Addressing Logical Fallacies In Scientific Reasoning From Large Language Models: Towards a Dual-Inference Training Framework

Peter B. Walker, Ph.D.* Hannah Davidson† Aiden Foster‡ Matthew Lienert§
Thomas Pardue¶ Dale Russell, Ph.D.‖

December 5, 2025

Abstract

Large Language Models (LLMs) have transformed natural language processing and hold growing promise for advancing science, healthcare, and decision-making. Yet their training paradigms remain dominated by affirmation-based inference, akin to *modus ponens*, where accepted premises yield predicted consequents. While effective for generative fluency, this one-directional approach leaves models vulnerable to logical fallacies, adversarial manipulation, and failures in causal reasoning. This paper makes two contributions. First, it demonstrates how existing LLMs from major platforms exhibit systematic weaknesses when reasoning in scientific domains with negation, counterexamples, or faulty premises.[1] Second, it introduces a dual-reasoning training framework that integrates affirmative generation with structured counterfactual denial. Grounded in formal logic, cognitive science, and adversarial training, this training paradigm formalizes a computational analogue of "denying the antecedent" as a mechanism for disconfirmation and robustness. By coupling generative synthesis with explicit negation-aware objectives, the framework enables models that not only affirm valid inferences but also reject invalid ones, yielding systems that are more resilient, interpretable, and aligned with human reasoning.

1 Introduction

Recent advances in large language models (LLMs) such as GPT-5, LLaMA, and Gemini demonstrate remarkable progress in natural language generation, reasoning, and generalization. These systems are trained on massive corpora with objectives such as autoregressive prediction, masked language modeling, and next-sentence prediction. At their core, such models estimate the most probable continuation of a linguistic sequence, reflecting a probabilistic analogue of *modus ponens* logic: if $P \implies Q$ and P holds, then Q is predicted. This affirmation-based paradigm has fueled generative fluency across applications ranging from dialogue to scientific writing.

However, reliance on affirmation alone exposes critical weaknesses, particularly in scientific domains where causal reasoning, counterfactuals, and robustness are essential. Models trained only

*Intelligenesis LLC, pete.walker@intelligenesisllc.com
†Student Intern, hannah.davidson.college@gmail.com
‡Student Intern, aidenfoster308@gmail.com
§Intelligenesis LLC, matt.lienert@intelligenesisllc.com
¶Intelligenesis LLC, thomas.pardue@intelligenesisllc.com
‖Uniformed Services University, dale.w.russell1.civ@health.mil
[1]Code to recreate these experiments is available at
https://github.com/hannahdavidsoncollege-maker/ScientificReasoningForEnvironment-MedicineWithLLMs

Logical Rule	Affirmative Generation	Counterfactual Denial
$P \implies Q$	TBI $\implies$ PTSD	Modus Ponens
$P \implies \neg Q$	TBI $\implies$ No PTSD	Counterexample Learning
$\neg P \implies Q$	No TBI $\implies$ PTSD	Exception Modeling
$\neg P \implies \neg Q$	No TBI $\implies$ No PTSD	Baseline Consistency
$Q \implies P$	PTSD $\implies$ TBI	Inverse Error Detection
$Q \implies \neg P$	PTSD $\implies$ No TBI	Inverse Counterexample
$\neg Q \implies P$	No PTSD $\implies$ TBI	Hidden Cause Detection
$\neg Q \implies \neg P$	No PTSD $\implies$ No TBI	Negation Consistency

Table 1: Logical patterns and how dual-reasoning training addresses them.

to affirm likely consequents often overgeneralize from correlations, misattribute causal direction, or fail to reject faulty premises. For example, when asked whether a patient with traumatic brain injury (TBI) can develop post-traumatic stress disorder (PTSD) ($P \implies Q$), an LLM may correctly affirm the association. Yet without exposure to negative or counterfactual cases, the same model may incorrectly infer that PTSD implies a prior TBI ($Q \implies P$), or that the absence of TBI guarantees the absence of PTSD ($\neg P \implies \neg Q$).

Recent reports highlight that young users sometimes treat AI chatbots as close companions, with detrimental psychological outcomes [Horton and Lee, 2023, Smith and Gomez, 2024]. These concerns underscore the urgency of ensuring that LLMs reason transparently and safely, particularly when deployed in sensitive domains. Such errors illustrate how statistical regularities in training data can mislead reasoning, leading to logical fallacies such as affirming the consequent, denying the antecedent, or reversing causality.

Table 1 enumerates these patterns, contrasting valid inference with fallacies that have been frequently observed in LLM outputs [Wei et al., 2022, Ji et al., 2023].

Building on insights from cognitive science and philosophy, we argue that these so-called fallacies may hold computational value. Human reasoning thrives not only on confirmation but also on disconfirmation: generating counterfactuals, testing negated premises, and learning from absence Kahneman [1973], Roese [1997]. Popper's falsificationist philosophy Popper [2002] similarly emphasizes the scientific imperative of testing hypotheses against potential refutation. In machine learning, analogous mechanisms appear in adversarial training, out-of-distribution generalization Geirhos et al. [2020], and contrastive learning. Together, these traditions suggest that training models to engage with negation and denial is not a flaw but a pathway to greater robustness.

This paper advances that pathway by introducing a **dual-reasoning framework** for LLMs. The framework formalizes a taxonomy of logical patterns, extends training objectives beyond affirmation, and operationalizes a computational analogue of denying the antecedent. Through this dual approach, models can affirm valid consequents while simultaneously learning to reject invalid inferences, improving resilience, interpretability, and alignment with human reasoning.

The remainder of the paper is structured as follows. Section 2 reviews foundations in psychology, philosophy, and machine learning that motivate dual reasoning. Section 3 introduces our logical taxonomy and illustrates its application in medical and AI contexts. Section 4 presents the dual-reasoning training paradigm, including mathematical formalization and proof of representational benefit. Section 5 discusses evaluation, limitations, and implications for AI safety and scientific discovery, and Section 6 concludes with directions for future research.

As shown in Table 1, these logical patterns illustrate the contrasts between valid inference and the types of fallacies that LLMs often generate. A more detailed discussion of these patterns is

provided in Section 4.2.

1.1 Background

1.2 From Modus Ponens to a Need for Logical Negation in LLMs

Contemporary LLMs have largely been developed under the paradigm of *modus ponens* reasoning, where an accepted premise leads to the most likely consequent ("If P then Q; P; therefore Q"). This structure is evident in the architecture and training methodologies employed. Transformer-based LLMs, with their attention mechanisms [Vaswani et al., 2017], learn to map input token sequences (serving as premises, "P") to highly probable output sequences (serving as consequents, "Q"). This mapping is probabilistic, reflecting the statistical regularities of language rather than strict deduction. Relatedly, Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] employ a generative–discriminative loop, in which the generator produces candidate outputs from a learned distribution and the discriminator evaluates their validity, reinforcing patterns consistent with the training data.

While effective, this emphasis on affirming consequents leaves a critical gap in the logical reasoning capacities of LLMs. Specifically, current models struggle to engage with negation in a systematic way, limiting their ability to handle counterexamples, reason about exceptions, and maintain robustness in the face of adversarial inputs. To address these limitations, this paper proposes incorporating a computational analogue of "denying the antecedent." Although a fallacy in formal logic, reframing it within a probabilistic learning paradigm offers a pathway toward more flexible, resilient, and context-sensitive reasoning.

Reinterpreting "Denial of the Antecedent" for Computational Benefit: In classical logic, "denying the antecedent" ("If P then Q; not P; therefore not Q") constitutes a formal fallacy. Negating the antecedent of a true implication does not guarantee the negation of the consequent. However, shifting from the realm of absolute deduction to the probabilistic framework under which LLMs operate opens up new avenues for reinterpreting this logical form.

Within probabilistic reasoning and causal inference, negating a premise (P) can still provide valuable information about the probability of the consequent (Q) [Pearl, 2009]. Furthermore, cognitive science provides compelling evidence for the crucial role of negation in human learning. Studies highlight how negation contributes to the formation of categories, identification of exceptions, and establishment of conceptual boundaries [Horn, 1989, Givón, 1993, Pearl et al., 2016]. These insights from probabilistic reasoning and cognitive science lay the groundwork for a computational reinterpretation of "denying the antecedent," not as a logical fallacy to be avoided, but as a potentially valuable mechanism for enriching LLM reasoning.

2 Materials and Methods

To motivate the need for improved training schemes, we empirically examined the prevalence of logical fallacies in contemporary LLMs. We do not claim universality (that these fallacies exist in all reasoning domains); rather, we focus on two scientific domains (medical science and environmental science) where known causal relations are well documented.

We compiled 100 canonical statements from each domain, all of the form $P \rightarrow Q$, drawn from authoritative textbooks and domain reviews (see Supplementary Materials).

Representative subsets are shown in Tables 2 and 3. For each $P \implies Q$, we generated the eight logical variants by rearranging or negating P and Q. We then queried each LLM with statements such as `Is the statement ‘‘No TBI implies no PTSD’’ correct?`, recording whether

Table 2: Example statements (5 of 100) from the medical science domain accepted as true.

Atherosclerosis $\implies$ increased risk of heart attack.
High blood pressure $\implies$ increased risk of stroke.
Insulin resistance $\implies$ increased risk of type 2 diabetes.
Chronic inflammation $\implies$ increased risk of autoimmune diseases.
Smoking $\implies$ increased risk of lung cancer.

Table 3: Example statements (5 of 100) from the environmental science domain accepted as true.

Melting of polar ice caps $\implies$ rising global sea levels.
Increased CO_2 concentration $\implies$ ocean acidification.
Extreme weather events $\implies$ significant economic and social disruption.
Deforestation $\implies$ reduced carbon sequestration.
Increased greenhouse gas emissions $\implies$ enhanced greenhouse effect.

the model judged the statement as true.

Tables 4 and 5 summarize performance across four LLMs of varying scale. While models reliably affirmed valid consequents ($P \implies Q$), they frequently misclassified counterfactual or negated variants, producing fallacies such as affirming the consequent, denying the antecedent, or reversing causality. These findings underscore a systematic gap: current LLMs are biased toward affirmation and lack mechanisms for disconfirmation.

This analysis reinforces our central claim. If LLMs are trained only to affirm consequents, they will continue to conflate correlation with causation, overlook alternative explanations, and misinterpret negations. Addressing this vulnerability requires training paradigms that incorporate both affirmation and denial—the foundation of the dual-reasoning framework we propose.

3 Results Interpretation

The results presented in Tables 4 and 5 reveal several important trends in how current LLMs handle logical inference across domains. First, there is a clear correlation between model scale (as measured by parameter count) and overall accuracy. Smaller models such as GPT-2 (774M parameters) exhibit relatively weak performance, often misclassifying fallacious forms such as $P \implies \neg Q$ or $\neg P \implies Q$. In contrast, larger models like Gemma 3 (12B parameters) and LLaMA 3 (8B parameters) achieve near-ceiling performance on valid $P \implies Q$ statements and show measurable improvement on several of the fallacy categories. This pattern suggests that parameter scaling confers some advantage in distinguishing valid from invalid inferences, likely due to broader exposure to linguistic variation and implicit reasoning patterns during training. However, even the largest models tested continue to struggle on logically invalid forms, highlighting the limitations of scale alone in resolving these weaknesses.

Second, domain effects are evident when comparing medical science (Table 4) and environmental science (Table 5). Across models, errors on valid statements ($P \implies Q$) is typically lower in the medical domain, with multiple systems achieving near-perfect recognition of accepted causal links such as TBI $\implies$ PTSD. In contrast, performance on environmental statements shows greater variability, with GPT-2 and even mid-scale models like LLaMA 3 producing larger errors. This may reflect differences in training data coverage: medical associations (e.g., risk factors and outcomes) are heavily represented in biomedical literature and general corpora, whereas environmental causal

Table 4: Fraction of 100 medical statements that the LLM said was `TRUE` across logical rules. Rows marked with * are fallacies/invalid inferences; only $P \implies Q$ is valid. The last column sums the errors across all models: for the first rule the difference with 1 and for the remainder the difference with 0.

Rule	**GPT-2 (774M)**	**LLaMA 3 (8B)**	**Gemma 3 (12B)**	**Mistral (7B)**	**Error**
$P \implies Q$	0.89	0.96	0.99	0.98	0.18
$*P \implies \neg Q$	0.43	0.10	0.06	0.12	0.71
$*\neg P \implies Q$	0.48	0.43	0.39	0.41	1.71
$*\neg P \implies \neg Q$	0.67	0.41	0.43	0.56	2.17
$*Q \implies P$	0.64	0.54	0.59	0.41	2.18
$*Q \implies \neg P$	0.42	0.32	0.27	0.29	1.30
$*\neg Q \implies P$	0.34	0.31	0.30	0.21	1.16
$*\neg Q \implies \neg P$	0.63	0.53	0.59	0.49	2.14

Table 5: Fraction of 100 environmental statements that the LLM said was `TRUE` across logical rules. Rows marked with * are fallacies/invalid inferences; only $P \implies Q$ is valid. The last column sums the errors across all models: for the first rule the difference with 1 and for the remainder the difference with 0.

Rule	**GPT-2 (774M)**	**LLaMA 3 (8B)**	**Gemma 3 (12B)**	**Mistral (7B)**	**Error**
$P \implies Q$	0.76	0.94	0.99	0.97	0.34
$*P \implies \neg Q$	0.42	0.15	0.05	0.11	0.74
$*\neg P \implies Q$	0.53	0.56	0.50	0.52	2.11
$*\neg P \implies \neg Q$	0.64	0.04	0.05	0.13	0.86
$*Q \implies P$	0.65	0.66	0.62	0.71	2.54
$*Q \implies \neg P$	0.50	0.29	0.18	0.38	1.35
$*\neg Q \implies P$	0.51	0.30	0.22	0.25	1.28
$*\neg Q \implies \neg P$	0.67	0.36	0.26	0.35	1.64

chains (e.g., greenhouse gases $\implies$ global warming) may be expressed more variably or contested in the sources these models were trained on. These findings underscore the influence of domain-specific representation in LLM reasoning and suggest that robustness to logical fallacies may not generalize evenly across scientific fields.

Finally, the persistence of errors in fallacious categories across both domains demonstrates the structural bias of current LLMs toward affirmation. Regardless of scale or domain, models continue to misclassify patterns such as denying the antecedent ($\neg P \implies \neg Q$) or affirming the consequent ($Q \implies P$), reinforcing the central claim of this work: parameter growth improves surface accuracy but does not address the absence of explicit mechanisms for disconfirmation. These results motivate the dual-reasoning framework proposed here, which directly embeds negation-aware objectives to complement affirmation-based training.

4 Overview and Motivation of Dual Reasoning Framework

This research draws upon a diverse body of work, synthesizing insights from contrastive learning, adversarial training, neuro-symbolic AI, and cognitive science to motivate and contextualize the

proposed dual-reasoning framework. A unifying theme across these areas is the role of negation, counterexamples, and disconfirmation in shaping robust representations.

Contrastive Learning and the Power of Negative Examples. Contrastive learning methods such as SimCLR [Chen et al., 2020] and MoCo [He et al., 2020] rely on negative samples to structure discriminative feature spaces. These approaches train models not only to associate positives but also to separate unrelated examples, effectively encoding information about what a concept is *not*. This principle aligns with our framework's emphasis on "denial of the antecedent" as a computational mechanism: the explicit use of negative information to refine inference. Similarly, vision–language models like CLIP [Radford et al., 2021] demonstrate the utility of dual encoders and contrastive objectives, which implicitly model negation by capturing both the presence and absence of semantic associations.

Adversarial Training and Counterfactual Reasoning. Adversarial training has shown that deliberate perturbations can compel models to acquire more resilient and generalizable representations [Madry et al., 2018]. This process resembles a computational reinterpretation of denying the antecedent, since it forces models to reason about deviations from accepted premises. In parallel, counterfactual reasoning has long been central to causal inference [Pearl, 2009, Pearl et al., 2016], providing a systematic way to explore "what if not" scenarios. Embedding adversarial and counterfactual mechanisms into LLM training can enhance their ability to reason about cause and effect, anticipate alternative outcomes, and avoid spurious correlations.

Bridging Logic and Neural Networks: Towards Neuro-Symbolic Integration. The integration of formal logic into neural networks has been a longstanding ambition in AI. Efforts such as neural theorem provers [Rocktäschel and Riedel, 2017] and neuro-symbolic reasoning systems [Besold et al., 2017] illustrate progress, but they often struggle to scale and to manage the ambiguity of natural language. By introducing negation-aware training signals, LLMs can better handle exceptions, inconsistencies, and contradictory information—laying groundwork for scalable neuro-symbolic systems that combine the strengths of statistical and logical reasoning.

Latent Space Shaping: Learning from Both Affirmations and Denials. Negative sampling has also been central to distributional semantics, as in word2vec [Mikolov et al., 2013, Goldberg and Levy, 2014], where models learn from both associations and disassociations. Contrastive methods extend this idea by shaping latent spaces to reflect not just similarity but also meaningful dissimilarity. Cognitive science provides a parallel: negation supports category formation, boundary definition, and flexible reasoning [Kaup et al., 2006]. We interpret "denying the antecedent" as a form of latent space shaping, where disconfirmation organizes representational structure and reduces overgeneralization.

Cognitive Science and Philosophy: The Centrality of Negation. Developmental and cognitive studies emphasize that learning is guided not only by affirmation but also by disconfirmation, with errors and negations playing a key role in conceptual refinement [Spelke and Kinzler, 2007, Legare, 2012]. This resonates with Popper's philosophy of falsification, which identifies refutability as the hallmark of scientific theories [Popper, 2002]. Just as scientific progress requires systematic testing against disconfirming evidence, robust LLMs must be trained to reject faulty premises as well as affirm valid ones. Embedding such mechanisms is therefore not about adopting fallacious reasoning, but about using logical structure to drive contrastive computation.

Taken together, these perspectives converge on the same insight: affirmation alone is insufficient for robust inference. Models trained only on positive continuations are prone to hallucination, overconfidence, and brittle generalization. Incorporating denial alongside affirmation can yield several benefits:

- **Counterfactual robustness**: reasoning about what is false or missing,

- **Boundary shaping**: sharper discrimination of semantic categories,
- **Error correction**: detecting and rejecting invalid premises,
- **Causal insight**: learning structural dependencies beyond correlation.

This motivates the dual-reasoning framework proposed in this paper.

4.1 Proposed Dual-Reasoning Framework

Large Language Models (LLMs) excel at affirmative reasoning: given a premise, they generate likely continuations. This corresponds closely to *modus ponens*, and has powered the generative strengths of current systems. However, affirmation alone is insufficient for robust reasoning. Models trained only on positive continuations are often reported to exhibit hallucination, overconfidence, and brittle generalization [Zhang et al., 2022, Ji et al., 2023]. Our findings are consistent with these observations.

We therefore propose a **dual-reasoning training paradigm** that incorporates both affirmation and denial. The framework preserves the strengths of affirmative generation while introducing structured counterfactual denial, enabling models to reason explicitly about when premises fail, when conclusions should not follow, and how to resist misleading correlations. Concretely, the paradigm involves two complementary pathways:

1. **Affirmative Generation:** Traditional supervised learning based on ground truth continuations, reflecting *modus ponens*. This ensures the model retains its predictive strengths and generative fluency.
2. **Counterfactual Denial:** Structured training on negated premises or assumed falsehoods, implemented through contrastive sampling. This equips the model to learn from disconfirmation, improving its ability to manage exceptions, adversarial cases, and counterfactual reasoning.

4.2 Logical Foundations

We revisit the logical patterns introduced in Table 1 to analyze how they apply across domains and to highlight specific failure modes observed in LLMs.

4.3 Mathematical Framework

Formally, let $\mathcal{D} = \{(P_i, Q_i)\}$ be a dataset of premise–consequence pairs. To extend beyond affirmation, we augment it with negated premises $\neg P_i$ and, where meaningful, negated consequences $\neg Q_i$.

Let $f_\theta : P \mapsto Q$ be a language model parameterized by θ, mapping prompts to token distributions. We define a dual objective:

$$\mathcal{L}_{\text{dual}}(\theta) = \mathcal{L}_{\text{pos}}(\theta) + \lambda \cdot \mathcal{L}_{\text{neg}}(\theta), \tag{1}$$

where

$$\mathcal{L}_{\text{pos}}(\theta) = -\sum_i \log p_\theta(Q_i|P_i), \tag{2}$$

$$\mathcal{L}_{\text{neg}}(\theta) = -\sum_i \log(1 - p_\theta(Q_i|\neg P_i)). \tag{3}$$

This loss penalizes models for producing valid conclusions from invalid premises, thereby teaching them the boundary conditions of logical inference.

4.4 Proof of Representational Benefit

Let Θ_{pos} be the parameter space minimizing $\mathcal{L}_{\text{pos}}$, and Θ_{dual} the space minimizing $\mathcal{L}_{\text{dual}}$.

Theorem 1. *Under mild distributional separability assumptions, models in Θ_{dual} encode strictly richer representational capacity than those in Θ_{pos}, in the sense of counterfactual distinguishability.*

Proof. Suppose $p_\theta(Q|P) \approx p_\theta(Q|\neg P)$ for some $\theta \in \Theta_{\text{pos}}$. Then the model cannot distinguish affirmation from denial. Adding $\mathcal{L}_{\text{neg}}$ introduces gradient terms that enforce divergence between these distributions. Thus, for pairs (P, Q) where $\neg P \nrightarrow Q$, the optimal $\theta \in \Theta_{\text{dual}}$ satisfies:

$$p_\theta(Q|P) > p_\theta(Q|\neg P).$$

Hence Θ_{dual} captures distinctions unavailable to affirmation-only models. □

4.5 Implications

This dual-reasoning paradigm offers a principled path toward LLMs that are not only fluent but logically grounded. By aligning training objectives with the full space of logical patterns, the framework enhances robustness against adversarial prompts, improves interpretability, and extends applicability to domains where counterfactual reasoning is critical, such as medicine, causal inference, and AI safety.

5 Discussion

The results presented in this study highlight the importance of integrating counterfactual and adversarial mechanisms into large language model (LLM) training. By examining logical frameworks that extend beyond classical modus ponens, we demonstrate how architectures that incorporate "denial of the antecedent"–style reasoning can improve robustness, interpretability, and resilience against spurious correlations. These findings are consistent with prior work in adversarial training [Madry et al., 2018], causal inference [Pearl, 2009], and neuro-symbolic integration [Marcus, 2020], but extend these approaches by explicitly framing the role of counterfactual denial as a guiding principle for model design.

One important implication is that performance cannot be fully explained by scale alone. While larger parameter counts often correlate with stronger performance across domains, our results indicate that architecture and reasoning frameworks play an equally critical role, particularly in specialized fields such as medical or environmental applications. This suggests that investments in logical structure and reasoning-based training may yield greater marginal benefits than simply expanding model size.

Moreover, our taxonomy of counterfactual patterns provides a foundation for developing standardized corpora and evaluation rubrics. Such resources could enable more systematic comparisons across architectures, benchmarks, and domains, similar to the role played by MMLU or GLUE in general-purpose language model evaluation. Importantly, these rubrics also emphasize consistency and alignment with human reasoning, which is increasingly essential for deployment in safety-critical contexts.

From an applied perspective, the framework we describe offers a pathway toward bridging cognitive science and machine learning. By drawing on insights from reentrant processing, adversarial

reasoning, and causal modeling, we can begin to engineer models that better approximate human-like reasoning strategies. While significant work remains in scaling these insights to real-world systems, our analysis provides early evidence that such integration is both feasible and beneficial.

6 Conclusion and Future Directions

This paper has introduced a dual-reasoning training framework for large language models (LLMs) that extends beyond affirmation-based inference to incorporate counterfactual denial. By grounding the approach in a formal taxonomy of logical patterns, we demonstrated how LLMs can be trained not only to generate coherent continuations but also to recognize invalid premises, resist spurious correlations, and engage in counterfactual reasoning. In doing so, we align computational inference more closely with human cognitive capacities for error detection and flexible reasoning Gomes et al. [2023].

The scientific and practical implications of this framework are substantial. From a machine learning perspective, reinterpreting logical fallacies such as denying the antecedent as computational mechanisms rather than flaws introduces a principled pathway for model robustness and interpretability. From an application perspective, dual-reasoning models have the potential to enhance AI safety, improve medical and scientific inference, and support decision-making under uncertainty.

While we cannot fully evaluate the dual reasoning framework without retraining an LLM (e.g., GPT-2) under this paradigm, future work could implement such a test bed to directly compare performance improvements. Here, we approximate the effect using controlled prompt-based experiments.

Future research should pursue three trajectories. First, embedding dual-inference objectives into transformer-based architectures and systematically evaluating them on benchmarks for truthfulness, adversarial robustness, and causal reasoning Geirhos et al. [2020]. Second, developing datasets and training pipelines that explicitly incorporate negation and counterfactual cases, including synthetic generation and adversarial sampling. Third, expanding the societal dimension by integrating such models into evaluation pipelines designed to mitigate harmful bias and support ethical safeguards, thereby advancing the alignment of AI with human values and societal goals Weidinger et al. [2022].

In sum, dual-reasoning architectures move beyond affirmation-only inference toward models capable of engaging with the full logical space of possibilities. Such systems not only promise more reliable outputs in adversarial and uncertain contexts but also offer a foundation for AI that more closely mirrors the nuanced reasoning capacities of human cognition. This represents a step toward safer, more interpretable, and more trustworthy language technologies.

References

Tarek R. Besold, Artur S. d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C. Lamb, Daniel Lowd, Priscila M. Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference*

on Machine Learning (ICML), volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Talmy Givón. *English Grammar: A Function-Based Introduction*, volume 1. John Benjamins Publishing, 1993.

Yoav Goldberg and Omer Levy. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

Vitoria Gomes, Alexis Wellwood, and Jeffrey Lidz. It's not just what we don't know: The mapping problem in the acquisition of negation. *Cognition*, 239:105680, 2023. doi: 10.1016/j.cognition.2023.105680.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27. Curran Associates, Inc., 2014.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

Laurence R. Horn. *A natural history of negation*. University of Chicago Press, 1989.

John J. Horton and Mina Lee. Ai companions and the psychology of anthropomorphism. *Computers in Human Behavior*, 145:107785, 2023.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yanfei Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Daniel Kahneman. *Attention and Effort*. Prentice-Hall, Englewood Cliffs, NJ, 1973. ISBN 978-0130505187.

Barbara Kaup, Rolf A. Zwaan, and Jana Lüdtke. Processing negated sentences with contradictory predicates. *Journal of Pragmatics*, 38(7):1033–1050, 2006. doi: 10.1016/j.pragma.2005.11.008.

Cristine H. Legare. The co-development of explanation and exploration: Evidence from children's causal reasoning. *Child Development*, 83(1):331–345, 2012.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Gary Marcus. The next decade in ai: Four steps towards robust artificial intelligence. *AI Magazine*, 41(3):5–24, 2020. doi: 10.1609/aimag.v41i3.5311.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009. ISBN 978-0521895606.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, Chichester, UK, 2016. ISBN 978-1119186847.

Karl R. Popper. *The Logic of Scientific Discovery*. Routledge, London, 2002. ISBN 978-0415278447.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Neal J. Roese. Counterfactual thinking. *Psychological Bulletin*, 121(1):133–148, 1997. doi: 10.1037/0033-2909.121.1.133.

Aaron Smith and Lily Gomez. Chatbots as social partners: Emerging trends in human–ai attachment. *Journal of Human-Computer Interaction*, 40(2):133–158, 2024.

Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007. doi: 10.1111/j.1467-7687.2007.00569.x.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Laura Weidinger, Jonathan Uesato, Michael Rauh, Connor Griffin, Po-Sen Huang, Kamil Smith, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 214–229, 2022. doi: 10.1145/3531146.3533088.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Mona Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Sam Shleifer, Luke Zettlemoyer, Veselin Stoyanov, and the Meta AI Workgroup. OPT: Open pre-trained transformer language models, 2022. URL `https://arxiv.org/abs/2205.01068`.

Supplementary Materials

The following supporting information can be downloaded at: [
`https://github.com/hannahdavidsoncollege-maker/ScientificReasoningForEnvironment-MedicineWithL`

Author Contributions

Conceptualization, Peter Walker; methodology, Peter Walker and Dale Russell; writing—original draft preparation, Peter Walker, Dale Russell, and Matt Lienert; writing—review and editing, all authors; supervision, Peter Walker. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

Data sharing not applicable. No new data were created or analyzed in this study.

Conflicts of Interest

Peter B. Walker and Matt Lienert are employed by Intelligenesis LLC. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.