines with longer sequences as at $N = 100$, several models roughly halve their $N = 20$ accuracy (e.g., DS-Llama-70B: 0.48; Qwen3-32B: 0.38) and at $N = 250$ only two models show above 50% accuracy (i.e., gpt-5 at 0.76 and o3 at 0.68) while the majority ones perform 0.20-0.30 accuracy.

Extraneous Load / Needle-to-hay Ratio (ρ) A characteristic U-shaped response is typical with performance usually reaches as low as $\rho \in [25, 50]\%$ and recovers as ρ increases. Recovered performance equals or exceeds that of small-ρ baseline in several cases (DS-Llama-70B: 0.61→0.64; gpt-5-mini: 0.84→0.86; gemini-2.5-flash-lite: 0.38→0.53). The strongest models show smooth variations (gpt-5: 0.97→0.89→0.91; o3: 0.93→0.89→0.88), indicating marginal sensitivity to distraction, while some models recover only partially or not at all (Phi-4-reasoning-plus: 0.59→0.45; EXAONE-Deep-32B: 0.47→0.32).

3.1 Load-sensitivity Regression

To quantify model-specific sensitivities of the accuracy to load dimensions and derive interpretable capacity thresholds for each model, we employ a regression-based approach that allows us to isolate the impact of each type of cognitive load (see Table 2).

[1] The gpt-5 family models were evaluated using the "medium" reasoning effort setting.

Table 2: Per-model quadratic-ρ GLM estimates with Wald z statistic for p-values alongside derived 50% load-capacity thresholds (see Section 3.1.3). The value $--$ for NT_{50} indicates that no real root exists in $[0, 1]$. "DS" abbreviates "DeepSeek" in the model names. $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$

Model	β_0	β_d	β_N	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

3.1.1 Regression Model Specification

We model the performance of LLMs using a binomial generalized linear model (GLM) with a logit link function:

$$\Pr\big(Y{=}1\big) = \sigma\big(\beta_0 + \beta_d\, d + \beta_N\, \log_{10} N + \beta_\rho\, \rho + \beta_{\rho^2}\, \rho^2\big),$$

where the binary outcome Y represents exact-match accuracy ($Y = 1$, when the model solves the puzzle correctly), $\sigma(\cdot)$ is the inverse logit function, and the coefficients β_d, β_N and β_ρ quantify sensitivity to intrinsic difficulty (ICL), task length (GCL), and distractor ratios (ECL), respectively. The inclusion of a quadratic term for ρ is motivated by the characteristic U-shape observed in the third panel of Figure 1 and based on an improved Akaike Information Criterion (AIC) value for 18 out of the 22 fitted models when included (see Appendix E). Since N ranges up to 250, we apply $\log_{10}$ to keep it at a similar scale as the other parameters of the regression.

3.1.2 Significance of Main Effects

In all models, β_d and β_N are significant and highly negative, confirming performance degradation with increased ICL and GCL. The quadratic term for ρ is also significant (except for two models) confirming the U-shaped response for most models: models typically perform worst at intermediate ρ values and recover as ρ approaches either extreme. Five models exhibit statistically insignificant coefficients for ρ terms, reflecting poor baseline performance for the smallest models and indifference to distraction for strong models (i.e., o3, gemini-2.5-pro, gemini-2.5-flash).

3.1.3 Capacity Points at 50% Accuracy

The GLM coefficients (Table 2) allow us to derive interpretable capacity thresholds. These represent the point at which a model's accuracy is predicted to drop to 50% when varying a single load parameter, while holding other load parameters at their estimated mean values:

ECL_{50} (Effective Context Length): Maximum number of statements a model can process while maintaining 50% accuracy. Large ECL_{50} values indicate superior context handling.

NT_{50} (Needle-to-hay Threshold): Minimum proportion of relevant information required to maintain 50% accuracy. Crucially, *small* values indicate greater robustness to distractors. If the estimated

NT_{50} is missing, then the model accuracy is not expected to cross the 50% threshold for any value $0 \leq \rho \leq 1$, under mean conditions for d and N.

ID_{50} (Intrinsic Difficulty): It is the maximum intrinsic complexity (number of interacting entities/attributes) that a model can handle while maintaining 50% accuracy. Negative values indicate failure to reach 50% accuracy even at the lowest difficulty setting under mean conditions for N and ρ.

Mathematically, these thresholds are derived by setting the logit in the GLM equation to zero (for $\Pr(y = 1) = 0.5$) and solving for the parameter of interest, e.g.:

$$\mathrm{ECL}_{50} = 10^{-(\beta_0 + \beta_d \bar{d} + \beta_\rho \bar{\rho} + \beta_{\rho^2} \bar{\rho}^2)/\beta_N}; \quad \mathrm{ID}_{50} = -(\beta_0 + \beta_N \overline{\log_{10} N} + \beta_\rho \bar{\rho} + \beta_{\rho^2} \bar{\rho}^2)/\beta_d.$$

For NT_{50}, we solve the quadratic equation $\beta_0 + \beta_d \bar{d} + \beta_N \overline{\log_{10} N} + \beta_\rho \rho + \beta_{\rho^2} \rho^2 = 0$ for ρ.

3.1.4 Model Capacity

The regression analysis and estimated capacity thresholds (Table 2) reveal clear variations among models that can be grouped into three classes:

Frontier/High-capacity Models: gpt-5 and o3 lead by a wide margin (ECL50 > 300), followed by gemini-2.5-pro, gpt-5-mini, o4-mini, and DS-R1-0528. The high baseline performance of gemini-2.5-pro (β_0 = 22.5) together with the large β_N of -9.15 is consistent with the uniquely large amount long context errors as N increases (as illustrated in the Appendix D).

Mid-capacity Models: DS-Llama-70B , Qwen3-32B, DS-Qwen-32B, QwQ-32B, Phi-4-reasoning, Phi-4-reasoning-plus, Qwen3-30B-A3B, gpt-5-nano-2025-08-07, and Qwen3-8B form a broad middle tier with good performance at moderate N and d values.

Low-capacity Models: DS-Qwen-7B, Phi-4-mini-reasoning, DS-Qwen-1.5B, and Qwen3-1.7B exhibit minimal effective context handling capacity failing to reach 50% accuracy even under mean context/distractor conditions, deteriorating rapidly under slightly increasing load.

3.1.5 Differential Sensitivity to Load Dimensions

The estimated coefficients further reveal distinct sensitivity profiles:

Sensitivity to context length (β_N): Universally negative and highly significant, larger models often show greater relative degradation compared to their higher baselines. Yet large ECL_{50} values for frontier models arise from the combined effect of β_0, β_N indicating comparisons are best made via ECL_{50} and not β_N in isolation.

Sensitivity to intrinsic difficulty (β_d): Negative across models with a narrow range, it suggests a more uniform effect. Despite some steep β_d values, high baselines (e.g., gemini-2.5-flash) yield large ID_{50} values unlike smaller models with similar β_d (e.g., Phi-4-reasoning-plus).

Sensitivity to information relevance (β_ρ and β_{ρ^2}): Confirms the U-shaped response, but NT_{50} values reveal nuanced distractor robustness variations masked by aggregate scores (e.g., DS-Llama-70B vs. Qwen3-32B). For frontier models, the absence of NT_{50} indicates achieving above 50% accuracy while for weak models the same absence indicates remaining below 50% accuracy.

3.2 Failure modes across models, length, and difficulty

We analyze error categories from the evaluation pipeline (see Appendix C) to identify failure modes and provide the complete distributions and per-model breakdowns in the Appendix D.

State-tracking mistakes dominate under load. Across models, the most common non-context failure is wrong final attribution in the last valid PoI sentence (valid-logic), consistent with mistracking sequential updates rather than formatting issues. For example, at N=250, Qwen3-32B has 2'541 valid-logic cases, DS-Llama-70B 2'465, and QwQ-32B 2'092. These logic errors also increase monotonically with d for nearly all models.

Long-context budget overflows are a prominent, model-specific failure at extreme N. The max-context errors grow sharply with N for some models: gemini-2.5-flash (280 errors in 350 samples at $N = 250$) or gemini-2.5-pro (268/350). OpenAI models also make these errors at $N = 250$ but at

much lower levels (gpt-5: 32/350; o3: 24/348). The high error counts for the Gemini models indicate relatively poor token efficiency when reasoning.

Instruction-following drift emerges under higher N and d, mainly in smaller models. While poi-logic stays near zero for most models, last-logic increases notably for compact models (e.g., Phi-4-mini-reasoning: 400 last-logic at $N = 250$); DS-Qwen-7B: 116), indicating that under load, models often fail to answer in the instructed format.

"Other" failures rise with sequence length in small and mid-tier models. At $N = 250$, DS-Qwen-1.5B has 605 "other" cases (often claiming the puzzle is unsolvable) and DS-Qwen-7B 461 indicating a shift from precise (but wrong) answers to non-answers as load grows.

4 Discussion

CogniLoad, by operationalizing CLT, enables a multi-dimensional evaluation of LLM reasoning, revealing nuanced failure patterns obscured by single-dimensional benchmarks. Our empirical results (Section 3) offer several key insights: task length (N) emerges as a dominant determinant, suggesting challenges in sustained, germane-like processing for long, intrinsically demanding tasks; models exhibit distinct sensitivities to intrinsic difficulty (d) versus extraneous load (ρ), with the latter surprisingly showing U-shaped performance curves, indicating particular difficulties with intermediate distractor densities; and estimated capacity thresholds provide concise "cognitive fingerprints" for diagnostic LLM evaluation. The limitations of our study are summarized as follows:

Nuances of the CLT-LLM Analogy While CLT provides a powerful analogous framework, it is crucial to acknowledge that "cognitive load" in LLMs manifests as computational constraints (e.g., attention saturation, representational bottlenecks) rather than biological working memory limitations. Our operationalization of N as a proxy for conditions demanding GCL, for example, is an abstraction. Future research should aim to bridge CLT concepts with direct, mechanistic measures of the underlying computational processes in LLMs to refine this analogy.

Scope of Reasoning and Generalizability CogniLoad focuses on sequential and pure deductive reasoning without requiring domain knowledge. While this reasoning type is fundamental to various subject areas (e.g., code, math), it is distinct from alternative reasoning paradigms like inductive, abductive, or analogical reasoning. Extending the CLT-grounded multi-dimensional evaluation to other reasoning types and evaluating it in other languages is a promising next step.

Beyond Accuracy and Main Effects The current evaluation relies on exact-match accuracy. Future iterations could incorporate richer metrics (e.g., step-wise reasoning fidelity, solution coherence, uncertainty of solutions) and systematically investigate interactions among d, N, ρ, which CogniLoad's factorial design supports. Reinforcement learning on verifiable rewards (Guo et al., 2025) presents a promising application of CogniLoad in LLM training, as its generated metadata enables precise verification of reasoning steps despite limited data of this kind.

Despite these considerations, by decomposing the "task difficulty" into principled, controllable dimensions inspired by CLT, CogniLoad provides a more insightful perspective than single-dimensional benchmarks. It allows a more differentiated understanding of LLM reasoning capabilities and limitations, paving the way for more targeted development of robust and generalizable AI systems.

5 Conclusion

We introduced **CogniLoad**, a novel synthetic benchmark grounded in CLT for multi-dimensional evaluation of LLM long-context reasoning. By independently controlling parameters for intrinsic cognitive load (d), extraneous cognitive load (ρ), and task length (N), CogniLoad offers unprecedented diagnostic precision. Our evaluations revealed task length as a dominant performance constraint and uncovered unique "cognitive fingerprints" of LLM sensitivities to different load types, providing actionable insights beyond single-dimensional benchmarks. CogniLoad offers a reproducible, scalable, and theoretically-grounded tool to systematically dissect LLM reasoning limitations and guide the development of more capable and robust AI systems. While human and artificial cognition are mechanistically distinct, applying frameworks such as CLT to AI evaluation can provide valuable perspectives for understanding and characterizing their operational differences and capabilities.

6 LLM DISCLOSURE

LLMs were used only in the early stages of writing the paper to refine phrasing and correct grammar. During coding they were used to update the chart formatting code, and to translate the manual implementation of the generation algorithm in Elixir to python for reproducibility.

7 REPRODUCIBILITY

The code for generating CogniLoad puzzles according to the algorithm described in this paper is provided at: `https://anonymous.4open.science/r/cogniload-292B/`. The dataset of the puzzles on which the LLMs were evaluated for the results presented in this paper is provided on HuggingFace: `https://huggingface.co/datasets/cogniloadteam/cogniload`

REFERENCES

C. An, S. Gong, M. Zhong, X. Zhao, M. Li, J. Zhang, L. Kong, and X. Qiu. L-eval: Instituting standardized evaluation for long context language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14388–14411, 2024.

Y. Bai, X. Lv, J. Zhang, H. Lyu, J. Tang, Z. Huang, Z. Du, X. Liu, A. Zeng, L. Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, 2024a.

Y. Bai, S. Tu, J. Zhang, H. Peng, X. Wang, X. Lv, S. Cao, J. Xu, L. Hou, Y. Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *arXiv preprint arXiv:2412.15204*, 2024b.

P. Chandler and J. Sweller. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332, 1991.

F. Chollet, M. Knoop, G. Kamradt, and B. Landers. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*, 2024.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

K. A. Ericsson and W. Kintsch. Long-term working memory. *Psychological review*, 102(2):211, 1995.

Gkamradt. Needle in a haystack - pressure testing llms, 2023. URL `https://github.com/gkamradt/LLMTest_NeedleInAHaystack`.

D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

G. S. Halford, W. H. Wilson, and S. Phillips. Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and brain sciences*, 21(6):803–831, 1998.

C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

W. Kryściński, N. Rajani, D. Agarwal, C. Xiong, and D. Radev. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*, 2021.

Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, and M. Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

F. Lieder and T. L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43:e1, 2020.

B. Y. Lin, R. L. Bras, K. Richardson, A. Sabharwal, R. Poovendran, P. Clark, and Y. Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*, 2025.

Z. Ling, K. Liu, K. Yan, Y. Yang, W. Lin, T.-H. Fan, L. Shen, Z. Du, and J. Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion. *arXiv preprint arXiv:2501.15089*, 2025.

J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

A. Modarressi, H. Deilamsalehy, F. Dernoncourt, T. Bui, R. A. Rossi, S. Yoon, and H. Schütze. Nolima: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*, 2025.

F. Paas, A. Renkl, and J. Sweller. Cognitive load theory and instructional design: Recent developments. *Educational psychologist*, 38(1):1–4, 2003.

M. Parmar, N. Patel, N. Varshney, M. Nakamura, M. Luo, S. Mashetty, A. Mitra, and C. Baral. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models. *arXiv preprint arXiv:2404.15522*, 2024.

U. Shaham, E. Segal, M. Ivgi, A. Efrat, O. Yoran, A. Haviv, A. Gupta, W. Xiong, M. Geva, J. Berant, et al. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*, 2022.

U. Shaham, M. Ivgi, A. Efrat, J. Berant, and O. Levy. Zeroscrolls: A zero-shot benchmark for long text understanding. *arXiv preprint arXiv:2305.14196*, 2023.

K. Sinha, S. Sodhani, J. Dong, J. Pineau, and W. L. Hamilton. Clutrr: A diagnostic benchmark for inductive reasoning from text. *arXiv preprint arXiv:1908.06177*, 2019.

M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

J. Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2): 257–285, 1988.

J. Sweller. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*, 22:123–138, 2010.

Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

K. Vodrahalli, S. Ontanon, N. Tripuraneni, K. Xu, S. Jain, R. Shivanna, J. Hui, N. Dikkala, M. Kazemi, B. Fatemi, et al. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*, 2024.

K. Yan, Z. Ling, K. Liu, Y. Yang, T.-H. Fan, L. Shen, Z. Du, and J. Chen. Mir-bench: Benchmarking llm's long-context intelligence via many-shot in-context inductive reasoning. *arXiv preprint arXiv:2502.09933*, 2025.

W. Yu, Z. Jiang, Y. Dong, and J. Feng. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*, 2020.

Y. Yu, Q.-W. Zhang, L. Qiao, D. Yin, F. Li, J. Wang, Z. Chen, S. Zheng, X. Liang, and X. Sun. Sequential-niah: A needle-in-a-haystack benchmark for extracting sequential needles from long contexts. *arXiv preprint arXiv:2504.04713*, 2025.

M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. H. Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, et al. Qmsum: A new benchmark for query-based multi-domain meeting summarization. *arXiv preprint arXiv:2104.05938*, 2021.

Y. Zhou, H. Liu, Z. Chen, Y. Tian, and B. Chen. GSM-infinite: How do your LLMs behave over infinitely increasing context length and reasoning complexity?, 2025. URL https://arxiv.org/abs/2502.05252.

Q. Zhu, F. Huang, R. Peng, K. Lu, B. Yu, Q. Cheng, X. Qiu, X. Huang, and J. Lin. Autologi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models. *arXiv preprint arXiv:2502.16906*, 2025.