Causal Investigations of Compound AI System Behaviors:
Applications in Synthetic Audio Detection and Air Combat Training Scenario Generation

Anonymous Submission

The state-of-the-art (SOTA) for evaluating multimodal AI systems revolves around the development and application of benchmarks that typically compare a model's performance on a certain task to the performance we would expect from a human (Li et al., 2024; Meng et al., 2024; Xun et al, 2025). While benchmarks show proficiency with respect to a certain task, they fail to explore architectural components or input parameters that contribute to observed outcomes. Moreover, modern AI systems go beyond models, e.g., LLMs and VLMs, and include autonomous agents, agentic AI workflows, and human-agent interactions (Volkova et al., 2024; Zaharia et al., 2024). Thus, they require a complete rethink of AI evaluation paradigms that moves beyond static task performance to encompass dynamic behavioral assessment, explainability of decision-making processes, and robustness across diverse operational contexts (Cohen et al., 2025; Volkova et al., 2025; Nguyen et al., 2025). This gap becomes particularly critical when deploying AI systems in high-stakes environments where understanding the causal relationships between system design and performance outcomes is essential for ensuring reliability, safety, and mission success.

Causal discovery is an approach for determining cause-and-effect relationships between variables and outcomes enabling discussion of causation between input parameters and outcomes instead of only identifying correlations (Pearl, 2009; Pearl & Mackenzie, 2018). Causal methods allow the operators to study emergent behaviors, i.e., first- and second-order effects, and makes AI systems more explainable – an increasingly desirable feature. Our work utilizes causal methods in two downstream tasks: a synthetic audio detection task and military training scenario generation task, to identify and explain relationships between compound AI system components e.g., model, agent, tool, inputs and outputs. Both tasks utilized a causal discovery pipeline that includes Structural Equation Modeling (SME) (Pearl, 2012) and Average Treatment Effect Estimation (ATE) (Rubin, 1974).

The synthetic audio detection task utilizes data from the Synthetic Audio Forensics Evaluation (SAFE) Challenge (Trapeznikov et al., 2025) held earlier this year. Participants were tasked with building models that could distinguish pristine samples from generated ones. Our causal discovery task aims to explain why certain models performed better than others and what factors influenced their performance across different audio samples. Our initial results show the relationships between positive ATE scores for performer analytics that performed well in the challenge with the winning team/analytics have the highest average ATE scores overall. Ongoing work is expanding this analysis to include results from a causal relationships analysis between audio sample features and model performance. At the workshop, we will present these results and discuss insights this analysis has afforded us.

The scenario generation task utilized large language models (LLMs) to transform prompts into training scenarios for combat pilots. We constructed a dataset designed for causal analysis to determine which prompt features led to specific scenario characteristics. Prompts were systematically designed with predetermined features, curated via meta-prompting, and fed into three LLMs (Gemma Team, 2025; Anthropic, 2024; OpenAI, 2024) utilizing different knowledge corpora via Retrieval Augmented Generation (Lewis et al., 2020). This approach generated six responses per prompt, enabling characterization of model and corpus selection impacts on scenario generation. Preliminary qualitative analysis reveals that Gemma 3 produces robust responses with accurate corpus retrieval but suffers from extended response times, Claude 3.5 Sonnet demonstrates higher abstention rates for uncomfortable prompts, while ChatGPT 4o-Mini exhibits the greatest tendency to extrapolate beyond corpus limitations.

Our causal investigation framework demonstrates the potential for moving beyond correlation-based AI evaluation toward mechanistic understanding of system behaviors in critical applications. Future work will expand the synthetic audio analysis to include comprehensive causal relationships between audio features and model performance, while the scenario generation task will incorporate quantitative causal discovery methods to systematically identify optimal model-corpus configurations for military training applications, ultimately advancing explainable AI for high-stakes domains.