universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Exploring the Capabilities of Large Language Models in Causal Inference

verfasst von | submitted by

Andres Vladimir Arostegui Arias

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 645
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Data Science
Betreut von | Supervisor:	RNDr. CSc. Katerina Schindlerova Privatdoz.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, RNDr. Katerina Schindlerova, CSc. Privatdoz., for her guidance and support throughout this thesis. Her expertise in causal inference and thoughtful feedback helped shape this work in meaningful ways. I am especially grateful for her willingness to explore new directions and for encouraging me to tackle challenging questions about causal reasoning in AI systems.

I am deeply thankful to my family in Ecuador, whose support and encouragement have been constant throughout my studies in Vienna. Despite the distance, they have always believed in me and made this journey possible.

I also want to acknowledge the friends I made during this Master's program. Their collaboration, insights, and companionship made this challenging journey more rewarding. The discussions we shared, both academic and personal, contributed significantly to my growth as a researcher and as a person.

This thesis represents work that has been prepared for publication, reflecting an ongoing effort to understand how modern AI systems reason about causality.

Use of AI Tools. In accordance with the University of Vienna's guidelines on good academic practice, the following AI-based tools were used during the preparation of this thesis: *Claude* (Anthropic) was used for writing assistance, including proofreading, text editing, and generating the German abstract (Kurzfassung). *ChatGPT* (OpenAI) was used for brainstorming and drafting support during early stages of writing. *GitHub Copilot* was used for coding assistance in the development of the evaluation pipeline and Streamlit application. All AI-generated content was reviewed, verified, and revised by the author. The intellectual contributions, research design, analysis, and conclusions are entirely my own work.

Abstract

Knowing what causes what matters for making good decisions. In economics, health, and engineering, the causal direction between two variables can determine which interventions work. But in many real-world systems, this direction is not fixed—it can reverse depending on the conditions. For example, below a certain threshold, inflation may drive interest rate changes; above it, the relationship flips. Detecting such regime-dependent causality from data alone is difficult, and most existing methods assume a single, stable causal direction.

This thesis poses a question, whether Large Language Models (LLMs) can identify causal direction from numerical data, and how they compare to established data-driven algorithms. We test GPT-5.2 against three causal discovery methods (ROCHE, LOCI, and LCUBE) on five bivariate datasets where the causal direction reverses between regimes. The datasets span synthetic validation data, automotive engineering, energy systems, and monetary policy. Each regime boundary comes from domain research, providing ground truth for evaluation.

To compare these different approaches fairly, we define a regime-level accuracy metric and test the LLM under four prompt configurations—varying whether variable names are visible and whether data is pre-segmented by regime.

The results show a clear gap. ROCHE and LCUBE achieve 80% mean accuracy; GPT-5.2 achieves 30%. The algorithms correctly detect regime-switching on 60% of datasets; the LLM achieves 0%. The LLM fails in two distinct ways: on most datasets, it ignores regime structure and predicts a single global direction; on economic datasets, it detects a switch but predicts both directions backwards. When meaningful variable names are provided, LLM accuracy improves on 2 of 5 datasets (from 50% to 100%), suggesting the model draws on word associations from its training data rather than analyzing the numerical patterns it receives.

These findings carry a practical message: current LLMs are not reliable for regime-dependent causal discovery from numerical data. For this task, data-driven algorithms remain the appropriate choice. To support further research, we release an open-source Python tool implementing the complete evaluation pipeline.

Kurzfassung

Die Frage nach Ursache und Wirkung ist in vielen Bereichen entscheidend. In der Wirtschaft, Medizin und Technik bestimmt die kausale Richtung zwischen zwei Variablen, welche Maßnahmen funktionieren. Aber in vielen realen Systemen ist diese Richtung nicht stabil. Sie kann sich je nach Bedingungen umkehren. Zum Beispiel: Unterhalb eines Schwellenwerts kann die Inflation die Zinsen beeinflussen. Oberhalb davon kehrt sich die Beziehung um. Es ist schwierig, solche Wechsel nur aus Daten zu erkennen. Die meisten Methoden gehen davon aus, dass die kausale Richtung konstant bleibt.

Diese Arbeit untersucht, ob große Sprachmodelle die kausale Richtung aus numerischen Daten erkennen können. Wir vergleichen sie mit etablierten datenbasierten Algorithmen. Wir vergleichen GPT-5.2 mit drei Methoden (ROCHE, LOCI und LCUBE) auf fünf Datensätzen mit zwei Variablen. Bei diesen Datensätzen wechselt die kausale Richtung zwischen verschiedenen Regimen. Die Datensätze umfassen synthetische Daten, Fahrzeugtechnik, Energiesysteme und Geldpolitik. Jede Regimegrenze stammt aus der Fachliteratur und dient als Grundwahrheit.

Für einen fairen Vergleich definieren wir eine Metrik auf Regimeebene. Wir testen das Sprachmodell unter vier Konfigurationen. Diese unterscheiden sich darin, ob Variablennamen sichtbar sind und ob die Daten vorab nach Regime getrennt wurden.

Die Ergebnisse zeigen deutliche Unterschiede. ROCHE und LCUBE erreichen eine mittlere Genauigkeit von 80%. GPT-5.2 erreicht 30%. Die Algorithmen erkennen den Regimewechsel bei 60% der Datensätze korrekt. Das Sprachmodell erreicht 0%. Das Sprachmodell scheitert auf zwei Arten: Bei den meisten Datensätzen ignoriert es die Regimestruktur und sagt eine einzige globale Richtung vorher. Bei wirtschaftlichen Datensätzen erkennt es einen Wechsel, sagt aber beide Richtungen vertauscht vorher. Wenn aussagekräftige Variablennamen gegeben werden, verbessert sich die Genauigkeit bei 2 von 5 Datensätzen (von 50% auf 100%). Das deutet darauf hin, dass das Modell auf Wortverbindungen aus seinen Trainingsdaten zurückgreift und nicht die numerischen Muster analysiert.

Die praktische Schlussfolgerung ist: Aktuelle Sprachmodelle sind nicht zuverlässig für regimeabhängige kausale Inferenz aus numerischen Daten. Für diese Aufgabe sind datenbasierte Algorithmen die bessere Wahl. Um weitere Forschung zu unterstützen, stellen wir ein Open-Source Python-Tool bereit, das die komplette Auswertungspipeline implementiert.

Diese deutsche Kurzfassung wurde mithilfe eines großen Sprachmodells (Claude, Anthropic) erstellt und anschließend vom Autor überprüft und überarbeitet.

Contents

Acknowledgements i

Abstract iii

Kurzfassung v

List of Tables xi

List of Figures xiii

List of Algorithms xv

Listings xvii

1. Introduction 1
- 1.1. Motivation and Context . . . 1
- 1.2. Problem Statement . . . 3
- 1.3. Scope and Assumptions . . . 3
- 1.4. Contributions . . . 4
- 1.5. Research Questions . . . 4
- 1.6. Thesis Roadmap . . . 5

2. Related Work 7
- 2.1. Causal Inference Methods . . . 7
  - 2.1.1. Pearl's Causal Inference Hierarchy . . . 7
  - 2.1.2. Data-Driven Causal Discovery Algorithms . . . 8
- 2.2. LLMs and Causality . . . 9
  - 2.2.1. Numerical Data Representation Challenges . . . 10
- 2.3. Regime-Dependent Causality . . . 10
- 2.4. Prompt Engineering for Causal Tasks . . . 11
- 2.5. Chapter Summary . . . 12

3. Datasets and Evaluation Framework 13
- 3.1. Synthetic Baseline Dataset: Stress-Fatigue . . . 13
- 3.2. Regime-Level Evaluation Framework . . . 13
  - 3.2.1. Regime-Level Accuracy Metric . . . 15
- 3.3. Real-World Datasets . . . 16
  - 3.3.1. Auto MPG and Horsepower . . . 16

3.3.2. Renewable Energy Share and CO_2 Emissions . . . 16
3.3.3. UK Short-Term and Long-Term Interest Rates . . . 16
3.3.4. US Inflation Rate and Federal Funds Rate . . . 19
3.4. Dataset Summary . . . 20
3.5. Data Preprocessing . . . 21
3.6. Chapter Summary . . . 21

4. Data-Driven Causal Discovery **23**
4.1. ROCHE Method . . . 23
4.1.1. Algorithm Foundation . . . 23
4.1.2. Implementation and Parameters . . . 23
4.2. LOCI Method . . . 24
4.2.1. Algorithm Foundation . . . 24
4.2.2. Implementation and Parameters . . . 24
4.3. LCUBE Method . . . 24
4.3.1. Algorithm Foundation . . . 24
4.3.2. Implementation and Parameters . . . 25
4.3.3. Comparison with LOCI and ROCHE . . . 25
4.4. Evaluation Workflow . . . 25
4.5. Experimental Setup . . . 26
4.5.1. Data Preparation . . . 26
4.5.2. Regime Application . . . 26
4.5.3. Testing Protocol . . . 27
4.5.4. Performance Metrics . . . 27
4.6. Results and Analysis . . . 27
4.6.1. Synthetic Dataset Results . . . 27
4.6.2. Real-World Dataset Results . . . 27
4.6.3. Method Comparison . . . 28
4.6.4. Regime-Switching Detection . . . 28
4.7. Chapter Summary . . . 28

5. LLM Evaluation Framework **31**
5.1. Motivation and Research Questions . . . 31
5.2. Model Selection . . . 31
5.3. Multi-Step Prompting Strategy . . . 31
5.3.1. Design Rationale . . . 32
5.3.2. Four-Step Architecture . . . 32
5.3.3. Prompt Chain Visualization . . . 32
5.3.4. Algorithm vs LLM Input Comparison . . . 34
5.3.5. Prompt Refinement Process . . . 34
5.4. Experimental Configurations . . . 35
5.4.1. Configuration Design . . . 35
5.4.2. Configuration Rationale . . . 36

5.5. Data Representation for LLMs . . . 36
5.5.1. Numerical Data Encoding . . . 36
5.5.2. Regime Information Encoding . . . 36
5.6. Evaluation Protocol . . . 37
5.6.1. Output Parsing . . . 37
5.6.2. Evaluation Criteria . . . 37
5.7. Comparison with Algorithmic Results . . . 38
5.8. Implementation Details . . . 38
5.9. Results . . . 38
5.9.1. Regime-Switching Detection . . . 39
5.9.2. Configuration Effects . . . 39
5.9.3. Comparison Summary . . . 40
5.10. Chapter Summary . . . 40

6. Experimental Results and Comparison **41**
6.1. Experimental Objective . . . 41
6.2. Performance Overview . . . 41
6.2.1. Baseline Algorithm Performance . . . 42
6.2.2. LLM Performance . . . 42
6.3. Dataset-Level Analysis . . . 42
6.3.1. Synthetic Baseline Results . . . 42
6.3.2. Real-World Dataset Results . . . 43
6.3.3. Performance Patterns . . . 43
6.4. Configuration Effects on LLM Performance . . . 43
6.4.1. Variable Naming Effects . . . 44
6.4.2. Data Segmentation Effects . . . 44
6.4.3. Configuration Comparison . . . 44
6.5. Regime-Switching Detection . . . 45
6.5.1. Detection Rates . . . 45
6.5.2. Failure Pattern Analysis . . . 45
6.6. LLM Reasoning Quality Analysis . . . 46
6.6.1. Successful Reasoning Examples . . . 46
6.6.2. Failure Mode Examples . . . 46
6.6.3. Inverted Failure Example . . . 46
6.6.4. Same-Direction vs Inverted Failures . . . 50
6.6.5. Common Failure Modes . . . 50
6.7. Method Complementarity Analysis . . . 50
6.7.1. Domain-Specific Observations . . . 51
6.7.2. Best Performers by Domain . . . 51
6.7.3. Algorithm Agreement Analysis . . . 52
6.8. Visualization of Results . . . 52
6.9. Summary of Findings . . . 52
6.9.1. Key Results . . . 52

6.9.2. Implications . . . 54
6.10. Chapter Summary . . . 54

7. Discussion and Conclusion **55**
7.1. Discussion of Findings . . . 55
7.1.1. Data-Driven Methods Outperform LLMs on Numerical Causality . 55
7.1.2. LLM Reliance on Semantic Knowledge . . . 55
7.1.3. The Inverted Failure Problem . . . 56
7.1.4. Regime-Switching as a Hard Problem for LLMs . . . 56
7.1.5. Value of Structured Prompting . . . 56
7.2. Answering the Research Questions . . . 57
7.2.1. RQ1: Algorithmic Performance Comparison . . . 57
7.2.2. RQ2: LLM Causal Reasoning Capabilities . . . 57
7.2.3. RQ3: Configuration Effects . . . 57
7.3. Contributions . . . 57
7.4. Limitations . . . 58
7.4.1. Dataset Scope . . . 58
7.4.2. Model Selection . . . 58
7.4.3. Ground Truth Challenges . . . 58
7.4.4. Prompt Sensitivity . . . 59
7.4.5. API Constraints . . . 59
7.5. Future Work . . . 59
7.5.1. Expanded Model Evaluation . . . 59
7.5.2. Multivariate Extension . . . 59
7.5.3. Hybrid Approaches . . . 59
7.5.4. Improved Prompting Strategies . . . 60
7.5.5. Richer Datasets . . . 60
7.6. Conclusion . . . 60

Bibliography **63**

A. Development of the Comparison Framework **67**
A.1. Final Application Architecture . . . 67
A.2. Key Design Decisions . . . 68
A.2.1. From Clustering to Literature-Based Thresholding . . . 68
A.2.2. Prompt Chain Optimization . . . 69
A.2.3. Configuration Preset System . . . 69
A.3. Batch Experiment Pipeline . . . 69
A.4. Development Timeline . . . 71

List of Tables

3.1. Summary of datasets with regime thresholds and ground-truth causal directions. 21

4.1. Comparison of data-driven causal discovery algorithms. 25
4.2. Regime-level accuracy by algorithm and dataset. 27
4.3. Regime-switching detection. "Correct" = detected reversal with correct directions. "Inverted" = detected reversal but directions backwards. "Same Dir" = inferred same direction for both regimes. 28

5.1. GPT-5.2 regime-level accuracy by dataset (baseline configuration). 39
5.2. Regime-switching detection across all methods. All datasets have true reversals. 39
5.3. Method comparison: mean accuracy and regime-switching detection. . . . 40

6.1. Overall method performance across five datasets (baseline configuration). Perfect = both regimes correct; Failed = both regimes incorrect. 41
6.2. Method performance by dataset using `segmented_anon_hint_explicit` configuration. GPT-5.2 predictions show inferred direction and confidence for low/high regimes. Anonymous variables shown as Variable_X/Variable_Y in LLM predictions. 42
6.3. Regime-dependent direction switching detection. All datasets exhibit true causal direction reversal between regimes. "Correct" = detected reversal with correct directions (100% accuracy); "Inverted" = detected reversal but directions are backwards (0% accuracy); "Same Dir" = infers identical direction for both regimes (50% accuracy). GPT-5.2 results shown for the baseline `segmented_anon_hint_explicit` configuration; other configurations may differ per dataset. 45
6.4. Method accuracy by domain category. *GPT-5.2 range reflects variation across configurations (anonymous vs. named variables). 51

A.1. Key modules and their roles in the evaluation pipeline. 68
A.2. Experimental configuration matrix. Each cell is an independently evaluated preset. 69
A.3. Development timeline of the `causal_inference_app` framework. 71

List of Figures

1.1. Comparative workflow: (a) a data-driven algorithm receives standardised numeric arrays per regime; (b) the LLM receives data samples and threshold context via a 4-step prompting chain. Both produce per-regime causal direction outputs. 2

3.1. Synthetic stress-fatigue dataset. Blue points represent the low regime ($S < 0.5$) where $F \to S$, and red points represent the high regime ($S \geq 0.5$) where $S \to F$. The purple dashed line marks the threshold at $S = 0.5$. . . 14
3.2. Auto MPG and horsepower relationship showing regime-dependent causal structure. The threshold at 140 HP (midpoint of 120–160 HP transition region) separates consumer-driven design (MPG $\to H$) from performance-driven engineering ($H \to$ MPG). 17
3.3. CO_2 emissions per capita vs renewable electricity share showing regime-dependent causal structure. Below 15% renewable share, emissions drive renewable investment; above 15%, renewables actively reduce emissions. . 18
3.4. UK short-term vs long-term interest rates showing regime-dependent causal structure. The threshold at 3.5% (midpoint of 3–4% range) separates policy-driven expectations ($S \to L$) from market-driven dynamics ($L \to S$). 19
3.5. US inflation rate vs federal funds rate showing regime-dependent causal structure. The threshold at 5.4% inflation (midpoint of 5.3–5.5% range) separates inflation-driven policy ($I \to F$) from policy-driven inflation control ($F \to I$). 20

5.1. Four-step prompting chain for regime-dependent causal inference. Each step builds on previous analysis while preventing premature commitment to a causal direction. 33
5.2. Comparative workflow: (left) baseline algorithms always receive standardized numeric arrays per regime; (right) LLM input varies by configuration—shown here is `segmented_anon_hint_explicit`, which provides per-regime statistics, data samples, and explicit threshold specification. Both produce per-regime causal direction outputs. 34

6.1. GPT-5.2 accuracy across five datasets and four experimental configurations. Darker cells indicate higher accuracy. Named configurations (left columns) generally outperform anonymous (right columns), while segmentation effects vary by dataset. 44

6.2. Successful LLM reasoning example: Emissions-Renewables dataset with `segmented_named_hint_explicit` configuration. The model correctly leverages domain knowledge about renewable energy infrastructure maturity, correctly identifying the direction reversal between regimes (X→Y in low regime, Y→X in high regime). 47
6.3. Failed LLM reasoning example (Same-Direction): Synthetic Stress-Fatigue dataset with `segmented_anon_hint_explicit` configuration. Ground truth is X→Y in low regime and Y→X in high regime (direction reversal). The model incorrectly predicts Y→X for *both* regimes, failing to detect the designed causal reversal. Without semantic variable names, the model defaults to interpreting statistical patterns as a "gating mechanism" rather than analyzing regime-specific causal asymmetries. 48
6.4. Inverted failure example: UK Interest Rates dataset with `segmented_named_hint_explicit` configuration. The model correctly detects that causal direction reverses between regimes but predicts the directions exactly backwards, achieving 0% accuracy. This "inverted" failure pattern appears on both economic datasets, suggesting the model applies incorrect domain knowledge about monetary policy mechanisms. 49
6.5. Regime-dependent causal structure for each dataset. Blue points represent the low regime and red points represent the high regime. Yellow dashed lines show regime-low fitted slopes with their causal directions, while green dashed lines show regime-high slopes. Purple dotted lines mark the threshold value. 53

A.1. Directory structure of the `causal_inference_app` framework. 67
A.2. Prompt log export structure. Each step's full prompt and LLM response is preserved for auditability. 70
A.3. Structured JSON output parsed from the LLM's Step 4 response. Per-regime directions are compared against ground truth to compute accuracy. 70

List of Algorithms

1. Regime-based causal direction inference 26

Listings

1. Introduction

1.1. Motivation and Context

Correlation tells us that two variables move together, but not why. Causal inference goes further: it identifies cause-and-effect relationships and the mechanisms that explain why events occur. Causation has direction: a change in X may cause a change in Y, but not necessarily the other way around. Understanding causal direction matters in economics, health, environmental science, and engineering because good decisions depend on knowing what causes what. For instance, knowing whether interest rate changes drive inflation or inflation forces central banks to react determines whether policy should target lending conditions or price stability [GMP25].

Large Language Models (LLMs) show potential in reasoning about such relationships. LLMs learn from large text datasets and can summarize, answer questions, and solve reasoning tasks. But when it comes to math and logic, they often rely on pattern matching rather than genuine reasoning, and this has been observed across multiple benchmarks and task types [SBV$^+$24, FPC$^+$23]. Small changes in how a problem is phrased can cause their performance to drop [MAS$^+$25]. This suggests that LLMs may struggle with genuine reasoning even when they perform well on simpler tasks.

In many fields, randomised controlled trials are either impossible or unethical, so researchers must work with observational data. Causal discovery methods address this by recovering directional relationships from data alone, without experimental manipulation. Over the past decade, identifiability theory has clarified when and how causal direction can be recovered from bivariate observations [ISV$^+$23], and benchmark datasets such as the Tübingen Cause–Effect Repository have provided common ground for method evaluation [MPJ$^+$16]. These advances rely on statistical assumptions about noise and functional form. LLMs offer a different approach: they draw on knowledge from large text collections, but whether they can reason about numerical patterns remains an open question.

Classical causal inference methods like structural causal models (SCMs), additive noise models (ANMs), or location–scale noise models (LSNMs) use mathematical assumptions to infer direction from observational data. These models are formulated as structural equations, and in modern causal inference are often described as structural causal models (SCMs) to emphasize their causal interpretation [KN21]. For instance, ANMs are a specific case where the effect is written as $Y = f(X) + N$, with the noise term independent of the cause. This setup is identifiable, meaning the correct direction can be recovered under certain conditions, such as non-linear functions or non-Gaussian noise [ISV$^+$23]. LSNMs extend this idea by also letting the variance depend on the cause, written as

Algorithm Input & Execution

```
# Global: standardize before regime
    split
scaler = StandardScaler()
df[[X, Y]] = scaler.fit_transform(df[[
    X, Y]])

# Per-regime: extract 1D arrays
x = regime_data[X].values  # 1D array
y = regime_data[Y].values  # 1D array

# Algorithms receive NO context:
# - No variable names/semantics
# - No threshold information
# - No regime membership labels

# Output: numeric score per regime
score = algorithm_infer(x, y)
# score > 0: X->Y, score < 0: Y->X
```

(a) Data-driven algorithms receive standardised numeric arrays per regime, with no contextual information.

LLM Input (Segmented Anonymous)

```
Data segmented by Variable_Y at
threshold 0.5:

REGIME 0 (Low):  Variable_Y ≤ 0.5
Variable_X: Range=[0.05, 0.48]
Variable_Y: Range=[0.12, 0.50]
Sample (100 rows):  [...]

REGIME 1 (High):  Variable_Y > 0.5
Variable_X: Range=[0.51, 0.92]
Variable_Y: Range=[0.55, 0.98]
Sample (100 rows):  [...]
```

LLM Output (Step 4 JSON)

```
{"threshold_variable":  "Variable_Y",
 "threshold_value":  0.5,
 "directions":  [
{"regime":  0, "direction":  "Y->X"},
{"regime":  1, "direction":  "X->Y"}]}
```

(b) LLM input varies by configuration. Shown here: `segmented_anon_hint_explicit`, which provides data samples and an explicit threshold.

Figure 1.1.: Comparative workflow: (a) a data-driven algorithm receives standardised numeric arrays per regime; (b) the LLM receives data samples and threshold context via a 4-step prompting chain. Both produce per-regime causal direction outputs.

$Y = f(X) + g(X)N$ [ISV+23].

Algorithms like LOCI, ROCHE, and LCUBE build on these ideas: LOCI uses location–scale noise assumptions, ROCHE adds robustness with Student's-t likelihood [TDNN23], and LCUBE uses Minimum Description Length (MDL) scores over dense functional classes [HSM25]. These methods rely on explicit structural assumptions, including noise independence and restricted functional forms, and require sufficient sample sizes for stable estimation. In practice, real-world data with heavy-tailed noise, outliers, or non-stationary mechanisms can reduce their reliability [TDNN23].

This raises a question: *Can LLMs infer causal direction from numeric or tabular data, and how do they compare to LOCI, ROCHE, and LCUBE?* Figure 1.1 illustrates the contrast: data-driven algorithms receive only standardised numeric arrays, while the LLM receives data samples together with threshold context through a multi-step prompting chain. Both produce per-regime causal direction outputs, but the information each method works with differs fundamentally.

1.2. Problem Statement

Most evaluations of LLMs in causal reasoning focus on text-based tasks, such as extracting cause–effect relations from narratives or scientific documents [RRWT24]. Broader benchmarks also assess mathematical or coding abilities [WAN24], but these settings do not typically involve recovering causal direction from raw numerical data. Some studies include synthetic or numerical causal tasks [RRWT24, DLA+25], yet these experiments often provide structured descriptions or supervised signals. Hybrid approaches further integrate LLMs into classical pipelines. For example, by answering conditional independence queries within the PC algorithm [CVD+24]. In such cases, however, the underlying structure learning is still carried out by the classical algorithm. This differs from tabular causal discovery, where direction must be inferred directly from observational numerical samples.

A remaining question is whether LLMs can detect *shifts* in causal direction. In many real systems, the causal relationship flips depending on the regime: below 140 horsepower, fuel efficiency goals drive engine design, but above that threshold, engine power dictates fuel consumption. Similarly, the relationship between interest rates and inflation reverses at around 5.4% inflation, where monetary policy shifts from reacting to inflation to controlling it [GMP25]. We are not aware of a study that compares LLMs with data-driven causal discovery algorithms on numerical bivariate data where causal direction changes across regimes.

While benchmarks such as CausalBench [WAN24] and Cross-Questions [SS24] test LLM causal reasoning, they rely on text-based or synthetic settings. The benchmarks we reviewed do not evaluate numerical bivariate data where the ground truth itself changes across regimes. This setting is harder, because the correct answer depends on which subset of the data is considered.

This thesis fills that gap by: (1) testing LLMs on numeric/tabular bivariate datasets from real-world domains (monetary policy, energy systems, and automotive engineering), and (2) evaluating whether they can identify regime-dependent direction shifts, comparing their results to three data-driven causal discovery methods—ROCHE, LOCI, and LCUBE. While the Tübingen Cause–Effect Repository [MPJ+16] was initially considered, it does not provide regime definitions or ground-truth annotations for causal direction reversals, making it unsuitable for evaluating regime-dependent causal inference.

1.3. Scope and Assumptions

To keep the study focused, the following scope is set:

- **Bivariate data only** – just pairs of variables, not full graphs.
- **Observational data only** – no interventional or experimental data.
- **Literature-based thresholding** – each dataset is split into two regimes using thresholds from domain literature (e.g., 140 horsepower, 5.4% inflation rate, 3.5% short-term interest rate).

- **Algorithms** – comparison is between LLMs and three data-driven causal discovery methods: ROCHE, LOCI, and LCUBE.
- **LLM model** – GPT-5.2 is the primary model, accessed via the Responses API. The tool supports other providers (Claude, Gemini), but results from those models are reserved for future work.

Non-goals. This thesis evaluates causal direction inference only, not formal interventions or counterfactual reasoning [KN21]. We compare LLMs and classical methods without claiming one approach is superior. The evaluation covers bivariate direction inference only, not multivariate graphs or causal effect estimation.

1.4. Contributions

This thesis contributes four things:

1. **We introduce a regime-dependent causal evaluation** comprising real-world bivariate pairs from monetary policy, energy systems, and automotive engineering domains, where literature-based thresholds define regime splits with documented causal direction changes. We also propose a regime-level accuracy metric that measures inferred causal directions across changing regimes.
2. **We provide a systematic comparison** of data-driven causal discovery algorithms (ROCHE, LOCI, LCUBE) against LLMs (GPT-5.2) on numerical bivariate datasets exhibiting regime-dependent causal direction switches.
3. **We design an evaluation framework** to assess causal reasoning capabilities, using multiple experimental configurations (raw/segmented data, named/anonymous variables), and characterize LLM failure modes, including reliance on semantic cues when statistical patterns are ambiguous (Chapter 6).
4. **We provide an open-source Python tool** that implements the complete evaluation pipeline for regime-dependent numerical data.

1.5. Research Questions

Four main research questions guide this study:

RQ1: Can large language models infer causal direction from numeric and tabular data without relying on additional algorithms?

RQ2: How do the causal judgments made by LLMs compare with those produced by established data-driven methods (ROCHE, LOCI, LCUBE) across different datasets and regime-dependent contexts?

RQ3: How do naming and segmentation configurations affect LLM performance on regime-dependent causal inference?

1.6. Thesis Roadmap

Here is our approach: we first define regimes using literature-based thresholds, then apply data-driven algorithms to each regime, and finally evaluate LLMs under controlled experimental configurations. Six chapters follow:

- **Chapter 2 (Related Work)** reviews classical causal inference methods including additive noise models, location–scale noise models, and the ROCHE, LOCI, and LCUBE algorithms. It also discusses how LLMs have been studied in the context of causality, with emphasis on their predominant focus on text-based tasks.
- **Chapter 3 (Datasets and Evaluation Framework)** introduces the datasets used in this study, including synthetic validation data and real-world bivariate pairs from automotive engineering, energy systems, and monetary policy. It also defines the literature-based thresholds used for regime segmentation and the regime-level accuracy metric.
- **Chapter 4 (Data-Driven Causal Discovery)** presents the application of ROCHE, LOCI, and LCUBE to the regime-segmented data, including parameter choices and per-regime baseline results.
- **Chapter 5 (LLM Evaluation Framework)** describes the evaluation of LLMs, including the multi-step prompting architecture, experimental configurations, and GPT-5.2 model selection.
- **Chapter 6 (Experimental Results)** compares the outputs of LLMs with data-driven algorithms, analysing performance patterns, reasoning quality, failure modes, and the effect of different experimental configurations.
- **Chapter 7 (Discussion and Conclusion)** summarizes the contributions, discusses limitations, and points to future research directions.

2. Related Work

2.1. Causal Inference Methods

Classical methods for inferring causal direction from observational data span a wide range of statistical and algorithmic techniques [SGS00]. These methods fall into three broad categories. Constraint-based approaches test conditional independencies to orient edges. Score-based approaches assign a numerical score to candidate causal models and select the best fit. Functional causal models rely on asymmetries in the data-generating process to determine direction from bivariate observations alone [MPJ$^+$16]. The algorithms used in this thesis (LOCI, ROCHE, and LCUBE) belong to the third category.

Granger causality tests whether past values of one time series improve the prediction of another. It captures temporal precedence but does not, on its own, establish structural causation [KN21]. For time-series data with both lagged and contemporaneous dependencies, the PCMCI framework extends conditional independence testing to handle these more complex temporal structures [RNK$^+$19]. To address nonlinearity and heteroscedasticity, *additive noise models* (ANMs) assume that the effect can be written as $Y = f(X) + N$ where the noise term N is independent of the cause; identifiability results show that, under suitable conditions such as nonlinearity and independent noise, the causal direction can be uniquely determined [ISV$^+$23]. *Location–scale noise models* (LSNMs) generalise ANMs by allowing both the mean and variance of the noise to depend on the cause ($Y = f(X) + g(X)N$). Under appropriate regularity conditions, location–scale noise models remain identifiable and allow causal direction to be recovered in the presence of heteroscedasticity [ISV$^+$23].

2.1.1. Pearl's Causal Inference Hierarchy

Pearl's ladder of causation [KN21] provides a useful hierarchy for framing causal reasoning, distinguishing between three levels:

1. **Association** (seeing): This level concerns statistical dependencies in observed data and is where traditional statistical models and many machine learning systems operate. Questions at this level ask "How would seeing X change my belief in Y?"

2. **Intervention** (doing): This level asks what would happen under hypothetical manipulations and is the target of many causal discovery and causal inference methods. Questions take the form "What would happen to Y if I set X to a particular value?"

3. **Counterfactual** (imagining): This level addresses alternative histories, asking what would have happened had a variable taken a different value. Questions ask "Would Y have occurred if X had not happened, given that X did happen?"

In the bivariate observational setting studied in this thesis, classical causal discovery methods use structural assumptions that are consistent with intervention-level semantics, such as in ANMs and LSNMs where mechanisms of the form $Y = f(X) + N$ or $Y = f(X) + g(X)N$ assume independence between the cause X and the noise term N [ISV+23]. In contrast, LLMs are trained on association-level objectives yet can articulate intervention-like statements; whether they can reliably infer causal direction from numerical structure, rather than generating linguistically plausible causal explanations, therefore remains an empirical question that this thesis investigates.

2.1.2. Data-Driven Causal Discovery Algorithms

Building on these models, the **LOCI** algorithm estimates the causal direction by fitting neural networks or feature maps to each direction and comparing the fit. It is valid under the assumptions of the location–scale noise model, but is sensitive to heavy-tailed noise and may misclassify when sample sizes are small or distributions are multimodal [ISV+23]. To improve robustness, the **ROCHE** method replaces the usual Gaussian likelihood with a Student's-t likelihood, providing better resistance to outliers and small-sample effects [TDNN23]. Both LOCI and ROCHE are sensitive to distributional assumptions and assume the absence of latent confounding. They use structural assumptions to determine causal direction in settings where other methods (e.g., the PC algorithm) struggle with heteroscedastic noise.

The **LCUBE** algorithm approaches bivariate causal direction identification using Minimum Description Length (MDL) scores over dense functional classes. Unlike LOCI and ROCHE, which rely on likelihood-based model comparison, LCUBE selects the causal direction that yields the most compressive description of the data, enabling flexible approximation of nonlinear relationships without fixed parametric forms [HSM25]. Specifically, LCUBE evaluates both candidate directions by computing the MDL score for each, choosing the direction whose model compresses the data most efficiently. The use of dense functional classes, such as cubic spline approximations with adaptively chosen knots, allows LCUBE to capture complex nonlinear relationships without imposing parametric restrictions. This MDL-based approach provides formal identifiability guarantees and has been evaluated on established benchmark datasets [HSM25].

Finally, there are methods based on conditional independence tests, such as the **PC algorithm** [SG91], which recover entire causal graphs by testing conditional independencies. These approaches require large sample sizes and can be sensitive to the choice of significance thresholds [SG91]. While powerful for multivariate problems, they are more complex than needed for bivariate analyses, and their conclusions depend on assumptions such as faithfulness and causal sufficiency [CVD+24].

2.2. LLMs and Causality

Large Language Models (LLMs) have been explored for causal inference. Surveys highlight that they can extract cause–effect relations from text and answer causal questions, but evaluations are still mostly text-based [RRWT24]. Large-scale evaluation studies report uneven performance across reasoning tasks, with strong results on familiar formats and weaker performance under perturbations [MAS+25]. Several benchmarks report failure cases under perturbation. For example, functional reasoning benchmarks reveal that while models can handle memorised or static examples, with substantial drops when numerical values are varied [SBV+24]. Analyses of GPT-4 and ChatGPT show that they handle undergraduate-level mathematics but fail on graduate-level problems, suggesting reliance on pattern retrieval rather than formal reasoning [FPC+23]. The GSM-Symbolic benchmark further shows that small changes in wording or numbers can cause large drops in accuracy, indicating pattern replication rather than logical reasoning [MAS+25].

These findings raise the question of whether LLMs operate at the association level of Pearl's hierarchy (detecting statistical patterns) or can genuinely approximate intervention-level reasoning. Evidence suggests the former: LLMs are trained on association-level objectives and optimise next-token prediction objectives and may not explicitly model invariant causal mechanisms.

In causal discovery, Cohrs et al. proposed *chatPC*, where LLMs act as conditional-independence oracles for the PC algorithm [CVD+24]. Independence tests are rephrased as prompts, and the yes/no answers are used to build a graph. The approach can recover small graphs, but results depend on prompt formulation. A voting scheme reduces errors but variability remains. The *CausalBench* benchmark explores causal queries across text, mathematics, and coding and includes synthetic evaluation settings [WAN24]. The *Cross-Questions* paper shows that carefully designed follow-up questions improve causal event attribution [SS24]. Finally, *LLM-initialised Differentiable Causal Discovery* demonstrates that using LLM outputs to initialise optimisation can help recover causal graphs [KHvdP+24]. Taken together, these works show that LLMs can support causal tasks but are variable in performance across settings when reasoning beyond their training distribution.

More recent work has addressed LLM limitations through structured reasoning approaches. The *Causal Chain of Prompting* (C2P) framework decomposes causal reasoning into subtasks inspired by constraint-based discovery, including variable extraction, independence identification, collider detection, and hypothesis evaluation [BAD+25]. C2P reports improved performance over unstructured prompting on symbolic and narrative evaluation tasks, suggesting that decomposed reasoning outperforms single-step causal queries. Healey and Kohane show that LLMs may violate basic temporal causal principles (e.g., cause-before-effect) in multimodal time series, even with structured inputs [HK24]. Meanwhile, Dong et al. demonstrate in their CARE framework that LLMs show sensitivity to variable naming and limited use of observational structure, motivating fine-tuning approaches for multivariate graph recovery [DLA+25]. These limitations motivate the focus in this thesis on bivariate numerical evaluation, where regime-dependent mechanisms

introduce additional structural complexity beyond static causal discovery settings.

2.2.1. Numerical Data Representation Challenges

A central challenge in applying LLMs to causal discovery lies in representing numerical observational data. Unlike symbolic or narrative tasks where causal relationships can be expressed verbally, bivariate numerical datasets require encoding distributional patterns, statistical dependencies, and regime-specific mechanisms that may not have natural language equivalents.

As discussed in Section 2.2, LLMs are sensitive to small numerical perturbations. For causal discovery specifically, independence query phrasing strongly influences results when LLMs serve as oracles [CVD+24]. These findings suggest that translating numerical causal discovery into language-based reasoning introduces sensitivity that differs from dedicated statistical methods.

Regime-dependent causality presents additional representational challenges. There is limited guidance on how to encode threshold-based regime switches, non-stationary mechanisms, or regime-dependent causal reversals in natural language. While structured frameworks like C2P report improved performance, there is limited agreement on how best to represent numerical tabular data with regime information. This motivates the standardised prompting framework used in this thesis.

2.3. Regime-Dependent Causality

In some real-world systems, causal relationships are not uniform across all observations. *Regime-dependent* causality refers to settings in which the causal direction between two variables changes across subsets of the data, where a *regime* corresponds to a subset of observations defined by a contextual condition or threshold such that the causal relationship differs across regimes

This setting differs from static bivariate causal discovery, since mixing regimes can violate the single-mechanism assumptions that underpin identifiability, including the existence of a fixed functional relationship and a stable noise structure [ISV+23]. As a result, methods that fit a single global model may average over conflicting signals and produce ambiguous or unstable causal estimates.

Examples of regime-dependent causal relationships appear across the domains studied in this thesis:

- **Monetary policy**: In low-inflation environments, central banks control inflation through interest rate adjustments (interest rates → inflation), but during high-inflation crises exceeding documented thresholds, this relationship may invert as policy becomes reactive (inflation → interest rates) [GMP25].

- **Energy systems**: Below a certain share of renewable electricity generation, rising per-capita emissions drive investment in renewable capacity (emissions →

renewables). Above that threshold, the growing renewable share begins to displace fossil generation and reduce emissions (renewables → emissions).

- **Automotive engineering**: The causal relationship between engine power and fuel efficiency shifts across market segments. Below roughly 140 horsepower, consumer demand for fuel economy shapes engine design choices (fuel efficiency → horsepower), whereas in performance segments above this threshold, horsepower requirements dictate fuel consumption (horsepower → fuel efficiency) [GHH+18].

Many causal discovery methods assume a single global causal mechanism, treating the entire dataset as one regime. From a theoretical standpoint, regime-dependent causality departs from the single-mechanism assumption that underpins most identifiability results. Standard ANM and LSNM proofs assume a fixed functional relationship $Y = f(X) + g(X)N$ across all observations [ISV+23]. When the causal direction reverses across regimes, a single global model receives conflicting signals, potentially producing aggregated or unstable causal scores. This motivates the regime-segmentation approach adopted in this thesis: by splitting the data at literature-based thresholds before applying causal inference, each subset can be analysed under the standard identifiable assumptions. Recent theoretical work has begun to formalise causal discovery in systems with switching causal relations, providing identifiability conditions for settings where the mechanism changes across regimes [WTB+25].

This thesis evaluates this gap by assessing whether data-driven methods and LLMs can recover regime-specific causal directions when regimes are defined using thresholds grounded in domain literature.

2.4. Prompt Engineering for Causal Tasks

Prompt design influences LLM performance. As discussed in Section 2.2, small wording changes or added irrelevant details can reduce accuracy on reasoning tasks. In causal discovery, Cohrs et al. note that results can change depending on how independence queries are phrased, and recommend averaging across multiple prompts to stabilise outcomes [CVD+24]. General prompt optimisation methods have been proposed, though they vary in computational cost and interpretability. Recent work proposes more targeted prompting strategies. For example, *Cross-Questions* use structured follow-up prompts to check assumptions such as temporal ordering or confounding and report improved causal attribution [SS24]. Similarly, *LLM-initialised Differentiable Causal Discovery* uses LLM outputs as priors to guide optimisation in causal graph recovery [KHvdP+24]. These studies suggest that structured prompting can improve reliability, but there is limited agreement on how to encode numeric tabular data or regime changes. This thesis uses standardised prompt templates to ensure consistent presentation and to minimise variation from prompt wording.

Building on insights from C2P and related structured approaches, this thesis employs a multi-step prompting procedure to query LLMs for causal direction in regime-dependent

settings. Instead of requesting causal direction in a single prompt, the approach decomposes the task into sequential steps: providing data context, requesting statistical analysis, asking for causal reasoning, and finally extracting a directional judgment. The details of this four-step architecture are presented in Chapter 5.

2.5. Chapter Summary

This chapter has reviewed the foundational concepts and prior work relevant to this thesis. Classical causal inference methods, including Granger causality, additive noise models, and location–scale noise models, provide structured approaches to inferring causal direction from observational data, with algorithms like LOCI, ROCHE, and LCUBE offering competitive performance under their respective assumptions. Pearl's causal hierarchy distinguishes between association, intervention, and counterfactual reasoning, highlighting that most machine learning systems, including LLMs, are trained using association-level objectives.

Recent work on LLMs and causality has demonstrated both strengths and limitations. While LLMs can articulate causal relationships and support causal discovery pipelines as oracles, their performance is sensitive to prompt phrasing and numerical perturbations, suggesting sensitivity to surface-level variations in input formulation. Structured prompting frameworks like C2P improve performance on symbolic tasks, but challenges remain in representing numerical data and encoding regime information.

Existing benchmarks for causal discovery, including the Tübingen pairs and reasoning-focused datasets like GSM-Symbolic, do not explicitly evaluate regime-dependent causal direction changes or the ability of methods to handle non-stationary mechanisms. While recent theoretical work has begun to address regime-switching causal models [WTB+25], empirical evaluation in real-world numerical settings remains limited. This thesis addresses three specific gaps in the literature:

1. **No systematic comparison** of LLMs versus data-driven methods on numerical bivariate data with regime-dependent causal direction switches.
2. **No evaluation framework** for testing whether methods can identify causal reversals across regimes defined by domain-specific thresholds.
3. **Limited understanding** of LLM failure modes in causal inference, particularly regarding their reliance on semantic knowledge versus statistical patterns.

The following chapters present the datasets, regime segmentation methodology, algorithm implementations, and experimental results that address these gaps.

3. Datasets and Evaluation Framework

This chapter describes the datasets used to evaluate causal discovery using data-driven algorithms and LLMs, and defines the evaluation framework for comparing methods. We focus on bivariate pairs exhibiting regime-dependent causal direction switches, where domain literature provides threshold values and ground-truth causal directions for each regime. The datasets cover automotive engineering, energy systems, monetary policy, and one synthetic validation dataset.

3.1. Synthetic Baseline Dataset: Stress-Fatigue

To test whether our method detects regime-dependent causality, we generated a synthetic dataset with known regime-switching behavior. Real-world data often does not provide a clear ground truth for the exact threshold where the causal direction changes. This synthetic dataset allows us to test whether the methods detect a change in causal direction when it is present.

The synthetic dataset contains $n = 1000$ samples of two variables: FatigueLevel (F) and StressLevel (S), with threshold $\tau = 0.5$ on S. The choice of variables is motivated by evidence of reciprocal associations between stress and fatigue in daily life [DDS+15]. We follow the meta-causal framework of [WTB+25] to model regime-dependent mechanisms. The generation procedure creates two regimes with opposite causal mechanisms:

- **Regime 1 (Low):** $S < 0.5$. Fatigue causes stress with negative relationship: $S = 0.5 - 0.3 \cdot F + \epsilon_1$, where $\epsilon_1 \sim \mathcal{N}(0, 0.05^2)$ and $F \sim \text{Uniform}(0.0, 0.5)$. Generated stress values are clipped to $[0, 0.49]$ to ensure regime separation. Causal direction: $F \to S$.
- **Regime 2 (High):** $S \geq 0.5$. Stress causes fatigue with positive relationship: $F = 0.2 + 0.6 \cdot S + \epsilon_2$, where $\epsilon_2 \sim \mathcal{N}(0, 0.05^2)$ and $S \sim \text{Uniform}(0.5, 1.0)$. Causal direction: $S \to F$.

We concatenate and shuffle data from both regimes to create a single dataset with regime-dependent causality. This setting allows us to test whether causal discovery methods and LLMs identify the direction change across regimes. The dataset provides a known ground truth benchmark for regime-dependent analysis.

3.2. Regime-Level Evaluation Framework

Many causal discovery methods fit a single global model to the entire dataset, which can miss important regime-dependent relationships where causal direction reverses under

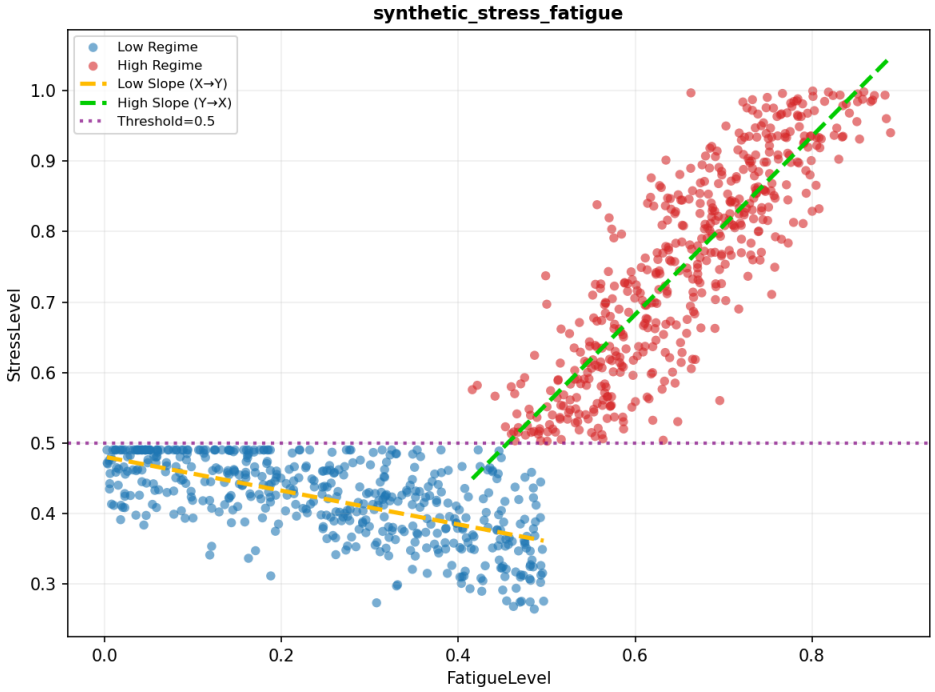

Figure 3.1.: Synthetic stress-fatigue dataset. Blue points represent the low regime ($S < 0.5$) where $F \to S$, and red points represent the high regime ($S \geq 0.5$) where $S \to F$. The purple dashed line marks the threshold at $S = 0.5$.

different conditions. Real-world causal relationships often exhibit regime-switching behavior driven by factors such as economic development stages, policy environments, or infrastructure maturity [WTB+25]. Global causal models that ignore this heterogeneity may produce misleading results by averaging across distinct causal regimes.

Our approach addresses these limitations by segmenting data according to well-established theoretical thresholds from domain literature. Unlike data-driven clustering methods that introduce additional algorithmic complexity, literature-based thresholds provide theoretically grounded regime boundaries with clear interpretability.

3.2.1. Regime-Level Accuracy Metric

To enable fair comparison between data-driven causal discovery algorithms and LLM-based causal reasoning, we define a regime-level accuracy metric that evaluates inferred causal directions against literature-based ground truth for each regime independently.

Given a pair of variables (X, Y) and a partition of observations into regimes $\{R_1, R_2, \ldots, R_k\}$, the objective is to assess whether a method can correctly infer the causal direction associated with each regime. The output space is binary, consisting of either $X \to Y$ or $Y \to X$. Ground truth causal directions are specified at the regime level based on domain literature and are used exclusively for evaluation.

For a dataset with k regimes and a causal discovery method or LLM reasoning method m, we define the regime-level accuracy as:

$$\mathrm{Acc}^{(m)}(X, Y) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}\left[\widehat{d}_i^{(m)} = d_i\right], \tag{3.1}$$

where d_i is the literature-based ground truth direction in regime R_i and $\widehat{d}_i^{(m)}$ is the inferred direction by method m. The indicator function $\mathbb{I}[\cdot]$ returns 1 if the condition is true and 0 otherwise.

In our experiments, we restrict to $k = 2$ regimes (low/high), reflecting the most common regime distinctions reported in domain literature. For two regimes, the metric simplifies to:

$$\mathrm{Acc}^{(m)}(X, Y) = \frac{1}{2}\Big(\mathbb{I}[\widehat{d}_{\text{low}}^{(m)} = d_{\text{low}}] + \mathbb{I}[\widehat{d}_{\text{high}}^{(m)} = d_{\text{high}}]\Big). \tag{3.2}$$

This metric captures whether each method correctly identifies the causal direction in both the low and high regimes. A score of 100% indicates perfect accuracy (both regimes correct), 50% indicates one regime correct and one incorrect, and 0% indicates both regimes incorrect. All methods—ROCHE, LOCI, LCUBE, and LLM prompting—are evaluated using this metric with identical regime definitions, ensuring that performance differences reflect method-specific inference behavior rather than informational or configuration advantages.

3.3. Real-World Datasets

We evaluate our methods on four real-world bivariate datasets with regime-dependent causality described in domain literature. For each dataset, domain research reports threshold values or transition ranges where the causal mechanism changes.

3.3.1. Auto MPG and Horsepower

This dataset contains city-cycle fuel consumption (MPG) and vehicle horsepower from the UCI Auto MPG repository [Qui93], originally compiled for the 1983 American Statistical Association Exposition. The relationship between fuel efficiency and engine power reflects trade-offs between performance, cost, and regulation in vehicle design [GHH+18]. These competing objectives motivate the consideration of heterogeneous design regimes.

Figure 3.2 shows two segments separated by a transition band between approximately 120 and 160 HP. In lower-horsepower vehicles ($H < 120$ HP), fuel economy constraints and cost sensitivity shape design decisions, and fuel efficiency targets constrain engine configuration choices (MPG $\rightarrow$ H). In higher-horsepower vehicles ($H > 160$ HP), performance goals drive design decisions, and engine power determines fuel consumption ($H \rightarrow$ MPG). The transition region reflects mixed design priorities.

For evaluation, we introduce a threshold of 140 HP to define low and high regimes. This value corresponds to the midpoint of the identified transition band.

3.3.2. Renewable Energy Share and CO_2 Emissions

This dataset contains country-level annual CO_2 emissions per capita and renewable electricity generation share, sourced from World Bank and OECD databases. The relationship between renewable energy deployment and macro-level outcomes is known to exhibit heterogeneous and context-dependent causal patterns [AO16]. We therefore model the relationship as depending on the level of renewable deployment.

In **low renewable penetration** ($R < 15\%$), renewable energy remains a small share of electricity generation. In this setting, high emissions and fossil fuel dependence may drive policy responses and investment in renewable capacity ($E \rightarrow R$). In **high renewable penetration** ($R \geq 15\%$), renewable electricity forms a meaningful share of the energy mix and can displace fossil fuel generation. In this regime, renewable share influences emissions levels ($R \rightarrow E$).

This regime-dependent reversal reflects the dual role of renewable energy as both a response to emissions pressure and a mechanism for emissions reduction.

For evaluation, we introduce a threshold of 15% renewable share to separate low and high deployment regimes. Figure 3.3 illustrates this regime-dependent relationship.

3.3.3. UK Short-Term and Long-Term Interest Rates

This dataset contains monthly interest rates from FRED (January 1978–February 2025, $n = 566$ months), capturing short-term and long-term UK government bond yields. The

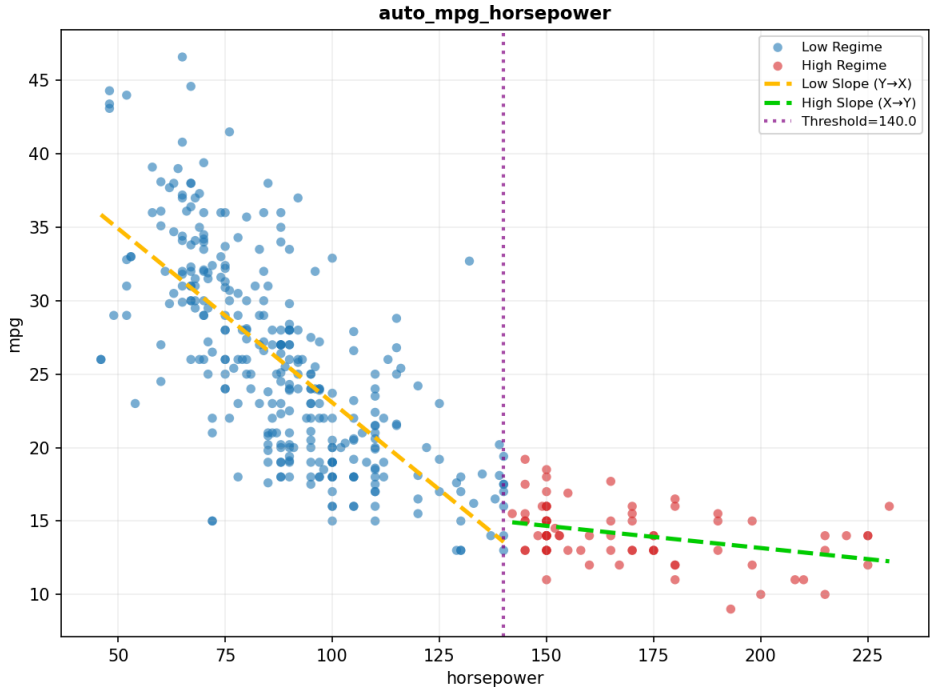

Figure 3.2.: Auto MPG and horsepower relationship showing regime-dependent causal structure. The threshold at 140 HP (midpoint of 120–160 HP transition region) separates consumer-driven design (MPG $\rightarrow H$) from performance-driven engineering ($H \rightarrow$ MPG).

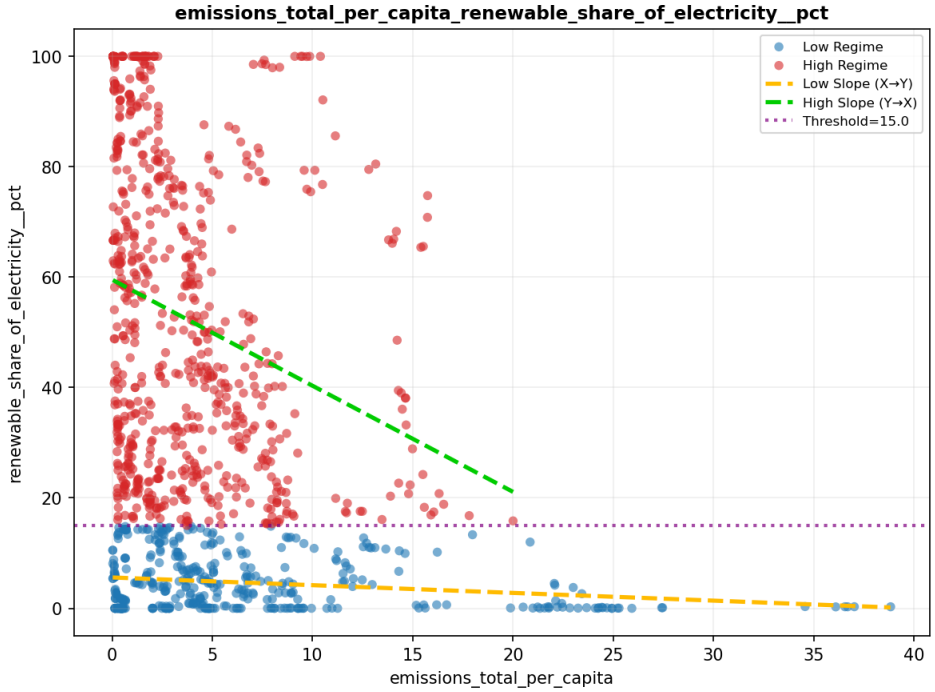

Figure 3.3.: CO_2 emissions per capita vs renewable electricity share showing regime-dependent causal structure. Below 15% renewable share, emissions drive renewable investment; above 15%, renewables actively reduce emissions.

term structure reflects interactions between monetary policy and market expectations [EM96]. We model the relationship as depending on the level of short-term interest rates.

Prior work discusses yield curve slope changes in the range of 3–4% short-term rates [EM96]. In periods with lower short-term rates ($S < 3.5\%$), policy rate movements tend to precede changes in long-term yields, consistent with a policy-to-market transmission channel ($S \to L$). In periods with higher short-term rates ($S \geq 3.5\%$), the yield curve flattens or inverts, and long-term yields tend to lead short-term rate adjustments, suggesting a reversed direction ($L \to S$).

For evaluation, we introduce a threshold of 3.5% (midpoint of the 3–4% transition range) to separate low and high short-rate regimes. Figure 3.4 shows this regime-dependent interest rate relationship.

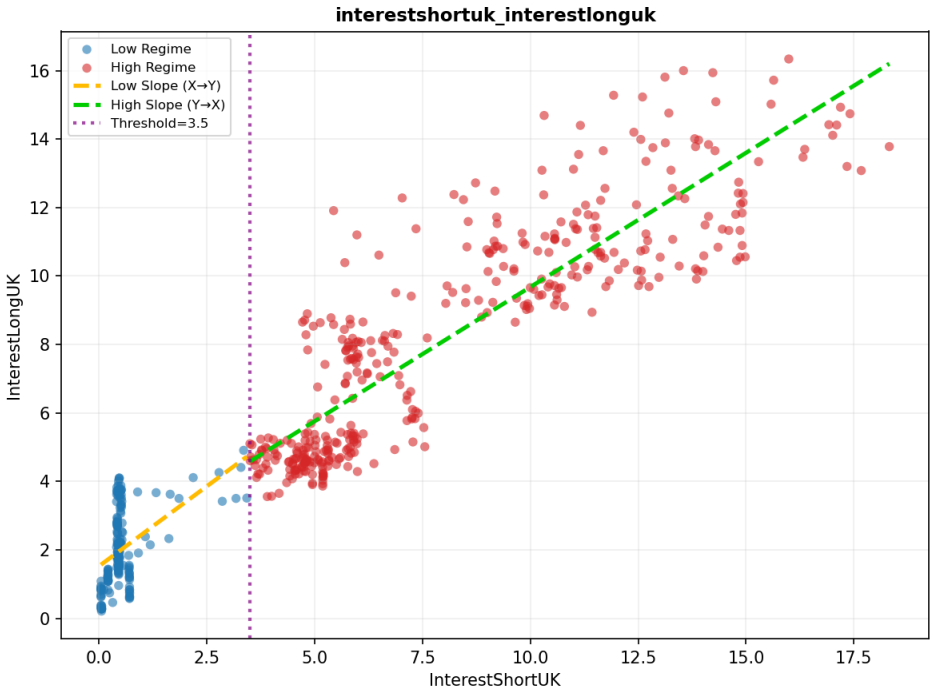

Figure 3.4.: UK short-term vs long-term interest rates showing regime-dependent causal structure. The threshold at 3.5% (midpoint of 3–4% range) separates policy-driven expectations ($S \to L$) from market-driven dynamics ($L \to S$).

3.3.4. US Inflation Rate and Federal Funds Rate

This dataset combines monthly Federal Reserve interest rate data with inflation measurements, sourced from the Federal Reserve Economic Data (FRED) portal. The interest rate variable is the Effective Federal Funds Rate, and inflation is measured using changes in the Consumer Price Index.

Prior work documents regime-switching behavior in the monetary policy transmission mechanism, with a transition zone around 5.3–5.5% inflation [GMP25]. In **low-inflation regimes** ($I \leq 5.4\%$), price movements tend to precede policy rate adjustments, consistent with a reactive monetary policy stance ($I \to F$). In **high-inflation regimes** ($I > 5.4\%$), policy rate changes tend to precede shifts in inflation, consistent with an active disinflation channel ($F \to I$).

For evaluation, we set a threshold of 5.4% (midpoint of the documented 5.3–5.5% transition range) to separate low and high inflation regimes [KNP04]. Figure 3.5 illustrates this regime-dependent monetary policy relationship.

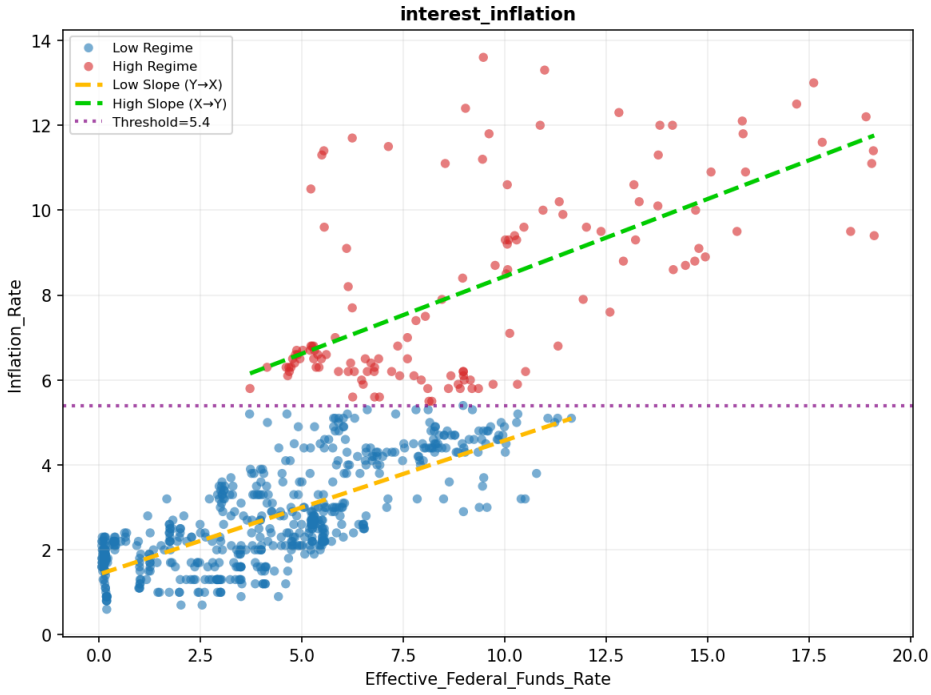

Figure 3.5.: US inflation rate vs federal funds rate showing regime-dependent causal structure. The threshold at 5.4% inflation (midpoint of 5.3–5.5% range) separates inflation-driven policy ($I \to F$) from policy-driven inflation control ($F \to I$).

3.4. Dataset Summary

Table 3.1 summarizes all datasets used in this thesis, including their regime thresholds and expected causal directions based on domain literature.

Table 3.1.: Summary of datasets with regime thresholds and ground-truth causal directions.

Dataset	Variables	Threshold	Low Regime	High Regime
Synthetic Stress-Fatigue	FatigueLevel, StressLevel	$S = 0.5$	$F \to S$	$S \to F$
Auto MPG-Horsepower	MPG, Horsepower	$H = 140$ HP	MPG $\to H$	$H \to$ MPG
Emissions-Renewables	CO_2/capita, Renewable %	$R = 15\%$	$E \to R$	$R \to E$
UK Interest Rates	Short Rate, Long Rate	$S = 3.5\%$	$S \to L$	$L \to S$
US Inflation-Fed Rate	Inflation, Fed Rate	$I = 5.4\%$	$I \to F$	$F \to I$

3.5. Data Preprocessing

All datasets underwent consistent preprocessing to ensure fair comparison across methods:

1. **Regime Definition**: Threshold values are fixed from domain literature prior to any analysis. Each observation is assigned to the low regime ($\leq \tau$) or high regime ($> \tau$) based on the threshold variable.

2. **Data Filtering**: Missing or incomplete observations are removed. No imputation is performed to preserve data integrity.

3. **Standardization**: Data is normalized using `StandardScaler()` before algorithm application, ensuring numerical stability across different variable scales.

4. **Data Sampling (LLM only)**: For large datasets, a maximum of 200 observations are randomly sampled to manage token usage for LLM evaluation while retaining sufficient distributional information. This sampling step applies only to the LLM evaluation protocol; baseline algorithms receive the full regime subsets.

3.6. Chapter Summary

This chapter has presented the five datasets used in this thesis and the evaluation framework for comparing causal discovery methods. The datasets include one synthetic validation dataset (Stress-Fatigue) and four real-world bivariate pairs from automotive engineering (Auto MPG-Horsepower), energy systems (Emissions-Renewables), and monetary policy (UK Interest Rates, US Inflation-Fed Rate). Each dataset exhibits regime-dependent causal direction switches with thresholds documented in domain literature. The regime-level accuracy metric (Equation 3.1) provides a unified framework for comparing data-driven algorithms (ROCHE, LOCI, LCUBE) and LLM-based reasoning on identical regime definitions.

4. Data-Driven Causal Discovery

This chapter tests three causal discovery algorithms (ROCHE, LOCI, and LCUBE) on the datasets from Chapter 3. These algorithms serve as baselines for comparing against Large Language Models in Chapter 5. All three methods infer causal direction from numerical data alone, without using variable names or domain knowledge.

4.1. ROCHE Method

ROCHE (Robust Causal Heteroscedastic Estimation) fits models in both directions and picks the better fit [TDNN23]. We use the authors' open-source implementation from GitHub,[1] specifically the `roche()` function.

4.1.1. Algorithm Foundation

ROCHE fits a location-scale noise model in both directions ($X \to Y$ and $Y \to X$) and compares how well each fits the data. The algorithm uses Student's-t distributions instead of Gaussian distributions. This choice makes ROCHE less sensitive to outliers and heavy-tailed noise.

The output is a score: positive values indicate $X \to Y$, negative values indicate $Y \to X$. Larger absolute values show stronger evidence for the detected direction.

4.1.2. Implementation and Parameters

We use hyperparameters from the original publication [TDNN23]:

- **Training epochs**: 5000
- **Learning rate**: 10^{-2} with cosine annealing to 10^{-6}
- **Random seed**: 711 (fixed for reproducibility)

Data is normalized using StandardScaler before running the algorithm. Each regime is processed separately. The algorithm receives no information about other regimes or ground truth labels.

[1] `https://github.com/quangdzuytran/ROCHE`

4.2. LOCI Method

LOCI (Location-scale Causal Inference) also fits location-scale noise models but uses Gaussian distributions instead of Student's-t [ISV+23]. We use the authors' open-source implementation from GitHub,[2] specifically the `loci()` function.

4.2.1. Algorithm Foundation

LOCI estimates both candidate causal directions and picks the one that better matches a location-scale noise assumption. It uses neural networks to model the relationship between variables. LOCI works well when its Gaussian noise assumptions hold, but can struggle with heavy-tailed distributions or small sample sizes [ISV+23].

4.2.2. Implementation and Parameters

We use hyperparameters from the original publication [ISV+23]:

- **Training epochs**: 5000
- **Learning rate**: 10^{-2} with cosine annealing to 10^{-6}
- **Random seed**: 711 (fixed for reproducibility)

As with ROCHE, each regime is processed separately without access to ground truth labels.

4.3. LCUBE Method

LCUBE uses Minimum Description Length (MDL) to identify causal direction [HSM25]. Unlike LOCI and ROCHE, which compare how well different noise models fit the data, LCUBE picks the direction that gives the shortest description of the data. We use the authors' open-source implementation from GitHub,[3] specifically the `infer_causal_direction()` function.

4.3.1. Algorithm Foundation

The MDL principle says that the best model is the one that compresses the data most. LCUBE applies this idea to causal discovery: it computes description lengths for both directions ($X \to Y$ and $Y \to X$) and picks the shorter one.

LCUBE does not assume a specific noise distribution. This makes it more flexible than LOCI (which assumes Gaussian noise) or ROCHE (which assumes Student's-t noise). The method uses cubic splines to approximate the causal function.

[2] `https://github.com/AlexImmer/loci`
[3] `https://github.com/suzi216/LCUBE`

4.3.2. Implementation and Parameters

LCUBE uses adaptive spline fitting. The number of knots depends on the data size:

$$\text{number of knots} = \max\left(1, \left\lfloor \frac{n}{40} \right\rfloor\right) \tag{4.1}$$

where n is the number of observations.

Unlike ROCHE and LOCI, LCUBE is fully deterministic and as such, it produces the same output every time given the same input. No random seed is needed.

4.3.3. Comparison with LOCI and ROCHE

Table 4.1 summarizes the three algorithms.

Table 4.1.: Comparison of data-driven causal discovery algorithms.

Feature	**LOCI**	**ROCHE**	**LCUBE**
Noise Model	Gaussian	Student's-t	None (MDL)
Optimization	Neural network	Neural network	Spline fitting
Outlier handling	Sensitive	Robust	Flexible
Deterministic	No	No	Yes

All three algorithms share a common evaluation pattern: they run independently on each regime subset (low/high) without access to ground truth labels or information from other regimes. Regime segmentation is done externally and identically for all methods.

4.4. Evaluation Workflow

All three methods share the same evaluation pipeline, described in Algorithm 1. The pipeline applies a literature-based threshold to split each dataset into two regimes, then runs the chosen algorithm independently on each regime.

Data: Dataset D, cause column X, effect column Y, method $M \in \{\text{ROCHE}, \text{LOCI}, \text{LCUBE}\}$, threshold variable v, threshold value τ
Result: List of (regime, direction, score) for low and high regimes
`// Step 1: Split into regimes using literature-based threshold`
$L \leftarrow \text{segment\_data}(D,\ v,\ \tau)$ `//` $L_i = 0$ `if` $D_i[v] \leq \tau$`, else` $L_i = 1$
`// Step 2: Standardize the full dataset before splitting`
$(D[X],\ D[Y]) \leftarrow \text{StandardScale}(D[X],\ D[Y])$;
`// Step 3: Run method independently on each regime`
$\mathcal{R} \leftarrow []$;
foreach $r \in \{0, 1\}$ **do**
  $(x_r,\ y_r) \leftarrow D[L{=}r,\ X],\ D[L{=}r,\ Y]$;
  if $M = LCUBE$ **then**
    (dir_str, score) $\leftarrow$ `infer_causal_direction`$(x_r,\ y_r)$;
    direction $\leftarrow$ "X causes Y" if dir_str = `"->"`, else "Y causes X";
  else
    `// ROCHE or LOCI: positive score means` X `causes` Y`; negative means` Y `causes` X
    score $\leftarrow$ `roche`$(x_r,\ y_r)$ or `loci`$(x_r,\ y_r)$;
    direction $\leftarrow$ "X causes Y" if score > 0, else "Y causes X";
  end
  append $(r,$ direction, score$)$ to $\mathcal{R}$;
end
return $\mathcal{R}$;

Algorithm 1: Regime-based causal direction inference

4.5. Experimental Setup

We tested all three algorithms on the five datasets from Chapter 3. Each dataset is split into two regimes (low/high) using the literature-based thresholds.

4.5.1. Data Preparation

The datasets used in this chapter are stored as pre-cleaned text files. Missing or incomplete observations were removed during dataset construction; no imputation is performed at any stage. The evaluation pipeline standardizes the two analysis variables using StandardScaler before running any algorithm. Each algorithm receives the full regime subset after standardization.

4.5.2. Regime Application

We used the threshold values from Table 3.1 (Chapter 3) to split each dataset into low and high regimes. Each algorithm runs separately on each regime subset.

4.5.3. Testing Protocol

For ROCHE and LOCI, we used fixed random seeds to ensure reproducibility. LCUBE is deterministic and produces the same output every run. Each algorithm processes one regime at a time with no information sharing between regimes.

4.5.4. Performance Metrics

We use the regime-level accuracy metric from Section 3.2:

$$\mathrm{Acc}^{(m)}(X, Y) = \frac{1}{2}\Big(\mathbb{I}[\widehat{d}_{\mathrm{low}}^{(m)} = d_{\mathrm{low}}] + \mathbb{I}[\widehat{d}_{\mathrm{high}}^{(m)} = d_{\mathrm{high}}]\Big) \tag{4.2}$$

where d_i is the ground truth direction and $\widehat{d}_i^{(m)}$ is the inferred direction. A score of 100% means both regimes were correctly identified.

4.6. Results and Analysis

Table 4.2 shows the accuracy of each algorithm on each dataset. ROCHE and LCUBE achieve 80% mean accuracy, while LOCI achieves 50%.

Table 4.2.: Regime-level accuracy by algorithm and dataset.

Dataset	**ROCHE**	**LOCI**	**LCUBE**
Stress-Fatigue (synthetic)	100%	100%	100%
MPG-Horsepower	100%	50%	100%
Emissions-Renewables	100%	0%	100%
UK Interest Rates	50%	50%	50%
US Inflation-Fed Rate	50%	50%	50%
Mean	**80%**	**50%**	**80%**

4.6.1. Synthetic Dataset Results

All three algorithms achieve 100% accuracy on the synthetic Stress-Fatigue dataset. This confirms that the regime-dependent causal reversal is detectable when the mechanism is known by construction.

4.6.2. Real-World Dataset Results

Engineering and Energy Datasets. ROCHE and LCUBE achieve 100% accuracy on both MPG-Horsepower and Emissions-Renewables. LOCI achieves 50% on MPG-Horsepower (correct in one regime) but 0% on Emissions-Renewables (inverted—it detected a reversal but got both directions backwards).

Economic Datasets. All three algorithms achieve 50% accuracy on both UK Interest Rates and US Inflation-Fed Rate. Each algorithm correctly identifies one regime but not

the other. Unlike the engineering and energy pairs, where at least two algorithms reach 100%, economic datasets challenge all three methods equally.

4.6.3. Method Comparison

ROCHE and LCUBE show the same pattern: they perform well on synthetic, engineering, and energy datasets (100%) but struggle with economic datasets (50%). LOCI performs worse overall, with particular difficulty on the Emissions-Renewables pair.

The economic datasets (UK Interest Rates, US Inflation-Fed Rate) prove hardest for all methods. No algorithm achieves more than 50% accuracy on these pairs. As shown in Section 4.6.4, all three algorithms infer the same causal direction for both regimes on these pairs, meaning none detects the regime reversal.

4.6.4. Regime-Switching Detection

Table 4.3 shows whether each algorithm correctly detected that the causal direction reverses between regimes. All five datasets have true causal reversals.

Table 4.3.: Regime-switching detection. "Correct" = detected reversal with correct directions. "Inverted" = detected reversal but directions backwards. "Same Dir" = inferred same direction for both regimes.

Dataset	**ROCHE**	**LOCI**	**LCUBE**
Stress-Fatigue	Correct	Correct	Correct
MPG-Horsepower	Correct	Same Dir	Correct
Emissions-Renewables	Correct	Inverted	Correct
UK Interest Rates	Same Dir	Same Dir	Same Dir
US Inflation-Fed Rate	Same Dir	Same Dir	Same Dir
Correct Rate	**60%**	**20%**	**60%**

ROCHE and LCUBE correctly detect regime-switching in 60% of cases. LOCI achieves only 20%, with one inverted detection (Emissions-Renewables). On economic datasets, all three algorithms infer the same direction for both regimes, failing to detect the reversal.

4.7. Chapter Summary

The three data-driven algorithms show clear patterns:

- **ROCHE and LCUBE** achieve 80% mean accuracy and correctly detect regime-switching in 60% of cases. They perform best on synthetic, engineering, and energy datasets.
- **LOCI** achieves 50% mean accuracy with only 20% correct regime-switching detection. It struggles most with the Emissions-Renewables dataset.

- **Economic datasets** (UK Interest Rates, US Inflation-Fed Rate) challenge all methods equally, with no algorithm exceeding 50% accuracy.

These results establish baselines for Chapter 5, where we test whether LLMs can match or exceed this performance using domain knowledge and reasoning.

5. LLM Evaluation Framework

This chapter tests whether Large Language Models can infer causal direction from numerical data. We describe the prompting strategy, experimental settings, and evaluation methods. We then compare LLM performance against the data-driven algorithms from Chapter 4.

5.1. Motivation and Research Questions

ROCHE, LOCI, and LCUBE use statistical methods to infer causal direction. LLMs work differently as they process text and generate responses based on patterns learned from training data. Prior work suggests LLMs struggle with formal causal reasoning, especially when context changes [MAS+25, CVD+24].

We test three questions:

- Can LLMs infer causal direction from numerical data?
- Does using variable names (vs. anonymous labels) improve performance?
- Can LLMs detect when causal direction reverses between regimes?

5.2. Model Selection

We use **GPT-5.2** (OpenAI, 2025) for all experiments. While our tool supports multiple LLM backends (OpenAI, Claude, Gemini), we use a single model to keep comparisons simple.

API setup. We access GPT-5.2 via OpenAI's Responses API with default settings. No temperature or randomness parameters are applied.

Why this approach. Our goal is to test whether LLMs can do regime-dependent causal inference "out of the box." We do not fine-tune or optimize the model. This gives a fair comparison: we test what the model can do with its default capabilities.

5.3. Multi-Step Prompting Strategy

We use a four-step prompting procedure. Breaking complex tasks into smaller steps tends to improve LLM performance compared to single-step prompts [BAD+25].

5.3.1. Design Rationale

Asking "What is the causal direction?" directly can cause the LLM to commit to an answer too early, before analyzing all the evidence. Our design separates hypothesis generation from regime analysis and final output.

Three principles guide the prompt design:

1. **Avoid early commitment.** The LLM should not conclude a direction until the final step.

2. **Define regime-switching clearly.** Direction reversal means $X \to Y$ becomes $Y \to X$, not just a change in strength.

3. **Stay neutral.** Ask for "plausible scenarios" for each direction, not which seems "more likely."

5.3.2. Four-Step Architecture

All configurations use the same four steps. Only the data format changes (raw vs. segmented, named vs. anonymous).

Step 1: Hypothesis Generation. The LLM receives a data summary and first produces a statistical overview (marginal distributions, correlation, notable patterns). It then generates competing causal hypotheses: "one plausible scenario where X could cause Y" and "one plausible scenario where Y could cause X," without assessing or preferring either direction.

Step 2: Regime Identification. The prompt gives the threshold variable and value explicitly. The LLM describes the two regimes (low/high) and considers whether the causal direction might reverse. The prompt clarifies: "Direction reversal means $X \to Y$ becomes $Y \to X$, not just different strengths."

Step 3: Causal Analysis. For each regime, the LLM states the causal direction, explains the mechanism in two to three sentences, assigns a confidence level (High, Moderate, or Low), and notes supporting evidence. It also answers: "Does the causal direction reverse between regimes?"

Step 4: JSON Output. The LLM outputs a structured JSON with the threshold variable, threshold value, and per-regime predictions. Each prediction includes a confidence label (High/Moderate/Low). The prompt says "Output ONLY the JSON" to enable automated parsing.

5.3.3. Prompt Chain Visualization

Figure 5.1 illustrates the complete four-step prompting chain used to query LLMs for regime-dependent causal inference. Each step builds on previous responses while introducing new reasoning requirements.

Step 1: Domain Understanding & Hypothesis Generation

```
Role: Data Scientist -- Explore
Potential Causal Directions

Instructions:
1. Statistical Overview:
  - Summarize marginal distributions
  - Describe correlation and patterns

2. Competing Causal Hypotheses:
  - For X → Y: Propose one plausible
scenario
  - For Y → X: Propose one plausible
scenario

IMPORTANT: Do not assess or prefer
one direction. Present both as viable
possibilities.
```

(a) Step 1 prevents anchoring by generating balanced hypotheses for both causal directions.

Step 2: Regime Identification

```
Role: Data Scientist -- Analyze
Causal Relationships Across Regimes

Segmentation from Domain Literature:
Data segmented at threshold_var =
value:
- Regime 0 (Low): threshold_var ≤
value
- Regime 1 (High): threshold_var >
value

Instructions:
1. For each regime, summarize
defining patterns
2. Assess if causal direction
reverses:
  - Reversal means X→Y switches to
Y→X
  - Only direction changes count, not
strength
3. Explain why mechanism could differ
or stay consistent
```

(b) Step 2 provides explicit threshold from domain literature and defines regime structure.

Step 3: Regime-Specific Causal Analysis

```
Role: Data Scientist -- Determine
Causal Directions and Mechanisms
Within Regimes

For each regime (Low and High):

1. Causal Direction: State X→Y
or Y→X
2. Mechanism: Explain why (2-3
sentences)
3. Confidence: High, Moderate, or
Low
4. Evidence: Note supporting data
patterns

Direction Switching:
- Does direction REVERSE between
regimes?
- If yes: Explain why mechanism flips
- If no: Explain why it remains
consistent
```

(c) Step 3 forces explicit per-regime reasoning before cross-regime comparison.

Step 4: JSON Output

```
Role: Data Scientist -- Extract and
Standardize Causal Analysis Results

Output ONLY valid JSON:

{
  "threshold_variable": "<name>",
  "threshold_value": <numeric>,
  "regimes": 2,
  "directions": [
    {"regime": 0,
     "direction": "<X->Y or Y->X>",
     "confidence":
"<High|Moderate|Low>"},
    {"regime": 1, ...}
  ]
}
```

(d) Step 4 extracts structured output for automated parsing and evaluation.

Figure 5.1.: Four-step prompting chain for regime-dependent causal inference. Each step builds on previous analysis while preventing premature commitment to a causal direction.

5.3.4. Algorithm vs LLM Input Comparison

Figure 5.2 contrasts the fundamentally different inputs received by data-driven algorithms and LLMs. While algorithms operate on standardized numeric arrays with no contextual information, LLMs receive structured textual descriptions with regime-specific statistics and explicit threshold specifications.

Algorithm Input & Execution

```
# Step 1: Standardize FULL dataset first
scaler = StandardScaler()
df[[X, Y]] = scaler.fit_transform(df[[X, Y]])

# Step 2: Split into regimes, then run
for regime in [low, high]:
    x = df[regime]['Variable_X'].values
    y = df[regime]['Variable_Y'].values

    # Algorithms receive NO context:
    # - No variable names/semantics
    # - No threshold information
    # - No regime membership labels

    # Output: numeric score per regime
    score = algorithm_infer(x, y)
    # score > 0: X->Y, score < 0: Y->X
```

(a) Algorithms receive standardized numeric arrays per regime. StandardScaler is fitted on the full dataset before splitting.

LLM Input (Segmented Anonymous)

```
Data segmented by Variable_Y at
threshold 0.5:

REGIME 0 (Low):  Variable_Y ≤ 0.5
Variable_X: Range=[0.05, 0.48]
Variable_Y: Range=[0.12, 0.50]
Sample (100 rows):  [...]

REGIME 1 (High):  Variable_Y > 0.5
Variable_X: Range=[0.51, 0.92]
Variable_Y: Range=[0.55, 0.98]
Sample (100 rows):  [...]
```

(b) LLMs receive per-regime statistics, data samples, and threshold context via 4-step prompting chain.

Figure 5.2.: Comparative workflow: (left) baseline algorithms always receive standardized numeric arrays per regime; (right) LLM input varies by configuration—shown here is `segmented_anon_hint_explicit`, which provides per-regime statistics, data samples, and explicit threshold specification. Both produce per-regime causal direction outputs.

5.3.5. Prompt Refinement Process

The multi-step prompting approach draws on findings that decomposing complex reasoning tasks into subtasks improves LLM performance [BAD+25]. The final four-step design resulted from iterative refinement. Key lessons from the development process:

- **Letting the LLM find thresholds failed.** When we asked the LLM to identify

regime boundaries itself, it produced thresholds that did not match the literature. We switched to giving explicit thresholds in Step 2.

- **Five steps became four.** An earlier design had separate steps for "domain understanding" and "hypothesis generation," but this was redundant. Merging them reduced token usage without losing reasoning depth.
- **Early assessment caused bias.** An "Initial Assessment" section led LLMs to conclude "the more plausible direction is..." before analyzing regimes, creating anchoring effects in subsequent steps. We removed it.
- **Strength vs. direction confusion.** LLMs sometimes called same-direction relationships with different strengths "regime-switching." We added an explicit clarification that only direction changes count as a reversal.
- **Variable name leakage in anonymized mode.** Step 2 prompts originally used the actual threshold variable name (e.g., "horsepower") even when the data was presented as "Variable_X" and "Variable_Y." This undermined the anonymous configurations. We added dynamic variable mapping so the threshold prompt uses "Variable_X" or "Variable_Y" when anonymization is active.
- **Context ownership and step labeling.** Earlier prompts labelled previous analysis generically ("Previous Analysis:") without making clear that the content was the model's own output. We changed this to "Your Previous Analysis (Steps 1–2):" with visual delimiters, which improved reasoning continuity across steps.

5.4. Experimental Configurations

We test four configurations that vary along two dimensions: (1) data format (raw vs. pre-segmented by regime), and (2) variable naming (actual names vs. anonymous labels). The naming axis follows prior work that compares abstract vs. meaningful variable names to test whether LLMs rely on semantic priors [HK24, DLA+25]. This 2×2 design lets us isolate the effects of each factor.

5.4.1. Configuration Design

1. **Raw + Named** (`raw_named_hint_explicit`): Raw data with actual variable names.
2. **Raw + Anonymous** (`raw_anon_hint_explicit`): Raw data with labels like "Variable_X, Variable_Y."
3. **Segmented + Named** (`segmented_named_hint_explicit`): Data pre-split by regime with actual variable names.
4. **Segmented + Anonymous** (`segmented_anon_hint_explicit`): Data pre-split by regime with anonymous labels.

All configurations give explicit threshold information. The only differences are data format and variable naming.

5.4.2. Configuration Rationale

Baseline comparison. The `segmented_anon_hint_explicit` configuration gives the fairest comparison to data-driven algorithms. Anonymous labels reduce reliance on semantic priors in variable names [HK24, DLA+25], and pre-segmented data matches how algorithms process data.

Testing semantic knowledge. The `segmented_named_hint_explicit` configuration adds actual variable names to test whether domain knowledge improves LLM performance.

Testing autonomous analysis. The raw configurations test whether LLMs can work with unsegmented data when given explicit threshold information.

5.5. Data Representation for LLMs

LLMs cannot process numerical data directly like algorithms do. A Transformer treats each number as a sequence of text tokens, not as a value in a number field. Small numerical perturbations can therefore change LLM conclusions even when the logical structure is unchanged [MAS+25]. We must convert data into text that fits within token limits while keeping enough information for causal inference.

5.5.1. Numerical Data Encoding

LLMs receive data samples rather than complete datasets. Complete datasets would exceed token limits, but summaries alone would lose distributional information. Our pipeline:

1. **Regime definition**: Threshold values come from literature (not estimated from data). Each observation goes into the low ($\leq \tau$) or high ($> \tau$) regime.
2. **Data filtering**: Missing values are removed. No imputation.
3. **Sampling**: For large datasets, we randomly sample up to **200 observations** (100 per regime in segmented mode) with a fixed random seed for reproducibility. This applies only to LLM evaluation since algorithms use all data.

For each regime, LLMs receive the value range (minimum and maximum) plus the sampled data points. Mean and standard deviation were disabled in the final configuration to reduce prompt length and to avoid giving the LLM pre-computed summary statistics that the data-driven algorithms do not receive.

5.5.2. Regime Information Encoding

Regime information is given explicitly:

- **Threshold variable**: Which variable defines the regime split
- **Threshold value**: The numeric boundary between low and high
- **Regime labels**: "Low" means $\leq$ threshold, "High" means $>$ threshold

We give thresholds explicitly because pilot tests showed that LLMs produced poor thresholds when asked to find them on their own.

5.6. Evaluation Protocol

LLMs are tested without fine-tuning. Each test gives the LLM a data description plus regime information from domain literature.

5.6.1. Output Parsing

Responses are parsed using Python's `json.loads()`. If parsing fails, the evaluation is marked as failed.

The JSON output includes:

- `threshold_variable`: The regime-defining variable
- `threshold_value`: The numeric threshold
- `directions`: Per-regime causal directions with confidence labels

Only the direction predictions count for accuracy. Confidence labels are for qualitative analysis only.

5.6.2. Evaluation Criteria

We evaluate LLM responses on:

- **Direction accuracy**: Does the LLM correctly infer $X \to Y$ or $Y \to X$?
- **Regime-switching detection**: Does the LLM correctly identify when direction reverses?

We follow prior LLM-based causal discovery evaluations that measure direction-level correctness against known ground truth [RRWT24, HK24]. We use the same regime-level accuracy metric as Chapter 4.

5.7. Comparison with Algorithmic Results

The evaluation framework is comparative, not optimization-focused. All baseline algorithms use default hyperparameters from their original implementations, and LLMs are not fine-tuned. This ensures that performance differences reflect inference quality rather than tuning effort.

For each dataset and configuration, the evaluation follows three stages:

1. **Method execution.** Algorithms (ROCHE, LOCI, LCUBE) run on each regime subset independently, producing one direction inference per regime. The LLM follows the four-step prompting chain, and the final JSON is parsed to extract per-regime directions.

2. **Accuracy scoring.** Each method's per-regime inferences are compared against ground truth using the regime-level accuracy metric from Chapter 4. We also record whether each method detected regime-switching.

3. **Results export.** The tool exports CSV files with accuracy scores, JSON summaries, and a ZIP archive of all LLM prompt logs (prompts sent and responses received for each step).

We do not apply statistical significance tests because algorithms produce deterministic output and each LLM configuration is run once per dataset.

5.8. Implementation Details

The evaluation tool is built in Python with a Streamlit interface. It supports:

- Dataset selection and LLM configuration switching
- Scatter plots and regime segmentation visualization before prompting
- Display of the full four-step reasoning chain during execution
- Export of results as CSV, JSON, and ZIP archives of prompt logs

Appendix A documents the full application architecture, key design decisions made during development, and the batch experiment pipeline.

5.9. Results

Table 5.1 shows GPT-5.2 accuracy on each dataset using the baseline configuration (`segmented_anon_hint_explicit`).

GPT-5.2 achieves 30% mean accuracy, compared to 80% for ROCHE and LCUBE and 50% for LOCI. The LLM fails completely (0%) on both economic datasets. Confidence labels (High/Moderate/Low) show that the model reports moderate confidence even on incorrect predictions.

Table 5.1.: GPT-5.2 regime-level accuracy by dataset (baseline configuration).

Dataset	**Low/High Predictions**	**Accuracy**
Stress-Fatigue (synthetic)	$Y \to X$ (Mod) / $Y \to X$ (Mod)	50%
MPG-Horsepower	$X \to Y$ (Low) / $X \to Y$ (Mod)	50%
Emissions-Renewables	$Y \to X$ (Mod) / $Y \to X$ (Low)	50%
UK Interest Rates	$Y \to X$ (Low) / $X \to Y$ (Mod)	0%
US Inflation-Fed Rate	$X \to Y$ (Low) / $Y \to X$ (Mod)	0%
Mean		**30%**

5.9.1. Regime-Switching Detection

Table 5.2 shows whether each method correctly detected regime-switching across all five datasets.

Table 5.2.: Regime-switching detection across all methods. All datasets have true reversals.

Dataset	**ROCHE**	**LOCI**	**LCUBE**	**GPT-5.2**
Stress-Fatigue	Correct	Correct	Correct	Same Dir
MPG-Horsepower	Correct	Same Dir	Correct	Same Dir
Emissions-Renewables	Correct	Inverted	Correct	Same Dir
UK Interest Rates	Same Dir	Same Dir	Same Dir	Inverted
US Inflation-Fed Rate	Same Dir	Same Dir	Same Dir	Inverted
Correct Rate	**60%**	**20%**	**60%**	**0%**

"Correct" means the method detected the reversal with correct directions (100% accuracy). "Inverted" means it detected a reversal but got both directions backwards (0% accuracy). "Same Dir" means it inferred the same direction for both regimes (50% accuracy).

ROCHE and LCUBE correctly detect regime-switching on 60% of datasets—all three non-economic datasets. LOCI detects only one (20%) and inverts the emissions dataset. GPT-5.2 achieves 0% correct detection: on three datasets it predicts the same direction for both regimes, and on both economic datasets it detects a reversal but gets both directions backwards.

5.9.2. Configuration Effects

Performance varies across configurations:

Variable naming matters. Named configurations outperform anonymous ones on most datasets. Emissions-Renewables shows 100% accuracy with named variables but only 50% with anonymous labels. US Inflation shows a similar pattern: 50% with named variables, 0% with anonymous labels. This indicates that domain-specific variable names activate semantic knowledge that improves causal reasoning, consistent with findings by [HK24] and [DLA+25].

Segmentation effects are mixed. Pre-segmented data does not consistently help. Auto MPG achieves 100% with raw+named but only 50% with segmented+named. The interaction between data format and naming varies by dataset.

UK Interest Rates fails across all configurations. This dataset achieves 0% accuracy regardless of naming or segmentation. Variable naming provides no benefit when the underlying causal structure exceeds the model's capability.

5.9.3. Comparison Summary

Table 5.3 compares all methods.

Table 5.3.: Method comparison: mean accuracy and regime-switching detection.

Method	Mean Acc	Switching Detection
ROCHE	80%	60%
LCUBE	80%	60%
LOCI	50%	20%
GPT-5.2	30%	0%

Data-driven methods outperform GPT-5.2 on both metrics. ROCHE and LCUBE achieve 80% accuracy and correctly detect regime-switching in 60% of cases. LOCI reaches 50% accuracy and detects switching in one dataset (20%). GPT-5.2 achieves only 30% accuracy and never correctly detects regime-switching.

5.10. Chapter Summary

Key findings:

- **GPT-5.2 achieves 30% mean accuracy**, compared to 80% for ROCHE and LCUBE and 50% for LOCI.
- **Regime-switching detection: 0%** for GPT-5.2, vs. 60% for ROCHE and LCUBE and 20% for LOCI.
- **Variable names help.** Named configurations outperform anonymous ones on some datasets.
- **Economic datasets are hardest.** All data-driven methods drop to 50% accuracy on UK Interest Rates and US Inflation, meaning no method detects regime-switching on these datasets. GPT-5.2 fails completely (0%).

These results show that current LLMs cannot match data-driven methods for regime-dependent causal discovery from numerical data.

6. Experimental Results and Comparison

This chapter presents the experimental results comparing data-driven causal discovery algorithms (ROCHE, LOCI, LCUBE) against LLM-based causal inference across all datasets and configurations. We analyze performance patterns, identify failure modes, and discuss implications for regime-dependent causal reasoning.

6.1. Experimental Objective

This section describes the experimental setup for evaluating regime-dependent causal discovery methods. We compare GPT-5.2 against three classical algorithms (ROCHE, LOCI, LCUBE) across five bivariate pairs with literature-based ground truth. Each method is evaluated on its ability to correctly infer causal direction within each regime, scored using the regime-level accuracy metric defined in Equation 3.1.

6.2. Performance Overview

We evaluate four methods (ROCHE, LOCI, LCUBE, and GPT-5.2) across five bivariate datasets with regime-segmented data. Baseline algorithms are reproducible across runs: ROCHE and LOCI use fixed random seeds to ensure consistent results, while LCUBE is fully deterministic. All three algorithms therefore produce consistent outputs for each dataset. GPT-5.2 is evaluated under the `segmented_anon_hint_explicit` configuration, which provides a fair comparison to algorithms by using anonymized variable names and pre-segmented regime data.

Table 6.1 summarizes overall performance across all five datasets.

Table 6.1.: Overall method performance across five datasets (baseline configuration). Perfect = both regimes correct; Failed = both regimes incorrect.

Method	**Mean Acc.**	**Perfect**	**Failed**
ROCHE	80%	60%	0%
LOCI	50%	20%	20%
LCUBE	80%	60%	0%
GPT-5.2	30%	0%	40%

Table 6.2.: Method performance by dataset using `segmented_anon_hint_explicit` configuration. GPT-5.2 predictions show inferred direction and confidence for low/high regimes. Anonymous variables shown as Variable_X/Variable_Y in LLM predictions.

Dataset	Threshold	ROCHE	LOCI	LCUBE	GPT-5.2 (Low/High)	GPT-5
Synthetic Stress-Fatigue	StressLevel @ 0.5	100%	100%	100%	Y→X (Mod) / Y→X (Mod)	50%
Auto MPG-Horsepower	horsepower @ 140.0	100%	50%	100%	X→Y (Low) / X→Y (Mod)	50%
Emissions-Renewables	renewable @ 15.0	100%	0%	100%	Y→X (Mod) / Y→X (Low)	50%
UK Interest Rates	short @ 3.5	50%	50%	50%	Y→X (Low) / X→Y (Mod)	0%
US Inflation-Fed Rate	inflation @ 5.4	50%	50%	50%	X→Y (Low) / Y→X (Mod)	0%
Mean		**80%**	**50%**	**80%**		**30%**

6.2.1. Baseline Algorithm Performance

ROCHE and LCUBE achieve the highest mean accuracy (80% each), with 60% of datasets achieving perfect accuracy (both regimes correct) and no complete failures. LOCI shows lower performance (50% mean accuracy) with sensitivity to distributional assumptions, achieving only 20% perfect accuracy while experiencing 20% complete failures. All three algorithms achieve 100% accuracy on the synthetic validation dataset, confirming that regime-dependent mechanisms are detectable when present.

6.2.2. LLM Performance

GPT-5.2 achieves 30% mean accuracy under the `segmented_anon_hint_explicit` configuration, falling substantially below algorithmic baselines. Most notably, GPT-5.2 achieves 0% perfect accuracy (never correctly identifying both regimes for any dataset) while experiencing 40% complete failures (0% accuracy on two datasets). This performance pattern suggests limitations in regime-dependent numerical causal reasoning even with structured four-step prompting and pre-segmented regime data.

6.3. Dataset-Level Analysis

Table 6.2 presents detailed results by dataset, showing method performance and GPT-5.2's regime-specific predictions.

6.3.1. Synthetic Baseline Results

The synthetic Stress-Fatigue dataset validates the experimental methodology. All three data-driven algorithms (ROCHE, LOCI, LCUBE) achieve perfect accuracy (100%), confirming that regime-dependent causal reversals are detectable under controlled conditions with clear statistical patterns. GPT-5.2 achieves only 50% accuracy, correctly identifying one regime while failing on the other, despite receiving explicit threshold information and pre-segmented data.

6.3.2. Real-World Dataset Results

Real-world datasets reveal varying challenges across domain complexity and method capabilities:

Auto MPG-Horsepower (Automotive Engineering): ROCHE and LCUBE achieve perfect accuracy (100%), demonstrating robust detection of manufacturing design regime switching. LOCI achieves 50% accuracy, showing sensitivity to the twist-region characteristics of this dataset. GPT-5.2 achieves 50% with identical predictions for both regimes.

Emissions-Renewables (Energy Systems): ROCHE and LCUBE again achieve perfect accuracy (100%), while LOCI fails completely (0%), suggesting strong sensitivity to distributional assumptions violated by this dataset. GPT-5.2 achieves 50% with consistent Y→X predictions across regimes.

UK Interest Rates (Monetary Policy): All methods struggle equally, with ROCHE, LOCI, and LCUBE each achieving 50% accuracy. GPT-5.2 achieves 0% accuracy on this dataset. Although it detects that a reversal occurs, it assigns both regime directions exactly opposite to the ground truth (inverted failure).

US Inflation-Fed Rate (Monetary Policy): All data-driven algorithms achieve 50% accuracy, indicating that economic datasets with subtle regime boundaries challenge statistical methods. GPT-5.2 achieves 0% (complete failure), detecting a direction switch but predicting both directions backwards (inverted).

6.3.3. Performance Patterns

Performance reveals systematic patterns across domain complexity and method capabilities. The synthetic validation confirms data-driven algorithms achieve perfect detection of controlled regime-switching, while GPT-5.2 reaches only 50% accuracy despite explicit regime information. Real-world datasets show varying challenges: automotive engineering (Auto MPG) enables consistent algorithmic detection (ROCHE/LCUBE 100%), while energy systems (Emissions-Renewables) expose divergent behaviors with LOCI failing completely (0%) and others succeeding. Economic datasets (interest rates, inflation) challenge all approaches equally, with mixed 50% performance across methods. LOCI shows difficulties across multiple datasets, suggesting sensitivity to distributional assumptions violated by real-world data.

6.4. Configuration Effects on LLM Performance

GPT-5.2 performance varies substantially across the four experimental configurations, revealing systematic effects of data representation and variable naming on causal reasoning quality.

6.4.1. Variable Naming Effects

Named configurations improve performance on 2 out of 5 datasets; on the remaining 3, naming has no effect on accuracy. On Emissions-Renewables, named variables achieve 100% accuracy in both raw and segmented settings, compared to 50% for anonymous in both cases. This pattern indicates that LLMs rely heavily on semantic associations from training data when inferring causal relationships, rather than analyzing statistical patterns in the numerical data.

6.4.2. Data Segmentation Effects

The effect of pre-segmented regime data is mixed and dataset-dependent. On the Synthetic dataset, segmented configurations (50% for both named and anonymous) outperform raw+named (0%), though raw+anonymous also reaches 50%. On Auto MPG, raw+named achieves 100% accuracy while all segmented configurations reach only 50%, showing that segmentation can hurt when the model already benefits from variable names. Overall, segmentation does not provide a consistent benefit across datasets.

6.4.3. Configuration Comparison

Figure 6.1.: GPT-5.2 accuracy across five datasets and four experimental configurations. Darker cells indicate higher accuracy. Named configurations (left columns) generally outperform anonymous (right columns), while segmentation effects vary by dataset.

The configuration effects demonstrate that LLM causal reasoning is highly sensitive to presentation format. The best LLM configuration (`segmented_named_hint_explicit`) achieves substantially higher accuracy than the worst (`raw_anon_hint_explicit`), but even optimal configurations fail to match data-driven algorithm performance.

6.5. Regime-Switching Detection

A key capability for regime-dependent causal discovery is detecting when causal direction actually reverses between regimes. Table 6.3 analyzes whether methods detect actual causal direction reversals. All five datasets exhibit true causal direction reversal between regimes by design.

Table 6.3.: Regime-dependent direction switching detection. All datasets exhibit true causal direction reversal between regimes. "Correct" = detected reversal with correct directions (100% accuracy); "Inverted" = detected reversal but directions are backwards (0% accuracy); "Same Dir" = infers identical direction for both regimes (50% accuracy). GPT-5.2 results shown for the baseline `segmented_anon_hint_explicit` configuration; other configurations may differ per dataset.

Dataset	**ROCHE**	**LOCI**	**LCUBE**	**GPT-5.2**
Synthetic Stress-Fatigue	Correct	Correct	Correct	*Same Dir*
Auto MPG-Horsepower	Correct	*Same Dir*	Correct	*Same Dir*
Emissions-Renewables	Correct	**Inverted**	Correct	*Same Dir*
UK Interest Rates	*Same Dir*	*Same Dir*	*Same Dir*	**Inverted**
US Inflation-Fed Rate	*Same Dir*	*Same Dir*	*Same Dir*	**Inverted**
Correct Rate	**60%**	**20%**	**60%**	**0%**

6.5.1. Detection Rates

ROCHE and LCUBE show consistent correct regime-switching detection (60% each), successfully identifying direction reversals on the synthetic, automotive, and energy datasets. LOCI performs moderately (20% correct rate), detecting switches but with one inverted detection on Emissions-Renewables where it predicts directions backwards. GPT-5.2 achieves 0% correct detection rate—it never correctly identifies both the reversal and the correct directions. When GPT-5.2 detects a reversal (UK Interest Rates, US Inflation), it predicts the directions backwards (inverted). On other datasets, it defaults to same-direction predictions.

6.5.2. Failure Pattern Analysis

The failure patterns reveal fundamental differences between algorithmic and LLM approaches. Data-driven algorithms fail to detect regime-switching primarily on economic datasets (UK Interest Rates, US Inflation-Fed Rate), where subtle statistical patterns may be obscured by noise or violate distributional assumptions. In contrast, GPT-5.2 exhibits two distinct failure modes:

1. **Same-direction failures:** On datasets where algorithms succeed (Synthetic, Auto MPG, Emissions-Renewables), GPT-5.2 defaults to a single global direction despite

explicit threshold information, achieving 50% accuracy by chance.

2. **Inverted failures:** On economic datasets (UK Interest Rates, US Inflation-Fed Rate), GPT-5.2 correctly detects that a reversal occurs but predicts the directions backwards, achieving 0% accuracy. This "inverted" pattern suggests the model applies incorrect domain knowledge about economic mechanisms.

This complementary failure pattern suggests that algorithms and LLMs rely on fundamentally different signals: algorithms detect statistical asymmetries in functional relationships, while LLMs draw on semantic associations from training data that may not align with regime-specific mechanisms.

6.6. LLM Reasoning Quality Analysis

Beyond quantitative accuracy, we analyze the qualitative characteristics of LLM reasoning to understand failure modes and reasoning patterns.

6.6.1. Successful Reasoning Examples

GPT-5.2 achieves highest accuracy on Emissions-Renewables with the `segmented_named_hint_explicit` configuration (100% accuracy). Figure 6.2 shows excerpts from the model's reasoning chain, demonstrating correct regime-specific causal analysis.

In this successful case, the model correctly leverages domain knowledge about renewable energy infrastructure maturity and its effects on emissions. The reasoning chain demonstrates proper regime-specific analysis: in the low-renewables regime, the model recognizes that emissions drive policy responses (X→Y), while in the high-renewables regime, it correctly identifies that renewable infrastructure can materially reduce emissions (Y→X).

6.6.2. Failure Mode Examples

On the Synthetic Stress-Fatigue dataset with `segmented_anon_hint_explicit` configuration, GPT-5.2 achieves only 50% accuracy despite the dataset's clear statistical patterns designed to exhibit regime-switching. Figure 6.3 shows the model's reasoning, which incorrectly concludes that no direction reversal occurs.

6.6.3. Inverted Failure Example

A distinct failure mode occurs on economic datasets where GPT-5.2 detects a reversal but predicts the directions backwards. Figure 6.4 shows this "inverted" pattern on the UK Interest Rates dataset.

This inverted failure case reveals that detecting a reversal is not sufficient—the model must also correctly identify which direction applies in which regime. On economic datasets, GPT-5.2 appears to apply plausible but incorrect domain reasoning about monetary policy mechanisms, demonstrating that semantic knowledge can lead to systematic errors when it conflicts with ground truth.

GPT-5.2 Step 3 Analysis (Emissions-Renewables, Named Variables)

```
REGIME 0 (Low renewables: Y ≤ 15%)
1) Causal Direction: X → Y (emissions per capita influence renewable share)
2) Mechanism: In the low-renewables regime, renewables are still ‘‘niche’’ and
largely determined by policy/investment constraints. Higher emissions tend
to increase policy salience (air quality, climate commitments) and can trigger
renewable support, so emissions pressure is more likely to move renewable share than
small changes in renewable share are to move total per-capita emissions.
3) Confidence: Moderate
4) Evidence: Y is tightly bounded near zero (mean 4.67, many 0% values); X has very
large dispersion and extreme upper tail (up to 38.84).

REGIME 1 (High renewables: Y > 15%)
1) Causal Direction: Y → X (renewable share influences emissions per capita)
2) Mechanism: Once renewables are a material share of generation, changing the
electricity mix plausibly displaces fossil generation and reduces total per-capita
emissions. In this regime, renewable shares vary widely (15-100), so Y has
meaningful room to affect system-level emissions outcomes.
3) Confidence: Moderate

Direction Switching: Does the causal direction reverse? Yes
Below 15% renewables, adoption is primarily a policy response to high-emissions
fossil dependence (X→Y dominates), while above 15% renewables are large enough that
marginal changes more directly alter the generation mix and therefore emissions
(Y→X dominates).
```

(a) Step 3 reasoning correctly identifies regime-specific mechanisms and direction reversal.

GPT-5.2 JSON Output (100% Accuracy)

```
{
  "threshold_variable": "renewable_share_pct",
  "threshold_value": 15.0,
  "regimes": 2,
  "directions": [
    {"regime": 0,
     "direction": "emissions→renewable",
     "confidence": "Moderate"},
    {"regime": 1,
     "direction": "renewable→emissions",
     "confidence": "Moderate"}
  ]
}
```

(b) Final JSON output matches ground truth for both regimes.

Figure 6.2.: Successful LLM reasoning example: Emissions-Renewables dataset with `segmented_named_hint_explicit` configuration. The model correctly leverages domain knowledge about renewable energy infrastructure maturity, correctly identifying the direction reversal between regimes (X→Y in low regime, Y→X in high regime).

GPT-5.2 Step 3 Analysis (Synthetic, Anonymous Variables) — FAILURE

```
REGIME 0 (Low):  Variable_Y ≤ 0.5
1) Causal Direction:  Y → X
2) Mechanism:  The regime is defined by Y, and X appears structurally capped in this
regime (X ∈ [0, 0.50]), consistent with a rule/assignment where low Y constrains the
allowable/assigned X. The strong heaping of Y at 0.49 suggests censoring near the
threshold, reinforcing that Y is driving a gating process.
3) Confidence:  High

REGIME 1 (High):  Variable_Y > 0.5
1) Causal Direction:  Y → X
2) Mechanism:  Once Y crosses 0.5, X shifts upward and rarely falls below ∼0.42,
consistent with a policy/activation mechanism where being in the high-Y state
enables or assigns higher X.
3) Confidence:  Moderate

Direction Switching:  Does the causal direction reverse?  No.
In both regimes, the most salient pattern is X's feasible range shifting
with Y-defined regime membership.  That is more consistent with a single
gating/assignment mechanism from Y to X than with X driving Y differently on each
side.
```

(a) Step 3 reasoning incorrectly concludes Y→X for both regimes, missing the designed reversal.

GPT-5.2 JSON Output (50% Accuracy)

```
{
  "threshold_variable":  "Variable_Y",
  "threshold_value":  0.5,
  "regimes":  2,
  "directions":  [
    {"regime":  0,
     "direction":  "Variable_Y→Variable_X",
     "confidence":  "High"},
    {"regime":  1,
     "direction":  "Variable_Y→Variable_X",
     "confidence":  "Moderate"}
  ]
}
```

(b) Final JSON output predicts same direction for both regimes—incorrect.

Figure 6.3.: Failed LLM reasoning example (Same-Direction): Synthetic Stress-Fatigue dataset with `segmented_anon_hint_explicit` configuration. Ground truth is X→Y in low regime and Y→X in high regime (direction reversal). The model incorrectly predicts Y→X for *both* regimes, failing to detect the designed causal reversal. Without semantic variable names, the model defaults to interpreting statistical patterns as a "gating mechanism" rather than analyzing regime-specific causal asymmetries.

GPT-5.2 Step 2 Analysis (UK Interest Rates, Named Variables) — INVERTED FAILURE

Regime 0 (Low short $\leq$ 3.5): "Short rate is compressed; long rate remains relatively free to move." High dispersion of long rates while short rates stay near ∼0.4-0.7, consistent with long rates moving "independently" of current short rates.

Regime 1 (High short $>$ 3.5): "Stronger co-movement in levels": when short is 5-12, long is typically also in a similar band (often ∼4-13).

(a) Step 2 correctly identifies distinct regime characteristics but misinterprets causal implications.

GPT-5.2 Step 3 Analysis (UK Interest Rates) — INVERTED FAILURE

Does the causal direction reverse? Yes.

"In the low regime, the short rate is constrained/inertial so long rates [...] can act as a signal that influences later short-rate moves (Y→X). In the high regime, policy is active and long yields more directly price the policy-rate path (X→Y)."

Ground truth: Short→Long (low regime), Long→Short (high regime)

(b) Step 3 detects reversal but predicts directions backwards—assigns Long→Short in low regime (ground truth: Short→Long) and Short→Long in high regime (ground truth: Long→Short).

Figure 6.4.: Inverted failure example: UK Interest Rates dataset with `segmented_named_hint_explicit` configuration. The model correctly detects that causal direction reverses between regimes but predicts the directions exactly backwards, achieving 0% accuracy. This "inverted" failure pattern appears on both economic datasets, suggesting the model applies incorrect domain knowledge about monetary policy mechanisms.

6.6.4. Same-Direction vs Inverted Failures

The two failure modes have different implications:

- **Same-direction failures** (50% accuracy): The model ignores regime structure entirely, defaulting to a single global direction. This occurs primarily on non-economic datasets with anonymous variables.
- **Inverted failures** (0% accuracy): The model correctly identifies regime-dependent structure but applies incorrect domain knowledge. This occurs on economic datasets even with named variables.

This distinction suggests that LLM performance depends on both the availability of semantic cues (variable names) and the accuracy of domain knowledge encoded during training. For economic relationships with complex or contested causal mechanisms, training data may encode plausible but incorrect associations.

6.6.5. Common Failure Modes

Analysis of GPT-5.2 reasoning chains reveals systematic failure patterns:

1. **Global direction anchoring:** The model often commits to a single causal direction early in reasoning and maintains it across regimes, ignoring explicit information about regime-dependent mechanisms.
2. **Confusing direction reversal with strength changes:** The model sometimes interprets regime-dependent causality as changes in relationship strength rather than direction reversal.
3. **Inverted domain knowledge:** On economic datasets, the model applies domain reasoning that produces exactly reversed predictions, suggesting training data encodes incorrect or contested causal relationships.
4. **Semantic over-reliance:** Named variable configurations show dramatically higher accuracy, indicating the model relies heavily on semantic associations rather than statistical pattern analysis.
5. **Insufficient numerical attention:** Despite receiving sampled data points, the model's reasoning chains rarely reference specific numerical patterns that distinguish regimes.

6.7. Method Complementarity Analysis

Different methods work better on different datasets. ROCHE and LCUBE achieve 100% accuracy on synthetic and engineering/energy datasets, but reach only 50% on economic data. LOCI fails on Emissions-Renewables (0%) but reaches 50% on economics. GPT-5.2

achieves 100% on two datasets with named variables (Auto MPG, Emissions) but achieves 0% on economic pairs and 50% on synthetic data. Because methods perform well on different datasets, combining them could improve results.

Table 6.4 summarizes method performance by domain category.

Table 6.4.: Method accuracy by domain category. *GPT-5.2 range reflects variation across configurations (anonymous vs. named variables).

Domain	ROCHE	LOCI	LCUBE	GPT-5.2
Synthetic	100%	100%	100%	0–50%*
Engineering	100%	50%	100%	50–100%*
Energy	100%	0%	100%	50–100%*
Economics	50%	50%	50%	0–50%*

6.7.1. Domain-Specific Observations

Synthetic validation: All data-driven algorithms achieve 100% accuracy, confirming correct regime-switching detection. GPT-5.2 reaches only 50%, failing to detect the switch even with explicit threshold information.

Engineering/Energy domains: ROCHE and LCUBE show robust performance (100%), while LOCI struggles on Emissions-Renewables (0%). GPT-5.2 achieves 100% with named variables but drops to 50% with anonymization. These domains have clear physical mechanisms that algorithms detect reliably.

Economics: All methods show reduced performance (50% for algorithms, 0–50% for GPT-5.2). Economic datasets present challenges beyond current method capabilities, with subtle regime boundaries and complex causal mechanisms that may be contested in the literature.

6.7.2. Best Performers by Domain

Performance varies systematically by domain type:

- **Synthetic (Controlled):** ROCHE, LOCI, LCUBE (100%); GPT-5.2 (50%)
- **Automotive Engineering:** ROCHE, LCUBE (100%); LOCI (50%); GPT-5.2 (50–100%)
- **Energy Systems:** ROCHE, LCUBE (100%); LOCI (0%); GPT-5.2 (50–100%)
- **Monetary Policy (UK):** All methods (50% or lower)
- **Monetary Policy (US):** ROCHE, LOCI, LCUBE (50%); GPT-5.2 (0–50%)

Engineering and environmental domains with clear statistical patterns enable robust algorithmic detection, while economic domains with noisy signals and subtle regime boundaries challenge all approaches equally.

6.7.3. Algorithm Agreement Analysis

Algorithm agreement patterns reveal dataset difficulty:

- **Full agreement (Synthetic):** ROCHE, LOCI, and LCUBE all achieve 100%, indicating clear statistical signals.
- **Partial agreement (Auto MPG, Emissions-Renewables):** ROCHE and LCUBE achieve 100% while LOCI diverges (50% on Auto MPG, 0% inverted on Emissions), suggesting LOCI's distributional assumptions are violated on these datasets.
- **Low agreement (Economic datasets):** All algorithms produce mixed results, indicating genuine difficulty in regime-dependent inference for these domains.

6.8. Visualization of Results

Figure 6.5 visualizes the regime-dependent causal structure of each dataset. Each panel shows the bivariate data partitioned by the literature-based threshold, with fitted slopes indicating the inferred causal direction within each regime. The direction reversals are visible as opposing slope signs between the low (blue) and high (red) regime segments.

6.9. Summary of Findings

6.9.1. Key Results

The experimental evaluation yields the following key findings:

1. **Data-driven algorithms outperform LLMs:** ROCHE and LCUBE achieve 80% mean accuracy compared to GPT-5.2's 30%, demonstrating that established causal discovery methods outperform current LLMs on regime-dependent numerical data.
2. **LLMs struggle with numerical causal reasoning:** Despite structured four-step prompting with explicit regime information, GPT-5.2 never achieves perfect accuracy on any dataset (0% perfect rate) and fails completely on 40% of datasets, indicating fundamental limitations in numerical pattern analysis.
3. **Zero correct regime-switching detection:** GPT-5.2 achieves 0% correct detection rate for regime-switching. When it detects a reversal (economic datasets), it predicts directions backwards (inverted failures). On other datasets, it defaults to same-direction predictions.
4. **Variable naming has major impact:** Named configurations dramatically outperform anonymous configurations, revealing LLM reliance on semantic associations rather than statistical inference.

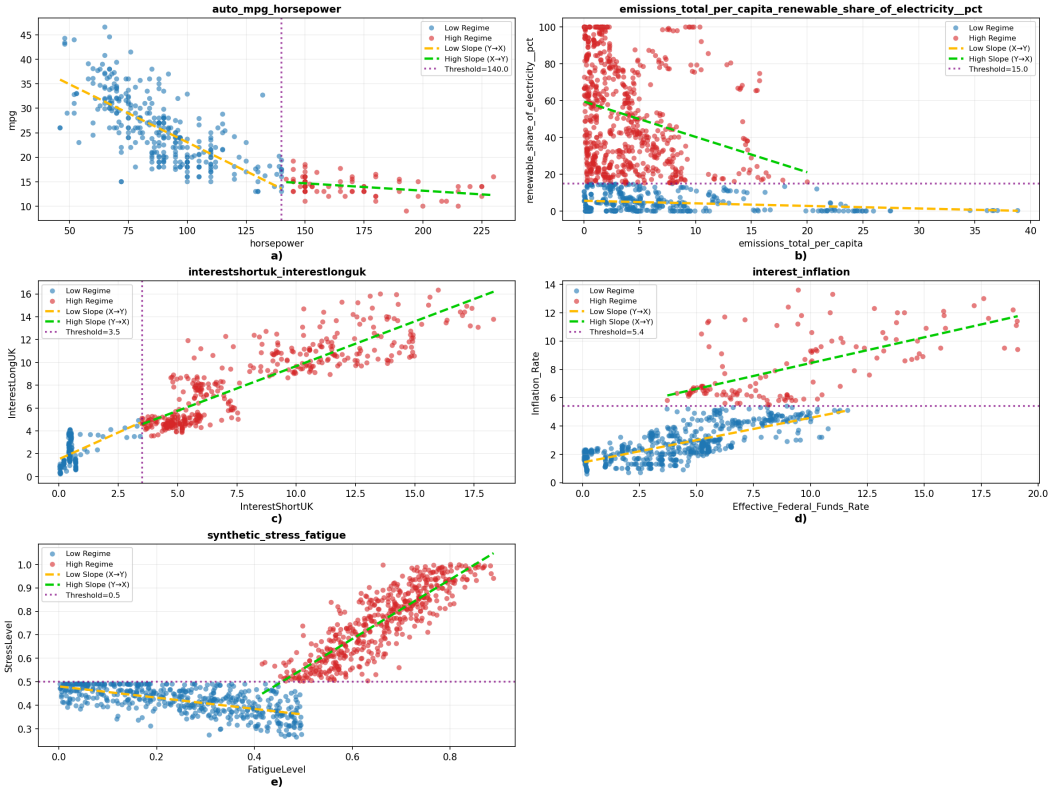

Figure 6.5.: Regime-dependent causal structure for each dataset. Blue points represent the low regime and red points represent the high regime. Yellow dashed lines show regime-low fitted slopes with their causal directions, while green dashed lines show regime-high slopes. Purple dotted lines mark the threshold value.

5. **Pre-segmentation effects are mixed:** Providing pre-segmented regime data improves LLM performance on some datasets (e.g., Synthetic) but can reduce it on others (e.g., Auto MPG), and never closes the gap with algorithmic methods.

6.9.2. Implications

These findings have important implications for causal inference methodology:

- **LLMs rely on semantic knowledge:** The strong naming effects suggest LLMs primarily draw on training data associations rather than analyzing presented numerical evidence for causal inference.
- **Inverted domain knowledge is problematic:** On economic datasets, LLMs apply plausible but incorrect reasoning, producing exactly reversed predictions. This suggests training data may encode contested or incorrect causal relationships for complex domains.
- **Current LLMs unsuitable for regime-dependent discovery:** The consistent underperformance indicates that current LLM architectures are not suitable for regime-dependent causal discovery tasks requiring numerical pattern recognition.
- **Hybrid approaches may be promising:** The complementary failure patterns (algorithms struggle on economic data; LLMs struggle on clear statistical patterns) suggest potential value in hybrid approaches combining LLM domain knowledge with algorithmic inference.

6.10. Chapter Summary

This chapter presented comprehensive experimental results comparing data-driven causal discovery algorithms (ROCHE, LOCI, LCUBE) against GPT-5.2 on five bivariate datasets with regime-dependent causal mechanisms. Key findings demonstrate that established algorithms substantially outperform LLMs, with ROCHE and LCUBE achieving 80% mean accuracy compared to GPT-5.2's 30%. The regime-switching detection analysis reveals that GPT-5.2 achieves 0% correct detection rate—when it detects reversals, it predicts directions backwards (inverted failures on economic datasets). Configuration effects show that LLM performance is highly sensitive to variable naming and data segmentation, indicating reliance on semantic associations rather than numerical pattern analysis. These results establish that current LLMs are not suitable replacements for data-driven methods in regime-dependent causal discovery tasks, while suggesting potential complementarity for hybrid approaches.

7. Discussion and Conclusion

This chapter discusses the implications of our experimental findings, acknowledges limitations, and proposes directions for future research on regime-dependent causal discovery.

7.1. Discussion of Findings

7.1.1. Data-Driven Methods Outperform LLMs on Numerical Causality

Our experiments demonstrate that data-driven causal discovery methods outperform LLMs on regime-dependent numerical datasets. ROCHE and LCUBE achieve 80% mean accuracy compared to GPT-5.2's 30%. All three algorithms achieve 100% accuracy on the synthetic validation dataset, confirming that regime-dependent mechanisms are detectable when present. In contrast, GPT-5.2 never achieves perfect accuracy on any dataset (0% perfect rate) and fails completely on 40% of datasets.

This performance gap indicates that despite their strong performance in text-based reasoning, current LLMs lack the numerical pattern recognition abilities needed for regime-dependent causal discovery. Data-driven methods remain more appropriate for causal discovery in numerical tabular datasets, particularly when regime-specific structure must be extracted from data alone.

7.1.2. LLM Reliance on Semantic Knowledge

Our results show that LLMs appear to rely on semantic knowledge about variable relationships rather than statistical pattern recognition in the data itself. Named variable configurations consistently outperform anonymous configurations across most datasets. On Emissions-Renewables, named variables achieve 100% accuracy while anonymous variables reach only 50%.

This dependency on semantic cues has important implications:

- LLMs draw on training data associations rather than analyzing presented numerical evidence
- Performance depends on whether relevant domain knowledge was encoded during training
- Anonymous numerical data—the standard input for causal discovery algorithms—is particularly challenging for LLMs

7.1.3. The Inverted Failure Problem

Our analysis reveals a notable failure mode: on economic datasets, GPT-5.2 correctly detects that causal direction reverses between regimes but predicts the directions exactly backwards. This "inverted" failure pattern produces 0% accuracy—worse than random guessing would achieve on a binary choice.

This finding suggests that LLM training data may encode plausible but incorrect causal relationships for complex domains like monetary policy. The model applies domain reasoning that produces coherent explanations but arrives at exactly the wrong conclusions. This is notable because the reasoning appears confident and well-justified, making errors difficult to detect without ground truth comparison.

7.1.4. Regime-Switching as a Hard Problem for LLMs

GPT-5.2 achieves 0% correct regime-switching detection rate across all five datasets. This is notable because:

- The model receives explicit threshold information in every configuration
- Pre-segmented regime data is provided in half of the configurations
- The four-step prompting chain explicitly asks about direction reversal

Despite these accommodations, the model either defaults to same-direction predictions (ignoring regime structure entirely) or produces inverted predictions (applying incorrect domain knowledge). This suggests that regime-dependent causal reasoning requires capabilities that current LLM architectures do not possess.

7.1.5. Value of Structured Prompting

Our four-step prompting architecture was designed to prevent anchoring bias and encourage regime-specific reasoning. While we did not conduct ablation studies comparing against simpler prompting approaches, the structured chain does enable:

- Explicit hypothesis generation for both causal directions
- Clear regime characterization before causal analysis
- Structured JSON output for automated evaluation

However, even with this optimized prompting strategy, LLM performance remains substantially below algorithmic baselines. This suggests that prompting improvements alone cannot overcome fundamental limitations in numerical causal reasoning.

7.2. Answering the Research Questions

7.2.1. RQ1: Algorithmic Performance Comparison

Research Question: How do ROCHE, LOCI, and LCUBE perform on regime-dependent bivariate pairs?

Answer: ROCHE and LCUBE achieve the best performance with 80% mean accuracy and 60% correct regime-switching detection. Both methods achieve perfect accuracy (100%) on synthetic validation, engineering (Auto MPG), and energy (Emissions-Renewables) datasets. LOCI shows lower performance (50% mean accuracy, 20% correct detection) with sensitivity to distributional assumptions—it fails completely on Emissions-Renewables (0%) and produces an inverted detection. All algorithms struggle on economic datasets (UK Interest Rates, US Inflation-Fed Rate), achieving only 50% accuracy with no correct regime-switching detection.

7.2.2. RQ2: LLM Causal Reasoning Capabilities

Research Question: Can LLMs reason about regime-dependent causal mechanisms?

Answer: Current LLMs show limited capability for regime-dependent causal inference. GPT-5.2 achieves only 30% mean accuracy with 0% correct regime-switching detection. When the model detects a reversal (economic datasets), it predicts directions backwards (inverted failures). Performance depends heavily on semantic variable names—with named variables, GPT-5.2 achieves 100% on two datasets (Auto MPG, Emissions), but with anonymous variables it never exceeds 50%. The model appears to rely on domain knowledge from training rather than analyzing numerical patterns in the presented data.

7.2.3. RQ3: Configuration Effects

Research Question: How do naming and segmentation configurations affect LLM performance?

Answer: Variable naming has substantial impact: named configurations outperform anonymous on most datasets. On Emissions-Renewables, named variables achieve 100% while anonymous reach only 50%. Pre-segmented regime data provides modest improvement over raw data, but the effect is inconsistent across datasets. The interaction between naming and segmentation is complex—on Auto MPG, raw named achieves 100% while segmented named reaches only 50%. Overall, no configuration enables LLMs to match algorithmic performance, and even the best configuration (`segmented_named`) achieves at most 50% on economic datasets.

7.3. Contributions

This thesis makes the following contributions:

1. **Regime-dependent causal evaluation:** We provide the first systematic evaluation of LLMs on regime-dependent causal discovery from numerical bivariate

data. The evaluation comprises real-world bivariate pairs from monetary policy, energy systems, and automotive engineering, with ground truth causal directions established using domain thresholds from peer-reviewed research. We also propose a regime-level accuracy metric for evaluating inferred causal directions across changing regimes.

2. **Systematic comparison:** We directly compare LLM causal reasoning (GPT-5.2) against established data-driven methods (ROCHE, LOCI, LCUBE) using identical regime definitions and accuracy metrics, enabling fair assessment of relative capabilities.

3. **Evaluation framework and failure mode analysis:** We designed a four-step prompting architecture that prevents anchoring bias and enables incremental causal reasoning. The 2×2 configuration design (segmented/raw × named/anonymous) systematically isolates effects of data representation and semantic information. Through this framework, we identify and characterize two distinct LLM failure modes: same-direction failures (ignoring regime structure) and inverted failures (detecting reversals but predicting directions backwards).

4. **Open-source infrastructure:** We release a reusable Python tool implementing the complete evaluation pipeline, supporting multiple LLM backends and enabling reproducible evaluation on regime-dependent numerical data. The application architecture and development process are documented in Appendix A.

7.4. Limitations

7.4.1. Dataset Scope

We evaluate on five bivariate pairs, limiting generalization to multivariate settings or larger causal graphs. While our datasets span multiple domains (synthetic, engineering, energy, economics), they represent a small sample of possible regime-dependent relationships. The binary regime assumption (exactly two regimes per dataset) may not capture more complex regime structures in real-world data.

7.4.2. Model Selection

Only GPT-5.2 is evaluated; results may differ for other LLMs. Claude, Gemini, and open-source models like LLaMA may exhibit different failure patterns or capabilities. Our tool architecture supports multiple backends, enabling future evaluation across model families. However, architectural differences make direct comparison challenging, and we focused on controlled single-model evaluation.

7.4.3. Ground Truth Challenges

Literature-based ground truth relies on interpretation of domain research. While we selected datasets with well-documented causal mechanisms, some relationships (particularly

in economics) may be contested in the literature. Threshold values represent midpoints of documented transition regions, introducing some arbitrariness. Different threshold choices could affect accuracy calculations.

7.4.4. Prompt Sensitivity

LLM results may be sensitive to specific prompt wording. While we refined prompts iteratively to address identified issues (anchoring, direction confusion), we did not exhaustively explore the prompting design space. Different prompting strategies (chain-of-thought, few-shot examples) might yield different results. Our findings apply to the specific four-step architecture used.

7.4.5. API Constraints

GPT-5.2 was accessed via OpenAI's Responses API without temperature control (not supported for this model tier). This limits our ability to assess output stability across multiple runs or to use temperature=0 for fully deterministic outputs. Results represent single-run evaluations per configuration.

7.5. Future Work

7.5.1. Expanded Model Evaluation

Future work should evaluate additional LLM families (Claude, Gemini, LLaMA) to determine if limitations are model-specific or general across architectures. Open-source models would enable fine-tuning experiments to assess whether causal reasoning can be improved through targeted training.

7.5.2. Multivariate Extension

Extending beyond bivariate pairs to multivariate graphs would test scalability. Regime-dependent DAG discovery presents additional challenges: identifying which edges exhibit regime-switching, handling multiple simultaneous regime definitions, and scaling prompting strategies to larger variable sets.

7.5.3. Hybrid Approaches

The complementary failure patterns suggest potential for hybrid approaches combining LLM semantic knowledge with algorithmic statistical inference. Possible directions include:

- Using LLMs to generate prior distributions over possible causal directions, then applying algorithms for data-driven refinement
- Using algorithms for regime-specific inference, then LLMs for mechanism interpretation

- Ensemble methods that combine algorithmic and LLM predictions with learned weighting

7.5.4. Improved Prompting Strategies

More sophisticated prompting approaches may improve LLM performance:

- Chain-of-thought prompting with explicit numerical reasoning steps
- Few-shot examples showing correct regime-dependent analysis
- Fine-tuning on synthetic regime-switching datasets with known ground truth
- Retrieval-augmented generation incorporating relevant domain literature

7.5.5. Richer Datasets

Expanding the dataset collection would improve generalizability:

- Additional domains beyond current coverage
- Time series data with temporal regime changes
- Datasets with more than two regimes
- Interventional data where available for stronger ground truth

7.6. Conclusion

This thesis provides the first systematic comparison of data-driven causal discovery algorithms and Large Language Models on regime-dependent numerical datasets. Our evaluation reveals a substantial performance gap: established algorithms (ROCHE, LCUBE) achieve 80% mean accuracy with 60% correct regime-switching detection, while GPT-5.2 achieves only 30% accuracy with 0% correct detection.

The failure mode analysis uncovers two distinct patterns. On non-economic datasets with anonymous variables, GPT-5.2 ignores regime structure entirely, defaulting to same-direction predictions. On economic datasets, it correctly detects that reversals occur but predicts directions backwards—an "inverted" failure suggesting problematic domain knowledge in LLM training. Both failure modes result from the model's reliance on semantic associations rather than statistical pattern analysis, as demonstrated by the dramatic performance improvement with named variables.

These findings establish that current LLMs are not suitable for regime-dependent causal discovery from numerical data. Data-driven methods remain the appropriate choice when regime-specific causal structure must be extracted from observations alone. However, the complementary failure patterns—algorithms struggle on economic data while LLMs struggle on clear statistical patterns—suggest potential value in hybrid approaches combining semantic knowledge with algorithmic inference.

The Python tool developed for this research enables reproducible evaluation of causal discovery methods on regime-dependent datasets. By releasing this infrastructure, we hope to facilitate future research comparing LLM capabilities against statistical baselines as both fields continue to evolve. The methodology introduced here—literature-based thresholding, four-step prompting, and 2×2 configuration design—provides a framework for systematic evaluation that can be extended to additional models, datasets, and causal discovery tasks.

Bibliography

[AO16] Aslan Alper and Ocal Oguz. The role of renewable energy consumption in economic growth: Evidence from asymmetric causality. *Renewable and Sustainable Energy Reviews*, 60:953–959, 2016. Publisher: Elsevier.

[BAD+25] Abdolmahdi Bagheri, Matin Alinejad, Mahdi Dehshiri, Kevin Bello, and Alireza Akhondi-Asl. C^2P: Featuring large language models with causal reasoning, 2025. arXiv preprint arXiv:2407.18069.

[CVD+24] Kai-Hendrik Cohrs, Gherardo Varando, Emiliano Diaz, Vasileios Sitokonstantinou, and Gustau Camps-Valls. Large language models for constrained-based causal discovery, 2024. arXiv preprint arXiv:2406.07378.

[DDS+15] Johanna M. Doerr, Beate Ditzen, Jana Strahler, Alexandra Linnemann, Jannis Ziemek, Nadine Skoluda, Christiane A. Hoppmann, and Urs M. Nater. Reciprocal relationship between acute stress and acute fatigue in everyday life in a sample of university students. *Biological Psychology*, 110:42–49, 2015. DOI: 10.1016/j.biopsycho.2015.06.009.

[DLA+25] Juncheng Dong, Yiling Liu, Ahmed Aloui, Vahid Tarokh, and David Carlson. Care: Turning llms into causal reasoning expert, 2025. arXiv preprint arXiv:2511.16016.

[EM96] Arturo Estrella and Frederic S Mishkin. The yield curve as a predictor of recessions in the united states and europe. *The Determination of Long-Term Interest Rates and Exchange Rates and the Role of Expectations. Basel: Bank for International Settlements*, 1996.

[FPC+23] Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. Mathematical capabilities of chatgpt, 2023. arXiv preprint arXiv:2301.13867.

[GHH+18] David Greene, Anushah Hossain, Julia Hofmann, Gloria Helfand, and Robert Beach. Consumer willingness to pay for vehicle attributes: What do we know? *Transportation Research Part A: Policy and Practice*, 118:258–279, 2018. Publisher: Elsevier.

[GMP25] Valeria Gargiulo, Christian Matthes, and Katerina Petrova. Monetary policy across inflation regimes. *European Economic Review*, 178:105109, 2025. DOI: 10.1016/j.euroecorev.2025.105109.

[HK24] Elizabeth Healey and Isaac S. Kohane. Can large language models reason about causal relationships in multimodal time series data? In *Causality and Large Models @NeurIPS 2024*, 2024. OpenReview: https://openreview.net/forum?id=Wmg7IZl32b.

[HSM25] Katerina Hlavackova-Schindler and Suzana Marsela. Identifying causal direction via dense functional classes, 2025. arXiv preprint arXiv:2509.00538.

[ISV+23] Alexander Immer, Christoph Schultheiss, Julia E. Vogt, Bernhard Schölkopf, Peter Bühlmann, and Alexander Marx. On the identifiability and estimation of causal location-scale noise models, 2023. arXiv preprint arXiv:2210.09054.

[KHvdP+24] Shiv Kampani, David Hidary, Constantijn van der Poel, Martin Ganahl, and Brenda Miao. LLM-initialized differentiable causal discovery. In *Causality and Large Models @NeurIPS 2024*, 2024. OpenReview: https://openreview.net/forum?id=2gENrLD5nn.

[KN21] Aditi Kathpalia and Nithin Nagaraj. Measuring causality: The science of cause and effect. *Resonance*, 26(2):191–210, 2021. Publisher: Springer.

[KNP04] Chang-Jin Kim, Charles R Nelson, and Jeremy Piger. The less-volatile us economy: a bayesian investigation of timing, breadth, and potential explanations. *Journal of Business & Economic Statistics*, 22(1):80–93, 2004. Publisher: Taylor & Francis.

[MAS+25] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2025. arXiv preprint arXiv:2410.05229.

[MPJ+16] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

[Qui93] R. Quinlan. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

[RNK+19] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019. Publisher: American Association for the Advancement of Science.

[RRWT24] Atul Rawal, Adrienne Raglin, Qianlong Wang, and Ziying Tang. Investigating causal reasoning in large language models. In *Causality and Large Models @NeurIPS 2024*, 2024. OpenReview: https://openreview.net/forum?id=EGWrfirmIM.

[SBV+24] Saurabh Srivastava, Annarose M B, Anto P V, Shashank Menon, Ajay Sukumar, Adwaith Samod T, Alan Philipose, Stevin Prince, and Sooraj Thomas. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap, 2024. arXiv preprint arXiv:2402.19450.

[SG91] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991. Publisher: Sage Publications Sage CA: Thousand Oaks, CA.

[SGS00] Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT Press, 2000.

[SS24] Sanyam Saxena and Sunita Sarawagi. Reasoning with a few good cross-questions greatly enhances causal event attribution in LLMs. In *Causality and Large Models @NeurIPS 2024*, 2024. OpenReview: https://openreview.net/forum?id=uEOUvr2xVo.

[TDNN23] Quang-Duy Tran, Bao Duong, Phuoc Nguyen, and Thin Nguyen. Robust estimation of causal heteroscedastic noise models, 2023. arXiv preprint arXiv:2312.10102.

[WAN24] ZEYU WANG. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Causality and Large Models @NeurIPS 2024*, 2024. OpenReview: https://openreview.net/forum?id=kbmGbm2L1P.

[WTB+25] Moritz Willig, Tim Nelson Tobiasch, Florian Peter Busch, Jonas Seng, Devendra Singh Dhami, and Kristian Kersting. Systems with switching causal relations: A meta-causal perspective, 2025. arXiv preprint arXiv:2410.13054.

A. Development of the Comparison Framework

This appendix documents the iterative development and final architecture of the `causal_inference_app` Python framework used throughout this thesis. The tool orchestrates both data-driven causal discovery algorithms (ROCHE, LOCI, LCUBE) and LLM-based inference (GPT-5.2) on regime-dependent bivariate datasets, providing a unified evaluation pipeline.

A.1. Final Application Architecture

The framework is a Dockerized Streamlit application. Figure A.1 shows the final directory layout, and Table A.1 maps key modules to their responsibilities.

Application Directory Structure

```
causal_inference_app/
+-- Dockerfile, docker-compose.yml    # Containerized deployment
+-- requirements.txt                  # Python dependencies
+-- app/
|   +-- main.py                       # Streamlit entry point
|   +-- components/
|   |   +-- dataset_panel.py          # Dataset selection UI
|   |   +-- segmentation_panel.py     # Threshold-based segmentation
|   |   +-- causal_panel.py           # Algorithm execution
|   |   +-- comparison_panel.py       # Ground truth evaluation
|   |   +-- batch_panel.py            # Batch experiment runner
|   |   +-- llm_panel.py             # LLM interaction UI
|   +-- utils/
|       +-- llm_config.py            # 12 preset configurations
|       +-- prompts.py               # 4-step prompt templates
|       +-- data_summarizer.py       # Sampling and anonymization
|       +-- ground_truth.py          # Ground truth lookup
|       +-- llm_client.py            # OpenAI/Claude/Gemini
+-- helpers/
|   +-- causal_param_optimization.py  # Dynamic algorithm loading
|   +-- visualization_segmentation.py # Regime visualization
+-- DATA/
|   +-- pairmeta_with_ground_truth.txt # Ground truth definitions
+-- ROCHE/causa/                       # ROCHE algorithm
+-- LOCI/causa/                        # LOCI algorithm
+-- LCUBE/causa/                       # LCUBE algorithm
```

Figure A.1.: Directory structure of the `causal_inference_app` framework.

Table A.1.: Key modules and their roles in the evaluation pipeline.

Module	Responsibility
`batch_panel.py`	Runs all methods across datasets; exports CSV/ZIP
`segmentation_panel.py`	Threshold-based regime splitting from metadata
`causal_panel.py`	Executes ROCHE, LOCI, LCUBE per regime
`comparison_panel.py`	Compares predictions against ground truth
`llm_config.py`	Defines 12 configuration presets
`prompts.py`	Four-step prompt chain with variable substitution
`data_summarizer.py`	Data sampling, anonymization, formatting for LLMs
`causal_param_optimization.py`	Dynamic `importlib` loading of algorithm modules

Algorithms are loaded dynamically at runtime via Python's `importlib`, allowing each algorithm's `causa/` submodule to be imported independently without modifying the host application. This design enables adding new algorithms by simply placing their module in the project root.

A.2. Key Design Decisions

The framework evolved through several critical architectural changes during development.

A.2.1. From Clustering to Literature-Based Thresholding

The most significant refactoring (January 12–13, 2026) replaced data-driven clustering with literature-based threshold segmentation for defining data regimes.

Before vs. After: Regime Assignment Logic

```
Before (data-driven clustering):
regimes = KMeans(n_clusters=2).fit_predict(data)
# Problem:  cluster IDs vary across runs; regime
# boundaries are data-dependent, not interpretable

After (literature-based thresholding):
threshold = metadata[‘‘threshold_val’’]   # e.g., 140.0 HP
regime = 0 if value <= threshold else 1
# Segment 0 = Low regime, Segment 1 = High regime
# Deterministic, interpretable, from domain literature
```

This change ensured: (1) regime definitions are interpretable and grounded in domain literature, (2) all methods are evaluated on identical regime splits, and (3) results are fully reproducible without clustering variability.

A critical normalization bug was also fixed during this refactoring. Previously, each regime subset was standardized independently after segmentation, causing inconsistencies between notebook experiments and the Streamlit application. The corrected pipeline fits `StandardScaler` on the *full dataset* before extracting regime subsets, ensuring that both

regimes share the same scaling parameters. Regime membership is still determined on the original (unscaled) values using literature-based thresholds.

A.2.2. Prompt Chain Optimization

The LLM evaluation protocol was refined from an initial five-step prompting chain to a streamlined four-step architecture (January 13, 2026). The redundant step—which repeated the dataset summary—was removed, reducing token usage by approximately 40% without affecting reasoning depth. The final four-step architecture is described in Chapter 5, Section 5.3.

An additional prompt refinement explicitly distinguished *direction reversal* (X→Y becomes Y→X) from mere changes in relationship *strength*. Early experiments revealed that GPT-5.2 would frequently report "the causal relationship weakens" rather than identifying an actual direction switch—the core phenomenon under study.

A.2.3. Configuration Preset System

A systematic preset system was developed to ensure every experimental combination was tested consistently. Two dimensions—data representation and variable naming—yield the four primary configurations used in the final experiments:

Table A.2.: Experimental configuration matrix. Each cell is an independently evaluated preset.

	Named Variables	**Anonymous Variables**
Raw Data	`raw_named`	`raw_anon`
Pre-segmented Data	`segmented_named`	`segmented_anon`

All configurations include an explicit threshold hint (suffix `_hint_explicit`), as preliminary experiments showed that omitting the threshold made the task intractable for the LLM. The `segmented_anon_hint_explicit` configuration serves as the baseline for the fairest algorithm comparison, since it mirrors algorithm operating conditions: anonymized variables and pre-segmented regime data.

A.3. Batch Experiment Pipeline

The batch experiment runner (`batch_panel.py`) automates evaluation across all datasets and configurations with a single button press. A "Run All LLM Configurations" mode iterates over all four presets, producing structured results and prompt log exports for reproducibility.

The final JSON output produced by the LLM in Step 4 follows a structured format that is automatically parsed for evaluation:

Prompt Log Export Structure

```
prompt_logs_export/
+-- raw_named_hint_explicit/
|   +-- synthetic_stress_fatigue/
|   |   +-- Step_1_Domain_Understanding.txt
|   |   +-- Step_2_Regime_Identification.txt
|   |   +-- Step_3_Causal_Analysis.txt
|   |   +-- Step_4_JSON_Output.txt
|   +-- auto_mpg_horsepower/
|   |   +-- Step_1 ... Step_4
|   +-- ...
+-- segmented_anon_hint_explicit/
|   +-- ...
+-- (4 config folders x 5 dataset folders)
```

Figure A.2.: Prompt log export structure. Each step's full prompt and LLM response is preserved for auditability.

LLM JSON Output Format (Step 4)

```
{
  ''threshold_variable'': ''Variable_Y'',
  ''threshold_value'': 0.5,
  ''regimes'': 2,
  ''directions'': [
    {''regime'': 0,
     ''direction'': ''Y→X'',
     ''confidence'': ''High''},
    {''regime'': 1,
     ''direction'': ''X→Y'',
     ''confidence'': ''Moderate''}
  ]
}
```

Figure A.3.: Structured JSON output parsed from the LLM's Step 4 response. Per-regime directions are compared against ground truth to compute accuracy.

Results are exported as a CSV file (`batch_results_YYYYMMDD.csv`) containing per-dataset, per-configuration accuracy scores for all methods—the authoritative data source for all tables and figures presented in Chapter 6.

A.4. Development Timeline

Table A.3 summarizes the chronological development milestones.

Table A.3.: Development timeline of the `causal_inference_app` framework.

Date	**Milestone**
September 2025	Repository initialization, thesis context added
October 2025	Data pair preprocessing, header formatting
November 2025	Unified `DataLoader` class for all datasets
Late November 2025	Full Docker application deployment
December 2025	Ground truth evaluation framework and UI improvements
January 6–7, 2026	Major refactor: clustering → threshold-based segmentation
January 13, 2026	LLM prompt optimization (5→4 steps), configuration presets
January 20, 2026	Segmented regime presets, LLM direction mapping fixes
January 30, 2026	Batch experiment runner, prompt log export, final experiments
February 2026	Results verification, thesis-code alignment audit