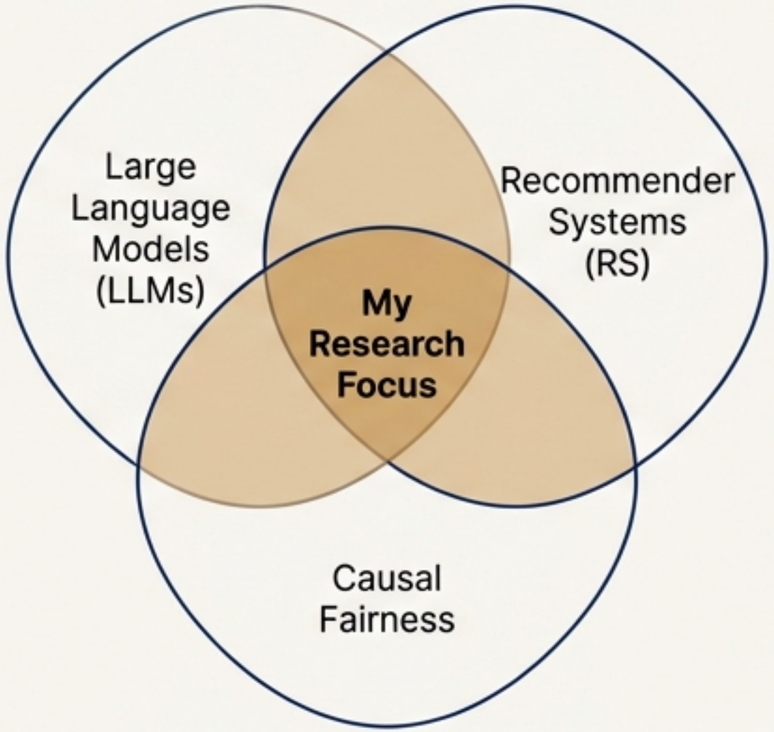
# My PhD Roadmap: Causal Fairness in LLM-Enhanced Recommender Systems

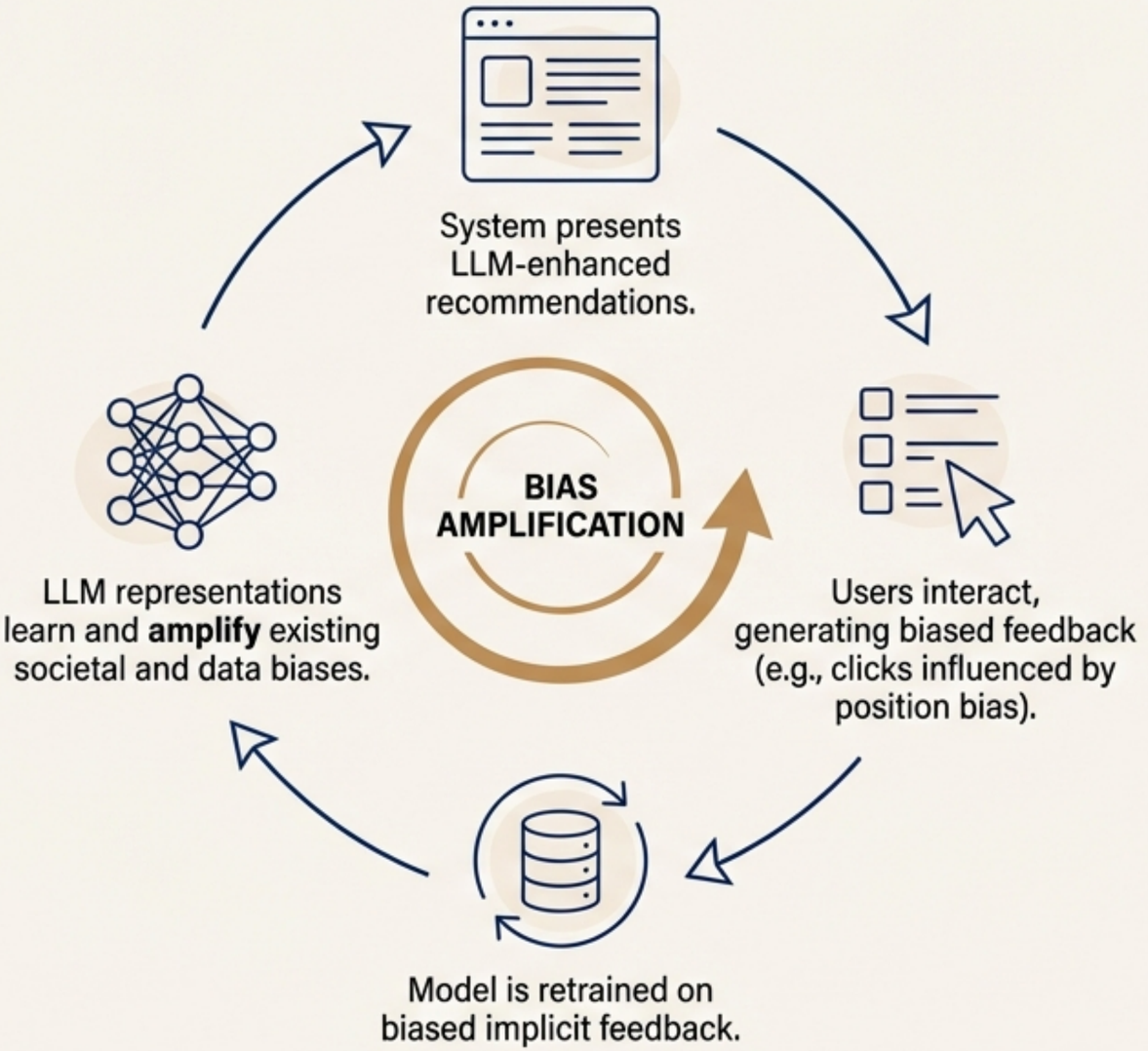
A 3-Year Plan for Analyzing and Mitigating Algorithmic Bias Amplification




## Thesis Statement

My research moves beyond traditional correlation-based metrics to establish a new paradigm for fairness in recommender systems. I will use **causal analysis** to model, measure, and mitigate how **Large Language Models (LLMs)** can amplify algorithmic bias. The goal is to develop a novel framework that builds more resilient, multi-stakeholder-aware, and trustworthy AI systems.

# The Problem: Bias Amplification in the LLM-RS Feedback Loop




## The Challenge: Why Current Systems Fail

**Representational Harm:** LLMs can perpetuate and amplify stereotypes in user and item embeddings, a key source of unfairness identified in the literature.

**The Feedback Loop:** This 'rich-get-richer' dynamic intensifies initial biases, reducing diversity and fairness over time. It can disproportionately favor certain providers and disadvantage others.

**Correlational Blinders:** Existing fairness metrics often measure symptoms (statistical disparities) but cannot identify the root causes, limiting the effectiveness of interventions. Causal analysis is required.

## My Contribution: A PhD by Publication

**Publication 1: Causal Analysis & Measurement**
*A Causal Framework for Quantifying Bias Amplification in LLM-Enhanced Recommender Systems.*

**Publication 2: LLM-Driven Mitigation**
*Breaking the Loop: An LLM-Based Re-ranking Strategy for Multi-Stakeholder Fairness.*

# The Roadmap: Year 1 - Building the Foundation



Phase 1: Core Fluency (Months 1-4)

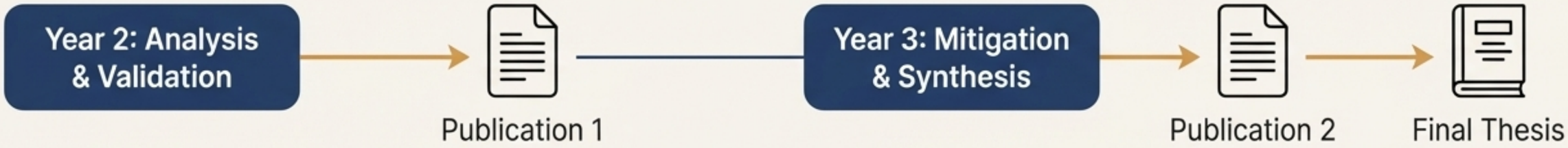
Phase 2: Fairness Theory (Months 5-8)

Phase 3: Framework Design (Months 9-12)

| Milestone | Study Plan (What to Learn) | Action Plan (How to Apply) |
|---|---|---|
| **M1.1: Core Domain Fluency** | • **Recommender Systems:** Traditional (Content-Based, Collaborative Filtering) vs. modern architectures.<br>• **LLMs:** Foundational architectures (BERT, GPT), fine-tuning, and prompt tuning. | • **System Setup:** Implement baseline RS models.<br>• **Dataset Selection:** Acquire and analyze a suitable fairness dataset (e.g., MovieLens, TREC Fair Ranking track). |
| **M1.2: Fairness Theory Deep Dive** | • **Core Concepts:** Individual vs. Group Fairness.<br>• **Key Metrics**: **Opportunity-based** fairness and **Exposure** Fairness (Expected Exposure, EUR, RUR) for provider utility. | • **Literature Review:** Complete comprehensive review on fairness in RS.<br>• **Bias Identification:** Quantify pre-existing popularity and representational biases in the chosen dataset. |
| **M1.3: Causal Framework Design** | • **Causal Inference**: Causal graphs, interventions, and counterfactuals to move beyond correlation.<br>• **LLM Bias Amplification:** How bias manifests in learned latent representations. | • **Proposal Draft**: Formalize the **causal graph** illustrating how LLM enhancement influences distributional and representational harms.<br>• **Proof-of-Concept**: Experiment using an LLM to generate embeddings 'disentangled from sensitive attributes'. |

**Key Year 1 Outcome:** A robust experimental testbed and a formalized PhD proposal, grounding all future work in theory and practice.

# The Roadmap: Years 2 & 3 - From Analysis to Impact




## Year 2: Empirical Validation & Publication 1

**Objective:** To empirically validate my causal models and quantify LLM-driven bias amplification.

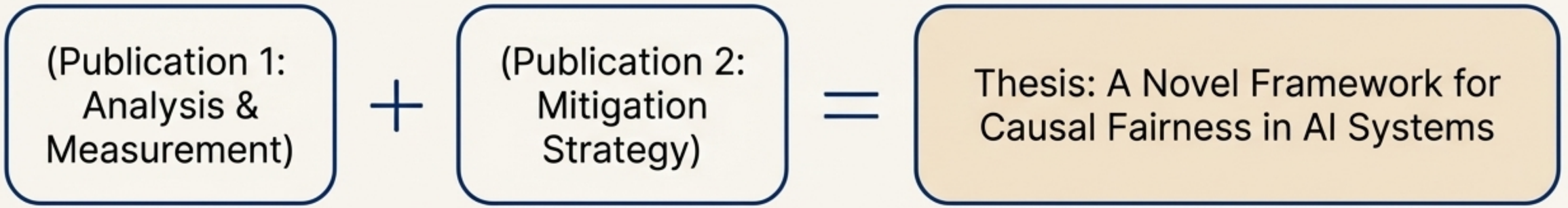
- **Causal Analysis**: Implement my framework to measure how LLM embeddings causally influence representational and distributional harms.
- **LLM Auditing:** Use LLMs to simulate interventions and audit user feedback, differentiating between 'unseen' and 'uninterested' items to better model exposure bias.
- **Quantify Amplification**: Measure how the feedback loop intensifies baseline biases over time, forming the core empirical result of my first paper.
- **Key Output: 🎯 Draft and submit Publication 1.**

## Year 3: Mitigation, Synthesis & Thesis Defense

**Objective:** To develop my novel mitigation strategy and synthesize all research into the final thesis.

- **Mitigation Design**: Develop an **LLM-driven, fairness-aware re-ranking module**. Design prompts that instruct the LLM to optimize for multi-stakeholder utility (e.g., user relevance subject to provider exposure constraints).
- **Validate Solution:** Evaluate the re-ranking strategy against baseline methods, demonstrating its ability to break the feedback loop while managing utility/fairness trade-offs.
- **Key Output: 🎯 Draft and submit Publication 2.**
- **Final Synthesis:** Integrate both publications into a cohesive thesis document for submission and defense.

# Thesis Synthesis & Contribution to Resilient AI




### The Synthesized Thesis

My final thesis will articulate a novel, end-to-end framework for achieving **causal fairness** in LLM-enhanced recommender systems. It will present a cohesive argument for using LLMs not just as a source of bias, but as a sophisticated tool for a just as a source of bias, but as a sophisticated tool for both its **causal analysis** and **its principled mitigation**, marking a significant departure from purely statistical approaches.

### Broader Impact: Fairness as Sociotechnical Cyber Resilience

This research frames algorithmic fairness not just as an ethical imperative, but as a **critical component of system resilience**. A fair system is inherently more robust because it is less vulnerable to:

- **Adversarial Manipulation**: Unfair models with predictable biases are prime targets for attacks designed to poison data or exploit stereotypes to manipulate system outcomes.
- **User Distrust & System Failure**: Biased results erode user trust, a key operational risk that threatens a system's viability and long-term user engagement.
- **Regulatory & Legal Risk**: As regulations evolve, a lack of demonstrable fairness constitutes a direct threat to a system's operational license and opens the door to legal liabilities.