# Addressing Temporal Bias and Causal Fragility in Large Language Models

### Background and Problem Definition

LLMs, such as GPT, PaLM, and LLaMA, have transformed natural language processing by generating coherent, contextually rich text. However, their training on static web-scraped data, representing a snapshot of knowledge at the time of collection, introduces two intertwined issues:

- **Temporal Bias**: LLMs encode knowledge from past states of the world, which becomes outdated as facts, causal relationships, policies, and cultural contexts evolve. This leads to:
    - Providing outdated or stale information as current.
    - Failing to recognize when previously true statements no longer hold.
    - Inability to reason about events post-training cutoff, limiting their utility in dynamic environments.
- **Causal Fragility**: LLMs implicitly learn causal relationships (e.g., "X causes Y") from statistical correlations in their training data, rather than through explicit causal understanding. This fragility manifests as:
    - Causal relationships changing or inverting over time (e.g., political alliances, economic dependencies).
    - Reliance on correlations rather than causal mechanisms, making reasoning brittle.
    - Lack of mechanisms to unlearn outdated causal links or update reasoning based on new causal evidence.

The research problem is to develop methods that mitigate temporal bias by enabling LLMs to incorporate new data continuously and enhance causal robustness by ensuring their reasoning adapts to changing causal structures.

### Review of Recent Literature

A review of 30–50 recent papers from NeurIPS, ICLR, ACL, and related venues (focusing on 2024–2025 publications) reveals several key themes and approaches:

### Temporal Generalization and Bias

- The paper *"Is Your LLM Outdated? A Deep Look at Temporal Generalization"* (arXiv, 2025) introduces the concept of temporal generalization, evaluating LLMs' ability to handle past, present, and future data. It found significant temporal biases, with performance declining over time, especially for powerful models in future generalization. The paper proposes **FreshBench**, a dynamically updated benchmark for testing LLMs' ability to predict future events, ensuring no data leakage or subjective bias. This work highlights the need for LLMs to adapt to evolving data environments, with open-source models showing better long-term adaptability than closed-source counterparts.

- Other resources, such as *"Data bias in LLM and generative AI applications"* (MOSTLY AI, 2023), discuss temporal or historical bias arising when training data is not representative of current contexts, emphasizing the need for human supervision to mitigate such issues.

**Causal Reasoning and Fragility**

- *"Causal Parrots: Large Language Models May Talk Causality But Are Not Causal"* (TMLR, 2023) argues that LLMs are not truly causal; they mimic causal reasoning by reciting patterns from training data. It introduces "meta SCMs" (Structural Causal Models) to explain why LLMs sometimes appear to reason causally, suggesting they are "causal parrots" with weak causal understanding. This directly addresses causal fragility, highlighting the brittleness of LLMs' causal reasoning when faced with changing patterns.
- *"Causal Reasoning and Large Language Models: Opening a New Frontier for Causality"* (TMLR, 2024) surveys LLMs' causal capabilities, showing they can generate text resembling causal arguments but often rely on correlations. It benchmarks LLMs on tasks like pairwise causal discovery (97% accuracy with GPT-4, 13 points gain over existing methods) and counterfactual reasoning (92%, 20 points gain), yet notes unpredictable failure modes, suggesting a need for robustness checks.
- *"The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective"* (arXiv, 2023) discusses endogeneity issues in LLM-simulated experiments, where prompt variations introduce confounding factors, leading to unreliable causal inferences. This underscores the fragility of LLMs in causal tasks, particularly in dynamic settings.

**Integrating Causality and Temporal Adaptation**

- *"Language Agents Meet Causality -- Bridging LLMs and Causal World Models"* (OpenReview, 2024) proposes a framework integrating causal representation learning (CRL) with LLMs to enable causally-aware reasoning. It learns a causal world model linked to natural language, acting as a simulator for LLMs to query, addressing both causal fragility and the need for dynamic reasoning.
- *"Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey"* (arXiv, 2024) explores how causal inference enhances LLMs' fairness, robustness, and explainability, while also noting LLMs' potential to aid causal discovery. It emphasizes the interplay between causal frameworks and LLMs, suggesting collective potential for advanced AI systems.
- The GitHub repository *CausalNLP_Papers* provides a comprehensive reading list, including papers like *"Can Large Language Models Build Causal Graphs?"* (arXiv, 2024) and *"Efficient Causal Graph Discovery Using Large Language Models"* (arXiv, 2024), which explore using LLMs for causal discovery, potentially addressing causal fragility by leveraging metadata and natural language.

**Empirical Findings and Benchmarks**

- Several papers, such as *"Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data"* (ACL Findings, 2024), evaluate LLMs' causal reasoning capabilities, finding limitations in handling complex causal tasks. Benchmarks like FreshBench and others mentioned in surveys provide frameworks, but comprehensive evaluation metrics for both temporal and causal generalization remain underdeveloped.
- The paper *"LLMs Are Prone to Fallacies in Causal Inference"* (arXiv, 2024) highlights specific failure modes, reinforcing the need for methods to improve causal robustness.

**Key Gaps and Future Directions**

Despite these advancements, several critical gaps remain unaddressed:

- **Continuous Updating of LLMs**: Current methods for updating LLMs, such as fine-tuning, are resource-intensive and not feasible for continuous application. Efficient, scalable methods to keep LLMs updated with new data without retraining from scratch are needed, as discussed in *"How would you bring an LLM up to date?"* (Ronan Laker, 2024).
- **Robust Causal Reasoning**: LLMs lack true causal understanding, relying on statistical correlations, making their reasoning fragile. Developing methods for LLMs to integrate explicit causal models (e.g., via CRL) or learn causal structures dynamically is essential, as suggested by *"Causality for Large Language Models"* (arXiv, 2024).
- **Adapting to Changing Causal Relationships**: Causal relationships can shift over time due to policy changes, technological advancements, or cultural shifts. LLMs need mechanisms to detect and adapt to these changes, which is currently underexplored, as noted in surveys like *"Bridging Causal Discovery and Large Language Models: A Comprehensive Survey of Integrative Approaches and Future Directions"* (arXiv, 2024).
- **Evaluation Metrics for Temporal and Causal Generalization**: While benchmarks like FreshBench exist, more comprehensive frameworks are needed to assess both temporal adaptability and causal robustness, ensuring LLMs can handle dynamic, real-world scenarios.
- **Integration with Domain Knowledge**: LLMs often lack domain-specific knowledge, crucial for accurate causal reasoning in fields like medicine or law. Methods to integrate domain knowledge or make LLMs more adaptable to different domains are needed, as highlighted in papers like *"Knowledge Graph Structure as Prompt: Improving Small Language Models Capabilities for Knowledge-based Causal Discovery"* (arXiv, 2024).
- **Handling Multimodality and Real-World Data**: Many real-world causal relationships involve multimodal data (e.g., text, images, sensor data). Research on how LLMs can handle causality in multimodal contexts is limited, as noted in surveys focusing on multimodality.

**Conclusion**

The research problem of addressing temporal bias and causal fragility in LLMs is multifaceted, requiring advancements in temporal generalization, causal reasoning, and integration with dynamic

causal models. Recent papers have made strides in evaluating temporal biases (e.g., FreshBench), understanding causal limitations (e.g., "Causal Parrots"), and proposing frameworks for causally-aware LLMs. However, significant gaps remain, particularly in continuous updating, robust causal understanding, and adapting to changing causal relationships. Future research must focus on these areas to enhance LLMs' reliability for real-world applications.

**Table: Summary of Key Papers and Findings**

| Paper Title | Venue/Year | Focus Area | Key Finding |
|---|---|---|---|
| Is Your LLM Outdated? A Deep Look at Temporal Generalization | arXiv, 2025 | Temporal Generalization | Significant temporal biases, introduces FreshBench for future event prediction. |
| Causal Parrots: Large Language Models May Talk Causality But Are Not Causal | TMLR, 2023 | Causal Fragility | LLMs mimic causal reasoning, weak "causal parrots" due to training data. |
| Causal Reasoning and Large Language Models: Opening a New Frontier for Causality | TMLR, 2024 | Causal Reasoning | LLMs perform well on benchmarks but have unpredictable failure modes. |
| Language Agents Meet Causality -- Bridging LLMs and Causal World Models | OpenReview, 2024 | Causal Integration | Proposes framework integrating CRL with LLMs for causally-aware reasoning. |
| Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey | arXiv, 2024 | Survey | Explores interplay, notes need for causal robustness and fairness. |

This table summarizes key papers reviewed, highlighting their contributions to understanding and addressing the research problem.

**Key Citations**

- Is Your LLM Outdated? A Deep Look at Temporal Generalization
- Causal Parrots: Large Language Models May Talk Causality But Are Not Causal
- Causal Reasoning and Large Language Models: Opening a New Frontier for Causality
- Language Agents Meet Causality -- Bridging LLMs and Causal World Models
- Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey
- Data bias in LLM and generative AI applications
- How would you bring an LLM up to date?
- CausalNLP_Papers reading list