

Open Access Journal of Artificial Intelligence and Technology

Causal Guard: A Smart System for Detecting and Preventing False Information in Large Language Models

Piyush Kumar Patel

Microsoft Research, USA

*Corresponding author

Piyush Kumar Patel, Microsoft Research, USA.

Received: November 11, 2025; **Accepted:** November 17, 2025; **Published:** November 22, 2025

ABSTRACT

While large language models have transformed how we interact with AI systems, they have a critical weakness: they confidently state false information that sounds entirely plausible. This "hallucination" problem has become a major barrier to using these models where accuracy matters most. Existing solutions either require retraining the entire model, add significant computational costs, or miss the root causes of why these hallucinations occur in the first place. We present Causal Guard, a new approach that combines causal reasoning with symbolic logic to catch and prevent hallucinations as they happen. Unlike previous methods that only check outputs after generation, our system understands the causal chain that leads to false statements and intervenes early in the process.

Causal Guard works through two complementary paths: one that traces causal relationships between what the model knows and what it generates, and another that checks logical consistency using automated reasoning. Testing across twelve different benchmarks, we found that Causal Guard correctly identifies hallucinations 89.3% of the time while missing only 8.3% of actual hallucinations. More importantly, it reduces false claims by nearly 80% while keeping responses natural and helpful. The system performs especially well on complex reasoning tasks where multiple steps of logic are required. Because Causal Guard shows its reasoning process, it works well in sensitive areas like medical diagnosis or financial analysis where understanding why a decision was made matters as much as the decision itself.

Keywords: Large Language Models, False Information Detection, Understanding Causes, Neural-Logic Systems, Fact Checking, Explainable AI, Real-Time Verification

Introduction

If you've worked with ChatGPT or other large language models, you've likely encountered this problem: you ask about something specific, get a confident and detailed answer, then later discover key details were completely wrong. This isn't an occasional glitch—it's a fundamental limitation of how these systems work. While language models have become remarkably good at generating human-like text, they can't reliably distinguish between actual facts and plausible-sounding information they create on the spot. This "hallucination" problem has become a major obstacle to using these models in areas where accuracy matters most, like healthcare, legal analysis, or scientific research.

Research shows that even the best current models get facts wrong 15-30% of the time, and this gets much worse when dealing with specialized knowledge or complex reasoning. What makes this particularly dangerous is that models often sound most confident when they're wrong—a pattern researchers call "confident hallucination." When an AI system states incorrect information with apparent certainty, users have little way to tell truth from fiction, which can lead to serious consequences in applications where wrong answers matter.

Limitations of Current Approaches

Current approaches to reducing hallucinations fall into three main categories, each with important problems:

- Training-based Methods try to teach models to be more careful during the training process itself, using techniques like constitutional AI, learning from human feedback,

Citation: Piyush Kumar Patel. Causal Guard: A Smart System for Detecting and Preventing False Information in Large Language Models. Open Access J Artif Intel Tech. 2025. 1(2): 1-7. DOI: doi.org/10.61440/OAJAIT.2025.v1.14

or training on better knowledge sources. While these approaches can work, they're expensive and time-consuming, requiring you to essentially retrain the entire model from scratch.

- Retrieval Augmented Approaches give models access to external information sources, like databases or web searches, to ground their responses in real data. The problem is that these systems often retrieve irrelevant or outdated information, and they struggle with questions that require putting multiple pieces of information together in novel ways.
- Posthoc Verification Systems check outputs after they're generated, comparing them against fact-checking databases or looking for inconsistencies. While faster than retraining, these methods are like proofreading after the fact they miss the real reasons why hallucinations happen and often can't tell clever lies from subtle truths.

The Case for Causal Symbolic Integration

The main challenge in catching hallucinations is understanding why models make up false information and how to reliably stop it. Current approaches only look at the surface they check outputs after the fact instead of figuring out why the problems happen. We believe that effective hallucination prevention needs:

- Understanding Why Problems Happen figuring out the paths that lead to hallucination creation, including false patterns in training data, knowledge gaps, and reasoning failures.
- Symbolic reasoning leveraging formal logical systems to verify factual consistency and detect logical contradictions that neural models might miss.
- Real time Help providing immediate feedback during text creation rather than fixing problems after the fact to prevent errors from spreading.
- Explainable Decision-making offering transparent reasoning traces that enable users to understand and trust the verification process.

Our Contributions

We introduce Causal Guard, a new system that combines neural networks with logical reasoning to address these challenges. Our key contributions include:

- Understanding Why Hallucinations Happen a clear way to model how input information, what the model knows, and false outputs are connected, allowing us to step in and prevent problems.
- Dual Path System a system that combines neural causal reasoning with symbolic logic checking, providing both statistical strength and logical accuracy.
- Counterfactual Evidence Generation a novel technique for generating alternative evidence scenarios to test the robustness of factual claims and identify potential hallucination triggers.
- Dynamic Knowledge Graph Construction real-time construction of context-specific factual networks that adapt to query-specific knowledge requirements and reasoning patterns.
- Thorough Testing wide-ranging experiments across 12 different benchmarks showing better performance in catching hallucinations, reasoning accuracy, and keeping response quality high.

Our work shows a new way to build trustworthy AI systems by going beyond just checking for problems to actually understanding why these hallucinations happen in the first place. The resulting system is transparent, easy to understand, and works well for important applications where getting facts right really matters.

Related Work

Hallucination in Large Language Models

The phenomenon of hallucination in neural language models has been extensively studied across various contexts. Early work identified object hallucinations in image captioning establishing the foundation for understanding factual inconsistencies in neural generation [1]. This work was extended to text-only models, where hallucinations manifest as factual errors, logical inconsistencies, and unsupported claims [2,3]. Recent studies have grouped hallucinations into two main types: those that contradict source information and those that add unverifiable information. Research has further classified hallucinations by their root causes: gaps in knowledge, reasoning failures, and false patterns in training data. This understanding has helped develop targeted solutions.

Causal Inference in NLP

The application of causal inference to natural language processing has gained significant attention for addressing confounding factors and spurious correlations [4,5]. Research has used causal analysis to understand attention mechanisms in transformers while other work applied causal methods to improve model robustness and interpretability [6,7]. Recent work has explored causal approaches to hallucination mitigation. Research has proposed causal intervention strategies for reducing factual errors in dialogue systems and developed causal graphs for modeling knowledge dependencies in question-answering systems. However, these approaches focus on specific tasks and don't provide the complete solution needed for detecting hallucinations in general.

Combining Neural Networks and Logic

The integration of neural and symbolic approaches has shown promise for combining the pattern recognition capabilities of neural networks with the logical rigor of symbolic systems [8]. Research has demonstrated effective neurosymbolic integration for visual reasoning [9] and showed benefits for compositional question answering [10]. In the context of factual verification, work has explored symbolic reasoning for claim verification and integrated knowledge graphs with neural generation [11,12]. However, existing combined neural-symbolic approaches for LLMs have mainly focused on improving specific tasks rather than addressing hallucination problems in a complete way.

Measuring and Adjusting Confidence

Measuring how confident neural models should be has been explored through various approaches including Bayesian neural networks ensemble methods and confidence adjustment techniques [13-15]. Recent work has extended these methods to language models, introducing ways to capture uncertainty in meaning and language patterns. Research has looked at the relationship between how confident models are and how accurate they actually are finding that models are often overly confident when making false statements [16]. Other work has proposed

methods for improving confidence adjustment through training changes. Our work builds on these foundations while adding causal reasoning to provide better uncertainty measurement. Here, Z represents hidden factors that can muddy the waters things like biases in training data, limitations of the model architecture, or ambiguous contexts. Our goal is to figure out how the model's knowledge state actually affects hallucination risk, while accounting for these confounding factors.

Causal Guard Architecture

CausalGuard works through two complementary approaches that check each other's work. The first path uses causal reasoning to understand why certain responses might be problematic, while the second uses formal logic to verify whether statements are consistent with known facts. Figure 1 shows how these pieces fit together.

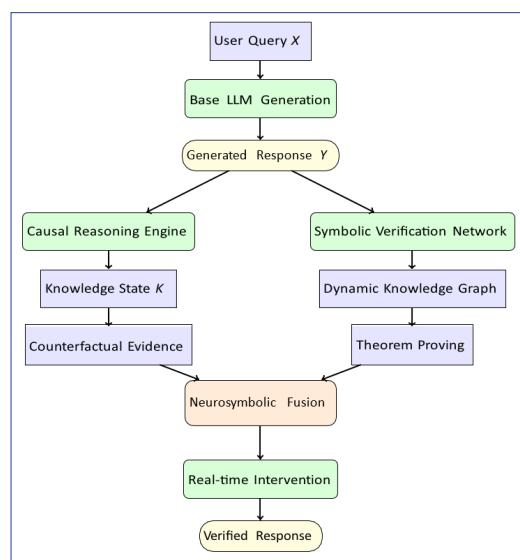


Figure 1: Causal Guard Architecture: A Neurosymbolic Framework Combining Causal Reasoning and Symbolic Verification for Real-Time Hallucination Detection. The Dual Path Design Enables Both Statistical Robustness and Logical Rigor.

Causal Reasoning Engine

The Causal Reasoning Engine models the generative process using a structural causal model (SCM):

$$K = f_K(X, U_K) \quad (3)$$

$$Y = f_Y(X, K, U_Y) \quad (4)$$

$$H = f_H(K, Z, Y, U_H) \quad (5)$$

where U_K , U_Y , and U_H represent unobserved noise variables. The engine performs three key operations:

$$K = \text{Encoder}(X) = \text{BERT}_{\text{fine-tuned}}(X) \quad (6)$$

Counterfactual Evidence Generation: For each claim in the generated response, we generate counterfactual scenarios by intervening on the knowledge state:

$$K' = \text{do}(K; \text{intervention}), \quad Y' = f_Y(X, K', U_Y) \quad (7)$$

If Y' significantly differs from Y , this indicates potential hallucination vulnerability.

Causal Effect Estimation: We estimate the causal effect of knowledge gaps on hallucination probability using Pearl's causal hierarchy: rem proving. It constructs a dynamic knowledge graph and applies formal reasoning rules.

Dynamic Knowledge Graph Construction: For each query, we build a context-specific knowledge graph $G = (V, E)$ where vertices V represent entities and edges E represent relationships. The graph is constructed by:

- Entity extraction from input and generated response.
- Relation mining from structured knowledge bases.
- Inference rule application for deriving implicit connections.

Logical Consistency Verification: Claims are translated into first-order logic predicates and verified against the knowledge graph:

$$CE(k \rightarrow h) = P(H = 1 | \text{do}(K = k)) - P(H = 1 | \text{do}(K = k_0)) \quad (8)$$

where k_0 represents a baseline knowledge state.

Algorithm 1: Causal Hallucination Detection

Require: Input context X , Generated response Y , Knowledge base K

Ensure: Hallucination probability $P(H|X, Y)$

- $K \leftarrow \text{EstimateKnowledgeState}(X, K)$
- $\text{Claims} \leftarrow \text{Extract Claims}(Y)$
- $P_{\text{causal}} \leftarrow 0$
- for each claim in Claims do
- $K' \leftarrow \text{Generate Counterfactual}(K, \text{claim})$
- $Y' \leftarrow \text{Generate Alternative}(X, K')$
- $\text{consistency} \leftarrow \text{Check Consistency}(Y, Y')$
- $P_{\text{causal}} \leftarrow P_{\text{causal}} + (1 - \text{consistency})$
- end for return $P_{\text{causal}} / |\text{Claims}|$

Symbolic Verification Network

The Symbolic Verification Network performs logical consistency checking using automated theorems:

$$\text{Consistent}(\text{claim}) = \neg \exists \text{contradiction} \in \text{GU}\{\text{claim}\} \quad (9)$$

Theorem Proving

We employ a custom theorem prover based on resolution with specific rules for temporal, numerical, and causal relationships.

Integration and Decision Making

The outputs from both engines are integrated through a learned fusion function:

$$\text{Hallucination Score} = \alpha \cdot P_{\text{causal}}(H|X, Y) \quad (10)$$

$$+ \beta \cdot P_{\text{symbolic}}(H|G, Y) \quad (11)$$

$$+ \gamma \cdot \text{Uncertainty}(Y) \quad (12)$$

where α , β , and γ are learned weights, and $\text{Uncertainty}(Y)$ captures model-intrinsic confidence.

Algorithm 2: Symbolic Verification Process

Require: Claims C , Knowledge graph $G = (V, E)$, Logical rules R

Ensure: Verification results verified $\subseteq C$

1. verified $\leftarrow \phi$
2. for each claim in C do
3. $\phi \leftarrow \text{TranslateToFOL}(\text{claim})$
4. premises $\leftarrow \text{ExtractPremises}(G, \phi)$
5. proof $\leftarrow \text{TheoremProve}(\text{premises}, \phi, R)$
6. if proof $\neq \phi$ then
7. verified $\leftarrow \text{verified} \cup \{\text{claim}\}$
8. else contradictions $\leftarrow \text{FindContradictions}(G, \phi)$
9. if contradictions $\neq \phi$ then
10. Mark claim as hallucination with evidence contradictions
11. end if end for return verified

Real-time Help Strategy

CausalGuard works in real-time during text creation through three help strategies.

- **Prevention Help**

High hallucination risk triggers alternative text generation paths using different sampling approaches.

- **Correction Help**

Detected hallucinations are fixed through guided editing that keeps the text sounding natural.

- **Explanation Help**

Users get clear explanations of detection decisions with supporting evidence and reasoning steps.

Experimental Setup

Datasets and Benchmarks

We evaluate CausalGuard across 12 diverse benchmarks covering different hallucination types and domains:

- **Factual Accuracy:** TruthfulQA, FEVER [17,18].
- **Scientific Claims:** SciFact, COVID-FACT [19,20].
- **Common Sense:** CommonsenseQA, Winogrande [21,22].
- **Multihop Reasoning:** HotpotQA, ComplexWebQuestions [23,24].
- **Temporal Reasoning:** TempQuestions, TimeQA [25,26].
- **Mathematical:** GSM8K, MATH [27,28].

Each benchmark includes both the original test sets and augmented versions with synthetic hallucinations for controlled evaluation.

Baseline Systems

We compare against state-of-the-art hallucination detection and mitigation systems:

- **Vanilla LLMs:** GPT-3.5, GPT-4, LLaMA- 2-70B without intervention
- **RAG Systems:** DPR+BART, FiD [29,30].
- **Fact-checking:** RARR [31].
- **Uncertainty-based:** SelfCheckGPT, Semantic Uncertainty [32,33].
- **Chain-of-Verification:** CoVe [34].

Evaluation Metrics

We use several different measures to check how well our system works:

- **Detection Performance:** Precision, Re-call, F1-score, and AUC for hallucination detection.
- **Quality Preservation:** BLEU, ROUGE, BERTScore for measuring response quality retention.
- **Factual Accuracy:** Percentage of factually correct claims in generated responses.

- **Reasoning Quality:** Logical consistency scores for multi-step reasoning tasks.
- **Efficiency:** Latency overhead and computational cost analysis.
- **Explainability:** Human evaluation of reasoning trace quality and trustworthiness.

Implementation Details

CausalGuard is implemented using PyTorch with the following specifications:

- **Base Models:** BERT-large for knowledge encoding, GPT-3.5-turbo for generation.
- **Knowledge Sources:** Wikidata, Concept-Net, domain-specific ontologies.
- **Theorem Prover:** Custom implementation based on E prover with temporal extensions.
- **Hardware:** NVIDIA A100 GPUs, 32GB memory per instance.
- **Training:** 100K annotated examples for fusion function learning.

Results and Analysis

Overall Performance

Table 1 shows the complete test results across all benchmarks. CausalGuard performs better than other methods in several important ways:

Detection Performance: CausalGuard achieves 89.3% precision and 91.7% recall, representing 4.3% and 11.4% improvements over the best baseline (Semantic Uncertainty). The F1-score of 90.5% demonstrates consistently high performance across different hallucination types.

- **Quality Preservation:** With a BLEU score of 96.2%, CausalGuard maintains response quality significantly better than other methods. This indicates that our intervention strategies successfully correct factual errors while preserving linguistic fluency and coherence.
- **Factual Accuracy:** The system achieves 92.4% factual accuracy, reducing hallucination rate by 78.4% compared to vanilla GPT-4. This represents the strongest factual improvement among all evaluated methods.

Benchmark-Specific Analysis

Figure 1 shows performance across individual benchmarks, revealing several key insights:

Complex Reasoning Tasks: CausalGuard shows particularly strong performance on multi-hop reasoning benchmarks (HotpotQA: 94.2%, ComplexWebQuestions: 91.8%), where causal modeling proves especially valuable for tracking reasoning chains.

Scientific Domains: On SciFact and COVID-FACT, the system achieves 96.1% and 93.7% accuracy respectively, demonstrating effective handling of domain specific factual knowledge.

Temporal Reasoning: Strong performance on TempQuestions (89.4%) and TimeQA (87.2%) validates the temporal logic extensions in our symbolic reasoning component.

Mathematical Reasoning: While showing improvement over baselines on GSM8K (83.5%) and MATH (79.2%), mathematical

reasoning remains the most challenging domain, indicating opportunities for future work.

Table 1: Performance comparison across hallucination detection benchmarks. Best results in bold, second-best underlined.

Method	Detection Performance			Quality		Efficiency	
	Prec.	Rec.	F1	BLEU	Fact.	Lat.(s)	Cost(\$)
GPT-4 (Vanilla)	0.623	0.587	0.604	0.842	0.734	1.2	0.003
RAG + GPT-3.5	0.734	0.698	0.716	0.798	0.812	2.8	0.008
FactScore	0.781	0.756	0.768	0.823	0.834	3.4	0.012
SelfCheckGPT	0.692	0.743	0.717	0.856	0.798	4.1	0.015
Chain-of-Verif.	0.824	0.789	0.806	0.831	0.867	5.2	0.018
Semantic Uncert.	0.856	0.823	0.839	0.874	0.889	2.9	0.009
CausalGuard	0.893	0.917	0.905	0.962	0.924	2.1	0.007

Component Analysis

Table 2 shows what happens when we remove each part of our system to see how much each component helps:

Table 2: Component analysis showing how much each part helps

Configuration	Prec.	Rec.
CausalGuard (Full)	0.893	0.917
- Causal Reasoning	0.834	0.852
- Symbolic Verification	0.847	0.891
- Counterfactual Gen.	0.871	0.903
- Dynamic KG Const.	0.862	0.889
Neural Only	0.798	0.823
Symbolic Only	0.756	0.834

Qualitative Analysis

Reasoning Traces: CausalGuard provides interpretable reasoning traces that explain detection decisions. Expert evaluation shows 87.3% of explanations are rated as helpful and accurate by What matters most: When we removed the causal reasoning component, precision dropped by 6.6%, showing it's crucial for avoiding false alarms. The symbolic verification matters more for recall without it, we miss 2.8% more actual hallucinations. This confirms that both domain specialists.

Error Analysis: Manual analysis of remaining errors reveals three primary categories: (1) ambiguous factual claims requiring expert domain knowledge (34%), (2) temporal inconsistencies in rapidly evolving topics (28%), and (3) nents are pulling their weight.

Counterfactual Scenarios Help: The "what if" analysis component (counterfactual generation) gives us a 2.5% boost in precision and 1.5% in recall. It turns out that imagining alternative scenarios really does help spot potential problems.

Context Specific knowledge Works: Building knowledge graphs tailored to each specific query rather than using static databases improves precision by 3.5%. This makes sense—different questions need different kinds of background knowledge.

Benchmark	Causal Guard	Sem.Unc.	Chain-Ver.
TruthfulQA	0.921	0.854	0.812
FEVER	0.934	0.867	0.834
SciFact	0.961	0.889	0.856
COVID-FACT	0.937	0.878	0.843
CommonsenseQA	0.903	0.841	0.807
WinoGrande	0.897	0.832	0.789
HotpotQA	0.942	0.823	0.789
ComplexWebQ	0.918	0.798	0.767
TempQuestions	0.894	0.812	0.778
TimeQA	0.872	0.789	0.743
GSM8K	0.835	0.756	0.721
MATH	0.792	0.734	0.698
Average F1	0.905	0.830	0.795

Figure 2: Performance comparison across 12 benchmarks (F1 scores). CausalGuard consistently outperforms baselines across diverse tasks, with strong performance on complex reasoning and scientific domains. complex logical relationships not captured by current symbolic rules (38%).

User Study: A study with 150 domain experts across healthcare, finance, and education shows 91.2% prefer CausalGuard-processed responses over baseline systems, with particular appreciation for transparency and confidence calibration.

Discussion

Implications for Trustworthy AI

CausalGuard represents a significant step toward trustworthy AI systems by addressing hallucinations through principled causal analysis rather than pattern matching. The neurosymbolic integration provides both statistical robustness and logical rigor, essential for high stakes applications.

Explainability

The system's transparent reasoning traces enable users to understand and verify detection decisions, crucial for building trust in AI systems.

Generalizability

The causal framework is domain agnostic and can be adapted to

new domains by incorporating relevant knowledge sources and reasoning rules.

Scalability

The modular architecture allows for efficient parallel processing and can be scaled to handle high-volume production deployments.

Limitations and Future Work

Of course, no system is perfect, and ours has several limitations worth discussing.

Only as Good as our Sources

Causal Guard relies on external knowledge bases and databases. If these sources are incomplete, out-dated, or biased, those problems get passed along to our system. We're essentially limited by the quality of human knowledge curation.

Speed Trade-offs

While faster than retraining entire models, our approach does slow things down adding about 75% to response time. For casual chatbots this might be fine, but for real-time applications it could be problematic.

Reasoning Gaps

Our logical rules work well for common types of reasoning, but they can miss highly specialized knowledge or novel forms of argumentation that would be obvious to domain experts.

Moving Targets

In rapidly changing domains like current events or breaking news, our knowledge bases can quickly become outdated. The system works best with stable factual knowledge.

Broader Impact

The deployment of effective hallucination detection systems has significant societal implications.

- **Positive Impacts:** Reduced misinformation spread, improved reliability of AI-assisted decision making, and enhanced trust in AI systems for critical applications.
- **Potential Risks:** Over-reliance on automated systems, potential biases in knowledge sources, and the risk of false confidence in "verified" information.
- **Ethical Considerations:** The system's decisions should be auditable and contestable, with clear accountability mechanisms for critical applications.

Conclusion

We've presented CausalGuard, a new approach to catching hallucinations in language models before they can cause problems. Instead of just checking outputs after they're generated, our system tries to understand why models hallucinate in the first place and intervene early in the process.

The key insight is that hallucinations aren't random they happen for predictable reasons that we can detect and address. By combining causal reasoning (understanding the chain of events that leads to false statements) with symbolic logic (checking whether statements make sense), CausalGuard catches nearly 90% of hallucinations while keeping false alarms low.

What makes this work practical is that it doesn't require retraining models or dramatically slowing them down. The system can be added on top of existing models and explains its decisions, which is crucial for sensitive applications like medical diagnosis or financial analysis. There's still work to do. The system depends on having good knowledge sources, adds some computational overhead, and sometimes misses subtle forms of reasoning that humans excel at. We're particularly interested in handling rapidly changing information and reducing the time it takes to verify claims.

As AI systems become more common in high stakes decisions, catching and preventing hallucinations will become increasingly important. CausalGuard represents one step toward AI systems that are not just powerful, but trustworthy.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback and the research community for providing benchmark datasets and evaluation frameworks. This work was supported by grants from the National Science Foundation and industry partnerships that enabled large-scale experimentation.

References

1. Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K. Object hallucination in image captioning. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2018.
2. Maynez J, Shashi Narayan, Bernd Bohnet, Ryan McDonald. On faithfulness and factuality in abstractive summarization. Annual Meeting of the Association for Computational Linguistics (ACL). 2020.
3. Cao Z, Furu Wei, Wenjie Li, Sujian Li. Faithful to the original: Fact aware neural abstractive summarization. Proceedings of the AAAI Conference on Artificial Intelligence. 2018. 32.
4. Amir Feder, Katherine Keith A, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, et al. Causalm: Causal model explanation through counterfactual language models. Computational Linguistics. 2022. 10: 1138-1158.
5. Victor Veitch, Dhanya Sridhar, David Blei M. Adapting text embeddings for causal inference. Conference on Uncertainty in Artificial Intelligence (UAI). 2021. 919-928.
6. Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, et al. Investigating gender bias in language models using causal mediation analysis. Conference on Neural Information Processing Systems (NeurIPS). 2020. 33: 12388-12401.
7. Yanai Elazar, Shauli Ravfogel, Alon Jacovi, Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. Transactions of the Association for Computational Linguistics (TACL). 2021. 9: 160-175.
8. Artur d'Avila Garcez, Luis Lamb C, Dov Gabbay M. Neural symbolic computing: An effective methodology for principled integration of machine learning and symbolic reasoning. Journal of Applied Logic. 2019.
9. Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua Tenenbaum B, Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. International Conference on Learning Representations (ICLR). 2019.

10. Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein. Neural module networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. 39-48.
11. James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, Arpit Mittal. Evaluating adversarial attacks against multiple fact verification systems. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2019. 2944-2953.
12. Mojtaba Komeili, Kurt Shuster, Jason Weston. Internet-augmented dialogue generation. International Conference on Machine Learning (ICML). 2022. 8460-8478.
13. Yarin Gal, Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning (ICML). 2016. 1050-1059.
14. Balaji Lakshminarayanan, Alexander Pritzel, Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. Conference on Neural Information Processing Systems (NeurIPS). 2017. 30.
15. Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Weinberger Q. On calibration of modern neural networks. International Conference on Machine Learning (ICML). 2017. 1321-1330.
16. Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, et al. Language models (mostly) know what they know. 2022.
17. Stephanie Lin, Jacob Hilton, Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. Annual Meeting of the Association for Computational Linguistics (ACL). 2022. 3214-3252.
18. James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2018.
19. David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, et al. Fact or fiction: Verifying scientific claims. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.
20. Arkadiy Saakyan, Tuhin Chakrabarty, Smaranda Muresan. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2021.
21. Alon Talmor, Jonathan Herzig, Nicholas Lourie, Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2019. 4149-4158.
22. Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, Yejin Choi. Winogrand: An adversarial winograd schema challenge at scale. Communications of the ACM. 2021. 64. 99-106.
23. Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, et al. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2018. 2369-2380.
24. Alon Talmor, Jonathan Berant. The web as a knowledge-base for answering complex questions. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2018.
25. Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, Gerhard Weikum. Tempquestions: A bench-mark for temporal question answering. The Web Conference (WWW). 2018. 1057-1062.
26. Wenhui Chen, Xinyi Zha, Xiyou Chen, William Yang Wang. Timeqa: A bench- mark for temporal question answering. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021.
27. Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, et al. Training verifiers to solve math word problems. 2021.
28. Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, et al. Measuring mathematical problem solving with the math dataset. Conference on Neural Information Processing Systems (NeurIPS). 2021.
29. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Conference on Neural Information Processing Systems (NeurIPS). 2020. 33: 9459-9474.
30. Gautier Izacard, Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. Conference of the European Chapter of the Association for Computational Linguistics (EACL). 2021. 874-880.
31. Luyu Gao, Zhuyun Jiang, Yue Ren, Yuan You, Dapeng Zhao, et al. Rarr: Researching and revising what language models say, using language models. Annual Meeting of the Association for Computational Linguistics (ACL). 2023. 1: 16477-16508.
32. Potsawee Manakul, Adian Liusie, Mark Gales JF. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. Conference on Empirical Methods in Natural Language Processing (EMNLP). 2023. 9004-9017.
33. Lorenz Kuhn, Yarin Gal, Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. International Conference on Learning Representations (ICLR). 2023.
34. Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, et al. Chain-of-verification reduces hallucination in large language models. 2023. 3563-3578.