10.48047/jocaaa.2025.34.11.14

Semantic Consistency Testing in Large Language Model-Based Conversational Systems

Yash Panjari
Independent Researcher, USA

Abstract

Large Language Models have transformed conversational AI systems, making human-computer interactions more natural and contextually informed through high-fidelity natural language understanding and generation abilities. But the probabilistic nature of such systems poses unprecedented challenges in guaranteeing semantic coherence in multi-turn dialogues, when traditional deterministic testing approaches are fundamentally unsuitable. This extensive survey discusses recent methods for testing semantic consistency in LLM-based conversational AI, highlighting essential gaps in current evaluation methods while introducing holistic frameworks for testing conversational coherence. The material discusses the transition from rule-based conversation systems to modern transformer-based models, emphasizing the sophistication of consistency evaluation needs over time. Some of the main contributions include adversarial conversation generation systems, graph-based knowledge tracking methods, and multi-modal consistency evaluation frameworks. The review lays the grounds for designing strong, industry-standard methodologies for conversational AI reliability testing, including both technical implementation problems and practical deployment requirements. With a thorough analysis of modern test paradigms and new metrics for evaluation, this work offers critical insights to practitioners looking to deploy effective consistency testing frameworks into production systems.

Keywords: Large Language Models, Conversational AI, Semantic Consistency, Software Testing, Dialogue Systems

1. Introduction

The emergence of Large Language Models (LLMs) in chat applications has revolutionized the face of human-computer interaction to its core. Modern language models show incredible ability to produce contextually appropriate responses for a wide range of domains through few-shot learning methods, where models are able to carry out tasks using very little task-specific training data [1]. These systems have created new paradigms for natural language generation and understanding with emergent capabilities that increase with model size and computational power. Deployment into production environments has, however, highlighted key issues around sustaining semantic coherence across longer conversations.

Conventional software testing practices, founded on deterministic input-output behavior, are insufficient when applied to the probabilistic and contextual nature of LLM-based systems. The large response space and emergent properties exhibited by these models call for new quality assurance methods that go beyond traditional functional testing paradigms [2]. Modern conversational systems need to handle sophisticated multi-turn dialogues while ensuring coherence over extended sequences of interactions, which calls for highly developed evaluation frameworks capable of judging both real-time response quality and long-term conversational consistency.

The challenge gets harder when the architectural sophistication of contemporary conversational systems that consist of many different components, such as retrieval models, knowledge repositories, and domain-specific reasoning modules, is taken into consideration. Open-domain conversational systems have to be able to balance engagement, knowledge usage, and personality coherence without repetitive or

10.48047/jocaaa.2025.34.11.14

inconsistent responses over a wide variety of conversation topics. The probabilistic nature of language generation adds variability that standard test methods cannot easily deal with, especially when measuring subjective factors like conversational naturalness and contextual relevance.

The value of semantic consistency testing is not merely of academic concern but has direct consequences for user trust, system reliability, and the commercial success of conversational AI applications. Data show that inconsistent answers are one of the main causes of user abandonment of AI-driven dialogue systems. Additionally, the release of semantically inconsistent systems in life-critical environments like healthcare, education, and customer service presents serious threats to the credibility of organizations and the safety of users.

Assessment difficulties in conversational AI stem from the multi-dimensionality of debate, which consists of high-quality debate that includes real correctness, logical coherence, emotional suitability, and contextual appropriateness. Unlike other software structures wherein correctness can be objectively checked, conversational AI structures need sophisticated assessment strategies that address the subjective and context-established nature of human discourse. The aggregate of retrieval-augmented era, reminiscence tactics, and personality modeling makes the testing situation even harder, necessitating sophisticated frameworks that may show sophisticated behavioral styles in an extensive variety of interaction contexts.

This extensive review discusses the state of semantic consistency testing in LLM-based conversational systems at present, discussing both theoretical models and practical implementation methodologies. The review includes conventional testing constraints, developing evaluation methods, and the creation of specialized frameworks aimed at solving the specific issues that probabilistic language models in conversational settings offer.

10.48047/jocaaa.2025.34.11.14

2. Literature Review and Background

The development of conversational AI testing has passed through a number of quite distinct periods, each defined by progressively advanced techniques for assessment and quality control. Initial dialogue systems, founded upon rule-based architectures and finite state machines, utilized fairly simple testing techniques that were centered mostly around coverage of pre-planned conversation routes and response accuracy within restricted domains. These systems were in highly controlled settings in which the patterns of conversation flows were deterministic, making extensive validation possible using exhaustive path testing and rule-checking methods.

Statistical and neural-based methods for dialog generation presented a dramatic change to test requirements that fundamentally altered the landscape of evaluation for conversational systems. The conventional methods of evaluation encountered considerable difficulty when they were used for neural conversation models, especially in their evaluation of response diversity as well as conversational naturalness. The introduction of diversity-inducing objective functions constituted an important step in overcoming the tendency of neural models to produce generic and repetitive responses that are unengaging in conversation [3]. These methods underscore the limitations of conventional similarity metrics in measuring conversational quality, calling for the creation of more advanced assessment frameworks that can better account for the subtleties in natural conversation.

The shift to neural dialogue systems brought novel challenges to the area of evaluation methodology, with systems yielding intricate behavioral profiles that challenged conventional testing paradigms. Statistical models of dialogue showcased the requirement for evaluation paradigms that can capture semantic coherence, contextual relevance, and fluency of conversation across protracted interaction sequences. The advent of attention-based neural architectures rendered testing paradigms even more challenging, necessitating the development of complex methodologies to examine long-range dependencies and contextual comprehension within multi-turn dialogue.

Recent developments in transformer-based language models have resulted in a tenfold increase in conversational AI testing complexity, with emergent behavior that cannot be anticipated by component-level testing methods. Large-scale language models show advanced reasoning and contextual understanding capabilities that require system-wide evaluation methodologies to measure system-wide performance for a wide range of interaction scenarios. The intricateness of such systems has unveiled the significance of holistic testing frameworks that will detect fine-grained semantic discrepancies, logical inconsistencies, and factual errors that might arise over long conversational windows.

Modern conversational AI evaluation frameworks have increasingly considered multi-dimensional assessment methods incorporating coherence, consistency, fluency, and relevance measures. Current assessment approaches acknowledge human-oriented evaluation principles based on user satisfaction, conversation naturalness, and contextual appropriateness as central quality determinants. The establishment of knowledge consistency measures tailored to identify factual contradictions in neural dialogue systems as a particularly advanced automated quality assurance in conversational applications is an important achievement.

The adversarial testing challenge in conversational AI has attracted much recent attention, with studies illustrating the susceptibility of language models to specially constructed inputs that can draw out adverse responses. Universal adversarial triggers are a specific and worrying vulnerability where particular input subsequences can provoke models into generating unwanted outputs without regard to the context of the conversation [4]. These results have underscored the essential need for strong testing paradigms with the

10.48047/jocaaa.2025.34.11.14

ability to detect possible failure modes under hostile environments and have prompted the creation of extensive analysis methods that measure system resilience in a wide range of difficult situations.

Graph-based conversation modeling has developed as a promising area for consistency assessment, with structured techniques for monitoring information propagation and identifying contradictions in multi-turn conversations. Such methodologies enable complex frameworks for representing conversational context, allowing semantic inconsistency detection to be performed that might otherwise escape through other evaluation methods. Combinations of knowledge graph representations with conversational AI systems provide potential solutions for automated consistency checks and quality assurance within large-scale deployment contexts.

Testing Era	Primary Methodology	Key Evaluation Focus
Rule-based Systems	Exhaustive path testing and rule verification approaches	Coverage of predefined conversation routes and response accuracy within constrained domains
Statistical & Neural Systems	Diversity-promoting objective functions and similarity-based metrics	Response diversity, conversational naturalness, and semantic coherence across interaction sequences
Transformer-based Models	Multi-dimensional assessment frameworks and adversarial testing methods	System-wide performance evaluation, emergent behavior detection, and resilience against universal adversarial triggers

Table 1: Evolution of Conversational AI Testing Methodologies [3, 4]

3. Semantic Consistency Testing in Large Language Model-Based Conversational Systems

Conventional software testing is based on deterministic inputs and outcomes, but large language models pose special testing challenges because they are probabilistic and have immense response spaces. Modern LLM-based systems have high response variation given the same input situations and consistency problems grow immensely as the length of conversation moves beyond early interaction phases. Extensive testing frameworks expressly developed to measure semantic coherence within multi-turn dialogue need to contend with the intrinsic randomness of neural language creation without compromising tough quality assurance standards.

The fundamental problem is in identifying subtle inconsistencies that human users easily perceive but that prevailing testing techniques often miss. Product conversational systems typically suffer from knowledge claim inconsistencies, whereby systems can claim expertise over particular domains during initial stages of conversation but later contradict earlier claims or depict knowledge gaps within prolonged dialogues. The failures of consistency are a matter of serious concern for conversational AI deployment and play a large role in leading to user abandonment while decreasing overall satisfaction across various application areas.

Sophisticated assessment methods call for new measures of evaluation that go beyond the conventional similarity-based metrics like BLEU scores or perplexity computations, which show weak correlation with human consistency judgments. Unsupervised evaluation methods for interactive dialogue systems have the potential to present viable alternatives to human evaluation, such that conversational quality is accessible in a scalable way without needing large-scale manual annotation [5]. These methods are based

10.48047/jocaaa.2025.34.11.14

on semantic coherence evaluation frameworks that involve factual consistency verification, logical reasoning chain validation, and contextual appropriateness scoring, demanding much higher computational resources compared to conventional metrics but offering more sound consistency assessment capabilities.

Automated test environments that can create adversarial conversation flows intended to reveal inconsistencies prove highly useful for the detection of likely failure modes that are missed by conventional testing methodologies. Graph representation strategies that monitor knowledge claims and identify contradictions across dialogue turns support advanced mechanisms for detecting semantic consistency within long-running conversational sessions. Specialized knowledge representation designs that support monitoring of logical consistency and factual accuracy across multi-turn conversations are needed to realize these detection systems.

Implementation methodologies include the creation of domain-specific testing harnesses that can mimic realistic patterns of conversation in a variety of domains and interactional contexts. Development of benchmark datasets for semantic consistency testing involves a wide coverage of types of inconsistencies, including temporal, factual, and logical contradiction situations. Industry standards for testing the reliability of conversational AI complement current functional testing methodologies with significant improvements to overall system reliability when implemented as part of extensive quality assurance programs.

The probabilistic character of LLMs creates a number of unique classes of semantic inconsistency for which specialized detection mechanisms are necessary. Temporal inconsistencies exist when systems present conflicting information on time-sensitive subjects within conversation histories. Factual inconsistencies are key failure modes in which systems present conflicting assertions of objective fact or exhibit differential knowledge of fact across conversational contexts. Logical inconsistencies exist when systems compromise underlying principles of reasoning or contradict established positions of dialogue.

Coherence testing by entailment-based methods provides advanced techniques for identifying semantic inconsistencies in neural dialogue systems [6]. Such approaches offer frameworks for evaluating logical interactions among conversational turns and allowing automated identification of contradictions not evident from surface-level inspection. Next-generation testing frameworks are needed to meet scalability challenges intrinsic to the evaluation of systems with exponentially increasing conversation state spaces, requiring smart sampling methods that achieve maximal coverage with computational tractability.

Reinforcement learning methods for adaptive test case guidance towards the most promising scenarios for revealing inconsistencies yield convergent testing strategies that reach optimal test case distributions. Such techniques present scalable solutions for real-time quality assurance in production conversational AI, allowing systematic detection of consistency vulnerabilities under various interaction patterns.

Human-in-the-loop testing continues to be essential for verifying consistency detection systems that are automated, with studies showing considerable benefits in detecting implicit knowledge-based subtle semantic inconsistencies, contextual subtlety, and domain-area specialties. Blended evaluation frameworks incorporating automated screening along with focused human review maximize consistency checking while minimizing workload on manual review, in contrast to completely human-based methods.

10.48047/jocaaa.2025.34.11.14

Inconsistency Type	Definition and Characteristics	Detection Method
Temporal Inconsistencies	Systems present conflicting information on time-sensitive subjects within conversation histories	Specialized detection mechanisms with temporal reasoning capabilities
Factual Inconsistencies	Systems present conflicting assertions of objective facts or exhibit differential knowledge across conversational contexts	Unsupervised evaluation methodologies and factual consistency verification frameworks
Logical Inconsistencies	Systems compromise underlying principles of reasoning or contradict established dialogue positions	Entailment-based coherence evaluation methods and automated contradiction detection

Table 2: Types and Detection Methods for Conversational AI Consistency Failures [5, 6]

4. Methodological Frameworks for Consistency Evaluation

Design of strong methodological frameworks for semantic consistency testing calls for a multi-dimensional strategy that accounts for technical and practical aspects. Recent literature has identified a number of major components critical for thorough consistency testing, such as adversarial conversation generation, graph-based knowledge monitoring, multi-modal consistency checking, and longitudinal coherence analysis. Current frameworks need to strike a balance between thoroughness of evaluation and computational cost while ensuring scalability for production use cases.

Adversarial conversation generation is an important aspect of sophisticated testing systems, with bespoke systems showing enhanced ability to unveil semantic inconsistencies over conventional random sampling methods. The systems utilize advanced natural language generation algorithms to generate conversation sequences that are built to probe LLM responses' possible vulnerabilities through reinforcement learning methods that optimize adversarial tactics via iterative refinement procedures. The method involves creating realistic dialogue contexts that test systematic various dimensions of model knowledge and reasoning ability while ensuring conversational naturalness.

Modern adversarial generation platforms use multi-objective optimization methods that optimize for conversation realism and inconsistency detection capability, resulting in solutions that are conversational-realistic yet have maximal exposure of consistency failure modes. Sophisticated implementations utilize hierarchical generation techniques that produce detailed conversation exploration over various conversational scenarios. Such systems prove to be very effective in exposing temporal reasoning inconsistencies, facts-knowledge gaps, and logical contradiction patterns that are typically found in long multi-turn conversations.

Graph-based knowledge tracking supports disciplined methods of tracking information flow and contradiction detection between dialogue turns, with contemporary examples building dynamic knowledge graphs capturing facts, claims, and mentions of relationships throughout conversations. Temporal reasoning algorithms are used to track changes in information over conversation histories, preserving graph structures to support contradiction detection over remotely separated conversational turns. Knowledge graph building utilizes advanced entity extraction algorithms that are highly accurate for factual claim detection and relationship identification over a wide range of conversational topics.

10.48047/jocaaa.2025.34.11.14

High-performance graph-based monitoring systems utilize distributed storage structures that can handle large conversation histories for large-scale deployment applications, with query processing sustaining fast response times for consistency checking requests. The creation of human-labeled multi-turn dialogue datasets offers critical groundwork for training and testing these knowledge tracking systems, with strong coverage of conversational patterns and consistency issues [7]. These kinds of systems incorporate temporal reasoning facilities that identify consistency errors occurring over long stretches of dialogue histories, offering robust long-term coherence monitoring.

Multi-modal consistency evaluation confronts the growing trend of conversational systems incorporating text, image, and audio inputs into a single system, with evaluation systems processing varied input combinations in conjunction while enforcing modality consistency. Such systems assess coherence while processing varied input formats within conversation sequences, making sure that textual outputs align with visual or auditory context data. Implementation entails the creation of cross-modal representation learning methods that are capable of finding contradictions among modality inputs, calling for distinctive neural architectures for holistic multi-modal understanding.

Longitudinal coherence assessment analyzes consistency over prolonged conversation sessions and multiple episodes of interaction, where systems retain consistent conversation models over prolonged interaction histories. Such implementations utilize episodic memory structures that can monitor long-term patterns of consistency while accommodating changing user preferences and contextual details. Learning end-to-end goal-oriented dialogue methods offers frameworks for retaining coherence over prolonged conversational contexts while seeking certain goals [8].

The deployment of these frameworks demands prudent regard for computational efficiency and scalability, with systems for monitoring consistency in real-time holding evaluation thoroughness against response latency demands for production deployment. Hierarchical evaluation methods utilize effective screening protocols followed by rigorous analysis for indication interactions. Distributed computing architectures permit parallel processing of numerous conversation threads, affording scalability needs for systems managing many simultaneous conversations while maintaining complete consistency evaluation coverage.

Benchmark dataset construction is an important element of methodological framework adoption, with thorough evaluation datasets including large sets of annotated conversation instances from various domains. Existing research places emphasis on developing large-scale datasets that cover the entire range of consistency issues, such as domain knowledge verification, temporal reasoning testing, logical inference testing, and factual correctness verification.

10.48047/jocaaa.2025.34.11.14

Framework Component	Primary Methodology	Key Capabilities
Adversarial Conversation Generation	Multi-objective optimization methods with reinforcement learning techniques for iterative refinement	Exposes temporal reasoning inconsistencies, factual knowledge gaps, and logical contradiction patterns in multi-turn conversations
Graph-based Knowledge Tracking	Dynamic knowledge graphs with temporal reasoning algorithms and distributed storage architectures	Tracks information flow and detects contradictions across widely separated conversational exchanges using manual multi-turn dialogue datasets
Longitudinal Coherence Evaluation	Episodic memory architectures with end-to-end goal-oriented dialog frameworks	Maintains consistency across extended conversation sessions while adapting to evolving user preferences and contextual information

Table 3: Key Components of Methodological Frameworks for Semantic Consistency Evaluation [7, 8]

5. Future Directions

This thorough review has considered the state of the art of semantic consistency testing in large language model-driven conversational systems, noting major successes and enduring challenges in the area. The discussion identifies that although classical testing approaches are unsuitable for probabilistic language models, the newer frameworks hold a lot of potential to serve the special needs of conversational AI testing. Industry trends forecast significant investment growth in conversational AI testing techniques, fueled largely by corporate need for trustworthy consistency evaluation tools across the range of application domains.

The research in specialized testing harnesses, adversarial conversation generation systems, and graph-based knowledge tracking methods constitutes significant progress toward strong consistency evaluation, with implementations to date exhibiting remarkable consistency detection performance across various conversational contexts. Still, large-scale problems must be overcome in scalability, real-time capability, and the incorporation of human-oriented evaluation metrics. The semantic coherence complexity of multi-turn dialogues requires ongoing innovation in theoretical paradigms and practical implementation strategies, with research funding commitments forecasted to grow dramatically to solve these ongoing issues.

Directions for future studies must focus on the creation of standardized measures of evaluation that will be widely accepted by the industry, with ongoing standardization initiatives undertaken by key technology firms and research institutions globally. The construction of benchmark datasets aimed at consistency evaluation will allow for comparison purposes and expedite the development of testing methodology. Language models built specifically for dialog applications show the promise of more natural and consistent conversational exchange, offering a basis for improving assessment frameworks that can evaluate both factual correctness and conversational naturalness [9].

In addition, the incorporation of cutting-edge technologies like neuro-symbolic reasoning and causal inference could bring new consistency detection and evaluation strategies with potential early studies showing the promise of accuracy gains over existing purely neural methods. Neuro-symbolic architectures hold special promise in logical consistency checking, performing higher accuracy in

10.48047/jocaaa.2025.34.11.14

identifying contradictions in reason compared to standard neural techniques. The computational overhead for such hybrid systems is still high, but current research efforts on optimizing them pursue minimizing the cost of processing while maximizing the quality of assessment.

Comprehensive consistency testing frameworks' real-world application involves interaction between standardization bodies, industry stakeholders, and academic researchers, with existing collaborative programs already engaging large research groups in various countries. Establishing industry standards for conversational AI reliability testing will make consistency assessment a standard part of the system development and deployment processes, with regulatory compliance schemes likely to mandate consistency testing for essential applications. In addition, the development of open-source testing tools and frameworks will make high-level evaluation tools available to organizations of all sizes and technical proficiency levels.

Future opportunities for semantic consistency testing are the possibility of incorporating these frameworks into the training loops of large language models themselves. Self-supervised consistency learning mechanisms have the potential to allow models to create internal consistency monitoring abilities, cutting down on the dependency on off-the-shelf evaluation systems while keeping comparable detection performance. Chain-of-thought reasoning methods present the possibility for language models to form more organized and consistent reasoning habits, laying building blocks for self-monitoring consistency evaluation abilities [10].

Moreover, extending consistency testing principles to new domains like multimodal conversational systems and embodied AI agents will necessitate ongoing innovation and adjustment of currently available methodologies. Consistency testing multimodal systems poses especially tricky challenges, with cross-modal inconsistencies running at high rates as opposed to single-modal systems. The creation of specialized testing frameworks for embodied AI agents is a central research area that will need future investment, with initial deployments to start in specialist fields like healthcare and education.

The final success of semantic consistency testing frameworks will be determined by their effectiveness in building user confidence, system dependability, and overall quality of human-computer dialogue interactions. As conversational AI systems find their way into more critical applications, with significant deployment growth across healthcare, financial services, and education sectors, the need for sound consistency evaluation will increase, rendering this an imperative area of ongoing research and development investment.

10.48047/jocaaa.2025.34.11.14

Future Direction	Implementation Approach	Key Challenge/Opportunity
Standardized Evaluation Metrics	Collaborative initiatives involving major technology companies and academic institutions worldwide	Industry-wide adoption of benchmark datasets specifically designed for consistency evaluation to facilitate comparative analysis
Advanced Technology Integration	Neuro-symbolic reasoning and causal inference methods with multi-objective optimization frameworks	Superior performance in logical consistency verification while managing substantial computational overhead requirements
Self-Supervised Learning Systems	Chain-of-thought reasoning approaches integrated into language model training processes	Development of internal consistency monitoring capabilities to reduce dependency on external evaluation systems
Multimodal and Embodied AI Testing	Specialized testing frameworks for cross-modal consistency and embodied agent applications	Managing elevated inconsistency rates in multimodal systems and deployment in critical domains like healthcare and education

Table 4: Emerging Directions in Semantic Consistency Evaluation for Large Language Models [9, 10]

Conclusion

The systematic review of semantic consistency testing of large language model-based conversational systems is both filled with tremendous progress and still beset by numerous challenges that characterize the present state of conversational AI assessment. While established testing paradigms are inadequate to probabilistic language models, new frameworks show strong promise for meeting the distinctive needs of conversational AI assessment, specifically through the creation of domain-specific testing harnesses, adversarial conversation generation programs, and graph-based knowledge tracking solutions. These developments show strong strides toward robust consistency assessment capabilities, but there are still serious challenges in scalability, real-time execution, and the incorporation of human-focused evaluation metrics. Future work should focus on the creation of standardized metrics of evaluation that would be widely adopted across the industry and supported by thorough benchmark datasets specifically created for the purpose of consistency evaluation. The convergence of cutting-edge technologies like neuro-symbolic reasoning and causal inference presents promising directions to consistency detection and evaluation, while the possibility of integrating consistency testing frameworks into the training procedures of large language models themselves holds prospects for self-supervised consistency learning methods. Final success with semantic consistency test suites will be gauged by their impact to increase user confidence, system dependability, and the general quality of human-computer conversational exchanges, and so this is a key area of continued investment development as conversational AI systems increasingly permeate high-stakes uses in healthcare, financial services, and education fields.

10.48047/jocaaa.2025.34.11.14

References

1. Tom B. Brown, et al., "Language Models are Few-Shot Learners," arXiv preprint 2020. https://arxiv.org/abs/2005.14165
2. Stephen Roller, et al., "Recipes for Building an Open-Domain Chatbot," Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021. https://aclanthology.org/2021.eacl-main.24.pdf
3. Jiwei Li, et al., "A Diversity-Promoting Objective Function for Neural Conversation Models," ResearchGate, 2016. https://www.researchgate.net/publication/305334287_A_Diversity-Promoting_Objective_Function_for_Neural_Conversation_Models
4. Eric Wallace, et al., "Universal Adversarial Triggers for Attacking and Analyzing NLP," arXiv, 2021. https://arxiv.org/abs/1908.07125
5. Shikib Mehri and Maxine Eskenazi, "Unsupervised Evaluation of Interactive Dialog with DialoGPT," ACL Antology, 2020. https://aclanthology.org/2020.sigdial-1.28.pdf
6. Nouha Dziri, et al., "Evaluating Coherence in Dialogue Systems using Entailment," University of Alberta, 2019. https://webdocs.cs.ualberta.ca/~zaiane/postscript/NAACL19.pdf
7. Yanran Li, et al., "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," arXiv preprint, 2017. https://arxiv.org/pdf/1710.03957
8. Antoine Bordes, et al., "LEARNING END-TO-END GOAL-ORIENTED DIALOG," arXiv preprint, 2017. https://arxiv.org/pdf/1605.07683
9. Romal Thoppilan, et al., "LaMDA: Language Models for Dialog Applications," arXiv preprint, 2022. https://arxiv.org/pdf/2201.08239
10. Jason Wei, et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv preprint, 2023. https://arxiv.org/pdf/2201.11903