N8 CIR
COMPUTATIONALLY INTENSIVE RESEARCH

University of Sheffield | IT Services

Evaluating AI Workflows using Causal Testing

Dr. Farhad Allian

University of Sheffield

farhad.allian@sheffield.ac.uk | www.farhadallian.co.uk

CITCOM

1. Testing Software Is Hard; Testing AI Is Harder

Modern research software is arguably becoming centered on AI development for scientific discovery. This puts software testing at the heart of trustworthy research as the best way to deliver transparency and reproducibility when using these black-box systems. However, their large parameter spaces, randomness, and lack of "ground truth" deem traditional testing approaches challenging. Instead, causal inference can be leveraged to analytically test an AI system's expected cause-effect relationships. This poster presents **causal-ai** – a causal-inference tool that evaluates AI workflows using existing training/inference data.

2. What Is Causal Testing?

Causal testing is a causal inference-driven framework that utilises **Directed Acyclic Graphs** (DAG) to model the system's expected input-output relationships. Each relationship then gets tested statistically against experimental data. An example DAG of an AI system could look like the following:

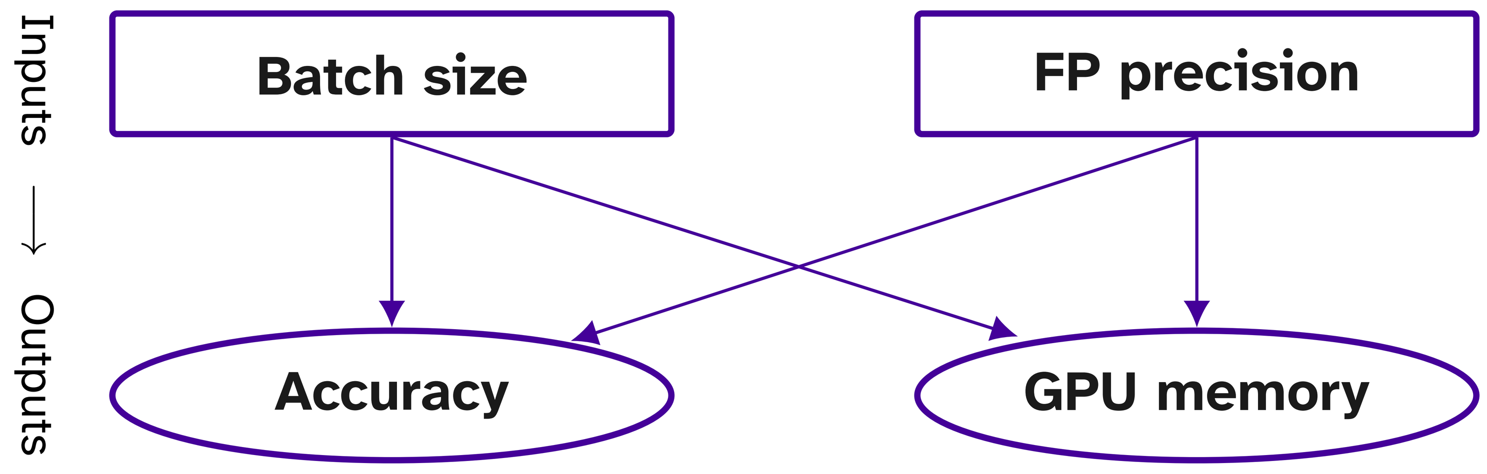

Figure 1: A DAG representing the expected cause-effect relationships of a simple AI system.

3. The causal-ai Tool

The `causal-ai` tool was developed to enable **robust** and **tractable** testing of AI workflows on HPC clusters. The pipeline can be summarised in 3 steps:

1. **Experimental Data Collection** – Log parameters and metrics during training or inference runtimes.
2. **Model Causal Relationships** – Encode the expected input-output relationships using a DAG.
3. **Evaluate Causal Test Cases** – Review the causal test cases and summaries that have passed/failed based on statistical tests.

The tool can be hosted centrally on HPC systems and readily loaded using `module load causal_ai`. Built-in visualisation features improve the accessibility and interpretability of the causal test results.

4. How to Evaluate a Workflow using Causal Test Cases

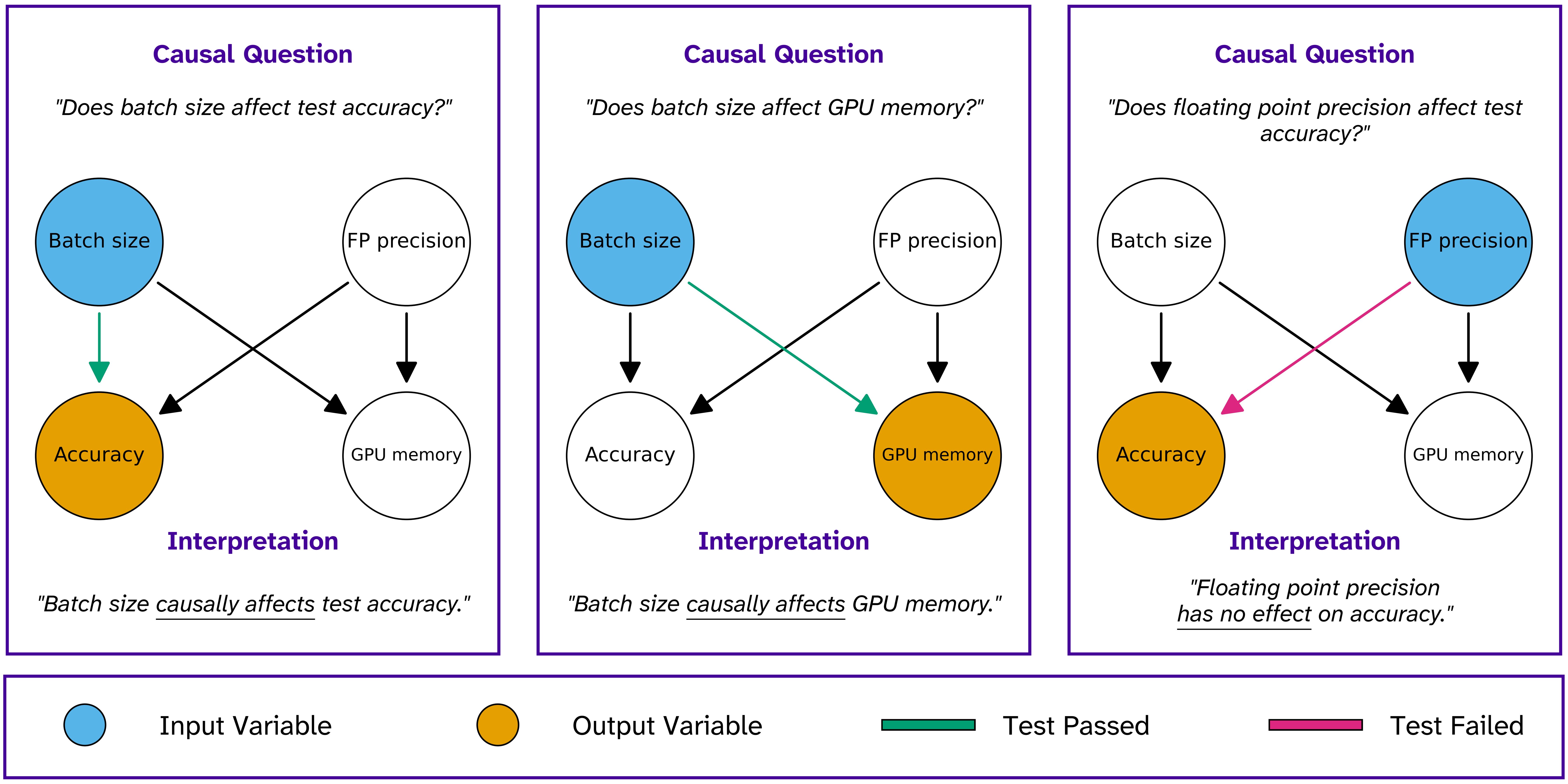

Figure 2: Sample causal test results identifying the effect of batch size and floating point precision on model accuracy and GPU memory in a domain adaptation digit classification pipeline.

5. Three Real-World Applications

This tool was applied to three real-world AI pipelines: digit classification, video domain adaptation and multi-source image classification. Each one uses the same number of observations and underlying DAG, which was validated on two different HPC clusters (N8/Bede and Stanage/Sheffield). Cross-cluster comparisons reveal platform-dependent hardware effects that a single baseline performance metric would obscure. Causal testing thus distinguishes **genuine effects** from **spurious artefacts**.

6. Future Of Research Software

AI models exhibit complex behaviours that traditional testing cannot adequately characterise, limiting confidence in their findings. Causal testing addresses this challenge by identifying the cause-effect relationships driving model behavior, providing a systematic approach to **verifiable research**. Because rigorous testing is rarely prioritised in research, research software engineers are uniquely positioned to bridge this gap, making AI-driven scientific discovery fundamentally **trustworthy** and **reproducible**.

7. Causal Test It Out For Yourself!

1. N8/Bede Case Study
2. Software, documentation & tutorials
3. Causal Testing Framework

Acknowledgments: This project was funded by EPSRC through N8/Bede: AI4Science. The `causal-ai` tool is built on the Causal Testing Framework developed by the CITCOM group (Dr. Neil Walkinshaw and Dr. Michael Foster) at University of Sheffield.