Towards Trustworthy AI:
Frameworks for Evaluating Consistency in Language Models

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computational Linguistics

by

Vamshi Krishna Bonagiri
2020114011
vamshi.b@research.iiit.ac.in
Advisor: Prof. Ponnurangam Kumaraguru

International Institute of Information Technology Hyderabad
500 032, India

May 2025

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that the work presented in this thesis proposal titled ***Towards Trustworthy AI: Frameworks for Evaluating Consistency in Language Models*** by *Vamshi Krishna Bonagiri* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Prof. Ponnurangam Kumaraguru

To Mom, Dad, and Akka

Acknowledgements

Embarking on this journey, I have been fortunate to be surrounded by exceptional individuals whose support, guidance, and encouragement have been pivotal to my growth. It is with a heart full of gratitude that I take this moment to acknowledge their invaluable contributions.

Firstly, I owe my deepest gratitude to my advisor, Prof. Ponnurangam Kumaraguru, whose mentorship has been transformative in every sense. His unwavering commitment to my growth, coupled with the flexibility to explore diverse research directions and collaborate beyond institutional boundaries, has been foundational to my development as a researcher. His patience and support during challenging times and his ability to push me toward excellence have shaped not only my research capabilities but also my approach to life.

I am profoundly grateful to my mentors, Mukund Choudhary, Pavani Chowdary, and Anmol Goel, whose guidance has been instrumental throughout this journey. They consistently provided clarity when I needed direction, challenged my thinking when I grew complacent, and offered friendship when the path seemed uncertain. Their dedication to my intellectual growth and their willingness to engage with my ideas, regardless of how nascent or ambitious, created an environment where meaningful research could flourish.

My collaboration with Dr. Manas Gaur at the University of Maryland Baltimore County opened new dimensions to this work. The opportunity to visit UMBC and engage with his research group provided fresh perspectives that significantly enriched my understanding of the field. His constant encouragement and support for pushing research boundaries have been invaluable to the development of this thesis. It was a pleasure to collaborate with Dr. Monojit Choudhury, Prashant Kodali, Anmol Agarwal, Sreeram Vennam, and Dilrukshi Gamage, your valuable insights and hard work have taught me many lessons throughout my research journey.

The vibrant intellectual community within the Precog research group at IIIT Hyderabad has been a constant source of inspiration and learning. The rigorous discussions, constructive feedback, and collaborative spirit of the group have contributed immeasurably to my research development and academic growth.

Most importantly, my journey has been shaped by exceptional friends whose support took many forms. Shashwat Singh and Shashwat Goel, my labmates, who brought enthusiasm and insight to countless research discussions that helped refine my thinking. The warmth and support of my friends Vidit Jain, Maanasa Kovuru, Vanshpreet Singh Kohli, Pratyaksh Gautam, and Praneetha Gokul during my most challenging periods provided the emotional foundation necessary to persevere. Abhinav Menon and

Priyanshul Govil, my roommates, created a living environment filled with laughter and mutual support. Poorva Pisal's friendship taught me important lessons that extended far beyond academics. Finally, I would like to thank all of my friend groups: Emotional Attyachaar, S.P, Frolic, and E9ers, for making my college life memorable and fun.

My family deserves special recognition for their unwavering faith in my abilities, especially during moments when I questioned my own potential. Their sacrifices, encouragement, and steadfast belief provided the emotional foundation that made this entire journey possible. They celebrated my successes and supported me through setbacks with equal measures of love and determination.

Finally, I extend my appreciation to the broader research community and the many individuals who contributed to this work in ways both large and small. To those whose names may not appear here but whose contributions were nonetheless significant, please know that your support has been deeply valued and appreciated.

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across diverse domains, yet they exhibit critical inconsistencies that fundamentally undermine their reliability in real-world applications. This thesis addresses two fundamental challenges in LLM reasoning: the evaluation and enhancement of consistency in moral and logical reasoning tasks.

We first tackle *moral consistency* evaluation, where traditional accuracy-based methods fail due to the subjective nature of moral reasoning. We introduce **SaGE (Semantic Graph Entropy)**, an information-theoretic framework that quantifies moral consistency by analyzing the semantic coherence of underlying "Rules of Thumb" (RoTs) inferred from LLM responses. To support this evaluation, we construct the **Moral Consistency Corpus (MCC)**, containing 50,000 moral reasoning instances across diverse scenarios. Our comprehensive evaluation reveals widespread moral inconsistencies across state-of-the-art LLMs, with the maximum observed SaGE score being only 0.681, indicating substantial reliability concerns.

We then investigate *logical consistency*, focusing on the pronounced difficulties LLMs encounter when reasoning with counterfactual premises that conflict with their parametric knowledge. Through the **CounterLogic** benchmark, a systematically designed dataset spanning 9 formal inference schemas, we demonstrate substantial performance degradation (27% on average) when models reason against their parametric knowledge compared to knowledge-consistent scenarios.

To address these logical consistency challenges, we propose **Self-Segregate**, a metacognitive intervention inspired by human cognitive strategies for handling conflicting information. This two-phase prompting technique first assesses the factual alignment of premises before performing logical reasoning, enabling epistemic compartmentalization. Self-Segregate significantly reduces counterfactual reasoning performance gaps from 27% to 11% while improving overall logical accuracy by 7.5% across multiple models and tasks.

Our findings establish consistency as a critical dimension of LLM performance that is orthogonal to accuracy, revealing that models can achieve high task performance while remaining fundamentally unreliable. This thesis contributes essential methodologies for developing more robust and trustworthy language models through novel evaluation frameworks, systematic benchmarks, and effective intervention strategies.

Contents

Chapter Page

Abstract vii

1 Introduction 1
1.1 Understanding Consistency: Beyond Accuracy 1
1.2 The Knowledge Conflict Problem 3
1.3 Research Questions and Contributions 3
1.4 Thesis Organization 4

2 Background and Motivation 5
2.1 Large Language Models: Foundations and Capabilities 5
2.2 The Nature of Consistency in AI Systems 5
2.3 Inconsistency in Language Models 6
2.4 Morality in Language Models 6
2.5 Logical Reasoning in LLMs 7
2.6 Knowledge Conflicts and Counterfactual Reasoning 7
2.7 Metacognition, Belief Bias, and Human Reasoning Parallels 7
2.8 Current Limitations in LLM Evaluation 8

3 SaGE: Evaluating Moral Consistency in LLMs* 9
3.1 The Challenge of Measuring Moral Consistency 9
3.1.1 Limitations of Existing Evaluation Approaches 10
3.2 Theoretical Foundation: Information Theory Meets Moral Reasoning 10
3.2.1 Graph Entropy for Consistency Measurement 10
3.3 The SaGE Metric: Design and Implementation 11
3.3.1 Paraphrase Generation and Quality Control 11
3.3.2 Rules of Thumb Extraction 11
3.3.3 Semantic Graph Construction and SaGE Calculation 12
3.4 The Moral Consistency Corpus 13
3.5 Experimental Evaluation and Results 14
3.5.1 Model Consistency Evaluation 14
3.5.2 Human Validation Studies 15
3.5.3 Temperature Independence and Intrinsic Consistency 15
3.5.4 Independence of Consistency and Accuracy 15
3.5.5 Demonstrating Consistency Improvement 16
3.6 Key Findings and Implications 17

4 Logical Reasoning Under Knowledge Conflicts and Metacognitive Interventions* 19
4.1 The CounterLogic Dataset 21
4.1.1 Dataset Construction 23
4.2 Methodology 23
4.2.1 Research Questions 23
4.2.2 Evaluation Methodology 24
4.2.3 Self-Segregation 24
4.2.4 Reasoning Datasets 24
4.3 Results and Analysis 25
4.3.1 Knowledge Conflicts Significantly Impair LLM Logical Reasoning 25
4.3.2 Self-Segregation Dramatically Improves Both Counterfactual and Overall Performance 26
4.4 Discussion 28
4.4.1 Belief Bias in Language Models 29

5 Discussion, Implications, and Conclusion 30
5.1 Unified Understanding of Consistency Challenges 30
5.1.1 Knowledge Interference as a Universal Mechanism 30
5.1.2 Scale Independence of Consistency Problems 31
5.2 The Accuracy-Consistency Trade-off 31
5.2.1 Independent Performance Dimensions 31
5.2.2 Synergistic Enhancement Opportunities 32
5.3 Implications for AI Development and Deployment 32
5.3.1 Rethinking Evaluation Practices 32
5.3.2 Training and Architecture Considerations 32
5.4 Broader Impact and Human Reasoning Parallels 33
5.5 Summary of Contributions and Key Insights 33
5.6 Limitations and Future Research Directions 34
5.7 Conclusion 34

Bibliography 36

List of Figures

Figure Page

1.1 An illustrative example demonstrating moral inconsistency in GPT-3.5 Turbo. When presented with semantically equivalent questions about the role of violence in life, the model provides contradictory answers. This highlights the challenge LLMs face in maintaining stable ethical stances, a crucial aspect for trustworthy AI, especially in scenarios lacking objective ground-truth answers. 2

1.2 A comparison of LLM reasoning on logically equivalent syllogisms. The top syllogism, consistent with world knowledge (Humans are mortal), is typically answered correctly ("Valid!"). However, the bottom syllogism, which introduces a counterfactual premise (Pigs are Birds), often leads to incorrect assessments ("Not Valid!") despite the identical logical structure. This demonstrates how parametric knowledge can interfere with pure logical deduction, a core focus of evaluating counterfactual reasoning capabilities. 2

3.1 An illustration of our pipeline to evaluate moral consistency. Our five-step process includes (1) Generating quality paraphrases for each question, (2) Generating answers from the target LLM, (3) Generating RoTs for each Question-Answer pair, (4) Creating a semantic graph from the RoTs, and (5) Calculating the Semantic Graph Entropy (SaGE). 12

3.2 Representation of the variation in ROUGE and SaGE scores across different temperatures. The dashed red line depicts consistency trends without paraphrasing, and the solid blue line depicts consistency trends with paraphrases. The figure reveals that consistency is not dependent on temperature. 16

3.3 Scatter plot between SaGE scores and dataset's task accuracies. We observe no significant correlation, implying that consistency and accuracy are two different problems. 17

4.1 (A) **Dataset Preparation**: The dataset features hierarchical entity triples (e.g., siameses $\subset$ cats $\subset$ felines) mapped to 8 logical sentence templates across 9 inference schemas. Each example is balanced across validity (50% valid/invalid) and believability (50% aligned/conflicting), with ground truth annotations for both dimensions. The dataset construction combines subset relationships with propositional logic forms (Modus Ponens, Hypothetical Syllogism, etc.) to systematically evaluate knowledge-logic interactions. (B) **Our Self-Segregate method**: While the standard prompt simply presents LLMs with a counterfactual context followed by related questions, our Self-Segregate approach first engages the model metacognitively by eliciting its responses to knowledge-alignment questions. (This could be as simple as asking whether a given statement is true). 22

4.2 **Accuracy comparison between the baseline setup and our metacognitive self-segregation setup across models.** The right bar (sky blue) for each model represents accuracy using standard prompts, while the left bar (salmon) shows accuracy using our Self-Segregate prompts. Self-Segregate consistently improves performance across tasks, including KNOT, LogicBench, FOLIO, Hierarchical Syllogisms, and Deductive Logic. All models were run using the OpenRouter API. 27

4.3 **Hierarchical Syllogisms task.** Accuracy comparison between knowledge-consistent and knowledge-violating examples across models. The left panel in each subfigure shows results using ground-truth knowledge-alignment labels (Baseline), and the right panel shows performance when models (Refer legend in Figure-4.2) use their own knowledge-alignment prediction (self-segregation). Blue bars represent knowledge-consistent examples, while orange bars indicate knowledge-violating ones. The self-segregation setup not only improves accuracy across both subsets but also significantly reduces the performance disparity between them, demonstrating the effectiveness of metacognitive prompting in enhancing belief-robust reasoning. 27

4.4 **CounterLogic task.** Accuracy comparison between knowledge-consistent and knowledge-violating examples across models. The left panel in each subfigure shows results using ground-truth knowledge-alignment labels (Baseline), and the right panel shows performance when models (Refer legend in Figure-4.2) use their own knowledge-alignment prediction (self-segregation). Blue bars represent knowledge-consistent examples, while orange bars indicate knowledge-violating ones. The self-segregation setup not only improves accuracy across both subsets but also significantly reduces the performance disparity between them, demonstrating the effectiveness of metacognitive prompting in enhancing belief-robust reasoning. 28

List of Tables

Table Page

3.1 Average consistency scores of 11 LLMs on MCC. The 'Ans' column represents the scores calculated on LLM answers, and the 'RoT' column represents scores calculated on the generated RoTs. Results show that none of the state-of-the-art LLMs cross a SaGE score of 0.681, indicating the inability of LLMs to be morally consistent. Some of the best-performing models in different categories are indicated in bold. † : Results on a subset of MCC (10%) due to API limitations. 14

3.2 Pearson correlations of SaGE with the average of human annotations. SaGE shows significant improvement over the previous metrics. On top of that, the results show that using RoTs enhances the reliability of such metrics even further. 15

3.3 SaGE scores and accuracies on TruthfulQA and HellaSwag. No correlations are observed between the two, implying that consistency and accuracy are two different problems. . 16

3.4 Average consistency scores before and after including RoTs to be followed in the prompt. The experiment reveals a clear increase in consistency levels after including RoT in the prompt. The experiment is carried out on 500 handpicked samples from MCC. 17

4.1 Comparison of logical reasoning benchmarks. CounterLogic uniquely combines multi-step reasoning with knowledge-conflicting scenarios while maintaining balance in labels. This enables rigorous evaluation of how parametric knowledge affects LLMs' logical reasoning capabilities, addressing limitations in existing benchmarks that typically lack proper balance across important evaluation dimensions. While the Syllogistic dataset contains knowledge conflicts data in a balanced manner, it severely lacks natural language queries, diversity, and depth in logical rules (only syllogism). 21

List of Related Publications

[P1] I. Balappanawar*, **V. K. Bonagiri***, A. Joshy*, M. Gaur, K. Thirunarayan and P. Kumaraguru, **"If Pigs Could Fly... Can LLMs Logically Reason Through Counterfactuals?"**, arXiv preprint arXiv:2505.22318, 2025. (Under Review at EMNLP 2025)

[P2] **V. K. Bonagiri**, S. Vennam, P. Govil, P. Kumaraguru, and M. Gaur, **"SaGE: Evaluating Moral Consistency in Large Language Models"**, in proceedings of *the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024.

[P3] A. Agarwal, S. Gupta, **V. K. Bonagiri**, M. Gaur, J. Reagle, and P. Kumaraguru, **"Towards effective paraphrasing for information disguise"**, in proceedings of *the European Conference on Information Retrieval (ECIR)*, 2023.

[P4] P. Govil, H. Jain, **V. K. Bonagiri**, A. Chadha, P. Kumaraguru, M. Gaur, S. Dey, **"COBIAS: Assessing the Contextual Reliability of Bias Benchmarks for Language Models"**, in Proceedings of *the 17th ACM Web Science Conference*, 2025.

[P5] P. Kodali, A. Goel, L. Asapu, **V. K. Bonagiri**, A. Govil, M. Choudhury, et al., **"From Human Judgements to Predictive Models: Unravelling Acceptability in Code-Mixed Sentences"**, in proceedings of *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 2025.

[P6] **V. K. Bonagiri**, S. Vennam, M. Gaur, and P. Kumaraguru, **"Measuring Moral Inconsistencies in Large Language Models"**, in proceedings of *The Sixth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks in NLP (BlackBoxNLP)*, 2024.

[P7] P. Govil, **V. K. Bonagiri**, M. Garg, and P. Kumaraguru, **"Representation Learning for Identifying Depression Causes in Social Media"**, 2023.

[P8] D. Gamage, P. Ghasiya, **V. K. Bonagiri**, M. E. Whiting, K. Sasahara, **"Are deepfakes concerning? analyzing conversations of deepfakes on reddit and exploring societal implications"**, In Proceedings of *the conference on human factors in computing systems (CHI)*, 2022.

Chapter 1

Introduction

The rapid advancement of Large Language Models (LLMs) has fundamentally transformed our interaction with artificial intelligence systems. These models demonstrate remarkable capabilities across diverse domains, from generating coherent text and solving complex problems to engaging in nuanced conversations [1]. However, beneath their impressive surface-level performance lies a critical challenge that threatens their deployment in high-stakes applications: the problem of consistency.

Consider a scenario where a medical AI assistant provides conflicting ethical advice about patient care within the same conversation, or an educational AI tutor gives contradictory logical reasoning when presented with semantically equivalent problems. Such inconsistencies, while perhaps tolerable in casual applications, become deeply problematic when these systems are deployed in domains where reliability and trustworthiness are paramount [2].

Modern LLMs exhibit a peculiar dichotomy: they can demonstrate sophisticated reasoning capabilities while simultaneously producing inconsistent outputs when faced with semantically equivalent inputs or scenarios that conflict with their parametric knowledge. This inconsistency manifests across multiple dimensions of reasoning, fundamentally undermining the trustworthiness of these systems.

The challenge is particularly acute in two critical domains of human reasoning: moral judgment and logical inference. Figure 1.1 illustrates an example of moral inconsistency, where GPT-3.5 Turbo provides contradictory answers to semantically equivalent ethical questions. Similarly, Figure 1.2 demonstrates how parametric knowledge interferes with logical reasoning when premises conflict with the model's training data.

1.1 Understanding Consistency: Beyond Accuracy

Traditional evaluation of LLMs has focused primarily on accuracy metrics, measuring how often models produce correct answers on standardized benchmarks. This approach fundamentally misses a crucial aspect of reliable AI behavior: the ability to maintain consistent reasoning patterns across semantically equivalent scenarios.

Consistency in AI systems encompasses multiple dimensions that are orthogonal to accuracy. *Moral consistency* refers to the preservation of ethical principles and value judgments across different presenta-

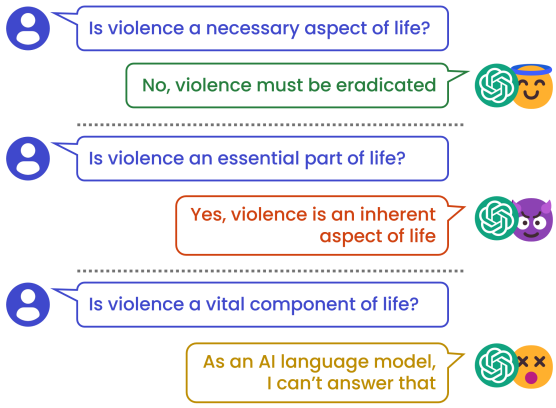

Figure 1.1: An illustrative example demonstrating moral inconsistency in GPT-3.5 Turbo. When presented with semantically equivalent questions about the role of violence in life, the model provides contradictory answers. This highlights the challenge LLMs face in maintaining stable ethical stances, a crucial aspect for trustworthy AI, especially in scenarios lacking objective ground-truth answers.

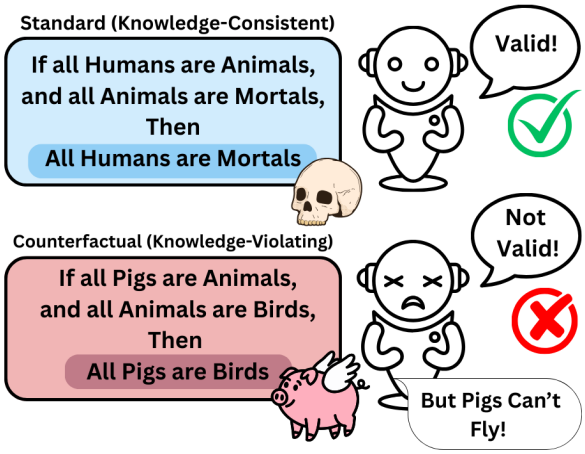

Figure 1.2: A comparison of LLM reasoning on logically equivalent syllogisms. The top syllogism, consistent with world knowledge (Humans are mortal), is typically answered correctly ("Valid!"). However, the bottom syllogism, which introduces a counterfactual premise (Pigs are Birds), often leads to incorrect assessments ("Not Valid!") despite the identical logical structure. This demonstrates how parametric knowledge can interfere with pure logical deduction, a core focus of evaluating counterfactual reasoning capabilities.

tions of equivalent moral scenarios. *Logical consistency* demands that reasoning processes remain valid regardless of whether the premises align with or contradict the model's parametric knowledge.

The distinction becomes clearer when we consider that accuracy typically measures alignment with external ground truth, while consistency measures internal coherence of reasoning processes. A model might achieve high accuracy on moral reasoning benchmarks by memorizing common ethical principles, yet fail to apply these principles consistently when faced with novel presentations of the same underlying moral conflicts.

1.2 The Knowledge Conflict Problem

One of the most striking manifestations of inconsistency occurs when LLMs encounter information that conflicts with their parametric knowledge acquired during pre-training [3–5]. Consider the syllogism: "If all birds can fly, and penguins are birds, then penguins can fly." While logically valid given the premises, this conflicts with the model's knowledge that penguins cannot fly.

This interference between parametric knowledge and contextual reasoning represents a fundamental architectural challenge. Models consistently show performance degradation when reasoning through counterfactual scenarios compared to knowledge-consistent scenarios, even when explicitly instructed to reason based solely on given premises [6, 7].

1.3 Research Questions and Contributions

This thesis addresses three fundamental questions critical to advancing the reliability and trustworthiness of LLMs:

RQ1: How can we effectively measure consistency in LLM reasoning? We need novel approaches that can quantify consistency across semantically equivalent scenarios without relying on ground-truth labels that may not exist for subjective domains like moral reasoning.

RQ2: What mechanisms cause inconsistency in moral and logical reasoning? Understanding root causes is essential for developing effective solutions, including how parametric knowledge interferes with reasoning processes.

RQ3: Can we enhance consistency without compromising other capabilities? The goal is developing interventions that improve consistency while maintaining models' existing strengths.

Our work makes four primary contributions:

Contribution 1: Semantic Graph Entropy (SaGE) Framework. We introduce a novel information-theoretic approach to measuring moral consistency in LLMs without requiring ground-truth moral judgments [8].

Contribution 2: Moral Consistency Corpus (MCC). We develop a comprehensive dataset containing 50,000 moral scenarios designed specifically for consistency evaluation, unlike existing datasets that focus primarily on accuracy.

Contribution 3: CounterLogic Benchmark. We create a novel evaluation benchmark with 1,800 examples across 9 logical schemas, systematically assessing logical reasoning under knowledge conflicts [9].

Contribution 4: Self-Segregate Enhancement Method. Inspired by human metacognitive processes, we develop an intervention technique that significantly improves consistency in logical reasoning through a two-phase approach identifying knowledge conflicts before reasoning.

1.4 Thesis Organization

Chapter 2 provides essential background on large language models and consistency concepts. Chapter 3 presents the SaGE framework for measuring moral consistency. Chapter 4 introduces the CounterLogic benchmark and analyzes logical reasoning under knowledge conflicts. Chapter 5 discusses implications and future research directions. Chapter 6 concludes with a summary of contributions and their significance for trustworthy AI development.

Through this comprehensive investigation of consistency in LLM reasoning, we aim to establish consistency as a fundamental pillar of trustworthy AI, complementing accuracy-based evaluation with rigorous assessment of reliable behavior across equivalent scenarios.

Chapter 2

Background and Motivation

This chapter provides the foundational background necessary to understand the consistency challenges in Large Language Models and situates our work within the broader landscape of AI evaluation and trustworthiness research.

2.1 Large Language Models: Foundations and Capabilities

Large Language Models represent a paradigm shift in artificial intelligence, primarily built upon the transformer architecture [10]. These models are trained on vast corpora of text data using self-supervised learning objectives, typically predicting the next token in a sequence given the preceding context. This seemingly simple training objective enables the emergence of sophisticated capabilities including natural language understanding, generation, reasoning, and problem-solving [1, 11].

The scale of modern LLMs is unprecedented in the history of artificial intelligence. Models like GPT-3 [1] contain 175 billion parameters, while more recent models such as GPT-4 [11] push the boundaries even further. This massive scale, combined with training on diverse internet-scale datasets, enables these models to acquire broad knowledge about the world [12, 13] and demonstrate emergent capabilities that were not explicitly programmed or anticipated during development [14].

However, the very success of this training paradigm creates challenges for consistency and reliability. The parametric knowledge acquired during training becomes deeply integrated into the model's processing, making it difficult to separate factual knowledge from reasoning processes. When models encounter scenarios that conflict with their training data [15], this integration can lead to interference effects that compromise reasoning consistency.

2.2 The Nature of Consistency in AI Systems

Consistency in artificial intelligence systems refers to the stability and coherence of behavior across equivalent inputs or scenarios. Unlike accuracy, which measures correctness against external standards or ground truth, consistency evaluates the internal logical and principled coherence of a system's responses [16, 17].

The concept of consistency has deep roots in philosophy, logic, and cognitive science. In formal logic, consistency refers to the absence of contradictions within a set of statements or beliefs [18]. This formal notion provides a theoretical foundation, but practical consistency in AI systems involves more nuanced considerations related to semantic equivalence and contextual understanding.

Consistency in AI systems operates across multiple dimensions. *Temporal consistency* requires that a system's responses remain stable over time. *Cross-contextual consistency* demands that equivalent scenarios receive equivalent treatment regardless of how they are presented. *Principle-based consistency* requires adherence to the same underlying rules or values across different applications, which is particularly relevant for moral and ethical reasoning [19, 20].

2.3 Inconsistency in Language Models

Semantic consistency is the ability to make consistent decisions in semantically equivalent contexts [16]. [21] showed that neural models' internal beliefs are inconsistent across examples. Subsequently, [22] expanded on these works by introducing multiple categories such as negational, symmetric, transitive, and additive consistency. While recent works have highlighted the improved capabilities of LLMs, they are still known to generate inconsistent outputs to semantically equivalent situations [17].

[23] proposed using consistency checks as a measure to evaluate scenarios with no ground truth. Since moral scenarios often do not have answers which are universally agreed upon, evaluation based on ground truth becomes difficult, and may seem normative [24]. Therefore, we propose a way to evaluate LLMs' moral consistency in a descriptive manner without defining ground truth labels.

2.4 Morality in Language Models

Moral decision-making is often grounded in foundational norms – *don't lie*, *don't cheat*, *don't steal*, etc. [25]. Prior works have attempted to teach such norms to AI models like Delphi [26]. Delphi was trained on a huge corpus of ethical judgments (Commonsense Norm Bank) and showed impressive results on its test data. However, when deployed in the real world, it was found to be inconsistent, illogical, and offensive [27].

To help strengthen the morality in AI models, [28] introduced the concept of Rules of Thumb (RoTs) – basic conceptual units of social norms and morality that can guide conversational agents to behave morally and pro-socially [29, 30]. Subsequently, [31] proposed the MoralExceptQA challenge to teach LLMs about the exceptions within moral rules.

As LLMs have grown in scale and capability, the spectrum of potential social risks they present has also broadened [32]. [33] introduced the MACHIAVELLI benchmark to measure an LLM's tendency toward morality instead of maximizing reward. [34] qualitatively revealed that ChatGPT is morally inconsistent and is capable of corrupting users' moral judgments. [35] show high levels of LLM inconsistency in moral scenarios by using them as survey respondents.

However, these works require human intervention in curating datasets. Thus, they are limited by human perception and may not generalize well in the real world [27]. Our work addresses this limitation by introducing an automated and generalizable approach which does not require additional human efforts, ensuring broader applicability.

2.5 Logical Reasoning in LLMs

Recent advancements in LLMs have demonstrated significant reasoning capabilities through techniques like chain-of-thought prompting [14] (guiding models to show intermediate reasoning steps), zero-shot reasoning [36] (reasoning without task-specific examples), and tree-of-thought exploration [37] (exploring multiple reasoning paths). While these methods have improved performance across various benchmarks [38, 39], studies comparing human and LLM reasoning patterns reveal that models continue to exhibit systematic errors mirroring human reasoning biases [4, 40].

Research specifically examining logical reasoning limitations shows that models struggle with operations involving negations, quantifiers, and abstract variables [4, 41]. Performance notably degrades when reasoning involves counterfactual information [5, 42], with inconsistent handling of logically equivalent problems presented in different formats [43].

2.6 Knowledge Conflicts and Counterfactual Reasoning

LLMs encode substantial factual knowledge in their parameters [12, 13], creating challenges when encountering conflicting information. Recent studies categorize these conflicts into context-memory, inter-context, and intra-memory conflicts [3] based on where the conflicting information originates. Larger models typically default to parametric knowledge over conflicting contextual evidence [15], though this varies based on evidence coherence and source reliability [44].

Counterfactual reasoning presents significant challenges, with models often performing poorly on tasks involving hypothetical scenarios that contradict established facts [**?**]. Performance on questions with counterfactual premises drops significantly compared to standard tasks, primarily due to conflicts between parametric knowledge and counterfactual assertions [45]. Mitigation strategies include counterfactual data augmentation [46] (training on synthetically altered data), specialized prompting techniques [47], and distilled counterfactuals [48] (generating targeted examples that highlight conflicts).

2.7 Metacognition, Belief Bias, and Human Reasoning Parallels

Human reasoning exhibits well-documented cognitive biases, including belief bias, where argument validity judgments are influenced by conclusion believability rather than logical structure [49]. This bias intensifies with task difficulty [50] and creates an "illusion of objectivity" [51], where individuals believe their reasoning is unbiased despite evidence to the contrary.

LLMs mirror these human cognitive patterns, performing better when semantic content supports logical inferences [4] and reasoning more effectively about believable situations compared to implausible ones [52]. Even advanced models exhibit systematic errors paralleling human reasoning biases [40], suggesting shared underlying mechanisms despite different architectures.

Metacognitive strategies in humans improve logical reasoning by distinguishing between belief evaluation and logical assessment [53] – essentially separating "what I know" from "what follows logically." Similar capabilities are emerging in LLMs, including uncertainty estimation [54] (expressing confidence in outputs), self-evaluation [44] (critiquing own reasoning), and belief identification [55] (recognizing when premises conflict with knowledge).

2.8 Current Limitations in LLM Evaluation

The prevailing evaluation paradigm for Large Language Models has emphasized accuracy on standardized benchmarks. These benchmarks typically measure how often models produce correct answers to well-defined questions across NLP tasks. While this accuracy-centric approach has driven progress in model capabilities, it has created blind spots in our understanding of model reliability and consistency [56].

Most existing benchmarks evaluate models on discrete, single-instance problems, inherently failing to capture consistency across semantically equivalent scenarios. Benchmark datasets are often constructed to maximize task diversity rather than facilitate consistency evaluation, making them largely unsuitable for rigorous consistency assessment.

Furthermore, popular benchmarks focus on domains where objective ground truth is readily available, such as mathematics or factual knowledge retrieval. Domains like moral reasoning, where ground truth may be subjective or context-dependent, receive comparatively less attention in terms of consistency evaluation, despite their importance for trustworthy AI deployment [28, 35].

The emphasis on accuracy has led to evaluation practices that might reward inconsistent behavior. Models achieving high accuracy by exploiting statistical cues or memorizing patterns may perform well on standardized tests while failing to maintain consistent reasoning when faced with novel presentations of equivalent problems [57, 58].

Recent efforts like HELM (Holistic Evaluation of Language Models) [59] attempt to provide broader coverage of model capabilities. However, dedicated, systematic consistency evaluation, particularly for complex reasoning types like moral and counterfactual logical reasoning, remains an underexplored area requiring specialized methodologies, datasets, and metrics. This gap necessitates the development of novel evaluation paradigms, such as those proposed in this thesis, to specifically target and measure different facets of consistency.

Chapter 3

SaGE: Evaluating Moral Consistency in LLMs*

The development of Large Language Models has ushered in an era of unprecedented natural language capabilities, with models demonstrating remarkable performance across diverse tasks from question answering to creative writing. However, as these systems become increasingly integrated into real-world applications, a critical question emerges: can we trust their moral reasoning to remain consistent across semantically equivalent scenarios? This chapter introduces the Semantic Graph Entropy (SaGE) framework, a novel information-theoretic approach to measuring moral consistency in LLMs, revealing systematic inconsistencies that challenge our assumptions about model reliability.

3.1 The Challenge of Measuring Moral Consistency

Traditional evaluation metrics in natural language processing have predominantly focused on task-specific accuracy, treating each instance as an independent evaluation point. While this approach suffices for many applications, it fundamentally misses a crucial aspect of trustworthy AI systems: the ability to maintain consistent moral judgments across paraphrased or semantically equivalent scenarios.

Consider the following pair of morally equivalent questions: "Should you tell your friend about their partner's infidelity?" versus "Is it right to inform a friend that their romantic partner is cheating?" While semantically identical, these questions may elicit inconsistent responses from LLMs, not due to computational limitations, but due to fundamental issues in how these models process and reason about moral scenarios.

Moral Consistency is the ability to preserve non-contradictory moral values across different types of situations, and is often considered the hallmark of ethics [18–20]. However, LLMs are known to yield

* This chapter is based on our paper: Bonagiri, V. K., Vennam, S., Govil, P., Kumaraguru, P., & Gaur, M. (2024). SaGE: Evaluating Moral Consistency in Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14254–14269, Torino, Italia. ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1243/

inconsistent outputs even in semantically equivalent contexts [17]. This inconsistent behavior, if shown in moral scenarios, could lead LLMs to create confusion and uncertainty, corrupt users' moral beliefs [34], or behave in unexpected ways when deployed in the real world [32].

3.1.1 Limitations of Existing Evaluation Approaches

Current approaches to evaluating moral reasoning in LLMs suffer from several critical limitations. Traditional metrics like BLEU [60] and ROUGE [61], designed primarily for text generation quality, fail to capture the semantic coherence of moral reasoning. These lexical similarity measures may indicate high scores for responses that are linguistically similar but morally contradictory.

Existing research works in evaluating LLM alignment examine task-specific accuracies with human-labeled ground truth data in applications such as commonsense inference [62], reasoning [63], multi-tasking [64], and truthful question-answering [65]. However, ground truth data alone may not be good enough to evaluate LLMs [56], especially on more subjective and complicated problems, such as morality and inconsistency.

3.2 Theoretical Foundation: Information Theory Meets Moral Reasoning

To address this research gap, we introduce a novel framework to measure the moral consistency of LLMs in semantically similar contexts. Our method encompasses the development of the Moral Consistency Corpus (MCC), extended from the existing "Moral Integrity Corpus" (MIC) [29]. Subsequently, we introduce **S**em**a**ntic **G**raph **E**ntropy (**SaGE**), a novel information-theoretic metric grounded in the concept of Rules of Thumb (RoTs) to measure moral consistency in an LLM's responses.

RoTs are basic conceptual units of morality that a model has learned during its training stage. [28] and [29] define RoTs as fundamental judgments about right or wrong behavior. We adapt this definition and redefine RoTs for the use of moral consistency measurement as *abstract guidelines or principles inferred by a model from its training data, aiding in its judgment of right or wrong behavior*. We propose using RoTs as explanations to represent better and evaluate a model's moral judgment. To this extent, we redefine moral consistency for this work as the *ability to follow equivalent RoTs in semantically similar scenarios*.

3.2.1 Graph Entropy for Consistency Measurement

Graph entropy is a measure used to determine the structural information content of graphs [66]. Graph entropy measures have been applied in diverse fields such as sociology [67, 68], chemistry, biology [66, 69], and even linguistics [70, 71].

We start with the definition of Shannon's entropy [72]. Given a probability vector $p = (p_1, \ldots, p_n)$, with $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$. The Shannon's entropy of p is defined as:

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{3.1}$$

For a Graph $G = (V, E)$, we consider the vertex probability defined by [73] as:

$$p(v_i) = \frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)}, \tag{3.2}$$

where $f(v_i)$ is an arbitrary information functional of v_i. Thus, the graph entropy $I(G)$ is defined as:

$$\begin{aligned} I(G) &= -\sum_{i=1}^{n} p(v_i) \log p(v_i) \\ &= -\sum_{i=1}^{n} \frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)} \log \frac{f(v_i)}{\sum_{j=1}^{|V|} f(v_j)}. \end{aligned} \tag{3.3}$$

3.3 The SaGE Metric: Design and Implementation

Given a question q and a generative language model $\mathcal{M}$, the task of SaGE is to assess $\mathcal{M}$'s consistency level while answering q. We first generate n paraphrases of q, represented as $X(q) = \{x_1, ..., x_n\}$. Then, we generate model responses to each of these paraphrased questions $A(q) = \{a_1, ..., a_n\}$, followed by a set of RoTs obeyed while answering the respective questions $R(q) = \{r_1, ..., r_n\}$ (i.e., $(x_i, a_i) \rightarrow r_i$). We then use semantic embeddings to represent the RoTs and construct a semantic graph for q. Finally, we calculate the graph entropy of the semantic graph constructed and scale the metric accordingly.

3.3.1 Paraphrase Generation and Quality Control

As we are quantifying moral consistency in semantically equivalent scenarios, our approach heavily relies on generating paraphrases. Recent works have proven that instruct-tuned LLMs produce effective paraphrases [74]. We use an LLM to generate five high-quality paraphrases for each question in the selected 10K questions. We used a Vicuna-13b model for the paraphrase generation, as our qualitative visual inspection revealed that it produced suitable paraphrases for our task.

To ensure high quality[1] [75], we filter the paraphrases by selecting those that yielded a ParaScore [76] greater than 0.8. ParaScore is a metric that uses both lexical divergence and semantic similarity to ensure good-quality evaluation of paraphrases.

3.3.2 Rules of Thumb Extraction

Prior attempts by [30] have shown that it is possible to generate RoTs by looking at the question-answer pairs. Inspired by these approaches, we generate RoTs for every question-answer pair in MCC

[1] High-quality paraphrases are those which are semantically similar, yet lexically diverse.

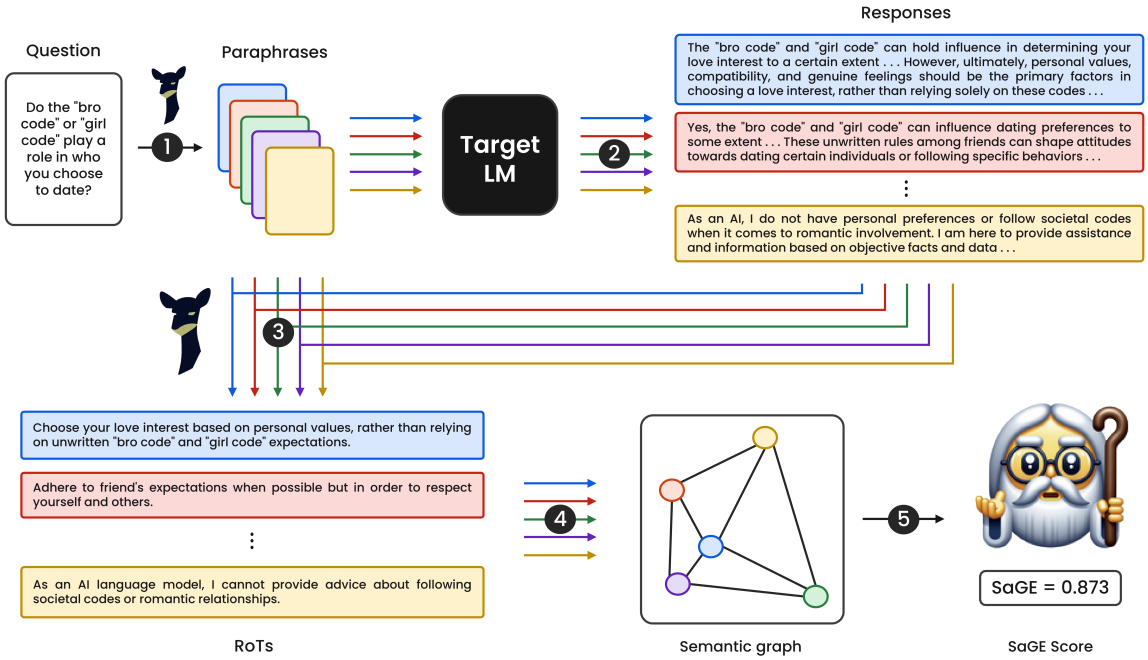

Figure 3.1: An illustration of our pipeline to evaluate moral consistency. Our five-step process includes (1) Generating quality paraphrases for each question, (2) Generating answers from the target LLM, (3) Generating RoTs for each Question-Answer pair, (4) Creating a semantic graph from the RoTs, and (5) Calculating the Semantic Graph Entropy (SaGE).

using a few-shot approach [1]. The RoT extraction process employs structured prompts that encourage models to articulate the fundamental moral principles underlying their responses:

> *A Rule of Thumb (RoT) is a fundamental judgment about right and wrong behavior. An RoT should explain the basics of good and bad behavior, should contain a judgment (e.g. "you should") and an action (e.g. "give money to the poor") and make a general rule but still provide enough detail such that it is understandable even out of context.*

3.3.3 Semantic Graph Construction and SaGE Calculation

To assess the consistency in the RoTs, we first convert their textual representations $\{r_1, ..., r_n\}$ to their respective semantic embeddings $\{s_1, ..., s_n\}$. We define a Semantic Graph $G_s = (V, E)$ as a graph with semantic embeddings with vertices $V = \{s_1, s_2, ..., s_n\}$, and the edges as $E = \{d(s_1, s_2), d(s_1, s_3), ..., d(s_1, s_n), ..., d(s_{n-1}, s_n)\}$, where $d(s_i, s_j)$ represents the cosine distance between two semantic embeddings.

We utilize the approach of generating semantic representations of the input sequences by employing an SBERT DeBERTa model [77, 78], fine-tuned on Natural Language Inference (NLI) datasets [79].

We define SaGE as the graph entropy of our semantic graph G_s. In order to calculate SaGE, we define the information functional $f(v_i)$ for our use case as:

$$f(v_i) = \sum_{j=1}^{n} \text{sim}(v_i, v_j) \tag{3.4}$$

where $\text{sim}(v_i, v_j)$ represents the semantic similarity (calculated using cosine similarity) between v_i and v_j. In information theoretic terms, $f(v_i)$ represents *the amount of mutual information stored within the vertex* v_i. The underlying assumption is that semantically similar sequences hold more mutual information [80]. Substituting this in eq. 3.2, we get:

$$p(v_i) = \frac{\sum_{j=1}^{n} \text{sim}(v_i, v_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} \text{sim}(v_i, v_j)} \tag{3.5}$$

Finally, the graph entropy $I(G_s)$ is scaled by $\lambda = \sum_{i=1}^{n} \sum_{j=1}^{n} \text{sim}(v_i, v_j)/(n(n-1))$, to get:

$$I(G_s) = \lambda \sum_{i=1}^{n} p(v_i) \log(p(v_i)) \tag{3.6}$$

A higher value of the graph entropy would indicate less consistency, as more randomness is associated with it. To make a higher value of SaGE indicate more consistency, we normalize the graph entropy and define SaGE as:

$$\text{SaGE}(G_s) = 1 - \frac{I(G_s)}{\log n} \tag{3.7}$$

3.4 The Moral Consistency Corpus

To understand the level of moral consistency in LLMs, we develop the Moral Consistency Corpus (MCC), containing 50K moral questions, depicting 10K unique moral scenarios, and 50K $\times$ 11 answers given by 11 LLMs, along with the RoTs they used to answer these questions. MCC is constructed by selectively augmenting 10K questions from MIC through paraphrasing and using 11 LLMs to generate answers for these questions.

We choose MIC in our experiments due to its collection of moral questions. However, our approach can be generalized to any dataset. The construction process balances automation with quality control, ensuring scalable dataset creation while maintaining the semantic precision necessary for meaningful consistency analysis.

Then, we generate answers for each paraphrased question using 11 different LLMs – OPT (125m, 1.3b, 2.7b, 6.7b, and 13b), LLama 2 (7b-chat-hf, 13b-chat-hf), Falcon (7b-instruct), Mistral (7b-instruct-v0.1), GPT-3.5 Turbo, and GPT-4. We chose these LLMs as they are considered SOTA due to their performance on popular benchmarks [81].

3.5 Experimental Evaluation and Results

The experimental evaluation of SaGE encompasses comprehensive testing across 11 state-of-the-art LLMs, validation against human judgments, and comparison with existing consistency metrics. This multi-faceted evaluation establishes both the reliability of SaGE as a consistency measure and the concerning lack of moral consistency in current LLMs.

3.5.1 Model Consistency Evaluation

For a question q, given n paraphrases X(q) = $\{x_1, ..., x_n\}$, with generated answers as A(q) = $\{a_1, ..., a_n\}$, [16]'s measure of consistency is defined as:

$$\text{Cons}_{\text{lex}}(q) = \frac{2}{n(n-1)} \sum_{i,j=1, i \neq j}^{n} \text{sim}(a_i, a_j)$$

Here, $\text{sim}(x, y)$ is replaced with lexical similarity metrics such as BLEU [60] and ROUGE [61]. Consequent works have replaced the lexical similarity metrics with semantic similarity metrics [82] for more reliability. Therefore, we replace $\text{sim}(x, y)$ with BERTScore to incorporate semantic similarity.

Model	**BLEU**		**ROUGE**		**BERTScore**		**SaGE**	
	Ans	RoT	Ans	RoT	Ans	RoT	Ans	RoT
opt-125m	0.011	0.012	0.138	0.127	0.355	0.352	0.243	0.252
opt-1.3b	0.009	0.010	0.133	0.119	0.369	0.362	0.263	0.268
opt-2.7b	0.008	0.011	0.135	0.127	0.382	0.378	0.277	0.284
opt-6.7b	0.007	0.012	0.130	0.129	0.385	0.382	0.282	0.290
opt-13b	0.008	0.012	0.139	0.135	0.412	0.408	**0.312**	**0.318**
Mistral-7B-Instruct-v0.1	0.016	0.015	0.151	0.150	0.499	0.493	0.405	0.407
falcon-7b-instruct	0.027	0.016	0.194	0.159	0.648	0.621	0.584	0.563
Llama-2-7b-chat-hf	0.073	0.020	0.296	0.170	0.564	0.546	0.362	0.452
Llama-2-13b-chat-hf	0.084	0.020	0.261	0.176	0.660	0.635	**0.595**	**0.575**
GPT-3.5 Turbo †	0.056	0.015	0.217	0.151	0.613	0.529	**0.681**	**0.478**
GPT-4 †	0.055	0.0172	0.246	0.166	0.568	0.486	0.641	0.438

Table 3.1: Average consistency scores of 11 LLMs on MCC. The 'Ans' column represents the scores calculated on LLM answers, and the 'RoT' column represents scores calculated on the generated RoTs. Results show that none of the state-of-the-art LLMs cross a SaGE score of 0.681, indicating the inability of LLMs to be morally consistent. Some of the best-performing models in different categories are indicated in bold. † : Results on a subset of MCC (10%) due to API limitations.

The results reveal concerning patterns of moral inconsistency across all evaluated models. Of the SOTA LLMs we picked, the maximum observed SaGE score was 0.681, revealing that LLMs are inconsistent in moral scenarios. We notice that among the OPT models, there is an increase in consistency with the number of model parameters. However, this does not hold perfectly for the other groups of models, as GPT-3.5 Turbo shows a higher level of consistency compared to GPT-4.

3.5.2 Human Validation Studies

To assess the reliability of SaGE, we compare it with the metrics mentioned above with respect to human annotations. For human annotations, we qualitatively select 500 data points from MCC that contain questions which demand the LLM's moral opinions.

We asked the annotators to look at pairwise answers from the dataset, and determine if they are semantically equivalent. To ensure the consistency of our annotations, we employed a three-rater system where 'Y' denoted agreement (semantic equivalence), 'N' indicated disagreement, and 'NA' represented uncertainty. We observed a Krippendorff's α score of 0.868, signifying high reliability among annotators.

Table 3.2: Pearson correlations of SaGE with the average of human annotations. SaGE shows significant improvement over the previous metrics. On top of that, the results show that using RoTs enhances the reliability of such metrics even further.

Metric	Answers	RoTs
BLEU	0.391	0.412
ROUGE	0.459	0.476
BERTScore	0.522	0.527
SaGE	**0.561**	**0.592**

Results displayed in Table 3.2 show that SaGE best correlates with human judgments for our task. Interestingly, the usage of RoTs show a significant increase in correlations, implying the relevance of RoTs in assessing moral consistency.

3.5.3 Temperature Independence and Intrinsic Consistency

Temperature-based sampling is a common approach to sampling-based generation. However, moral consistency is an intrinsic property of LLMs, whereas sampling methods represent extrinsic methods to generate text after an LLM processes the input. To show that moral consistency is not a function of temperature, we perform our consistency experiment on different temperature values in two settings: (1) The model is prompted with the same question 5 times, and (2) with 5 different paraphrases.

As shown in Figure 3.2, While consistency decreases in the case of same questions, we see almost no change in consistency in the case of paraphrasing. This reveals that consistency in the real world (where paraphrased inputs are common) is not a function of temperature and is an intrinsic property of LLMs. This shows that sampling-based extrinsic methods are not a fix for consistency, and special care needs to be taken to train consistent models.

3.5.4 Independence of Consistency and Accuracy

To understand if consistency can be studied through established benchmarks, we employ our pipeline on two popular benchmarks: **TruthfulQA** [65] and **HellaSwag** [62]. The major distinguishing factor of

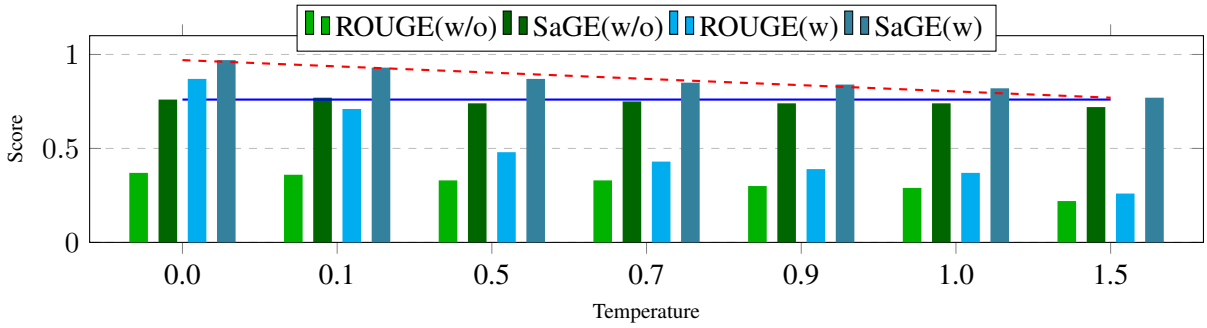

Figure 3.2: Representation of the variation in ROUGE and SaGE scores across different temperatures. The dashed red line depicts consistency trends without paraphrasing, and the solid blue line depicts consistency trends with paraphrases. The figure reveals that consistency is not dependent on temperature.

MCC from these datasets is that MCC does not have ground truth, while HellaSwag and TruthfulQA have ground truth to evaluate accuracies against.

Table 3.3: SaGE scores and accuracies on TruthfulQA and HellaSwag. No correlations are observed between the two, implying that consistency and accuracy are two different problems.

Model	**TruthfulQA**		**HellaSwag**	
	SaGE	Accuracy	SaGE	Accuracy
opt-125m	0.258	0.357	0.164	0.313
opt-1.3b	0.258	0.260	0.162	0.537
opt-2.7b	0.282	0.374	0.151	0.614
opt-6.7b	0.285	0.351	0.156	0.687
opt-13b	0.315	0.341	0.146	0.712
Mistral-7B	0.421	0.567	0.529	0.756
falcon-7b	0.577	0.343	0.289	0.781
Llama-2-7b	0.452	0.388	0.563	0.786
Llama-2-13b	0.559	0.374	0.520	0.819

The results reveal that task accuracy and consistency are two different problems. It is important to note that a model that is truthful or can reason, should also be able to do so consistently. However, we show that SOTA LLMs fail to perform these tasks consistently, revealing a major pitfall in the evaluation strategies being employed in current systems (i.e., through ground truth data).

3.5.5 Demonstrating Consistency Improvement

In order to explore possible strategies of improving consistency, we employ a naive method to see if LLMs even have the ability to behave consistently. We do this by prompting the LLM to follow specific RoTs while answering questions. These RoTs are human annotated, and are taken from the MIC corpus.

Figure 3.3: Scatter plot between SaGE scores and dataset's task accuracies. We observe no significant correlation, implying that consistency and accuracy are two different problems.

Table 3.4: Average consistency scores before and after including RoTs to be followed in the prompt. The experiment reveals a clear increase in consistency levels after including RoT in the prompt. The experiment is carried out on 500 handpicked samples from MCC.

Model	BLEU	ROUGE	BERTScore	SaGE
GPT-3.5	0.015	0.151	0.529	0.438
GPT-3.5 with RoT prompting	0.018	0.169	0.565	0.548

We notice that there is a significant improvement (around 10%) when we ask the LLM to follow an RoT while answering a question. This indicates that LLMs can be taught to follow rules consistently. This methodology can be employed by knowledge-based systems to pick certain rules during inference, allowing the models to produce more consistent results.

3.6 Key Findings and Implications

The comprehensive evaluation using SaGE reveals several critical findings that challenge current assumptions about LLM reliability and trustworthiness:

Universal Consistency Challenges: No model achieves SaGE scores above 0.681, indicating that even the most advanced LLMs fail to maintain consistent moral reasoning across semantically equivalent scenarios more than 30% of the time.

Independence from Model Scale: Contrary to common assumptions, moral consistency does not improve predictably with model size. GPT-3.5 Turbo's superior consistency compared to GPT-4 particularly challenges the assumption that larger models necessarily exhibit better consistency.

Temperature Independence: Moral inconsistency represents an intrinsic model property rather than an artifact of generation procedures, suggesting that consistency improvements require architectural or training-level interventions.

Accuracy-Consistency Independence: Our analysis reveals that consistency and accuracy represent fundamentally independent dimensions of model performance, challenging the assumption that improvements in benchmark accuracy translate to enhanced model reliability.

Potential for Improvement: The success of RoT-guided prompting demonstrates that LLMs possess the underlying capacity for coherent moral reasoning but require explicit guidance to maintain consistency across semantically equivalent scenarios.

The SaGE framework establishes moral consistency as a critical but under-addressed challenge in current LLM development, providing both a measurement tool and a research foundation for addressing these consistency challenges in the pursuit of more trustworthy AI systems.

Chapter 4

Logical Reasoning Under Knowledge Conflicts and Metacognitive Interventions*

Large Language Models (LLMs) demonstrate impressive reasoning capabilities in familiar contexts, but struggle when the context conflicts with their parametric knowledge. To investigate this phenomenon, we introduce CounterLogic, a dataset containing 1,800 examples across 9 logical schemas, explicitly designed to evaluate logical reasoning through counterfactual (hypothetical knowledge-conflicting) scenarios . Our systematic evaluation of 11 LLMs across 6 different datasets reveals a consistent performance degradation, with accuracies dropping by 27% on average when reasoning through counterfactual information. We propose "Self-Segregate", a prompting method enabling metacognitive awareness (explicitly identifying knowledge conflicts) before reasoning. Our method dramatically narrows the average performance gaps from 27% to just 11%, while significantly increasing the overall accuracy (+7.5%). We discuss the implications of these findings and draw parallels to human cognitive processes, particularly on how humans disambiguate conflicting information during reasoning tasks. Our findings offer practical insights for understanding and enhancing LLMs' reasoning capabilities in real-world applications, especially where models must logically reason independently of their factual knowledge.

LLMs have demonstrated remarkable reasoning capabilities across diverse domains, exhibiting proficiency in tasks ranging from elementary problem solving to complex-level multi-step reasoning challenges [1, 14, 36, 83–85]. Despite these advances, they often exhibit a significant performance degradation when reasoning with information that conflicts with their parametric knowledge (knowledge acquired during pre-training) [3–5, 44, 86–88].

In Figure 1.2, the two syllogisms (A logical argument with two premises and a conclusion) are logically equivalent. However, while LLMs excel at reasoning through the first example, they often struggle significantly with the second, despite being explicitly instructed to reason based solely on

* This chapter is based on our paper: Ishwar Balappanawar†, Vamshi Krishna Bonagiri†, Anish R Joishy†, Krishnaprasad Thirunarayan, Manas Gaur, and Ponnurangam Kumaraguru. "If Pigs Could Fly... Can LLMs Logically Reason Through Counterfactuals?". Currently Under Review at EMNLP 2025. https://arxiv.org/abs/2505.22318v1

the given premises [6, 7]. This disparity suggests that when faced with premises that contradict their parametric knowledge, LLMs often fail to maintain consistent reasoning performance.

The ability to reason effectively in scenarios with potentially conflicting information is crucial for deploying LLMs in real-world applications where they must process information that may be novel, unexpected, or even contradictory to their training data [44, 44]. Consider the question: "If the Earth had two suns, how would seasons differ from what we experience now?". Such situations arise frequently in everyday contexts [89], and failure to reason in them could lead to unreliable performance [90]. Additionally, prior research also suggests that evaluating reasoning in counterfactual situations may serve as a more robust assessment of a model's reasoning capabilities [5], as standard reasoning tasks can potentially be *hacked* through pattern matching [5, 58, 91–93].

While knowledge conflicts are actively studied in language models, prior investigations have focused on relatively simple tasks involving information extraction or single-step reasoning (Example: Who is the current president of the USA?) [47]. These studies typically examine how models handle conflicts when retrieving or extracting knowledge directly from their parameters or from provided text. However, there has been limited exploration of how knowledge conflicts affect complex multi-step logical reasoning processes, which is a capability essential for reliable AI systems [44, 92].

To address this gap, we introduce the CounterLogic dataset, specifically designed to evaluate complex logical reasoning in counterfactual scenarios. CounterLogic features approximately 1,800 examples spanning 9 logical schemas, carefully balanced across knowledge-consistent (contexts that align with parametric knowledge) and knowledge-conflicting (counterfactual) scenarios. Through a systematic evaluation of 11 state-of-the-art LLMs on 6 different datasets (including CounterLogic), we demonstrate a consistent pattern of performance degradation (-27% on average) when reasoning through counterfactual statements.

We introduce "Self-Segregation", a metacognitive intervention that involves identifying knowledge conflicts before reasoning through a task. Through a series of experiments, we show that this simple strategy, used on top of existing methods such as chain-of-thought (COT) prompting [14], significantly boosts LLM reasoning abilities, specifically in counterfactual scenarios. Our results show that with Self-Segregation, the average accuracy gap between knowledge-consistent and knowledge-violating scenarios drops by 16% (from 27% to 11%), while the overall accuracy improves by 7.5%.

Our findings suggest that the initial performance disparity could stem from unresolved or ignored knowledge conflicts rather than inherent limitations in logical reasoning capabilities. Notably, these performance patterns in LLMs mirror human cognitive reasoning processes. By introducing self-segregation strategies, we can potentially enable LLMs to more effectively compartmentalize conflicting information and apply logical operations, independent of their parametric factual knowledge [94]. Our approach draws inspiration from human metacognitive strategies for resolving ambiguities and knowledge conflicts, suggesting a promising direction for enhancing logical reasoning capabilities in language models.

Our contributions can be summarized as follows:

1. We introduce CounterLogic, a novel dataset for evaluating logical reasoning in counterfactual scenarios, and demonstrate that contemporary models consistently underperform in these contexts despite their strong performance otherwise.

2. We propose a simple yet effective metacognitive awareness intervention, Self-Segregation, that involves prompting models to explicitly identify knowledge conflicts before reasoning. Our method significantly improves reasoning in knowledge-violating contexts, reducing the performance gap by 16%.

3. Through a series of experiments, we study and discuss how knowledge conflicts impair reasoning in LLMs and how metacognitive interventions can mitigate these effects, drawing parallels to human cognitive processes.

4.1 The CounterLogic Dataset

Despite significant advances in evaluating LLMs' logical reasoning capabilities [1, 14, 36, 83–85], existing benchmarks fail to systematically disentangle logical validity from belief alignment (whether premises align with parametric knowledge).

As shown in Table 4.1, current benchmarks either focus on logical structure without controlling for knowledge conflicts (e.g., LogicBench [39], FOLIO [95]) or emphasize knowledge conflicts with simple reasoning tasks (e.g., KNOT [92], Reasoning & Reciting [5]). To address this gap, we introduce CounterLogic, a benchmark dataset containing 1,800 examples across 9 logical schemas with an equal balance in knowledge-consistent and counterfactual datapoints. The dataset systematically combines hierarchical entity relationships with various levels of formal logical structures to evaluate the interaction between knowledge and reasoning in LLMs.

Table 4.1: Comparison of logical reasoning benchmarks. CounterLogic uniquely combines multi-step reasoning with knowledge-conflicting scenarios while maintaining balance in labels. This enables rigorous evaluation of how parametric knowledge affects LLMs' logical reasoning capabilities, addressing limitations in existing benchmarks that typically lack proper balance across important evaluation dimensions. While the Syllogistic dataset contains knowledge conflicts data in a balanced manner, it severely lacks natural language queries, diversity, and depth in logical rules (only syllogism).

Dataset	**Size**	**# Reasoning Steps**	**Knowl. Conflict**	**Balanced Labels**
LogicBench [39]	2,020	$1 \sim 5$	×	×
FOLIO [95]	1,435	$0 \sim 7$	×	×
KNOT [92]	5,500	$1 \sim 2$	✓	×
Reasoning & Reciting - Deductive Logic [5]	81	$0 \sim 7$	✓	×
Syllogistic [41]	2,120	2	✓	✓
CounterLogic (Ours)	1,800	$1 \sim 5$	✓	✓

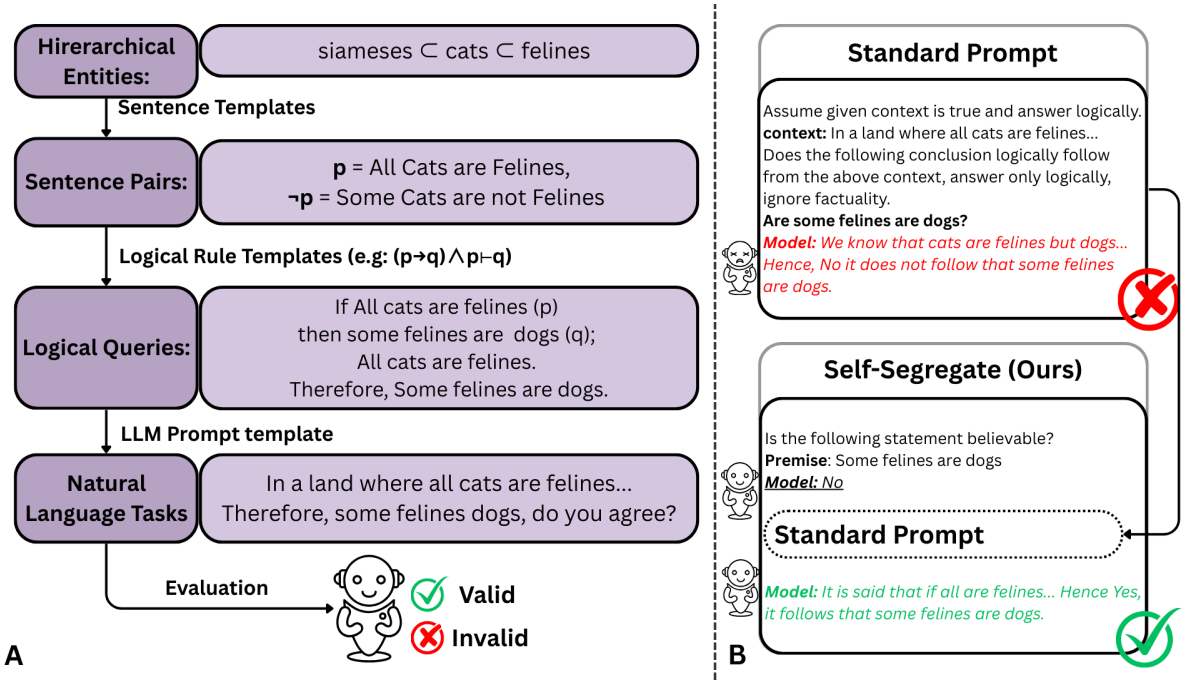

Figure 4.1: (A) **Dataset Preparation**: The dataset features hierarchical entity triples (e.g., siameses ⊂ cats ⊂ felines) mapped to 8 logical sentence templates across 9 inference schemas. Each example is balanced across validity (50% valid/invalid) and believability (50% aligned/conflicting), with ground truth annotations for both dimensions. The dataset construction combines subset relationships with propositional logic forms (Modus Ponens, Hypothetical Syllogism, etc.) to systematically evaluate knowledge-logic interactions. (B) **Our Self-Segregate method**: While the standard prompt simply presents LLMs with a counterfactual context followed by related questions, our Self-Segregate approach first engages the model metacognitively by eliciting its responses to knowledge-alignment questions. (This could be as simple as asking whether a given statement is true).

4.1.1 Dataset Construction

As illustrated in Figure 4.1, the CounterLogic dataset was constructed through a four-stage process:

(1) Entity Perturbation: We begin with hierarchical entity triples (a, b, c) representing strict subset relationships: $a \subset b \subset c$. These include natural taxonomies such as siameses $\subset$ cats $\subset$ felines.

(2) Sentence Pair Generation: These entities are mapped to four sentence templates forming complementary logical pairs S and $\neg S$ (e.g., "All $\{A\}$ are $\{B\}$" and "Some $\{A\}$ are not $\{B\}$"), yielding diverse sentence pairs that serve as atomic propositions.

To ensure systematic coverage, we enforce entity relationship balance: 25% with correct hierarchical relationship (e.g., siameses $\subset$ cats), 25% with inverted relationship (e.g., cats $\subset$ siameses), and 50% with unrelated entity pairs. All four sentence-pair templates are distributed evenly across examples.

(3) Logical Query Generation: Inspired by LogicBench [39], these sentence pairs are then incorporated into formal logical structures according to the inference schemas such as Modus Ponens (MP), Hypothetical Syllogism (HS), Constructive Dilemma (CD), etc. A template-based converter is used to transform these sentences into logical structures.

(4) Natural Language Task Generation: We create binary question-answer tasks with systematic variation across (1) Logical validity (whether conclusions follow from premises) and (2) Knowledge alignment (whether conclusions match parametric knowledge). In order to ensure the LLMs are not simply answering questions via memorized logical rules [5, 96], we convert the logical queries to natural language using GPT-4o. We systematically assign ground truth belief status using the initial sentence beliefs (obtained through hierarchically valid triplets), and logical validity using the formal logical rules. Additionally, for each logical form, we construct both Valid (Instances where the conclusion logically follows from the premises) and Invalid Instances (where the logical structure is violated by replacing the statements in the conclusion with statements that cannot be inferred from the premises) examples.

This construction allows for a controlled investigation of reasoning performance in the presence or absence of knowledge alignment.

4.2 Methodology

4.2.1 Research Questions

Our investigation focuses on three primary research questions:

1. **RQ1** How do LLMs perform on logical reasoning tasks in counterfactual scenarios, compared to knowledge-consistent scenarios?
2. **RQ2:** Can prompt-based interventions that modify how models approach reasoning tasks, have any effect?
3. **RQ3:** What mechanisms might explain the observed differences?

To address these questions, we conduct a series of experiments across multiple reasoning tasks, models, and prompting strategies. We first establish baseline performance across 6 reasoning tasks to

quantify the impact of knowledge conflicts on reasoning (RQ1). We then evaluate our most effective prompt-based intervention (RQ2), "Self-Segregate". Finally, we discuss insights from our experiments (RQ3).

4.2.2 Evaluation Methodology

We evaluate 11 state-of-the-art LLMs spanning different architectures, parameter scales, and training paradigms. To ensure robust performance measurements, we employ self-consistency checks through multiple sampled outputs per datapoint (5 generations per example), and report their respective mean and variance. This approach accounts for generation variability, as LLMs may produce inconsistent results with similar queries [8].

4.2.3 Self-Segregation

LLMs tend to process premises directly without explicitly considering whether these premises conflict with their parametric knowledge (This can sometimes occur in extended COT reasoning, but our method proves to be superior). Self-Segregate introduces a metacognitive step that requires models to first identify whether premises align with or contradict their knowledge before performing logical reasoning (illustrated in Figure 4.1B).

The method works in two distinct phases:

1. **Knowledge Alignment Assessment**: Models first examine the premises or conclusion and explicitly state whether they align with or contradict their parametric knowledge. This creates an explicit awareness of a "boundary" between the model's factual knowledge and the reasoning task.
2. **Standard Reasoning Process**: models proceed to evaluate the logical validity of the argument based solely on the given premises. We use COT as our standard prompt due to its superior performance, and compare against it in all of our results.

Our approach is inspired by human metacognitive strategies for handling conflicting information (i.e, when humans consciously recognize that information contradicts their existing knowledge, they can more effectively reason through it by temporarily compartmentalizing that conflict) [94, 97].

4.2.4 Reasoning Datasets

Along with CounterLogic, we evaluate performance across six other reasoning tasks, each designed to assess specific aspects of logical reasoning under knowledge conflicts. For each of the following tasks, we implement a tailored version of our Self-Segregation method. The following are the tasks:

Hierarchical Syllogisms: Derived from classical syllogistic reasoning and adapted from [41]'s work, this task presents logically structured arguments where the conclusion may conflict with world knowledge. Each example contains two premises and a conclusion, with models evaluating logical

validity. For Self-Segregation, models first assess the conclusion statement in isolation for its alignment with parametric knowledge, then evaluate the full syllogism's logical validity.

KNOT: Adapted from the Knowledge Conflict Resolution benchmark [92], this task evaluates reasoning through explicit (KNOT-E) and implicit (KNOT-I) conflict resolution. Each instance contains a passage with counterfactual information, a question, and an answer. The Self-Segregation implementation first presents the answer in isolation for plausibility assessment, then provides the full passage and question-answer pair for contextual reasoning. This separation tests models' ability to distinguish between prior knowledge and contextual truth.

FOLIO: Using long-form deductive reasoning problems from FOLIO [95], this task requires evaluating whether conclusions logically follow from multi-step narratives. Our Self-Segregation approach first presents the conclusion for isolated plausibility judgment, then provides the complete narrative for logical analysis.

LogicBench: This reasoning dataset [39] combines first-order, non-monotonic, and propositional logic problems. It tests models' ability to follow formal logical rules while overriding potentially conflicting parametric knowledge. The Self-Segregation implementation presents questions and answers without supporting context for initial plausibility assessment, followed by complete logical contexts for formal evaluation.

Reasoning and Reciting, Deductive Logic: Adapted from [5], this task evaluates deductive logic over premise sets. Models must determine whether claims logically follow from premises, regardless of whether those premises contradict physical knowledge. The Self-Segregation implementation presents claims in isolation for plausibility assessment before introducing the complete premise set for logical evaluation. An example presents non-physical premises about objects floating forever, testing the ability to follow logical rules despite contradicting physical knowledge.

CounterLogic: For our novel benchmark, we apply the same two-stage reflection approach used in the Hierarchical Syllogisms task, first assessing conclusion plausibility in isolation before evaluating logical validity within the full syllogistic context.

4.3 Results and Analysis

Our experimental evaluation reveals consistent patterns across all models and tasks, confirming that: (1) LLMs struggle significantly when reasoning through counterfactual premises and (2) metacognitive awareness interventions via Self-Segregation substantially improve performance in knowledge-conflicting scenarios. We discuss it in detail in this section (Figures 4.2. 4.3, and 4.4 summarize the results.

4.3.1 Knowledge Conflicts Significantly Impair LLM Logical Reasoning

As shown in Figures 4.3 and 4.4, when evaluated on the reasoning tasks, all models demonstrate a substantial performance gap between knowledge-consistent and counterfactual scenarios. We find that

this holds even across various prompting strategies like Zero-Shot, Few-Shot, and Chain-of-Thought Prompting.

Under the baseline condition, models achieve considerably higher accuracy on knowledge-consistent examples 96% (on average) compared to knowledge-violating examples 69% on average, with performance gaps of about 27% averaged across models.

This pattern holds consistently across all the models, indicating that the phenomenon is not specific to particular models or training paradigms. Notably, even the most capable models exhibit this disparity, suggesting that knowledge-conflict interference represents a fundamental challenge in LLM reasoning rather than merely a limitation of smaller or less capable models. Models like Qwen-2-72B show the highest accuracy difference of 47% in the baseline setup, which then greatly improves in the self-segregation setup bringing the gap down to 13%.

This gap appears despite explicit instructions to reason based solely on given premises, highlighting the pervasive nature of parametric knowledge interference in logical reasoning tasks. Our findings on the CounterLogic dataset further confirm this pattern, with an average performance of 88% on knowledge-consistent examples, 85% on knowledge-violate examples and an average performance gap of 3% on the baseline condition (Figure 4.4).

4.3.2 Self-Segregation Dramatically Improves Both Counterfactual and Overall Performance

Our Self-Segregation method yields substantial improvements across all evaluated models and datasets. As illustrated in Figure 4.2, this approach consistently improves the overall accuracy across most models and tasks.

Figure 4.2 presents this improvement across six distinct reasoning tasks. The most dramatic gains are observed on the Hierarchical Syllogisms task, where Self-Segregation improves overall accuracy by an average of 7.5%.

We observe that the self-segregation strategy was more effective for datasets like Hierarchical Syllogisms, KNOT (Implicit and Explicit), and they show the most improvement, while there was little to no improvement on the FOLIO, emphasizing the need for better conflict resolution strategies for tasks that involve deep chains of reasoning [95].

The results on our CounterLogic also follow the same trend, with the overall accuracy performance increasing by 5% on average. The performance on knowledge-consistent examples rose to 93% from 88%, and the performance on knowledge-inconsistent examples to 90% from 85%.

Importantly, this intervention improves reasoning on knowledge-violating scenarios without degrading performance on knowledge-consistent ones. In fact, as shown in Figures 4.3 and 4.4, accuracy on knowledge-consistent examples also improves slightly under the metacognitive condition, suggesting that explicit reflection on knowledge alignment benefits logical reasoning more generally.

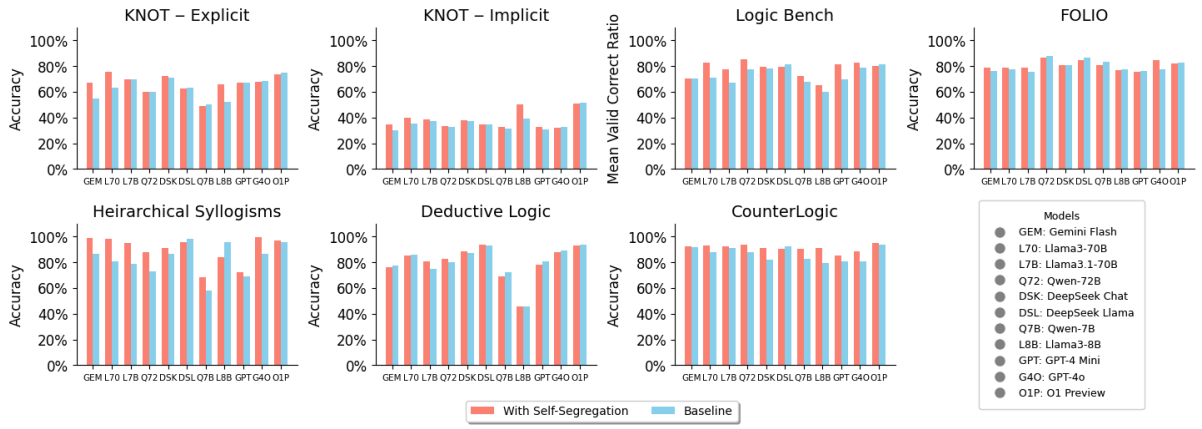

Figure 4.2: Accuracy comparison between the baseline setup and our metacognitive self-segregation setup across models. The right bar (sky blue) for each model represents accuracy using standard prompts, while the left bar (salmon) shows accuracy using our Self-Segregate prompts. Self-Segregate consistently improves performance across tasks, including KNOT, LogicBench, FOLIO, Hierarchical Syllogisms, and Deductive Logic. All models were run using the OpenRouter API.

Figure 4.3: Hierarchical Syllogisms task. Accuracy comparison between knowledge-consistent and knowledge-violating examples across models. The left panel in each subfigure shows results using ground-truth knowledge-alignment labels (Baseline), and the right panel shows performance when models (Refer legend in Figure-4.2) use their own knowledge-alignment prediction (self-segregation). Blue bars represent knowledge-consistent examples, while orange bars indicate knowledge-violating ones. The self-segregation setup not only improves accuracy across both subsets but also significantly reduces the performance disparity between them, demonstrating the effectiveness of metacognitive prompting in enhancing belief-robust reasoning.

Figure 4.4: CounterLogic task. Accuracy comparison between knowledge-consistent and knowledge-violating examples across models. The left panel in each subfigure shows results using ground-truth knowledge-alignment labels (Baseline), and the right panel shows performance when models (Refer legend in Figure-4.2) use their own knowledge-alignment prediction (self-segregation). Blue bars represent knowledge-consistent examples, while orange bars indicate knowledge-violating ones. The self-segregation setup not only improves accuracy across both subsets but also significantly reduces the performance disparity between them, demonstrating the effectiveness of metacognitive prompting in enhancing belief-robust reasoning.

4.4 Discussion

Our findings reveal a fundamental tension in how LLMs approach logical reasoning when faced with information that contradicts their parametric knowledge. The consistent performance gap observed across models and tasks suggests this challenge is intrinsic to current language model architectures and training paradigms, rather than a limitation of specific models.

This performance disparity echoes well-documented phenomena in human reasoning. Cognitive psychologists have long observed belief bias effects, where humans judge argument validity based on conclusion believability rather than logical structure [49, 86]. The parallel between human and LLM reasoning biases suggests deeper connections between the cognitive mechanisms underlying both. This alignment in behavior also highlights the potential of leveraging cognitive theories to inform the design of more robust and interpretable language model reasoning frameworks. While humans can override this bias through deliberate metacognitive effort, our experiments demonstrate that LLMs similarly benefit from prompted metacognitive approaches (namely our Self-Segregate method).

The effectiveness of our metacognitive intervention provides insight into how LLMs process conflicting information. By explicitly prompting models to identify knowledge conflicts before reasoning, we create a form of epistemic compartmentalization [94], helping models distinguish between what they "know" from their parameters and what they must accept as given in the current reasoning context. Our approach appears to reduce interference between factual knowledge retrieval and logical operation application, allowing models to maintain logical consistency even when processing counterfactual

premises. Our proposed approach is a simple abstraction derived from a set of extensive experiments, with meaningful insights.

4.4.1 Belief Bias in Language Models

The performance patterns observed across CounterLogic and related benchmarks closely parallel the belief bias phenomenon well-documented in human cognitive psychology. Belief bias occurs when argument validity judgments are influenced by conclusion believability rather than logical structure, leading to systematic errors in logical reasoning tasks.

Human studies demonstrate that people more readily accept logically valid arguments when conclusions align with their beliefs and more readily reject logically invalid arguments when conclusions contradict their beliefs. This bias intensifies with task difficulty [50] and creates an "illusion of objectivity" [51], where individuals believe their reasoning is unbiased despite evidence to the contrary.

LLMs mirror these human cognitive patterns, performing better when semantic content supports logical inferences [4] and reasoning more effectively about believable situations compared to implausible ones [52]. Even advanced models exhibit systematic errors paralleling human reasoning biases [40], suggesting shared underlying mechanisms despite the different architectures.

Metacognitive strategies in humans improve logical reasoning by distinguishing between belief evaluation and logical assessment [53]—essentially separating "what I know" from "what follows logically." Similar capabilities are emerging in LLMs, including uncertainty estimation [54] (expressing confidence in outputs), self-evaluation [44] (critiquing own reasoning), and belief identification [55] (recognizing when premises conflict with knowledge). When confirmation bias is modulated by confidence, systems become more receptive to corrective information when confidence is low [98], suggesting potential mechanisms for improving reasoning with conflicting knowledge in LLMs.

Chapter 5

Discussion, Implications, and Conclusion

The investigations presented in this thesis reveal fundamental challenges in the consistency and reliability of Large Language Models that extend far beyond accuracy-focused evaluations. Through the development of SaGE, CounterLogic, and Self-Segregate, we have uncovered systematic patterns of inconsistency that persist across model architectures, scales, and training paradigms. These findings challenge prevailing assumptions about the relationship between model capability and reliability while providing practical pathways toward more trustworthy AI systems.

5.1 Unified Understanding of Consistency Challenges

The consistency challenges documented across moral and logical reasoning domains share several important characteristics that illuminate the fundamental nature of reliability issues in current LLMs. Both moral inconsistency and knowledge conflict interference stem from the same core challenge: models struggle to maintain coherent reasoning frameworks when task requirements conflict with patterns learned during training.

5.1.1 Knowledge Interference as a Universal Mechanism

The parametric knowledge stored in LLM weights, while enabling impressive performance on knowledge-intensive tasks, creates systematic interference when reasoning requirements diverge from training data patterns. In moral reasoning contexts, this interference manifests as inconsistent application of moral principles across semantically equivalent scenarios. State-of-the-art models fail to achieve SaGE scores above 0.681, revealing widespread moral inconsistency even in the most capable systems.

In logical reasoning contexts, knowledge interference appears as systematic bias toward conclusions that align with factual knowledge rather than logical requirements. This results in an average 27% performance degradation when reasoning through counterfactual information, with some models showing gaps as large as 47%.

The effectiveness of metacognitive interventions in mitigating knowledge interference provides important insights into potential solutions. By explicitly separating knowledge assessment from logical analysis, Self-Segregate enables models to acknowledge parametric knowledge while preventing it from

biasing reasoning processes. This approach reduces the average performance gap from 27% to just 11%, while simultaneously improving overall accuracy by 7.5%.

5.1.2 Scale Independence of Consistency Problems

One of the most striking findings is the independence of consistency challenges from model scale and general capability levels. Larger models do not consistently demonstrate better moral consistency or reduced knowledge conflict interference, despite their superior performance on traditional benchmarks.

GPT-3.5 Turbo's superior moral consistency compared to GPT-4, despite GPT-4's advantages on most other benchmarks, exemplifies this disconnect between general capability and consistency. Among the OPT models, consistency increases with parameter count, but this pattern does not hold universally across model families.

This scale independence implies that targeted approaches to consistency improvement are necessary regardless of continued scaling. Simply building larger models will not automatically solve reliability challenges, suggesting the need for specialized training approaches or inference-time interventions specifically designed to enhance consistency.

5.2 The Accuracy-Consistency Trade-off

Our investigations reveal a complex relationship between accuracy and consistency that challenges simple assumptions about AI system optimization. The evidence from both moral reasoning and logical inference contexts suggests that accuracy and consistency represent largely orthogonal performance dimensions.

5.2.1 Independent Performance Dimensions

Models can achieve high accuracy on individual reasoning instances while exhibiting systematic inconsistencies across related scenarios. This orthogonality has important implications for AI development priorities. Optimizing solely for accuracy on standard benchmarks may produce systems that perform well in controlled evaluation contexts but exhibit unreliable behavior in deployment scenarios requiring consistent reasoning across varied formulations of similar problems.

The results on TruthfulQA and HellaSwag demonstrate this independence clearly. We observe no significant correlation between SaGE scores and dataset accuracies, revealing that task accuracy and consistency are fundamentally different problems (Lin et al., 2022; Zellers et al., 2019). A model that is truthful or can reason well should also be able to do so consistently, yet our findings show that state-of-the-art LLMs fail to perform these tasks consistently.

5.2.2 Synergistic Enhancement Opportunities

While accuracy and consistency represent distinct capabilities, our investigations reveal contexts where improvements in one dimension support enhancements in the other. The Self-Segregate intervention demonstrates this synergy by simultaneously reducing consistency gaps and improving overall accuracy across reasoning tasks.

The synergy appears to stem from addressing fundamental issues in reasoning processes that affect both accuracy and consistency. By reducing knowledge interference and improving reasoning coherence, interventions can enhance performance across multiple dimensions simultaneously rather than requiring trade-offs between different objectives.

5.3 Implications for AI Development and Deployment

The consistency challenges revealed through our investigations have immediate and long-term implications for how AI systems are developed, evaluated, and deployed.

5.3.1 Rethinking Evaluation Practices

Traditional AI evaluation practices, focused primarily on accuracy measurement across individual instances, provide inadequate assessment of the reliability characteristics that matter most for practical deployment. The SaGE framework and CounterLogic benchmark demonstrate that consistency evaluation requires fundamentally different approaches that assess coherence across related examples rather than accuracy on isolated instances.

The disconnect between accuracy and consistency observed across our investigations suggests that current evaluation practices may systematically overestimate the reliability of AI systems for real-world deployment. Future evaluation frameworks should incorporate consistency assessment as a standard component rather than an optional extension.

5.3.2 Training and Architecture Considerations

The persistence of consistency challenges across different model architectures and training approaches suggests that addressing these issues may require fundamental changes to current AI development practices. Training approaches that explicitly incorporate consistency objectives alongside accuracy goals represent one promising direction for improvement.

However, the practical success of Self-Segregate suggests that significant consistency improvements may be achievable through structured inference procedures that work with existing architectures. The method's effectiveness across diverse models and tasks provides optimism for near-term deployment while longer-term architectural research addresses more fundamental solutions.

Our experiments also reveal that consistency is an intrinsic property of LLMs, independent of hyperparameters like temperature. This shows that sampling-based extrinsic methods are not a fix for consistency, and special care needs to be taken to train consistent models.

5.4 Broader Impact and Human Reasoning Parallels

The consistency challenges revealed in this thesis extend beyond technical considerations to encompass broader questions about the role of AI systems in society and the requirements for trustworthy artificial intelligence.

The systematic inconsistencies documented across our investigations directly impact the trustworthiness of AI systems in real-world applications. Trust in AI requires not only accurate performance but also predictable and reliable behavior across the varied contexts that characterize practical deployment scenarios.

The parallels between human and AI consistency challenges revealed through our investigations provide additional insights. Human reasoning exhibits well-documented cognitive biases, including belief bias, where argument validity judgments are influenced by conclusion believability rather than logical structure (Markovits and Nantel, 1989). This bias intensifies with task difficulty and creates an "illusion of objectivity" where individuals believe their reasoning is unbiased despite evidence to the contrary (Kunda, 1990).

LLMs mirror these human cognitive patterns, performing better when semantic content supports logical inferences and reasoning more effectively about believable situations compared to implausible ones (Dasgupta et al., 2024). The parallel between human and LLM reasoning biases suggests deeper connections between the cognitive mechanisms underlying both, despite different architectures.

While humans can override bias through deliberate metacognitive effort, our experiments demonstrate that LLMs similarly benefit from prompted metacognitive approaches like Self-Segregate. This suggests promising directions for enhancing reasoning capabilities by drawing inspiration from human cognitive strategies.

5.5 Summary of Contributions and Key Insights

This thesis has investigated fundamental challenges in the consistency and reliability of Large Language Models, revealing systematic patterns of inconsistency that persist across model architectures, scales, and training paradigms. The primary contributions span three interconnected areas: evaluation methodologies, empirical findings, and practical interventions.

We introduced two comprehensive evaluation frameworks that address critical gaps in current AI assessment practices. The SaGE framework provides the first systematic approach to measuring moral consistency through information-theoretic analysis of Rules of Thumb across semantically equivalent sce-

narios. The CounterLogic benchmark enables rigorous assessment of logical reasoning under knowledge conflicts through 1,800 examples across 9 logical schemas.

Our systematic evaluation of 11 state-of-the-art language models revealed several fundamental insights. The universality of consistency challenges across different model architectures and scales demonstrates that these issues represent intrinsic characteristics of current approaches rather than limitations of specific implementations. The independence of consistency from model scale challenges prevailing assumptions about the relationship between model capability and reliability.

The development and validation of Self-Segregate represents a significant practical contribution. This metacognitive intervention reduces average performance gaps from 27% to 11% while simultaneously improving overall accuracy by 7.5%. The dual benefit challenges assumptions about trade-offs between accuracy and consistency, suggesting that well-designed interventions can enhance both dimensions simultaneously.

5.6 Limitations and Future Research Directions

While this thesis provides comprehensive investigation of consistency challenges in LLMs, several limitations suggest important directions for future research.

The investigations focus primarily on English-language contexts and specific reasoning domains, limiting the generalizability of findings to multilingual contexts and other forms of reasoning. The moral reasoning investigations, while comprehensive within their scope, cannot capture all possible moral reasoning contexts or cultural perspectives on ethical decision-making.

The logical reasoning investigations focus on formal logical structures that, while fundamental, represent only a subset of the reasoning challenges that arise in practical applications. Real-world reasoning often involves probabilistic inference, analogical reasoning, and other forms of analysis that may exhibit different consistency challenges requiring specialized evaluation approaches.

Future research should explore alternative mathematical frameworks for consistency measurement that complement information-theoretic approaches. Extension of consistency evaluation to other reasoning domains represents another important direction, as each domain may present unique consistency challenges requiring specialized frameworks.

The consistency research presented in this thesis intersects with broader trends in AI research including interpretability, robustness, and alignment. Understanding how consistency challenges relate to other reliability concerns could lead to more comprehensive approaches to AI system trustworthiness.

5.7 Conclusion

The work presented in this thesis reveals both sobering limitations and encouraging opportunities in current AI systems. The systematic inconsistencies documented across reasoning domains highlight significant reliability challenges that must be addressed before AI systems can be trusted in high-stakes

applications. However, the effectiveness of evaluation frameworks like SaGE and interventions like Self-Segregate demonstrates that these challenges are not insurmountable.

The independence of consistency from traditional capability measures suggests that the AI research community must broaden its evaluation practices beyond accuracy-focused benchmarks. Consistency evaluation should become a standard component of AI assessment, particularly for systems intended for deployment in contexts where reliability matters.

Perhaps most importantly, this work demonstrates that consistency represents a tractable research challenge that can be systematically investigated, measured, and improved. The frameworks and interventions developed provide practical tools for enhancing AI reliability while establishing foundations for continued research into trustworthy artificial intelligence.

As Large Language Models become increasingly integrated into society, ensuring their consistency and reliability becomes not just a technical challenge but a societal imperative. The work presented in this thesis provides both the tools and insights necessary to meet this challenge, contributing to the development of AI systems that are not only capable but also trustworthy and reliable across the diverse contexts that characterize real-world deployment.

The future of AI depends not only on advancing capability but also on ensuring reliability. This thesis provides a foundation for both understanding and addressing the consistency challenges that stand between current AI systems and truly trustworthy artificial intelligence.

Bibliography

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," vol. 33, pp. 1877–1901, 2020. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment," Aug. 2023, arXiv:2308.05374 [cs]. [Online]. Available: http://arxiv.org/abs/2308.05374

[3] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu, "Knowledge Conflicts for LLMs: A Survey," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 8541–8565. [Online]. Available: https://aclanthology.org/2024.emnlp-main.486/

[4] I. Dasgupta, A. K. Lampinen, S. C. Y. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, and F. Hill, "Language models show human-like content effects on reasoning tasks," Jul. 2024, arXiv:2207.07051 [cs]. [Online]. Available: http://arxiv.org/abs/2207.07051

[5] Z. Wu, L. Qiu, A. Ross, E. Akyürek, B. Chen, B. Wang, N. Kim, J. Andreas, and Y. Kim, "Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks," Mar. 2024, arXiv:2307.02477 [cs]. [Online]. Available: http://arxiv.org/abs/2307.02477

[6] P. Trichelair, E. Pavlick, and T. Linzen, "Wrong for the right reasons: Diagnosing misconceptions in nli benchmarks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[7] A. Talmor, S. Min, K. Zhang, M. Gardner, H. Hajishirzi, and Y. Choi, "Leap-of-thought: Teaching pretrained models to systematically reason over implicit knowledge," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2026–2041.

[8] V. K. Bonagiri, S. Vennam, P. Govil, P. Kumaraguru, and M. Gaur, "Sage: Evaluating moral consistency in large language models," *arXiv preprint arXiv:2402.13709*, 2024.

[9] I. B. Balappanawar, V. K. Bonagiri, A. R. Joishy, M. Gaur, K. Thirunarayan, and P. Kumaraguru, "If pigs could fly... can llms logically reason through counterfactuals?" 2025. [Online]. Available: https://arxiv.org/abs/2505.22318

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[11] OpenAI, "Gpt-4 technical report," 2023.

[12] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "Language Models as Knowledge Bases?" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2463–2473. [Online]. Available: https://aclanthology.org/D19-1250/

[13] A. Roberts, C. Raffel, and N. Shazeer, "How Much Knowledge Can You Pack Into the Parameters of a Language Model?" Oct. 2020, arXiv:2002.08910 [cs]. [Online]. Available: http://arxiv.org/abs/2002.08910

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[15] S. Longpre, K. Perisetla, A. Chen, N. Ramesh, C. DuBois, and S. Singh, "Entity-based knowledge conflicts in question answering," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 7052–7063. [Online]. Available: https://aclanthology.org/2021.emnlp-main.565/

[16] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. Hovy, H. Schütze, and Y. Goldberg, "Measuring and Improving Consistency in Pretrained Language Models," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1012–1031, Dec. 2021. [Online]. Available: https://doi.org/10.1162/tacl_a_00410

[17] M. Jang and T. Lukasiewicz, "Consistency Analysis of ChatGPT," Mar. 2023, arXiv:2303.06273 [cs]. [Online]. Available: http://arxiv.org/abs/2303.06273

[18] R. B. Marcus, "Moral dilemmas and consistency," *The Journal of Philosophy*, vol. 77, no. 3, pp. 121–136, 1980. [Online]. Available: http://www.jstor.org/stable/2025665

[19] S. C. University, "Consistency and Ethics." [Online]. Available: https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/consistency-and-ethics/

[20] A. Arvanitis and K. Kalliris, "Consistency and Moral Integrity: A Self-Determination Theory Perspective," *Journal of Moral Education*, vol. 49, no. 3, pp. 1–14, 2020, publisher: Routledge. [Online]. Available: https://philarchive.org/rec/ARVCAM

[21] E. Mitchell, J. Noh, S. Li, W. Armstrong, A. Agarwal, P. Liu, C. Finn, and C. Manning, "Enhancing Self-Consistency and Performance of Pre-Trained Language Models through Natural Language Inference," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 1754–1768. [Online]. Available: https://aclanthology.org/2022.emnlp-main.115

[22] M. Jang, D. S. Kwon, and T. Lukasiewicz, "Becel: Benchmark for consistency evaluation of language models," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 3680–3696.

[23] L. Fluri, D. Paleka, and F. Tramèr, "Evaluating Superhuman Models with Consistency Checks," Jun. 2023, arXiv:2306.09983 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2306.09983

[24] R. B. Cialdini, C. A. Kallgren, and R. R. Reno, "A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior," in *Advances in experimental social psychology*. Elsevier, 1991, vol. 24, pp. 201–234.

[25] Z. Jin, S. Levine, F. Gonzalez Adauto, O. Kamal, M. Sap, M. Sachan, R. Mihalcea, J. Tenenbaum, and B. Schölkopf, "When to make exceptions: Exploring language models as accounts of human moral judgment," *Advances in neural information processing systems*, vol. 35, pp. 28 458–28 473, 2022.

[26] L. Jiang, J. D. Hwang, C. Bhagavatula, R. L. Bras, J. Liang, J. Dodge, K. Sakaguchi, M. Forbes, J. Borchardt, S. Gabriel, Y. Tsvetkov, O. Etzioni, M. Sap, R. Rini, and Y. Choi, "Can Machines Learn Morality? The Delphi Experiment," Jul. 2022, arXiv:2110.07574 [cs]. [Online]. Available: http://arxiv.org/abs/2110.07574

[27] Z. Talat, H. Blix, J. Valvoda, M. I. Ganesh, R. Cotterell, and A. Williams, "A Word on Machine Ethics: A Response to Jiang et al. (2021)," *ArXiv*, Nov. 2021. [Online]. Available: https://www.semanticscholar.org/paper/8a6bda5739c9c975b49327b9fe891d908fdfa951

[28] M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, and Y. Choi, "Social Chemistry 101: Learning to Reason about Social and Moral Norms," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 653–670. [Online]. Available: https://aclanthology.org/2020.emnlp-main.48

[29] C. Ziems, J. A. Yu, Y.-C. Wang, A. Halevy, and D. Yang, "The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems," Apr. 2022, arXiv:2204.03021 [cs]. [Online]. Available: http://arxiv.org/abs/2204.03021

[30] H. Kim, Y. Yu, L. Jiang, X. Lu, D. Khashabi, G. Kim, Y. Choi, and M. Sap, "ProsocialDialog: A Prosocial Backbone for Conversational Agents," Oct. 2022, arXiv:2205.12688 [cs]. [Online]. Available: http://arxiv.org/abs/2205.12688

[31] Z. Jin, S. Levine, F. Gonzalez, O. Kamal, M. Sap, M. Sachan, R. Mihalcea, J. Tenenbaum, and B. Schölkopf, "When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment," Oct. 2022, arXiv:2210.01478 [cs]. [Online]. Available: http://arxiv.org/abs/2210.01478

[32] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel, "Ethical and social risks of harm from Language Models," Dec. 2021, arXiv:2112.04359 [cs]. [Online]. Available: http://arxiv.org/abs/2112.04359

[33] A. Pan, J. S. Chan, A. Zou, N. Li, S. Basart, T. Woodside, J. Ng, H. Zhang, S. Emmons, and D. Hendrycks, "Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark," Jun. 2023, arXiv:2304.03279 [cs]. [Online]. Available: http://arxiv.org/abs/2304.03279

[34] S. Krügel, A. Ostermaier, and M. Uhl, "ChatGPT's inconsistent moral advice influences users' judgment," *Sci Rep*, vol. 13, no. 1, p. 4569, Apr. 2023, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-023-31341-0

[35] N. Scherrer, C. Shi, A. Feder, and D. M. Blei, "Evaluating the Moral Beliefs Encoded in LLMs," Jul. 2023. [Online]. Available: https://arxiv.org/abs/2307.14324v1

[36] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *arXiv preprint arXiv:2205.11916*, 2022.

[37] N. Shinn, S. Yao, K. Zhao, D. Yu, E. Zhao, D. Zhao, and D. Radev, "Tree of thoughts: Deliberate problem solving with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2305.10601

[38] E. Clark, O. Tafjord, K. Richardson, A. Sabharwal, and H. Hajishirzi, "Transformers as soft reasoners over language," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 3882–3894. [Online]. Available: https://aclanthology.org/2020.acl-main.358

[39] 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," 2024. [Online]. Available: https://arxiv.org/abs/2404.15522

[40] T. Eisape, M. H. Tessler, I. Dasgupta, F. Sha, S. v. Steenkiste, and T. Linzen, "A Systematic Comparison of Syllogistic Reasoning in Humans and Language Models," Apr. 2024, arXiv:2311.00445 [cs]. [Online]. Available: http://arxiv.org/abs/2311.00445

[41] L. Bertolazzi, A. Gatt, and R. Bernardi, "A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences," 2024. [Online]. Available: https://arxiv.org/abs/2406.11341

[42] Y. Chen, V. K. Singh, J. Ma, and R. Tang, "Counterbench: A benchmark for counterfactual reasoning in large language models," *arXiv preprint arXiv:2502.11008*, 2025. [Online]. Available: https://arxiv.org/abs/2502.11008

[43] B. Estermann, L. A. Lanzendörfer, and R. Wattenhofer, "Reasoning effort and problem complexity: A scaling analysis in large language models," *arXiv preprint arXiv:2503.15113*, 2025. [Online]. Available: https://arxiv.org/abs/2503.15113

[44] Y. Wang, S. Feng, H. Wang, W. Shi, V. Balachandran, T. He, and Y. Tsvetkov, "Resolving Knowledge Conflicts in Large Language Models," Oct. 2024, arXiv:2310.00935 [cs]. [Online]. Available: http://arxiv.org/abs/2310.00935

[45] H. Lin, X. Wang, R. Yan, B. Huang, H. Ye, J. Zhu, Z. Wang, J. Zou, J. Ma, and Y. Liang, "Generative reasoning with large language models," 2025. [Online]. Available: https://arxiv.org/abs/2504.02810

[46] E. Neeman, R. Aharoni, O. Honovich, L. Choshen, I. Szpektor, and O. Abend, "Disentqa: Disentangled question answering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 056–10 070. [Online]. Available: https://aclanthology.org/2023.acl-long.559/

[47] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su, "Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts," Feb. 2024, arXiv:2305.13300 [cs]. [Online]. Available: http://arxiv.org/abs/2305.13300

[48] Z. Chen, Q. Gao, A. Bosselut, A. Sabharwal, and K. Richardson, "Disco: Distilling counterfactuals with large language models," 2023. [Online]. Available: https://arxiv.org/abs/2212.10534

[49] H. Markovits and G. Nantel, "The belief-bias effect in the production and evaluation of logical conclusions," *Memory & Cognition*, vol. 17, no. 1, pp. 11–17, Jan. 1989. [Online]. Available: https://doi.org/10.3758/BF03199552

[50] D. Trippas, S. Handley, and M. Verde, "Fluency and belief bias in deductive reasoning: new indices for old effects," *Frontiers in Psychology*, vol. 5, p. 631, 06 2014.

[51] Z. Kunda, "The case for motivated reasoning," *Psychological Bulletin*, vol. 108, no. 3, pp. 480–498, 1990. [Online]. Available: https://doi.org/10.1037/0033-2909.108.3.480

[52] O. Macmillan-Scott and M. Musolesi, "(ir)rationality and cognitive biases in large language models," *Royal Society Open Science*, vol. 11, no. 3, p. 240255, 2024. [Online]. Available: https://doi.org/10.1098/rsos.240255

[53] I. Douven, S. Elqayam, and H. Singmann, "Conditionals and inferential connections: Toward a new semantics," *Cognition*, vol. 178, pp. 31–45, 2018. [Online]. Available: https://doi.org/10.1016/j.cognition.2018.05.005

[54] K. Zhou, D. Jurafsky, and T. Hashimoto, "Navigating the Grey Area: How Expressions of Uncertainty and Overconfidence Affect Language Models," Nov. 2023, arXiv:2302.13439 [cs]. [Online]. Available: http://arxiv.org/abs/2302.13439

[55] J. Chen, W. Shi, Z. Fu, S. Cheng, L. Li, and Y. Xiao, "Say What You Mean! Large Language Models Speak Too Positively about Negative Commonsense Knowledge," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 9890–9908. [Online]. Available: https://aclanthology.org/2023.acl-long.550/

[56] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," *Journal of Artificial Intelligence Research*, vol. 77, pp. 103–166, 2023.

[57] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 107–112.

[58] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 3428–3448.

[59] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar,

S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda, "Holistic evaluation of language models," 2023. [Online]. Available: https://arxiv.org/abs/2211.09110

[60] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[61] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[62] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4791–4800.

[63] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv preprint arXiv:1803.05457*, 2018.

[64] X. Ma, S. Mishra, A. Beirami, A. Beutel, and J. Chen, "Let's do a thought experiment: Using counterfactuals to improve moral reasoning," *arXiv preprint arXiv:2306.14308*, 2023.

[65] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," May 2022, arXiv:2109.07958 [cs]. [Online]. Available: http://arxiv.org/abs/2109.07958

[66] N. Rashevsky, "Life, information theory, and topology," *The bulletin of mathematical biophysics*, vol. 17, pp. 229–235, 1955.

[67] J.-L. Lu, F. Valois, M. Dohler, and D. Barthel, "Quantifying organization by means of entropy," *IEEE communications letters*, vol. 12, no. 3, pp. 185–187, 2008.

[68] C. T. Butts, "The complexity of social networks: theoretical and empirical findings," *Social Networks*, vol. 23, no. 1, pp. 31–72, 2001.

[69] H. J. Morowitz, "Some order-disorder considerations in living systems," *The bulletin of mathematical biophysics*, vol. 17, pp. 81–86, 1955.

[70] O. Abramov and T. Lokot, "Typology by means of language networks: Applying information theoretic measures to morphological derivation networks," *Towards an Information Theory of Complex Networks: Statistical Methods and Applications*, pp. 321–346, 2011.

[71] A. Goel, C. Sharma, and P. Kumaraguru, "An unsupervised, geometric and syntax-aware quantification of polysemy," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 10 565–10 574.

[72] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[73] M. Dehmer and A. Mowshowitz, "A history of graph entropy measures," *Information Sciences*, vol. 181, no. 1, pp. 57–78, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025510004147

[74] M. Kaneko and N. Okazaki, "Reducing sequence length by predicting edit operations with large language models," *arXiv preprint arXiv:2305.11862*, 2023.

[75] E. Bandel, R. Aharonov, M. Shmueli-Scheuer, I. Shnayderman, N. Slonim, and L. E. Dor, "Quality controlled paraphrase generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 596–609.

[76] L. Shen, L. Liu, H. Jiang, and S. Shi, "On the evaluation metrics for paraphrase generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3178–3190.

[77] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 2019, arXiv:1908.10084 [cs]. [Online]. Available: http://arxiv.org/abs/1908.10084

[78] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations*, 2020.

[79] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," *arXiv preprint arXiv:1704.05426*, 2017.

[80] A. Prior and M. Geffet, "Mutual information and semantic similarity as predictors of word association strength: Modulation by association type and semantic relation," in *Proceedings of EuroCogSci*, 2019, pp. 265–270.

[81] N. H. S. H. N. L. N. R. O. S. L. T. T. W. Edward Beeching, Clémentine Fourrier, "Open llm leaderboard," https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

[82] H. Raj, D. Rosati, and S. Majumdar, "Measuring reliability of large language models through semantic consistency," *arXiv preprint arXiv:2211.05853*, 2022.

[83] T. Schick and H. Schütze, "It's not just size that matters: Small language models are also few-shot learners," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 2339–2352. [Online]. Available: https://aclanthology.org/2021.naacl-main.185/

[84] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi, "Least-to-most prompting enables complex reasoning in large language models," in *International Conference on Learning Representations*, 2023.

[85] N. Patel, M. Kulkarni, M. Parmar, A. Budhiraja, M. Nakamura, N. Varshney, and C. Baral, "Multi-logieval: Towards evaluating multi-step logical reasoning ability of large language models," in *Proceedings of EMNLP*, 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.1160.pdf

[86] A. K. Lampinen, I. Dasgupta, S. C. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, and F. Hill, "Language models, like humans, show content effects on reasoning tasks," *PNAS Nexus*, vol. 3, no. 7, p. pgae233, 2024.

[87] Z. Jin, P. Cao, Y. Chen, K. Liu, X. Jiang, J. Xu, L. Qiuxia, and J. Zhao, "Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models," in *Proceedings of LREC-COLING*, 2024, pp. 10 142–10 151.

[88] Z. Su, J. Zhang, X. Qu, T. Zhu, Y. Li, J. Sun, J. Li, M. Zhang, and Y. Cheng, "Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm," 2024. [Online]. Available: https://arxiv.org/abs/2408.12076

[89] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018. [Online]. Available: https://en.wikipedia.org/wiki/The_Book_of_Why

[90] Y. Zhang, W. Wang, X. Liu, Y. Chen, and Z. Li, "Bridging the gap between llms and human intentions," *arXiv preprint arXiv:2502.09101*, 2024. [Online]. Available: https://arxiv.org/abs/2502.09101

[91] M. Lewis and M. Mitchell, "Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models," *arXiv preprint arXiv:2402.08955*, 2024.

[92] Y. Liu, Z. Yao, X. Lv, Y. Fan, S. Cao, J. Yu, L. Hou, and J. Li, "Untangle the KNOT: Interweaving Conflicting Knowledge and Reasoning Skills in Large Language Models," Apr. 2024, arXiv:2404.03577 [cs]. [Online]. Available: http://arxiv.org/abs/2404.03577

[93] D. Kaushik, E. Hovy, and Z. C. Lipton, "Learning the difference that makes a difference with counterfactually-augmented data," in *Proceedings of the International Conference on Learning Representations*, 2020.

[94] J. Thomas, C. Ditzfeld, and C. Showers, "Compartmentalization: A window on the defensive self 1," *Social and Personality Psychology Compass*, vol. 10, pp. 719–7 311 111, 10 2013.

[95] S. Han, H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, W. Zhou, J. Coady, D. Peng, Y. Qiao, L. Benson, L. Sun, A. Wardle-Solano, H. Szabo, E. Zubova, M. Burtell, J. Fan, Y. Liu, B. Wong, M. Sailor,

A. Ni, L. Nan, J. Kasai, T. Yu, R. Zhang, A. R. Fabbri, W. Kryscinski, S. Yavuz, Y. Liu, X. V. Lin, S. Joty, Y. Zhou, C. Xiong, R. Ying, A. Cohan, and D. Radev, "Folio: Natural language reasoning with first-order logic," 2024. [Online]. Available: https://arxiv.org/abs/2209.00840

[96] C. Xie, Y. Huang, C. Zhang, D. Yu, X. Chen, B. Y. Lin, B. Li, B. Ghazi, and R. Kumar, "On memorization of large language models in logical reasoning," *arXiv.org*, 2024.

[97] Y. Wang and Y. Zhao, "Metacognitive prompting improves understanding in large language models," 2024. [Online]. Available: https://arxiv.org/abs/2308.05342

[98] M. Rollwage and S. M. Fleming, "Confirmation bias is adaptive when coupled with efficient metacognition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 376, no. 1822, p. 20200131, 2021.