Novelty Assessment Report

Paper: CogniLoad: A Synthetic Natural Language Reasoning Benchmark With Tunable Length, Intrinsic Difficulty, and Distractor Density
PDF URL: https://openreview.net/pdf?id=0Sex2H5Jnn
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

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.

Core Task Landscape

This paper addresses: **Long-Context Reasoning with Controllable Cognitive Load Dimensions**

A total of **23 papers** were analyzed and organized into a taxonomy with **10 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Cognitive Load Theory Foundations and Benchmarking**
- **Context Management and Memory Optimization Frameworks**
- **Application Domains and Task-Specific Implementations**

Complete Taxonomy Tree

- Long-Context Reasoning with Controllable Cognitive Load Dimensions Survey Taxonomy
- Cognitive Load Theory Foundations and Benchmarking
  - Synthetic Benchmark Design with Parametric Control ★ (2 papers)
  - [0] CogniLoad: A Synthetic Natural Language Reasoning Benchmark With Tunable Length, Intrinsic Difficulty, and Distractor Density (Anon et al., 2026) View paper
  - [15] seqBench: A Tunable Benchmark to Quantify Sequential Reasoning Limits of LLMs (Mohammad Ramezanali, 2025) View paper
  - Cognitive Load Mechanisms and Interference Effects (3 papers)
  - [4] Unable to forget: Proactive lnterference Reveals Working Memory Limits in LLMs Beyond Context Length (Chupei Wang, 2025) View paper
  - [13] Don't Think of the White Bear: Ironic Negation in Transformer Models Under Cognitive Load (Logan Mann, 2025) View paper
  - [16] Cognitive Load Limits in Large Language Models: Benchmarking Multi-Hop Reasoning (Adapala, 2025) View paper
  - Human-Model Cognitive Alignment Studies (3 papers)
  - [8] Context limitations make neural language models more human-like (Brassard, 2022) View paper
  - [14] Text comprehension (An Chi Guo, 2018) View paper
  - [17] CHÃ□P CHÃ□NH Tá°¢-PHÆ¯Æ NG PHÃ□P HIá»□U QUả°¢ TRONG Dá° Y Ká»¸ NÃ□NG NGHE-HIá»□U TIá°¾NG ANH CHO Lá»□P 10 (PT Hiá»□n, 2025) View paper
- Context Management and Memory Optimization Frameworks
  - Active Memory and Workspace Management (2 papers)
  - [3] SAGE: Self-evolving Agents with Reflective and Memory-augmented Abilities (Xuechen Liang, 2025) View paper
  - [7] Cognitive Workspace: Active Memory Management for LLMs--An Empirical Study of Functional Infinite Context (An, 2025) View paper
  - Context Compression and Retrieval-Augmented Generation (2 papers)
  - [9] ContextBot: Improving Response Consistency in Crowd-Powered Conversational Systems for Affective Support Tasks (Yao Ma, 2023) View paper
  - [11] BRIEF-Pro: Universal Context Compression with Short-to-Long Synthesis for Fast and Accurate Multi-Hop Reasoning (Gu, 2025) View paper
  - Positional and Structural Context Reorganization (2 papers)
  - [5] Resonant constraint propagation for large language models through latent manifold interference (Annabel, 2025) View paper
  - [10] RePo: Language Models with Context Re-Positioning (Huayang Li, 2025) View paper
  - Inference-Time Cognitive Load Optimization (2 papers)
  - [12] Cognitive Load-Aware Inference: A Neuro-Symbolic Framework for Optimizing the Token Economy of Large Language Models (Zhang, 2025) View paper
  - [18] Cognitive Overload Attack:Prompt Injection for Long Context (Upadhayay, 2024) View paper

- Application Domains and Task-Specific Implementations
  - Long-Horizon Planning and Action Representation (2 papers)
  - [2] The cognitive bandwidth bottleneck: Shifting long-horizon agent from planning with actions to planning with schemas (Xu, 2025) View paper
  - [6] RoboCerebra: A Large-scale Benchmark for Long-horizon Robotic Manipulation Evaluation (HAN SongHao, 2025) View paper
  - Multi-Hop and Cross-Document Reasoning (4 papers)
  - [19] Robust Long-Context Multilingual Retrieval and Reasoning Enabled by Combined Neural and Symbolic Techniques (SB Nezhad, 2016) View paper
  - [20] ReCogLab: a framework testing relational reasoning & cognitive hypotheses on LLMs (A Liu, n.d.) View paper
  - [21] CEA: Context Engineering Agent for Enhanced Reliability and Sustainability in Deep Research Systems (S HUANG, n.d.) View paper
  - [22] Large Reasoning Models: A Survey of Techniques, Applications, and Future Challenges in Structured AI Reasoning (Qing Li, n.d.) View paper
  - Adaptive Cognitive Control and Multimodal Integration (2 papers)
  - [1] Sparks of cognitive flexibility: self-guided context inference for flexible stimulus-response mapping by attentional routing (Thorat, 2025) View paper
  - [23] When Multimodal Models â□□Burn Outâ□□: Diagnosing and Healing Modality Fatigue via MAD+ MAC (J Yang, n.d.) View paper

Narrative

Core task: long-context reasoning with controllable cognitive load dimensions. This emerging field examines how language models handle extended inputs under varying cognitive demands, drawing inspiration from human cognitive load theory. The taxonomy organizes research into three main branches: Cognitive Load Theory Foundations and Benchmarking, which develops theoretical frameworks and evaluation protocols for measuring cognitive strain; Context Management and Memory Optimization Frameworks, which addresses architectural strategies for efficient information retention and retrieval; and Application Domains and Task-Specific Implementations, which explores how cognitive load principles manifest across diverse problem settings. Representative works span from foundational studies on human-like context limitations (Context Limitations Human-like[8]) to recent frameworks that explicitly model cognitive constraints (Cognitive Bandwidth Bottleneck[2], Cognitive Workspace[7]). The field reflects growing recognition that scaling context windows alone is insufficient without understanding the qualitative dimensions of reasoning difficulty.

Particularly active lines of work explore synthetic benchmark design with parametric control over task complexity, enabling systematic study of how models degrade under increasing cognitive demands. CogniLoad[0] sits squarely within this benchmarking thrust, offering controllable dimensions to isolate specific load factors in long-context scenarios. Its emphasis on parametric control aligns closely with seqBench[15], which similarly provides structured evaluation of sequential reasoning capabilities. Meanwhile, neighboring efforts like SAGE[3] and Cognitive Load-Aware Inference[12] pursue complementary angles: the former develops adaptive reasoning strategies, while the latter optimizes inference under explicit load constraints. A central tension across these branches concerns whether cognitive load should be treated as an intrinsic property of tasks (as in benchmark design) or as a dynamic resource managed by the system (as in memory optimization frameworks). Open questions remain about transferring insights from controlled synthetic settings to real-world applications where multiple load dimensions interact unpredictably.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. seqBench: A Tunable Benchmark to Quantify Sequential Reasoning Limits of LLMs

Authors: Mohammad Ramezanali, Paolo Santi | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â□¦ to probe model performance under more complex long-horizon reasoning regimes. The â□¦ complexity dimensions allows for targeted analysis of their distinct impacts on LLM reasoning â□¦

Relationship Analysis

Both papers belong to the Synthetic Benchmark Design with Parametric Control category, creating benchmarks with independently tunable parameters for systematic evaluation of cognitive load dimensions in reasoning tasks. They overlap in their approach to controlling task complexity through multiple independent dimensions: CogniLoad manipulates intrinsic difficulty (d), distractor density (ρ), and task length (N) in natural language logic puzzles, while seqBench controls logical depth (L), backtracking count (B), and noise ratio (N) in spatial pathfinding tasks. The key difference is that CogniLoad explicitly grounds its design in Cognitive Load Theory with three distinct load types (intrinsic, extraneous, germane), whereas seqBench focuses on sequential reasoning limits through pathfinding scenarios with emphasis on backtracking complexity and uses a different task domain (grid navigation vs. logic puzzles).

Contributions Analysis

Overall novelty summary. The paper introduces CogniLoad, a synthetic benchmark that applies Cognitive Load Theory to evaluate long-context reasoning in LLMs through independently tunable parameters: intrinsic difficulty, distractor-to-signal ratio, and task length. Within the taxonomy, it resides in the 'Synthetic Benchmark Design with Parametric Control' leaf, which contains only two papers total. This represents a relatively sparse research direction focused specifically on benchmarks with systematic parametric control over cognitive load dimensions, suggesting the paper enters a nascent but well-defined area of investigation.

The taxonomy reveals that CogniLoad's parent branch, 'Cognitive Load Theory Foundations and Benchmarking', encompasses three distinct research directions. Its sibling leaves include 'Cognitive Load Mechanisms and Interference Effects' (examining phenomena like proactive interference and context saturation) and 'Human-Model Cognitive Alignment Studies' (comparing model limitations with human cognitive constraints). These neighboring areas provide complementary perspectives—mechanistic studies investigate specific cognitive phenomena, while alignment research grounds model behavior in human baselines—but neither offers the systematic parametric control that defines CogniLoad's methodological contribution.

Among the three contributions analyzed, the first two appear relatively novel within the limited search scope. 'Grounding LLM evaluation in Cognitive Load Theory' examined 10 candidates with zero refutations, while 'CogniLoad benchmark with independent cognitive load control' examined 8 candidates, also with zero refutations. However, the third contribution, 'Automatic puzzle generation and evaluation algorithm', shows more substantial prior work: among 10 candidates examined, 3 were identified as potentially refuting. This suggests that while the theoretical framing and benchmark design may be distinctive, the technical implementation of puzzle generation has more established precedents in the examined literature.

Based on the analysis of 28 candidate papers from semantic search, CogniLoad appears to occupy a relatively novel position in systematically operationalizing cognitive load theory for LLM evaluation. The sparse population of its taxonomy leaf and low refutation rates for core contributions suggest meaningful differentiation from prior work. However, this assessment is constrained by the limited search scope and does not constitute an exhaustive survey of all potentially relevant benchmarking literature.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Grounding LLM evaluation in Cognitive Load Theory

Description: The authors establish a theoretical foundation for evaluating large language models by mapping benchmark parameters to the three types of cognitive load from Cognitive Load Theory: intrinsic load (ICL), extraneous load (ECL), and germane load (GCL). This provides a principled framework for understanding LLM reasoning limitations.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. United Minds or Isolated Agents? Exploring Coordination of LLMs under Cognitive Load Theory

URL: View paper

Brief Assessment

United Minds Isolated[32] applies Cognitive Load Theory to multi-agent LLM coordination and collaboration, not to benchmark design or evaluation frameworks. The candidate focuses on managing cognitive load through agent specialization and communication, while the original paper creates a benchmark with tunable parameters mapped to CLT dimensions.

---

2. Understanding Review Helpfulness through Diagnosticity and Cognitive Load: Comparative Analysis of LLM and ML Models on Restaurant Reviews

URL: View paper

Brief Assessment

Review Helpfulness Diagnosticity[36] applies Cognitive Load Theory to analyze review helpfulness using LLMs, not to evaluate LLM reasoning capabilities themselves. The candidate focuses on restaurant review analysis rather than benchmarking LLM performance.

---

3. Improving Young Learners with Copilot: The Influence of Large Language Models (LLMs) on Cognitive Load and Self-Efficacy in K-12 Programming Education

URL: View paper

Brief Assessment

Copilot Young Learners[31] applies Cognitive Load Theory to evaluate educational outcomes in K-12 programming instruction, not to evaluate LLM reasoning capabilities or benchmark design. The contexts are fundamentally different: educational pedagogy versus AI model evaluation.

---

4. Addressing educational overload with generative AI through dual coding and cognitive load theories

URL: View paper

Brief Assessment

Educational Overload Dual[35] applies Cognitive Load Theory to health professions education and multimodal content design, not to evaluating large language models or their reasoning capabilities.

---

5. Developing the PsyCogMetricsâ□¢ AI Lab to Evaluate Large Language Models and Advance Cognitive Scienceâ□□A Three-Cycle Action Design Science Study

URL: View paper

Brief Assessment

PsyCogMetrics AI Lab[33] applies Cognitive Load Theory to platform design for usability (minimizing intrinsic/extraneous load, maximizing germane load), not to benchmark parameter design for evaluating LLM reasoning capabilities as the original paper does.

---

6. The cognitive impacts of large language model interactions on problem solving and decision making using EEG analysis

URL: View paper

Brief Assessment

Cognitive Impacts EEG[34] focuses on measuring neural dynamics during LLM interactions using EEG signals for emotion classification, not on establishing a theoretical framework for evaluating LLMs through Cognitive Load Theory's three load types (intrinsic, extraneous, germane).

---

7. Cognitive overload: Jailbreaking large language models with overloaded logical thinking

URL: View paper

Brief Assessment

Cognitive Overload Jailbreaking[30] applies Cognitive Load Theory to jailbreak attacks (adversarial prompting), not to benchmark design or systematic evaluation of reasoning capabilities. The original paper creates a diagnostic benchmark with tunable parameters, while the candidate focuses on exploiting cognitive overload for security vulnerabilities.

---

8. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry

URL: View paper

Brief Assessment

Cognitive Ease Cost[39] examines cognitive load effects on students using LLMs for learning tasks, not LLM evaluation frameworks. The candidate focuses on human cognitive load during LLM usage, while the original develops a benchmark to evaluate LLM reasoning capabilities using CLT principles.

---

9. â□¦ of the cognitive theory of multimedia learning in the design and evaluation of an AI educational video assistant utilizing large language models

URL: View paper

Brief Assessment

Multimedia Learning AI[37] applies cognitive load theory to educational video design, not to LLM evaluation or benchmarking. The contexts address fundamentally different applications of CLT.

---

10. Research on the Integration of Multimodal Large Language Models (MLLM) and Augmented Reality (AR) for Smart Navigation with Real-Time Cross-Language Interaction and Cognitive Load Balancing Strategies

URL: View paper

Brief Assessment

MLLM AR Navigation[38] focuses on reducing cognitive load in AR navigation systems through adaptive content delivery, not on evaluating LLM reasoning capabilities using CLT as a theoretical framework for benchmark design.

Contribution 2: CogniLoad benchmark with independent cognitive load control

Description: The authors present CogniLoad, a novel synthetic benchmark that enables independent manipulation of intrinsic difficulty (d), distractor density (ρ), and task length (N). This factorial design allows precise diagnosis of LLM failure modes across distinct cognitive load dimensions.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Use of Eye-Tracking Technology to Investigate Cognitive Load Theory

URL: View paper

Brief Assessment

Eye-Tracking Cognitive Load[28] focuses on measuring cognitive load types using eye-tracking physiological methods in multimedia learning contexts, not on creating controllable synthetic benchmarks for LLM evaluation with factorial parameter manipulation.

2. Cognitive Load Traces as Symbolic and Visual Accounts of Deep Model Cognition

URL: View paper

Brief Assessment

Cognitive Load Traces[27] proposes a mid-level interpretability framework for analyzing internal model dynamics during inference, not a benchmark for evaluating models. The original paper presents a synthetic benchmark with controllable parameters for systematic evaluation, while the candidate focuses on measuring and visualizing cognitive load through internal signals like attention entropy and KV-cache utilization.

3. The Impact of Simple, Brief, and Adaptive Instructions within Virtual Reality Training: Components of Cognitive Load Theory in an Assembly Task

URL: View paper

Brief Assessment

VR Training Instructions[26] focuses on instructional design within virtual reality assembly tasks, not on creating benchmarks for evaluating LLM reasoning capabilities with controllable cognitive load parameters.

4. Trainee perception of cognitive load during observed faculty teaching of procedural skills

URL: View paper

Brief Assessment

Trainee Perception[29] focuses on measuring cognitive load during medical procedural skills training (colonoscopy), not on creating synthetic benchmarks for evaluating LLMs with controllable cognitive load dimensions.

5. Cognitive Load-Aware Inference: A Neuro-Symbolic Framework for Optimizing the Token Economy of Large Language Models

URL: View paper

Brief Assessment

Cognitive Load-Aware Inference[12] focuses on optimizing LLM inference efficiency through cognitive load management during the generation process, not on creating benchmarks that independently control cognitive load dimensions for evaluation purposes. The candidate addresses inference optimization, while the original contribution is about benchmark design for systematic evaluation.

6. Cognitive Load Limits in Large Language Models: Benchmarking Multi-Hop Reasoning

URL: View paper

Brief Assessment

Cognitive Load Limits[16] focuses on multi-hop reasoning tasks with context saturation and attentional residue mechanisms, not on synthetic logic puzzles with factorial control over intrinsic difficulty (d), distractor density (ρ), and task length (N) as independent parameters.

7. Development and validation of a theory-based questionnaire to measure different types of cognitive load

URL: View paper

Brief Assessment

Theory-Based Questionnaire[25] develops a psychometric instrument to measure cognitive load types in human learners, not a synthetic benchmark for evaluating LLMs. The candidate focuses on validating self-report scales for educational psychology research, while the original paper presents a computational benchmark with tunable parameters for testing language models.

8. Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load

URL: View paper

Brief Assessment

Instructional Design Effects[24] focuses on measuring cognitive load in human learning contexts using questionnaires, not on creating synthetic benchmarks for LLM evaluation. The candidate addresses educational psychology and instructional design optimization, while the original paper develops a computational benchmark for testing language models.

Contribution 3: Automatic puzzle generation and evaluation algorithm

Description: The authors develop an algorithmic framework for automatically generating and evaluating natural-language logic puzzles with tunable parameters. This enables scalable, reproducible benchmarking of reasoning capabilities across different models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Enigmata: Scaling Logical Reasoning in Large Language Models with Synthetic Verifiable Puzzles

URL: View paper

Prior Art Analysis

Enigmata[45] demonstrates prior work on automatic generation and evaluation algorithms for logic puzzles. The candidate paper presents a comprehensive framework with 36 puzzle tasks, each equipped with an automated generator-verifier pair that can produce unlimited examples with controllable difficulty and instant correctness verification. This directly refutes the novelty claim of the original

paper's algorithmic framework for automatically generating and evaluating natural-language logic puzzles with tunable parameters, as Enigmata[45] provides a more extensive implementation of the same concept across multiple puzzle categories.

Evidence

Evidence 1 - **Rationale**: Both papers describe automated generation and evaluation systems for puzzles. Enigmata[45] explicitly presents a generator-verifier architecture that produces puzzles with controllable difficulty and automatic verification, which is the same core concept as the original paper's contribution. - **Original**: we provide an algorithm for the automatic randomized generation and evaluation of puzzle instances, enabling large-scale and reproducible comparison of llm capabilities. - **Candidate**: every task is backed by an automatedgenerator-verifier pair: the generator can vary both the quantity and the difficulty of the puzzles it produces, while the verifier instantly checks the correctness.

Evidence 2 - **Rationale**: Both papers describe algorithmic frameworks for puzzle generation with tunable parameters. Enigmata[45] provides auto-generators for 30 tasks and auto-verifiers for all 36 tasks, demonstrating a comprehensive prior implementation of automatic puzzle generation and evaluation. - **Original**: 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... - **Candidate**: phase ii: auto-generator and verifier development. second, we equip 30 tasks inenigmata with custom puzzle instance auto-generators, which enable automatic data scaling to generate training and evaluation data tailored to o1-like complex reasoning research. in addition, all 36 puzzle tasks have corr...

Evidence 3 - **Rationale**: Both papers implement tunable difficulty parameters in their puzzle generation algorithms. Enigmata[45] explicitly describes identifying and controlling difficulty variables in the auto-generator, which is the same approach as the original paper's tunable parameter system. - **Original**: intrinsic difficulty (d) for d ∈ {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 (≈ (d + 1)d)... - **Candidate**: phase iii: sliding difficulty control.for each puzzle task, we identify key variables that control difficulty, such as grid size and blank cell count inbinario. these variables serve as parameters in our auto-generator to create puzzle instances across difficulty levels.

---

2. MARVEL: Multidimensional Abstraction and Reasoning through Visual Evaluation and Learning

URL: View paper

Brief Assessment

MARVEL[49] focuses on visual abstract reasoning puzzles with fixed patterns and manual curation (770 puzzles), not automatic generation with tunable parameters for natural-language logic puzzles.

---

3. LLM-ARC: Enhancing LLMs with an Automated Reasoning Critic

URL: View paper

Brief Assessment

LLM-ARC[44] focuses on a neuro-symbolic framework for logical reasoning using answer set programming and critic feedback, not on automatic generation of logic puzzles or benchmark creation algorithms.

---

4. SATBench: Benchmarking LLMs' Logical Reasoning via Automated Puzzle Generation from SAT Formulas

URL: View paper

Prior Art Analysis

SATBench[48] demonstrates that automated puzzle generation and evaluation algorithms for logic benchmarks existed prior to the original paper. SATBench[48] presents a fully automated generation process that samples SAT formulas, translates them into natural language puzzles using LLMs, and validates them through both LLM-based and solver-based consistency checks. The original paper's claim to novelty in developing an algorithmic framework for automatic generation and evaluation is refuted by SATBench[48]'s earlier implementation of similar automated generation, validation, and evaluation mechanisms for logic puzzles.

Evidence

Evidence 1 - **Rationale**: Both papers claim automated generation and evaluation of logic puzzles. SATBench[48] explicitly states its generation process is 'fully automated' with quality ensured through 'llm-based and solver-based checks,' directly paralleling the original paper's claim of providing 'an algorithm for the automatic randomized generation and evaluation of puzzle instances.' - **Original**: we provide an algorithm for the automatic randomized generation and evaluation of puzzle instances, enabling large-scale and reproducible comparison of llm capabilities. - **Candidate**: our generation process is fully automated and features adjustable difficulty levels by varying the number of clauses in sat formulas. we ensure the quality of the 2100 generated logical puzzles through llm-based and solver-based checks, with human validation showing passing rates above 90%.

Evidence 2 - **Rationale**: Both papers describe systematic, multi-stage algorithmic approaches to puzzle generation with tunable parameters. The original paper uses d, n, and ρ parameters, while SATBench[48] uses SAT formula sampling with adjustable clause numbers, demonstrating prior work in parameterized automatic puzzle generation. - **Original**: 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... - **Candidate**: the generation method is divided into three stages: sat formula sampling (section 3.1), llm-based story generation (section 3.2), and consistency validation (section 3.3). in the evaluation phase, we assess the correctness of the reasoning trace (section 3.4).

Evidence 3 - **Rationale**: Both papers implement automated evaluation algorithms. The original paper uses exact-matching with lexical variants, while SATBench[48] uses 'llm-based and solver-based consistency checks' with human validation, showing that automated evaluation mechanisms for logic puzzles existed prior to the original work. - **Original**: we evaluate each puzzle by exact-matching of the queried attribute value in the output of modelm with accommodating minor phrasing and common lexical variants. - **Candidate**: our objective is to create logical puzzles derived from boolean satisfiability (sat) formulas, ensuring the quality of the dataset through both llm-based and solver-based consistency checks. we further validate each llm-involved process with human review.

Evidence 4 - **Rationale**: Both papers describe systematic approaches to controlling puzzle difficulty through tunable parameters. SATBench[48]'s use of variable and clause counts to control difficulty demonstrates that automated, parameterized puzzle generation with difficulty control existed before the original paper's contribution. - **Original**: 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. - **Candidate**: automation and difficulty controlthe sat problem can be solved using a sat solver, which provides a soundness guarantee and allows for an automated and scalable solution. to systematically generate problems with varying levels of difficulty, we can sample formulas that differ in the number of boolea...

---

5. AutoLogi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models

URL: View paper

Prior Art Analysis

AutoLogi[40] demonstrates that prior work exists for automatic generation and evaluation of logic puzzles. Both papers present algorithmic frameworks for automatically generating natural-language logic puzzles with controllable parameters and automated evaluation mechanisms. The candidate paper explicitly describes a three-stage automated pipeline (puzzle formulation, format & verifiers generation, and data augmentation) that generates logic puzzles with tunable difficulty levels and uses program-based verification for evaluation. This directly parallels the original paper's claim of developing 'an algorithmic framework for automatically generating and evaluating natural-language logic puzzles with tunable parameters.' The candidate paper was published as a preprint in February 2025, predating any potential publication of the original paper, and provides detailed technical implementation of automated puzzle generation with controllable complexity parameters.

Evidence

Evidence 1 - **Rationale**: Both papers claim to provide automated methods for generating logic puzzles. The candidate explicitly describes automated generation with controllable difficulty, directly overlapping with the original's claim of automatic generation with tunable parameters. - **Original**: we provide an algorithm for the automatic randomized generation and evaluation of puzzle instances, enabling large-scale and reproducible comparison of llm capabilities. - **Candidate**: we introduce a novel method for automatically synthesizing openended logic puzzles to construct a reasoning benchmark named autologi. our approach offers three key advantages: (1) as is illustrated in figure 1, the open-ended format requires models to construct complete solutions from scratch, signi...

Evidence 2 - **Rationale**: While the original claims to be 'the first' to control cognitive load dimensions, the candidate demonstrates a complete automated pipeline with controllable difficulty parameters, challenging the novelty of automated generation with tunable parameters. - **Original**: we introduce cogniload, the first benchmark designed to independently control these three dimensions of cognitive load, while scaling to arbitrarily long contexts. - **Candidate**: the proposed method follows a three-stage pipeline: information extraction from corpora for open-ended question synthesis, program-based verifier generation using advanced llms, and dataset augmentation for difficulty balance.

Evidence 3 - **Rationale**: Both papers describe parameterized puzzle generation with controllable difficulty. The candidate's constraint-based difficulty control mechanism demonstrates prior implementation of tunable puzzle generation algorithms. - **Original**: 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... - **Candidate**: stage 3 performs data augmentation through two complementary techniques, reduction and expansion, to construct a dataset with balanced difficulty distribution, enabling better discrimination of models' logical reasoning capabilities. in the reduction process, we randomly select and remove one logica...

Evidence 4 - **Rationale**: Both papers describe automated evaluation procedures. The candidate's cross-validation approach for verifying puzzle solutions demonstrates a sophisticated automated evaluation algorithm that predates the original paper's claims. - **Original**: we evaluate each puzzle by exact-matching of the queried attribute value in the output of modelm with accommodating minor phrasing and common lexical variants. - **Candidate**: to ensure the reliability of the evaluation mechanism, we propose a cross-validation method to check the correctness of llm-generated verifiers and traversal function. we run the traversal function to examines all possible combinations within the value ranges defined in background (domain space), an...

---

6. Puzzle-Level Generation With Simple-Tiled and Graph-Based Wave Function Collapse Algorithms

URL: View paper

Brief Assessment

Puzzle-Level Generation[46] focuses on procedural generation of spatial puzzle levels (Strimko, Flow) using WFC algorithms, not natural-language logic puzzles with tunable cognitive load parameters for LLM benchmarking.

---

7. Steamroller Problems: An Evaluation of LLM Reasoning Capability with Automated Theorem Prover Strategies

URL: View paper

Brief Assessment

Steamroller Problems[47] focuses on evaluating LLMs using existing ATP reasoning strategies on a specific theorem-proving domain, not on developing algorithmic frameworks for automatically generating and evaluating natural-language logic puzzles with tunable parameters.

---

8. LLM Reasoners: New Evaluation, Library, and Analysis of Step-by-Step Reasoning with Large Language Models

URL: View paper

Brief Assessment

LLM Reasoners[42] focuses on evaluating reasoning chains across diverse reasoning approaches (CoT, ToT, RAP) using AutoRace for evaluation, not on generating logic puzzles with tunable parameters for benchmarking.

---

9. PuzzleBench: A Fully Dynamic Evaluation Framework for Large Multimodal Models on Puzzle Solving

URL: View paper

Brief Assessment

PuzzleBench[41] focuses on multimodal visual puzzle generation for evaluating large multimodal models (LMMs) on image-based tasks, while the original paper develops natural-language logic puzzles for evaluating LLMs' reasoning capabilities. The domains, modalities, and evaluation targets are fundamentally different.

---

10. Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

URL: View paper

Brief Assessment

Illusion of Thinking[43] focuses on critiquing experimental design flaws in existing benchmarks rather than proposing new generation algorithms. It does not present an automatic puzzle generation framework.

Appendix: Text Similarity Detection

Textual similarity detection checked 29 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. United Minds or Isolated Agents? Exploring Coordination of LLMs under Cognitive Load Theory

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] CogniLoad: A Synthetic Natural Language Reasoning Benchmark With Tunable Length, Intrinsic Difficulty, and Distractor Density View paper
- [1] Sparks of cognitive flexibility: self-guided context inference for flexible stimulus-response mapping by attentional routing View paper
- [2] The cognitive bandwidth bottleneck: Shifting long-horizon agent from planning with actions to planning with schemas View paper
- [3] SAGE: Self-evolving Agents with Reflective and Memory-augmented Abilities View paper
- [4] Unable to forget: Proactive lnterference Reveals Working Memory Limits in LLMs Beyond Context Length View paper
- [5] Resonant constraint propagation for large language models through latent manifold interference View paper
- [6] RoboCerebra: A Large-scale Benchmark for Long-horizon Robotic Manipulation Evaluation View paper
- [7] Cognitive Workspace: Active Memory Management for LLMs--An Empirical Study of Functional Infinite Context View paper
- [8] Context limitations make neural language models more human-like View paper
- [9] ContextBot: Improving Response Consistency in Crowd-Powered Conversational Systems for Affective Support Tasks View paper
- [10] RePo: Language Models with Context Re-Positioning View paper
- [11] BRIEF-Pro: Universal Context Compression with Short-to-Long Synthesis for Fast and Accurate Multi-Hop Reasoning View paper
- [12] Cognitive Load-Aware Inference: A Neuro-Symbolic Framework for Optimizing the Token Economy of Large Language Models View paper
- [13] Don't Think of the White Bear: Ironic Negation in Transformer Models Under Cognitive Load View paper
- [14] Text comprehension View paper
- [15] seqBench: A Tunable Benchmark to Quantify Sequential Reasoning Limits of LLMs View paper
- [16] Cognitive Load Limits in Large Language Models: Benchmarking Multi-Hop Reasoning View paper
- [17] CHÃ□P CHÃ□NH Tá°¢-PHÆ¯Æ NG PHÃ□P HIá»□U QUá°¢ TRONG Dá° Y Ká»¸ NÃ□NG NGHE-HIá»□U TIá°¾NG ANH CHO Lá»□P 10 View paper
- [18] Cognitive Overload Attack:Prompt Injection for Long Context View paper
- [19] Robust Long-Context Multilingual Retrieval and Reasoning Enabled by Combined Neural and Symbolic Techniques View paper
- [20] ReCogLab: a framework testing relational reasoning & cognitive hypotheses on LLMs View paper
- [21] CEA: Context Engineering Agent for Enhanced Reliability and Sustainability in Deep Research Systems View paper
- [22] Large Reasoning Models: A Survey of Techniques, Applications, and Future Challenges in Structured AI Reasoning View paper
- [23] When Multimodal Models â□□Burn Outâ□□: Diagnosing and Healing Modality Fatigue via MAD+ MAC View paper
- [24] Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load View paper
- [25] Development and validation of a theory-based questionnaire to measure different types of cognitive load View paper
- [26] The Impact of Simple, Brief, and Adaptive Instructions within Virtual Reality Training: Components of Cognitive Load Theory in an Assembly Task View paper
- [27] Cognitive Load Traces as Symbolic and Visual Accounts of Deep Model Cognition View paper
- [28] Use of Eye-Tracking Technology to Investigate Cognitive Load Theory View paper
- [29] Trainee perception of cognitive load during observed faculty teaching of procedural skills View paper
- [30] Cognitive overload: Jailbreaking large language models with overloaded logical thinking View paper
- [31] Improving Young Learners with Copilot: The Influence of Large Language Models (LLMs) on Cognitive Load and Self-Efficacy in K-12 Programming Education View paper
- [32] United Minds or Isolated Agents? Exploring Coordination of LLMs under Cognitive Load Theory View paper
- [33] Developing the PsyCogMetricsâ□¢ AI Lab to Evaluate Large Language Models and Advance Cognitive Scienceâ□□A Three-Cycle Action Design Science Study View paper
- [34] The cognitive impacts of large language model interactions on problem solving and decision making using EEG analysis View paper
- [35] Addressing educational overload with generative AI through dual coding and cognitive load theories View paper
- [36] Understanding Review Helpfulness through Diagnosticity and Cognitive Load: Comparative Analysis of LLM and ML Models on Restaurant Reviews View paper
- [37] â□¦ of the cognitive theory of multimedia learning in the design and evaluation of an AI educational video assistant utilizing large language models View paper
- [38] Research on the Integration of Multimodal Large Language Models (MLLM) and Augmented Reality (AR) for Smart Navigation with Real-Time Cross-Language Interaction and Cognitive Load Balancing Strategies View paper
- [39] Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry View paper
- [40] AutoLogi: Automated generation of logic puzzles for evaluating reasoning abilities of large language models View paper
- [41] PuzzleBench: A Fully Dynamic Evaluation Framework for Large Multimodal Models on Puzzle Solving View paper
- [42] LLM Reasoners: New Evaluation, Library, and Analysis of Step-by-Step Reasoning with Large Language Models View paper
- [43] Comment on The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity View paper
- [44] LLM-ARC: Enhancing LLMs with an Automated Reasoning Critic View paper
- [45] Enigmata: Scaling Logical Reasoning in Large Language Models with Synthetic Verifiable Puzzles View paper
- [46] Puzzle-Level Generation With Simple-Tiled and Graph-Based Wave Function Collapse Algorithms View paper
- [47] Steamroller Problems: An Evaluation of LLM Reasoning Capability with Automated Theorem Prover Strategies View paper
- [48] SATBench: Benchmarking LLMs' Logical Reasoning via Automated Puzzle Generation from SAT Formulas View paper
- [49] MARVEL: Multidimensional Abstraction and Reasoning through Visual Evaluation and Learning View paper